DESTATIS
wissen. nutzen.

Forum der Bundesstatistik

ANONYMISIERUNG WIRTSCHAFTS-STATISTISCHER EINZELDATEN

Beiträge zum Workshop am 20./21. März 2003 in Tübingen

Gerd Ronning und Roland Gnoss

Band 42

Statistisches Bundesamt

Bibliographische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über http://dnb.ddb.de abrufbar.

Herausgeber: Statistisches Bundesamt, Wiesbaden

Fachliche Informationen
zu dieser Veröffentlichung:

Gruppe I A, Roland Sturm
Tel.: 06 11 / 75 25 80
Fax: 06 11 / 75 39 50
roland.sturm@destatis.de

Allgemeine Informationen
zum Datenangebot:

Informationsservice,
Tel.: 06 11 / 75 24 05
Fax: 06 11 / 75 33 30
info@destatis.de
www.destatis.de

Veröffentlichungskalender
der Pressestelle:
www.destatis.de/presse/deutsch/cal.htm

Erscheinungsweise: einmalig
Erschienen im Oktober 2003
Buch mit CD-ROM: EUR 16,80 [D] zzgl. Versandkosten
Bestellnummer: 1030442- 03900
ISBN: 3-8246-0699-2

Recyclingpapier aus 100 % Altpapier.

Vertriebspartner: SFG – Servicecenter Fachverlage GmbH
Postfach 43 43
72774 Reutlingen
Tel.: 0 70 71 / 93 53 50
Fax: 0 70 71 / 93 53 35
www.s-f-g.com
destatis@s-f-g.com

Vorwort

Die Verbesserung des Zugangs der Wissenschaft zu Mikrodatenfiles der amtlichen Statistik ist für die statistischen Ämter eine wichtige Aufgabe. In den Jahren 1988 bis 1990 sind Verfahren zur faktischen Anonymisierung von Mikrodatenfiles entwickelt worden. Sie bilden die Basis für die in den statistischen Ämtern heute praktizierten Techniken bei haushalts- und personenbezogenen Statistiken. Mittlerweile ist die Weitergabe von faktisch anonymisierten Einzeldaten an die Wissenschaft – soweit Personen- und Haushaltserhebungen betroffen sind – geübte Praxis.

Die Wissenschaft, insbesondere die empirische Wirtschafts- und Sozialforschung, benötigt aber auch Einzeldatensätze aus dem Bereich der amtlichen Unternehmens- und Betriebserhebungen. Mit einer schlichten Übertragung der Anonymisierungsroutinen von Haushalts- und Personendaten auf Unternehmen und Betriebe ist es nicht getan. Die Probleme, die sich zur Wahrung des Statistikgeheimnisses bei der Nutzung von Mikrodaten aus Unternehmenserhebungen stellen, sind in vielem anders als bei Mikrodaten aus Personen- und Haushaltserhebungen. Deshalb müssen Methoden zur Anonymisierung gefunden werden, die der speziellen Situation bei Unternehmensdaten Rechnung tragen. Dies ist das Ziel eines gemeinschaftlichen Forschungsprojekt der statistischen Ämter und des Instituts für Angewandte Wirtschaftsforschung, Tübingen. Vom Bundesministerium für Bildung und Forschung wird diese Untersuchung finanziell gefördert. Die Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik hat in ihrem Abschlussbericht vom März 2001 ausdrücklich ein solches Projekt empfohlen.

Im vorliegenden Band der Schriftenreihe „Forum der Bundesstatistik" sind sämtliche Referate und Koreferate eines Workshops zusammengetragen, der im Rahmen dieses Forschungsprojektes am 20. und 21. März 2003 an der Universität Tübingen ausgerichtet wurde.

Der Herausgeber bedankt sich bei allen Teilnehmern des Workshops, der dem Projekt wichtige Impulse für die weitere Forschungsarbeit gegeben hat.

Der Dank gilt auch den Organisatoren vom Institut für angewandte Wirtschaftsforschung und dem Statistischen Bundesamt, die sehr dazu beigetragen haben, dass der Workshop die in ihn gesetzten Erwartungen erfüllen konnte.

Tübingen/Wiesbaden, im August 2003

Prof. Dr. Gerd Ronning
Institut für Angewandte Wirtschaftsforschung

Dr. Roland Gnoss
Statistisches Bundesamt

Inhalt

Seite

Beilage

CD „Public Use File“ Mikrodaten der Kostenstrukturerhebungen im Bergbau und Verarbeitenden Gewerbe (KSE) 1999 für kleine und mittlere Unternehmen (KMU) – siehe 3. Umschlagseite des Bandes –

Roland Gnoss *)

Das Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" – Ziele, Ablauf, Beteiligte

Die faktische Anonymisierung wirtschaftsstatistischer Daten ist in den letzten Jahren stärker in das Interesse der statistischen Ämter und der Wissenschaft gerückt. Nicht zuletzt durch die Arbeiten der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (KVI) wurde der Weg für das Forschungsprojekt geebnet, mit dessen ersten Forschungsergebnissen sich dieser Workshop näher befassen will. Zur Einstimmung in die folgenden Beiträge möchte ich auf die Hintergründe unseres Projektes eingehen, etwas zur allgemeinen Problematik der Anonymisierung von Unternehmensdaten sagen und Ihnen die Struktur des Projektes in wenigen Zügen darstellen.

1 Faktisch anonymisierte Einzeldaten als privilegierter Datenzugang

Die Verbesserung des Zugangs der Wissenschaft zu Mikrodatenfiles der amtlichen Statistik ist für die statistischen Ämter eine wichtige Aufgabe. Ein solcher Zugang kann nur im Rahmen der geltenden Rechtsvorschriften erfolgen. In der Folge des Bundesverfassungsgerichtsurteils zur Volkszählung 1983 (Stichwort „Recht auf informationelle Selbstbestimmung") wurde im Gesetz über die Statistik für Bundeszwecke (BStatG) vom 22. Januar 1987 der Wissenschaft eine privilegierte Form der Nutzung von statistischen Mikrodaten eingeräumt. Das bedeutet, Forschungseinrichtungen können Einzelangaben erhalten, *wenn sie nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können* (vgl. § 16 (6) BStatG).[1] Hierfür prägte sich in der Bundesstatistik der Begriff der **faktischen Anonymität** von Einzeldaten ein.

Es galt, diesen Begriff der faktischen Anonymität mit Inhalt zu füllen, also Kriterien für den *unverhältnismäßig großen Aufwand einer Deanonymisierung* zu standardisieren. In den Jahren 1988 bis 1990 sind unter der Leitung von Professor Dr. Walter Müller, Universität Mannheim, mit finanzieller Unterstützung des Bundesforschungsministeriums Verfahren zur faktischen Anonymisierung von Mikrodatenfiles für personenbezogene Daten entwickelt worden.[2] Sie bilden die Basis für die in den statistischen Ämtern noch heute praktizierten Techniken bei haushalts- und personenbezogenen Statistiken.

Die Ergebnisse dieses Anonymisierungsprojektes führten dazu, dass zunächst mit den Daten des Mikrozensus, später auch mit der Einkommens- und Verbrauchsstichprobe faktisch anonymisierte Mikrodatenfiles erstellt wurden. Dies war eine große Verbesse

---

*) Dr. Roland Gnoss, Statistisches Bundesamt, Wiesbaden.

1) Vor Inkrafttreten des neuen Bundesstatistikgesetzes konnten der Wissenschaft nur absolut anonymisierte Daten übermittelt werden. Die statistischen Ämter konnten daher der Wissenschaft statistische Daten nur in Einzelfällen nach mühsamen und arbeitsaufwendigen Verfahren der Anonymisierung zur Verfügung stellen.

2) Vgl. W. Müller et al., Die faktische Anonymität von Mikrodaten, Band 19 der Schriftenreihe Forum der Bundesstatistik, Wiesbaden 1991.

rung der Datennutzung für die Wissenschaft. Gleichwohl: Die Produktionszeiten waren lang und die Kosten hoch, so dass nur ein kleiner Teil der Wissenschaft die finanziellen Möglichkeiten hatte, das vorhandene Angebot zu nutzen.

Ein Durchbruch gelang durch eine neuerliche finanzielle Unterstützung des Bundesforschungsministeriums. Dadurch wurde es möglich, den Mikrozensus, die Einkommens- und Verbrauchsstichprobe, das Europäische Haushaltspanel und die Zeitbudgeterhebung für verschiedene Jahre als faktisch anonymisiertes Mikrodatenfile der Wissenschaft zum Preis von jeweils 65,– € zur Verfügung zu stellen.

Heute ist die Weitergabe von faktisch anonymisierten Einzeldaten an die Wissenschaft – soweit Personen- und Haushaltserhebungen betroffen sind – geübte Praxis. Daneben werden weitere Möglichkeiten des Datenzugangs in den Forschungsdatenzentren der statistischen Ämter angeboten.

2 Anonymisierung von Unternehmensdaten

In der Vergangenheit wurden nur Datensätzen aus Haushalts- und Personenerhebungen betrachtet. Die Wissenschaft verlangt heute auch Einzeldatensätze aus dem Bereich der Unternehmens- und Betriebserhebungen. Methoden zu finden, solche Einzeldatensätze zu anonymisieren ist Aufgabe des Projektes, in dessen Rahmen wir auf diesem Workshop austauschen wollen.

Die Aufgabe, die wir uns hier vorgenommen haben, ist nicht trivial. Mit einer schlichten Übertragung der Anonymisierungsroutinen von Haushalts- und Personendaten auf Unternehmen und Betriebe ist es nicht getan. Die Sensibilität von Mikrodaten aus Unternehmenserhebungen ist wesentlich höher einzustufen als bei Personen- und Haushaltserhebungen. Die bekannteste amtliche Bevölkerungsstichprobe, der Mikrozensus, ist mit seinem Auswahlsatz von 1 % relativ klein, allerdings mit ca. 800 000 Befragten eine Mammuterhebung, in der die einzelne Person in der Stichprobe untergeht.

Den Unternehmenserhebungen liegen dagegen wesentlich kleinere Grundgesamtheiten zugrunde, so dass Einzelfälle wesentlich häufiger vorkommen als bei Personenerhebungen. Auch sind die Stichprobenauswahlsätze bei Unternehmenserhebungen wesentlich größer. Bestimmte Teilpopulationen (d.h. Wirtschaftszweige) müssen hier sogar voll erhoben werden. Mikrodaten aus Unternehmenserhebungen verdienen also besondere Aufmerksamkeit.

In den neu eingerichteten Forschungsdatenzentren der statistischen Ämter sollen faktisch anonymisierte Einzeldaten aus Unternehmenserhebungen ein wichtiger Baustein werden. Von der Wissenschaft ist das Vorhaben einhellig begrüßt worden. Die bereits oben erwähnte Kommission (KVI) empfiehlt in ihrem Abschlussgutachten ausdrücklich unser Projekt. Das Bundesministerium für Bildung und Forschung übernimmt einen wesentlichen Teil der Projektfinanzierung, den Rest tragen die beteiligten statistischen Ämter.

3 Struktur und Stand der Projektarbeiten

Folgende Projektpartner beteiligen sich an den Forschungsarbeiten: Von Seiten der amtlichen Statistik: das Statistische **Bundesamt**, das Statistische Landesamt **Berlin**, das Landesamt für Datenverarbeitung und Statistik **Nordrhein-Westfalen**, das Statistische Landesamt **Schleswig-Holstein**[3] und das Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit/Nürnberg (**IAB**). Von Seiten der Wissenschaft das Institut für Angewandte Wirtschaftsforschung (**IAW**), das diesen Workshop gemeinsam mit dem Statistischen Bundesamt ausrichtet.

Aus Sicht der statistischen Ämter muss das Projekt **zwei Ziele** erreichen:

(1) Ein **ausreichender Schutz der Einzelangaben** muss gewährleistet sein und

(2) Die **Analysemöglichkeiten** der anonymisierten Daten müssen weitestgehend **erhalten** bleiben.

Aus Sicht der künftigen Nutzer von Einzeldaten ist vor allem der zweite Aspekt wichtig. Niemandem wäre damit gedient, wirkungsvoll anonymisierte Daten zu erhalten, die keinen wissenschaftlichen Nutzen haben. Die Wissenschaft ist in allen Phasen des Projektes eng eingebunden: Das Institut für Angewandte Wirtschaftsforschung in Tübingen (IAW) ist Projektpartner der statistischen Ämter. Professor Ronning, Inhaber des Lehrstuhls für Statistik und Ökonometrie an der Universität Tübingen und Direktor des IAW, ist der wissenschaftliche Leiter des Projektes. Er hat zu seiner Unterstützung einen Wissenschaftlichen Begleitkreis (WBK) eingerichtet, der sich aus empirisch arbeitenden Mikroökonomen und Sozialwissenschaftlern zusammensetzt.

Die Schwerpunkte der Arbeitsteilung im Projekt ergeben sich wie folgt:

a) die **statistischen Ämter** führen die Anonymisierungen durch und beurteilen deren Schutzwirkung,

b) das **IAW** , unterstützt durch den **WBK**, untersucht und bewertet das Analysepotential der anonymisierten Einzeldaten.

Die im Projekt eingesetzten Erhebungen wurden von den statistischen Ämtern so gewählt, dass sowohl den Interessen der empirischen Forschung als auch den für die Anonymisierung wesentlichen Eigenschaften der Befragungen Rechnung getragen wurde.

Berücksichtigt wurden somit wirtschaftsbereichsübergreifende und Bereichsstatistiken, Statistiken mit großem und mit kleinerem Merkmalskanon, Statistiken zu bereits länger beobachteten und zu neueren Thematiken. Das Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit beteiligt sich mit der Untersuchung seiner Betriebsdaten an dem Projekt.

Es ist wichtig darauf hinzuweisen, dass die Einbeziehung von bestimmten Erhebungen in das Projekt keine Vorentscheidung für ein späteres Angebot der statistischen Ämter an faktisch anonymisierten Unternehmensdaten darstellt.

---

3) An dieser Stelle ist zu erwähnen, dass auch die übrigen statistischen Ämtern ebenfalls Projektpartner sind, da sie dem Projekt ihre Einzeldaten zur Verfügung gestellt und somit eine wichtige Voraussetzung für die Durchführung des Forschungsvorhabens erfüllt haben.

Aufbauend auf den Projektergebnissen gehen die statistischen Ämter davon aus, unter Berücksichtigung der vorhanden Kapazitäten einen breiten Kanon von Erhebungen bedarfsorientiert anzubieten. Die untersuchten Anonymisierungsmethodiken sollen deshalb in **ein Kompendium praxisrelevanter Verfahren** überführt werden, das den statistischen Ämtern Standardverfahren für die Anonymisierung von Unternehmensdaten bietet. Die möglichen Einzelmaßnahmen müssen auf die Eigenschaften der jeweils zu anonymisierenden Erhebung abgestimmt werden. Die zur Verfügung stehenden Anonymisierungsmaßnahmen sind daher im Projekt zu beschreiben, und es ist darzulegen, in welchen Fällen welche Maßnahmen zu ergreifen und in welcher Reihenfolge die einzelnen Schritte sinnvoller Weise durchzuführen sind.

Aufgabe des zweitägigen Workshops wird es sein, einem breiteren Kreis von Nutzern den Projektstand vorzustellen und wichtige Anregungen für den weiteren Projektverlauf aus Nutzersicht zu bekommen. Die Projektarbeiten laufen inzwischen über ein Jahr. Der Workshop markiert sozusagen die Halbzeit.

Der Workshop ist thematisch in **zwei Blöcken** gegliedert. Im ersten Block werden wir unser Hauptaugenmerk auf die **Anonymisierung** lenken, im zweiten wird die **Datennutzung** im Vordergrund stehen.

Frau Dr. Wirth wird uns zunächst einen Überblick über Deanonymisierungsversuche geben, also Szenarien, wie man sich einen unerlaubten Deanonymisierungsversuch vorstellen könnte.

Der Vortrag von Herrn Dr. Wiegert befasst sich mit den technischen Möglichkeiten des Einsatzes von Zusatzwissen bei Deanonymisierungsversuchen.

Herr Dr. Vorgrimler wird uns schließlich etwas über die Verfügbarkeit und Qualität dieses Zusatzwissen berichten.

Das Ko-Referat von Herrn Dr. Blien wird uns im Anschluss daran zur weiteren Diskussion ermuntern.

Herr Höhne wird uns im weiteren Verlauf in seinem Referat einen Überblick über den „state of the art" der Anonymisierungsmethodiken verschaffen. Frau Dr. Brand wird das Ko-Referat hierzu halten.

Der Beitrag von Professor Domingo-Ferrer über „Konzepte von Schutzwirkung und Analysepotential" rundet den ersten Themenblock ab und zeigt, dass wir im Rahmen des Projektes auch mit der internationalen Wissenschaft in Kontakt sind. Das Ko-Referat hierzu hält Frau Gießing.

Die Referate im zweiten Block befassen sich überwiegend mit der Nutzung anonymisierter Daten.

Professor Pohlmeier wird über den Einfluss von Anonymisierungsverfahren auf die Eigenschaften von ökonometrischen Schätzern berichten. Professor Ronning wird uns in einem Ko-Referat seine bisherigen Erkenntnisse zu der gleichen Thematik darlegen.

Herr Professor Wagner wird an einem konkreten Beispiel – nämlich dem Monatsbericht im Verarbeitenden Gewerbe – seine Erfahrungen mit der Arbeit mit nicht-ano-

nymisierten Material darlegen. Warum das für unser Projekt wichtig ist, werden wir dann sicher sehen. Im Hinblick auf die Anonymisierbarkeit von Panelerhebungen könnte der Monatsbericht in unserem Projekt noch besondere Bedeutung bekommen. Herr Dr. Strotmann wird hierzu das Ko-Referat halten und damit noch weitere Aspekte in die Diskussion bringen.

Herr Rosemann wird uns abschließend am Beispiel von zwei weiteren Datensätzen, nämlich der Kostenstrukturerhebung und der Umsatzsteuerstatistik seine Ergebnisse einer vergleichenden Analyse präsentieren. Den Part des Ko-Referenten übernimmt hier Herr Dr. Licht vom ZEW.

Das Tagungsprogramm und die Zusammensetzung der Teilnehmer des Workshops lassen eine interessante Auseinandersetzung mit der Thematik erwarten. Wir sind gespannt, ob diese Erwartungen am Ende auch erfüllt werden.

Der Erfolg des gesamten Projektes hängt davon ab, ob es gelingt, die Interessen der Statistik als auch der Wissenschaft zu erfüllen. Daher ist es den statistischen Ämtern sehr wichtig, die Projektarbeiten im engen Dialog mit der interessierten Wissenschaft durchzuführen. Herr Professor Ronning als wissenschaftlicher Leiter des Projektes, die Projektmitarbeiter des IAW und die Wissenschaftler im Wissenschaftlichen Begleitkreis sind Beleg dafür. Nicht zuletzt soll auch dieser Workshop dabei helfen, die Sicht der Nutzer noch intensiver in unsere Arbeiten einfließen zu lassen. Wir benötigen das wissenschaftliche Urteil der Nutzer und ihre Einschätzung unserer Forschungsergebnisse. In diesem Sinne wünsche ich allen Teilnehmern eine erfolgreiche Veranstaltung und fruchtbare Diskussionen.

Heike Wirth *)

Szenarien für Angriffe auf wirtschaftsstatistische Einzeldaten – Ein Überblick

1 Problemstellung

Daten der amtlichen Statistik stellen seit langem eine wichtige Datenquelle für viele Bereiche und Zwecke der empirischen Wirtschafts- und Sozialforschung dar. Im Zentrum des Interesses stehen hierbei vor allem Mikrodaten[1], da diese eine flexible, den jeweiligen Analyseinteressen angemessene Datenaufbereitung ermöglichen und eine Vielzahl von Fragestellungen nur unter Verwendung von Mikrodaten angemessen untersucht werden kann.[2] In diesem Kontext mutet das Thema des vorliegenden Beitrags „Szenarien für Angriffe auf wirtschaftsstatistische Einzeldaten" befremdlich und fast wie ein Rückfall in die Datenschutzdiskussion der achtziger Jahre an. Dies um so mehr, als zwischen amtlicher Statistik und Wissenschaft ein Konsens dahingehend besteht, dass das Interesse der Forschung nicht auf eine Reidentifikation von amtlichen Mikrodaten gerichtet ist, sondern auf eine ausschließliche Nutzung dieser Daten für wissenschaftliche Zwecke (Sturm 2002, S. 102). Darüber hinaus ist bis zum gegenwärtigen Zeitpunkt – weder national noch international – kein einziger Verstoß von Seiten der Wissenschaft gegen das Reidentifikationsverbot bekannt geworden (Wagner 2003a).

Unabhängig hiervon unterliegen die von der amtlichen Statistik erhobenen Daten dem Statistikgeheimnis, welches auch bei einer Übermittlung von Mikrodaten an die Wissenschaft in Form von § 16 (6) BstatG 1987 zur Anwendung kommt. Danach dürfen die statistischen Ämter Einzelangaben dann an Hochschulen sowie sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung übermitteln, wenn eine Reidentifizierung nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten und Arbeitskraft möglich ist. Man spricht hier von „*faktischer* Anonymität", da die Möglichkeit einer Deanonymisierung im Unterschied zur „*absoluten* Anonymität" nicht mit Sicherheit ausgeschlossen werden muss.[3] Vielmehr ist im Sinne einer Güterabwägung zwischen den Nutzungsbedürfnissen der Forschung einerseits und den berechtigten Interessen der Be-

---

*) Dr. Heike Wirth, Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim.

1) Der Begriff Mikrodaten steht für Informationen über die Elementareinheiten (Personen, Haushalte, Betriebe, Unternehmen, Verbände etc.) einer Erhebung. Andere gebräuchliche Bezeichnungen hierfür sind Einzelangaben oder auch Individualdaten.

2) Zu Nutzungspotenzialen von wirtschaftsstatistischen Mikrodaten vgl. z.B. Wagner (1999, 2003b).

3) Mit der Einführung des Konzepts der faktischen Anonymität wurde die bis dahin geltende gesetzliche Regelung (§11(5) BStatG 1980) ergänzt, welche eine Datenübermittlung nur bei absoluter Anonymität vorsah: „Einzelangaben, die so anonymisiert werden, dass sie Auskunftspflichtigen oder Betroffenen nicht mehr zuzuordnen sind, dürfen vom Statistischen Bundesamt und von den Statistischen Ämtern der Länder übermittelt werden" (Statistisches Bundesamt 1981, S. 404). Diese vom Gesetzgeber ursprünglich mit dem Ziel eingeführte Regelung, „der Wissenschaft und anderen Stellen in gewissem Umfang Daten zur eigenen Aufbereitung unter Wahrung des Datenschutzes zur Verfügung zu stellen" (Deutscher Bundestag 1986, S. 21), hatte sich in der Praxis nicht bewährt. Eine Datenweitergabe war in der Regel mit solch umfangreichen datenmodifizierenden Anonymisierungsmaßnahmen verbunden, dass das wissenschaftliche Analysepotenzial dieser Daten erheblichen Einschränkungen unterlag.

fragten an einer Geheimhaltung ihrer Daten andererseits ein hinreichend hohes Maß an Schutz vor einer Reidentifizierung zu gewährleisten, wobei aber ein Restrisiko hinzunehmen ist.

Grundlage für die Operationalisierung der faktischen Anonymität ist das Kriterium des „unverhältnismäßig hohen Aufwands", d.h. es ist von einem Rationalkalkül auszugehen, bei welchem der für eine Reidentifikation zu erbringende Aufwand ins Verhältnis zu dem erwarteten Nutzen zu setzen ist. Dabei hat der Gesetzgeber nicht präzisiert, wie die faktische Anonymität konkret umzusetzen ist. Deshalb wurde unmittelbar nach der Novellierung des BStatG eine umfassende Untersuchung (Müller et al. 1991) durchgeführt, um eine Regelung für die praktische Umsetzung der faktischen Anonymisierung zu finden. Zu diesem Zweck wurden die im Wissenschaftsbereich vorliegenden Randbedingungen (Zusatzwissen, Reidentifikationsmotivationen und -strategien etc.), die für ein hypothetisches Reidentifikationsvorhaben von Bedeutung sein könnten, eingehend untersucht. Hieraus wurden zentrale „Angriffsszenarien" abgeleitet, für welche eine detaillierte Analyse des Reidentifikationsrisikos und des damit verbundenen Aufwands[4] vorgenommen wurde. Auf Basis dieser Ergebnisse wurden Empfehlungen für die Umsetzung der faktischen Anonymisierung erarbeitet, die spezifischen Risikokonstellationen gezielt entgegenwirken bei einem zugleich möglichst schonenden Eingriff in das Analysepotenzial der Daten.

Der Schwerpunkt dieser Untersuchung lag auf Personen- und Haushaltsdaten. Die allgemeinen Überlegungen bezüglich der im Wissenschaftsbereich vorliegenden Randbedingungen sind jedoch auch auf wirtschaftsstatistische Daten übertragbar. Dabei werden die aus den Projektergebnissen abgeleiteten Anonymisierungsmaßnahmen (Stichprobenziehung, Ausprägungsvergröberung, Unterdrückung differenzierender Regionalmerkmale etc.) hinsichtlich ihrer Schutzwirkung für wirtschaftsstatistische Daten aber als nicht ausreichend erachtet.[5] Dieser Einschätzung zufolge ist für Wirtschaftsdaten aufgrund der anderen Datenstruktur, des umfangreicheren, öffentlich zugänglichen Zusatzwissens und des potenziell größeren Nutzens einer Reidentifikation ein höheres Ge-

4) Bei der theoretischen Analyse wurde deutlich, dass der Nutzen einer Reidentifikation innerhalb des Wissenschaftsbereichs als sehr gering einzuschätzen ist. Allerdings bereitet eine empirische Quantifizierung eines solchen subjektiven Nutzens Schwierigkeiten. Deshalb erfolgte die Bewertung im Wesentlichen durch die Gegenüberstellung des für eine Reidentifikation zu betreibenden Aufwands im Vergleich zu den Kosten einer alternativen Informationsbeschaffung.

5) Bei dieser Argumentation wird möglicherweise übersehen, dass bereits die für Bevölkerungs- und Haushaltsdaten empfohlene Basis-Anonymisierungsregel, nach welcher im Mikrodatenfile univariat keine Merkmalsausprägung ausgewiesen werden darf, die in der Grundgesamtheit nicht mindestens 5 000 Fälle umfasst (Müller et al. 1991), bei einer Anwendung auf wirtschaftsstatistische Daten aufgrund deren erheblich kleineren Grundgesamtheiten den Informationsgehalt dieser Daten so stark reduzieren würde, dass eine Deanonymisierung mit an Sicherheit grenzender Wahrscheinlichkeit auszuschließen wäre. Als unerwünschter Nebeneffekt wären diese Daten jedoch für Analysezwecke gleichfalls weitgehend untauglich.

fährdungspotenzial als für Bevölkerungsdaten zu erwarten (Brand et al. 1999; Gnoss/Sturm 2002), welches die Anwendung von komplexeren, datenmanipulierenden Anonymisierungsverfahren erfordert.[6]

Allerdings lassen sich aus der These des höheren Reidentifikationspotenzials keine Kriterien für eine Abschätzung der Schutzwirkung von Anonymisierungsmaßnahmen bzw. für die Operationalisierung der faktischen Anonymität ableiten. Das in der internationalen Literatur für eine Abschätzung von Reidentifikationsrisiken häufig herangezogene Uniqueness-Konzept[7] erscheint hierfür allenfalls bedingt geeignet. Bei diesem Verfahren wird der Anteil der reidentifikationsgefährdeten Datensätze – vereinfacht dargestellt – anhand der Wahrscheinlichkeit bestimmt, ob ein im Mikrodatenfile in seiner Ausprägungskombination einzigartiger Fall auch in der Population einzigartig ist. Dies verführt leicht dazu, alle einzigartigen Fälle eines Mikrodatenfiles als reidentifikationsgefährdet einzustufen und die Anonymisierungsmaßnahmen hieran zu orientieren. Eine solche Vorgehensweise wird in der Tendenz jedoch eher in Richtung einer absoluten Anonymisierung der Daten wirken als in Richtung faktischer Anonymisierung, mit entsprechend stärkeren Einschränkungen des Analysepotenzials als es nach der Gesetzeslage notwendig ist. Denn Einzigartigkeit in einem Mikrodatenfile ist zwar eine notwendige, aber keinesfalls eine hinreichende Bedingung für eine erfolgreiche Reidentifikation. Wie die empirischen Überprüfungen für Haushalts- und Personendaten gezeigt haben (Müller et al. 1991) – und dieser Befund ist, mit Einschränkungen, sicherlich auch auf wirtschaftsstatistische Einzeldaten übertragbar –, führen ausschließlich am Kriterium der Einzigartigkeit orientierte Abschätzungen des Reidentifikationsrisikos zu einer erheblichen Überschätzung des real bestehenden Risikos.[8] Gleichfalls ist zu berücksichtigen, dass für die faktische Anonymisierung nicht der Anteil potenziell reidentifikationsgefährdeter Einzeldatensätze maßgeblich ist, sondern vielmehr der für die Reidentifikation dieser Fälle notwendige Aufwand im Verhältnis zu dem hieraus zu ziehenden Nutzen.

Folgt man dieser Argumentation, ist es unvermeidlich, die potenziellen Reidentifikationsrisiken von wirtschaftsstatistischen Einzeldaten in Verbindung mit den nutzen- und kostenrelevanten Randbedingungen zu thematisieren und in weiterführenden Analysen einer empirischen Überprüfung zu unterziehen. Aus den hieraus gewonnenen Erkenntnissen lassen sich dann Kriterien für die Operationalisierung der faktischen Anonymisierung dieser Daten erarbeiten.

Vor diesem Hintergrund wird im Folgenden ein Überblick über ausgewählte Angriffsszenarien für wirtschaftsstatistische Einzeldaten gegeben. Hierfür wird in weiten Teilen auf die für Bevölkerungsdaten vorliegenden Überlegungen zur Logik der wissenschaftlichen

6) Unabhängig von dem Argument des höheren Reidentifikationsrisikos ist ein direkter Transfer der für bevölkerungsstatistische Daten empfohlenen Anonymisierungsregeln auf wirtschaftsstatistische Einzeldaten auch deshalb nicht sinnvoll, weil es sich bei ersteren vorwiegend um kategoriale, bei letzteren hingegen vor allem um stetige Merkmale handelt. Anonymisierungsmaßnahmen, die bei Personendaten eine hohe Schutzwirkung entfalten, ohne das Analysepotenzial zu sehr zu beeinträchtigen, wirken sich bei stetigen Merkmalen unter Umständen erheblich ungünstiger auf das Analysepotenzial aus (vgl. Unterarbeitsgruppe Anonymisierungsmethodik 2002a). Als Beispiel für die Anonymisierung von nicht-amtlichen Unternehmensdaten siehe Gottschalk (2002).

7) Für eine ausführliche Darstellung des Uniqueness-Ansatz siehe Brand (2000).

8) In ähnlicher Weise führen auch Reidentifikationsexperimente auf Basis von teilweise oder vollständig synthetisch generierten Daten zu einer Überschätzung der realen Reidentifikationsrisiken (Müller et al. 1991, S. 306 ff.).

Datennutzung zurückgegriffen, da diese unabhängig vom Datentyp sind und die zentralen Randbedingungen für hypothetische Reidentifikationsversuche innerhalb des Wissenschaftsbereichs erfassen.

2 Ausgewählte Angriffsszenarien auf wirtschaftsstatistische Einzeldaten

Wie einführend dargestellt, ist für die Operationalisierung der faktischen Anonymität von einem Rationalkalkül auszugehen, bei welchem das Verhältnis zwischen Kosten und Nutzen einer Reidentifikation zugrunde gelegt wird. Die Quantifizierung des potenziellen Nutzens von Reidentifikationen dürfte – insbesondere wenn nicht-monetäre Elemente zu berücksichtigen sind – auf erhebliche Probleme stoßen. Empirische Kriterien für die faktische Anonymität werden sich daher eher über die Kostenseite finden lassen, d.h. im Vergleich des Aufwands für einen Reidentifikationsversuch zu den Kosten einer alternativen Informationsbeschaffung (Unterarbeitsgruppe 2002b). Dies bedeutet jedoch nicht, dass die Nutzenseite außer Acht gelassen werden kann. Denn inwieweit – abgesehen von dem zu betreibenden Aufwand – überhaupt Reidentifikationsversuche zu erwarten sind, steht in einem unmittelbaren Zusammenhang mit dem potenziellen Nutzen der reidentifizierten Daten. Hierbei kann zwischen zwei grundlegend unterschiedlichen Nutzenkategorien unterschieden werden:

(1) *Wissenschaftsorientierter Nutzen*: Da nach § 16 (6) BStatG anonymisierte Mikrodaten nur an unabhängige Institutionen der Wissenschaft für Forschungszwecke weitergegeben werden dürfen, ist primär zu klären, welche wissenschaftlich begründeten Motive für eine Reidentifizierung von wirtschaftsstatistischen Einzeldaten bestehen könnten.

(2) *Wissenschaftsfremder Nutzen*: In der Literatur ist gelegentlich eine gewisse Ambivalenz diesbezüglich festzustellen, dass einerseits der Forschung keine Reidentifikationsabsichten unterstellt werden, andererseits jedoch auf berufsfremde Motive als mögliche Anreize für Reidentifikationsversuche verwiesen wird. Insgesamt scheint berufsfremden Motiven (Diskreditierung der amtlichen Statistik, ökonomische Anreize, Enthüllungsjournalismus etc.) in der allgemeinen Reidentifikationsdiskussion eine wesentlich größere Bedeutung beigemessen zu werden als wissenschaftsorientierten Motiven. In der internationalen Diskussion (vgl. z.B. Elliot/Dale 1999) ist dies vor allem dadurch bedingt, dass es sich hierbei häufig um die Bereitstellung von *Public-Use-Files* handelt und dementsprechend eine Vielzahl hypothetischer Risikokonstellationen zu berücksichtigen ist. Im Unterschied hierzu handelt es sich bei der Diskussion über die faktische Anonymisierung in Deutschland um die Bereitstellung von *Scientific-Use-Files* auf Basis eines vom Gesetzgeber explizit eingeführten *Wissenschaftsprivilegs*. Deshalb sind außerberufliche Nutzenkomponenten zwar zu thematisieren, aber sie sollten die Diskussion keinesfalls dominieren.

Eine detaillierte Analyse der Logik der wissenschaftlichen Datennutzung – und welche Motive sich hieraus für Deanonymisierungsversuche ergeben könnten – findet sich bei Müller et al. (1991). Dort werden auch wissenschaftsfremde Motivationslagen ausführlich diskutiert. Die dort vorliegenden Überlegungen bilden die Basis für die folgenden

Ausführungen. Zunächst werden kurz beruflich motivierte Nutzenkomponenten erörtert und zwei hieraus abgeleitete Risikoszenarien für wirtschaftsstatistische Einzeldaten diskutiert. Anschließend werden nichtberufliche Nutzenkomponenten betrachtet und ein spezifisches Szenario, welchem bei der Diskussion potenzieller Reidentifikationsrisiken von wirtschaftsstatistischen Daten eine herausgehobene Bedeutung zukommt, einer detaillierteren Analyse unterzogen.

2.1 Beruflich motivierte Nutzenkomponenten von Reidentifikationsversuchen

Wie bei Müller et al. (1991, S. 133 ff.) ausführlich dargestellt, ist das Interesse der quantitativ orientierten Wirtschafts- und Sozialforschung auf die Beschreibung und Erklärung von sozialen, wirtschaftlichen oder politischen Tatbeständen ausgerichtet. Für die empirische Analyse werden dabei vorwiegend Einzelangaben benötigt, da nur diese die volle Flexibilität bei der Datenumformung und -neubildung sowie den zur Anwendung kommenden Methoden bieten. Das Forschungsinteresse richtet sich jedoch nicht auf identifizierbare Einzelfälle, sondern erfordert ganz im Gegenteil eine Abstraktion weg vom Individualfall hin zu allgemeinen Zusammenhängen. Der Einzelfall interessiert hier ausschließlich in seiner Eigenschaft als Merkmalsträger.[9] Sollen beispielsweise der Beitrag wachsender bzw. schrumpfender Betriebe zur Arbeitsplatzentwicklung in einzelnen Regionen, Industrien etc. oder die Unterschiede im Gründungs- und Schließungsgeschehen zwischen Branchen analysiert werden, steht nicht der Einzelbetrieb im Zentrum des Forschungsinteresses. Das Ziel besteht vielmehr in dem Nachweis und der Überprüfung von Zusammenhängen, für welche eine Identifizierbarkeit der Einzelbetriebe weder notwendig noch in irgendeiner Form für das Analyseziel nützlich ist.

Individualfälle sind im Verlauf des Forschungsprozesses nur insofern von Bedeutung, als die Daten bei ihnen erhoben werden. So ist es bis zum Abschluss der Datenerhebungsphase häufig unerlässlich, dass die auskunftsgebenden Einheiten identifizierbar sind, insbesondere um bei unvollständig oder inkonsistent ausgefüllten Fragebögen Rücksprache halten zu können.[10] Ebenso ist für eine Datenverknüpfung auf Individualebene, wie z.B. bei der Zusammenführung von Panelwellen, eine Identifizierbarkeit der Befragungseinheiten notwendig, allerdings nicht notwendigerweise in Form von direkten Identifikatoren (d.h. Name und Anschrift). Stattdessen kann auch ein Schlüssel verwendet werden, der eine Zusammenführung der anonymen Daten auf Einzelfallebene ermöglicht.

Bei der Übermittlung von wirtschaftsstatistischen Einzeldaten an die Wissenschaft liegt jedoch eine völlig andere Situation vor. Hier wird die Datenerhebung nicht vom Forscher oder in seinem Auftrag durchgeführt. Es handelt sich vielmehr um Daten, bei welchen die Erhebungs- und Aufbereitungsphase bei der Weitergabe bereits abgeschlossen ist. Eine Identifizierbarkeit der Einzelfälle ist daher vollkommen unnötig.

9) Siehe hierzu auch die Ausführung des Bundesverfassungsgerichts: „(...) der Wissenschaftler ist regelmäßig nicht an der einzelnen Person interessiert, sondern an dem Individuum als Träger bestimmter Merkmale" (BverfG 65,1 (IV, 5) vom 15.12.1983).

10) In der empirischen Sozialforschung werden die meisten Erhebungen von professionellen Umfrageinstituten durchgeführt. Die Forscher – als Auftraggeber – erhalten die Einzeldatensätze nach Abschluss der Datenerhebung und Datenaufbereitung üblicherweise in bereits anonymisierter Form, d.h. ohne Namen und Anschriften.

Allerdings kann aus der obigen Argumentation, nach welcher die Identität von Einzelfällen während des Forschungsprozesses nur im Kontext der Datenerhebung oder Datenzusammenführung von Relevanz ist, auch die These abgeleitet werden, dass wirtschaftsstatistische Einzeldaten möglicherweise genau für diese Zwecke von einem wissenschaftlichen Nutzen sein könnten (Müller et al. 1991, S. 144 ff.). Im Folgenden soll die Plausibilität dieser These am Beispiel von zwei hypothetischen Angriffsszenarien in den Bereichen 'Datenerhebung' und 'Datenverknüpfung' einer näheren Betrachtung unterzogen werden.

2.1.1 Szenario 1: Reidentifikation wirtschaftsstatistischer Einzeldaten mit dem Ziel, die reidentifizierten Daten für eine eigene Erhebung zu verwenden

Im ersten Szenario wird eine Situation betrachtet, bei welcher Wirtschaftsdaten als Basis für eine eigene Erhebung genutzt werden sollen: Angenommen, ein Forscher interessiert sich für eine spezifische Population von Unternehmen (z.B. Anteil der teilzeitbeschäftigten Personen größer 30 Prozent, Anteil der in Forschung und Entwicklung tätigen Personen größer 15 Prozent) und plant eine Erhebung bei dieser Population durchzuführen. Es erweist sich jedoch als schwierig, eine geeignete Auswahlgrundlage für die Stichprobenziehung zu finden, da die relevanten Merkmale in keinem Unternehmensregister erfasst sind. In diesem Fall wäre es vorstellbar, dass ein zur Verfügung stehendes amtliches Mikrodatenfile für die Stichprobenziehung nützlich sein könnte. D.h., dass der Versuch unternommen wird, wirtschaftsstatistische Einzeldaten zu reidentifizieren, um die hierdurch gewonnenen Informationen für die eigene Erhebung zu nutzen.

Obgleich dieses Szenario zunächst als ein durchaus plausibles Motiv für einen Reidentifikationsversuch erscheint, ist das Eintreten eines solchen Falls wenig realistisch, wie die folgenden Überlegungen zeigen.

Damit überhaupt ein Reidentifikationsversuch erfolgen kann, müssen zunächst folgende Bedingungen gegeben sein:

(1) Die für die Stichprobenziehung relevanten Merkmale müssen in den amtlichen Mikrodaten enthalten sein.[11] Dies kann nicht als selbstverständlich unterstellt werden.

(2) Es muss Zusatzwissen mit einer Reihe von Überschneidungsmerkmalen[12] zum Mikrodatenfile für eine große Anzahl von Fällen zur Verfügung stehen, die ebenfalls im Mikrodatenfile enthalten sind.

(3) Es muss ein leistungsfähiges Deanonymisierungsverfahren verfügbar sein, welches eine massenhafte Reidentifikation ermöglicht.

Sind diese Randbedingungen gegeben, kann mittels einer einfachen Überschlagsrechnung eine grobe Kosten-Nutzen-Abschätzung dieser Risikokonstellation vorgenommen werden. Zu diesem Zweck wird unterstellt: (A) Dem Forscher steht ein sehr effizien-

11) Sofern das Mikrodatenfile nicht nur die interessierenden Stichprobenmerkmale enthält, sondern auch die im spezifischen Forschungskontext eigentlich interessierenden Informationen, können die Daten direkt für die Bearbeitung der Forschungsfrage genutzt werden und es besteht keinerlei Notwendigkeit für Reidentifikationsversuche.

12) Die Überschneidungsmerkmale sind nicht deckungsgleich mit den für die Stichprobeziehung relevanten Merkmalen. Wären sie dies, würden sich Reidentifikationsversuche erübrigen, da die für die Stichprobenziehung benötigten Informationen in Form des Zusatzwissens zur Verfügung stehen würden.

tes Reidentifikationsverfahren zur Verfügung, mit welchem etwa jedes siebte Unternehmen, welches sowohl im Zusatzwissen als auch im Mikrodatenfile enthalten ist, reidentifiziert werden kann.[13] (B) Das Mikrodatenfile stellt eine Vollerhebung dar und jedes zehnte Unternehmen weist die für die Stichprobenziehung relevanten Merkmalsausprägungen auf. Unter diesen schon sehr optimalen Bedingungen müsste ein Forscher, um nur 10 Fälle für seine Stichprobe zu erhalten, über Zusatzwissen für mehr als 700 Unternehmen verfügen. Um den Stichprobenumfang auf 500 Fälle zu erhöhen, müsste das Zusatzwissen bereits mehr als 35 000 Unternehmen enthalten.

Nun liegen für Unternehmen im Unterschied zur allgemeinen Bevölkerung zwar eine Vielzahl von Informationen in Form von allgemein zugänglichen Datenbanken vor (Brand 2000, S. 110 ff.). Selbst wenn man jedoch hypothetisch unterstellt, dass diese Datenbanken Überschneidungsmerkmale zu amtlichen Mikrodaten enthalten, die sich für eine massenhafte Reidentifizierung eignen würden, ist der Zugang zu diesen Überschneidungsmerkmalen mit Kosten in nicht unbeträchtlicher Höhe verbunden. Für eine Nutzung des Komplettbestands der Hoppenstedt Bilanzdatenbank mit rund 7000 Unternehmen sind bspw. 13 500 € zu veranschlagen (Unterarbeitsgruppe Anonymisierungsmethodik 2002, S. 11). Damit würde schon der Erwerb von Zusatzwissen bei diesem Szenario mit einem nicht geringen Kostenfaktor zu Buche schlagen.[14] Da weiterhin nicht von einer direkten Vergleichbarkeit der Überschneidungsmerkmale auszugehen ist, käme als zusätzlicher Kostenfaktor die für eine adäquate Datenaufbereitung benötigte Zeit hinzu. Gleichfalls ist der Zeitaufwand für die Implementation des Reidentifikationsverfahrens sowie für die letztendliche Durchführung der Reidentifikationsversuche als Kostenfaktor einzubeziehen. Deutlich geringere Kosten wären allerdings dann zu veranschlagen, wenn das für die Reidentifikation benötigte Zusatzwissen nicht erst erworben werden muss, sondern in Form von früher erworbenen Datenbankbeständen oder eigenen Erhebungen bereits zur Verfügung steht.

Aber auch unter dieser verschärften Randbedingung ist die Realitätsnähe dieses Szenarios fragwürdig, da der wissenschaftliche Nutzen einer auf Basis von Reidentifikationen gewonnenen Stichprobe schon aus zwei Gründen äußerst begrenzt ist: (1.) Unternehmen mit seltenen Ausprägungskombinationen in den Überschneidungsmerkmalen weisen eine höhere Reidentifikationswahrscheinlichkeit als Durchschnittsunternehmen auf. Die reidentifizierten Unternehmen würden daher keine Zufallsauswahl der interessierenden Population darstellen, sondern wären durch einen vermutlich erheblichen Selektionsbias charakterisiert. Dieser Selektionsbias würde sich auf eine auf diesen Fällen aufbauenden Eigenerhebung übertragen. Daher ist sehr zweifelhaft, ob derart gewonnene Daten

13) Eine Erfolgswahrscheinlichkeit dieser Höhe ist unrealistisch und nach den bislang vorliegenden Befunden auf Basis realer Daten nur unter sehr spezifischen, eher experimentell angelegten Randbedingungen erreichbar (Müller/Blien/Wirth 1995, S. 145; Bender/Brand/Bacher 2001, S. 380). Diese Randbedingungen beinhalten (1.), dass die Daten keine direkten Identifikatoren (Namen, Anschrift) enthalten, sonst aber in keiner Weise anonymisiert wurden; (2.) dass für die gesuchten Fälle „Teilnahmekenntnis" vorliegt. D.h. bekannt ist, welche der gesuchten Unternehmen an der amtlichen Befragung teilgenommen haben; bzw. (3.) dass das Mikrodatenfile dem Forscher in Form einer Vollerhebung zur Verfügung steht. Mit anderen Worten: Mit der hypothetisch unterstellten Größenordnung wird das potenziell höhere Reidentifikationsrisiko von Wirtschaftsdaten einkalkuliert und eine Abschätzung nach der sicheren Seite vorgenommen.

14) Eine nähere Quantifizierung ist an dieser Stelle nicht möglich. Hierfür wäre u.a. zu klären, welche Datenbanken welche Überschneidungsmerkmale zu amtlichen Wirtschaftsdaten aufweisen und zu welchen genauen Kosten diese Informationen zu beziehen sind.

für die Untersuchung der eigentlich interessierenden Fragestellung noch brauchbar wären. (2.) Bei Verwendung von Primärerhebungen werden bei Erstpublikationen üblicherweise sowohl die Auswahlgrundlage als auch die Stichprobenziehung dokumentiert. Geschieht dies nicht, werden die Forschungsergebnisse in aller Regel eher skeptisch aufgenommen. Werden die Details der Stichprobenziehung hingegen bekannt gemacht, ist mit der *Einleitung eines Strafverfahrens* zu rechnen, da eine Verletzung der Geheimhaltung nach §203 StGB mit einer Geldstrafe oder Freiheitsstrafe bis zu zwei Jahren geahndet werden kann (Statistisches Bundesamt 1988, S. 25).[15]

Zusammenfassend erscheint es sehr unwahrscheinlich, dass ein Forscher bereit wäre, für eine kleine, mit hoher Wahrscheinlichkeit stark verzerrte Stichprobe Kosten und Zeit in Reidentifikationsversuche zu investieren sowie das Risiko einer Strafverfolgung auf sich zu nehmen und nicht stattdessen eine alternative und wissenschaftlich solide Informationsbeschaffung präferiert.

2.1.2 Szenario 2: Reidentifikation wirtschaftsstatistischer Einzeldaten mit dem Ziel der Datenverknüpfung

Das zweite Szenario, bei welchem hypothetisch ein wissenschaftsorientiertes Interesse an einer Reidentifikation unterstellt werden könnte, ist die Verknüpfung von unterschiedlichen Datenbeständen. Vorstellbar wäre etwa, dass die in einer eigenen Erhebung vorliegenden Informationen nicht ausreichen, um eine bestimmte Fragestellung zu untersuchen, andererseits in einem amtlichen Mikrodatenfile (z.B. in der Kosten- und Strukturerhebung) die zusätzlich benötigten Angaben erfasst sind. In diesem Fall könnte eine Verknüpfung der beiden Datenbestände angestrebt werden mit dem Ziel, das Analysepotenzial der eigenen Daten zu erweitern.

Die zentralen Voraussetzungen für dieses Szenario sind: (1.) die eigenen Daten sind *nicht anonym* und sie enthalten (2.) eine Reihe von Überschneidungsmerkmalen zum Mikrodatenfile; (3.) die Auswahlpopulation der eigenen Erhebung überschneidet sich zumindest teilweise mit jener des Mikrodatenfiles, d.h. eine gewisse Zahl von Unternehmen hat an beiden Erhebungen teilgenommen.

Weiterhin werden wiederum optimale Reidentifikationsbedingungen dergestalt unterstellt, dass das amtliche Mikrodatenfile als Vollerhebung verfügbar ist und es möglich wäre, ca. jeden siebten Fall, der sowohl im Mikrodatenfile als auch in der eigenen Datenbasis enthalten ist, zu reidentifizieren: Unter der Annahme der eigene Datenbestand umfasst 3 000 Fälle, könnte die Datenbasis dann für ca. 420 Fälle erweitert werden. Dies mag auf den ersten Blick wie eine sehr große Zahl wirken. In Hinblick auf die eigentliche Motivation der Datenverknüpfung ist es jedoch eine sehr schmale Datenbasis. Hinzu kommt, dass diese erweiterte Datenbasis – wie schon im ersten Szenario – aufgrund des nach Ausprägungskombinationen variierenden Reidentifikationsrisikos einen erheblichen Selektionsbias aufweisen würde. Mit anderen Worten, der wissenschaftliche Nutzen einer derart gewonnenen Informationsbasis erscheint sowohl hinsichtlich des zu erwartenden Umfangs wie auch bezüglich der inhaltlichen Verwertbarkeit dieser Daten als gering.

15) Diese Sanktionsmöglichkeit ergibt sich aus der in § 16 (6) BStatG enthaltenen Eingrenzung des Empfängerkreises von faktisch anonymen Daten auf Amtsträger und für den öffentlichen Dienst besonders Verpflichtete. Diese Einschränkung wurde explizit mit dem Ziel eingeführt, bei einer „unbefugten Offenbarung" eine strafrechtliche Belangung zu ermöglichen (Deutscher Bundestag 1986, S. 21 f.).

Der in diesem Szenario aus einer Reidentifikation zu ziehende Nutzen ist jedoch auch grundsätzlich in Frage zu ziehen: Sofern für eine Verknüpfung von zwei unterschiedlichen Datenfiles nicht für beide Files direkte Identifikatoren zur Verfügung stehen, kann nur mittels Überschneidungsmerkmalen verknüpft werden. Umgekehrt können auch anonyme Datenfiles verknüpft werden, sofern sie geeignete Überschneidungsmerkmale aufweisen. Für eine Datenverknüpfung ist eine Reidentifikation deshalb nicht nur nicht notwendig, sondern ergibt auch keinen Sinn, da sich die Verknüpfung durch die Reidentifikation in keiner Weise qualitativ verbessert. Eine qualitative Verbesserung im Sinne einer Sicherstellung, dass die verknüpften Datensätze tatsächlich zu ein- und demselben Unternehmen gehören, wäre nur möglich, wenn beide Datenbestände jeweils direkte Identifikatoren enthalten. In diesem Fall erübrigt sich aber eine Reidentifikation.

Auf der Basis der zwei hier betrachteten Szenarien lassen sich sicherlich weitere beruflich motivierte Angriffsvarianten ableiten. Allerdings zeigt die obige Diskussion auf, dass es schwer fällt, ausgehend von der beruflichen Interessenlage eines Wissenschaftlers plausible Motive für Reidentifikationsversuche überzeugend zu konstruieren: Selbst in den Grenzfällen, bei welchen ein berufsorientiertes Interesse an einer Reidentifikation unterstellt werden könnte, ergibt die nähere Betrachtung, dass der wissenschaftliche Nutzen einer solchen Handlung als eher gering einzuschätzen ist.

2.2 Nichtberuflich motivierte Nutzenkomponenten von Reidentifikationsversuchen

Das BStatG differenziert explizit zwischen einer wissenschaftsorientierten und einer beliebigen Datennutzung. Diese Differenzierung kommt darin zum Ausdruck, dass der Empfängerkreis von faktisch anonymen Daten nach § 16 (6) BStatG auf die Wissenschaft beschränkt ist und darüber hinaus nicht nur dem allgemein geltenden Reidentifikationsverbot gemäss § 21 BStatG unterliegt, sondern auch dem Geheimhaltungsgebot nach § 16 (1) BStatG. Letzteres bedeutet: Für Nutzer von faktisch anonymen Daten gelten die gleichen strengen Regeln bezüglich der Geheimhaltung wie für Personen, die mit der Durchführung der amtlichen Statistik betraut sind (vgl. auch Fußnote 15)). Vor diesem Hintergrund ist es zumindest diskutabel, ob eine Einbeziehung wissenschaftsfremder Motive für die Operationalisierung der faktischen Anonymität überhaupt erforderlich ist.

Wie bereits oben erwähnt, kommt man nicht jedoch umhin zu konstatieren, dass bei der öffentlichen Diskussion potenzieller Reidentifikationsrisiken wissenschaftsbedingten Motiven im Vergleich zu allgemein menschlichen Motiven eine eher niedrigere Priorität zukommt. Dies mag daran liegen, dass bei letzteren der Phantasie im Prinzip keine Grenzen gesetzt sind: Das Spektrum der hypothetischen Motivlagen reicht von der „Diskreditierung der amtlichen Statistik" über die „Erzielung monetärer Vorteile" oder der „ökonomischen Schädigung Dritter" bis hin zu eindeutig kriminellen Aktivitäten (vgl. Müller et al. 1991, S. 151 ff.). Deshalb werden außerberufliche Motive zumindest angesprochen. Das Ziel kann jedoch auch nicht in der Konstruktion von Szenarien bestehen, bei welchen hochkriminelle Motivationslagen (z.B. Erpressung von Unternehmen) unterstellt werden, um dann hieraus die Schlussfolgerung zu ziehen, dass die angewandten Anonymisierungsmaßnahmen die Unmöglichkeit eines solchen Vorgehens gewährleisten müssen. Dies wäre sicherlich nicht im Sinne von § 16 (6) BStatG.

Vor diesem Kontext wird im Folgenden nicht eine Vielzahl unterschiedlichster wissenschaftsfremder Szenarien diskutiert. Stattdessen wird die für Wirtschaftsdaten zentrale These aufgegriffen, nach welcher Unternehmensdaten Informationen enthalten können, die bspw. für Konkurrenzunternehmen von Interesse sind. Dieser These zufolge ist der aus einer Reidentifikation dieser Daten zu ziehende monetäre Nutzen deutlich höher anzusetzen als für Bevölkerungsdaten und rechtfertigt einen dementsprechend höheren Reidentifikationsaufwand.

2.2.1 Szenario 3: Konkurrenzbeobachtung [16]

Wie bei Brand (2000, S. 37) ausgeführt, sind bei Unternehmensdaten im Unterschied zu Bevölkerungsdaten Risiken zu berücksichtigen, die sich aufgrund der Wettbewerbssituation im Unternehmensbereich ergeben. Beispiele hierfür wären etwa das „Auskundschaften" der Innovationskraft und Entwicklungstätigkeit eines Marktkonkurrenten oder die Verwendung von reidentifizierten Daten durch Consulting-Firmen.[17] Derartige Risikokonstellationen spielen primär zwar nur dann eine Rolle, *wenn wirtschaftsstatistische Daten allgemein zugänglich zur Verfügung stehen würden*, d.h. in Form eines *Public-Use-File*. Hypothetisch kann jedoch auch unterstellt werden, dass ein Forscher aus monetären Interessen wirtschaftsstatistische Einzeldaten dazu benutzen könnte, um gezielt an geheime Unternehmensinformationen zu gelangen. Diese Informationen könnten beispielsweise an interessierte Kreise (z.B. Marktkonkurrenten) verkauft oder für eigene geschäftliche Aktivitäten (z.B. in Form von Unternehmensberatungen) genutzt werden. Vorstellbar wäre eine Situation, bei der im Auftrag eines Unternehmens im Mikrodatenfile gezielt nach Konkurrenzunternehmen gesucht wird. Vice versa könnte auch versucht werden, einzelne interessant erscheinende Unternehmen im Mikrodatenfile zu reidentifizieren und für diese Informationen interessierte Konkurrenten zu suchen.

Im Unterschied zu wissenschaftsorientierten Szenarien ist bei diesem Szenario von einem direkten monetären Nutzen einer erfolgreichen Reidentifikation auszugehen, der unter Umständen sehr hoch sein kann. Eine Quantifizierung dieses Nutzens wirft allerdings insofern Probleme auf, als zum einen nicht bekannt ist, welche Ausgaben Unternehmen ganz allgemein aufwenden, um an geheime Konkurrenzinformationen zu gelangen. Zum anderen liegen keine Anhaltspunkte dafür vor, welchen Wert Unternehmen wirtschaftsstatistischen Einzeldaten im Kontext der Konkurrenzbeobachtung beimessen würden.

Auf einer argumentativen Ebene sind für eine erste Einschätzung des potenziellen Nutzens dieser Daten für eine Konkurrenzbeobachtung folgende Punkte zu berücksichtigen:

---

16) Wenngleich im Folgenden von „geheimen Unternehmensinformationen" gesprochen wird, erscheint die Bezeichnung „Konkurrenzbeobachtung" treffender als „Wirtschaftsspionage". Das Interesse der Wirtschaftsspionage konzentriert sich auf Konstruktionspläne, Fertigungsmethoden, Herstellungsverfahren und Grundlagenforschung; Kundenkarteien und Geschäftsunterlagen, Preislisten, Preispolitik, Absatzplanungen, Bezugsquellen, Angebote, Soft- und Hardwarelösungen; Unterlagen über Serviceleistungen, etc. (Wessing II/Verjans 2002). Es handelt sich hierbei also vorwiegend um Sachverhalte, die eher nicht Gegenstand der amtlichen Statistik sind.

17) Für typische Nutzungsmöglichkeiten von Firmeninformationen siehe Corsten (1999).

Die übermittelten Mikrodaten müssen Informationen[18] enthalten, die im Sinne der Konkurrenzbeobachtung von hohem Interesse[19] sind und nicht über andere legale Wege[20] in Erfahrung gebracht werden können. Inwieweit und für welche wirtschaftsstatistischen Erhebungen eine solche Situation gegeben ist, wäre zu klären. Grosse Unternehmen, die i.d.R. eine professionelle Konkurrenzbeobachtung betreiben, werden aus reidentifizierten amtlichen Daten vermutlich wenig neue Informationen über Konkurrenten ziehen können. Welche Bedeutung der Konkurrenzbeobachtung bei kleineren Unternehmen zukommt, ist schwer abzuschätzen. Allerdings liegt für kleinere Unternehmen auch weniger Zusatzwissen vor, so dass sich die Reidentifikation von Konkurrenzunternehmen erheblich schwieriger gestaltete (vgl. Vorgrimler 2003).

(1) Konkurrenzbeobachtung dient primär der Gewinnung von Marktvorteilen. Es ist anzunehmen, dass die hierfür herangezogenen Daten eine hohe Aktualität aufweisen müssen: Wenn die eigene Strategie am Verhalten der Konkurrenten ausgerichtet werden soll, sind Daten, deren Erhebungszeitpunkt unter Umständen bereits ein Jahr zurückliegt, möglicherweise schon überholt und für die eigene Ausrichtung allenfalls bedingt brauchbar. Dies bedeutet, dass sich mit zunehmenden Alter der Daten der aus einer Reidentifikation zu erzielende monetäre Gewinn verringert.

(2) Damit reidentifizierte Daten für eine Konkurrenzbeobachtung überhaupt von Interesse sein könnten, müssen sie in einer sehr spezifischen Weise *zuverlässig* sein: Es muss sicher gestellt sein, dass die Daten tatsächlich zu dem vermuteten Unternehmen gehören und nicht zu irgendeinem anderen Unternehmen. Diese Art von Zuverlässigkeit ist bei reidentifizierten Daten allerdings in keiner Weise gewährleistet. Selbst unter optimalen Randbedingungen kann ein „Reidentifizierer" nie absolut sicher sein, ob die auf Basis von Überschneidungsmerkmalen vorgenommenen Zuordnungen korrekt sind. Es gibt keine Möglichkeit für eine Verifizierung, ohne den Reidentifikationsversuch offen zu legen. Unter empirischen Bedingungen ist damit zu rechnen, dass der weitaus überwiegende Anteil der vorgenommenen Zuordnungen falsch ist und diese nicht von korrekten Zuordnungen zu unterscheiden sind.[21] Auch unter der Annahme, dass bei wirtschaftsstatistischen Daten das Verhältnis von korrekten zu falschen Zuordnungen im Sinne eines Reidentifizierers „günstiger" aus-

18) Ein Beispiel hierfür könnten etwa Produktinformationen sein. Hierunter fallen alle Informationen, die zu einem Produkt gehören, z.B. technische Angaben, Produzent, Lieferant, Lieferbedingungen und Preis (Corsten 1999).

19) Eine im Sinne der Konkurrenzbeobachtung als hochsensibel eingeschätzte Erhebung ist in Frankreich etwa der „Survey on industrial production prices" (Lequiller 1992).

20) Typische Quellen für Unternehmensinformationen sind etwa Verbände, Nachschlagewerke, Messen, Unternehmensdatenbanken, face-to-face Kommunikation (d.h. Kontakte mit Unternehmensmitarbeitern), Golfplatz, Presse etc. (Corsten 1999).

21) Dies war eine wesentliche Erkenntnis der experimentellen Überprüfungen im Rahmen der faktischen Anonymisierung von Haushalts- und Personendaten. Diese Ergebnisse wurden kürzlich in einer Studie von Bender et al. (2001) bestätigt, bei welcher gleichfalls Reidentifikationsexperimente mit realen Daten durchgeführt wurden. Die Datenkonstellation in dieser Studie war denkbar günstig, da schon bei der Erhebung des Datenfiles, das als Zusatzwissen eingesetzt wurde, hohe Qualitätsmaßstäbe angelegt wurden. Weiterhin lag für alle im Zusatzwissen aufgenommenen Einzeldatensätze „Teilnahmekenntnis" vor, d.h. es wurde hier ein hochriskantes Angriffsszenario untersucht. Dennoch konnten von den gesuchten Fällen nur maximal 14 Prozent korrekt zugeordnet werden, während minimal 86 Prozent falsch zugeordnet wurden. Eine Differenzierung zwischen korrekten und falschen Zuordnungen ist mit dem in der Studie verwendeten Verfahren nicht möglich (mündliche Auskunft von S. Bender).

fällt, ist von einem nicht zu unterschätzenden Unsicherheitsfaktor auszugehen (Vorgrimler 2003). In der Konsequenz bedeutet dies: Beabsichtigt ein Unternehmen für die Konkurrenzbeobachtung auf reidentifizierte Datensätze zurückzugreifen, läuft es Gefahr, seine Entscheidungen nicht nur auf einer *unsicheren*, sondern *möglicherweise völlig falschen Informationsgrundlage* zu treffen. Die ursprüngliche angestrebte Intention, einen Marktvorsprung zu gewinnen, kann sich auf diese Weise ins Gegenteil verkehren. Berücksichtigt man diesen Sachverhalt, wird der eingangs als hoch eingeschätzte monetäre Nutzen von Reidentifikationsversuchen für dieses Szenario doch relativiert.

3 Abschließende Bemerkungen

Die vorangegangenen Ausführungen erheben keinen Anspruch auf Vollständigkeit. Allerdings zeigt auch schon dieser kurze Überblick, dass es schwer fällt, annähernd realistische Angriffsszenarien für wirtschaftsstatistische Einzeldaten der amtlichen Statistik zu konstruieren, die auch einer näheren Betrachtung standhalten.

Für die weiteren Überlegungen ist zu berücksichtigen, dass die Argumentation, soweit sie sich auf Reidentifikationsrisiken bezog, im Wesentlichen auf den für Bevölkerungsdaten vorliegenden empirischen Befunden beruht, wobei die spezifischen Randbedingungen möglichst an die Besonderheiten von wirtschaftsstatistischen Daten angepasst wurden. Wenngleich auf Basis dieser argumentativen Analyse noch keine Kriterien für die faktische Anonymisierung von wirtschaftsstatistischen Einzeldaten ableitbar sind, zeigt sie einige wichtige Punkte auf, die einer näheren empirischen Überprüfung bedürfen:

(a) Dies betrifft die Annahme bei den Wissenschaftsszenarien, dass eine massenhafte Reidentifikation trotz des potenziell umfangreich zur Verfügung stehenden Zusatzwissens auch für wirtschaftsstatistische Einzeldaten nicht möglich ist.

(b) Unter der Annahme, dass eine massenhafte Reidentifikation nicht möglich ist, konzentriert sich das Risikopotenzial bei wirtschaftsstatistischen Einzeldaten auf die Reidentifikation von einzelnen Unternehmen, d.h. vor allem auf wissenschaftsfremde Motive. Sofern wissenschaftsfremden Motiven wie etwa der Konkurrenzbeobachtung bei der Diskussion zur faktischen Anonymisierung von Wirtschaftsdaten eine erhöhte Bedeutung beigemessen wird, wäre zu überprüfen mit welcher Wahrscheinlichkeit gezielte Suchen zu korrekten bzw. falschen Zuordnungen führen und welche Möglichkeiten ein Reidentifizierer hätte, zwischen diesen zu unterscheiden.

(c) Weiterhin wäre es für eine realistische Einschätzung der Reidentifikationsrisiken wichtig, das im Unternehmensbereich zur Verfügung stehende Zusatzwissen nicht nur in Hinblick auf mögliche Überschneidungsmerkmale zu amtlichen Daten zu betrachten, sondern gleichfalls zu analysieren, welcher zusätzliche Informationsgewinn aus einer Reidentifikation zu erzielen wäre.

Abschließend ist anzumerken, dass die gegenwärtig für eine faktische Anonymisierung zur Diskussion stehenden Wirtschaftsdaten beträchtlich hinsichtlich ihres Merkmalsumfangs und damit vermutlich auch bezüglich möglicher Reidentifikationsrisiken variieren. Dies bedeutet, dass man nicht umhinkommen wird, die letztendlich zur Anwendung kommenden Anonymisierungsmaßnahmen datenfilespezifisch anzupassen. Die in Hin-

blick auf das wissenschaftliche Analysepotenzial dieser Daten am wenigsten wünschenswerte Lösung wäre, auf Basis des risikoreichsten Datenfiles Anonymisierungsmaßnahmen zu erarbeiten und diese eins-zu-eins auf weniger riskante Datenfiles zu übertragen.

Literaturhinweise

Bender, S.; Brand R.; Bacher, J. (2001): Re-identifying register data by survey data: An empirical study. In: Statistical Journal of the United Nations ECE 18, pp. 373 – 381.

Brand, R. (2000): Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos. Beiträge zur Arbeitsmarkt und Berufsforschung 237.

Brand; R.; Bender, S.; Kohaut, S. (1999): Möglichkeiten der Erstellung eines Scientific-Use Files aus dem IAB-Betriebspanel. In: Statistisches Bundesamt (Hrsg.): Möglichkeit einer wissenschaftlichen Nutzung von Unternehmensdaten aus der amtlichen Statistik, Stuttgart: Metzler Poeschel, S. 148 – 165.

Corsten, R. (1999): Das gläserne Unternehmen? Firmeninformationen in kommerziellen Online-Archiven. Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft. Band 20.

Deutscher Bundestag (1986): Gesetzentwurf der Bundesregierung. Entwurf eines Gesetzes über die Statistik für Bundeszwecke. Drucksache 10/5345. Sachgebiet 29.

Elliot, M.; Dale, A. (1999): Scenarios of attack: the data intruder's perspective on statistical disclosure risk. In: Netherlands Official Statistics, pp. 6 – 10.

Gnoss, R. (1999): Möglichkeiten und Grenzen der Bereitstellung wirtschaftsstatistischer Einzeldaten der amtlichen Statistik für die Wissenschaft. In: Statistisches Bundesamt (Hrsg.): Möglichkeit einer wissenschaftlichen Nutzung von Unternehmensdaten aus der amtlichen Statistik, Stuttgart: Metzler Poeschel, S. 18 – 29.

Gottschalk, S. (2002): Anonymisierung von Unternehmensdaten – Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels. ZEW: Discussion Paper No. 02 – 23.

Lequiller, F. (1992): Confidentiality in Statistics: The case of the French survey on industrial production prices. In: International Seminar on Statistical Confidentiality. September 8 – 10. Dublin, Ireland. Session 6. Proceedings.

Müller, W.; Blien, U.; Knoche, P.; Wirth, H. (1991): Die faktische Anonymität von Mikrodaten. Forum der Bundesstatistik, Band 19.

Müller, W.; Blien; Wirth, H. (1995): Identification Risks of Microdata. Evidence from experimental studies. In: Sociological Methods & Research. Vol. 24 (2), pp. 131 – 157.

Statistisches Bundesamt (1981): Das Arbeitsgebiet der Bundesstatistik 1981. Kohlhammer. Stuttgart.

Statistisches Bundesamt (1988): Das Arbeitsgebiet der Bundesstatistik 1988. Kohlhammer. Stuttgart.

Sturm, R. (2002): Wirtschaftsstatistische Einzeldaten für die Wissenschaft. In: Wirtschaft und Statistik, Heft 2, S. 101 – 109.

Unterarbeitsgruppe Anonymisierungsmethodik (2002a): Anonymisierungsmethodik und Anonymisierungsstrategie. Arbeitspapier.

Unterarbeitsgruppe Anonymisierungsmethodik (2002b): Aspekte faktischer Anonymisierung. Arbeitspapier.

Vorgrimler, D. (2002): Analyse des Zusatzwissens. Statistisches Bundesamt – Gruppe IA. Arbeitpapier.

Vorgrimler, D. (2003): Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios. (In diesem Band S. 40 ff.)

Wagner, G. (2003a): Innovative Solutions in Providing Access to Microdata: The Case of the German Socio-Economic Panel Study, in: Eurostat News (Theme 1: General Statistics), 19th CEIES Seminar „Innovative Solutions in Providing Access to Microdata", 2003, pp. 133 – 138.

Wagner, J. (1999): Nutzung von betrieblichen Einzeldaten aus der amtlichen Statistik durch externe Wissenschaftler – Modelle, Erfahrungen, Perspektiven –. In: Statistisches Bundesamt (Hrsg.): Möglichkeit einer wissenschaftlichen Nutzung von Unternehmensdaten aus der amtlichen Statistik, Stuttgart: Metzler Poeschel, S. 9 – 17.

Wagner, J. (2003b): Arbeiten mit Einzeldaten der amtlichen Statistik am Beispiel des Monatsberichts im Verarbeitenden Gewerbe. (In diesem Band S. 140 ff.)

Wessing II/Verjans (2002): Betriebsspionage und Maßnahmen zum Schutz von Unternehmensgeheimnissen. In: Der Syndicus. Wirtschaftsstrafrecht. Juli/August 2002. URL: http://www.der-syndikus.de/briefings/wi/wi_005.htm.

Rolf Wiegert *)

Matching Verfahren und die Re-Identifikation faktisch anonymisierter Einzeldaten

Einleitung

Die amtliche Statistik in Deutschland bemüht sich intensiv darum, der Wissenschaft anonymisierte Mikrodatenfiles mit ökonomischen Daten aus Ihren verschiedenen Quellen zur Verfügung zu stellen. Das Forschungsprojekt *Anonymisierung wirtschaftsstatistischer Einzeldaten* begleitet mit seinen Aktivitäten dieses Vorhaben. Dazu gab es in jüngster Zeit zahlreiche Vor-Überlegungen und -Arbeiten, die man anhand der folgenden Publikationen im Wesentlichen verfolgen kann: Albrecht, Weidmann (2002); Brand, Bender, Kohaut (1999); Bustros, Berigan (1999); Müller, Blien, Knoche, Wirth (1991); Vries de, Nobel, J. (1999); Westergard-Nielsen (1999); Wiegert, R. (1993, 1999, 2003).

Betrachtet man die Risiken einer möglichen De-Anonymisierung der anonymisierten Files durch so genannte *Datenangreifer* so zeigen sich spezifische Probleme, die noch untersucht und zuverlässig geklärt werden müssen, um die Datensicherheit mit ihrer prinzipiellen Priorität in der amtlichen Statistik nicht zu gefährden. Zur Messung der Qualität anonymisierter Dateien siehe Höhne (2002).

Die zur Bearbeitung einer beliebigen Fragestellung in einer Gesamtheit zusammengefassten ökonomischen Einheiten, deren relevante Merkmalswerte in einem anonymisierten Mikrodatenfile publiziert werden sollen, müssen vor Abgabe an potenzielle wissenschaftliche Nutzer soweit maskiert werden, dass sie als faktisch anonymisiert gelten können (vgl. dazu auch Müller et al. 1991). Damit ist gemeint, dass nur ein enormer Aufwand, der in keinem realistischen Verhältnis zu einem möglichen Nutzen stünde, erforderlich ist, um eine De-Anonymisierung einzelner (oder aller) Einheiten und Daten des anonymisierten Files durchzuführen. Dennoch bleibt wegen des Datenschutzes vorsorgend zu berücksichtigen, dass De-Anonymisierungsversuche selbst für faktisch anonymisierte Daten nicht ausgeschlossen werden können. Möglicherweise werden aus verschiedenen Motiven und mit unterschiedlichen Methoden, bevorzugt auf Computern, solche Versuche angestellt werden, um sich durch automatisiertes Matching in unterschiedlichen Datenbeständen rekonstruierte Informationen zu verschaffen, die in Struktur und Werten möglichst ähnlich den Originaldaten sind. Die für eine intendierte Re-Identifizierung der Einheiten erforderlichen Zusatzinformationen können u.U. aus kommerziellen Datenbanken oder aus sonstwie compiliertem Datenmaterial beschafft werden. Deshalb wird zum Schutz der faktischen Anonymität bei der weiteren Projektarbeit das Problemfeld gezielter Re-Identifikationsversuche untersucht.

Welche Möglichkeiten eröffnen sich derzeit durch Anwendung von Computern und geeigneter Software, um automatisierte De-Anonymisierungsversuche, so genannte *Massenfischzüge*, durchzuführen? Die folgenden Abschnitte beschäftigen sich mit diesem speziellen Problem und seinen verschiedenen Aspekten (siehe dazu auch Brand 2000).

---

*) Dr. Rolf Wiegert, Akad. Dir. a.D., Universität Tübingen, wissenschaftlicher Berater des Instituts für Angewandte Wirtschaftsforschung (IAW), Tübingen.

Für die weitere Projektarbeit ist es erforderlich, die bereits konkret vorliegenden, anonymisierten Mikrodaten einem solchen computerisierten Validitätstest zu unterziehen. Als Vergleichsmaterial können zunächst die originalen Daten herangezogen werden, die ein Optimum an Qualität möglicher Hilfsinformationen darstellen. Über Besseres kann kein Angreifer verfügen. Wenn die verwendeten Anonymisierungen solchen simulativen[1] Experimenten mit Hilfe von Matching-Prozeduren standhalten, so ist ihre Resistenz gegenüber automatisierten De-Anonymisierungsversuchen, die sich vergleichbarer oder analoger Algorithmen bedienen, nachgewiesen.

Matching – Definitionen, Begriffe und Abgrenzungen

- Ein Datensatz (record) einer statistischen Einheit mit k definierten Merkmalen sei als Zeilenvektor mit k Komponenten definiert. Die Merkmale können unterschiedlich skaliert sein, kategorial oder metrisch.[2] Die Komponenten sind die zugehörigen Merkmalswerte. Durch tabellarische Anordnung der Zeilenvektoren ergeben sich die Merkmalsvariablen als Spalten. Ordnet man den Spalten (absteigend) nach ihrer informationellen Bedeutung für eine spezifische De-Anonymisierung Rangziffern zu, so werden die Variablen 1,...,n als so genannte *Link-Variable* bezeichnet, über die das Matching abgewickelt wird. n wird als Verfahrensparameter festgesetzt.
- Als Link-Variable sind solche Variable zu favorisieren, welche die verglichenen Datensätze möglichst deutlich trennen, d.h. bei denen gleiche Merkmalswerte relativ selten sind. Zum Beispiel ist eine kategoriale Variable mit nur 2 Merkmalausprägungen weniger gut geeignet als eine mit 10 – 12 verschiedenen Ausprägungen.
- *Matching* ist ein paarweiser Vergleich der Komponenten der Link-Variablen von Datensätzen zu Einheiten in einer Datengesamtheit. Die dem Paarvergleich zugrundegelegten Kriterien sind entsprechend der Skalierung der Merkmale zu definieren. Bei Kategorien wird auf Übereinstimmung geprüft, bei metrischen Komponenten durch zusammengesetzte oder einfache Distanzen, die komparativ behandelt werden oder unterhalb vorgegebener Schwellen liegen.
- Ein Match zweier Datensätze liegt vor, wenn z.B die Kriterien für eine Mindestanzahl von Werten der Link-Variablen erfüllt sind oder eine zweckmässig auszuwählende Distanz zweier Objekte (Datensätze) unterhalb einer bestimmter Schranken liegt; man ordnet damit die beiden Datensätze einander zu und wertet sie als zu einer Einheit gehörend.

---

1) Simulativ deshalb, weil bei dieser Vorgehensweise zur Kontrolle der Verfahren sowohl Originalwerte als auch maskierte Werte der jeweiligen Einheit vorliegen. Ordnungsnummern der Originale und der anonymisierten Werte für a posteriori Kontrollzwecke sind bekannt, werden aber beim Matching nicht verwendet. Das ist anders, wenn der Datenangreifer nur Hilfsinformationen einsetzen kann. Er weiß dann nichts über die Zuordnungen. Gerade diese sollen beim Matching über Abgleiche in den Link-Variablen, bei unterschiedlicher Sortierung, erst ermittelt werden.

2) Ein Vorkommen von Alpha-Informationen (Namen, Adressen, u.a.) in den Variablen wird im Kontext der De-Anonymisierung von ökonomischen Mikrodaten nicht betrachtet.

In dem nachfolgenden, zweigeteilten Überblicks-Schema werden verschiedene Verfahren aufgeführt, die in in diesem Kontext von Interesse sind.

Schema 1: Gliederung relevanter Matching Verfahren

(a) Record Linkage Verfahren

Verbindung von Datensätzen gemäss der Erfüllung von Link-Kriterien

Exact Matching Deterministic Matching (All-Or-None-Method)	**Probabilistic Matching** Stochastic Matching
Zusammenführung von Daten mit hinreichend genauer Übereinstimmung in bestimmtenVariablen, den Matching - (Link-) Variablen	Zusammenführung von Daten mit Fehlern, (Maskierungen) in den Matching- (Link-) Variablen; Entscheidungskriterien für Match/Nonmatch, entsprechend den Fehlern, Maskierungen erforderlich. erforderlich.

(b) Statistical Matching

Data Fusion (Data Merging, Mass Imputation)

Datenzusammenführung bei wenigen / keinen identischen Einheiten aber strukturellen Ähnlichkeiten

Ähnliche Einheiten, ähnliche Muster von Daten werden fusioniert nach Ähnlichkeitskriterien, z.B. verschiedenen Distanzmaßen; Daten aus verschiedenen Quellen zu vergleichbaren oder ähnlichdefinierten Einheiten werden zu einem neuen Datensatz vereinigt.

- Die Record Linkage Verfahren unter (a) sind eigentlich für das Verbinden von Informationen zu gleichen Einheiten in verschiedenen Dateien entwickelt worden, z.B. Zusammenführung von Patientendaten, Telefondaten, Alpha-Informationen etc. mit kategorial skalierten Link-Variablen. Für die wesentlich neuartige Frage einer Re-Identifikation anonymisierter Daten-Files mit Hilfe von zusätzlichen Informationen zeichnete sich ab, dass Verfahren des Record Linkage auch für ein solches Vorhaben verwendungsfähig sein könnten. Wenn metrisch skalierte Variable dabei auftreten, muss der eigentliche Matching Prozess grundlegend geändert werden. Eine wesentliche Voraussetzung des Record Linkage, dass beide Datenblöcke (anonymisierte und zusätzliche Daten) Informationen zu gleichen Einheiten enthalten, wird jedoch beim De-Anonymisieren mit Auswahl von mutmaßlich *passenden* Hilfsinfor-

mationen nicht immer erfüllbar sein. In der Praxis stellen sich Deanonymisierungsversuche als ein experimentelles Vorgehen *per trial and error* mit verschiedenen Methoden, unterschiedlichen Hilfsinformationen und variierten Parametereinstellungen dar.

- Probabilistische Matching Verfahren berücksichtigen wie z.B. das Modell von Fellegi und Sunter, im Unterschied zu den exakten Matching-Verfahren, Unsicherheiten der Zuordnungen aufgrund der Paarvergleichskriterien und der Maskierungen bzw. zusätzlicher Datenfehler. Es ist möglich, Schätzungen für Wahrscheinlichkeiten der theoretisch vorausgesetzten Matchings aufgrund empirischer Häufigkeiten der Zuordnungen im kriteriellen Paarvergleich anzugeben. Für potenzielle De-Anonymisierungsversuche sind Modelle dieser Art relevant. Unumgängliche Voraussetzung für ihre Anwendung ist jedoch die Konstruktion valider Paarvergleichskriterien, die sich insbesondere bei metrisch skalierten Merkmalen als nicht unproblematisch herausstellen.

- Die unter (b) im Schema aufgeführten Methoden des statistischen Matchens werden für Datenimputation und Datenmerging verwendet, die sich vom Matching durch schwächere Voraussetzungen hinsichtlich des Vorhandenseins gleicher Einheiten unterscheiden. Sie werden u.a. im Rahmen des Data-Minings verwendet (näheres in Rässler 2002).

- Für die Anwendung der Verfahren bei experimentellen Re-Identifikationen (De-Anonymisierungen) sollten zwei Gesamtheiten von Datensätzen vorhanden sein:

  1. Anonymisiertes Datenmaterial, dem, abhängig vom angewendeten Anonymisierungsverfahren, zusätzliche Informationen, z.B. zu Art, Herkunft, Zeitpunkt etc. beigefügt sind

  2. Hilfsinformationen, die in möglichst enger Relation zum anonymisierten Datenmaterial stehen

- Bei anonymisierten Daten fehlen alle die Einheiten identifizierenden Angaben. Die Variablenwerte jeder Einheit, insbesondere die Link-Variablenwerte sind so maskiert,dass prima vista keine Zuordnungen im Hilfsmaterial erkennbar sind.

- Die Zusammenstellung mutmaßlich brauchbarer Hilfsinformationen ist das zentrale Problem dieser Vorgehensweise. Ein potenzieller Datenangreifer könnte bereits durch diese unvermeidbaren Vorarbeiten mit ihren erheblichen Schwierigkeiten abgeschreckt werden und damit die faktische Anonymität der publizierten Datenfiles nicht antasten, weil der Ausgang des Experimentes für den Datenangreifer völlig ungewiss und das Risiko, Mühe und Kosten vergeblich aufzuwenden, hoch ist (siehe auch Bethleham et al. 1990).

- Für den Angreifer ist es erforderlich, Daten zu finden, die zeitlich, räumlich, sachlich vergleichbar abgegrenzt sind und zusätzlich gleiche/ähnliche Einheiten und Variablen enthalten. Die Einheiten im Hilfsmaterial müssen voll identifiziert sein und ihre Datensätze bereits in EDV tauglicher Form vorliegen. Selbst bei einem voll computergestützten experimentellen Vorgehen ist das eine hohe Hürde für De-Anonymisierungsversuche.

Matching Verfahren und Matching Software – Überblick

Das Modell von Fellegi und Sunter, FS-Modell (1969) zum Matchen von gemeinsamen Elementen in Datensätzen, kann als eine anwendungsfähige Modellbildung im Bereich des Probabilistischen Matchings angesehen werden. Es wurde entwickelt, um z.B. Patientendaten, Adressenbestände u.ä. Daten zusammen zu führen; damit für Link-Variable, die in Engramme zerlegbare Alpha-Informationen und kategoriale Merkmalswerte enthielten. Paarweiser Vergleich, Abfrage auf Gleichheit/Ungleichheit ist bei derartigen Links unproblematisch. Im Gegensatz zu den Methoden des exakten Matchings (0 bei Non-Match, 1 bei Match) bei solchem Datenmaterial berücksichtigt das Fellegi-Sunter-Modell zusätzliche Fehlereinflüsse in den Link-Variablen, durch deren Wirkungen den potenziellen Paarbildungen Wahrscheinlichkeiten zugeordnet werden. Paare von Datensätzen werden Verbindungen (Links) genannnt, die mit hoher Wahrscheinlichkeit Daten/ Informationen zur gleichen Einheit enthalten; Non-Links und Unentscheidbarkeiten sind mit entsprechenden Wahrscheinlichkeiten ausgestattet. Das Fellegi Sunter Modell setzt voraus, dass die Paarvergleiche in den einzelnen Link-Variablen voneinander unabhängig sind.[3] Eine Erweiterung und Verbesserung des FS-Modells mit Hilfe von Q-Abhängigkeiten findet sich in Schürle (2003).

Weil die Wahrscheinlichkeiten i.A. unbekannt sind, müssen sie aufgrund empirischer Paarvergleiche und bestimmter Kriterien für einen Link im konkreten Anwendungsfall geschätzt werden. Dies geschieht mit Hilfe des Expectation-Maximisation-Algorithmus, EM-Algorithmus (Dempster, Laird, Rubin 1977). Er ermöglicht es, unter der Hypothese bedingter Unabhängigkeit und nach Eingabe der empirischen Häufigkeiten, Maximum-Likelihood-Schätzer für die unbekannten Wahrscheinlichkeiten iterativ zu berechnen. Näheres zum EM-Algorithmus findet sich z.B. in Wu (1983).

Dem FS-Modell werden zur Gewinnung eines anwendungsfähigen Matching-Modells zwei zusätzliche Module hinzugefügt:

1. Der EM-Algorithmus
2. Eine dem zu vergleichenden Datenmaterial angemessene Methodik und effektive Kriterien für den Paarvergleich in den Link-Variablen.

Dadurch entsteht ein modular aufgebautes Modell, in dem der eigentliche FS-Ansatz eine Art von theoretischem Hintergrund darstellt, vor dem Ergebnisse geschätzt und probabilistisch interpretiert werden können. Weitere Einzelheiten zur Theorie des FS-Modells finden sich im Anhang zu diesem Beitrag.

Es wurde, im 2. Abschnitt oben, schon darauf hingewiesen, dass bei alphabetischen und kategorialen Link-Variablen keine Probleme auftreten. Schwierig wird es, wenn die Link-Variablen wie hauptsächlich bei der De-Anonymisierung ökonomischer Daten metrisch skaliert vorliegen. Werden für die Paarvergleiche keine angemessenen Vergleichskrite-

3) Das ist die Voraussetzung einer bedingten Unabhängigkeit der Werte der Link-Variablen voneinander. Zur Erläuterung: Bei Vornamen z.B. besteht eine Abhängigkeit mit dem Geschlecht, so dass die entsprechenden Paarvergleiche nicht unabhängig voneinander sind, sondern sich als bedingt abhängig in den Wahrscheinlichkeiten ausprägen. Für das FS-Modell ergeben sich damit im, Gegensatz zur Grundannahme, unterschiedliche aprori Wahrscheinlichkeiten. Die Schätzung derart zusammenhängender Wahrscheinlichkeiten sind im Modell nicht berücksichtigt. Eine Arbeit von Schürle, Tübingen zur Erweiterung des FS-Modells auch für Abhängigkeiten ist in Vorbereitung.

rien konstruiert, so sind die Ergebnisse aus der FS-Modellanwendung nicht valide. Distanzen, wie z.B. die Euklidische Distanz sind wegen ihres additiven Aufbaus aus Einzeldistanzen (gewichtet, ungewichtet) problematisch. Die durch sie bewirkte Abbildung in $\mathbf{R}$ ist surjektiv [4] und deshalb nicht in wünschenswertem Masse eindeutig wirksam für das Matching. Trotz einiger Probleme wird dieses Modell, z.B. im Bureau of the Census in den USA von Winkler (1988, 1995) und von Yancey, Winkler, Creecy (2002) und von Sebé, Domingo-Ferrer et al. (2001) für das Matching bei der theoretischen Untersuchung von De-Anonymisierungen eingesetzt, allerdings nicht bei Anonymisierungsverfahren allgemein, sondern spezifisch bei addiver Zufallsfehler-Überlagerung und perturbierenden Verfahren in den Link-Variablen. In den USA existieren dazu ältere Software-Routinen, die jedoch nicht ohne Zusatzwissen und Kommentierung portabel sind. Auch die bei Domingo-Ferrer vorhandene SAS-Software wurde zwar für Forschungsarbeiten, jedoch noch nicht für ein praktisches Problem mit einer sehr grossen Zahl von Datensätzen und vielen Variablen eingesetzt. An der Universität Tübingen, FB Wirtschaftswissenschaften existiert ein vollständiges Programm für das oben skizzierte FS-Modell, jedoch ausschließlich für Linkvariable, die nicht-metrisch skaliert sind.

An diesem Punkt muss einerseits noch Programmierungsarbeit geleistet werden, andererseits sollte vorrangig untersucht werden, wie metrisch skalierte Link-Variable beim Matching effektiv zu behandeln sind. Denn auch die in der Literatur vorgeschlagenen Vergleichsmaße für die Ähnlichkeiten von Mustern leiden unter demselben Manko einer Surjekivität bei Abbildung auf reelle Zahlen und einer daraus resultierenden Vielzahl unplausibler Zuordnungen. Der zuletzt genannte Grund ist jedoch der entscheidende für die meisten De-Anonymisierungsprobleme bei Wirtschaftsdaten.

Es ist deshalb nicht zweckmäßig, vor der Lösung der eigentlichen Paarvergleichsprobleme bei metrischen Variablen, De-Anonymisierungsexperimente mit unzureichenden Methoden anzustellen. Erst daran anschließend sind Versuche mit dem FS-Modell in sich stimmig und mehr als nur ein blindes Probieren.

Weil diese Schwierigkeiten bei metrischen Link-Variablen auftreten und weil auch die Thematik noch relativ neu ist, gibt es keine Matchingroutinen in den Standardsoftwarepaketen wie SAS, SPSS, STATA, S-Plus-Library oder R, die genuin für solche Aufgaben erstellt wurden. Über die algorithmischen Grundlagen der Routinen, die z.B. in SPSS verwendet werden, siehe Pokropp (2002). Es gab jedoch Versuche, die vorhandenen divisiblen Clusterroutinen in den Software-Paketen, für eine vollständige Zerlegung in Cluster zusammengehöriger Datensätze zu verwenden. Auch bei diesem Ansatz werden komplexe Distanzmasse, die Mehrfach- oder nahezu gleiche Zuweisungen produzieren, verwendet. Probabilistische Ansätze werden im Kontext nicht behandelt.

4) Surjektivität ist die Eigenschaft einer Abbildung Mehrfachzuweisungen verschiedener Urbilder auf gleiche Bildelemente vorzunehmen. Komplexe Distanzmasse können diese Eigenschaft haben. Z.B in dem simplen Fall, wenn die Einzelabstände gleich, aber als Komponenten permutiert wurden; durch unterschiedliche Gewichtungen kann speziell dies eingeschränkt werden, aber die Surjektivität nicht generell verhindert werden. Auch wenn die Distanzmasse nur in ein sehr enges Intervall (approximativ gleich abbilden), so wird eine Interpretation der Distanz-Maßzahlen problematisch.

In Artikeln von Bacher (2001) und Bacher, Brand, Bender (2002) werden interessante Matching-Versuche unter Verwendung von Standard-Clusterroutinen (nearest neighour) an Datenmaterial des Institutes für Arbeitsmarkt und Berufsfoschung (IAB) und des Max Plank Institutes für Humanentwicklung (MPI) dargestellt. Die Vergleichbarkeit nach Maß und Zahl (Kommensurabilität) der verwendeten (unterschiedlich skalierten) 14 Variablen wurde durch Transformation und Gewichtung derselben erreicht. Nach Aussagen der Autoren muss der potentielle Datenangreifer über ein große Menge von Zusatzinformationen zu den Daten verfügen.[5] Die Anwendung derartiger Methoden erscheint, den erzielten Ergebnissen nach zu urteilen, durchaus auch für Re-Identikationsversuche im Projekt in Frage zu kommen. Es wurde die Standardsoftware SPSS und darin die Prozedur Quick Cluster benutzt. Wie sich allerdings ein solches Vorgehen bei beispielsweise ca. 17 000 Einheiten der Daten der probeanonymisierten KSE mit rd. 20 Variablen bewährt, ist noch offen und muss deshalb zur Absicherung der faktischen Anonymität der probe-anonymisierten Datensätze noch näher untersucht werden.

Die Verwendung weiterer schon vorhandener Modelle wie Bayessche Schätzungen sind nicht sehr erfolgversprechend, weil sie Hypothesen über die zugrundeliegenden Matchingwahrscheinlichkeiten oder sonstige Vorkenntnisse über das Zustandekommen der Fehler in den Daten erfordern, die in der Praxis meistens fehlen. In einer Arbeit von Paaß et al. (1985) sind solche Untersuchungen mit Bayesschen Ansätzen vorgelegt worden. Es existieren dazu jedoch, soweit dem Autor zzt. bekannt, keine Standard-Routinen.

Eine neuerdings vorgelegte Arbeit von R. Lenz (2003) verwendet einen interessanten graphentheoretischen Ansatz, um auf der Grundlage von gewichteten, bipartiten Graphen und mit Hilfe von Suchalgorithmen innerhalb des Cartesischen Produktes von zwei bel. Datensatz-Mengen die Vektorpaare mit minimaler Distanz zu finden. Das Verfahren wird im Moment erprobt, deshalb liegen noch keine Ergebnisse praktischer Matchingversuche vor. Der originelle Ansatz, ist sehr überzeugend, wenngleich auch bei ihm wieder beim eigentlichen Matching die Surjektion durch die angewendeten Distanzmaße Schwierigkeiten bereitet. Dennoch sollen experimentelle De-Anonymisierungsversuche mit dieser Methode im Projekt gemacht werden. In einer früheren Arbeit von Chegireddy und Hamacher (1987) zur Bestimmung von k-besten perfekten Matchings werden Suchalgorithmen für das Auffinden k-bester Distanzen in Graphen verwendet. Welche konkreten Vergleichskriterien die Autoren verwenden, ist im Text nicht näher erläutert. Testergebnisse wurden nicht publiziert.

Weiter scheint die Anwendung von Neuronalen Netzen bei der De-Anonymisierung möglich zu sein. Veröffentlichungen oder Erfahrungsberichte liegen momentan noch nicht vor. Im ZEW Mannheim laufen derzeit Versuche mit einem solchen Ansatz im Rahmen des Projektes in Kooperation mit dem Statistischen Bundesamt.

Matching Verfahren spezifischer Art werden auch in der Muster- oder Bilderkennung eingesetzt. Dabei werden Bildelelemente anhand geometrischer Grundmustermuster klassifiziert und über geometrische Ähnlichkeitsmasse miteinander in Beziehung gesetzt. Inwieweit solche Verfahren für De-Anonymisierungen bei Datensätzen und möglicherweise Mustern von Datensätzen, d.h. multidimensionalen Punktmustern einsetzbar sind, konnte noch nicht geprüft werden. Eine Anwendungsmöglichkeit erscheint nicht

5) Bacher, Brand, Bender, S. 605: The intruder must have a lot additional information about a reasonable number of persons in the disseminated file in order to be successful.

ausgeschlossen. Die Bild oder Mustererkennung bedient sich auch der Korrespondensanalyse, um ähnliche Strukturen und Muster in Bildmaterial und Farbbildern auf Ähnlich- oder Unähnlichkeit zu analysieren.

Wie der Überblick in diesem Abschnitt zeigt, der keine Ansprüche auf Vollständigkeit erhebt[6], gibt es einige Arbeiten und Ansätze, die für De-Anonymisierungsexperimente bei der weiteren Projektarbeit in Betracht kommen, jedoch ist das Matching von metrisierten Link-Variablen nicht so trivial wie sie beim ersten Hinschauen erscheint. Sowohl probabilistische wie nicht-probabilistische Ansätze rekurrieren in der Regel auf Paarvergleiche. Sie beruhen auf relativ einfachen Distanzmaßen, die durch Zusammenziehung auf eine eindimensionale reelle Zahlen Unterschiede sowie Gleichheiten zu messen und darzustellen versuchen. Dabei ist die mögliche Vieldeutigkeit der praktizierten Abbildung in den reellen Zahlen und erforderliche Normierungen der Variablen und Distanzen wegen möglichen Verzerrungen der Datenmuster in der Praxis einschränkend zu beurteilen. Bei der De-Anonymisierung von Datensätzen scheinen diese simplen Strategien nicht hinreichend zu sein. Wenn dem so ist, so gehen für die faktische Anonymität durch computerisierte Matching Verfahren nicht allzu große Gefährdungen aus. Dennoch muss diese Hypothese mit praktischen Versuchen im Projekt, speziell an bereits anonymisierten Daten, konfrontiert werden.

An diesem Punkt im Projekt muss einerseits noch erhebliche Programmierarbeit geleistet werden und andererseits untersucht werden, inwieweit die vorhandenen Matching Routinen für Re-Identifikationsversuche der probe-anonymisierten Daten eingesetzt werden können, und zu welchen Resultaten sie für die faktische Anonymität führen. Wenn die Resultate keinen Erfolg signalisieren, ist das von Vorteil für die faktische Anonymität. Dies auch im Hinblick auf das Re-Identifikationsrisiko und seinen Zusammenhang mit der faktischen Anonymität. Wie oben einzeln ausgeführt, sind verschiedene Verfahren und ihre Möglichkeiten zu testen und die Ergebnisse zu bewerten.

Als Fazit aus den bisher vorliegenden Untersuchungen und Erfahrungen ergibt sich das Erfordernis, einige der genannten Verfahren am bis jetzt vorliegenden, mit verschiedenen Anonymisierungsmethoden maskierten Datenmaterial auf seine De-Identifizierungsfähigkeit zu testen und die Ergebnisse für eine praktikable Definition der faktischen Anonymität der Mikrodaten bereit zu stellen. Auch die im folgenden Abschnitt diskutierte heuristischen Methode soll dazu herangezogen werden.

Matching bei metrischen Link-Variablen

Wie der vorangehende Überblick zeigt, gibt es einige Arbeiten und Ansätze, die für De-Anonymisierungsexperimente in Betracht kommen, jedoch ist die Frage des Matchings bei metrisierten Link-Variablen nicht so trivial wie sie beim ersten Hinschauen erscheint. Sowohl probabilistische wie nicht-probabilistische Ansätze rekurrieren in der Regel auf Paarvergleichsmetriken beruhend auf meist additiven Distanzmaßen, die durch Zusam-

6) Die Anzahl von Einzelpublikationen über Record Linkage und verwandte Verfahren ist in den letzten Jahren stark angewachsen. Es werden deshalb hier nur die Arbeiten behandelt, die auf Standard-Software oder zugängliche Routinen zurückgreifen, die einem potentiellen Angreifer, ohne zu großen Aufwand treiben zu müssen, zugänglich sein könnten. Alle angewandten Arbeiten in diesem Bericht zu untersuchen, wäre zu aufwendig. Falls dennoch erforderlich, muss es einer späteren Phase des Projektes vorbehalten bleiben.

menziehung auf eine eindimensionale reelle Zahlen Unterschiede sowie Gleichheiten zu messen und darzustellen versuchen. Dabei ist die mögliche Surjektivität der praktizierten Abbildung in den reellen Zahlen hinderlich. Bei der De-Anonymisierung von Datensätzen scheinen diese einfachen Strategien nicht hinreichend zu sein. Wenn dem so ist, so gehen für die faktische Anonymität von computerisierten Matching Verfahren keine Gefährdungen aus. Dennoch muss diese Hypothese mit praktischen Versuchen im Projekt konfrontiert werden.

Das Matching und das eventuelle Auffinden von zusammengehörigen Datensätzen für eine Einheit kann, einmal anders betrachtet, als eine Art von Siebverfahren unter Verwendung einer k-dim. Intervallschachtelung dargestellt werden. Das hat den Vorteil, dass additive Distanz-Maße entbehrlich werden. Zugleich lässt sich erreichen, dass sich die Anzahl noch zu behandelnder Einzelvergleiche im Ablauf des Verfahrens reduziert.

Es folgt eine Skizze zu einem derartigen nicht-probabilistischen, heuristischen Ansatz, der darauf verzichtet additive Distanzmaße für den Paarvergleich bei metrischen Link-Variablen explizit zu verwenden. Stattdessen werden nur Einzeldistanzen abgefragt und danach deren logischen Aussagewerte zu einer logischen Gesamtaussage zusammengefasst.

Zwei Mengen k-dimensionaler Objekte (Datensätze) des $\mathbf{R}^n$ liegen als M_x, M_y bezeichnet und in unterschiedlicher Sortierung vor. M_x enthält die originalen Datensätze, M_y die anonymisierten Records. Es wird definiert: $C_{xy} := M_x \times M_y$, die Menge von Paaren solcher Objekte.

Folgende Relation R zwischen den Elementen von M_x, M_y auf C_{xy} wird erklärt:

$$\text{R: } \vec{x} \overset{R}{\square} \vec{y} \Leftrightarrow \underset{i \in \{1..k\}}{\forall} |x_i - y_i| \leq q_i |y_i| \quad \text{mit } 0 < q_i \leq 1 \text{ und } \gamma_i \in \mathbf{R}$$

Durch die Relation R wird auf C_{xy} eine Teilmenge von Datensatzpaaren generiert. R *siebt* Teilmengen aus M_x, M_y aus, deren Elemente in Relation stehen. Auf $C_{xy} \times C_{xy}$ entsteht durch die Eigenschaft der Elemente des Paar-Paares, bezüglich R gematcht zu sein, eine Äquivalenz mit Klasseneinteilung in genau 2 Klassen: gematchte und ungematchte Paare.

Bei den Abfragen für die einzelnen Komponenten-Paarvergleiche in den Link-Variablen geht man sukzessive vor , d.h. man *siebt* nach der 1. Link-Variablen, dann in der gewonnenen Teilmenge nach der 2. Link-Variablen u.s.w. Auf diese Art verringert man die Zahl der erforderlichen Abfragen je Datensatz erheblich und die Speicherung der Matches muss lediglich für die Indexpaare der Elemente in C_{xy} vorgenommen werden. Sowohl Speicherplatz- wie Rechenzeitanforderungen vermindern sich durch diese Eigenschaft.

Die q_i sind Parameter und beliebig einregelbar. Die Relation R hat bis auf Symmetrie keine weiteren qualifizierenden Eigenschaften. Obwohl auch in den einzelnen Komponentenpaarvergleichen mit einem einfachen Abstandsmass abgefragt wird, ergibt sich gegenüber komplexen Distanzen der Vorteil, dass die Einzeldistanzen nicht zu einem Wert additiv komprimiert werden müssen. Ordnet man den Einzelabfragen wahr (1) oder falsch (0) zu, so reduziert sich der Vergleich auf die Erzeugung von Vektoren mit Mustern aus wahr und falsch. Man kann über solche (0,1) - Muster Äquivalenzen von Datensätzen parametrisch festlegen und über sie zugleich mögliche Unentscheidbarkeiten bei der simultanen Betrachtung aller Einzelvergleiche in der Anwendung ausschliessen.

Als Ergebnis eines Siebdurchlaufes mit einer durch die q_i fixierten k-dim. Intervallschachtel erhält man eine Menge von gematchten Datensatzpaaren in einer Untermenge des Cartesichen Produktes, die als Kandidaten für Zugehörigkeit zur jeweils derselben Einheit in Frage kommen. Das Verfahren kann rekursiv unter Verwendung einer an den bereits erreichten Sachstand angepassten Parameter-Adaption laufen. Anstelle der Abfragen über prozentuale Abweichungen lassen sich selbstverständlich auch andere Formen über Rangziffern, Quotienten, Messziffern, etc. alternativ in den Algorithmus einbauen.

Die vom Verfahren in simulativen Experimenten unter wechselnden Bedingungen erzeugten Matches sind zur Kontrolle der Wirksamkeit dieser Form des Record Linkage mit Hilfe der Originaldaten nachzuprüfen. Man erhält so Aufschluss darüber, in welchem Ausmass das Verfahren tatsächlich Einheiten identifizieren kann, also wirksam sein und damit für das Computermatching bei metrischen Link-Variablen verwendet werden kann.

Die Festsetzung der Parameter q_i zu Beginn wird erleichtert, wenn man die Spannweiten der einzelnen Links betrachtet und zusätzlich die Histogramme jeder Link-Variablen im anonymisierten und originalen Datensatz zum Vergleich heranzieht. Man kann mit diesem Sieb experimentell – bei wechselnden Parameterkonstellationen vorgehen – und, u.a. zur Vermeidung von zu grossem Aufwand, Teilmengen untersuchen, wobei man zunächst nur wenige anonymisierte Datensätze den Vergleichsdaten zuordnet. Das Verfahren eignet sich infolge seiner einfachen Struktur und Überschaubarkeit für das erforderliche experimentelle Vorgehen bei den Versuchen eine Menge von Datensätzen zu de-anonymisieren.

Die Methode wurde zu Testzwecken im Programmpaket Mathcad implementiert. Die Testläufe ergaben bei Testdaten, dass nur niedrige Prozentsätze für richtige Matches auftraten. Die Abhängigkeit der Matchquote vom angewendeten Anonymisierungsverfahren wurde deutlich sichtbar. Der nächste Schritt wird sein, die bis jetzt faktisch anonymisierten Daten zur KSE (Kostenstrukturberichterstattung) mit diesem Verfahren versuchsweise und über die Auswahl von Teilmengen zur Begrenzung der Rechenzeit zu de-anonymisieren, Als Hilfsinformation stehen entweder die Originalwerte oder geeignete Hilfsinformationen zur Verfügung.

Literaturhinweise

Albrecht, M.; Weidmann, J. (2002): Das nichtamtliche Datenangebot über Unternehmen – Qualität und Verwendung, in: Unternehmen in der Statistik, Konzepte, Strukturen, Dynamik, Schriftenreihe Forum der Bundesstatistik, Bd. 39.

Bacher, J. (2001): Statistisches Matching – Anwendungsmöglichkeiten, Verfahren und ihre praktische Umsetzung in SPSS, in: ZA-Informationen 51.

Bacher, J.; Brand, R.; Bender, S. (2002): „Re-identifying register data by survey data using cluster analysis: an empirical study ", in: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, S. 53 – 74.

Bethlehem, J. G.; Keller, W. J.; Pannekoek, J. (1990): Disclosure Control of Microdata, in: Journal of the American Statistical Association, 85/409, S. 38 – 45.

Brand, R.; Bender, S.; Kohaut, S. (1999): Möglichkeiten der Erstellung eines Scientific-Use-Files aus dem IAB-Betriebspanel. In: Spektrum der Bundesstatistik, Bd. 14, Statistisches Bundesamt.

Brand, R. (2000): Anonymität von Betriebsdaten. Verfahren zur Erfassung und Maßnahmen zur Senkung des Re-Identifikationsrisikos, Beiträge zur Arbeitsmarkt- und Berufsforschung 237, Nürnberg: Bundesanstalt für Arbeit.

Bustros, J.; Berigan, J. (1999): Access to Statistics Canada Microdata files: the Canadian experience, in: Kooperation zwischen Wissenschaft und amtlicher Statistik – Praxis und Perspektiven –, Schriftenreihe Forum der Bundesstatistik, Bd. 34.

Chegireddy, CH.; Hamacher, H. W. (1987): Algorithms for Finding k-best Perfect Matchings, in: Discrete Applied Mathematics (18), S. 155 – 166.

Dandekar, R. A.; Domingo-Ferrer, J. und Sebé, F. (2002): LHS-Based Hybrid Microdata vs Rank Swapping and Micro-Aggregation for Numeric Micro data Protection, 2001. In: Domingo-Ferrer, Josep (Ed): Inference Control in Statistical Data Bases – From Theory to Practice. Springer, 2002.

Dempster, A. P.; Laird, N. M; Rubin, D. B. (1977): Maximum likelihood from incomplete data via the EM-algorithm, in: Journal of the Royal Statistical Society, Series B 39, S. 1 – 38.

Fellegi, I. P.; Sunter, A. B. (1969): A theory for record linkage, in: Journal of the American Statistical Association 1969, 64/328, S. 1183 – 1210.

Höhne, J. (2002): Messung der Qualität einer anonymen Datei. Arbeitspapier der Projektgruppe Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten.

Lenz, R. (2003): A Graph Theoretical Approach to Record Linkage, Working paper for the ECE/Eurostat work session on statistical data confidentiality, Working Paper, Nr. 35 submitted by Statistisches Bundesamt, Wiesbaden.

McLachlan, G. J.; Krishnan, T. (1997): The EM-Algorithm and Extensions, Wiley New York.

Müller, W.; Blien, U.; Knoche, P.; Wirth, H.; et al. (1991): „Die faktische Anonymität von Mikrodaten" Schriftenreihe Forum der Bundesstatistik, Band 19.

Paaß, G.; Wauschkuhn, U. (1985): „Datenzugang, Datenschutz und Anonymisierung. Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten", R. Oldenbourg – Verlag, München.

Pokropp, F. (2002): „Was SPSS rechnet – Grundlagen der angewandten Statistik mit SPSS", Shaker-Verlag, Aachen.

Rässler, S. (2002): „Statistical Matching – A Frequnetist Theory, Practical Applications, and Alternative Baysian Approaches, Lecture Notes in Statistics", Bd. 168, Springer-Verlag, New York, Berlin, u.a.

Schafer, J. L. (2000): EM Algorithm and Data Augmentation, in: Analysis of Incomplete Multivariate Data, S. 38 – 69, Boca Raton.

Schürle, J. (2003): „A Method for Consideration of Conditional Dependencies in the Fellegi and Sunter Model of Record Linkage", in: (to appear.

Sebé, F.; Domingo-Ferrer, J.; Mateo-Sanz, J. M und Torra, V. (2001): Post-Masking Optimization of the Trade-off between Information Loss and Disclosure Risk in masked Micro-Data Sets. In: Domingo-Ferrer, Josep (Ed) (2002): Inference Control in Statistical Data Bases – From Theory to Practice. Springer.

de Vries, W.; Nobel, J. (1999): „Statistik, Geheimnisse und Empfindungen", Oder: Ein kurzer Überblick über die lange und komplexe Geschichte, wie in den Niederlanden statistische Mikrodaten offiziell für externe Forschungszwecke zur Verfügung gestellt wurden, in: Kooperation zwischen Wissenschaft und amtlicher Statistik – Praxis und Perspektiven –, Schriftenreihe Forum der Bundesstatistik, Bd. 34.

Westergard-Nielsen, N. (1999): Linking employer-employee data – the Danish experience, in: Kooperation zwischen Wissenschaft und amtlicher Statistik – Praxis und Perspektiven –, Schriftenreihe Forum der Bundesstatistik Bd. 34

Wiegert, R. (1999): Möglichkeiten verstärkter Nutzung von Verwaltungsdaten für die Wirtschaftsstatistik und zur Entlastung der Wirtschaft von statistischen Berichtspflichten, in: Kooperation zwischen Wissenschaft und amtlicher Statistik – Praxis und Perspektiven –, Schriftenreihe Forum der Bundesstatistik, Bd. 34.

Wiegert, R. (2002): Einleitung, in: Unternehmen in der Statistik, Konzepte, Strukturen, Dynamik, Schriftenreihe Forum der Bundesstatistik, Bd. 39.

Wiegert, R.; Chlumsky, J. (1993): Qualität statistischer Daten, Schriftenreihe Forum der Bundesstatistik Bd. 25.

Winkler, W. E. (1988): Using the EM-algorithm for weight computation in the Fellegi-Sunter model of record linkage, in: Proceedings of the Survey Research Method Section, American Stat. Association, S. 667 – 671.

Winkler, W. E. (1995): Matching and Record Linkage, in: B. Cox et al., Business survey methods, New York, Wiley, S. 355 – 384.

Wu, C. F. J (1983): "On the convergence properties of the EM-algorithm", in: Annals of Statistics, 11, S. 129 – 142.

Yancey, W. E.; Winkler, W. E.; Creecy, R. H. (2002): Disclosure Risk Assessment in Perturbative Microdata Protection, US Bureau of the Census, Research Report Series (Statistics #2002 – 01).

Anhang

Grundlagen des Modells von Fellegi und Sunter

Zwei Datensätze A, B aus einer Menge von Daten zu gleichen Einheiten liegen vor. Sie sollen anhand von k=1,...n Link-Variablen bzw. deren realisierten Werten in A, B verglichen werden. Es wird dazu auf Gleichheit oder Ungleichheit in den Werten geprüft. Es gibt drei Möglichkeiten der Entscheidung:

- Übereinstimmung, Match,
- keine Übereinstimmung, Non-Match,
- keine Entscheidung, neutral.

Sofern die Link-Variablen kategorial skaliert sind und Kriterien für die drei Arten der Entscheidung festgelegt wurden, insbesondere wann Neutralität besteht, so ergeben sich keine Schwierigkeiten (bis auf die Fälle von Datenzwillingen). Falls jedoch metrisch skalierte Link-Variable ins Spiel kommen, werden diese Entscheidungen problematisch, weil z.B. komplexe Distanzmaße Surjektionen (Mehrdeutigkeiten) oder nahezu Surjektionen in $\mathbf{R}$ erzeugen, und außerdem die Entscheidung für einen Match über eine (komplexe) Minimaldistanz bei anonymisierten Daten nicht zuverlässig im Sinne eines realen Matches sein kann. Durch Hinzunahme weiterer Link-Variablen kann in besonderen Fällen eine Verbesserung der Situation erreicht werden, jedoch gilt das nicht generell.

Die Paare von gleichen bzw. ungleichen Paaren werden folgenderart definiert:

$$A \times B := \{ (a,b) \mid a \in A, b \in B \} = M \cup N$$

mit den folgenden Bedeutungen:

$$M := \{ (a,b) \in A \times B \mid a = b \}$$

$$N := \{ (a,b) \in A \times B \mid a \neq b \}.$$

Der Vektor, bezeichnet mit G(a,b), sei ein Vektor mit dem Ergebnis der Vergleiche hinsichtlich der n Link-Variablen für a, b, zwei beliebigen Datensätzen aus A bzw. B .

$$G(a, b) := (g^1(a,b),....,g^n(a,b))$$

Die Menge aller möglichen Realisationen diseses Vektors für alle Elemente aus A bzw. B der Vergleichsraum wird mit M_G bezeichnet. Die einzelnen Elemente dieser Menge mit $M_G(k)$ mit $k = 1,...,|M_G|$.

Die Verteilungen der Vergleichsvektoren zu M, N existieren und sind i.a. verschieden. Es wird zusätzlich vorausgesetzt, dass die Realisationen der Matches in den einzelnen Komponenten der Link-Variablen voneinander stochastisch unabhängig sind. Die Verteilungen sind über folgende Wahrscheinlichkeiten definiert für $k \in \{1,.....,|M_G|\}$:

$$P(M_G(k) \mid M) \quad \text{und} \quad P(M_G(k) \mid N)$$

Diese Wahrscheinlichkeiten sind unbekannt. Deshalb müssen sie bei einer praktischen Anwendung geschätzt werden. Die Wahrscheinlichkeit der neutralen (unentscheidbaren) Fälle soll minimiert werden unter der Nebenbedingung, dass P(Anz. Matches | N)

und P(Anz. Non-Matches | M) vorab fixierte Werte einhalten. Eine Vielzahl neutraler Entscheidungen, die Nachprüfungen von Hand erfordern, wird damit vermieden.

Die dazu passende Optimierungsregel für einen Test auf Qualität des betrachteten Matchens lautet:

P(Anzahl Neutralfälle) Minimum

unter den Restriktionen:

P(Anzahl Matches |N) = μ und P(Anzahl Non-Matches |M) = λ.

μ und λ werden durch vorangehende Testrechnungen oder empirisch festgelegt und bei Bedarf parametrisch variiert.

Mit dem EM-Algorithmus (Expectation-Maximisation-Algorithm; Dempster, Laird, Rubin 1977) werden die unbekannten P(Anz. Matches | M), P(Anz. Non-Matches | N) mit Hilfe der empirischen Auszählungswerte des Matchens als Maximum-Likelihood-Schätzer iterativ aus vorgegebenen Anfangswerten bestimmt. Der EM-Algorthmus ist i.A. numerisch stabil und konvergiert zu einem stationären Punkt der Likelihoodfunktion, der in der Regel zugleich ein lokales Maximum ist. Näheres zum EM-Algorithmus siehe in Wu (1983), McLachlan, Krishnan (1997) und Schafer (2000).

Man erhält nach den Berechnung zu den Matches und Wahrscheinlichkeitsschätzungen analog wie bei einem Test zur Qualitätskontrolle statistische Aussagen über die Zuverlässigkeit der Match- bzw. Non-Matches unsgesamt. Ihre Zuverlässigkeiten hängen strikt vom gewählten Vergleichskriterium des Matchens und der Erfüllung der weiteren Testvoraussetzungen ab. Damit zeigt sich eine kritische Eigenschaft, wenn man das FS-Modell für De-Ananymisierung mit Hilfe von Hilfinformationen betreiben will. Dandekar, Domingo-Ferrer et al. (2002) weist explizit auf diesen Punkt hin: *However, it can be generalized for any perturbative method provided that a distance between the original and the masked value can be defined.* Ein Datenangreifer weiß nicht, ob überhaupt Matches zwischen anonymisierten Daten und Hilfsdaten apriori existieren und wie er eine Distanzmessung angemessen definieren kann. In einer solchen Situation produziert das Verfahren möglicherweise nur Artefakte.

Daniel Vorgrimler *)

Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios

Zur Beurteilung, inwieweit ein anonymisierter Datensatz dem Kriterium der faktischen Anonymität[1] genügt, ist es notwendig, die Re-Identifikationsmöglichkeiten eines potenziellen Datenangreifers abzuschätzen. Die in der Literatur vorliegenden Abschätzungen der Re-Identifikationsmöglichkeiten gehen i.d.R. von einem vorgegebenen Zusatzwissen aus, das sich an dem zu schützenden Datensatz orientiert. Solche für die statistischen Ämter als „worst case" zu bezeichnenden Szenarien blenden das Problem der Generierung von kompatiblem Zusatzwissen[2] aus. Die Schwierigkeiten, die in diesem Prozess liegen, liefern einen Schutz, der bei realistischer Betrachtung der Re-Identifikationsmöglichkeiten nicht vernachlässigt werden darf. Im folgenden Beitrag werden ausgehend von einem extern generierten Zusatzwissen die Re-Identifikationsmöglichkeiten eines Datenangreifers – innerhalb eines bestimmten Szenarios – abgeschätzt und die Kompatibilität zwischen Zieldatensatz und Zusatzwissen analysiert. Darüber hinaus wird die Anonymisierungswirkung verschiedener Anonymisierungsmethoden getestet. Die Vorgehensweise ist an dieser Reihenfolge der Zielsetzung orientiert, wobei der Beitrag durch ein Fazit abgerundet wird.

1 Das Angriffsszenario

Für einen „Angriff" auf einen Mikrodatensatz sind zahlreiche Szenarien vorstellbar (vgl. Wirth 2003). Grundsätzlich sind diese zum einen von der angewandten Technik und zum anderen von der Motivation des Angreifers abhängig. Als Angriffstechnik kann zwischen einem Einzelangriff und einem Massenfischzug unterschieden werden (vgl. Elliot; Dale 1999, S. 7). Zur Vereinfachung wird ein Datenangreifer angenommen, der von 41 ihm nicht näher bekannten Unternehmen die Kenntnis hat, dass diese zur Kostenstrukturerhebung der amtlichen Statistik (KSE) meldepflichtig sind. Das Interesse des Angreifers an den Unternehmen kann aus finanziellen oder sonstigen Gründen geweckt worden sein. Eine solche Situation wurde im Projekt „faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" simuliert.

---

*) Dr. Daniel Vorgrimler, Statistisches Bundesamt, Wiesbaden.
Der Autor dankt Marija Kurtschanowa für ihre umfangreichen Recherchearbeiten.

1) Ausgehend vom Wortlaut des § 16 Abs. 6 des Bundesstatistikgesetzes gilt ein Datensatz als faktisch anonym, wenn der potenzielle Datenangreifer aus rationalem Kalkül die Kosten der De-Anonymisierung höher einschätzt als den Nutzen, den er aus einem erfolgreichen „Angriff" erwartet (unverhältnismäßig hoher Aufwand).

2) Unter Zusatzwissen wird im Folgenden jedes Wissen eines Datenangreifers über ein gesuchtes Unternehmen verstanden, dass dieser aus einer externen Quelle (d.h. nicht aus der amtlichen Erhebung) generiert hat. Merkmale (z.B. Angaben über Anzahl der Beschäftigten) die sowohl im Zusatzwissen als auch in der amtlichen Erhebung beinhaltet sind, werden Überschneidungsmerkmale genannt.

Als Ausgangslage des Datenangreifers ergibt sich daher folgendes Angriffsszenario:

- als Angriffstechnik wird der Einzelangriff gewählt,
- der Datenangreifer besitzt Teilnahmekenntnis der Unternehmen, was die Re-Identifikationswahrscheinlichkeit deutlich erhöht,[3]
- darüber hinaus besitzt er jedoch über die gesuchten Unternehmen nur das allgemein zugängliche Wissen (d.h. er besitzt keine Spezialkenntnisse über die gesuchten Unternehmen).

Das für die Re-Identifikation benötigte Zusatzwissen muss der Datenangreifer über eigene Recherche (z.B. im Internet) selbst generieren. Hier wurde die weitere Annahme getroffen, dass der Datenangreifer das Zusatzwissen nur über kostenfreie Quellen bezieht.[4]

2 Die Zieldatensätze

Als Ziel des Datenangriffs dient die Kostenstrukturerhebung (KSE) in der Form, wie sie als Testdatensatz im Projekt zur faktischen Anonymisierung wirtschaftstatistischer Einzeldaten verwendet wird (vgl. Opfermann et al. 2002, sowie Anhang). Um die Schutzwirkung von verschiedenen Anonymisierungsmaßnahmen abschätzen zu können, wird nicht nur versucht, im „originalen" Datensatz Merkmalsträger zu re-identifizieren, sondern es werden als Vergleich Re-Identifikationsexperimente mit Datensätzen durchgeführt, die in den folgenden Varianten probeanonymisiert wurden (zu den Methoden vgl. Höhne 2003):

a) die stetigen Merkmale wurden mikroaggregiert, wobei jedes Merkmal separat behandelt wurde, die diskreten Merkmale blieben unbehandelt,

b) die stetigen Merkmale wurden wie unter a) mikroaggregiert, bei den diskreten Merkmalen wurde die Wirtschaftsklassifikation nur zweistellig ausgewiesen (Mikroaggregation Variante 1),

c) die stetigen Merkmale wurden wie unter a) mikroaggregiert, bei den diskreten Merkmalen wurde auf die Regionalkennung verzichtet und dafür die Wirtschaftsklassifikation vierstellig ausgewiesen (Mikroaggregation Variante 2),

d) durch die Anwendung des vom Statistischen Landesamt Berlin entwickelte Verfahren SAFE,

e) die stetigen Merkmale wurden wie unter a) mikroaggregiert, die diskreten Merkmale siedlungsstruktureller Kreistyp (BBR 9)[5] und Wirtschaftsklassifikation (WZ 93) wurden mit dem Verfahren Post Randomisation Method (PRAM) behandelt,

---

3) Die KSE ist bei Unternehmen mit mehr als 500 Beschäftigten eine Vollerhebung, wodurch sich der Einwand wieder relativiert (vgl. Statistisches Bundesamt, Erläuterung zum Erhebungsvordruck).

4) Vgl. Abschnitt 4.3, Fußnote 12).

5) Der siedlungsstrukturelle Kreistyp gliedert die Kreise der Bundesrepublik Deutschland nach ihrer Siedlungsstruktur. Die am dichtesten besiedelten Kreise erhalten dabei die 1, während die ländlichsten Kreise die Ausprägung 9 erhalten.

f) durch Rankswapping und

g) die stetigen Merkmale durch das Verfahren „Latin Hypercube Sampling" (LHS), wobei die diskreten Merkmale unbehandelt blieben.

3 Das Zusatzwissen

Die Generierung des Zusatzwissens erfolgte über Internetrecherchen, Auswertungen von Unternehmenspublikationen, Anfragen bei den Unternehmen und Recherchen in kostenfreien Datenbanken der IHK.

Nach Elliot/Dale lässt sich das Zusatzwissen in vier Kategorien einteilen:

1. Leicht zugängliches Zusatzwissen von hoher Qualität (prime keys),
2. Leicht zugängliches Zusatzwissen von geringer Qualität (background keys),
3. Schwer zugängliches Zusatzwissen von hoher Qualität (critical keys),
4. Schwer zugängliches Zusatzwissen von niedriger Qualität (inefficient keys).

Die Qualität des Zusatzwissens hängt u.a. davon ab, wie stark es einen Datensatz differenziert, wie stabil das Zusatzwissen über die Zeit ist und wie hoch die Wahrscheinlichkeit von Messfehlern ist.

Ein Datenangreifer wird zunächst versuchen, die „prime keys" und – falls sie ihm zur Verfügung stehen – die „critical keys" zur Re-Identifikation zu verwenden. Reicht dies nicht zur eindeutigen Re-Identifizierung eines Datensatzes, wird er zusätzlich versuchen, über „background keys" zum Erfolg zu gelangen. Eine Re-Identifizierung, die über „background keys" erfolgt, ist jedoch mit großer Unsicherheit behaftet. Auf der anderen Seite kann eine durch prime keys erfolgte Re-Identifikation durch background keys bestätigt oder in Frage gestellt werden.

Versucht man, die Merkmale der KSE den 4 Kategorien zuzuordnen, so sind ex ante in der KSE als „prime keys" die Branchenzugehörigkeit und die Regionalkennung zu nennen. Diese sind leicht zugänglich und zumindest bei der Regionalkennung mit relativ hoher Sicherheit behaftet. Des Weiteren differenziert die Branchezugehörigkeit den Datensatz sehr stark (wie sich aber zeigen wird, ist dieses Merkmal stärker mit Fehler behaftet als die Regionalkennung). Das Merkmal „tätige Inhaber" ist nicht eindeutig klassifizierbar. Wird ein Unternehmen mit vier tätigen Inhabern gesucht, so ist dies ein „prime/critical key" da nur wenige Unternehmen vier tätige Inhaber in der KSE ausweisen (von 16 918 sind das lediglich 110). Dagegen gibt es viele Unternehmen ohne tätigen Inhaber. In diesem Fall wäre das Merkmal bestenfalls ein „background key", mit dem eine Re-Identifikation bestätigt oder verworfen werden kann.

Von der Publizitätspflicht der Unternehmen ist es u.a. abhängig, in welche Kategorie die Merkmale „Beschäftigte" und „Umsatz" einzuordnen sind. Bei publizitätspflichtigen Unternehmen ist dieses Merkmal leicht und mit hoher Qualität verfügbar (prime key). Bei kleineren und nicht publizitätspflichtigen Unternehmen sind die Ausprägungen dieser Merkmale dagegen schwieriger (oder gar nicht) und mit geringerer Qualität in Erfahrung zu bringen (background/inefficient keys).

Die Merkmale Handelsumsatz und Forschung und Entwicklung (FuE) sind „background keys", da diese meistens nur als ja/nein Ausprägung zu erhalten sind und sie daher den Datensatz nur gering differenzieren. Interessant sind diese Merkmale bei kleineren Unternehmen, die seltener Handel und FuE betreiben. Ist es darüber hinaus möglich, genaue Werte über FuE sowie den Handel zu bekommen, so sind diese Merkmale als „critical keys" zu bezeichnen.

Wie gesehen hängt die Eignung eines Merkmals als Überschneidungsmerkmal von der Struktur des zu suchenden Unternehmens ab. Daraus ergibt sich, dass ein Datenangreifer seine Strategie jeweils speziell auf ein bestimmtes Unternehmen abstimmen muss. Wird daher im Folgenden eine Suchstrategie beschrieben, so ist diese nur als idealtypisch anzusehen, von der je nach Einzelfall mehr oder weniger stark abgewichen werden muss.

4 Ergebnisse der Re-Identifikationsversuche am Originaldatensatz

Die Ergebnisse der Re-Identifikationsversuche, die mit Hilfe der Statistiksoftware SAS durchgeführt wurden, lassen sich in drei Gruppen einteilen:

- Versuche, bei denen Unternehmen *eindeutig und richtig* zugeordnet wurden,
- Versuche, bei denen Unternehmen eindeutig jedoch *falsch* zugeordnet wurden
- und Versuche, bei denen das gesuchte Unternehmen *nicht eindeutig* zugeordnet werden konnte.

Problematisch für die Datensicherheit sind nur die richtig zugeordneten Versuche. Die Möglichkeit, dass Re-Identifikationsversuche zu eindeutigen jedoch fehlerhaften Zuordnungen führen können, steigert sogar die Datensicherheit, da es einem Angreifer nicht möglich ist, zwischen einer richtigen und falschen Zuordnung zu unterscheiden. Für den Datenangreifer verbleibt somit immer die Unsicherheit, inwieweit seine gemachte Zuordnung der Realität entspricht.[6)]

4.1 Ergebnis über alle Unternehmensgrößen

Von den insgesamt 41 zu re-identifizierenden Unternehmen konnten 19 dem Originaldatensatz richtig zugeordnet werden (vgl. Tabelle 1). Weitere 12 wurden falsch zugeordnet und 10 Unternehmen konnten überhaupt nicht zugeordnet werden. Die Wahrscheinlichkeit, dass ein Datenangreifer ein Unternehmen re-identifizieren kann, liegt somit in dieser Simulation bei ca. 46 %. Da der Datenangreifer die Richtigkeit seiner Zuordnung nicht überprüfen kann, liegt das Risiko einer Falschzuordnung bei ca. 38 % (12 von 31 Fällen, vgl. Tabelle 2)[7)]. Beide Wahrscheinlichkeiten sind – wie später noch genauer analysiert wird – stark von der Größe des gesuchten Unternehmens abhängig.

6) Der Abgleichtest auf Richtigkeit der gemachten Zuordnung erfolgte in der Simulation über das Unternehmensregister. Diese Möglichkeit hat ein Datenangreifer nicht, so dass er die Richtigkeit seiner Zuordnung nicht überprüfen kann. Zu der Schutzwirkung des Risikos vgl. Höhne; Sturm; Vorgrimler (2003).

7) Zu beachten ist allerdings, dass dem Datenangreifer diese Wahrscheinlichkeit nicht bekannt ist.

Tabelle 1: Gesamtergebnis der Re-Identifikationsversuche

	alle Unternehmen		kleine Unternehmen [1]		mittlere Unternehmen [2]		große Unternehmen [3]		sehr große Unternehmen [4]	
	Anzahl	Anteil (%)	Anzahl	Anteil (%)	Anzahl	Anteil (%)	Anzahl	Anteil (%)	Anzahl	Anteil (%)
Gesuchte Unternehmen	41	100	9	100	13	100	15	100	4	100
eindeutige Zuordnungen	31	75,6	7	77,7	8	61,5	12	80	4	100
richtige Zuordnungen	19	46,4	1	11,1	5	38,5	9	60	4	100
falsche Zuordnungen	12	29,3	6	66,6	3	23	3	20	0	0
keine Zuordnungen	10	24,4	2	22,3	5	38,5	3	20	0	0
nicht identifiziert (gesamt) [5]	22	53,7	8	88,9	8	61,5	6	40	0	0

1) Weniger als 100 Beschäftigte.
2) 100 – 999 Beschäftigte.
3) 1 000 – 4 999 Beschäftigte.
4) Über 5 000 Beschäftigte.
5) Als nicht identifiziert gelten alle Unternehmen, die falsch zugeordnet wurden und alle, bei denen keine eindeutige Zuordnung gelang.

Quelle: eigene Darstellung

Ein Datenangreifer kann aus verschiedenen Gründen bei einer Re-Identifikation scheitern. So ist es möglich, dass einem Unternehmen ein statistischer Zwilling im Datensatz gegenübersteht, so dass auch bei bestmöglichem Zusatzwissen das Unternehmen nicht eindeutig re-identifiziert werden kann. Des Weiteren ist es möglich, dass entweder das Zusatzwissen fehlerhaft ist, oder der Datenangreifer selbst in seinen Bemühungen einen Fehler begangen hat. In diesen Fällen ist eine Re-Identifikation zwar prinzipiell möglich, jedoch aufgrund „menschlichen Versagens" (z.B. durch fehlerhafte Dateneingabe oder fehlerhaften Annahmen bei der Informationssuche) gescheitert. Ein Ziel der durchgeführten Simulation war, die Schutzwirkung, die von einem solchen „Versagen" ausgeht, zu bewerten. Auf die Bewertung dieser Schutzwirkung wird oftmals bei Analysen des Re-Identifikationsrisikos verzichtet. Dadurch werden aber die Möglichkeiten für einen Datenangreifer ein Unternehmen zu re-identifizieren überschätzt. Die Wahrscheinlichkeit für „menschliches Versagen" ist immer gegeben. Daher geht auch immer ein gewisser Schutz durch sie aus. Allerdings ist eine Quantifizierung dieser Wahrscheinlichkeit nahezu unmöglich. Mit Simulationen kann man einen Eindruck gewinnen, wie groß die Schutzwirkung aufgrund der fehlerhaften Suche ist.

Tabelle 2: Falschzuordnungsquoten [1)]

Alle Unternehmen	38,7
Kleine Unternehmen	85,7
Mittlere Unternehmen	37,5
Große Unternehmen	25,0
Sehr große Unternehmen	0,0

1) Falschzuordnungsquote: Anteil der falschen Zuordnungen an den eindeutigen Zuordnungen.

Quelle: eigene Darstellung

Bei 7 der 22 nicht identifizierten Unternehmen (vgl. Tabelle 1) liegen im Datensatz statistische Zwillinge vor, wobei die Wahrscheinlichkeit für solche Fälle mit der Größe der Unternehmen sinkt. Bei den restlichen 15 Unternehmen wäre es möglich gewesen, bei genauerem und/oder größerem Zusatzwissen die Unternehmen zu re-identifizieren. Die Ursache für die fehlerhafte Suche liegt in diesen Fällen entweder bei einem Fehler im Suchprozess oder bei der Inkompatibilität der Erhebung mit dem recherchierten Zusatzwissen. Auf die Inkompatibilität wird in Kapitel 5 näher eingegangen.

Da zu vermuten ist, dass das Risiko einer Re-Identifikation von der Unternehmensgröße abhängig ist, ist ein solches Gesamtergebnis wenig aussagekräftig. Vielmehr müssen die Versuche in den jeweiligen Kontext der unterschiedlichen Unternehmensgrößen gestellt werden. Um eine solche Analyse zu ermöglichen, wurde der KSE-Datensatz in unterschiedliche Größenkategorien eingeteilt und anschließend aus den verschiedenen Kategorien zufällig Unternehmen ausgewählt, die re-identifiziert werden sollten.[8)] Im Folgenden werden die Ergebnisse für die verschiedenen Größenklassen der Unternehmen näher erläutert.

4.2 Analyse der Re-Identifikationsversuche bei Unternehmen mit weniger als 100 Beschäftigten

In der KSE sind über 9 000 Unternehmen mit weniger als 100 Beschäftigten vertreten. Mehr als die Hälfte aller Unternehmen fallen demnach in diese Größenkategorie.[9)] Um ein Unternehmen in einer Kategorie, die so dicht besetzt ist, re-identifizieren zu können, ist sehr exaktes Zusatzwissen notwendig. Ein Merkmal kleinerer Unternehmen ist jedoch, dass über sie nur sehr wenige Informationen in der Öffentlichkeit vorhanden sind. Daher ist es wenig verwunderlich, dass bei 9 Versuchen lediglich ein Unternehmen richtig re-identifiziert werden konnte (bei 6 falschen und 2 unmöglichen Zuordnungen). Das re-identifizierte Unternehmen hatte eine sehr seltene Merkmalsstruktur, wodurch die Re-

8) Bei den Re-Identifikationsversuchen wäre damit prinzipiell als Vorabinformation eine grobe Beschäftigtengröße des Unternehmens bekannt gewesen. Diese Information blieb aber bei den durchgeführten Versuchen unberücksichtigt. Das Merkmal „Beschäftigte" wurde nur dann als Überschneidungsmerkmal verwendet, wenn das Wissen über die Beschäftigtenanzahl auch tatsächlich aus einer externen Quelle stammte.

9) Diese Unternehmen beschäftigen ca. 9,3 % aller Personen, die bei Unternehmen der KSE-Erhebung beschäftigt sind.

Identifikation möglich wurde. Es betreibt FuE, weist eine Handelstätigkeit auf und es sind Inhaber im Unternehmen tätig. Eine solche Kombination gibt es bei Unternehmen mit weniger als 100 Beschäftigten nur 160 Mal.[10] Diese 160 Fälle mit dem WZ-Viersteller und dem BBR 9 kombiniert, ergibt 129 einmalige Fälle, so auch das gesuchte Unternehmen. Alle anderen Versuche, ein Unternehmen zu re-identifizieren, schlugen fehl. Es ist daher sehr wahrscheinlich, dass eine Re-Identifizierung eines kleinen Unternehmens nur bei sehr spezifischen Bedingungen möglich ist. Und da – wie noch gezeigt wird – einige dieser Merkmale, die zu diesen Bedingungen führen, fehlerhaft sind, ist eine Re-Identifizierung kleiner Unternehmen von großen Unsicherheiten begleitet. Dies zeigt sich auch in der hohen Falschzuordnungsquote (vgl. Tabelle 2). Zuordnungen sind i.d.R. überhaupt erst möglich, wenn man als Datenangreifer bereit ist, subjektive Annahmen über das gesuchte Unternehmen zu verwenden. Dadurch ist das Zusatzwissen mit höheren Fehlerraten behaftet, wodurch die vielen falschen Zuordnungen entstanden sind.

Die Erfahrungen der Re-Identifikationsexperimente bei kleinen Unternehmen zeigen, dass ein hoher Schutz aufgrund einer hohen Dichte der Merkmale und dem schlechteren Zusatzwissen (sowohl qualitativ als auch quantitativ) gegeben ist. Über 9 000 Unternehmen – und damit mehr als die Hälfte der Erhebung – können aus dieser Sicht als faktisch anonym gelten und dies, ohne Maßnahmen zur Anonymisierung zu treffen.[11]

4.3 Analyse der Re-Identifikationsversuche bei Unternehmen mit 100 bis 999 Beschäftigten

Von den 13 zu re-identifizierenden Unternehmen mit 100 bis 999 Beschäftigten konnten 5 richtig zugeordnet werden. 3 Unternehmen wurden falsch zugeordnet. Bei den restlichen 5 Unternehmen war keine Zuordnung möglich (vgl. Tabelle 1). 5 Fehlversuche waren auf Fehler im Suchprozess bzw. im Zusatzwissen zurückzuführen. Bei den restlichen 3 Unternehmen kann selbst bei bestmöglichem Zusatzwissen keine Eindeutigkeit generiert werden.

Die genaue Analyse der Unterschiede in den Unternehmensgrößen zwischen re-identifizierten und nicht re-identifizierten Unternehmen liefert ein Indiz dafür, ab welcher Unternehmensgröße ein Datensatz nicht mehr per se als faktisch anonym zu bezeichnen ist. Die re-identifizierten Unternehmen hatten eine durchschnittliche Größe von 481 Beschäftigten. Durchschnittlich beschäftigen alle Unternehmen dieser Größenklasse 289 Personen. Selbst das kleinste der fünf re-identifizierten Unternehmen weist mit 372 tätigen Personen einen höheren Wert für dieses Merkmal aus. Von den gesuchten Unternehmen dieser Klasse hatten einerseits 6 weniger als 250 tätige Personen, von denen keines re-identifiziert werden konnte. Andererseits waren von den 7 Versuchen bei Unternehmen mit mehr als 250 Beschäftigten 5 erfolgreich. Irgendwo in diesem Größenbereich scheint demnach die Trennlinie zwischen faktisch anonymen und faktisch nicht

10) Hier wird deutlich, dass eine Kategorisierung des Zusatzwissen (vgl. Abschnitt 3) immer von der speziellen Struktur des Unternehmens abhängt. Die genannten Merkmale sind bei einer anderen Struktur nicht sehr hilfreich (höchsten als „background keys" zu verwenden), hier jedoch haben sie den Charakter von „prime keys".

11) Ausgenommen die Maßnahmen zur formalen Anonymisierung (streichen eindeutiger Identifikatoren wie Namen und Adresse) und das Umkodieren des Kreisschlüssel in den Siedlungsstrukturellen Kreistyp.

anonymen Datensätzen zu liegen[12]. Nimmt man die Grenze von 250 tätigen Personen, so könnten zusätzlich zu den rund 9 000 kleinen – in Abschnitt 4.2 behandelten – Unternehmen 3 906 weitere Unternehmen als faktisch anonym betrachtet werden (insgesamt 13 319 Datensätze, das entspricht mehr als ¾ aller Unternehmen der KSE-Erhebung).

4.4 Analyse der Re-Identifikationsversuche bei Unternehmen mit 1 000 bis 4 999 Beschäftigten

Von den 15 zu re-identifizierenden Unternehmen dieser Größenkategorie konnten 9 einem Datensatz in der KSE richtig zugeordnet werden. 3 Unternehmen konnten nicht und 3 wurden falsch zugeordnet (vgl. Tabelle 1). Dies zeigt, dass auch bei Unternehmen dieser Größe der Datenangreifer immer ein Restrisiko der Falschzuordnung trägt (25 %, vgl. Tabelle 2). 5 der 6 fehlerhaften Versuche kamen aufgrund von Fehlern im Suchprozess und/oder im Zusatzwissen zustande, in einem Fall war die Re-Identifikation nicht möglich, weil sich 3 Datensätze – selbst bei optimalem Zusatzwissen – zu stark ähnelten. Das größte nicht re-identifizierte Unternehmen hatte gut 2 300 Beschäftigte, die 3 Unternehmen dieser Größenklasse mit mehr als 2 300 Beschäftigten wurden alle re-identifiziert.

Die höhere Trefferquote war begleitet durch einen leichteren Zugang, bei gleichzeitig geringeren Bedarf an Zusatzwissen. So haben bei einigen Unternehmen bereits sehr grobe Angaben ausgereicht, um eine Zuordnung zu einem Datensatz in der KSE herzustellen. Dies gilt im besonderen Maße für Unternehmen, bei denen die Klassifikation nach dem WZ 93 auf vier Stellen möglich war.

Hauptproblem beim Generieren des Zusatzwissens war bei diesen Unternehmen,[13] die Zahlen desjenigen Unternehmens zu finden, das auch tatsächlich in die KSE gemeldet hat. Die KSE definiert ein Unternehmen als „die kleinste Einheit, die aus bilanz- und/oder steuerrechtlichen Gründen Bücher führt und bilanziert" (Statistisches Bundesamt, Kostenstrukturerhebung – Erläuterung zum Erhebungsvordruck). Dies hat zur Folge, dass nicht die Werte eines Konzerns einschließlich aller rechtlich selbständigen Töchter in die Erhebung Eingang finden, sondern alle „Konzernteile", die selbständig Bücher führen, finden sich in der KSE separat wieder. Erste Aufgabe beim Generieren von Zusatzwissen ist es daher, das „richtig" passende Unternehmen zu finden.[14]

4.5 Analyse der Re-Identifikationsversuche bei Unternehmen mit mindestens 5 000 Beschäftigten

Zu allen 4 Unternehmen mit mehr als 5 000 Beschäftigten konnten eindeutige und richtige Zuordnungen zu einem Datensatz der KSE hergestellt werden (vgl. Tabelle 1). Dabei herrschte nur bei einem Unternehmen Unsicherheit bezüglich der Richtigkeit der Zuord

12) Dieses Ergebnis wird dadurch erhärtet, dass bei einer Recherche in einer kostenpflichtigen Unternehmensdatenbank von den ausgewählten Unternehmen mit weniger als 250 Beschäftigten lediglich ein Unternehmen zusätzlich re-identifiziert werden konnte.

13) Eingeschränkt gilt dies auch bei Unternehmen mit 100 bis 999 Beschäftigten.

14) Zum Begriff des Unternehmens vgl. Voy, K. (2002).

nung. Dagegen waren die anderen 3 nicht nur eindeutig, sondern mit nur geringen Unsicherheiten behaftet. Nötig war zur Re-Identifikation dieser Unternehmen lediglich die vierstellige Wirtschaftsklassifikation und ungefähre Angaben über Umsatz und Beschäftigung.

4.6 Fazit der Größenbetrachtung

Die Analyse der Re-Identifikationsversuche in Abhängigkeit von der Größe der Unternehmen zeigt, dass unter dem betrachteten Szenario kleine Unternehmen (evtl. bis 250 Beschäftigte) so schwierig zu re-identifizieren sind, dass sie per se als faktisch anonym gelten könnten. Verfügt ein Datenangreifer nicht über Teilnahmenkenntnis (entgegen der Annahme in der Simulation), sinkt das Re-Identifikationsrisiko zusätzlich. Daher erhärtet die Tatsache, dass es sich bei der KSE bis zu dieser Unternehmensgröße um eine 38 % Stichprobe handelt (vgl. KSE-Beschreibung), das Gesagte. Bei Unternehmen mit mehr als 250 Beschäftigten verbessern sich die Möglichkeiten der Re-Identifikation beträchtlich. Allerdings muss ein Datenangreifer zumindest bis zu einer Größe von etwa 2 300 Beschäftigten immer noch mit fehlerhaften Zuordnungen rechnen. Die 7 Unternehmen, die eine höhere Anzahl von Beschäftigten hatten, konnten dagegen mit akzeptablem Aufwand alle richtig zugeordnet werden. Daraus lässt sich der Vorschlag ableiten, dass die Maßnahmen zur Anonymisierung der KSE in Abhängigkeit zur Größe der Unternehmen getroffen werden sollten. Bei Unternehmen bis 250 Beschäftigten könnte dabei weitgehend auf Anonymisierungsmaßnahmen verzichtet werden. Bei Unternehmen bis etwa 2 000 Beschäftigten kann berücksichtigt werden, dass die Re-Identifikation eines Unternehmens mit dem Risiko der Falschzuordnung behaftet ist. Für noch größere Unternehmen (das sind 256) werden wohl weitgehende Schutzmaßnahmen vonnöten sein.

5 Analyse der „natürlichen" Abweichungen zwischen Zieldatensatz und Zusatzwissen

Im vorigen Kapitel wurde eine nicht gelungene Zuordnung dadurch erklärt, dass entweder zu dem Unternehmen mindestens zwei nahezu identische Datensätze passten oder dass Fehler im Suchprozess aufgetreten sind. Um diese Fehler näher zu analysieren, sollen in diesem Abschnitt die Abweichungen zwischen Zusatzwissen und Zieldatensatz untersucht werden.

Folgende Merkmale wurden in unterschiedlicher Häufigkeit als Überschneidungsmerkmale verwendet:

- Siedlungsstruktureller Kreistyp (BBR 9),
- Wirtschaftsklassifikation (WZ 93),
- Umsatz,
- Beschäftigte,
- tätige Inhaber (in der Ausprägung ja / nein),
- Handelsumsatz (in der Ausprägung ja / nein),
- Aufwand für Forschung und Entwicklung (in der Ausprägung ja / nein).

- **Siedlungsstruktureller Kreistyp (BBR 9)**

Das Merkmal „Siedlungsstruktureller Kreistyp" (BBR 9) wurde für alle Re-Identifikationsversuche als Ausgangspunkt verwendet. Dabei wurde angenommen, dass dem Angreifer die Kodierung dieses Merkmals bekannt ist. Er kann daher mit relativ wenig Aufwand zu einem beliebigen deutschen Ort den passenden Kreistyp finden. Unter dieser Annahme ist dieses Merkmal das sicherste und am leichtesten zugängliche der o.g. Merkmale und so konnte jedem gesuchten Unternehmen der richtige BBR 9-Schlüssel zugewiesen werden. Durch die neun Ausprägungen garantiert er bereits eine erste sichere Differenzierung des Datensatzes. Tabelle 3 zeigt, wie sich die Merkmalsträger auf die unterschiedlichen Ausprägungen verteilen.

Tabelle 3: Verteilung der Unternehmen nach dem Siedlungsstrukturellen Kreistyp

Siedlungsstruktureller Kreistyp	Häufigkeit	Häufigkeit in %
1	2 920	17,26
2	2 994	17,70
3	1 379	8,15
4	486	2,87
5	788	4,66
6	4 199	24,82
7	1 987	11,74
8	1 488	8,80
9	677	4,00

Quelle: eigene Darstellung

- **Wirtschaftsklassifikation (WZ 93)**

Die Wirtschaftsklassifikation ist ein besonders wichtiges Schlüsselmerkmal. Die Unternehmen der KSE verteilen sich bei einem vierstelligen WZ 93 auf rund 250 verschiedene Klassen. Bei einem zweistelligen Ausweis des WZ 93 sind es immerhin noch 26 so genannte Abteilungen, auf die sich die Unternehmen verteilen. Aufgrund der hohen Differenzierungswirkung und der theoretisch leichten Verfügbarkeit dieses Merkmals, ist die WZ 93 als „Prime Key" Variable zu bezeichnen. Allerdings ist die Wirtschaftsklassifizierung eines Unternehmens oftmals mit Problemen bei der Verfügbarkeit und der Genauigkeit verbunden. Tabelle 4 zeigt für das empirische Fallbeispiel, für die verschiedenen Größenklassen und insgesamt, wie oft die WZ 93 als Überschneidungsmerkmal verwendet werden konnte, und wie viele Inkompatibilitäten zwischen der aus dem Zusatzwissen generierten Klassifikation und der Klassifikation, die in der KSE verwendet wird, aufgetreten sind. Eine Zuordnung des gesuchten Unternehmens zu einer vierstelligen WZ 93 Klasse war demnach 29 Mal bei 41 Versuchen möglich. Davon waren jedoch 7 Zuordnungen falsch. Dies bedeutet, dass nur in etwas mehr als die Hälfte der Fälle auf Basis eines richtigen vierstelligen WZ 93 Codes Re-Identifikationsversuche durchgeführt wer-

den konnte. Diese Quote erhöht sich bei den Gruppen der dreistelligen Wirtschaftsklassifikation.[15] Da ein Re-Identifikationsversuch, besonders bei mittleren und kleinen Unternehmen, ohne eine Wirtschaftsklassifizierung äußerst schwierig ist, wurde zu jedem Unternehmen mindestens eine „passende" zweistellige WZ 93 Abteilung „gesucht". Daraus folgt, dass einerseits alle Unternehmen einer Abteilung zugeordnet sind, andererseits die Fehlerquote in diesem Bereich beträchtlich ist (immerhin sind 9 der 41 Kategorisierungen falsch).

Tabelle 4: Inkompatibilitäten zwischen KSE und Zusatzwissen beim Merkmal WZ 93

Gliederungstiefe	Unternehmensgröße	insgesamt	richtig	falsch
vierstellig	bis 100 Beschäftigte	5	2	3
	100 – 999 Beschäftigte	11	9	2
	1 000 – 4 999 Beschäftigte	9	7	2
	über 5 000 Beschäftigte	4	4	0
	insgesamt	**29**	**22**	**7**
dreistellig	bis 100 Beschäftigte	6	3	3
	100 - 999 Beschäftigte	12	11	1
	1 000 – 4 999 Beschäftigte	12	10	2
	über 5 000 Beschäftigte	4	4	0
	insgesamt	**34**	**28**	**6**
zweistellig	bis 100 Beschäftigte	9	5	4
	100 – 999 Beschäftigte	13	11	2
	1 000 – 4 999 Beschäftigte	15	12	3
	über 5 000 Beschäftigte	4	4	0
	insgesamt	**41**	**32**	**9**

Quelle: eigene Darstellung

Die Abhängigkeit der Verfügbarkeit von Zusatzwissen von der Unternehmensgröße kommt ebenfalls in der Tabelle zum Ausdruck. Während die Unternehmen mit mehr als 5 000 Beschäftigten problemlos einer vierstelligen WZ 93 Klasse zugeordnet werden konnte, konnten die Unternehmen mit weniger als 100 Beschäftigten oftmals nur einer Abteilung (d.h. zweistellig) zugeordnet werden und auch dies nur mit einer hohen Fehlerquote. Die richtige vierstellige Codierung war für die kleinen Unternehmen nur zweimal vorhanden. Bei der hohen Dichte, die der Datensatz bei den kleinen Unternehmen aufweist, ist es aber eine notwendige Voraussetzung, dass die Unternehmen einen richtigen vierstelligen Code aufweisen. Bei den größeren Unternehmen ist das nicht unbedingt nötig. Dies ist auch der Grund, warum diese Unternehmen trotz ihrer Größe einen relativ geringen Anteil an richtigen Klassen (d.h. vierstelligen Codes) aufweisen (nur 7 von 15, dagegen 9 von 13 bei den mittleren Unternehmen). In einigen Fällen war die Suche nach einer vierstelligen Codierung nicht notwendig, da eine eindeutige Re-Identifizierung bereits nach einer zweistelligen Codierung gelungen war.

15) Bemerkung: Bei einem Übergang von einem vierstelligen auf einen dreistelligen Code werden auf der einen Seite Fehler reduziert (und zwar dann, wenn der Fehler lediglich bei der vierten Stelle auftritt), auf der anderen Seite kommen neue Unternehmen hinzu, deren Klassifikation evtl. auf der dritten Stelle fehlerhaft ist. Aufgrund dieser beiden unterschiedlichen Effekte kann nicht von der Fehleranzahl der einen Gliederungsebene auf die Fehleranzahl einer anderen geschlossen werden.

Wie bereits erwähnt ist die Klassifizierung für einen Datenangreifer eine wichtige Basis, von der er seine Re-Identifikationsversuche startet. Ist diese Basis bereits fehlerhaft, so wird er *nicht* den von ihm gewünschten Erfolg haben. Die Inkompatibilitäten, die in der Branchenklassifizierung der Unternehmen auftauchen, sind daher ein Hauptgrund für die falschen Zuordnungen.

- **Umsatz**

Bei dem Merkmal Umsatz stellt sich die Frage, inwieweit es überhaupt ein Überschneidungsmerkmal ist. Es kann vielmehr ein Hauptziel eines Datenangreifers sein, den Umsatz der Unternehmen zu enthüllen, in dem Fall wäre er dann kein Überschneidungsmerkmal, sondern ein Zielmerkmal.

Im verfügbaren Zusatzwissen spiegelt sich wider, dass Unternehmen ihren erzielten Umsatz lieber geheim halten. In nur 22 der 41 Fälle stand das Merkmal Umsatz als Überschneidungsmerkmal zur Verfügung.[16] Bei den 9 kleinen Unternehmen stand nur einmal eine Angabe zum Umsatz zur Verfügung und bei den 13 mittleren fünfmal. Bei den publizitätspflichtigen größeren Unternehmen standen dagegen i.d.R. Angaben zum Umsatz zur Verfügung. Die Angaben zum Umsatz müssen aber in zwei Kategorien unterteilt werden. Zum einen gibt es kategoriale Angaben, zum anderen gibt es „genaue" Angaben über die Umsatzhöhe. Bei Töchterunternehmen, von denen lediglich Werte ihrer Mütter bekannt waren, konnten nur Umsatzobergrenzen angegeben werden. Im Folgenden laufen solche Angaben unter kategorialen Angaben. Von den 22 Umsatzangaben standen 12 lediglich als kategoriale Angaben zur Verfügung. Davon waren 2 fehlerhaft. Von den restlichen 10 „genauen" Angaben zum Umsatz haben lediglich 2 beim Zusatzwissen und der KSE exakt den gleichen Wert. Zwei hatten Abweichungen von weniger als 10 %. Die restlichen 6 Angaben hatten höhere Abweichungen aufgewiesen, 2 davon waren wegen ihren extremen Abweichungen unbrauchbar. Die Konsequenz aus einem solchen Ergebnis ist, dass ein Angreifer den Umsatz nicht als „prime key" verwenden kann. Vielmehr stellt sie ein „background key" dar, mit der eine Entscheidung erhärtet werden kann. Die Gründe für die schlechte Eignung des Merkmals „Umsatz" sind vielschichtig. Der wichtigste, dass Unternehmen nur im geringeren Maße bereit sind Angaben zum Umsatz zu machen, wurde bereits genannt. Ebenso wurde als Problem bereits die unterschiedlichen Abgrenzungen eines Unternehmens genannt. Des Weiteren muss bedacht werden, dass der Umsatz eine sich sehr schnell veränderte Größe darstellt. In einigen Fällen liegen aber im Zusatzwissen Umsatzangaben vor, die zeitlich nicht dem Erhebungsjahr der KSE entsprechen.

- **Beschäftigte**

Auskunftsfreudiger sind die Unternehmen bei der Anzahl der Beschäftigten. So stehen in 33 Fällen Angaben zur Anzahl der Beschäftigten zur Verfügung. Bei allen 19 Unternehmen mit mehr als 1 000 Beschäftigten war dieses Wissen aus dem Zusatzwissen generierbar. Aber auch kleinere Unternehmen tun sich leichter die Zahl der Beschäftigten zu veröffentlichen. So stehen immerhin bei 5 der 9 kleinen und bei 9 der 13 mittleren Unternehmen Angaben zu diesem Merkmal zur Verfügung. In 13 Fällen handelt es sich um

16) In einigen Fällen liegt dies aber auch daran, dass die Re-Identifikation bereits erfolgreich abgeschlossen wurde, ohne dass nach dem Merkmal Umsatz im Zusatzwissen gesucht werden musste.

korrekte kategoriale Angaben, in 20 Fällen um „exakte". Diese Werte lassen im Verhältnis zum Umsatz auf eine höhere Qualität des Zusatzwissens schließen. Allerdings sind einige kategoriale Angaben derartig grob, dass sie den Datensatz nur unzureichend differenzieren. Bei den „exakten" Angaben stimmen 3 Angaben des Zusatzwissens mit denen der KSE überein. Weitere 6 haben eine Abweichung von weniger als 10 % und nur 2 sind aufgrund ihrer hohen Abweichung gänzlich unbrauchbar. Daher ergibt auch die Abweichungsanalyse der exakten Beschäftigtenangaben, dass sich dieses Merkmal besser für Re-Identifikationsversuche eignet als das Merkmal Umsatz.

Aufgrund der bisherigen Analyse erfolgt ein idealtypischer Datenangriff zunächst über die Merkmalskombination WZ 93 und BBR 9. Anschließend werden die Unternehmen gesucht, die von der Zahl der Beschäftigten am besten passen. Mit dem Merkmal Umsatz wird das Ergebnis bestätigt oder verworfen.

- **Tätige Inhaber**

Rund 4 000 Unternehmen geben in der KSE mindestens einen tätigen Inhaber an. Hat das gesuchte Unternehmen einen tätigen Inhaber und ist dieses Wissen im Zusatzwissen vorhanden, dann ist dies eine wertvolle Information, die den Datensatz ähnlich stark differenziert wie der BBR 9. Diese Information ist aber erstens nur schwierig zu bekommen und zweitens sehr unsicher (critical / inefficient key). Die Information „keine tätigen Inhaber" ist dagegen recht leicht zu erhalten (z.B. wenn es sich um ein Tochterunternehmen eines großen Konzerns handelt), diese Information differenziert den Datensatz jedoch nur sehr wenig (besonders bei großen Unternehmen).

Insgesamt wurde im Fallbeispiel elfmal die Information „tätige Inhaber vorhanden" verwendet. Davon sind die Informationen nur fünfmal korrekt. Diese 5 richtigen Fälle sind viermal direkt an einer korrekten Re-Identifizierung beteiligt. Dies zeigt, wie wertvoll diese Information ist. Allerdings ist die falsche Annahme eines tätigen Inhabers in mehreren Fällen für eine falsche Re-Identifikation verantwortlich.

In 11 Fällen wurde angenommen, dass kein Inhaber in dem jeweiligen Unternehmen mitarbeitet. Dies war zwar jedes Mal richtig, hatte aber nur eine geringe Differenzierungswirkung und ist in keinem Fall hauptverantwortlich für eine Re-Identifizierung.

Diese Ergebnisse zeigen, dass dieses Merkmal zwar theoretisch sehr bedeutsam sein kann, allerdings die Fehlerquote zu hoch ist, als dass sich ein Datenangreifer auf eine Re-Identifikation auf dieser Basis verlassen könnte. Daher erscheint dieses Merkmal höchstens dazu geeignet, zuvor getroffene Entscheidungen zu bestätigen.[17]

17) Hat z.B. ein als „identifiziert" vermutetes Unternehmen einen tätigen Inhaber angegeben und verfügt der Angreifer über diese Information, dann kann er seine Entscheidung bestätigt sehen.

- **Handelstätigkeit der Unternehmen**

Rund 8 000 Unternehmen und damit in etwa die Hälfte aller Unternehmen der KSE-Erhebung, weisen eine Handelstätigkeit auf.[18] Daraus folgt, dass der Angreifer durch das Wissen einer Handelstätigkeit, die Hälfte aller in Frage kommenden Unternehmen ausschließen kann.

Im Fallbeispiel steht dieses Merkmal elfmal zur Verfügung (davon siebenmal in der Ausprägung „ja"), zehnmal sind die Angaben korrekt. Die falsche Angabe liegt als Ausprägung „ja" vor. Tatsächlich benutzt wurde das Merkmal aber nur achtmal, da bei den großen Unternehmen, mit mehr als 1 000 Beschäftigten die Unternehmen entweder bereits zuvor eindeutig re-identifiziert wurden, oder aber die in Frage kommenden restlichen Unternehmen alle einen Handelsumsatz aufwiesen.

Da dieses Merkmal nur geringfügig differenziert, ist der Nutzen für einen Datenangreifer nur beschränkt. Allerdings ist das Merkmal relativ zuverlässig im Zusatzwissen auffindbar, wodurch es dem Datenangreifer bei der endgültigen Festlegung, welcher Datensatz in der KSE dem gesuchten Unternehmen entspricht, helfen kann.

- **Forschungstätigkeit**

Etwa 4 700 Unternehmen der KSE-Erhebung geben an, Ausgaben für FuE zu tätigen. Wie beim Handel steigt auch bei diesem Merkmal die Wahrscheinlichkeit für Forschungstätigkeit mit der Größe des Unternehmens. Daher ist dieses Merkmal besonders wertvoll, wenn es bei kleinen Unternehmen in der Ausprägung „ja" vorliegt.[19] Bei publizitätspflichtigen Unternehmen ist in manchen Fällen eine ungefähre Abschätzung des Volumens der Forschungstätigkeit ableitbar. Dies kann den Wert einer Information entscheidend erhöhen.

Insgesamt liegt im Fallbeispiel 19 Mal eine Information über die Forschungstätigkeit des Unternehmens vor. Da es wesentlich einfacher ist herauszufinden, dass ein Unternehmen forscht, verwundert es nicht, dass in 17 Fällen die Information in der Ausprägung „ja" vorliegt. Von diesen 17 Fällen sind 14 richtig. Die beiden Fälle, in der die Ausprägung „nein" vorliegt, sind ebenfalls richtig. Einschränkend muss zu diesem „guten" Ergebnis gesagt werden, dass in 9 der 17 richtigen Fälle mit der Ausprägung „ja" die jeweiligen Unternehmen mehr als 1 000 Beschäftigte haben, ein Bereich, in der durch die Publizitätspflicht der Unternehmen relativ leicht die Forschungstätigkeit abgeschätzt werden kann, in der aber der Großteil aller Unternehmen forscht. Diese Informationen haben daher nicht den Wert wie bei kleineren Unternehmen. In 4 Fällen liegt die Information eines „hohen" Forschungs- und Entwicklungsaufwandes vor. In zwei Fällen ist dies sogar die für die korrekte Re-Identifikation entscheidende Aussage.

18) Dabei nimmt die Wahrscheinlichkeit für Handelstätigkeit mit der Größe der Unternehmen zu. So haben von 756 Unternehmen mit mehr als 1 000 Beschäftigten lediglich 181 keine Handelstätigkeit. Dagegen stehen den 3 652 kleinen Unternehmen (< 100 Beschäftigten) die Handel betreiben 5 761 Unternehmen gegenüber, die keinen Handel betreiben.

19) Nur ca. 1 500 der knapp 10 000 kleinen Unternehmen geben Forschungstätigkeiten an.

6 Auswirkung von Anonymisierungsmaßnahmen auf die Möglichkeit der Re-Identifikation

Tabelle 5 zeigt die Anzahl der „erfolgreichen" Re-Identifikationsversuche bei verschieden anonymisierten KSE-Datensätze insgesamt und nach Größenklassen aufgeteilt. Den datenveränderten Verfahren ist gemeinsam, dass sie ihre größte Wirkung bei den großen Unternehmen entfalten. Kleinere Unternehmen werden dagegen weniger stark anonymisiert. Diese Unternehmen haben jedoch aufgrund ihrer geringen Größe bereits einen signifikanten Schutz (vgl. Abschnitt 4) und müssen daher weniger stark anonymisiert werden. Ebenfalls ist bei allen Verfahren aufgrund ihrer Anwendung eine größere Unsicherheit bei der Re-Identifikation zu verzeichnen. Dies erhöht die Sicherheit der Datensätze zusätzlich, auch wenn es nicht in der Tabelle zum Ausdruck kommt. Im Folgenden werden die Mikroaggregationsverfahren (inkl. SAFE) separat von den „Tauschverfahren" (Rankswapping / LHS) betrachtet.

Tabelle 5: Vergleich der Anzahl der erfolgreichen Re-Identifikationen

Nr.	Datensatz [1]	richtige Identifikationen				
		insgesamt	klein [2]	mittel [3]	groß [4]	sehr groß [5]
–	Original	19	1	5	9	4
a	Mikroaggregation	19	1	5	9	4
b	Mikro Variante 1	14	0	4	6	4
c	Mikro Variante 2	12	1	3	5	3
d	SAFE	15	1	3	7	4
e	Mikro PRAM	13	1	3	6	3
f	Rankswapping	7	1	3	2	1
g	LHS	6	0	0	5	1

1) Zu den verschiedenen Datensätze vgl. Abschnitt 2.
2) Weniger als 100 Beschäftigte.
3) 100 – 999 Beschäftigte.
4) 1 000 – 4 999 Beschäftigte.
5) Über 5 000 Beschäftigte.

Quelle: eigene Darstellung

6.1 Mikroaggregationsverfahren

Die 5 untersuchten KSE-Datensätze, bei denen die stetigen Merkmale mit der Mikroaggregation behandelt wurden, unterscheiden sich lediglich in der jeweiligen Anonymisierung der diskreten Merkmale. Die jeweils unterschiedliche Schutzwirkung, ist daher auf Unterschiede bei den diskreten Merkmalen zurückzuführen. Das gewählte Mikroaggregationsverfahren behandelte jedes stetige Merkmal separat (vgl. Höhne 2003). Dadurch wurde die „Veränderung" der Werte auf ein Minimum reduziert, was sich sehr gut auf den Erhalt des Analysepotenzials ausgewirkt hat (vgl. Rosemann 2003). Da der Datensatz relativ gut erhalten bleibt, ist aber die Schutzwirkung eher gering, so dass kein zusätzliches Unternehmen aufgrund der Mikroaggregation geschützt wird. Bei den großen Unternehmen wird die Schutzwirkung dadurch erhöht, dass ihre Werte am stärks-

ten verändert werden. So hat ein Datenangreifer zwar immer noch die Möglichkeit die großen Unternehmen zu re-identifizieren, er muss aber damit rechnen, dass die Werte die er erhält von den „wahren" Werten abweichen, wodurch sich die Information als unbrauchbar erweist. Im konkreten Fall sind von den acht Informationen (Umsatz und Beschäftigte bei den vier sehr großen Unternehmen) zwei unbrauchbar (mit Abweichungen über 10 %), drei nur bedingt brauchbar (Abweichungen deutlich unter 10 %) und drei Werte brauchbar (keine Abweichungen). Welche Informationen dies sind, muss der Datenangreifer abschätzen, was für ihn ein zusätzliches Risiko darstellt (vgl. Höhne; Sturm; Vorgrimler 2003).

Bei der zusätzlichen Anonymisierung der diskreten Merkmale zeigt sich, dass durch den Verzicht auf den BBR9 eine etwas höhere Schutzwirkung erzielt wird (**Mikroaggregation Variante 2**) als bei der Reduzierung der Wirtschaftsklassifikation auf zwei Stellen (**Mikroaggregation Variante 1**). Das liegt darin begründet, dass der BBR 9 in allen Fällen richtig vorliegt, während nicht alle Unternehmen richtig vierstellig klassifizierbar waren. Gelang im Originaldatensatz eine Re-Identifizierung auf der Basis einer zweistelligen Wirtschaftsklassifikation, so wurde dieser Datensatz durch eine Reduzierung der Wirtschaftsklassifikation auf zwei Stellen nicht zusätzlich geschützt. Im Gegensatz dazu bedeutet es für alle Unternehmen einen zusätzlichen Schutz, wenn auf den Ausweis des BBR 9 verzichtet wird.

Beim Verfahren **„SAFE"** werden neben den stetigen auch die diskreten Merkmale anonymisiert, wenn durch die Kombination der diskreten Merkmale Zellen entstehen, die mit weniger als drei Unternehmen besetzt sind (vgl. Höhne 2003). Im Extremfall bedeutet dies, dass die diskreten Merkmale überhaupt nicht behandelt werden und zwar dann, wenn alle Zellen mit mindestens drei Unternehmen besetzt sind. Durch diese Bedingung werden in diesem Verfahren die diskreten Merkmale am „schwächsten" anonymisiert. Daher verwundert es nicht, dass die verwendete Variante „SAFE" in dieser Simulation den geringsten Schutz bietet.[20)]

Beim Verfahren **PRAM**[21)] werden die Ausprägungen der diskreten Merkmale mit einer vorgegebenen Wahrscheinlichkeit verändert. Beim probenanonymisierten Zieldatensatz wurden die Ausprägungen der Merkmale BBR 9 und WZ 93 mit 20 % Wahrscheinlichkeit verändert. D.h., hatte ein Merkmalsträger vor der Anwendung des Verfahrens eine Ausprägung beim BBR 9 von 6, so hat er diese Ausprägung nach dem Verfahren nur noch mit einer Wahrscheinlichkeit von 80 %. Da eine Bandbreite – innerhalb derer sich der Wert verändern kann – beim BBR 9 von 2 Stufen der Kategorien gewählt wurde, hat nach der Anwendung von PRAM der Merkmalsträger in diesem Beispiel mit 20 % Wahrscheinlichkeit die Ausprägung 4,5,7 oder 8. Bei der Wirtschaftsklassifizierung wurde eine Bandbreite von 20 Klassen gewählt. Ohne Kenntnis dieser Anonymisierungsmaßnahme, wirkt die Anonymisierung für einen Datenangreifer wie eine zusätzliche Inkompatibilität zwischen Zieldatensatz und Zusatzwissen. In diesem Fall kann er nur noch 13 Unternehmen eindeutig und richtig re-identifizieren (vgl. Tabelle 5).

---

20) Bei anderen im Rahmen von SAFE verwendeten Mikroaggregationsverfahren, die alle Merkmale gleichzeitig anonymisieren, wird jeder Datensatz entweder dreimal oder überhaupt nicht vorkommen. Dadurch wird eine absolute Anonymität nahezu erreicht.

21) Das Verfahren PRAM wurde mit Hilfe der Anonymsierungssoftware Mu-Argus angewendet.

Kennt ein Datenangreifer die Parameter der Anonymisierung durch PRAM, so kann er diese Information für seine Re-Identifikationsversuche verwenden. Hat er z.B. die Information, dass das gesuchte Unternehmen für den BBR 9 die Ausprägung 6 aufweist, so wird er alle Unternehmen betrachten, die eine Ausprägung zwischen 4 und 8 aufweisen. Die Anonymisierung wirkt hier nur noch als Vergröberung der Information und der Datenangreifer kennt die Wahrscheinlichkeit, mit der eine bestimmte Ausprägung im Zieldatensatz enthalten ist (vgl. Tabelle 6).

Tabelle 6: PRAM: Wahrscheinlichkeit der Ausprägung im probeanonymisierten Datensatz in Abhängigkeit zur Originalausprägung

Originalausprägung	Wahrscheinlichkeit der Ausprägung im probeanonymisierten Datensatz								
	1	2	3	4	5	6	7	8	9
1	80	10	10						
2	6,6	80	6,6	6,6					
3	5	5	80	5	5				
4		5	5	80	5	5			
5			5	5	80	5	5		
6				5	5	80	5	5	
7					5	5	80	5	5
8						6,6	6,6	80	6,6
9							10	10	80

Quelle: eigene Darstellung

Wird davon ausgegangen, dass ein Datenangreifer die Parameter der Anonymisierungsmaßnahme kennt, so kann er von den 6 zuvor nicht re-identifizierten Unternehmen 2 wieder eindeutig zuordnen. Bei zwei weiteren Unternehmen kommen jeweils zwei Datensätze in Frage, wobei jeweils ein Datensatz besser zum Zusatzwissen kompatibel ist als der andere, wodurch ein Datenangreifer diesen dem Unternehmen – allerdings unter großen Unsicherheiten – zuordnen kann. Die beiden restlichen Unternehmen konnten auch mit dem Wissen über die Anonymisierung nicht eindeutig zugeordnet werden.

6.2 Tauschverfahren

- **Rankswapping**

Beim Rankswapping werden alle Merkmale einzeln anonymisiert. Dabei werden sie zunächst sortiert und anschließend innerhalb eines angegebenen Austauschsbereichs[22] miteinander getauscht. Dieser Bereich beträgt bei dem getesteten Datensatz 1 % der benachbarten Merkmalsträger. Durch diesen Prozess hat das Verfahren eine geringere Schutzwirkung, je höher die Merkmalsdichte ist. Im dicht besetzten Bereich unter 1 000 Beschäftigten werden die Merkmalsausprägungen immer mit sehr ähnlichen getauscht, was zur Folge hat, dass der entstandene Datensatz dem ursprünglichen sehr ähnlich ist.

22) Austauschbereich: der Bereich, innerhalb derer die Ausprägungen getauscht werden. Je größer dieser Bereich, desto stärker die Anonymisierung und desto größer der Informationsverlust.

Aus diesem Grund waren die Re-Identifikationsversuche bei den kleinen und mittleren Unternehmen relativ erfolgreich. Je größer die Unternehmen werden, desto stärker werden ihre Merkmalsausprägungen durch die Anwendung des Verfahrens verändert. Dadurch erklärt sich die hohe Schutzwirkung bei den Unternehmen mit mehr als 1 000 Beschäftigten. Die Datensätze der 4 Unternehmen mit mehr als 5 000 Beschäftigten wurden durch das Verfahren so stark verändert, dass der Nutzen des einen re-identifizierten Unternehmens gleich Null ist.

Rankswapping bietet gerade bei den großen Unternehmen einen sehr hohen Schutz. Ein faktisch anonymer Datensatz kann als fast erreicht bezeichnet werden. Einzig kritischer Bereich sind die Unternehmen mit 250 bis 1 000 Beschäftigten. Der zu Beginn dieses Abschnitts erwähnte Effekt, dass die großen Unternehmen geschützt werden, trifft für das Rankswapping in besonderer Weise zu.

- **Latin Hypercube Sampling (LHS)**

Liegen die Möglichkeiten der richtigen Zuordnung für einen Datenangreifer beim Rankswapping bei den kleineren und mittleren Unternehmen, sind es beim LHS die größeren Unternehmen, die theoretisch re-identifiziert werden können. Dies liegt zum einen an der Nichtbehandlung der diskreten Merkmale und zum anderen daran, dass für die großen Unternehmen bei den stetigen Merkmale bereits ungefähre Werte zur Re-Identifikation ausreichen. Allerdings stehen im Fallbeispiel den insgesamt 6 erfolgreichen Versuchen auch 3 gegenüber, die zu eindeutigen jedoch falschen Zuordnungen führten. Dies verdeutlicht, dass durch das Verfahren das Risiko des Datenangreifers, eine falsche Re-Identifikation als richtig anzunehmen, erhöht wird. Des Weiteren liegt der zusätzliche Schutz, den das Verfahren liefert, darin, dass die enthüllten Merkmalswerte sehr weit von den „wahren" Werten entfernt sind. So ist bei allen 7 Fällen zumindest eines der Merkmale „Beschäftigte" und „Umsatz" derartig verfälscht, dass die Information als unbrauchbar eingestuft werden muss. Allerdings weiß der Datenangreifer nicht, welche Werte unbrauchbar sind, sondern nur, dass eine für ihn große Risiko besteht, unbrauchbare Informationen zu erhalten. Dies alles führt beim LHS zum Fazit, dass zwar innerhalb besonderer Konstellationen (bei den größeren Unternehmen) richtige Zuordnungen möglich sind, jedoch aufgrund der zusätzlichen Risiken, die ein Datenangreifer eingeht und der z.T. stark verfälschten Informationen, der Datensatz als sicher einzustufen ist.

7 Fazit

Die Simulationen haben gezeigt, dass das Re-Identifikationsrisiko sehr stark von der Unternehmensgröße abhängig ist. Bis zu einer Größe von 250 Beschäftigten ist die Wahrscheinlichkeit einer richtigen Zuordnung sehr gering. Bis ca. 2 000 Beschäftigten setzt sich ein Datenangreifer immer noch der Gefahr von Falschzuordnungen aus. Im Gegensatz dazu waren die sehr großen Unternehmen relativ leicht zu re-identifizieren. Daraus lässt sich folgern, dass bei der Wahl der Anonymisierungsmethode die Größe der zu schützenden Unternehmen beachtet werden sollte.

Des Weiteren hat sich gezeigt, das Inkompatibilitäten zwischen Zusatzwissen und Zieldatensatz die Möglichkeiten einen Datensatz zu re-identifizieren deutlich verringern. Die meisten Fehlversuche sind auf solche Inkompatibilitäten zurückzuführen. Durch Anonymisierungsmaßnahmen lassen sich die Möglichkeiten der Re-Identifikation weiter ver-

ringern. Bei den Mikroaggregationsverfahren ist dabei der Schutz davon abhängig, welche und wie viele diskrete Merkmale der veröffentlichte Datensatz enthält. Bei den Tauschmethoden (Rankswapping / LHS) wird eine sehr hohe Sicherheit des Datensatzes gewährleistet. Bei allen Methoden steigt für einen Datenangreifer die Unsicherheit, inwieweit die erzielte Zuordnung richtig ist.

Die Simulation der Einzelangriffe hat gezeigt, dass es möglich ist, einen faktisch anonymem KSE-Datensatz zu erstellen. Inwieweit dieses Ergebnis im Rahmen eines Massenfischzuges bestand hat, müssen weitere Arbeiten zeigen. Auch muss untersucht werden, wie stark das Analysepotenzial durch die Anonymisierungsmethoden beeinträchtigt wird.

Literaturhinweise

Elliot, M.; Dale, A. (1999): Scenarios of attack: the data intruder's perspective on statistical disclosure risk. In: Netherlands Official Statistics, S. 6 – 10.

Opfermann, R.; Hennchen, O.; Czarkowski, P. (2002): Beschreibung der Kostenstrukturerhebung im Verarbeitenden Gewerbe; internes Papier im Anonymisierungsprojekt.

Höhne, J. (2003): Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten, Beitrag auf dem Anonymisierungsworkshop am 20.03.2003 in Tübingen. (In diesem Band S. 69 ff.)

Höhne, J.; Sturm, R.; Vorgrimler, D. (2003): Konzept zur Beurteilung der Schutzwirkung faktischer Anonymisierung, in: Wirtschaft und Statistik, Heft 4, S. 287 – 292.

Rosemann, M. (2003): Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik, Beitrag auf dem Anonymisierungsworkshop am 21.03. 2003 in Tübingen. (In diesem Band S. 154 ff.)

Statistisches Bundesamt (o. J.): Kostenstrukturerhebung – Erläuterungen zum Erhebungsvordruck.

Voy, K. (2002): Weiterentwicklung in der amtlichen Unternehmensstatistik – Der Unternehmensbegriff. In: Statistisches Bundesamt (Hrsg.): Unternehmen in der Statistik, Forum der Bundesstatistik, Bd. 39, S. 68 – 94.

Wirth, H. (2003): Szenarien für Angriffe auf wirtschaftsstatistische Einzeldaten – Ein Überblick, Beitrag auf dem Anonymisierungsworkshop am 20.03.2003 in Tübingen. (In diesem Band S. 11 ff.)

Anhang

Merkmale der Projektdaten „KSE-Erhebung"

1. Wirtschaftszweig (WZ 93)
2. Regionalbezug (BBR-Schlüssel, sog. „Neuner-Kategorie")
3. Beschäftigtengrößenklasse
4. Tätige Inhaber
5. Angestellte und Arbeiter
6. Teilzeitbeschäftigte
7. Teilzeitbeschäftigte umgerechnet in Vollzeiteinheiten
8. Tätige Personen insgesamt
9. Umsatz aus eigenen Erzeugnissen
10. Umsatz aus Handelsware
11. Gesamtumsatz (entspricht nicht der Summe aus 9. und 10.)
12. Anfangsbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen am Umsatz aus eigenen Erzeugnissen
13. Endbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen am Umsatz aus eigenen Erzeugnissen
14. Bestandveränderung an unfertigen/fertigen Erzeugnissen
15. Gesamtleistung/Bruttoproduktionswert
16. Anfangsbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und Betriebsstoffen gemessen am Umsatz aus eigenen Erzeugnissen
17. Endbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und Betriebsstoffen gemessen am Umsatz aus eigenen Erzeugnissen
18. Verbrauch an Rohstoffen
19. Energieverbrauch
20. Anfangsbestand an Handelsware gemessen am Umsatz aus Handelsware
21. Endbestand an Handelsware gemessen am Umsatz aus Handelsware
22. Einsatz an Handelsware
23. Bruttogehalts- und -lohnsumme
24. Gesetzliche Sozialkosten
25. Sonstige Sozialkosten
26. Kosten für Leiharbeitnehmer
27. Kosten für Lohnarbeiten
28. Kosten für Reparaturen
29. Mieten und Pachten
30. Sonstige Kosten
31. Fremdkapitalzinsen
32. Kosten insgesamt
33. Bruttowertschöpfung zu Faktorkosten
34. Nettowertschöpfung zu Faktorkosten
35. Gesamtaufwendungen für innerbetriebliche Forschung und Entwicklung
36. Anzahl der für Forschung und Entwicklung eingesetzten Lohn- und Gehaltsempfänger

Uwe Blien *)

Die Re-Identifikationsproblematik bei wirtschaftsstatistischen Einzeldaten

Einige allgemeine Gesichtspunkte in der Diskussion der Beiträge von Daniel Vorgrimler, Rolf Wiegert und Heike Wirth

1 Einführung

Welche Besonderheiten sind bei der Beurteilung der Anonymität wirtschaftsstatistischer Daten zu beachten? In diesem Papier erfolgt eine knappe Diskussion an Hand wichtiger Aspekte der Papiere von Daniel Vorgrimler, Rolf Wiegert und Heike Wirth (alle in diesem Band). Diese werden vor den Hintergrund eines allgemeineren Kontexts gestellt.

Diskussionsbasis ist die Definition faktischer Anonymität gemäß §16 Bundesstatistikgesetz. Ein Datenbestand ist dann faktisch anonym, wenn der für die Deanonymisierung notwendige Aufwand unverhältnismäßig groß wird. Ein Datenangreifer muss diesen Aufwand betreiben, um einen Datensatz in einem Mikrodatenfile einer Firma mit bekannter Identität zuzuordnen. Bei der Abwägung der „Unverhältnismäßigkeit" wird unterstellt, dass der Angreifer rational vorgeht.

In dem Papier von Heike Wirth werden Angriffsszenarien auf wirtschaftsstatistische Einzeldaten vor dem Hintergrund dieses unterstellten Rationalkalküls eines potentiellen Datenangreifers untersucht. Daniel Vorgrimler simuliert Re-Identifikationsversuche mit Original- und mit durch verschiedene Verfahren anonymisierten Daten. Rolf Wiegert schließlich ist um die Diskussion der Zuordnungsmethode bemüht. Beginnen wir allgemein mit dem Zuordnungsproblem und arbeiten uns dann zu den konkreteren Fragestellungen vor.

2 Das Zuordnungsproblem

Auf welche Weise wird ein Datenangreifer eine Zuordnung vornehmen? Zunächst könnte man sich vorstellen – und Rolf Wiegert geht ebenfalls davon aus – dass ein Datenangreifer für zwei Datenbestände den Weg des einfachen Matching wählt. Der eine Datenbestand spielt dabei die Rolle des anonymisierten Mikrodatenfiles, der andere jene des nicht anonymen Identifikationsfiles. Beide Datenbestände haben eine bestimmte Anzahl von Variablen gemeinsam und der Datenangreifer muss zumindest Grund zu der Vermutung haben, dass auch zumindest einige Firmen, auf die sich die Daten beziehen, in beiden Files gemeinsam enthalten sind. Die einfachst mögliche Form des Re-Identifikationsversuchs ist dann der Abgleich beider Datenbestände auf identische Merkmale.

Die beschriebene Vorgehensweise entspricht jener von Vorgrimler. Er weist darauf hin, dass dabei das Problem „statistischer Zwillinge" ignoriert wird. Es könnten Fälle auftreten, die verschiedene Identitäten haben, aber gleiche Merkmalsausprägungen. Tatsäch-

*) Uwe Blien, Institut für Arbeitsmarkt- und Berufsforschung (IAB), Nürnberg.

lich erhält er in seinen Versuchen eine Reihe von Fehlzuordnungen, die durch diese Konstellation bedingt sind. Die Möglichkeit der Fehlzuordnung trägt erheblich zum Schutz der Daten bei.

Geht der Angreifer davon aus, dass die Daten in beiden Fällen nicht identisch abgebildet werden, wird er u.U. eine Verfahrensweise einschlagen, die tolerant gegen Abweichungen ist. Er geht dann vom einfachen zum statistischen Matching über. Betrachten wir das Zuordnungsproblem an Hand von Abbildung 1 etwas genauer. Da hier zur Demonstration angenommen wird, dass nur zwei metrische Variablen A und B in beiden Files gemeinsam enthalten sind, lassen sich die Fälle (Records, Datensätze) in ein flächiges Diagramm eintragen. Man kann annehmen, dass Datensätze, die in beiden Datenbeständen identische Werte aufweisen, auch von der gleichen Firma stammen. Dies ist jene Situation die in der Abbildung mit (2) bezeichnet ist. Finden sich solche identischen Paare im Wege des deterministischen Matching, kann der Datensatz als „identifiziert" betrachtet werden. Ein derartiges Verfahren kann sehr einfach realisiert werden, z.B. mit der Prozedur Proximities im Programmpaket SPSS.

Abbildung 1
Das Zuordnungsproblem bei zwei metrischen Merkmalen

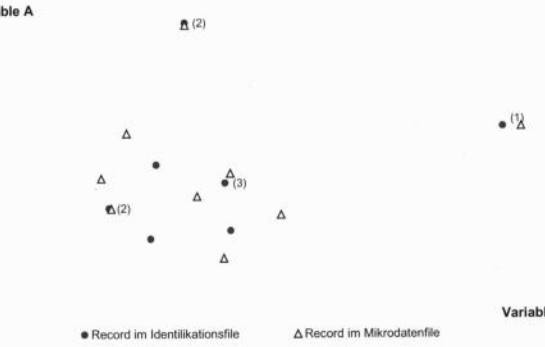


Abbildung 1 erlaubt eine Veranschaulichung verschiedener Strategien eines Datenangreifers und der Probleme, die das Re-Identifikationsvorhaben ihm bereitet. Für die Mehrzahl der Datensätze mit kleinen Werten bei beiden Variablen ist eine Zuordnung offensichtlich nicht sinnvoll. Hier liegen mehrere Fälle im Mikrodatenfile nicht weit von solchen des Identifikationsfiles. Der Augenschein spricht jedoch dagegen, dass sie Paare sind, die von jeweils einer Firma stammen.

Eine Ausnahme von dieser Aussage betrifft ein Paar von Merkmalskombinationen, das nur einen geringen internen Abstand aufweist. Kann man in diesem Fall erwarten, der in der Abbildung 1 mit (3) bezeichnet wurde, dass beide Datensätze die gleiche Firma beschreiben? Zum Vergleich sei ein anderes Paar herangezogen, das ganz rechts im Diagramm liegt und durch (1) markiert wurde. Auch in diesem Fall tritt eine Abweichung zwischen den Merkmalswerten im Identifikations- und jenen des Mikrodatenfiles auf. Im Unterschied zu dem anderen gefundenen Paar liegen jedoch alle anderen Datensätze weit entfernt. Hier wird ein Datenangreifer mit einer viel höheren Sicherheit davon ausgehen können, dass er richtig „identifiziert" hat, wenn er beide Datensätze zuordnet.

Man kann also folgern, dass die Schwierigkeit korrekter Identifikation u.a. von zwei Parametern abhängt, vom Abstand im Merkmalsraum, den zwei Datensätze zueinander und zu anderen Datensätzen haben.

Eine technisch überlegene Lösung würde darauf beruhen, dass Distanzen zwischen den Fällen berechnet und dann Fälle zugeordnet werden, die zueinander die nächsten Nachbarn sind oder innerhalb einer bestimmten, als kritisch betrachteten Distanz liegen. Auf diese Weise gehen z.B. Bacher, Brand, Bender (2002) vor. Derartige Methoden sind mit vielen Standard-Statistikprogrammpaketen zu realisieren, z.B. mit SPSS. Die Zuordnung kostet nicht viel Zeit und Energie. Dies ist ein wichtiges Kriterium im Rahmen des Konzepts der faktischen Anonymität.

Trotzdem ist auch diese Art der Zuordnung problematisch, da sie erneut nicht berücksichtigt, ob zugeordnete Paare im Merkmalsraum isoliert sind. Werden mehr als zwei Variablen für die Zuordnung verwendet, stehen graphische Methoden nicht zur Verfügung. Ein potentieller Datenangreifer könnte sich nicht sicher sein, dass er auch die richtigen Fälle zugeordnet hat. Die Methode von Bacher, Brand, Bender (2002) schließt kein Verfahren ein, Fehlzuordnungen von richtigen Matches zu unterscheiden.

Zudem müsste der Angreifer bei seiner Vorgehensweise berücksichtigen, ob die Daten aus einer Stichprobe kommen oder ob sie die Grundgesamtheit repräsentieren. Die Situation von Abbildung 1 wäre ganz unterschiedlich zu beurteilen, wenn die eingezeichneten Fälle nicht die Population umfassen, sondern eine 1 % Stichprobe. Dann wäre die tatsächliche Verteilung 100 Mal dichter als die beobachtete. Alle Schlüsse des Angreifers wären mit viel höheren Unsicherheiten belastet.

Zur Beurteilung der Wahrscheinlichkeit einer korrekten Zuordnung müsste ein stochastischer Ansatz verwendet werden. Ziel wäre die Beurteilung der Wahrscheinlichkeit, dass die beiden Datensätze von einer identischen Firma stammen. Zu diesem Zweck schlägt Wiegert den Ansatz von Fellegi und Sunter (1969) vor, worauf noch einzugehen sein wird. Statt dessen könnte man auf die Arbeiten von Paaß (Paaß 1988; Paaß, Wauschkuhn 1985) zurückgreifen, der eine stimmige, bis heute maßgebliche Theorie des Re-Identifikationsproblem entwickelt hat, die die Wahrscheinlichkeit einer korrekten Zuordnung im Rahmen eines Bayes-Konzepts beschreibt.

Anschaulich kann man sich den Ansatz so verdeutlichen, dass zunächst beurteilt wird, wie wahrscheinlich es ist, dass ein Firma mit einer bestimmten Merkmalskombination P(k) in einem File vorkommt. Anschließend muss für jedes File ein Fehlerprozess spezifiziert werden, in dem zum Beispiel auf Erkenntnisse der empirischen Sozialforschung zur

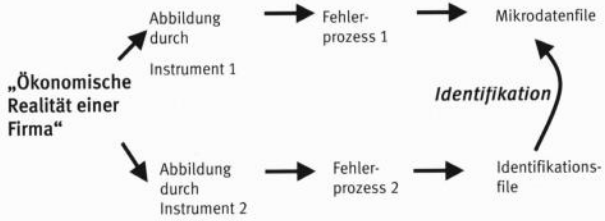
Reliabilität von Daten zurückgegriffen wird. Die Definition der Fehlerprozesse erlaubt die Beurteilung, mit welcher Wahrscheinlichkeit zwei unterschiedliche Datensätze von der gleichen Firma stammen, in dem folgender Bayes-Ansatz verwendet wird:

$$P(k|z) = \frac{P(z|k)P(k)}{\sum P(z|m)P(m)}$$

Dabei ist $P(k|z)$ die a posteriori Wahrscheinlichkeit, dass der Datensatz des Mikrodatenfiles mit den Merkmalen k von der gleichen Firma kommt wie der Datensatz des Identifikationsfiles mit den Merkmalen z. m bezeichnet andere Datensätze im File und $P(z|k)$ ist die entsprechende a priori Wahrscheinlichkeit.

Das eigentliche technische Problem ist die Abschätzung der bedingten Wahrscheinlichkeit $P(z|k)$, in der Aussagen über die Wahrscheinlichkeit enthalten sind, dass die Daten mit einem bestimmten Fehler abgebildet werden. Paaß und Wauschkuhn beurteilten diese Wahrscheinlichkeiten als relativ niedrig. Das volle Problem kann jedoch Abbildung 2 entnommen werden. Die Daten der beiden Files unterscheiden sich nicht nur wegen der beiden Fehlerprozesse, sondern auch aus anderen Gründen, weswegen im Mannheimer Anonymisierungsprojekt (Müller, Blien, Knoche, Wirth et al. 1991) der umständliche Begriff der „Dateninkompatibilitäten" geprägt wurde.

Abbildung 2
Das Zuordnungsproblem bei Dateninkompatibilitäten



Die Daten im Identifikations- und im Mikrodatenfile entstammen in der Regel ganz unterschiedlichen Quellen. Für sie gelten differierende Abbildungsregeln: Firmen antworten anders, wenn sie von der offiziellen Statistik gefragt werden („Instrument 1" in der Abbildung) als in Untersuchungen, z.B. der Marktforschung („Instrument 2"). Das Fragebogendesign unterscheidet sich jeweils, weitere Unterschiede treten auf.

Wichtig ist vor allem, dass über den Zusammenhang der Resultate von Instrument 1 und von Instrument 2 empirisch ermittelte Aussagen nur schwer möglich sind, da dem Datenangreifer, der die Methode von Paaß benutzt, kein File zur Verfügung steht, das die

Information aus dem Identifikations- und dem Mikrodatenfile zusammen enthält. Eine solche Trainingsstichprobe müsste er aber haben, um die Methode eichen zu können. Aus diesen Gründen – und wegen ihrer Komplexität – ist die Methode von Paaß eher für Laborexperimente geeignet als dass sie ein realer Angreifer anwenden könnte.

3 Verschiedene Matchingmethoden u.a. aus dem Bereich des Record Linkage

Bisher wurden drei verschiedene Methoden der Deanonymisierung identifiziert, die auch in der Literatur behandelt werden:

- Einfacher Datenbankabgleich (deterministisches Matching, vgl. Vorgrimler).
- Suche nach ähnlichen Datensätzen (suche nach dem nächsten Nachbarn im anderen File).
- Statistisches Matching (Bayes'scher Ansatz oder ein ähnlich komplexes Verfahren).

Diese Methoden sind für Tests der Datenanonymität zugänglich, wie im Mannheimer Anonymisierungsprojekt zumindest für die erste und die dritte demonstriert wurde. Heike Wirth weist darauf hin, dass für die Abschätzung der Re-Identifikationsmöglichkeit amtlicher Wirtschaftsdaten das in bestimmten Diskussionen unter Statistikern verbreitete Uniquenesskonzept (Bethlehem, Keller, Pannekoek 1990) kaum eine Rolle spielt. Bei diesem Konzept werden alle im Mikrodatenfile oder in der Grundgesamtheit einzigartigen Datensätze als potentiell identifizierbar betrachtet. Dabei wird jedoch das Auftreten von Dateninkompatibilitäten ignoriert. In Abbildung 1 sind alle Records des Mikrodatenfiles einzigartig, ohne dass in den meisten Fällen eine Zuordnung möglich wäre.

Rolf Wiegert diskutiert verschiedene Möglichkeiten, das oben benannte Methodenspektrum für Deanonymisierungsversuche zu erweitern. Er setzt sich vor allem mit zwei Verfahrensklassen auseinander, mit den Techniken des Record Linkage, die u. a. von Fellegi und Sunter (1969) und Nachfolgern entwickelt wurden und mit einer einfacheren Methode der Intervallschachtelung.

Verfahren des Record Linkage wurden entwickelt, um Datensätze zusammenzuführen, für die direkte Identifikatoren vorliegen, d.h. z.B. Namen, Adressen, Sozialversicherungsnummern etc. Diese Identifikatoren sind jedoch häufig abweichend angegeben und es treten Fehler auf. Die Leistung der Verfahren von Fellegi und Sunter und Nachfolgern besteht nun darin, die Wahrscheinlichkeit abzuschätzen, dass Datensätze aus zwei verschiedenen Files von der gleichen Firma stammen. Oberhalb einer bestimmten Schwelle wird angenommen, dass Identität gegeben ist, unterhalb einer zweiten Schwelle geht man von Nicht-Identität aus. Dazwischen liegt ein „Graubereich", innerhalb dessen ein Experte entscheiden muss, ob ein Match vorliegt oder nicht.

Auf den ersten Blick macht die Methode relativ schwache Voraussetzungen, da keine Annahmen über Fehlerprozesse in den beiden Files notwendig sind. Trotzdem benötigt der Datenangreifer keine Trainingsstichprobe, d.h. Fälle, von denen Daten in beiden Files enthalten sind und die er identifizieren kann. Die schwachen Voraussetzungen sind eine große Überraschung: Wie gelingt die „Quadratur des Kreises"? Die Lösung basiert darauf, dass mehrere Variablen zur Verfügung stehen, deren Fehler voneinander unabhängig sind. In diesem Fall können Zuordnungen über mehrere Variablen hinweg vorge-

nommen und optimiert werden. Wiegert weist darauf hin, dass dieser Abschätzungsprozess in jüngerer Zeit typischerweise mit dem EM-Algorithmus vorgenommen wird (vgl. dazu Jaro 1989).

Allerdings kann die Unabhängigkeitsannahme auch zu falschen Abschätzungen von Wahrscheinlichkeiten führen. Da sie relativ häufig verletzt sein wird, sind die Voraussetzungen der Methode doch nicht so unproblematisch, wie es zunächst den Anschein hatte. Belin und Rubin (1995) zeigen, dass die Verletzung der Unabhängigkeitsannahme zu stark optimistischen Annahmen über die Häufigkeit von Falschzuordnungen führt.

Ob Methoden des Record Linkage tatsächlich ein geeignetes Instrument für einen Datenangreifer darstellen könnten, ist demnach unklar. Wiegert weist auf einige Probleme neben der Unabhängigkeitsannahme hin. Demnach ist z.B. eine zusätzliche Schwierigkeit, dass die Verfahren für Variablen mit nominalem Skalenniveau entwickelt wurden und die Übertragung auf metrische Variablen nicht so einfach gelingt.

In der Beurteilung von Wiegerts Ausführungen ist festzuhalten, dass ihm der Verdienst zukommt, auf die Methoden des Record Linkage hingewiesen zu haben. Allerdings ist unklar, welche Leistungen derartiger Techniken im vorliegenden Zusammenhang erbringen können. Tatsächlich wäre es empfehlenswert, ein Projekt durchzuführen, bei dem unter realistischen Bedingungen abgetestet wird, in wie weit sich die Methoden für einen Datenangriff eignen. Hier scheint eine Forschungslücke zu bestehen.

Das zweite von Wiegert diskutierte Verfahren, das er Methode der „Intervallschachtelung" nennt, ist für Variablen mit Intervallskalenniveau gedacht, die – wie alle empirische Daten – Fehler aufweisen. Der Angreifer matcht alle Datensätze im Mikrodatenfile mit jenen des Identifikationsfiles, die für eine Variable innerhalb einer Toleranzschwelle äquivalente Werte aufweisen. Anschließend nimmt er sich für die zugeordneten Datensätze die nächste Variable vor. Auf diese Weise werden die Variablen einzeln und sequentiell abgearbeitet.

Wiegert verspricht sich von Experimenten mit dieser Methode eine Beurteilung des Deanonymisierungsrisikos von Wirtschaftsdaten. In der Tat ist das vorgeschlagene Verfahren allgemein einsetzbar, nicht einmal ein Statistikprogramm ist nötig. Trotzdem ist es in bestimmtem Grade fehlertolerant, anders als der einfache Datenbankabgleich. Offensichtlicher Nachteil gegenüber einer Zuordnung durch Distanzminimierung ist jedoch, dass eine bestimmte Abweichung in einer Variablen, z.B. durch eine Fehlkodierung, bereits zum Ausschluss aus den zugeordneten Sätzen führt. Gemeinsamkeit der Verfahren der Intervallschachtelung und der Distanzminimierung ist, dass der Datenangreifer keinen Aufschluss darüber erhält, mit welcher Wahrscheinlichkeit seine Zuordnungen richtig sind. Dazu sei an die Anmerkung von Heike Wirth erinnert, nach der bei den Experimenten von Bacher et al. überwiegend Fehlzuordnungen auftreten (vgl. die Fußnote 21) im Beitrag von Heike Wirth).

Schließlich sei zu den Ausführungen von Wiegert noch angemerkt, dass er als Identifikationsfile die nicht-anonymen Originaldaten verwendet. Diese dienen als Ersatz für mögliches Zusatzwissen. Betrachtet man Abbildung 2, wird klar, dass diese Vorgehensweise unterstellt, dass potentielles Zusatzwissen den gleichen Datengenerierungsprozess durchläuft, wie das angezielte Mikrodatenfile. Es unterstellt weiterhin, dass die Reliabilität der Mikrodaten genau 1 beträgt. Dieses Kriterium ignoriert mit anderen Worten die

„natürliche Schutzwirkung", die darin besteht, dass ein potentieller Datenangreifer eben nicht auf Daten der amtlichen Statistik zurückgreifen kann. Stützt man die Beurteilung der Datenanonymität auf Zuordnungsexperimente zwischen anonymisierten und Originaldaten, verwendet man unberechtigterweise sehr harte Forderungen für Anonymisierungsmaßnahmen.

4 Besonderheiten wirtschaftsstatistischer Daten in Zuordnungsexperimenten

Das bisher gefundene Resultat bezüglich der Erfolgsaussichten von Re-Identifikationsversuchen sieht ungünstig für den Angreifer und gut für den Interessenten wirtschaftsstatistischer Daten aus. Einfache, leicht realisierbare Methoden der Deanonymisierung (z.B. der einfache Datenbankabgleich) sind anscheinend nicht erfolgsträchtig, während komplexe, potentiell erfolgsversprechende Methoden (z.B. Paaß' Diskriminanzanalytische Methode) in erster Linie auf Laborexperimente beschränkt sind und kaum eine wirkliche Gefahr bedeuten.

Liest man nun das sehr instruktive Papier von Vorgrimler, so wird dieser Eindruck sofort zurechtgerückt. Er verwendet die einfachste mögliche Zuordnungsmethoden des deterministischen Matching mit Zusatzwissen, das er sich aus öffentlich zugänglichen Quellen zusammengesucht hat. Von 41 willkürlich aus der Kostenstrukturerhebung der amtlichen Statistik herausgegriffenen Firmen konnten nicht weniger als 19 richtig zugeordnet werden, 12 wurden falsch zugeordnet und für 10 wurde kein Match gefunden. Die Ergebnisse hängen vor allem von der Unternehmensgröße ab: Kleine Unternehmen werden ganz überwiegend falsch, sehr große Firmen immer (!) richtig zugeordnet.

Vorgrimler unterstellt Response Knowledge, d. h. der Angreifer weiß, dass die gesuchten Firmen im Mikrodatenfile enthalten sind. Wie die Analysen der Mannheimer Studie mit Personendaten zeigten, ist dies eine sehr wichtige Bedingung, die dem Angreifer die Arbeit stark erleichtert. Doch während sie bei Personendaten i.d.R. nicht gegeben ist, trifft sie bei wirtschaftsstatistischen Daten häufig zu. Die Kostenstrukturerhebung enthält praktisch alle deutschen Großunternehmen, lediglich bei kleineren Firmen ist ihr Auswahlsatz deutlich kleiner als eins.

Das Beispiel der Arbeit von Vorgrimler zeigt, wie wichtig empirische Re-Identifikationsexperimente sind. Diese können häufig aus praktischen Gründen nicht durchgeführt werden, so dass nur wenige derartige Experimente publiziert wurden. Die spezielle Bedeutung der Versuche von Vorgrimler liegt darin, dass die Ergebnisse erheblich von jenen abweichen, die mit Personendaten erzielt worden sind. Das Risiko der Identifikation ist für große Firmen offensichtlich deutlich höher als für Personen, wenn man die parallelen Experimente aus dem Mannheimer Projekt als Vergleichsgrundlage verwendet.

Grund für diesen Unterschied ist, dass ein Teil der Firmen offensichtlich völlig „isoliert" im Merkmalsraum steht. Im Sinne der mathematischen Informationstheorie ist der Informationsgehalt des zugehörigen Datensatzes derart hoch, dass ganz wenige Variablen ausreichen, um eine Zuordnung vorzunehmen. Für Großunternehmen gilt, dass ihre Daten selbst erheblich gestört werden müssten, damit eine Deanonymisierung nicht mehr möglich wäre. Die Tests zur Schutzwirkung von Anonymisierungsverfahren zeigen, dass es einschneidender Eingriffe in den Informationsgehalt der Daten bedarf, um hier Abhilfe zu schaffen.

5 Besonderheiten von Angriffsszenarien im Falle wirtschaftsstatistischer Daten

Wirth untersucht Angriffsszenarien auf wirtschaftsstatistische Daten, indem sie das Rationalkalkül eines Datenangreifers einer näheren Analyse unterzieht. Zum Teil fallen diese Szenarien parallel zu jenen aus, die bei Personen- bzw. Haushaltsdaten zu entwickeln sind. Der Autorin ist bei ihren Ergebnissen im Allgemeinen zuzustimmen, die kenntnisreichen Analysen tragen zum Wissensstand bezüglich der Datenanonymität bei.

Bei den wenigen zu kritisierenden Details ist anzuführen, dass die Wahl der Begriffe verbessert werden könnte: Wirth unterscheidet berufliche von nichtberuflichen Re-Identifikationsmotiven. Kriterium für die Unterscheidung ist jedoch der Wissenschaftsbezug. Wenn ein Wissenschaftler im Nebenberuf z.B. Unternehmer ist, wird er aus dieser Konstellation u.U. ein spezifisches Interesse an deanonymisierten Daten entwickeln, das man als solches nicht als „nicht-beruflich" bezeichnen kann.

Wirth schließt sich Brand (2000, S. 37) insofern an, als es spezifische Motivlagen gibt, die bei Unternehmensdaten wirksam werden und diskutiert insbesondere die „Konkurrenzbeobachtung" als möglichen Anlass für Datenangriffe. Sie weist die Praktikabilität eines so begründeten Datenangriffs zurück mit dem Argument, dass die Unsicherheiten von Zuordnungen eine ernste Gefahr nicht erst aufkommen ließen. Dieses Argument basiert jedoch auf den Erfolgsaussichten von Re-Identifikationsversuchen mit Haushaltsdaten, während Vorgrimler gezeigt hat, dass für Teilpopulationen von Firmen höhere Risiken zu verzeichnen sind.

Eine einfache Folgerung aus Wirths Ergebnissen ist, dass bei der Weitergabe von wirtschaftsstatistischen Daten auf die definitive Trennung der Rolle des Wissenschaftlers von anderen beruflichen Aktivitäten zu achten ist. Institute, die in einer Neben- oder Hauptfunktion auch kommerzielle Interessen verfolgen, müssten vom Bezug der Daten ausgeschlossen werden. Allerdings stellt sich in diesem Zusammenhang die Frage, ob Datenbestände wie jene der Kostenstrukturerhebung überhaupt Informationen – insbesondere für die eher gefährdeten Großunternehmen – beinhalten, die nicht aus öffentlich zugänglichen Quellen erhältlich sind.

6 Fazit

Als Fazit sei noch einmal betont, dass der Test der Methoden des Record Linkage für die Beurteilung des Deanonymisierungsrisikos interessant wäre. Außerdem seien die Besonderheiten zusammengefasst, die wirtschaftsstatistische Daten auszeichnen, und die bei der Entscheidung über die Anonymisierung zu beachten sind:

- Spezielle Motivlagen für Datenangreifer treten auf, dies geht in das Rationalkalkül des Datenangreifers ein.
- Das Risiko der Re-Identifikation streut stärker als bei Personendaten.
- Es ist im Falle von Großunternehmen deutlich höher als bei Personendaten.

Diese Situation erschwert die Entscheidung über die zu wählenden Anonymisierungsstrategien. Insbesondere die Differenzierung des Risikos verlangt eine abgestufte Vorge-

hensweise. Denn „die im Hinblick auf das wissenschaftliche Analysepotential dieser Daten am wenigsten wünschenswerte Lösung wäre, auf Basis der risikoreichsten Datenfiles Anonymisierungsmaßnahmen zu erarbeiten" (Wirth).

Literaturhinweise

Bacher, Johann; Brand, Ruth; Bender, Stefan (2002): "Re-identifying register data by survey data using cluster analysis: an empirical study", in: International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems.

Belin, Thomas R.; Rubin, Donald B. (1995): "A Method for Calibrating False-Match Rates in Record Linkage", in: Journal of the American Statistical Association 90/430.

Bethlehem, Jelke G.; Keller, Wouter J.; Pannekoek, Jeroen (1990): "Disclosure Control of Microdata", in: Journal of the American Statistical Association, 85/409, S. 38 – 45.

Blien, Uwe; Wirth, Heike; Müller, Michael (1992): "Disclosure risk for microdata stemming from official statistics", in: Statistica Neerlandica 46/1, S. 69 – 82.

Brand, Ruth (2000): „Anonymität von Betriebsdaten. Verfahren zur Erfassung und Maßnahmen zur Senkung des Reidentifikationsrisikos", Beiträge zur Arbeitsmarkt- und Berufsforschung 237, Nürnberg: Bundesanstalt für Arbeit.

Fellegi, Ivan P.; Sunter, Alan B. (1969): "A theory for record linkage", in: Journal of the American Statistical Association 1969, 64/328, S. 1183 – 1210.

Larsen, Michael D. (1997): "Modeling Issues and the Use of Experience in Record Linkage", Proceedings of an Internationa Workshop and Exposition on Record Linkage Techniques 1997, S. 95 – 105, Arlington (VA).

Müller, Walter; Blien, Uwe; Knoche, Peter; Wirth, Heike et al. (1991): „Die faktische Anonymität von Mikrodaten" (Schriftenreihe Forum der Bildungsstatistik, Band 19, herausgegeben vom Statistischen Bundesamt), Wiesbaden: Metzler-Poeschel.

Paaß, Gerhard (1988): "Disclosure Risk and Disclosure Avoidance for Microdata", in: Journal of Business and Economic Statistics 6/4, S. 487 – 500.

Paaß, Gerhard; Wauschkuhn, Udo (1985): „Datenzugang, Datenschutz und Anonymisierung. Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten", München, Wien, R. Oldenbourg.

Winkler, William (1995): "Matching and Record Linkage", in: Cox, Brenda G. et al.(Hrsg.): "Business Survey Methods", New York etc., John Wiley and Sons.

Jörg Höhne *)

Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten

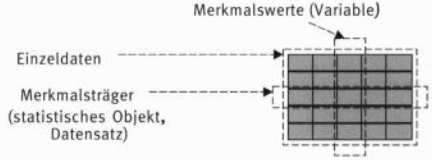
1 Einführung

Im Rahmen des Projektes „faktische Anonymisierung von wirtschaftsstatistischen Einzeldaten" sollen die Möglichkeiten der Bereitstellung von anonymisierten Einzeldaten für die Wissenschaft exemplarisch an einigen Statistiken untersucht werden. „Möglichkeiten" bedeutete dabei, dass sich das Projekt an den aktuellen methodischen und technischen Möglichkeiten orientiert. Der folgende Beitrag soll einen kurzen Überblick über die derzeitigen aktuellen Methoden geben und herausarbeiten, warum bestimmte Verfahren im Rahmen des Projektes einer näheren Untersuchung unterzogen werden.

1.1 Begriffsbestimmung

Die Definition von anonymen Einzeldaten im Rahmen der amtlichen Statistik ist aus dem „Gesetz über die Statistik für Bundeszwecke" abgeleitet. In § 16 (6) BStatG werden Daten als faktisch anonym eingestuft, wenn die Einzelangaben nur mit einem unverhältnismäßig hohen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können. Damit sind nur Re-Identifikationen zu betrachten. Unter einer Re-Identifikation wird die eindeutige und richtige Zuordnung von Einzelangaben eines Auskunftsgebenden zu seinem Identifikationsmerkmalen (Name/Adresse)[1] verstanden.

Abbildung 1: Mikrodatendatei



Unter Mikrodaten verstehen wir im Folgenden eine Datensammlung in rechteckiger Struktur mit Informationen über einzelne statistische Objekte (siehe Abbildung 1). Dabei entspricht jede Zeile der Datei einem statistischen Objekt (werden synonym auch Merkmalsträger, Datensatz oder Record genannt). Statistische Objekte sind bei wirt-

*) Jörg Höhne, Statistisches Landesamt Berlin.

1) Auch die Zuordnung zu eindeutigen Kennnummern muss damit vermieden werden. Diese erlauben, sofern sie allgemein zugänglich sind, eine unmittelbar mögliche, eindeutige Zuordnung zum Namen und zur Adresse.

schaftsstatistischen Einzeldaten in der Regel Unternehmen und Betriebe. Jede Spalte der Tabelle enthält für alle Objekte gleichartige Informationen / Werte über ein Merkmal (auch Variable genannt).

Andere Datensammlungen (z.B. hierarchische Datenbanken, nicht relationale Datenbanken, Summentabellen) werden hier nicht betrachtet. Die Merkmale/Variablen der Mikrodatendatei lassen sich in Schlüsselmerkmale und sensible Merkmale unterteilen. Schlüsselmerkmale sind Informationen über den Merkmalsträger, die auch auf anderem Wege (ohne Mikrodatendatei) für ein gesuchtes Objekt beschaffbar sind (z.B. aus Unternehmensdatenbanken, Internet u.Ä.). Sensible Merkmale sind die Informationen, die aus anderen Quellen kaum oder nicht beschaffbar sind, und wegen derer der Datenangreifer einen Angriff versucht. Dabei wird versucht, die extern beschaffte Information über die Schlüsselmerkmale (Zusatzwissen) mit dem Datenbestand abzugleichen und eine eindeutige Zuordnung (Reidentifikation) zu erreichen, die es ermöglicht, sensible Informationen abzuleiten. Schlüsselmerkmale können auch sensible Informationen sein, wenn es dem einzelnen Datenangreifer nicht gelingt, die Information extern zu beschaffen. Ein Datenangreifer könnte sich Informationen über die Branchenzugehörigkeit, die Region und die Anzahl der Beschäftigten eines Unternehmens z.B. aus Unternehmensdatenbanken oder dem Internet beschaffen (Zusatzwissen über Schlüsselmerkmale) und durch einen Datenangriff auf eine Mikrodatendatei des Verarbeitenden Gewerbes versuchen Informationen über Auslandsumsätze, Löhne usw. (sensible Merkmale) zu erhalten.

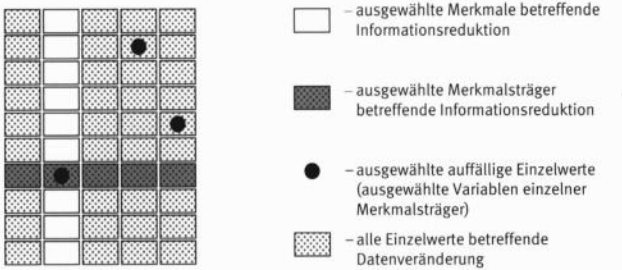
Unter Verletzung der Anonymität wird im Folgenden die nutzbringende Reidentifikation verstanden. Die Anonymität gilt dann als verletzt, wenn es dem Datenangreifer gelingt, eine eindeutige Zuordnung von bei ihm vorhandenem Wissen über das auszuspähende Objekt (Zusatzwissen über Schlüsselmerkmale) zu einem Datensatz der Mikrodaten vorzunehmen und aus diesem Datensatz nutzbringende, zusätzliche Information zu gewinnen. Diese auf dem Kosten-Nutzen-Aspekt basierende Definition von Anonymität ist erforderlich, da sich durch eine reine Beschränkung auf den technischen Begriff der Reidentifikation (eindeutige und richtige Zuordnung) ganze Gruppen von Anonymisierungsverfahren als unbrauchbar herausstellen, wenn man den Aspekt der Verhältnismäßigkeit des Aufwandes und der Qualität der gewonnenen Information vernachlässigt.

Aus dieser Herangehensweise an Anonymität ergeben sich verschiedene Effekte, durch die die Anonymisierungsansätze bei den Einzeldaten Anonymität erreichen:

- Verhinderung der eindeutigen Zuordnung von Merkmalsträgern.
- Verhinderung eines Informationsgewinns bei erfolgter Zuordnung.
- Reduzierung des Nutzens des Informationsgewinns (z.B. Unsicherheit der Information).

Alle Anonymisierungsverfahren führen Veränderungen an der Mikrodatendatei durch. Auf Grund der verschiedenen Effekte beim Anonymisieren gibt es auch verschiedene Ansätze der Veränderung der Mikrodaten (siehe Abbildung 2). Alle Ansätze lassen sich in die zwei Kategorien Informationsreduktion und Informationsveränderung (Verschweigen oder Notlüge) einteilen.

Abbildung 2: Ansätze der Veränderung von Mikrodaten



Die Bearbeitung ausgewählter Merkmale reduziert dateiübergreifend die Möglichkeit der eindeutigen Zuordnung, aber auch den potentiellen Informationsgewinn. Betrifft das hohe Deanonymisierungsrisiko nur einen kleinen Teil der Merkmalsträger, so ist auch eine gezielte Informationsreduktion für diese Merkmalsträger möglich, um sowohl die Zuordnungsmöglichkeit als auch den Informationsgewinn für diese zu reduzieren. Manchmal ist auch die Bearbeitung von nur einzelnen auffälligen Merkmalswerten ausreichend, um dieses Ziel zu erreichen. Eine alle Einzelwerte betreffende Datenveränderung (z.B. Zufallsüberlagerung) zielt in erster Linie auf eine Reduzierung des potentiellen Informationsgewinns, aber auch auf eine Reduzierung der Möglichkeit der richtigen Zuordnung.

1.2 Datenbeispiel

Die Beispielrechnungen zu den einzelnen Verfahren werden an einem einzelnem Datenbeispiel (orientiert am Monatsbericht im Verarbeitenden Gewerbe) illustriert, welches für alle Verfahren verwendet wird. Die erläuterten Beispielrechnungen sind im Anhang enthalten. Das Beispiel stellt keine übliche Grundgesamtheit dar, sondern ist für die verständliche Darstellung der Verfahren bewusst klein gewählt worden. Aussagen zur Qualität der Anonymisierungsverfahren lassen sich deshalb mit dem Beispiel nicht illustrieren.

Das Beispiel enthält einige Geheimhaltungsprobleme, die durch die verschiedenen Verfahren gelöst werden sollen. Generell können die Merkmale Region, Branche aber auch die Beschäftigtengrößenklasse als extern beschaffbare Information angesehen werden. Sie sind somit reidentifizierende Merkmale für einen Datenangriff. Die Merkmale Inlandsumsatz, Auslandsumsatz, Arbeitsstunden sowie Löhne und Gehälter stellen die schützenswerte Information dar, deren Ermittlung aus einem anonymisierten Datenbestand verhindert werden sollte.

Betrieb Nr.	Region		BGK	tätige Personen	davon Arbeiter	Umsatz	Umsatz Ausland	Arbeitsstunden	Löhne	Gehälter
		WZ93				Inland				
01	A	34	60	100	55	1 398 447	504 978	17 014	333 226	378 010
02	B	28	50	67	52	1 539 804	1 774 106	27 351	368 806	147 080
03	A	17	40	25	18	755 355	55 374	5 601	50 827	31 611
04	C	28	40	41	29	666 218	127 993	7 973	117 360	74 237
05	A	29	40	40	29	906 228	0	13 990	169 741	83 100
06	C	15	70	432	265	20 179 473	15 673	129 338	1 621 344	1 057 728
07	A	15	60	150	54	5 858 483	787 084	18 764	233 587	979 086
08	A	29	50	79	33	2 312 524	742 063	12 423	178 753	386 450
09	B	24	50	70	25	2 327 518	70 739	8 870	96 609	305 100
10	A	29	60	114	74	337 997	3 439 738	24 262	472 537	455 115
11	B	15	80	813	632	27 129 609	927 095	232 612	3 311 073	1 733 821
12	A	29	50	62	27	2 689 396	1 384 067	10 128	203 667	369 226

WZ93– Klassifikation der Wirtschaftszweige
BGK – Beschäftigtengrößenklasse
Hierbei handelt es sich um ein von den tätigen Personen abhängiges Merkmal mit:

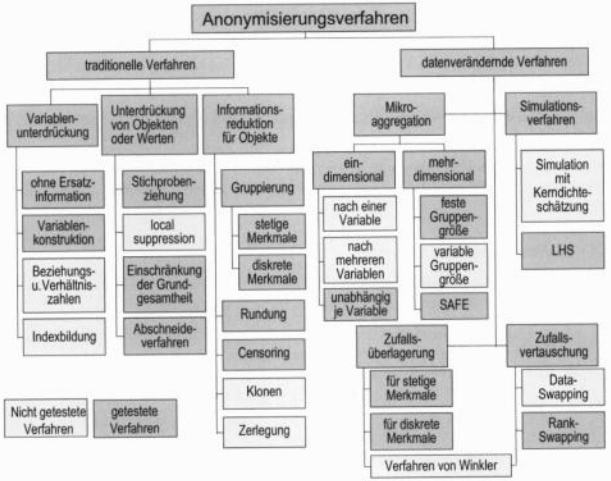
BGK	tätige Personen
40	20 bis 49
50	50 bis 99
60	100 bis 199
70	200 bis 499
80	500 bis 999

Ist es dem Datenangreifer möglich, über alle drei reidentifizierenden Merkmale Informationen zu erlangen, so können bereits fast alle Merkmalsträger zugeordnet werden. Außerdem würde, die Information, dass das gesuchte Unternehmen, das mit Abstand Größte in der Region ist, ausreichen um die Betriebe 06 und 11 zu erkennen. Die Information, dass nicht exportiert wird, reicht, um die Firma 05 zu erkennen.

1.3 Verfahrenstests

Im Rahmen des Projektes wurde eine Reihe von Verfahren getestet. Dabei bestand die Aufgabe darin, aus der Gesamtmenge aller der Projektgruppe bekannten Verfahren technisch praktikable und qualitativ erfolgversprechende Verfahren auszuwählen. Die Auswahl der Verfahren erfolgte mit dem Ziel die Anzahl der im Projekt zu testenden Verfahren auf eine realisierbare Menge zu reduzieren. Eine qualitative Bewertung der Verfahren ist damit noch nicht verbunden. Deshalb spielten Qualitätskriterien bei der Auswahl der Verfahren nur eine untergeordnete Rolle. Die Qualitätsbewertung erfolgt erst im Rahmen der Verfahrenstests.

Abbildung 3: Übersicht Anonymisierungsverfahren



Kriterien zur Verfahrensauswahl waren:

- Leichte Handhabbarkeit des Verfahrens
  Da die Verfahren später in den statistischen Ämtern mit dem vorhandenen Personal und den technischen Möglichkeiten einsetzbar sein müssen, ist die leichte Handhabbarkeit unumgänglich. Dieses Kriterium setzt eine verfügbare programmtechnische Realisierung und/oder verständliche Verfahrensbeschreibung voraus.
- Erfolgsaussichten der Verfahren
  Hier wurde auf Bewertungen in der Literatur zurückgegriffen. Bei einigen Verfahren sind bereits in der Literatur einige Nachteile bekannt, die durch andere Verfahren im Sinne einer Weiterentwicklung bereits gelöst wurden. Andere Verfahren befinden sich noch in einer theoretischen Entwicklungsstufe, so dass sie noch nicht praktikabel einsetzbar ist.
- Repräsentative Vertretung der Verfahrensgruppen
  Die Auswahl sollte aus allen Verfahrensgruppen Verfahren enthalten. Da Maße zur Datenqualität im Hinblick auf Analysefähigkeit und Datensicherheit noch nicht endgültig definiert wurden, soll keine Verfahrensgruppe generell ausgeschlossen werden, da jede Verfahrensgruppe einen anderen Ansatz der Veränderung der Einzeldaten repräsentiert.

- Abhängigkeit zwischen Verfahren
  Bei einigen Verfahren ist eine wirkungsvolle Anonymisierung nur durch das Verwenden von mehreren Verfahren gleichzeitig zu erzielen. Hier müssen natürlich alle untereinander abhängigen Verfahren berücksichtigt werden.

Die von der Projektgruppe ausgewählten Verfahren sind in Abbildung 3 gekennzeichnet.

2 Traditionelle Anonymisierungsverfahren

Die traditionellen Verfahren basieren in der Regel auf dem Verschweigen von kritischen Informationen. Das kann Merkmale, komplette Merkmalsträger (Objekte) oder aber den Abbau von Unterschieden zwischen einzelnen Merkmalsträgern (Informationsreduktion) beinhalten.

2.1 Variablenunterdrückung

Die Unterdrückung von Variablen ist mit und ohne Ersatzinformationen möglich. Folgende Möglichkeiten existieren (Datenbeispiele siehe Anhang-Abbildungen A2 – A5):

- Variablenunterdrückung ohne Ersatzinformation
  Existiert im Datenbestand ein Merkmal, dass sehr kritisch für die Reidentifikation ist, so kann es entfernt werden. Im Datenbeispiel würde es z.B. die Wirtschaftsklassifikation WZ 93 betreffen, da sie für 3 Unternehmen eindeutig ist.
- Variablenunterdrückung mit Ersatzinformation über Variablenkonstruktion
  Durch die Bildung von Linearkombinationen aus den kritischen Merkmalen (z.B. Summen, Durchschnitte) kann der Informationsverlust reduziert werden.

  Im Datenbeispiel könnte die Ausweisung von „Umsatz insgesamt" und „Löhne und Gehälter" z.B. die Erkennung des nicht exportierenden Unternehmens verhindern. Das Unternehmen, in dem die Gehälter die Löhne stark übersteigen, ist dann auch nicht mehr erkennbar.
- Variablenunterdrückung mit Ersatzinformation über Beziehungs- und Verhältniszahlen
  Beziehungs- und Verhältniszahlen verhindern das Erkennen von stark dominierenden Unternehmen, während sie die strukturellen Informationen erhalten. Merkmalsträger mit auffälligen Strukturen bleiben dadurch aber ungeschützt (z.B. Unternehmen ohne Auslandsumsatz).

  Möglich wären beim Datenbeispiel Verhältniszahlen wie „Umsatz je tätige Person", „Durchschnittslohn" und „Durchschnittsgehalt".
- Variablenunterdrückung mit Ersatzinformation über Indexbildung bei Zeitreihen und Paneldaten
  Indexzahlen verhindern wie Verhältniszahlen das Erkennen von stark dominierenden Unternehmen, während sie die zeitlichen Trendinformationen erhalten. Dabei wird für alle Werte nur die Veränderung zu einem Basiszeitraum ausgewiesen.

Ziel der Variablenunterdrückung bei Schlüsselmerkmalen ist die Reduzierung der Anzahl der eindeutigen Schlüsselvariablenkombinationen, d.h. der in der Mikrodatendatei nur einmal auftretenden Kombinationen von Merkmalsausprägungen der Schlüsselvariablen, und damit der Möglichkeit der eindeutigen Zuordnung. Bei der Entfernung von sensiblen Merkmalen sinkt automatisch der Nutzen der Reidentifikation.

Die Verfahren der Variablenunterdrückung sind für die statistische Analyse des verbleibenden Datenbestandes ohne Auswirkungen, da die verbleibenden Informationen nicht verändert wurden. Sie schränken jedoch die potentiellen Analysemöglichkeiten ein, da keine Analysen mehr möglich sind, die die entfernten Informationen betreffen. Hier ist immer genau zu prüfen, ob die verbleibenden Informationen für die ökonomischen Fragestellungen / Modelle noch ausreichend sind.

2.2 Unterdrückung von Objekten oder Werten

Bei der Unterdrückung von Objekten (Merkmalsträgern) handelt es sich um das Herausnehmen von ganzen Merkmalsträgern, um die Möglichkeit der Reidentifikation dadurch zu reduzieren, dass die Objekte nicht mehr im Datenbestand sind. Sind die Objekte nur auf Grund einzelner Merkmalswerte auffällig im Datenbestand, so besteht die Möglichkeit auch nur diese zu unterdrücken. Verfahrensvarianten sind:

- Stichprobenziehung
  Es wird eine Stichprobe generiert, wodurch eine Unsicherheit erzeugt wird, ob das gesuchte Objekt noch im Datenbestand ist oder nicht.
- Einschränkung der Grundgesamtheit
  Merkmalsträger, für die ein erhöhtes Risiko der Beschaffung von Zusatzwissen besteht, werden aus dem Datensatz entfernt. Sie können dadurch nicht mehr über den Datenbestand deanonymisiert werden (z.B. Entfernung publizitätspflichtiger Unternehmen, Entfernung von in der Öffentlichkeit stehenden Personen).
- Abschneideverfahren
  Merkmalsträger, die wegen ihrer Größe ein erhöhtes Reidentifikationsrisiko haben (besonders groß oder klein) werden aus dem Datenbestand entfernt.
- Unterdrückung einzelner Werte (local suppression)
  Einzelne auffällige Merkmalswerte werden im Datenbestand entfernt und durch eine entsprechende Kennzeichnung ersetzt.
- Unterdrückung einzelner Werte mit Einschätzung neuer Werte
  Einzelne auffällige Merkmalswerte werden im Datenbestand entfernt und durch eine entsprechende Schätzung ersetzt. Es wird z.B. eine Regressionsfunktion gebildet, mit der dann Schätzungen für die unterdrückten Werte erzeugt werden.

(Siehe Anhang-Abbildungen A6 – A9.)

Die Stichprobenziehung berührt die inferenzstatistischen Eigenschaften nicht, wenn das Stichprobendesign bekannt ist. Wird der Auswahlsatz in allen Schichten bzw. Klumpen konstant gehalten, weichen die Mittelwerte und die Stichprobenkovarianzmatrix lediglich aufgrund des Einflusses des größeren Zufallsfehlers von denen des Ausgangsdatensatzes ab. Ebenso bleiben die uni- und multivariaten Verteilungen asymptotisch erwartungstreu erhalten. Sind die Auswahlsätze in den einzelnen Schichten bzw. Klum-

pen verschieden, weichen die entsprechenden Statistiken der Sub-Stichprobe systematisch von denen der gesamten Originalstichprobe ab. Innerhalb der Schichten/Klumpen sind die Abweichungen aber nur auf die Realisierung des Zufallsfehlers zurückzuführen.

Unabhängig davon tritt jedoch folgendes Problem auf: Ein kleinerer Stichprobenumfang führt zu einer größeren Varianz der Schätzer. Das kann bei Statistiken, die bereits auf kleinen Stichprobenerhebungen basieren, die Brauchbarkeit der über eine Substichprobe anonymisierten Daten stark vermindern.

Einschränkung der Grundgesamtheit und Abschneideverfahren führen zu einem systematischen Entfernen von Teilgesamtheiten. Für diese sind keine Aussagen mehr möglich. Sofern dem Nutzer die Verkleinerung der Grundgesamtheit bekannt ist und die entfernten Teilgesamtheiten keinen wichtigen Beitrag zur empirischen Beurteilung der ökonomischen Fragestellungen liefern, bestehen für die Analyse der Restgesamtheit keinerlei Probleme.

Die Unterdrückung einzelner Werte (local suppression) hat gegenüber der vollständigen Entfernung der Merkmalsträger den Vorteil, dass die unkritischen Informationen erhalten bleiben. Die unterdrückten Werte können wie „missing-values" bei der Datenerhebung behandelt werden. Eine Variante ist die Einschätzung neuer Werte. Diese setzt ein entsprechendes Schätzmodell voraus. Gute Schätzungen sind jedoch nur möglich, wenn die unterdrückten Variablen mit anderen korreliert sind. Das Einschätzen von unterdrückten Werten erhöht bei guten Schätzungen aber wieder das Deanonymisierungsrisiko. Für die Regressionsanalyse sind Einschätzungen jedoch mit dem Risiko verbunden, dass sie zu einer systematischen und massiven Überschätzung des Bestimmtheitsmaßes führen, wenn die zu analysierenden Modelle auch für die Einschätzung der Werte verwendet wurden.

2.3 Informationsreduktion für Objekte

Bei der Informationsreduktion für Objekte (Merkmalsträger) handelt es sich um die Veränderung der Merkmalswerte von einzelnen Merkmalsträgern, um die Möglichkeit der Reidentifikation dadurch zu reduzieren, dass die Objekte nicht mehr eindeutig im Datenbestand sind. Verfahrensvarianten sind:

- Gruppierung
  Bei stetigen Merkmalen werden Intervalle gebildet, die dann als Intervallklassen immer einen Bereich von möglichen Werten beschreiben. Es wird dann der stetige Wert durch die entsprechende Klasse ersetzt (Beispiele: Beschäftigtengrößenklassen oder Umsatzgrößenklassen).
  Bei diskreten Merkmalen werden neue Kategorien eingeführt, die dann mehrere diskrete Merkmale umfassen und an Stelle der originalen diskreten Merkmale verwendet werden (Beispiele: Klassifikation der Wirtschaftszweige WZ 93 oder Regionalschlüssel).
- Rundung
  Stetige Merkmale werden so stark gerundet, dass dadurch sowohl die eindeutige Zuordnung als auch die Brauchbarkeit der Information für den Datenangreifer stark reduziert wird.

- Censoring (top-, bottomcoding)
  Merkmalsträger, die wegen ihrer Größe ein erhöhtes Reidentifikationsrisiko bei einem Merkmal haben (besonders groß oder klein), werden verändert, indem die kritischen Merkmalswerte (oberhalb/unterhalb eines Grenzwertes) auf die festgelegten Grenzwerte gesetzt werden.
- Klonen
  Einzelne kleine Merkmalsträger, die wegen ihrer seltenen diskreten Merkmalskombinationen auffällig sind, werden anonymisiert, indem gleichartige künstliche Merkmalsträger erzeugt werden. Die künstlichen Merkmalsträger haben die gleichen diskreten Merkmale und ähnliche stetige Merkmale.
- Zerlegung
  Einzelne große Merkmalsträger, die wegen der Größe ihrer stetigen Merkmalswerte auffällig sind, werden anonymisiert, indem ihre stetigen Merkmalswerte auf mehrere künstliche Merkmalsträger nach einem geheimen Verteilungsschlüssel verteilt werden.

(Siehe Anhang-Abbildungen A10 – A15.)

Die Informationsreduktion für Merkmalsträger beinhaltet eine Merkmalsvergröberung. Sie wird angewendet, um die Anzahl der vorhandenen Schlüsselvariablenkombinationen zu reduzieren. Damit reduziert sich die Anzahl der einzigartigen Ausprägungskombinationen, die wiederum den Aufwand zur Gewinnung einer eindeutigen Zuordnung erhöht. Gleichzeitig steigt der Informationsverlust in den Daten und die Unsicherheit des Datenangreifers wird erhöht.

Klonen und Zerlegung bewirken eine Vervielfältigung von Merkmalsträgern. Auch hiermit wird die dem einzelnen Merkmalsträger zuzuordnende Information reduziert. Unter dem Aspekt der geringsten Beeinflussung des Gesamtbestandes wird Klonen für kleine und Zerlegung für große Merkmalsträger angewandt.

3 Datenverändernde Verfahren

Datenverändernde Verfahren sind durch das systematische Verändern der Merkmalswerte gekennzeichnet. Während die im vorigen Abschnitt beschriebenen traditionellen Verfahren versuchen durch eine Informationsreduktion („Verschweigen") Anonymität herzustellen, generieren datenverändernde Verfahren neuen Merkmalswerte („Notlüge"). Dabei steht nicht mehr der Erhalt der Merkmalswerte des einzelnen Objektes im Vordergrund sonder der Erhalt der statistischen Eigenschaften der Mikrodatendatei. Das kann einerseits dadurch erfolgen, dass die neuen Merkmalswerte durch feste Transformationsvorschriften aus den originalen Merkmalswerten generiert werden (Mikroaggregation und Vertauschung), oder aber die Merkmalswerte durch stochastische Einflüsse überlagert werden.

3.1 Mikroaggregation

Grundprinzip von Mikroaggregationen ist die Gruppierung von möglichst ähnlichen Merkmalswerten und deren Vereinheitlichung durch das Ersetzen der gruppierten Merkmalswerte durch ihren Durchschnittswert. Fast alle Verfahren dieser Gruppe (bis auf SAFE) lassen sich nur für stetige Variablen sinnvoll anwenden.

Alle Gruppierungsverfahren gehen von Gruppengrößen von mindestens 3 Werten aus, denn bei nur zwei Merkmalsträgern ist das Risiko der Reidentifikation des einen Merkmalsträgers bei Kenntnis der Werte des anderen Merkmalsträgers noch vorhanden.

Die Gruppierung (Durchschnittsbildung) erfolgt üblicherweise über alle Merkmalswerte gleichzeitig (außer dem dritten Verfahren unter 3.1.1.). Es ist aber auch immer möglich, die Verfahren für Teilmengen der Merkmale separat durchzuführen (Blockung der Merkmale).

Mikroaggregationsverfahren reduzieren die Möglichkeit der eindeutigen Zuordnung der Merkmalsträger, weil durch die Vereinheitlichung innerhalb der Gruppen mehrere Merkmalsträger gleiche Merkmalswerte erhalten. Gleichzeitig erzeugt die Durchschnittsbildung eine Unsicherheit in den Daten, die den Wert der Information für den Datenangreifer reduziert. Für die Analysefähigkeit sind Mikroaggregationsverfahren durch eine systematische Verzerrung der Varianz/Standardabweichung gekennzeichnet. Die Gesamtvarianz im Datenbestand wird dabei um die entfernte Varianz innerhalb der Gruppen reduziert. Deshalb besteht das Ziel der Verfahren darin, möglichst ähnliche Merkmalsträger zu gruppieren, um den Varianzverlust zu minimieren.

3.1.1 Eindimensionale Mikroaggregation

Für die eindimensionalen Mikroaggregationen findet für die Bestimmung der Ähnlichkeit eine Transfomation an einem eindimensionalen Maß statt. Dabei gibt es folgende Varianten (für eine detailliertere Beschreibung vgl. Mateo-Sanz, J.M. und Domingo-Ferrer, J. 1998):

- Mikroaggregation nach einer Variable
  Es wird eine dominierende Variable herausgesucht und der Datenbestand danach sortiert. Danach werden absteigend immer drei benachbarte Merkmalsträger in einer Gruppe zusammmengefasst und alle ihre stetigen Merkmalswerte durch den Durchschnitt der Werte ersetzt. (Die dominierende Variable sollte dabei mit möglichst vielen weiteren Merkmalen stark korreliert sein.)
- Mikroaggregation nach mehreren Variablen
  Die Sortierung erfolgt an Hand von Hilfsvariablen. Die Hilfsvariablen sind dabei z.B. die Hauptkomponente (als eine durch Transformation gebildete Variable mit möglichst hoher Korrelation zu den anderen Variablen) oder die Z-Scores (als die Summe der standardisierten Originalvariablen).
- unabhängige eindimensionale Mikroaggregation
  Die einzelnen stetigen Merkmale werden unabhängig voneinander bearbeitet. Sie werden sortiert und die drei benachbarten Werte durch ihren Durchschnitt ersetzt. Danach werden sie an die Originalposition zurücksortiert und das nächste Merkmal bearbeitet.

(Siehe Anhang-Abbildungen A1 – A18.)

Die unabhängige eindimensionale Mikroaggregation hat gegenüber den anderen Mikroaggregationsverfahren ein zusätzliches Sicherheitsrisiko, weil für jeden Merkmalswert eine Ober- und Untergrenze der originalen Wertes durch die Durchschnitte der benachbarten Gruppen verfügbar ist. Diese Werte können jedoch den Bereich des originalen Wertes sehr stark einschränken.

3.1.2 Mehrdimensionale Mikroaggregation

Bei den mehrdimensionalen Mikroaggregationen findet für die Bestimmung der Ähnlichkeit eine mehrdimensionale Bestimmung des Abstandes zwischen den Merkmalsträgern statt. Dabei gibt es folgende Varianten (Datenbeispiele siehe Anhang-Abbildungen A19 – A23):

- mehrdimensionale Mikroaggregation mit fester Gruppengröße
  Es werden die beiden Merkmalsträger herausgesucht, die den größten Abstand untereinander haben (euklidischer Abstand der normierten Werte). Danach werden diesen beiden jeweils die zwei dichtesten Merkmalsträger dazugruppiert. Die verbleibenden, noch nicht gruppierten Merkmalsträger werden wieder analog behandelt. (für eine detailliertere Beschreibung vgl. Mateo-Sanz, J.M. und Domingo-Ferrer, J. 1998)
- mehrdimensionale Mikroaggregation mit variabler Gruppengröße
  Ziel der Mikroaggregationen mit variabler Gruppengröße ist eine noch stärkere datenorientierte Gruppenbildung durch die Möglichkeit von Gruppen größer als 3. Dabei werden Gruppen bis 5 angestrebt. Gruppen mit einer Größe vom doppelten der minimalen Gruppengröße oder größer (>5) lassen sich ohne Qualitätsverlust weiter teilen.

  Bei diesen Verfahren werden die einzelnen Objekte nach dem Kriterium der größten Ähnlichkeit ggf. auch Gruppen zugeordnet, die bereits 3 oder mehr Elemente haben. Erreichen Gruppen eine Größe von mehr als 5 Elementen, werden sie durch hierarchische Anwendung des Verfahrens wieder geteilt (für eine detailliertere Beschreibung vgl. Mateo-Sanz, J.M. und Domingo-Ferrer, J. 1998).
- SAFE Verfahren des Statistischen Landesamtes Berlin (siehe Höhne 2003)
  Der Algorithmus wurde ursprünglich für die Tabellengeheimhaltung entwickelt. Die Gruppenbildung orientiert sich deshalb an einer möglichst hochwertigen Abbildung der originalen ein- bis dreidimensionalen Verteilungstabellen und ist nicht abstandsorientiert wie bei den anderen Mikroaggregationsverfahren. Das Verfahren ermöglicht auch die Vereinheitlichung von diskreten Merkmalen. Es werden folgende Schritte vorgenommen:
  - Geheimhaltung auf Basis diskreter Variablen
    Es wird nach dem Kriterium minimaler Fehler in den Randsummen eine diskrete Basisdatei erstellt, in der alle Ausprägungskombinationen diskreter Variablen mit mindestens drei Einheiten besetzt sind.
  - Zuordnung der Merkmalswerte der stetigen Variablen
    Dabei werden die originalen Sätze zu den Merkmalskombinationen der diskreten Variablen nach folgenden Kriterien zugeordnet:
    - größte Ähnlichkeit in den diskreten Variablen,
    - Verschieben der möglichst kleinsten Merkmalswerte der stetigen Variablen.
  - Bearbeitung der Dominanzen und des Problems der merkmalsbezogenen Fallzahlen unter 2.
  - Optimierung und Qualitätssicherung der Ergebnisse.
  - Trippelbildung bei den stetigen Variablen und Berechnung der anonymisierten Einzelwerte.

SAFE reduziert das Risiko der eindeutigen Zuordnung von Einheiten, da immer mehrere gleiche Einheiten vorhanden sind. Außerdem wird durch die Mittelwertbildung eine Unsicherheit in den Daten induziert, die eine weitere Schutzwirkung hat. Die Bearbeitung von Dominanzen und merkmalsbezogenen Fallzahlproblemen erhöht zwar die Schutzwirkung (z.B. im Vergleich zur reinen Mikroaggregation), geht aber in der Regel mit einem zusätzlichen Qualitätsverlust einher. Auf diese Schritte kann bei der Erstellung anonymisierter Einzeldaten ggf. verzichtet werden.

3.2 Zufallsüberlagerung

Grundprinzip der Zufallsüberlagerung ist die Erzeugung von neuen Merkmalswerten. Diese werden durch ein stochastisches Verfahren aus den originalen Werten generiert. Zufallsüberlagerungen erzeugen ihre Schutzwirkung dadurch, dass die Sicherheit in den Daten reduziert wird, d.h. nur mit einer gewissen Wahrscheinlichkeit handelt es sich bei zugeordneten Merkmalsträgern um eine richtige Zuordnung. Bei den erhaltenen Informationen sind diese wiederum nur mit einer gewissen Wahrscheinlichkeit für einen Datenangreifer brauchbar.

3.2.1 Zufallsüberlagerung für stetige Merkmale

Für stetige Merkmale gibt es eine ganze Reihe von Verfahren (Kim/Winkler/Sullivan; für einen Überblick vgl. Brand 2000). Bei stetigen Merkmalen wird der neue Merkmalswert in der Regel aus dem originalen und einer Zufallszahl durch Addition oder Multiplikation generiert. Um die Größe der Zufallszahl und damit die Schutzwirkung abhängig von dem originalen Merkmalswert zu gestalten, sind verschiedene Varianten von linearen und nichtlinearen Transformationen der Zufallszahlen bekannt.

Bei der Analysefähigkeit gibt es zwischen den einzelnen Verfahren erhebliche Unterschiede. Die einfache additive Überlagerung mit normalverteilten Zufallsfehlern und das Verfahren von Kim erlauben die konsistente Schätzung von Erwartungswerten, Varianz-Kovarianz-Matrizen und Parametern und wichtigen Statistiken von OLS-Regressionen, sofern der Erwartungswert der Überlagerungen Null ist und ihre Varianz bekannt ist. Beim Verfahren von Kim sind diese Schätzungen auch ohne die Kenntnis der Varianz der Störterme möglich. Das Verfahren von Kim erlaubt ebenfalls die Schätzung von OLS-Regressionen in nicht zufällig ausgewählten Teilstichproben (Teilmassen). Diese Verfahren erhalten aber nicht die univariaten Verteilungen. Das Verfahren von Sullivan erhält dagegen näherungsweise die univariaten Verteilungen, kann aber nicht für die Analyse von Teilmassen eingesetzt werden.

(Siehe Anhang-Abbildung A24.)

3.2.2 Zufallsüberlagerung für diskrete Merkmale

Diskrete Merkmale werden durch die Definition von Übergangswahrscheinlichkeiten randomisiert. Dabei werden die Merkmale mit bei der Anwendung festzulegenden Übergangswahrscheinlichkeiten in andere Ausprägungen transformiert. Bekannt ist das Verfahren PRAM (Post Randomisation Method siehe Willenborg, L. und T. de Waal 2001). Dabei werden die diskreten Merkmalswerte wie bei der in Erhebungen verwendeten Randomisierung von Antworten (sog. Randomised Response Technique) verändert.

Die veröffentlichten Werte entsprechen nur noch mit einer im Verfahren festgelegten Wahrscheinlichkeit den Werten im Originaldatensatz. Die Richtigkeit der Zuordnung von Merkmalsträgern ist dadurch mit einer gewissen Unsicherheit behaftet.

Die Kenntnis der Übergangswahrscheinlichkeiten ermöglicht bei der Analyse die konsistente Schätzung der univariaten Verteilungen. Invariante PRAM sollten die univariaten Verteilungen erhalten, d.h. bei invarianten PRAM werden die univariaten Verteilungen nicht systematisch verzerrt. Systematische Untersuchungen zu den Auswirkungen auf multivariate Verfahren (z.B. Regressionen) liegen nicht vor.

(Siehe Anhang-Abbildungen A25 – A26.)

3.3 Zufallsvertauschungen

Grundprinzip der Zufallsvertauschungen ist die Verschiebung von Merkmalsausprägungen zwischen verschiedenen Merkmalsträgern. Diese Verschiebungen werden durch stochastische Faktoren beeinflusst, damit sie nicht reproduziert und somit rückgängig gemacht werden können. Bei der Zufallsvertauschung von Merkmalswerten entsteht die Anonymität dadurch, dass die Ausprägungskombinationen nicht mehr originalgetreu sind. Sowohl das Auffinden der gesuchten Merkmalskombinationen als auch dann erhaltene Zusatzinformationen sind nur mit einer gewissen Unsicherheit fehlerfrei. Zufallsvertauschungen sind sowohl für diskrete als auch stetige Merkmalsträger möglich.

Da weder künstliche Werte generiert werden noch die Anzahl der Merkmalsausprägungen verändert wird, bleiben die univariaten Verteilungen und somit auch Mittelwerte und Varianzen erhalten. Multivariate Verteilungen und Korrelationen zwischen den Variablen werden gestört. Rank-Swapping versucht den Erhalt der Rangkorrelationen.

3.3.1 Einfaches Data-Swapping

Beim einfachen Data-Swapping (siehe auch Boyd, M., Vickers, P. 1999) werden die Merkmalsträger inhaltlich, d.h. anhand ausgewählter diskreter Merkmale, gruppiert. Die übrigen Merkmalswerte werden dann innerhalb der Gruppen für jedes Merkmal getrennt zufällig getauscht. (Siehe Anhang-Abbildung A27)

3.3.2 Rank-Swapping

Nach einem Sortieren der Datensätze nach einem Merkmal werden die Merkmalswerte in einem festgelegtem Nachbarschaftsbereich (Anteil des Datenbestandes) zufällig getauscht. Damit werden möglichst ähnliche Merkmalswerte getauscht und so die Rangstatistiken möglichst erhalten. Die Bearbeitung erfolgt für jedes Merkmal getrennt. (Siehe Anhang-Abbildung A28.) – (Siehe auch Dalenius, T. und Reiss, S.P. 1982.)

3.3.3 Verfahren von Winkler

Das Verfahren von Winkler (siehe auch Kim, J.J., Winkler, W.E. 1995) stellt eine Kombination aus Zufallsüberlagerung und Zufallsvertauschung dar. Merkmalsträger, die durch Zufallsüberlagerung nicht genügend anonymisiert werden können, werden dabei nachträglich noch einem Data-Swapping unterzogen.

Hier werden somit die Vorteile der beiden Verfahren gekoppelt, indem die höhere Datensicherheit, die durch Data-Swapping erreicht wird, mit der höheren Datenqualität der Zufallsüberlagerung verbunden wird, da nur bei den Problemfällen Data-Swapping eingesetzt wird. (Siehe Anhang-Abbildung A29.)

3.4 Simulationsverfahren

Grundprinzip der Simulationsverfahren ist die Erzeugung von synthetischen Merkmalsträgern. Diese werden durch ein stochastisches Verfahren generiert. Typisch für Simulationsverfahren ist, dass die Anzahl der synthetischen Merkmalsträger nicht mit dem originalen Datenbestand übereinstimmen muss. Somit lassen sich auch durchaus viel kleinere/größere Testdaten erzeugen. Die Testdaten lassen sich dann nicht mehr auf die Originaldaten zurückführen.

3.4.1 Resampling

Das Resamplingverfahren (siehe auch Fienberg, S.E. 1997) basiert auf der Idee, die mehrdimensionale Kerndichte des gesamten Datenbestandes zu schätzen. Mit Hilfe dieser Dichte wird dann die gewünschte Anzahl der synthetischen Datensätze erzeugt.

Das Verfahren ist bei stetigen Variablen nur sehr schwer einsetzbar. Es gibt auch noch keine konkreten Erfahrungen mit dieser Methode. Der Einfluss auf ökonometrische Schätzungen ist noch ungeklärt.

3.4.2 Latin Hypercube Sampling

Beim Latin Hypercube Sampling (siehe Dandekar, R.A., M. Cohen und N. Kirkendall 2002) erfolgt zuerst ausgehend von der Anzahl (n) an gewünschten synthetischen Datensätzen eine Simulation der eindimensionalen Merkmalswerte. Diese werden mit Hilfe der geglätteten empirischen Verteilungsfunktion oder einer theoretischen Verteilungsfunktion für die einzelnen Variablen aus gleichverteilten Zufallswerten erzeugt. Diese synthetischen Merkmalswerte werden in einem zweiten Schritt durch folgendes Swapping-Verfahren so umgeordnet, dass die Rangkorrelationen optimiert werden. Die synthetischen Merkmalswerte bilden zusammen eine Datenmatrix V mit n Sätzen und m synthetischen Merkmalen. Es erfolgt dann die Bestimmung der beiden Rangkorrelationsmatrizen T für den originalen Datenbestand und C für die synthetische Datenmatrix V. Danach erfolgt eine Anpassung der Rangkorrelation von V an T über folgenden Algorithmus:

a) Bestimmung der unteren Dreiecksmatrizen Q und P derart, dass $Q*Q'=C$ und $P*P'=T$ gilt.

b) Bestimmung von $R = V*(P*Q^{-1})'$ (mit ' - transponiert und $^{-1}$ - inverse Matrix).

c) Spaltenweise Ermittlung der Rangpositionen in R.

d) Umsortieren der Werte in den Spalten in V derart, dass die Rangpositionen mit R übereinstimmen.

In Erweiterungen des LHS wird die Anpassung der Rangkorrelationen so lange wiederholt, bis der Abstand zwischen C und T sich nicht mehr als ein akzeptables Toleranzniveau verringert.

Beim LHS werden somit synthetische Merkmalswerte durch das Verfahren optimal zu synthetischen Merkmalsträgern kombiniert. Damit haben die erzeugten Sätze keinen direkten Bezug mehr zu den Originaldaten.

Die Qualität der verwendeten univariaten Verteilungen bestimmt dabei, wie gut die Mittelwerte, Varianzen und die univariaten Verteilungen erhalten bleiben. Die Rangkorrelationen werden durch den Anpassungsalgorithmus gut reproduziert. Für Teilmassen sind diese Aussagen nur möglich, wenn vor der Anwendung des Verfahren der Datenbestand entsprechend gruppiert wurde und das Verfahren auf jeden Teilbestand getrennt angewendet wird. (Siehe Anhang-Abbildungen A30 – A32.)

4 Fazit

In diesem Beitrag wurde ein systematischer Überblick über die Verfahren zur Anonymisierung von Einzeldaten gegeben. Die Verfahren wurden in Ihrer Funktionsweise beschrieben und die Eigenschaften wurden – soweit bekannt – zusammengefasst dargestellt. Der Schwerpunkt der Darstellung lag dabei auf den für die weiteren Arbeiten im Projekt ausgewählten verfahren. Die beschriebenen Verfahren bieten umfangreiche Möglichkeiten Anonymisierungen von wirtschaftsstatistischen Einzeldaten vorzunehmen. Aufbauend auf diesen Arbeiten können dann die weiteren Aufgaben im Projekt bearbeitet werden. Hierzu sollten zunächst Bewertungskriterien festgelegt werden, mit denen es möglich ist, die Qualität der Verfahren bezüglich des Erreichens der Anonymität und des Erhalts der Analysefähigkeit zu messen. Anschließend sind Grenzwerte für das Erreichen der Anonymität festzulegen, mit denen die Entscheidung möglich ist, ob die Verfahren für die Anonymisierung von wirtschaftsstatistischen Einzeldaten ausreichen. Unter der Menge der genügend anonymisierenden Verfahren sollen die Verfahren mit dem besten Erhalt der Analysefähigkeit im Projekt bestimmt werden. Bei den nicht genügend anonymisierenden Verfahren sind diejenigen Verfahren, die noch eine gute Analysequalität aufweisen, darauf zu untersuchen, ob ggf. die Kombination mit weiteren Anonymisierungsverfahren eine höhere Datensicherheit gewährleistet. Da die Verfahrensqualität durchaus von der Art des Mikrodatenmaterials abhängen kann, werden verschiedene Datenbestände gleichzeitig untersucht.

Literaturhinweise

Boyd, M.; Vickers, P. (1999): Record Swapping – A Possible Disclosure Control Approach for the 2001 UK Census, Beitrag zur: Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, 8. – 10. März 1999, Thessaloniki, Griechenland.

Brand, R. (2000): Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos, in: Beiträge zur Arbeitsmarkt- und Berufsforschung 237,

Brand, R. (2001): Microdata Protection through Noise Addition, in: Domingo-Ferrer, Josep (Hrsg.), Inference Control in Statistical Data Bases – From Theory to Practice, Springer, 2002.

Corsini, V.;, Franconi, L.; Pagliuca, D. und Seri, G. (1998): An application of microaggregation methods to Italian business surveys, in: Statistical data protection Proceedings of the conference, Eurostat 1999.

Dalenius, T. und Reiss, S. P. (1982): Data-swapping: A Technique for Disclosure Control, in: Journal of Statistical Planning and Inference 6, S. 73 – 85

Dandekar, R. A.; Cohen, M. und Kirkendall, N. (2002): Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique, 2001, in: Domingo-Ferrer, Josep (Hrsg.): Inference Control in Statistical Data Bases – From Theory to Practice, Springer.

Dandekar, R. A.; Domingo-Ferrer, J. und Sebé, F. (2002): LHS-Based Hybrid Microdata vs Rank Swapping and Microaggregation for Numeric Microdata Protection, 2001, in: Domingo-Ferrer, Josep (Hrsg.): Inference Control in Statistical Data Bases – From Theory to Practice, Springer.

Domingo-Ferrer, J. und Mateo-Sanz, J. M. (2001): An empirical comparison of SDC methods for continuous microdata in terms of information loss and disclosure risk, Working Paper at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality 14 – 16 March 2001.

Domingo-Ferrer, J. (Hrsg., 2002): Inference Control in Statistical Data Bases – From Theory to Practice, Springer.

Evers, K. und Höhne, J. (1999): SAFE – Ein Verfahren zur Anonymisierung und statistischen Geheimhaltung wirtschaftsstatistischer Einzeldaten, in: Spektrum Bundesstatistik, Band 14, Wiesbaden, S. 136 – 147.

Fienberg, S. E. (1997): Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, Technical Report No. 668, Carnegie Mellon University, Pittsburgh.

Höhne, J. (2003): SAFE – Ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben, in: Berliner Statistik, Statistische Monatsschrift, Nr. 3 2003, Berlin S. 96 – 107.

Kim, J. J. and Winkler, W. E. (1995): Masking Microdata Files, American Statistical Association, Proceedings of the Section on Survey Research Methods, S. 114 – 119.

Mateo-Sanz, J. M. und Domingo-Ferrer, J. (1998): A Comparative Study of Microaggregation Methods, auf der Homepage von Domingo-Ferrer URL: http://www.etse.urv.es/~jdomingo/.

Mateo-Sanz, J. M. und Domingo-Ferrer, J. (1998): A method for data-oriented multivariate microaggregation, in: Statistical data protection, Proceedings of the conference, Eurostat 1999.

Müller, W.; Blien, U.; Knoche, P.; Wirth, H. u.a. (1991): Die faktische Anonymität von Mikrodaten, in: Statistisches Bundesamt (Hrsg.), Forum der Bundesstatistik, Band 19.

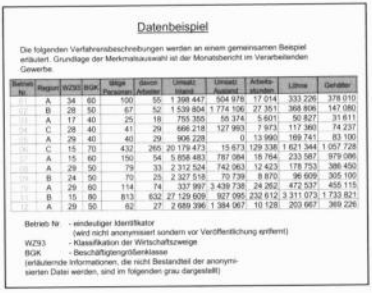
Ronning, G.; Brand, R.; Höhne, J.; Rosemann, M.; Wiegert, R. (2002): Anonymisierungsverfahren – Überblick und erste Bewertung –, Arbeitspapier der Projektgruppe: Faktische Anonymisierung von wirtschaftsstatistischen Einzeldaten.

Willenborg, L. und de Waal, T. (2001): Elements of Statistical Disclosure Control, Springer Lecture Notes in Statistics 155, Springer.

Anhang

Datenbeispiele für Anonymisierungsverfahren

Abbildung A1



Datenbeispiel

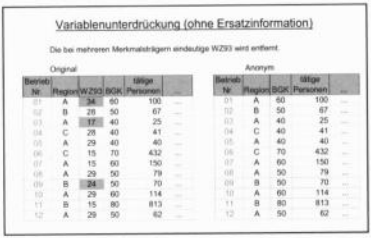
Die folgenden Verfahrensbeschreibungen werden an einem gemeinsamen Beispiel erläutert. Grundlage der Merkmalsauswahl ist der Monatsbericht im Verarbeitenden Gewerbe.

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	davon Arbeiter	Umsatz Inland	Umsatz Ausland	Arbeitsstunden	Löhne	Gehälter
01	A	34	60	100	55	1 398 447	504 978	17 014	333 226	378 010
02	B	28	50	67	52	1 539 804	1 774 106	27 351	368 806	147 080
03	A	17	40	25	18	755 355	55 374	5 601	50 827	31 611
04	C	28	40	41	29	686 218	127 993	7 973	117 380	74 237
05	A	29	40	40	29	906 228	0	13 990	169 741	83 100
06	C	15	70	432	265	20 179 473	15 673	129 338	1 621 344	1 057 728
07	A	15	60	150	54	5 858 483	787 084	18 764	233 587	979 096
08	A	29	50	79	33	2 312 524	742 063	12 423	178 753	386 450
09	B	24	50	70	25	2 327 518	70 739	8 870	96 609	305 100
10	A	29	60	114	74	337 997	3 439 738	24 262	472 537	455 115
11	B	15	80	813	632	27 129 609	927 095	232 612	3 311 073	1 733 821
12	A	29	50	62	27	2 689 396	1 384 067	10 128	203 667	369 226

Betrieb Nr. - eindeutiger Identifikator (wird nicht anonymisiert sondern vor Veröffentlichung entfernt)
WZ93 - Klassifikation der Wirtschaftszweige
BGK - Beschäftigtengrößenklassen
(erläuternde Informationen, die nicht Bestandteil der anonymisierten Daten werden, sind im folgenden grau dargestellt)

Traditionelle Verfahren – Variablenunterdrückung

Abbildung A2



Variablenunterdrückung (ohne Ersatzinformation)

Die bei mehreren Merkmalsträgern eindeutige WZ93 wird entfernt.

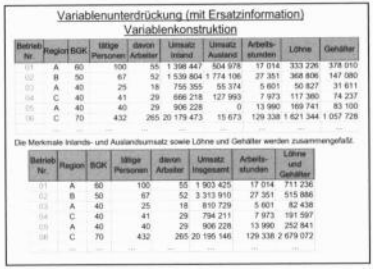
Original

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	...
01	A	34	60	100	...
02	B	28	50	67	...
03	A	17	40	25	...
04	C	28	40	41	...
05	A	29	40	40	...
06	C	15	70	432	...
07	A	15	60	150	...
08	A	29	50	79	...
09	B	24	50	70	...
10	A	29	60	114	...
11	B	15	80	813	...
12	A	29	50	62	...

Anonym

Betrieb Nr.	Region	BGK	tätige Personen	...
01	A	60	100	...
02	B	50	67	...
03	A	40	25	...
04	C	40	41	...
05	A	40	40	...
06	C	70	432	...
07	A	60	150	...
08	A	50	79	...
09	B	50	70	...
10	A	60	114	...
11	B	80	813	...
12	A	50	62	...

Abbildung A3



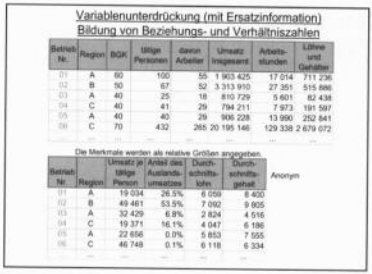
Variablenunterdrückung (mit Ersatzinformation)
Variablenkonstruktion

Betrieb Nr.	Region	BGK	tätige Personen	davon Arbeiter	Umsatz Inland	Umsatz Ausland	Arbeitsstunden	Löhne	Gehälter
01	A	60	100	55	1 398 447	504 978	17 014	333 226	378 010
02	B	50	67	52	1 539 804	1 774 106	27 351	368 806	147 080
03	A	40	25	18	755 355	55 374	5 601	50 827	31 611
04	C	40	41	29	686 218	127 993	7 973	117 380	74 237
05	A	40	40	29	906 228	0	13 990	169 741	83 100
06	C	70	432	265	20 179 473	15 673	129 338	1 621 344	1 057 728
...	...	...	...	...	...	...	...	...	...

Die Merkmale Inlands- und Auslandsumsatz sowie Löhne und Gehälter werden zusammengefaßt.

Betrieb Nr.	Region	BGK	tätige Personen	davon Arbeiter	Umsatz Insgesamt	Arbeitsstunden	Löhne und Gehälter
01	A	60	100	55	1 903 425	17 014	711 236
02	B	50	67	52	3 313 910	27 351	515 886
03	A	40	25	18	810 729	5 601	82 438
04	C	40	41	29	794 211	7 973	191 597
05	A	40	40	29	906 228	13 990	252 841
06	C	70	432	265	20 195 146	129 338	2 679 072
...	...	...	...	...	...	...	...

Abbildung A4



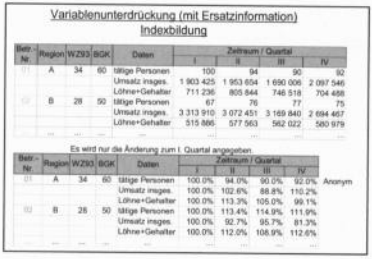
Variablenunterdrückung (mit Ersatzinformation)
Bildung von Beziehungs- und Verhältniszahlen

Betrieb Nr.	Region	BGK	tätige Personen	davon Arbeiter	Umsatz insgesamt	Arbeitsstunden	Löhne und Gehälter
01	A	60	100	55	1 903 425	17 014	711 236
02	B	50	67	52	3 313 910	27 351	515 886
03	A	40	25	18	810 729	5 601	62 436
04	C	40	41	29	794 211	7 973	191 597
05	A	40	40	29	906 228	13 990	252 841
06	C	70	432	265	20 195 146	129 338	2 679 072
...	...	...	...	...	...	...	...

Die Merkmale werden als relative Größen angegeben.

Betrieb Nr.	Region	Umsatz je tätige Person	Anteil des Auslandsumsatzes	Durchschnittslohn	Durchschnittsgehalt	Anonym
01	A	19 034	26.5%	6 059	8 400	
02	B	49 461	53.9%	7 092	9 805	
03	A	32 429	6.8%	2 824	4 516	
04	C	19 371	16.1%	4 047	6 186	
05	A	22 656	0.0%	5 853	7 555	
06	C	46 748	0.1%	6 118	6 334	
...	...	...	...	...	...	

Abbildung A5



Variablenunterdrückung (mit Ersatzinformation)
Indexbildung

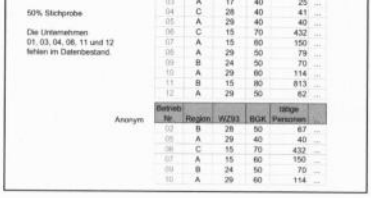
Betr.-Nr.	Region	WZ93	BGK	Daten	Zeitraum / Quartal I	II	III	IV
01	A	34	60	tätige Personen	100	94	90	92
				Umsatz insges.	1 903 425	1 953 654	1 690 006	2 097 546
				Löhne+Gehälter	711 236	805 844	746 518	704 488
02	B	28	50	tätige Personen	67	76	77	75
				Umsatz insges.	3 313 910	3 072 451	3 169 840	2 694 467
				Löhne+Gehälter	515 886	577 563	562 022	580 979
...	...	...	...	...	...	...	...	...

Es wird nur die Änderung zum I. Quartal angegeben.

Betr.-Nr.	Region	WZ93	BGK	Daten	Zeitraum / Quartal I	II	III	IV	Anonym
01	A	34	60	tätige Personen	100.0%	94.0%	90.0%	92.0%	
				Umsatz insges.	100.0%	102.6%	88.8%	110.2%	
				Löhne+Gehälter	100.0%	113.3%	105.0%	99.1%	
02	B	28	50	tätige Personen	100.0%	113.4%	114.9%	111.9%	
				Umsatz insges.	100.0%	92.7%	95.7%	81.3%	
				Löhne+Gehälter	100.0%	112.0%	108.9%	112.6%	
...	...	...	...	...	...	...	...	...	

Traditionelle Verfahren – Unterdrückung von Merkmalsträgern oder Werten

Abbildung A6



Unterdrückung von Merkmalsträgern
Stichprobenziehung

Original

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	...
01	A	34	60	100	...
02	B	28	50	67	...
03	A	17	40	25	...
04	C	28	40	41	...
05	A	29	40	40	...
06	C	15	70	432	...
07	A	15	60	150	...
08	A	29	50	79	...
09	B	24	50	70	...
10	A	29	60	114	...
11	B	15	80	813	...
12	A	29	50	62	...

50% Stichprobe

Die Unternehmen 01, 03, 04, 08, 11 und 12 fehlen im Datenbestand.

Anonym

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	...
02	B	28	50	67	...
05	A	29	40	40	...
06	C	15	70	432	...
07	A	15	60	150	...
09	B	24	50	70	...
10	A	29	60	114	...

Abbildung A7

Unterdrückung von einzelnen Werten
(local suppression / Blanking)

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt	Arbeits-stunden	Löhne und Gehälter	Original
01	A	34	60	100	1 903 425	17 014	711 236	
02	B	28	50	67	3 313 910	27 351	515 886	
03	A	17	40	25	810 729	5 601	82 438	
04	C	28	40	41	794 211	7 973	191 597	
05	A	29	40	40	906 228	13 990	252 841	
06	C	15	70	432	20 195 146	129 338	2 679 072	
...	...	...	...	...	...	...	...	

Unterdrückung von eindeutigen WZ-Klassifikationen und dominanten Umsätzen und Löhnen.

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt	Arbeits-stunden	Löhne und Gehälter	Anonym
01	A	*	60	100	1 903 425	17 014	711 236	
02	B	28	50	67	3 313 910	27 351	515 886	
03	A	*	40	25	810 729	5 601	82 438	
04	C	28	40	41	794 211	7 973	191 597	
05	A	29	40	40	906 228	13 990	252 841	
06	C	15	70	432		* 129 338	*	
...	...	...	...	...	...	...	...	

Abbildung A8

Einschätzung von unterdrückten Werten (Blanking and Imputation)

Bestimmung von Schätzmodellen für einzelne Werte. (z.B. Regressionsfunktionen)
Umsatz = tätige Personen* 36 747.55 + 224 342.00
Löhne = tätige Personen* 6 140.73 + 76 635.73
Ersetzung von dominanten Umsätzen und Löhnen durch Schätzwerte

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt	Arbeits-stunden	Löhne und Gehälter	Anonym
01	A	*	60	100	1 903 425	17 014	711 236	
02	B	28	50	67	3 313 910	27 351	515 886	
03	A	*	40	25	810 729	5 601	82 438	
04	C	28	40	41	794 211	7 973	191 597	
05	A	29	40	40	906 228	13 990	252 841	
06	C	15	70	432	16 099 282	129 338	2 729 429	
...	...	...	...	...	...	...	...	

Abbildung A9

Einschränkung der Grundgesamtheit und
Abschneideverfahren / Herausnahme von Ausreißern

Die Unternehmen 06 und 11 fehlen im anonymen Datenbestand.

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	...
01	A	34	60	100	...
02	B	28	50	67	...
03	A	17	40	25	...
04	C	28	40	41	...
05	A	29	40	40	...
07	A	15	60	150	...
08	A	29	50	79	...
09	B	24	50	70	...
10	A	29	60	114	...
12	A	29	50	62	...

Traditionelle Verfahren – Informationsreduktion für Objekte

Abbildung A10

Gruppierung von stetigen Merkmalen (Klassierung)

Wenn an Stelle der „tätigen Personen" nur die Beschäftigtengrößenklasse (BGK) veröffentlicht wird, ist für die Unternehmen nur ein Intervall zu den „tätigen Personen" bekannt. Mehrere Unternehmen gehören dann dem gleichen Intervall an.

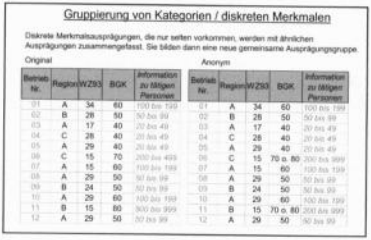
Original

Betrieb-Nr.	Region	WZ93	BGK	tätige Personen	...
01	A	34	60	100	...
02	B	28	50	67	...
03	A	17	40	25	...
04	C	28	40	41	...
05	A	29	40	40	...
06	C	15	70	432	...
07	A	15	60	150	...
08	A	29	50	79	...
09	B	24	50	70	...
10	A	29	60	114	...
11	B	15	80	813	...
12	A	29	50	62	...

Anonym

Betrieb-Nr.	Region	WZ93	BGK	Information zu tätigen Personen	...
01	A	34	60	100 bis 199	...
02	B	28	50	50 bis 99	...
03	A	17	40	20 bis 49	...
04	C	28	40	20 bis 49	...
05	A	29	40	20 bis 49	...
06	C	15	70	200 bis 499	...
07	A	15	60	100 bis 199	...
08	A	29	50	50 bis 99	...
09	B	24	50	50 bis 99	...
10	A	29	60	100 bis 199	...
11	B	15	80	500 bis 999	...
12	A	29	50	50 bis 99	...

Abbildung A11



Gruppierung von Kategorien / diskreten Merkmalen

Diskrete Merkmalsausprägungen, die nur selten vorkommen, werden mit ähnlichen Ausprägungen zusammengefasst. Sie bilden dann eine neue gemeinsame Ausprägungsgruppe.

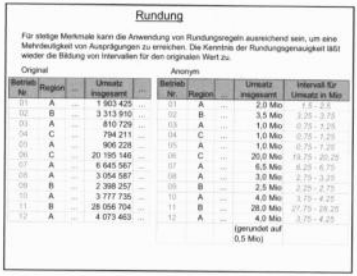
Original

Betrieb Nr.	Region	WZ93	BGK	Information zu tätigen Personen
01	A	34	60	100 bis 199
02	B	28	50	50 bis 99
03	A	17	40	20 bis 49
04	C	28	40	20 bis 49
05	A	29	40	20 bis 49
06	C	15	70	200 bis 499
07	A	15	60	100 bis 199
08	A	29	50	50 bis 99
09	B	24	50	50 bis 99
10	A	29	60	100 bis 199
11	B	15	80	500 bis 999
12	A	29	50	50 bis 99

Anonym

Betrieb Nr.	Region	WZ93	BGK	Information zu tätigen Personen
01	A	34	60	100 bis 199
02	B	28	50	50 bis 99
03	A	17	40	20 bis 49
04	C	28	40	20 bis 49
05	A	29	40	20 bis 49
06	C	15	70 o. 80	200 bis 999
07	A	15	60	100 bis 199
08	A	29	50	50 bis 99
09	B	24	50	50 bis 99
10	A	29	60	100 bis 199
11	B	15	70 o. 80	200 bis 999
12	A	29	50	50 bis 99

Abbildung A12



Rundung

Für stetige Merkmale kann die Anwendung von Rundungsregeln ausreichend sein, um eine Mehrdeutigkeit von Ausprägungen zu erreichen. Die Kenntnis der Rundungsgenauigkeit läßt wieder die Bildung von Intervallen für den originalen Wert zu.

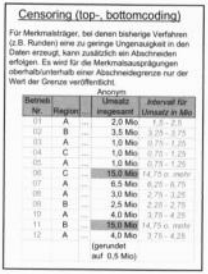
Original

Betrieb Nr.	Region	...	Umsatz insgesamt	...
01	A	...	1 903 425	...
02	B	...	3 313 910	...
03	A	...	810 729	...
04	C	...	794 211	...
05	A	...	906 228	...
06	C	...	20 195 146	...
07	A	...	6 645 567	...
08	A	...	3 054 587	...
09	B	...	2 398 257	...
10	A	...	3 777 735	...
11	B	...	28 056 704	...
12	A	...	4 073 463	...

Anonym

Betrieb Nr.	Region	...	Umsatz insgesamt	Intervall für Umsatz in Mio
01	A	...	2,0 Mio	1,5 - 2,5
02	B	...	3,5 Mio	3,25 - 3,75
03	A	...	1,0 Mio	0,75 - 1,25
04	C	...	1,0 Mio	0,75 - 1,25
05	A	...	1,0 Mio	0,75 - 1,25
06	C	...	20,0 Mio	19,75 - 20,25
07	A	...	6,5 Mio	6,25 - 6,75
08	A	...	3,0 Mio	2,75 - 3,25
09	B	...	2,5 Mio	2,25 - 2,75
10	A	...	4,0 Mio	3,75 - 4,25
11	B	...	28,0 Mio	27,75 - 28,25
12	A	...	4,0 Mio (gerundet auf 0,5 Mio)	3,75 - 4,25

Abbildung A13



Censoring (top-, bottomcoding)

Für Merkmalsträger, bei denen bisherige Verfahren (z.B. Runden) eine zu geringe Ungenauigkeit in den Daten erzeugt, kann zusätzlich ein Abschneiden erfolgen. Es wird für die Merkmalsausprägungen oberhalb/unterhalb einer Abschneidegrenze nur der Wert der Grenze veröffentlicht.

Anonym

Betrieb Nr.	Region	...	Umsatz insgesamt	Intervall für Umsatz in Mio
01	A	...	2,0 Mio	1,5 - 2,5
02	B	...	3,5 Mio	3,25 - 3,75
03	A	...	1,0 Mio	0,75 - 1,25
04	C	...	1,0 Mio	0,75 - 1,25
05	A	...	1,0 Mio	0,75 - 1,25
06	C	...	15,0 Mio	14,75 o. mehr
07	A	...	6,5 Mio	6,25 - 6,75
08	A	...	3,0 Mio	2,75 - 3,25
09	B	...	2,5 Mio	2,25 - 2,75
10	A	...	4,0 Mio	3,75 - 4,25
11	B	...	15,0 Mio	14,75 o. mehr
12	A	...	4,0 Mio (gerundet auf 0,5 Mio)	3,75 - 4,25

Abbildung A14

Zerlegung von großen Merkmalsträgern

Existieren nur vereinzelt besonders große Einheiten, so besteht die Möglichkeit durch Aufteilung der Werte auf zwei oder mehrere Einheiten die auffällig großen Einheiten zu verstecken. Die stetigen Merkmale werden nach einem zufälligen und geheimen Anteil aufgeteilt, die diskreten Merkmale bleiben identisch.

Original

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt	Arbeits-stunden	Löhne und Gehälter
01	A	34	60	100	1 903 425	17 014	711 236
02	B	28	50	67	3 313 910	27 351	515 886
03	A	17	40	25	810 729	5 601	82 438
04	C	28	40	41	794 211	7 973	191 597
05	A	29	40	40	906 228	13 990	252 841
06	C	15	70	432	20 195 146	129 338	2 679 072
...	...	...	...	...	...	...	...

Anonym

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt	Arbeits-stunden	Löhne und Gehälter	
01	A	34	60	100	1 903 425	17 014	711 236	
02	B	28	50	67	3 313 910	27 351	515 886	
03	A	17	40	25	810 729	5 601	82 438	
04	C	28	40	41	794 211	7 973	191 597	
05	A	29	40	40	906 228	13 990	252 841	
06a	C	15	60	130	6 058 544	38 801	803 722	30% von 06
06b	C	15	70	302	14 136 602	90 537	1 875 350	70% von 06
...	...	...	...	...	...	...	...	

Abbildung A15

Klonen von kleinen Merkmalsträgern

Existieren nur vereinzelt eindeutige kleine Einheiten, so besteht die Möglichkeit, durch Klonen (Erzeugung ähnlicher Einheiten) die Eindeutigkeit zu verhindern. Die stetigen Merkmale werden mit einem zufälligen und geheimen Faktor verändert, die diskreten Merkmale bleiben identisch.

Original

Betriebs-Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt	Arbeits-stunden	Löhne und Gehälter
01	A	34	60	100	1 903 425	17 014	711 236
02	B	28	50	67	3 313 910	27 351	515 886
03	A	17	40	25	810 729	5 601	82 438
04	C	28	40	41	794 211	7 973	191 597
05	A	29	40	40	906 228	13 990	252 841
06	C	15	70	432	20 195 146	129 338	2 679 072

Anonym

Betriebs-Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt	Arbeits-stunden	Löhne und Gehälter
01	A	34	60	100	1 903 425	17 014	711 236
02	B	28	50	67	3 313 910	27 351	515 886
03	A	17	40	25	810 729	5 601	82 438
04	C	28	40	41	794 211	7 973	191 597
05	A	29	40	40	906 228	13 990	252 841
06	C	15	70	432	20 195 146	129 338	2 679 072
06a	A	17	40	24	770 193	5 321	78 316 (u) 95% von 03

Datenverändernde Verfahren – Mikroaggregation

Abbildung A16

Eindimensionale Mikroaggregation nach einer Variable

Der Datenbestand wird nach dem ausgewählten Merkmal (dominierend, hoch korreliert) absteigend sortiert und dann für jeweils drei benachbarte Sätze alle stetigen Werte durch den jeweiligen Durchschnitt ersetzt.

Beispiel 1:
Es werden nur die Merkmale Arbeitsstunden und Lohnsumme bearbeitet.

Beispiel 2:
Es werden nur die Merkmale Inlands- und Auslandsumsatz bearbeitet.

• - originaler Merkmalsträger ■ - 3 anonyme Merkmalsträger vertikale Linien - Gruppengrenzen

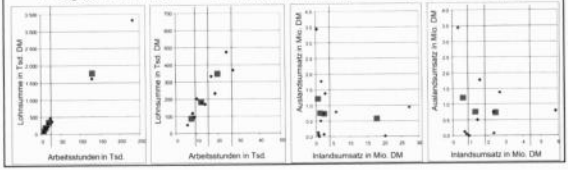


Abbildung A17

Eindimensionale Mikroaggregation nach mehreren Variablen

Es wird ein von mehreren Variablen abhängiges „künstliches" Merkmal gebildet (z.B. Z-Score als Summe der normierten Einzelwerte). Der Datenbestand wird nach diesem Merkmal absteigend sortiert und dann für jeweils drei benachbarte Sätze alle stetigen Werte durch den jeweiligen Durchschnitt ersetzt.

Beispiel 1:
Es werden nur die Merkmale Arbeitsstunden und Lohnsumme bearbeitet (Z-Score).

Beispiel 2:
Es werden nur die Merkmale Inlands- und Auslandsumsatz bearbeitet (Z-Score).

• - originaler Merkmalsträger ■ - 3 anonyme Merkmalsträger

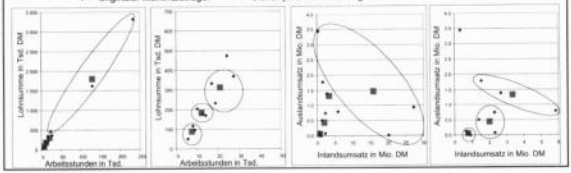


Abbildung A18

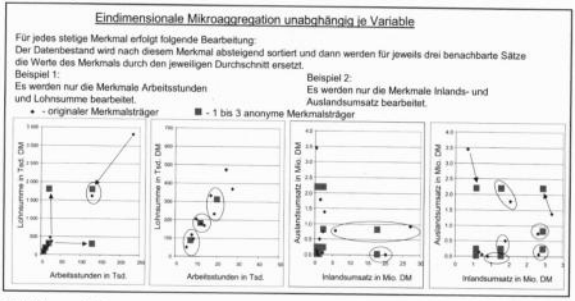


Abbildung A19

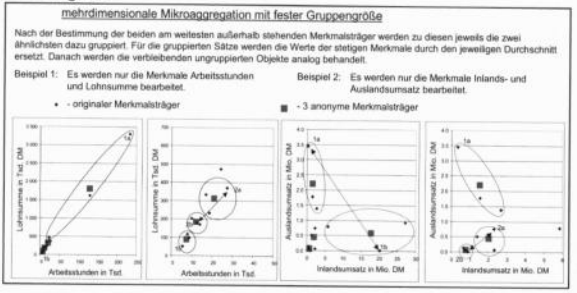


Abbildung A20

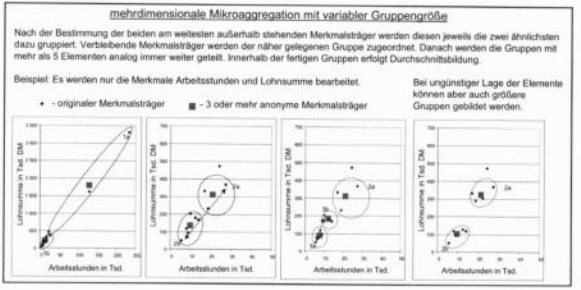


Abbildung A21

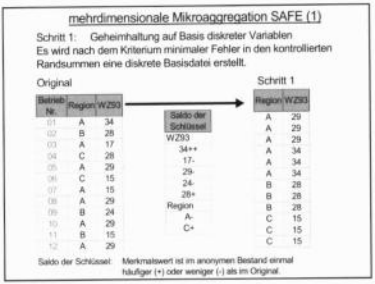


Abbildung A22

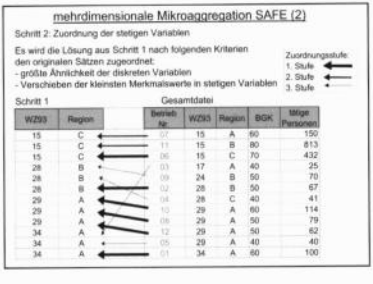


Abbildung A23

Datenverändernde Verfahren – Zufallsüberlagerung

Abbildung A24

Zufallsüberlagerung für stetige Merkmale

Generierung einer normalverteilten Zufallszahl mit dem Erwartungswert 0 und der Standardabweichung 200 000. Jedem originalen Umsatzwert, wird eine entsprechende Zufallszahl dazuaddiert.

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt	*Umsatz original*
01	A	34	60	100	1 460 705	*1 903 425*
02	B	28	50	67	3 355 041	*3 313 910*
03	A	17	40	25	918 041	*810 729*
04	C	28	40	41	743 497	*794 211*
05	A	29	40	40	1 355 569	*906 228*
06	C	15	70	432	19 938 428	*20 195 146*
07	A	15	60	150	6 953 886	*6 645 567*
08	A	29	50	79	3 610 630	*3 054 587*
09	B	24	50	70	2 292 144	*2 398 257*
10	A	29	60	114	3 990 655	*3 777 735*
11	B	15	80	813	28 119 621	*28 056 704*
12	A	29	50	62	3 673 977	*4 073 463*

Abbildung A25

Zufallsüberlagerung für diskrete Merkmale (1)

Post-Randomisierung (Post Randomisation Method (PRAM))

Bestimmung von Übergangswahrscheinlichkeiten mit denen eine diskrete Merkmalsausprägung in eine andere geändert wird. Invariante PRAM erhalten die eindimensionalen Verteilungen.

invariante PRAM am Beispiel der Region:

Region	Anzahl	Anteil	Anteil der zu verändernden Werte: 60% = 6/10 daraus ergibt sich: (Anteile)		Wahrscheinlichkeit	kumulierte Wahrscheinlichkeit
			Wert bleibt	4/10	0.40	0.40
A	7	7/12	wird zu A	6/10 * 7/12 = 7/20	0.35	0.75
B	3	3/12	wird zu B	6/10 * 3/12 = 3/20	0.15	0.90
C	2	2/12	wird zu C	6/10 * 2/12 = 2/20	0.10	1.00

Übergangswahrscheinlichkeiten:
(z.B. p_{AB} - Wahrscheinlichkeit, dass ein Merkmalsträger aus der Region A auf Region B geändert wird.)
$p_{AA} = P(Xn='A'|Xa='A') = 0.75$
$p_{AB} = P(Xn='B'|Xa='A') = 0.15$
$p_{AC} = P(Xn='C'|Xa='A') = 0.10$
...

oder Arbeitstabelle:

Entscheidungsregel für Zufallszahl z	neue Region
z < 0,40	alte Region
0,40 <= z < 0,75	'A'
0,75 <= z < 0,90	'B'
0,90 <= z	'C'

Abbildung A26

Zufallsüberlagerung für diskrete Merkmale (2)

Post-Randomisierung (Post Randomisation Method (PRAM))

Entscheidungsregel für Zufallszahl z	neue Region
z < 0,40	alte Region
0,40 <= z < 0,75	'A'
0,75 <= z < 0,90	'B'
0,90 <= z	'C'

Betrieb Nr.	Region	Zufallszahl	alte Region	WZ93	BGK	tätige Personen	Umsatz insgesamt	...
01	A	0.664	A	34	60	100	1 903 425	...
02	A	0.627	B	28	50	67	3 313 910	...
03	A	0.545	A	17	40	25	810 729	...
04	B	0.766	C	28	40	41	794 211	...
05	C	0.953	A	29	40	40	906 228	...
06	C	0.016	C	15	70	432	20 195 146	...
07	A	0.115	A	15	60	150	6 645 567	...
08	A	0.568	A	29	50	79	3 054 587	...
09	A	0.430	B	24	50	70	2 398 257	...
10	B	0.796	A	29	60	114	3 777 735	...
11	B	0.361	B	15	80	813	28 056 704	...
12	A	0.461	A	29	50	62	4 073 463	...

Datenverändernde Verfahren – Zufallsvertauschung

Abbildung A27

Zufallsvertauschung durch Data-Swapping

Die Daten werden nach inhaltlichen Gesichtpunkten gruppiert. Innerhalb der Gruppen erfolgt ein zufälliges Vertauschen der Werte.

Beispiel: Data-Swapping des Umsatzes innerhalb der Gruppen der Beschäftigtengrößenklasse

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt		Umsatz alt	Arbeits-stunden	Löhne und Gehälter
03	A	17	40	25	906 228	←	810 729	5 601	82 438
04	C	28	40	41	810 729	←	794 211	7 973	191 597
05	A	29	40	40	794 211	←	906 228	13 990	252 841
12	A	29	50	62	3 313 910	←	4 073 463	10 128	572 893
02	B	28	50	67	4 073 463	←	3 313 910	27 351	515 886
08	A	29	50	79	2 398 257	←	3 054 587	12 423	565 203
09	B	24	50	70	3 054 587	←	2 398 257	8 870	401 709
01	A	34	60	100	3 777 735	←	1 903 425	17 014	711 236
07	A	15	60	150	6 645 567	←	6 645 567	18 764	1 212 673
10	A	29	60	114	1 903 425	←	3 777 735	24 262	927 652
06	C	15	70	432	28 056 704	←	20 195 146	129 338	2 679 072
11	B	15	80	813	20 195 146	←	28 056 704	232 612	5 044 894

Abbildung A28

Zufallsvertauschung durch Rank-Swapping

Die Daten werden nach dem zu bearbeitenden Merkmal sortiert. Die Merkmalswerte werden in der Nachbarschaft zufällig getauscht.

Beispiel: Rank-Swapping des Umsatzes innerhalb der Nachbarschaft von 3 Sätzen.

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt		Umsatz alt	Arbeits-stunden	Löhne und Gehälter
04	C	28	40	41	906 228	↑	794 211	7 973	191 597
03	A	17	40	25	1 903 425	↑	810 729	5 601	82 438
05	A	29	40	40	794 211	↓	906 228	13 990	252 841
01	A	34	60	100	810 729	↓	1 903 425	17 014	711 236
09	B	24	50	70	3 777 735	↑	2 398 257	8 870	401 709
08	A	29	50	79	3 313 910	↕	3 054 587	12 423	565 203
02	B	28	50	67	3 054 587	↕	3 313 910	27 351	515 886
10	A	29	60	114	2 398 257	↓	3 777 735	24 262	927 652
12	A	29	50	62	20 195 146	↑	4 073 463	10 128	572 893
07	A	15	60	150	28 056 704	↑	6 645 567	18 764	1 212 673
06	C	15	70	432	4 073 463	↓	20 195 146	129 338	2 679 072
11	B	15	80	813	6 645 567	↓	28 056 704	232 612	5 044 894

Abbildung A29

Verfahren von Winkler

Daten, die nach einer Zufallsüberlagerung keinen ausreichenden Schutz erhalten, werden durch Zufallsvertauschung weiter anonymisiert. Andere Merkmalsträger bleiben von der Zufallsvertauschung unberührt.

Beispiel: Zufallsüberlagerung des Umsatzes mit Vertauschung der dominanten Werte

Betrieb Nr.	Region	WZ93	BGK	tätige Personen	Umsatz insgesamt		Umsatz alt	Arbeits-stunden	Löhne und Gehälter
01	A	34	60	100	1 460 705		1 903 425	17 014	711 236
02	B	28	50	67	3 355 041		3 313 910	27 351	515 886
03	A	17	40	25	918 041		810 729	5 601	82 438
04	C	28	40	41	743 497		794 211	7 973	191 597
05	A	29	40	40	1 355 569		906 228	13 990	252 841
06	C	15	70	432	28 119 621	▲	20 195 146	129 338	2 679 072
07	A	15	60	150	6 953 886		6 645 567	18 764	1 212 673
08	A	29	50	79	3 610 630		3 054 587	12 423	565 203
09	B	24	50	70	2 292 144		2 398 257	8 870	401 709
10	A	29	60	114	3 990 655		3 777 735	24 262	927 652
11	B	15	80	813	19 938 428	▼	28 056 704	232 612	5 044 894
12	A	29	50	62	3 673 977		4 073 463	10 128	572 893

Datenverändernde Verfahren – Simulationsverfahren

Abbildung A30

Latin Hypercube Sampling (LHS) (1)

1. Vorgabe der Anzahl n der Objekte für den simulierten Datenbestand.
2. Bestimmung einer Datenmatrix V mit n Sätzen mit m synthetischen Merkmalen aus der (empirischen) Verteilungsfunktion der Merkmale.
   (Die eindimensionalen Verteilungen sind somit erwartungstreu.)
3. Bestimmung der beiden Rangkorrelationsmatrizen
   T - für den orginalen Datenbestand und
   C - für die synthetische Datenmatrix V
4. Anpassung der Rangkorrelation von V an T über folgenden Algorithmus:
   a) Bestimmung der unteren Dreiecksmatrizen Q und P derart, dass $Q*Q^t=C$ und $P*P^t=T$ gilt.
   b) Bestimmung von $R = V*(P^t*Q^{-1})^t$ (mit ^{t} - transponiert und $^{-1}$ - inverse Matrix)
   c) spaltenweise Ermittlung der Rangpositionen in R
   d) Umsortieren der Werte in den Spalten in V derart, dass die Rangpositionen mit R übereinstimmen.

In Erweiterungen des LHS werden die Schritte 3 und 4 so lange wiederholt, bis der Abstand zwischen C und T sich nicht mehr als ein akzeptables Toleranzniveau verringert.

Abbildung A31

Latin Hypercube Sampling (LHS) (2)

Bei Vorgabe von 9 Objekten für den simulierten Datenbestand ergibt sich:

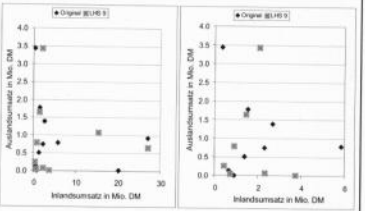


Abbildung A32

Latin Hypercube Sampling (LHS) (3)

Bei Vorgabe von 9 Objekten für den simulierten Datenbestand ergibt sich aus dem Datenbeispiel:

tätige Personen	davon Arbeiter	Umsatz Inland	Umsatz Ausland	Arbeits-stunden	Löhne	Gehälter	Anonym
41	20	447 404	253 655	6 392	66 068	80 146	
68	61	725 643	42 140	16 006	333 226	326 475	
30	26	906 228	787 084	8 571	110 443	45 820	
100	54	1 445 566	1 644 093	18 784	403 383	375 082	
338	201	2 054 951	3 439 738	95 342	1 238 408	1 031 514	
62	30	2 327 518	70 739	25 292	223 614	147 080	
126	46	3 745 758	5 224	12 945	169 741	386 450	
813	632	15 405 810	1 079 419	232 612	3 311 073	1 733 821	
76	29	27 129 609	663 035	10 128	187 058	629 772	

Ruth Brand *)

Koreferat zum Beitrag „Methoden der Anonymisierung wirtschaftsstatistischer Einzeldaten"

Einleitung

Der vorliegende Beitrag „Methoden der Anonymisierung wirtschaftsstatistischer Einzeldaten" von Jörg Höhne liefert einen systematischen Überblick über die verschiedenen Methoden, die zur Anonymisierung von Einzeldaten verwendet werden können. Damit stellt er die in der Literatur eher unverbunden nebeneinander stehenden Ansätze erstmals in einen gemeinsamen Kontext. Dieses ist eine zentrale Voraussetzung für die Untersuchung der Möglichkeiten der Anonymisierung wirtschaftsstatistischer Einzeldaten. Gleichzeitig zeigt der Aufsatz, welche Verfahren im Projekt „Anonymisierung wirtschaftsstatistischer Einzeldaten" (vgl. Gnoss 2003) näher betrachtet werden. Im folgenden werden zunächst die wesentlichen Ergebnisse des Aufsatzes kurz zusammengefasst. Anschließend wird auf einen zentralen Aspekt, die Auswahl der Verfahren näher eingegangen.

Zentrale Ergebnisse des Aufsatzes

Einleitend wird im vorliegenden Beitrag ein kurzer Überblick über zentrale Begrifflichkeiten im Kontext der Betrachtung von Anonymisierungsverfahren gegeben. Dabei wird auch auf die wichtige Unterscheidung zwischen einer technisch möglichen Zuordnung und einer Verletzung der faktischen Anonymität gemäß § 16 (6) BStatG hingewiesen. Da die faktische Anonymität eine datensatzindividuelle Abwägung von Kosten und Nutzen einer Zuordnung beinhaltet, ist eine Verletzung der Anonymität nur gegeben, wenn dem Angreifer eine „nutzbringende" Zuordnung (Re-Identifikation) gelingt[1]. Hieraus wird abgeleitet, dass die Schutzwirkung der einzelnen Verfahren sowohl auf einer Verhinderung einer eindeutigen Zuordnung als auch auf der Verringerung bzw. Verhinderung des Informationsgewinns für einen Angreifer beruht.

Für die systematische Darstellung der Verfahren werden dann folgende Kriterien herangezogen:

- Art der Informationsveränderung und
- Anzahl bzw. Anteil der Merkmalsträger, deren Angaben verändert werden.

Anhand dieser Kriterien werden die Verfahren dann in traditionelle Verfahren und „datenverändernde" Verfahren eingeteilt. Als traditionelle Verfahren werden hierbei die Verfahren bezeichnet, die vor allem eine Informationsreduktion durch das Zusammenfassen oder Unterdrücken von Merkmalswerten bzw. Merkmalsträgern beinhalten. Die

*) Ruth Brand, Statistisches Bundesamt, Wiesbaden.

1) Auch international werden nur Zuordnungen, bei denen ein Angreifer Informationen gewinnt, als Re-Identifikationen angesehen. Dieses dient hier vor allem dem Ausschluss von Selbstidentifikationen und des Spezialfalls, dass ein Angreifer über alle im Datensatz vorhandenen Merkmale bereits verfügt. – Vgl. z.B. Mokken et. al (1992).

datenverändernden Verfahren sind durch das systematische Ersetzen der Angaben der Auskunftgebenden durch andere Werte mittels verschiedener Algorithmen gekennzeichnet. Dabei steht nicht der Erhalt des vom Auskunftgebenden (Berichtspflichtigen) angegebenen Einzelwerts im Mittelpunkt. Vielmehr soll eine Datendatei erzeugt werden, deren statistische Eigenschaften möglichst ähnlich zu denen des originalen Einzeldatenmaterials sind, deren Einzelwerte sich aber von denen des originalen Einzeldatenmaterials unterscheiden. Im englischen werden diese Verfahren häufig als „perturbation" oder „masking methods" bezeichnet (z.B. Kim 1986; Willenborg/de Waal 2001).

Anschließend werden die einzelnen Verfahren überblicksartig vorgestellt, wobei der Schwerpunkt auf den für das Projekt ausgewählten Verfahren liegt. Dabei wird ihre Wirkungsweise anhand eines instruktiven Beispiels illustriert. Allerdings lässt das Beispiel keine Rückschlüsse auf die Eigenschaften der Verfahren bei realen Anwendungen zu, da der Beispieldatensatz naturgemäß wesentlich kleiner ist als reale Erhebungen. Entsprechend werden die zentralen Eigenschaften hauptsächlich auf Basis der in der Literatur zu findenden Ergebnisse referiert. Vergleicht man diese Abschnitte mit den eingangs gewählten Kriterien für die systematische Darstellung der Verfahren, zeigt sich, dass hier auch für bekannte Verfahren noch einige Lücken bestehen. Während für einige Verfahren sowohl die Schutzwirkung im Sinne von Verringerung der Zuordnungswahrscheinlichkeiten als auch die Auswirkungen auf einige Verfahren der beschreibenden und der schließenden Statistik in der Literatur ausführlich beschrieben werden, liegen für andere Verfahren nur kursorisch Ergebnisse vor. Zudem wird der im Beitrag des Autors eingangs genannte Aspekt der durch die Verfahren eingeführten Verringerung des Nutzens für einen Angreifer in der Regel nicht betrachtet, da hierzu keine Ergebnisse in der einschlägigen Literatur zu finden sind.

Insgesamt wird mit dem Beitrag eine wichtige Lücke geschlossen. Auch international liegen keine Beiträge vor, die die Neu- und Weiterentwicklungen von Anonymisierungsverfahren Verfahren ähnlich umfassend beschreiben und systematisch einordnen. Allerdings wäre auf Basis der genannten Kriterien auch eine andere Einordnung einzelner Verfahren möglich. So können die Verfahren, bei denen Variablen unterdrückt werden, sicherlich auch in die Kategorie „Informationsreduktion für Objekte" eingeordnet werden.

Kriterien zur Beurteilung von Anonymisierungsverfahren

Im Folgenden werden einige allgemeine Überlegungen zur Auswahl von Anonymisierungsverfahren vorgestellt werden. Höhne (2003) nennt in seinem Aufsatz als Gründe für die Auswahl der Verfahren im Projekt „Anonymisierung wirtschaftsstatistischer Einzeldaten":

- Leichte Handhabbarkeit des Verfahrens,
- Erfolgsaussichten der Verfahren,
- Repräsentative Vertretung der Verfahrensgruppen,
- Abhängigkeit zwischen Verfahren.

Diese Kriterien sind für die Aufgabenstellung des Projekts, die Untersuchung der Möglichkeiten der Anonymisierbarkeit wirtschaftsstatistischer Einzeldaten, sicherlich sinn-

voll, da sie eine pragmatische Herangehensweise ermöglichen. Für eine allgemeine Beurteilung von Verfahren sollte aber das Ziel der Anonymisierung, nämlich die Erstellung faktisch anonymer Daten, die wissenschaftliche Analysen nicht unangemessen einschränken, auch explizit im Zentrum der Betrachtung stehen. Hierauf aufbauend könnte z.B. folgender Kriterienkatalog angelegt werden:

- Re-Identifikationsrisiko bzw. Schutzwirkung mit den zu untersuchenden Komponenten:
  1. Untersuchung von Zuordnungswahrscheinlichkeiten zwischen den anonymisierten Daten und dem Ausgangsmaterial bei gegebenem Zusatzwissen durch Matching-Experimente und einfache Abgleiche.
  2. Analyse des Nutzens der Information, die ein Angreifer bei gelungener Zuordnung gewinnen kann, z.B. durch Verwendung von Abstandsmaßen, Schwankungsintervallen und Streuungsmaßen zur Messung der Abweichung der anonymisierten Werte von den Originalwerten.
- Analysemöglichkeiten der Wissenschaft mit den zu untersuchenden Komponenten:
  1. Veränderung der multivariaten Verteilung, z.B. durch empirische Vergleiche der Verteilung.
  2. Auswirkungen der Anonymisierung auf uni- und multivariate statistische Analysen, insbesondere auch auf ökonometrische Modelle, durch theoretische und empirische Analysen.
  3. Entwicklung von Kenngrößen (z.B. Erhalt der ersten und zweiten Momente, Anzahl der veränderten Merkmalswerte).

Zur Bestimmung von Zuordnungswahrscheinlichkeiten sind im Kontext der Anonymisierung von Einzeldaten bereits umfangreiche Arbeiten durchgeführt worden (z.B. Müller et. al 1991; Winkler 1998; Domingo-Ferrer/Mateo/Torres 2003). Dagegen sind systematische Analysen des Nutzens der bei gelungener Zuordnung gewonnenen Informationen nicht zu finden und Untersuchungen zu den verbleibenden Analysemöglichkeiten finden sich bisher nur für einzelne Verfahren. Dieses ist zum Teil sicherlich auf die Abhängigkeit der Wirkungen der Verfahren von den Ausgangsdaten zurückzuführen. Allerdings ist auch zu konstatieren, dass sich auch in der wissenschaftlichen Diskussion bisher keine klaren, allgemein akzeptierten Kriterien für die für Analysen notwendigen Eigenschaften herauskristallisiert haben. Dieses zeigt sich auch an der durchaus kontroversen Diskussion um die Entwicklung von Scores zur Beurteilung von Anonymisierungsverfahren (siehe auch Domingo-Ferrer/Mateo/Torres 2003 sowie Giessing 2003).

Die Beurteilung der Analysemöglichkeiten kann nicht allein auf Basis des in den statistischen Ämtern vorhandenen Wissens über die Nachfrage nach Daten erfolgen. Der Grund hierfür ist, dass den Ämtern in der Regel nicht bekannt ist, welche Analysen die Nutzer der Daten mit diesen durchführen. Zudem ist zu beachten, dass für unterschiedliche Analyseverfahren unterschiedliche Anonymisierungsverfahren geeignet sein können (vgl. z.B. Lechner/Pohlmeier 2003).

Die Verfahren, die aufgrund der obigen Kriterien generell zur Erreichung der Ziele einer Anonymisierung geeignet sind, sind dann auf ihre praktische Anwendbarkeit hin zu überprüfen. Dabei sind zwei Aspekte wichtig:

1. Die Überschaubarkeit der Verfahren und ihrer Eigenschaften (Methodentransparenz).
2. Die technische und personelle Durchführbarkeit.

Methodentransparenz ist eine zentrale Voraussetzung für die Akzeptanz von Verfahren in der Praxis. Transparenz der Methoden muss dabei sowohl für den Anwender in den Statistischen Ämtern und bei anderen datenproduzierenden Institutionen als auch für den Nutzer der anonymisierten Daten vorhanden sein. Daten, die mit Verfahren anonymisiert werden, die für die Mehrzahl der potentiellen Nutzer nicht durchschaubar sind, finden in der Regel nur eine sehr geringe Nutzerzahl. Verfahren, bei denen der Datenproduzent das Erreichen der faktischen Anonymität nicht prüfen kann, sind dagegen bei den Datenproduzenten nicht anwendbar. Mit letzterem ist der zweite Punkt, die technische und personelle Durchführbarkeit bei den Datenproduzenten, verbunden. Diese schließt aber auch die Programmverfügbarkeit und Handhabbarkeit sowie das Vorhandensein von ausreichend qualifiziertem Personal bei den Datenproduzenten ein. Hierbei sind dann auch die mit der Anonymisierung verbundenen Kosten zu beachten. Verfahren, die nur mit einem hohen Aufwand und damit hohen Kosten verbunden sind, können zu Preisen für anonymisierte Daten führen, die für die potentiellen Nutzer nicht mehr tragbar sind.

Zusammenfassung und Ausblick

Der referierte Beitrag gibt erstmals einen systematischen Überblick über die verschiedenen Methoden, die zur Anonymisierung von wirtschaftsstatistischen Einzeldaten verwendet werden können. Er stellt die in der Literatur eher unverbunden nebeneinander stehenden Ansätze in einen gemeinsamen Kontext. Die Beschreibung der Eigenschaften zeigt aber auch, dass die beschriebenen Verfahren noch nicht abschließend untersucht worden sind. Insbesondere die abschließende Bewertung der Verfahren durch die Nutzer und eine hierauf aufbauende Ableitung von Kriterien für die Einschränkungen des Analysepotentials müssen leider offen bleiben. Erste Hinweise werden auch für die Auswahl von Anonymisierungsverfahren in der Praxis gegeben. Für die Entwicklung von Strategien, wie wirtschaftsstatistische Daten anonymisiert werden können, sind aber vertiefende Untersuchungen mit Erhebungen aus der amtlichen Statistik notwendig, da nur diese eine fundierte Einschätzung der Eigenschaften der Verfahren und ihre Anwendbarkeit ermöglichen.

Literaturhinweise

Gießing, S. (2003): Koreferat zum Beitrag „Konzepte zur Bewertung von anonymisierten Datensätzen" von J. Domingo-Ferrer, J. M. Mateo und A. Torres.(In diesem Band S. 111 ff.)

Gnoss, R. (2003): Einführung – Das Projekt der statistischen Ämter: Ziele, Ablauf, Beteiligte – Problematik des Datenschutzes. (In diesem Band S. 6 ff.)

Höhne, J. (2003): Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten.(In diesem Band S. 69 ff.)

Kim, J. J. (1986): A Method for Limiting Disclosure in Microdata based on Random Noise and Transformation, in: American Statistical Association (Hrsg.): Proceedings of the Section on Survey Research Methods 1986, Washington D.C., pp. 303 – 308.

Lechner, S. und Pohlmeier, W. (2003): Anonymisierungsmethoden und ökonometrische Modelle. (In diesem Band S. 115 ff.)

Mokken, R. J.; Kooiman, P.; Pannekoek, J. und Willenborg, L.C.R.J. (1992): Disclosure Risks for Microdata, Statistica Neerlandica 46, S. 49 – 67.

Müller, W.; Blien, U.; Knoche, P.; Wirth, H. u.a. (1991): Die faktische Anonymität von Mikrodaten, in: Statistisches Bundesamt (Hrsg.): Schriftenreihe Forum der Bundesstatistik, Band 19, Stuttgart, Metzler-Poeschel.

Domingo-Ferrer, J.; Mateo, J. M. und Torres, A. (2003): Concepts for the Evaluation of Anonymized Data. (In diesem Band S. 100 ff.)

Winkler, W. (1998): Re-Identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata, Research in Official Statistics, Vol 1 (2), pp. 50 – 69.

Willenborg, L.C.R.J. und de Waal, T. (2001): Elements of Statistical Disclosure Control, Lecture Notes in Statistics Bd. 155, Springer-Verlag, Heidelberg, New York.

Josep Domingo-Ferrer/Josep M. Mateo-Sanz/Àngel Torres *)

Concepts for the Evaluation of Anonymized Data

Abstract: We present in this paper some criteria for empirical comparison of SDC methods for continuous microdata. Based on re-identification experiments, we try to optimize the tradeoff between information loss and disclosure risk. SDC methods compared include additive noise, distortion by probability distribution, microaggregation, resampling, rank swapping and a novel approach based on lossy compression. Generic information loss measures (not targeted to specific data uses) are defined, and two approaches to empirical re-identification are used: Euclidean record linkage and probabilistic record linkage. Some weighting schemes to aggregate information loss and disclosure risk measures are discussed and empirical results are given for one of them.

Keywords: Statistical disclosure control, Continuous microdata, Record linkage, Re-identification experiments, Information loss measures.

1 Introduction

This paper describes an empirical approach to evaluating the protection of statistical microdata. This work was started in the context of the U.S. Census Bureau contract OTTILIE-R (Optimizing the Tradeoff between Information Loss and dIsclosurE risk for continuous microdata) and continued during the European project CASC (Computational Aspects of Statistical Confidentiality). The idea is to define information loss and disclosure risk measures and then construct a score which combines both types of measures. Specifically, the following steps have been taken in this work:

- *Literature analysis.* Literature on SDC for microdata has been analyzed to identify those methods which are relevant for protecting continuous data. In addition, SDC of continuous microdata based on lossy compression has been introduced.
- *Test data.* Test data have been obtained from publicly available microdata files.
- *Disclosure risk assessment.* Two record linkage algorithms have been used to establish the disclosure risk associated to a particular SDC method. In addition, an interval disclosure measure has been defined.
- *Metrics definition.* Information loss actually depends on the data uses to be supported by masked data. Since data uses did not fall within the scope of OTTILIE-R, we have defined a battery of generic, robust information loss metrics which try to capture structural differences between the original and masked data files.
- *Empirical work.* Experiments carried out are directed to obtaining t-uples of the form *(method, parms, risk, loss)*, where *parms* are the input parameters to *method*, *risk* is the percent of re-identified records in the test data set and *loss* is the information loss. The obtained t-uples can be aggregated to rank methods; depending on the aggregation used, different method rankings are conceivable.

*) Dr. Josep Domingo-Ferrer/Dr. Josep M. Mateo-Sanz/Àngel Torres, Universitat Rovira i Virgili, Tarragona, Catalonia, Spain, E-mail jdomingo@etse.urv.es.

Section 2 reviews relevant SDC methods for the protection of continuous microdata. Section 3 lists information loss measures which have been taken into account in experimentation. Section 4 describes record linkage approaches to assessing disclosure risk. Section 5 reports on actual comparison results. Section 6 discusses alternative score constructions for aggregating information loss and disclosure risk. Section 7 is a conclusion.

2 Relevant SDC methods for continuous microdata

Sampling methods consist of publishing a sample of records from the original microdata set instead of publishing the whole original microdata set. Sampling methods are suitable for categorical microdata, but their adequacy for continuous microdata is less clear in a general disclosure scenario. The reason is that such methods leave a continuous variable V unperturbed for all individuals in the sample. Thus, if variable V is present in an external administrative public file, unique matches with V' (the version of variable V restricted to the published sample) are likely, since it is unlikely that a continuous variable (even one truncated due to digital representation) takes exactly the same value for more than one individual. Thus, we will concentrate in what follows on perturbative methods, which have the additional advantage of allowing the entire microdata set to be released.

Perturbative methods distort the microdata set before publication. Perturbative methods considered in our work are a subset of those making sense for continuous microdata:

- *Additive noise* (Noise*p* for short). Gaussian noise is added to the original data to get the masked data (Kim 1986). If the standard deviation of the original variable is s, noise is generated using a $N(0,ps)$. Values of p considered in the experiments below are 0.01, 0.02, 0.04, 0.06, 0.08 up to 0.2 with 0.02 increments.
- *Data distortion by probability distribution* (Distr for short,(Liew/Choi/Liew 1985)). For each variable in the original variable, the best fitted distribution is found; then the fitted distribution is used to generate the masked data set. There are no parameters.
- *Resampling*. Originally proposed for protecting tabular data (Domingo-Ferrer/Mateo-Sanz 1999; Heer 1993), resampling can also be used for microdata. Let V be an original variable in a dataset with n records. Take with replacement t independent samples $X_1, ..., X_t$ of size n of the values of V. Independently rank each sample (using the same ranking criterion for all samples). Finally, for $j=1$ to n, compute the j-th value v'_j of the masked variable V' as the average of the j-th ranked values in $X_1, ..., X_t$. Resampling has been tested for $t=1$ (Resamp1) and $t=3$ (Resamp3).
- *Microaggregation*. Records are clustered into small aggregates or groups of size at least k (Defays/Nanopoulos 1993; Domingo-Ferrer/Mateo-Sanz 2002). Rather than publishing a variable for a given individual, the average of the values of the variable over the group to which the individual belongs is published. Variants of microaggregation considered include: individual ranking (MicIR*k*); microaggregation on projected data using z-scores projection (Mic*Zk*) and principal components projection (MicPCP*k*); microaggregation on unprojected multivariate data considering two variables at a time (Mic2mul*k*), three variables at a time (Mic3mul*k*), four variables at a time (Mic4mul*k*) or all variables at a time (Micmul*k*). Values of k between 3 and 10 have been considered.

- *Lossy compression* (JPEG*q*). This method is new and proposed by these authors for continous data. The idea is to regard a numerical microdata file as an image (with rows being records and columns being variables). Lossy compression, and more specifically the JPEG algorithm (Joint Photographic Experts Group, Standard IS 10918-1 (ITU-T T.81), http://www.jpeg.org), is then used on the image, and the compressed image is interpreted as a masked microdata file. Depending on the lossy compression algorithm used, appropriate mappings between variable ranges and color scales will be needed. The JPEG quality q has been taken as a parameter with values from 5 % up to 100 % with 5 % increments.
- *Rank swapping* (Rank*p*). Although originally described only for ordinal variables, this method can be used for any numerical variable (Moor 1996). First values of variable V_i are ranked in ascending order; then each ranked value of V_i is swapped with another ranked value *randomly* chosen within a restricted range (e.g. the rank of two swapped values cannot differ by more than p % of the total number of records). The following values of p have been considered in experimentation: 1, 2, 3, 4, 5, 6, 7 and 10.

3 Information loss measures

To evaluate the information loss caused by an SDC method on a continuous microdata set, we want to assess how different the masked data set is from the original data set. We will say there is little information loss if the structure of the masked data set is very similar to the structure of the original data set. In fact, the motivation for preserving the structure of the data set is to ensure that the masked data set will be analytically valid and interesting. We can actually try several complementary ways to assess the preservation of the structure of the original data set:

1. Compare the data in the original and the masked data sets. The more similar the SDC method to the identity function, the less impact (but the higher the disclosure risk!).
2. Compare some statistics computed on the original and the masked data sets.

Let X and X' be the original and the masked data set. Let V and V' be the covariance matrices of X and X', respectively; similarly, let R and R' be the correlation matrices. Table 1 summarizes the measures proposed. In this table, p is the number of variables, n the number of records, and components of matrices are represented by the corresponding lowercase letters (e.g. x_{ij} is a component of matrix X). Regarding X –X' measures, it also makes sense to compute those on the averages of variables rather than on all data (see the $\bar{X} - \bar{X}'$ row in Table 1). Similarly, for V – V' measures, it is also sensible to compare only the variances of the variables, *i.e.* to compare the diagonals of the covariance matrices rather than the whole matrices (see the S – S' row in Table 1).

Table 1: Information loss measures

	Mean square error	Mean abs. Error	Mean variation
X-X'	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
$\bar{X}-\bar{X}'$	$\frac{\sum_{j=1}^{p}(\bar{x}_j-\bar{x}'_j)^2}{p}$	$\frac{\sum_{j=1}^{p}\lvert\bar{x}_j-\bar{x}'_j\rvert^2}{p}$	$\frac{\sum_{j=1}^{p}\frac{\lvert\bar{x}_j-\bar{x}'_j\rvert}{\lvert\bar{x}_j\rvert}}{p}$
V-V'	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
S-S'	$\frac{\sum_{j=1}^{p}(v_{jj}-v'_{jj})^2}{p}$	$\frac{\sum_{j=1}^{p}\lvert v_{jj}-v'_{jj}\rvert}{p}$	$\frac{\sum_{j=1}^{p}\frac{\lvert v_{jj}-v'_{jj}\rvert}{\lvert v_{jj}\rvert}}{p}$
R-R'	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i<j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$

4 Disclosure risk measures

The assessment of the quality of an SDC method cannot be limited to information loss; disclosure risk is another magnitude that should be measured. The method that optimizes the tradeoff between both magnitudes subject to some user requirements turns out to be the best option.

Literature on disclosure risk is basically related to sampling methods, in which a sample of the original data set is published. Disclosure risk here is measured as the probability that a sample unique is a population unique (Skinner/Marsh/Openshaw/Wymer 1994). If the size of the sample is similar to the size of the whole population, such a probability can be dangerously high; in that case, an intruder who locates a unique value in the released sample could be almost sure that there is a single individual in the population with that value. This could lead to identification of that individual.

The uniqueness property as stated above is no longer relevant for perturbative methods, since in this case the whole microdata set is published, but with some distortion. There is not much literature on disclosure risk that can be used for a broad class of perturbative methods; disclosure risk measures tend to be method-specific (measures described in (Adam/Wortmann 1989) are still up-to-date). Empirical methods, like record

linkage techniques, provide a more unified approach to disclosure risk assessment for perturbative methods. We briefly describe below two approaches to record linkage and one measure of interval disclosure.

4.1 Distance-based record linkage

This approach to record linkage is described in (Pagliuca/Seri 1998) for the specific case of microaggregation masking and using the Euclidean distance. However, it can be generalized for any perturbative method provided that a distance between the original and the masked value can be defined. As in any record linkage context, it is assumed that an intruder has an external data set containing as key variables some of the same variables present in the released masked data set. The intruder is assumed to try to link the masked data set with the external data set.

Linkage then proceeds by computing the distances between records in the original and the masked data sets. The distances used are standardized to avoid scaling problems. For each record in the masked data set, the distance to every record in the original data set is computed. Then the "nearest" and "second nearest" records in the original data set are considered. A record in the masked data set is labelled as "linked" when the nearest record in the original data set has the same record number is the corresponding original record). A record in the masked data set is labelled as "linked to 2nd nearest" when the second nearest record in the original data set has the same record number. In all other cases, a record in the masked data set is labelled as "not linked". The percent of "linked" and "linked to 2nd nearest" is a measure of disclosure risk.

4.2 Probabilistic record linkage

In (Jaro 1989), a probabilistic record linkage method was described and illustrated on the 1985 Census of Tampa, Florida. The matching algorithm uses the linear sum assignment model to "pair" records in the two files to be matched (the original file and the masked file in our case). The percent of correctly paired records is a measure of disclosure risk.

Although less simple than the Euclidean method described in the previous section, this approach is attractive because it only requires the user to provide two probabilities as input: one is an upper bound of the probability of a false match, and the other an upper bound of the probability of false non-match. The Euclidean method above requires rescaling variables as well as an assumption on the weight of variables when computing a distance: for instance, in the proposal of (Pagliuca/Seri 1998), all variables have the same weight.

The U.S. Census Bureau implementation of probabilistic record linkage provided by W. Winkler (U.S. Bureau of Census 2000; Winkler 1998) has been used (with some additions) in the experimentation.

4.3 Interval disclosure

For a record in the masked data set, take a rank interval centered on the values of that record as follows: each variable is independently ranked and a rank interval is defined around the value the variable takes on each record; the ranks of values within the inter-

val for a variable around record r should differ less than p % of the total number of records and the rank in the center of the interval should correspond to the value of the variable in record r. Then the measure is the proportion of original values which fall into the interval centered around their corresponding masked value. A 100 % proportion means that an intruder is completely sure that the original value lies in the interval around the masked value (interval disclosure). Values of p ranging between 1 % and 10 % have been considered for experimentation.

5 Comparison results

A microdata set was constructed using the Data Extraction System (DES) of the U.S. Census Bureau (http://www.census.gov/DES). 13 continuous variables were chosen and 1 080 records were selected so that there were not many repeated values for any of the variables (in principle, one would not expect repeated values for a continuous variable, but there were repetitions in the data set). Table 2 contains a ranking of methods described in Section 2 (the parameter values described in that section were tried for each method). The Information Loss column (IL) is computed by averaging the mean variations of X-X', $\overline{X}$ - $\overline{X}$ ',V-V',S-S' and the mean absolute error of R-R'; the resulting average has been multiplied by 100. The Distance Linkage Disclosure risk column (DLD) contains the average percent of linked records using distance-based record linkage; the average is computed over the number of key variables that the intruder is assumed to know (we have considered knowledge of 1 up to 7 variables). Similarly, the Probabilistic Linkage Disclosure risk column (PLD) is the average percent of correctly paired records using probabilistic linkage. The Interval Disclosure (ID) column contains the average percent of original values falling in the intervals around their corresponding masked values (averages have been computed over all parameter values, i.e. 1 % to 10 % with 1 % increments). Finally, the column Score has been used to rank Table 2 and has been computed as

$$\text{Score} = 0.5 \cdot \text{IL} + 0.125\ \text{DLD} + 0.125 \cdot \text{PLD} + 0.25 \cdot \text{ID}.$$

The rationale of the above weighting is to give equal weight to information loss (0.5) and to disclosure risk. The 0.5 weight of disclosure risk is equally divided among ID (0.25) and record linkage. The 0.25 weight of record linkage is equally divided among both approaches to record linkage. The correlation between DLD and PLD is actually 0.962, so both approaches are very similar. The (IL,DLD), (IL,PLD) and (IL,ID) correlations are −0.605, -0.551 and −0.807; thus, the lower the information loss, the higher the disclosure risk, as one would expect. The IL Rank, DLD Rank, PLD Rank and ID Rank columns contain the ranking of each method with respect to IL, DLD, PLD and ID; the lower the rank, the better a method performs (i.e. lower information loss and disclosure risk).

Table 2: Comparison results

Method	IL	DLD	PLD	ID	Score	IL Rank	DLD Rank	PLD Rank	ID Rank
Rank15	19.01	1.19	0.15	35.05	18.44	53	6	7	21
Rank19	22.95	0.93	0.08	28.04	18.61	59	2	2	2
Rank16	20.91	1.39	0.11	32.18	18.69	56	8	5	16
Rank13	16.77	2.17	0.12	40.35	18.76	48	12	6	28
Rank14	19.72	1.92	0.07	37.00	19.36	55	10	1	25
Rank11	14.32	2.43	0.25	47.81	19.45	44	13	14	39
Rank12	16.37	2.50	0.25	43.73	19.46	47	14	11	35
Rank20	25.81	0.69	0.09	26.83	19.71	64	1	3	1
Rank18	25.74	0.95	0.09	29.25	20.31	63	4	4	6
Rank10	13.37	3.90	0.38	53.17	20.51	41	24	17	45
Rank17	25.12	1.52	0.20	30.95	20.51	61	9	9	10
Rank09	11.66	5.01	0.52	57.58	20.91	38	37	29	49
Rank08	11.60	6.07	0.85	63.37	22.51	37	39	39	56
Rank07	9.25	7.51	1.08	68.71	22.87	30	41	43	63
Rank06	7.87	9.02	2.79	73.80	23.86	26	43	56	71
Mic3mul07	11.06	19.34	4.70	72.34	26.62	36	68	65	69
Rank05	6.78	16.80	13.60	78.89	26.91	22	58	70	77
Mic3mul09	13.46	19.22	3.44	69.91	27.04	42	67	60	65
Mic3mul10	14.84	17.99	3.44	68.61	27.25	46	64	59	62
Mic4mul04	12.14	19.76	6.67	71.85	27.33	39	69	68	68
Mic4mul05	14.50	17.43	5.45	69.09	27.39	45	61	66	64
Mic3mul08	13.51	20.81	4.15	70.68	27.54	43	71	63	66
Mic4mul08	18.89	17.78	3.35	62.84	27.80	52	62	58	55
Mic3mul06	10.24	20.41	13.90	74.00	27.91	33	70	71	72
Mic4mul07	19.36	17.10	2.08	64.41	28.18	54	60	53	58
Mic4mul06	17.91	17.82	3.98	66.41	28.28	50	63	62	60
Mic4mul09	21.35	15.93	2.00	61.66	28.33	58	57	52	54
Mic4mul10	22.98	16.85	2.37	60.56	29.03	60	59	55	51
Mic3mul05	9.73	23.78	18.29	76.59	29.27	31	76	73	74
Mic3mul04	7.45	23.49	22.75	79.14	29.29	24	75	75	79
Mic4mul03	10.69	22.88	16.69	76.89	29.51	35	74	72	75
Rank04	5.90	22.77	22.78	84.12	29.67	20	73	76	86
Micmul03	27.67	14.26	1.88	57.23	30.16	65	54	50	47
Micmul04	31.74	13.72	1.38	52.44	30.86	67	53	48	44
Mic3mul03	6.29	29.70	29.06	82.95	31.23	21	79	80	85
Micmul05	35.12	11.73	1.14	48.43	31.27	70	46	44	41
Micmul07	37.68	13.20	1.20	43.46	31.50	72	52	45	34
Micmul06	38.77	13.00	1.22	45.76	32.60	73	50	46	37
Micmul08	41.53	13.12	0.99	42.66	33.19	75	51	42	32
Rank03	5.07	31.73	36.92	89.53	33.50	18	80	83	93
Mic2mul10	10.68	49.38	27.29	77.43	34.28	34	86	78	76
Micmul10	44.69	14.66	0.50	40.41	34.34	76	55	27	29
Noise0.16	32.56	15.65	4.66	64.39	34.91	68	56	64	57
Micmul09	45.98	12.82	0.85	40.99	34.95	79	49	40	30

Table 2: Comparison results

Method	IL	DLD	PLD	ID	Score	IL Rank	DLD Rank	PLD Rank	ID Rank
Mic2mul09	9.93	51.03	33.04	78.94	35.21	32	87	81	78
Mic2mul08	8.55	54.31	33.70	79.77	35.22	27	88	82	80
Mic2mul07	7.53	54.72	37.41	81.40	35.63	25	89	84	83
Noise0.12	25.24	22.21	22.39	71.58	36.09	62	72	74	67
Noise0.1	21.14	27.70	29.03	75.20	36.46	57	78	79	73
Mic2mul06	7.03	56.38	42.00	82.89	36.54	23	90	86	84
JPEG080	33.97	19.13	6.93	66.35	36.83	69	65	69	59
Noise0.14	35.13	19.21	6.24	67.62	37.65	71	66	67	61
Noise0.18	41.12	11.96	3.52	60.95	37.73	74	47	61	52
Noise0.08	17.43	36.06	39.76	79.84	38.15	49	82	85	81
Rank02	2.90	47.26	57.47	94.56	38.18	11	85	90	96
JPEG070	44.92	9.66	2.34	57.28	38.28	77	44	54	48
Noise0.2	45.97	10.01	0.97	57.63	38.77	78	45	41	50
Mic2mul05	5.88	58.97	56.84	85.40	38.77	19	92	89	88
JPEG085	29.47	23.85	24.48	72.80	38.98	66	77	77	70
Mic2mul04	4.90	61.53	60.69	87.26	39.54	17	94	91	89
JPEG090	18.17	35.37	46.98	80.87	39.60	51	81	87	82
Noise0.06	13.03	45.54	56.22	84.16	40.28	40	84	88	87
Mic2mul03	3.28	66.97	64.79	90.51	40.74	15	95	92	94
Noise0.04	8.93	58.51	65.28	88.95	42.18	28	91	94	90
JPEG075	50.45	12.67	2.90	61.27	42.49	80	48	57	53
JPEG095	9.06	60.11	66.56	89.23	42.67	29	93	96	92
Resamp3	3.15	67.90	67.63	96.81	42.72	14	96	97	97
Rank01	2.34	69.19	66.35	99.54	43.00	9	97	95	106
JPEG065	57.77	7.02	1.90	53.87	43.47	81	40	51	46
Noise0.02	4.24	77.34	71.32	94.42	44.31	16	99	98	95
Resamp1	3.11	75.42	71.85	98.36	44.56	13	98	99	99
MicPCP03	69.62	3.16	0.77	38.41	44.90	84	17	38	26
JPEG055	63.70	5.57	1.26	49.70	45.13	83	38	47	42
Noise0.01	2.57	85.19	74.13	97.03	45.46	10	100	103	98
JPEG100	3.06	87.14	73.03	99.14	46.34	12	101	100	101
MicIR10	1.19	97.37	74.07	99.12	46.81	8	102	102	100
MicIR08	1.03	97.84	74.07	99.29	46.83	6	108	101	103
MicIR09	1.14	97.96	74.40	99.24	46.93	7	109	104	102
MicIR06	0.87	97.66	75.28	99.51	46.93	5	106	105	105
MicIR05	0.69	97.58	75.99	99.58	46.94	3	104	106	107
MicIR03	0.45	97.39	78.96	99.79	47.22	1	103	107	109
MicIR04	0.64	97.63	79.78	99.67	47.41	2	105	108	108
MicIR07	0.81	97.79	88.06	99.42	48.49	4	107	109	104
MicPCP04	78.84	3.43	0.62	36.00	48.92	87	19	32	23
JPEG050	73.20	4.26	0.67	47.96	49.21	86	31	36	40
JPEG060	71.24	7.66	1.52	51.71	49.69	85	42	49	43
MicPCP05	82.55	3.94	0.69	34.10	50.38	88	25	37	20
MicPCP07	89.28	4.02	0.62	32.56	53.36	91	27	33	17

Table 2: Comparison results

Method	IL	DLD	PLD	ID	Score	IL Rank	DLD Rank	PLD Rank	ID Rank
MicPCP09	**90.78**	4.54	**0.25**	31.40	**53.84**	94	**34**	12	**13**
MicPCP06	**90.26**	3.37	**0.50**	33.42	**53.97**	93	**18**	26	**19**
MicZ03	**90.25**	3.16	**0.61**	35.71	**54.52**	92	**16**	31	**22**
JPEG035	**88.80**	3.65	**0.44**	43.20	**55.71**	90	**20**	23	**33**
JPEG045	**87.55**	4.15	**0.67**	46.78	**56.07**	89	**30**	35	**38**
MicZ04	**94.94**	3.70	**0.53**	33.04	**56.26**	96	**21**	30	**18**
MicPCP08	**96.93**	3.97	**0.34**	32.04	**57.02**	97	**26**	16	**14**
MicPCP10	**97.82**	4.13	**0.46**	31.19	**57.28**	98	**29**	24	**11**
JPEG040	**90.99**	3.72	**0.66**	44.98	**57.29**	95	**22**	34	**36**
MicZ07	**102.87**	4.27	**0.38**	30.53	**59.65**	99	**32**	20	**9**
MicZ06	**103.92**	3.88	**0.41**	30.43	**60.10**	100	**23**	21	**8**
MicZ05	**104.06**	4.03	**0.42**	31.30	**60.41**	101	**28**	22	**12**
MicZ08	**107.92**	4.55	**0.52**	29.60	**61.99**	102	**35**	28	**7**
MicZ10	**109.79**	4.83	**0.38**	28.20	**62.59**	103	**36**	18	**3**
MicZ09	**110.91**	4.35	**0.38**	28.36	**63.14**	105	**33**	19	**4**
Distr	**58.62**	43.05	**64.88**	88.98	**65.04**	82	**83**	93	**91**
JPEG030	**110.48**	3.02	**0.48**	41.79	**66.12**	104	**15**	25	**31**
JPEG025	**155.15**	2.13	**0.25**	38.76	**87.56**	106	**11**	13	**27**
JPEG020	**164.91**	1.36	**0.29**	36.11	**91.69**	107	**7**	15	**24**
JPEG015	**202.66**	1.10	**0.15**	32.06	**109.50**	108	**5**	8	**15**
JPEG010	**269.38**	0.93	**0.22**	28.44	**141.94**	109	**3**	10	**5**

6 Alternative score constructions

In Torres (2003), an alternative score construction is proposed which can be used instead of the one proposed in the previous section. Like in the previous score, information loss and disclosure risk are both assigned 0.5 weight. The differences are in the way of information loss and disclosure risk are computed.

IL is computed as the 100 times the average of IL1, IL2 and IL3, where IL1 is the mean variation of X-X'; IL2 is the average of the mean variation of $\overline{X} - \overline{X}'$ and the mean variation of S-S'; and IL3 is the average of the mean variation of V-V' and the mean absolute error of R-R'. Thus, IL1 represents the discrepancy between raw original and raw protected data, IL2 represents the discrepancy between univariate statistics for original and protected data, while IL3 represents the discrepancy between bivariate statistics for original and protected data. The philosophy here is to give equal weight to those three discrepancies.

Disclosure risk assessment does not use probabilistic record linkage (only distance-based record linkage DLD is used), which allows larger experiments to be carried out in less time. Two kinds of interval disclosure are distinguished: ID1, which is based on ranks and exactly corresponds to ID as defined in Sections 4.3 and used in Section 5, and ID2, which is based on standard deviations. ID2 is defined as the average percent of original values falling in the intervals around their corresponding masked values, where

intervals are centered on the actual masked values (not on their rank) and interval widths are p % of the standard deviation of the variable.
Thus, this score can be formalized as:

$$\text{Score2} = 0.5 \cdot ((\text{IL1}+\text{IL2}+\text{IL3})/3) + 0.25 \cdot \text{DLD} + 0.125 \cdot \text{ID1} + 0.125 \cdot \text{ID2}$$

Results obtained with this new score are similar to those reported in Section 5 and Table 2 regarding the relative ranking of families of methods. Rank swapping scores best, closely followed by microaggregation. The differences show up in the best scoring parameterizations: while in Table 2, Rank 15 was best, the best method with Score2 is Rank07. Also, with Score2 microaggregation Micmul 16 and Micmul 10 appear as the fourth and eighth best scoring methods, respectively.

7 Conclusions

There is a rich array of methods for microdata disclosure limitation. A set of proposals for continuous microdata have been identified and described in this paper. Measures for assessing information loss have also been described. Experimental results presented in Table 2 are self-explanatory. One thing that stands out is that rankswapping with parameter around 10 % is a very good option; next follows multivariate microaggregation taking groups of three or four variables at a time; for microaggregation, the group size has no significant effect. Data distortion by probability distribution turns out to perform very poorly. For most methods, performance depends on parameter choice, even if some methods are more parameter-dependent than other.

Changing the way the score is constructed can change the experimental results to a limited extent: the relative ranking between families of methods (rank swapping, microaggregation, etc.) is likely to stay the same as long as the score strikes a balance between information loss and disclosure risk.

Acknowledgments

This work was partly funded by the U.S. Bureau of the Census under contract OTTILIE-R (ref. no OBLIG-2000-29158-0-0) and by the European Commission under project CASC (ref. no. IST-2000-25069). Thanks go to Francesc Sebé for his help in automating the probabilistic record linkage software and running the experiments. The contribution of Sarah Giessing as Ko-Referent is also gratefully acknowledged.

References

Adam, N. R.; Wortmann, J. C., (1989): Security-control methods for statistical databases: a comparative study, ACM Computing Surveys, vol. 21(4), pp. 515 – 556.

Defays, D.; Nanopoulos, P., (1993): Panels of enterprises and confidentiality: the small aggregates method, in Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa: Statistics Canada, pp. 195 – 204.

Domingo-Ferrer, J.; Mateo-Sanz, J. M., (1999): On resampling for statistical confidentiality in contingency tables, Computers & Mathematics with Applications, vol 38, pp. 13 – 32.

Domingo-Ferrer, J.; Mateo-Sanz, J. M., (2002): Practical Data-Oriented Microaggregation for Statistical Disclosure Control, IEEE Transactions on Knowledge and Data Engineering, vol 14(1), pp. 189 – 201.

Heer, G. R., (1993): A bootstrap procedure to preserve statistical confidentiality in contingency tables, in Proceedings of the International Seminar on Statistical Confidentiality (ed. D. Lievesley), Luxemburg: Office for Official Publications of the European Communities, pp. 261 – 271.

Jaro, M. A., (1989): Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association, vol. 84, pp. 414 – 420.

Joint Photographic Experts Group: Standard IS 10918-1 (ITU-T T.81) http://www.jpeg.org.

Kim, J. J. (1986): A method for limiting disclosure in microdata based on random noise and transformation, in Proc. of the ASA Sect. on Survey Res. Meth., pp. 303 – 308.

Liew, C. .; Choi, U. J.; Liew, C. J. (1985): A data distortion by probability distribution, ACM Transactions on Database Systems, vol. 10, pp. 395 – 411.

Moore, R., (1996): Controlled data swapping techniques for masking public use microdata sets, U. S. Bureau of the Census (unpublished manuscript).

Pagliuca, D.; Seri, G., (1998): Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey, Esprit SDC Project, Deliverable MI-3/D2.

Skinner, C.; Marsh, C.; Openshaw, S.; Wymer, C. (1994): Disclosure Control for Census Microdata, Journal of Official Statistics, vol. 10, pp. 31 – 51.

Torres, A. (2003): Contributions to Microaggregation for Statistical Data Protection, Ph.D. Dissertation (in Catalan), Polytechnical University of Catalonia, Barcelona, 2003. Advisors: J. Domingo-Ferrer and J. M. Mateo-Sanz.

U. S. Bureau of the Census (2000): Record Linkage Software: User Documentation. Available from U. S. Bureau of the Census.

Winkler, W. (1998): Re-identification methods for evaluating the confidentiality of analytically valid microdata, in Statistical Data Protection, Luxembourg: Office for Official Publications of the European Communities, 1999. Journal version in Research in Official Statistics, vol. 1(2), pp. 50 – 69.

Sarah Gießing *)

Kommentar zum Beitrag „Concepts for the Evaluation of Anonymized Data"

1 Einleitung

Der Beitrag von Domingo-Ferrer, Mateo und Torres stellt Konzepte und Ergebnisse eines umfangreichen empirischen Vergleichs von Methoden zur Anonymisierung von Einzeldaten vor. Die Studie wurde überwiegend in 2000 durchgeführt. Bis dahin waren zwar auf dem Gebiet der statistischen Geheimhaltung von Einzeldaten sehr viele Methoden und Varianten von Methoden vorgeschlagen und zum Teil auch in der Praxis eingesetzt worden. Umfassende, vergleichende empirische Untersuchungen zur qualitativen Bewertung der Verfahren lagen jedoch nicht vor. Domingo-Ferrer und Kollegen haben mit ihren Arbeiten auf diesem Gebiet Pionierarbeit geleistet. Mit ihren bahnbrechenden und richtungsweisenden Untersuchungen haben sie das Fundament für zukünftige Studien gelegt. Weitere Verfahrensvergleiche, die zum Teil auf dem selben, zum Teil auf leicht modifizierten Ansätzen beruhen, folgten (Yancey et al. 2002; Dandekar/Domingo-Ferrer et al. 2002; Torres 2003).

In diesem Beitrag (Abschnitt 2) soll die Methodik der oben genannten Studien einander gegenübergestellt werden. Abschließend (Abschnitt 3) werden die wichtigsten aus diesem Vergleich gewonnenen Erkenntnisse zusammengefasst und Hinweise für ein sinnvoll erscheinendes, weiteres Vorgehen bei der Entwicklung von Verfahren zur Bewertung von Anonymisierungsmethoden gegeben.

2 Verfahrensvergleich

Im Folgenden soll kurz auf Unterschiede und Gemeinsamkeiten der 2000 von Domingo-Ferrer und Kollegen durchgeführten Studie und den oben genannten, später durchgeführten Vergleichsuntersuchungen hingewiesen werden. Auf entsprechende, im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" (hier kurz: „deutsches Anonymisierungsprojekt") geplante Untersuchungen soll ebenfalls hingewiesen werden.

Untersuchte Anonymisierungsmethoden

Der im Beitrag von Domingo-Ferrer et al. beschriebene Verfahrensvergleich umfasst u.a. Varianten der Mikro-Aggregation, des Rank Swappings sowie einfache Zufallsüberlagerung. In Dandekar/Domingo-Ferrer et al. (2002) werden Varianten der Mikro-Aggregation mit Latin-Hypercube-Sampling (LHS) nach Dandekar et al. (2002) verglichen. Die Studie (Yancey et al. 2002) vergleicht Varianten von Rank-Swapping mit Varianten des in Roque (2000) vorgeschlagenen Zufallsüberlagerungsverfahrens. Das deutsche Anonymisierungsprojekt bezieht in den Vergleich die Methoden Rank-Swapping, Mikro-Aggregation, SAFE (Höhne 2003) und LHS ein.

*) Sarah Gießing, Statistisches Bundesamt, Wiesbaden

Testdaten

Die von Domingo-Ferrer und Kollegen durchgeführten Untersuchungen basieren auf einem Datensatz mit 1 080 Sätzen, der dreizehn stetige Variablen einer haushalts- und personenbezogenen Erhebung umfasst („Domingo-Datensatz"). In der Studie (Dandekar/Domingo-Ferrer et al. 2002) wurden die Verfahren zusätzlich an 1 080 Sätzen des Commercial Building Energy Consumption Survey mit 2 diskreten und 11 stetigen Variablen getestet. Die Untersuchungen von Yancey et al. (2002) basieren neben dem Domingo-Datensatz auf einem umfangreicheren, knapp 60 000 Sätze umfassenden personen- und haushaltsbezogenen Datensatz. Das Anonymisierungsprojekt wird ausschließlich Datensätze der Wirtschaftsstatistik betrachten, u.a. aus der Kostenstrukturerhebung im Verarbeitenden Gewerbe (ca. 17 000 Datensätze) und der Umsatzsteuerstatistik (ca. 2,8 Mill. Sätze).

Messung des Re-identifikationsrisikos

Zur Bewertung des Re-identifikationsrisikos werden im Beitrag von Domingo-Ferrer et al. wahrscheinlichkeitsbasiertes Record-Linkage nach Jaro (1989) in der Implementierung von (Winkler 1998), sowie distanzbasiertes Record-Linkage nach Pagliuca et al. (1998) verwendet. Zusätzlich wird ein Maß für das Risiko, dass Werte näherungsweise offengelegt werden können, verwendet. Hierfür wird der Anteil der Fälle, in denen ein Originalwert in ein „enges" Intervall um den entsprechenden anonymisierten Wert fällt, ermittelt. In ihrem Beitrag schlagen Domingo-Ferrer at al. vor, die Grenzen dieser Intervalle auf Basis von Rängen zu berechnen oder alternativ dazu (Torres 2003) statt der Ränge Standardabweichungen zu verwenden. Letzteres Vorgehen erscheint für Anwendungen im Bereich der Wirtschaftsstatistik mit den hier üblichen, stark schiefen Verteilungen sinnvoller. Als Gesamtmaß für das Re-identifikationsrisiko wird ein gewichteter Mittelwert aus den Teilmaßen ermittelt.

Das Statistische Bundesamt plant im Verlauf des deutschen Anonymisierungsprojekts auf Record Linkage Ansätzen beruhende Methoden zur Messung des Re-identifikationsrisikos zu entwickeln und entsprechende Software zu implementieren. Berücksichtigt werden dabei sowohl das Risiko der exakten, als auch der näherungsweisen Offenlegung von Einzelangaben.

Maße für den Informationsverlust

In ihrem Beitrag schlagen Domingo-Ferrer et al. vor, zur Messung des Informationsverlusts die (relativen) Abweichungen zwischen Original- und anonymisiertem Datensatz zum einen in den Daten selbst (Maß IL1), zum anderen bei den statistischen Kenngrößen Mittelwert (IL2), Varianz (IL3), Kovarianz (IL4) und Korrelation (IL5) zu ermitteln. Der Gesamtinformationsverlust wird als ungewichteter Mittelwert der Teilmaße IL1 bis IL5 berechnet. Torra (2003) verwendet dazu einen gewichteten Mittelwert. Yancey et al. (2002) diskutieren eine alternative Skalierung für die Ermittlung der Abweichung in den Daten selbst (d.h. alternative Berechnung von IL1). Darüber hinaus stellen sie die Einbeziehung von IL1 und IL4 grundsätzlich in Frage und empfehlen Weiterentwicklung des Maßes, um den Informationsverlust auf der Ebene von Teilpopulationen messen zu können.

Score-Verfahren

Im Beitrag von Domingo-Ferrer et al. werden die untersuchten Anonymisierungsverfahren anhand eines Scores, der als ungewichteter Mittelwert der Maße für Informationsverlust und Re-identifikationsrisiko ermittelt wird, bewertet.

Vorstellbar wäre hier natürlich auch die Verwendung eines gewichteten Mittelwerts. Denkbar wäre aber auch ein Verfahren, bei dem einerseits Anonymisierungsmethoden, die nicht gewisse Mindestanforderungen hinsichtlich des Re-identifikationsrisikos erfüllen, grundsätzlich als ungeeignet bewertet werden, bei dem andererseits bei solchen Anonymisierungsmethoden, die den Anforderungen hinsichtlich des Re-identifikationsrisikos genügen, das Maß für das Re-identifikationsrisiko nicht mehr in den Score-Wert mit eingeht. Auf diese Weise würde erreicht, dass unter den im Hinblick auf das Re-identifikationsrisiko zulässigen Verfahren dasjenige mit dem geringsten Informationsverlust am günstigsten erscheint, und nicht dasjenige, mit dem ein „optimales" Gleichgewicht zwischen Informationsverlust und Re-identifikationsrisiko erreicht wird.

3 Ergebnisse und Schlussfolgerungen

Ein Vergleich der Ergebnisse der in diesem Kommentar einander gegenübergestellten Untersuchungen weist darauf hin, dass kleinere Veränderungen in den Gewichten, die für die Ermittlung der Maße für Informationsverlust und Re-identifikationsrisiko verwendet werden, nur zu geringfügigen Verschiebungen innerhalb des Rankings der Anonymisierungsmethoden führt. Die Veränderungen wirken sich vor allem auf die als günstig erscheinenden Parametrisierungen der einzelnen Verfahren aus.

Zu ganz entscheidenden Veränderungen im Ranking kann jedoch die Verwendung eines anderen Test-Datensatzes führen (Yancey 2002). Das spricht dafür, dass ein Ranking entweder für jede zu anonymisierende Statistik individuell durchgeführt werden sollte oder zumindest für bestimmte Klassen von Daten. Kriterien für eine entsprechende Klassenbildung könnten die Zahl der Sätze bzw. Merkmale und die bei den Merkmalen vorliegenden Verteilungen sein.

Literaturhinweise

Dandekar, R.; Cohen, M.; Kirkendall, N. (2002): Sensitive Microdata Protection Using Latin Hypercube Sampling Technique, In: Inference Control in Statistical Databases Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316).

Dandekar, R.; Domingo-Ferrer, J.; Sebe, F. (2002): LHS-Based Hybrid Microdata vs. Rank Swapping and Microaggregation for Numeric microdata Protection, In: Inference Control in Statistical Databases, Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316).

Höhne, J. (2003): SAFE – ein Verfahren zur Geheimhaltung und Anonymisierung statistischer Einzelangaben, in: Berliner Statistik, Statistische Monatsschrift Nr. 3/2003.

Jaro, M. A. (1989): Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida, Journal of the American Statistical Association, vol. 84, pp. 414 – 420.

Pagliuca, D.; Seri, G. (1998): Some Results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey, Esprit SDC Project, Deliverable MI-3/D2.

Roque, G. M. (2000): Masking Microdata Files with Mixtures of Multivariate Normal Distributions, Unveröffentlichte Dissertation, Department of Statistics, University of California-Riverside.

Torres, A. (2003): Contributions to Microaggregation for Statistical Data Protection, Ph.D. Dissertation (in Catalan), Polytechnical University of Catalonia, Barcelona, 2003. Advisors: J. Domingo-Ferrer and J. M. Mateo-Sanz.

Winkler, W. (1998): Re-identification methods for evaluating the confidentiality of analytically valid microdata, in: Statistical Data Protection, Luxembourg: Office for Official Publications of the European Communities, 1999. Journal version in Research in Official Statistics, vol. 1(2), pp. 50 – 69.

Yancey, W E.; Winkler, W.; Creecy, R. H. (2002): Disclosure Risk Assessment in Perturbative Microdata Protection, In: Inference Control in Statistical Databases Domingo-Ferrer (Ed.), Springer (Lecture notes in computer science; Vol. 2316).

Sandra Lechner/Winfried Pohlmeier *)

Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten

Zusammenfassung

Die Anonymisierung von sensiblen Individualdaten führt zu einem Konflikt zwischen dem Ziel der Minimierung des Reidentifikationsrisikos und der Qualität ökonometrischer Schätzungen. Der durch Anonymisierung bedingte Verlust an Effizienz und/oder der Konsistenz eines Schätzers wirft die grundsätzliche Frage auf, inwieweit anonymisierte Individualdaten überhaupt für die wissenschaftliche Nutzung geeignet sind.

Deshalb gehen wir in dieser Arbeit der Frage nach, welchen Einfluss Anonymisierungsverfahren auf die Eigenschaften von ökonometrischen Schätzern haben. Zunächst untersuchen wir die Auswirkungen gängiger Anonymisierungsverfahren auf lineare ökonometrische Schätzer in endlichen Stichproben. Im zweiten Schritt untersuchen wir, inwieweit sich die Selektionseffekte durch Anonymisierung aufgrund von Data Blanking mit Hilfe von semiparametrischen Verfahren korrigieren lassen. Die quantitative Evidenz beruht auf Monte-Carlo-Simulationen und einer illustrativen Anwendung für einen Querschnitt der Kostenstrukturerhebung.

JEL Klassifikation: C81, C21, C24, C25

Schlüsselwörter: Mikroaggregation, stochastische Überlagerung, Data Blanking, IV-Schätzung, semiparametrisches Selektionsmodell

1 Einleitung

In den letzten dreißig Jahren ist die Nachfrage nach Mikrodaten durch die empirische Wirtschaftsforschung stark angestiegen. Diese Nachfrage, die sich ursprünglich auf Haushalts- und Personaldaten bezog, erweiterte sich rasch auf Firmendaten. Individualdaten und hier insbesondere Firmendaten beinhalten oftmals sensible Informationen, deren Vertraulichkeit im Interesse der Beobachtungseinheit, aber auch im Interesse der datenerhebenden Institution und der Datennutzer es zu schützen gilt. Damit stehen die datenerhebenden Institutionen vor einem Konflikt zwischen dem Ziel der Gewährleistung einer maximalen Vertraulichkeit der Daten und dem Ziel der Weitergabe maximaler Information.

Um die Möglichkeit der Reidentifikation individueller Angaben aus Mikrodaten zu minimieren und die von der datenerhebenden Institution gemachte Vertraulichkeitszusage zu garantieren, werden in der Praxis unterschiedliche Anonymisierungverfahren verwendet[1)], die sich im Ausmaß der Anonymisierung und ihres Effektes auf die Effizienz und die Konsistenz des verwendeten ökonometrischen Schätzverfahrens unterscheiden. Im Allgemeinen stellt ein Anonymisierungsverfahren nichts anderes als einen Datenfilter

*) Sandra Lechner, Universität Konstanz und Prof. Winfried Pohlmeier, Universität Konstanz, Center of Finance and Econometrics, Zentrum für Europäische Wirtschaftsforschung.

1) Siehe z.B. Gottschalk (2002) für eine Übersicht.

dar, der den wahren datengenerierenden Prozess verändert. Für den empirischen Wirtschaftsforscher ergibt sich hieraus die Frage, inwieweit sich der wahre datengenerierende Prozess auf der Grundlage der anonymisierten (gefilterten) Daten schätzen lässt. Letztlich stellt die Anonymisierung von Individualdaten den Nutzer vor die grundsätzliche Frage, wie erheblich der durch die Anonymisierung bedingte Verlust an Information ist und unter welchen Umständen überhaupt konsistente Schätzungen des wahren datengenerierenden Prozesses möglich sind. Selbst wenn ein gegebenes Anonymisierungsverfahren nicht zum Verlust der Konsistenzeigenschaft des Schätzverfahrens führt, stellt sich die Frage der Relevanz der erzielten empirischen Ergebnisse, denn statistisch als insignifikant gefundene Zusammenhänge mögen schlichtweg das Resultat des Informationsverlustes durch Anonymisierung sein. Ist der durch Anonymisierung bedingte Informationsverlust erheblich, wird der anonymisierte Datensatz für den mit statistischen Inferenzmethoden arbeitenden wissenschaftlichen Nutzer unbrauchbar.

In dieser Arbeit untersuchen wir deshalb den Einfluss von Anonymisierungsverfahren auf die Eigenschaften von ökonometrischen Schätzern. Im Mittelpunkt unseres Interesses stehen insbesondere Auswirkungen von Anonymisierungsverfahren auf die Eigenschaften ökonometrischer Schätzer in endlichen Stichproben. Anhand von Monte-Carlo Simulationen sollen dabei die Auswirkung der Anonymisierungsverfahren auf die ökonometrische Schätzung quantifiziert werden.

Die Arbeit ist wie folgt aufgebaut: In Abschnitt 2 arbeiten wir die Konsequenzen der faktischen Anonymisierung durch Mikroaggregation und der stochastische Überlagerungen für die Schätzung des linearen Regressionsmodells heraus.[2] Wir zeigen anhand einer Monte-Carlo-Studie, dass selbst im einfachen Fall des linearen Modells Anonymisierung nicht unproblematisch ist und den Nutzen für den Anwender erheblich einschränken kann. In Abschnitt 3 stellen wir einen zweistufigen semiparametrischen Selektionsschätzer vor, der der Selektionsverzerrung durch Data Blanking oder partielle Aggregation Rechnung trägt. Dieses Verfahren beruht auf dem semiparametrischen Schätzer von Klein und Spady (1993) für binäre Auswahlmodelle in der ersten Stufe und dem semiparametrischen Reihenapproximationsschätzer von Newey (1999). Abschnitt 4 illustriert am Beispiel einer Regression, basierend auf den Daten der Kostenstrukturerhebung des Statistischen Bundesamtes, inwieweit sich Schätzergebnisse aufgrund der verwendeten Anonymisierungsmethode im Vergleich zu einer Analyse auf Basis der Originaldaten ändern. Abschnitt 5 gibt einen Ausblick auf die zukünftige Forschung.

2) Weitere Verfahren jenseits der faktischen Anonymisierung (z.B. resampling, data swapping) werden in Brand (2000) vorgestellt.

2 Klassische Anonymisierungsverfahren: Einige Konsequenzen

Um den Effekt der Anonymisierung auf die Eigenschaften des KQ-Schätzers zu bewerten, gehen wir vom linearen Regressionsmodell unter vollen idealen Bedingungen aus:

$$Y = X\beta + \varepsilon\,, \tag{2.1}$$

mit

$$E[\varepsilon] = 0,$$
$$V[\varepsilon] = \sigma^2 I_N,$$
$$\operatorname{plim} \frac{1}{N} X'X = \Sigma_{XX},$$

wobei X eine $N \times K$ – Regressionsmatrix fester erklärender Variablen und Y der $N \times 1$-Vektor der abhängigen Variablen ist. Das idealtypische Design für die Originaldaten erlaubt uns, die Auswirkungen des Anonymisierungsverfahrens auf die stochastischen Eigenschaften verschiedener Schätzer gegenüber dem Idealfall des besten linearen unverzerrten Schätzers zu vergleichen.

Mikroaggregation im linearen Modell

Bei der Mikroaggregation werden die Variablenausprägungen durch vorher ermittelte Mittelwerte von jeweils ähnlichen Datensätzen ersetzt (Paaß and Wauschkuhn 1984). Hier sei nur der Fall der listenweisen Aggregation betrachtet, bei der jeweils die Variablen von A Beobachtungen zu entsprechenden Gruppenmittelwerten zusammengefasst werden. In diesem Fall ergeben sich $M = N/A$ unterschiedliche (aggregierte) Beobachtungen. Zur Vereinfachung der Notation sei angenommen, dass M ganzzahlig ist. Üblicherweise werden in der Praxis $A = 3, 4$ oder 5 Beobachtungen zu einer aggregierten Beobachtung zusammengefasst.

Da von einer Zufallsstichprobe unabhängig und identisch verteilter Beobachtungen ausgegangen wird, kann zur Vereinfachung der Notation ohne Verlust der Allgemeingültigkeit angenommen werden, dass die Mikroaggregation gemäß der Reihenfolge der Beobachtungen im Datensatz erfolgt. Hierzu sei die $N \times N$ – blockdiagonale Matrix

$$D = I_M \otimes \frac{1}{A} \iota\iota' \tag{2.2}$$

definiert, wobei ι ein A-dimensionaler Vektor von Einsen ist. Das lineare Regressionsmodell auf der Grundlage der mikroaggregierten Daten ergibt sich durch Prämultiplikation von D mit dem auf den Originaldaten basierenden Modell (2.1):

$$Y^* = X^*\beta + \varepsilon^* \tag{2.3}$$

mit $Y^* = DX$, $X^* = DX$ und $\varepsilon^* = D\varepsilon$.

Der gewöhnliche KQ-Schätzer für das mikroaggregierte Modell $\hat{\beta}_A$ hat die Form

$$\hat{\beta}_A = \left(X^{*\prime}X^*\right)^{-1} X^{*\prime}Y^* = \left(XDX\right)^{-1} X'DY, \tag{2.4}$$

wobei zur Berechnung des rechten Terms in (2.4) die Symmetrie und Idempotenz von D verwendet wurden. Offensichtlich bleibt durch die listenweise Mikroaggregation die Unverzerrtheit des KQ-Schätzers erhalten. Mit dem exogenen Datenfilter D verwenden wir hier das denkbar einfachste Aggregationsschema mit gleicher Gewichtung über alle Beobachtungen und gleichem Aggregationsniveau für alle Gruppen. Ein Aggregationsschema, basierend auf den exogenen Variablen $D = D(X)$, stellt nur eine unwesentliche Erweiterung dar. Wenn jedoch das Gewichtungsschema der Aggregation von der abhängigen Variablen abhängt, $D = D(Y, X)$, ist der KQ-Schätzer für die aggregierten Daten nichtlinear mit unbekannten Verteilungseigenschaften.

Schon die einfache exogene Aggregation führt jedoch zu einem Informationsverlust, so dass der KQ-Schätzer auf Grundlage der anonymisierten Daten gegenüber dem gewöhnlichen KQ-Schätzer $\hat{\beta}$ an Effizienz verliert (Beweis siehe Anhang):

$$V\left[\hat{\beta}_A\right] - V\left[\hat{\beta}\right] > 0,$$

wobei das Ungleichheitszeichen für die positive Definitheit der Differenz der beiden Varianz-Kovarianzmatrizen steht. Der durch Aggregation bedingte Effizienzverlust kann für den Fall $k = 1$ leicht verdeutlicht werden. Es sei

$$X = (X_1, X_2, \ldots, X_N)' = (X_{11}, X_{21}, \ldots, X_{A1}, X_{12}, X_{22}, \ldots, X_{A2}, \ldots, X_{AM})'$$

der Vektor der erklärenden Variablen, wobei die Doppelindizierung a,m die Beobachtung a in Gruppe m bezeichnet. Der Vergleich der Präzisionen beider Schätzer

$$V\left[\hat{\beta}\right]^{-1} - V\left[\hat{\beta}_A\right]^{-1} = \frac{1}{\sigma^2} \sum_{a=1}^{A} \sum_{m=1}^{M} \left(X_{am} - \bar{X}_m\right)^2$$

zeigt, dass der Effizienzverlust durch Aggregation besonders klein ist, wenn die Aggregation über möglichst homogene Gruppenmitglieder erfolgt, d.h. wenn die Variation innerhalb der Gruppen (within group variation) gegen null geht.

Die Mikroaggregation führt zu einer Verzerrung des herkömmlichen Schätzers für die Varianz des Fehlerterms und somit der Standardfehler des KQ-Schätzers von β. Sofern das Aggregationsniveau bekannt ist, ergibt sich ein unverzerrter Schätzer für σ^2 wie folgt:

$$\sigma^2 = \frac{1}{M - K} e^{*\prime} e^{*},$$

wobei $e^{*\prime} e^{*}$ die Summe der quadrierten Fehler des KQ-Schätzers auf der Grundlage der aggregierten Beobachtungen ist (Beweis siehe Anhang). Da $M < N$, führt eine Ignorierung der wahren Freiheitsgrade des Modells zu einer Unterschätzung der Standardfehler, so dass die auf der aggregierten Datenbasis erzielten t-Werte des KQ-Schätzers überhöht ausgewiesen werden.

Bootstrap-Aggregation

Bei der einfachen Mikoaggregation erscheint jede anonymisierte Beobachtungseinheit A – mal im Datensatz. Alternativ kann jedoch auch für jede Beobachtung i des Originaldatensatzes per Zufallsziehung mit Zurücklegen eine (möglichst homogene) Gruppe i zusammengestellt werden und die Mittelwerte der Kovariate dieser Gruppe i als anonymisierte Beobachtungseinheit verwendet werden. Die Idee für diese Art von Mikroaggregation hat gewisse Ähnlichkeiten mit dem Bootstrap-Verfahren, da durch Ziehung aus der Stichprobe künstlich neue Datensätze gezogen werden, über die dann aggregiert wird. Die Struktur des gewöhnlichen KQ-Schätzers auf Grundlage einer Bootstrap-Aggregation ist äquivalent zum Schätzer $\hat{\beta}_A$. Jedoch ist die $N \times N$ – Aggregationsmatrix D bei der Bootstrap-Aggregation eine Zufallsmatrix der Form

$$D = \frac{1}{B}(I_n + S_1 + S_2 + \ldots + S_{B-1}), \tag{2.5}$$

S_b stellt hierbei eine $N \times N$ Selektionsmatrix dar, die jeweils in einer Zeile an einer zufällig ausgewählten Position eine Eins und sonst Nullen enthält. Prämultiplikation des Originalmodells (2.1) mit D liefert das lineare Modell auf der Grundlage von N verschiedenen Gruppenmittelwerten. Da D nun eine Zufallsmatrix ist, haben wir es trotz fester X – Variablen mit einem Modell mit stochastischen Regressoren zu tun.

$$V\left[\hat{\beta}_B\right] = E\,V\left[\hat{\beta}_B | D\right] = \sigma^2 E\left[(X'DX)^{-1}\right] \tag{2.6}$$

Gegenüber der einfachen Mikroaggregation wird das Reidentifikationssrisiko durch die Bootstrap-Aggregation weiter verringert, da aus der zufälligen Aggregation die Wahrscheinlichkeit, eine korrekte Schlussfolgerung über die Originaldaten zu ziehen, weiter reduziert wird. Die Wahrscheinlichkeit, dass eine aggregierte Beobachtung i genau der Originalbeobachtung entspricht, beträgt N^{-B}.

Für die Standardfehler sollte bei einer Bootstrap-Aggregation der heteroskedastierobuste Varianz-Kovarianzmatrix-Schätzer verwendet werden, denn Regressoren bei einer Bootstrap-Aggregation können als gewogenes Mittel aus Originalbeobachtung und arithmetischem Mittel über alle Beobachtungen mit Gewichtungsfaktor 1/B und 1-1/B formuliert werden, wobei die Stichprobenvariation über einen heteroskedastischen Fehlerterm aufgefangen wird. Ersetzt man nämlich den stochastischen Aggregationsfilter D durch seinen Erwartungswert und einer zufälligen Abweichung ζ mit Erwartung $E[\zeta] = 0$,

$$D = E[D] + \zeta,$$

ergibt sich aus dem bootstrap-aggregierten Regressionsmodell

$$Y^* = E[D]X\beta + \omega,$$

wobei $\omega = \zeta X \beta + \varepsilon$ ein heteroskedastischer Fehlerterm ist. Die Regressormatrix $E[D]X$ ist das gewogene Mittel aus Originalbeobachtung und arithmetischem Mittel über alle Beobachtungen.

Stochastische Überlagerung

Als Alternative zur Mikroaggregation wird oftmals die stochastische Überlagerung verwendet. Dieses Verfahren ist besonders bei Paneldatensätzen attraktiv, wenn die stochastische Überlagerung multiplikativ und zeitlich konstant ist. Das loglineare Modell ist in diesem Fall nur durch einen stochastischen Individualeffekt vom loglinearen Modell auf Basis der Originaldaten verschieden. Differenzbildung oder Within-Transformation des loglinearen Modells beseitigen den Einfluss der multiplikativen stochastischen Überlagerung. Allerdings setzt dieses Verfahren voraus, dass der wahre datengenerierende Prozess tatsächlich loglinear ist. Die Überprüfung der funktionalen Form der Erwartungswertfunktion mit Hilfe von Tests auf funktionale Form ist nicht mehr trivial, weil entsprechende Annahmen über Art der stochastischen Überlagerung als beizubehaltende Hypothese berücksichtigt werden müssen (z.B. Nullhypothese: wahres Modell ist linear, beizubehaltende Hypothese: stochastische Überlagerung ist multiplikativ).

Im Folgenden sei von einer additiven stochastischen Überlagerung oder einem loglinearen Modell mit multiplikativer stochastischer Überlagerung ausgegangen. Zur abhängigen Variablen und zum Vektor der erklärenden Variablen werden unabhängig identisch verteilte Störgrößen hinzu addiert

$$\begin{aligned} Y_i^* &= Y_i + v_i, \\ X_i^* &= X_i + u_i, \end{aligned} \tag{2.7}$$

so dass das verfügbare Modell auf der Grundlage der stochastisch überlagerten Beobachtungen die Form

$$Y^* = X^*\beta + \omega \tag{2.8}$$

mit $\omega = \varepsilon + v - u\,\beta$ annimmt. Stochastische Überlagerung führt zu einem klassischen Fehler-in-den-Variablen-Modell. Aufgrund der stochastischen Überlagerung sind Fehlerterm und Regressoren miteinander korreliert, so dass der gewöhnliche KQ-Schätzer inkonsistent ist:

$$\operatorname{plim}\hat{\beta}_{EIV} = \left(I_K - \left(Q + \Sigma_{uu}\right)^{-1}\Sigma_{uu}\right)\beta \neq \beta\,, \tag{2.9}$$

wobei $\Sigma_{uu} = E\left[u_i u_i'\right]$ die Varianz-Kovarianz-Matrix des Fehlertermvektors u_i bezeichnet. Definieren wir $\kappa_{XX} = \left(\operatorname{plim}\frac{1}{N}X^{*\prime}X^{*}\right)^{-1}\operatorname{plim}\frac{1}{N}X'X = \left(Q + \Sigma_{uu}\right)^{-1}Q$ als die Zuverlässigkeitsmatrix im Sinne einer multivariaten Erweiterung des Zuverlässigkeitskoeffizienten (reliability ratio) von Fuller (1987, S. 3), erhalten wir

$$\operatorname{plim}\hat{\beta}_{EIV} = \kappa_{XX}\beta\,. \tag{2.10}$$

Anders als beim Fehler-in-den-Variablen-Modell ist jedoch hier der datengenerierende Prozess bekannt, so dass die asymptotische Verzerrung des KQ-Schätzers $\hat{\beta}_{EIV}$ leicht korrigiert werden kann, sofern Q und Σ_{uu} bzw. κ_{XX} bekannt sind. Der korrigierte unverzerrte KQ-Schätzer $\hat{\beta}_{CEIV}$ weist die Form

$$\hat{\beta}_{CEIV} = \left(I_K - \left(Q + \Sigma_{uu}\right)^{-1}\Sigma_{uu}\right)^{-1}\hat{\beta}_{EIV} = \kappa_{XX}^{-1}\hat{\beta}_{EIV} \qquad (2.11)$$

auf. In der Praxis könnte dieser korrigierte Fehler-in-der-Variablen-Schätzer ohne großen Aufwand für die datenerhebende Institution und ohne Erhöhung des Reidentifikationsrisikos implementiert werden. Als einzige zusätzliche Information müsste dem Datennutzer die Kovarianzmatrix Σ_{uu} bereitgestellt werden. Da das Reidentifikationsrisiko nicht unbedingt mit der Annahme unkorrelierter Anonymisierungsstörgrößen steigt, kann Unkorreliertheit vorausgesetzt werden, so dass die Information über die Varianzen der Störgrößen ausreicht. Ein konsistenter Schätzer des Terms $Q + \Sigma_{uu}$ ist die empirische Momentenmatrix der anonymisierten Regressoren

$$\operatorname{plim}\frac{1}{N}X^{*'}X^{*} = Q + \Sigma_{uu},$$

so dass ein verfügbarer korrigierter Fehler-in-dem-Variablen-Schätzer $\tilde{\beta}_{CEIV}$ die Form

$$\tilde{\beta}_{CEIV} = \left(I_k - \left(\frac{1}{N}X^{*'}X^{*}\right)^{-1}\Sigma_{uu}\right)^{-1}\hat{\beta}_{EIV} \qquad (2.12)$$

aufweist.

Als Alternative zum korrigierten Fehler-in-dem-Variablen-Schätzer wäre auch die Bereitstellung von Instrumentvariablen für die anonymisierten Variablen denkbar, in dem die wahren Variablen mit anderen Anonymisierungsstörgrößen stochastisch überlagert werden. Dieser zweite Satz von stochastisch überlagerten Variablen weist alle Eigenschaften von validen Instrumenten eines Instrumentvariablenschätzers auf. Da somit sowohl die Unkorreliertheit zwischen Instrumenten und Fehlerterm als auch die Korrelation zwischen Instrumenten und anonymisierten Regressoren garantiert ist, werden die notwendigen Verteilungsannahmen der IV-Schätzers per Datenkonstruktion erfüllt. Durch Bereitstellung von Instrumentvariablen steigt allerdings das Reidentifikationsrisiko, da anonymisierte Regressoren und Instrumente gemeinsam mehr Information über die wahren Merkmalsausprägungen liefern.

Monte-Carlo-Evidenz

Mit Hilfe einer einfachen Monte-Carlo-Studie soll im Folgenden die quantitative Auswirkung von Mikroaggregation und stochastischer Überlagerung illustriert werden. Hierzu soll das lineare Modell

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

geschätzt werden. Der Fehlerterm ε wird als unabhängig, identisch $t-$verteilt mit 4 Freiheitsgraden unterstellt, so dass $V[\varepsilon] = 2$. Die beiden erklärenden Variablen werden aus einer bivariaten Normalverteilung der Form

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}\begin{pmatrix} 1 & .4 \\ .4 & 1 \end{pmatrix}\right)$$

gezogen. Die Verwendung zweier, relativ stark korrelierter Regressoren ist vor allem im Kontext des in Abschnitt 3 untersuchten semiparametrischen Selektionsschätzers von Interesse, da Multikollinearität zwischen Regressoren und Kontrollfunktion in kleinen Stichproben von Bedeutung sein kann. Für alle Simulationen werden als wahre Parameterwerte für die Regressionskoeffizienten die Werte $\beta_0 = .5$, $\beta_1 = 1$, $\beta_2 = -1$ verwendet.

Basierend auf einem Monte-Carlo Design für stochastische Regressoren werden Schätzungen auf Datensätzen mit Beobachtungsumfang $N = 120$, 1200 und 3600 durchgeführt.[3] Die Auswertung der Monte-Carlo Schätzungen beruht auf $R = 1000$ Replikationen. Mit dem idealtypischen KQ-Schätzer auf Grundlage der Originaldaten $\hat{\beta}$ wird der KQ-Schätzer $\hat{\beta}_A$ unter Verwendung mikroaggregierter Daten des Aggregationsniveaus $A = 3$, 4 und 5 verglichen. Des Weiteren wird der bootstrap-aggregierte Schätzer $\hat{\beta}_B$ untersucht, wobei wir über 3 Beobachtungen $(B = 3)$ aggregieren.

Für die mit stochastischer Überlagerung anonymisierten Daten verwenden wir als Anonymisierungsstörgröße für Y, X_1 und X_2 jeweils unabhängig und identisch normalverteilte Zufallsvariablen mit Erwartungswert 0 und Varianz 0.25. Untersucht werden im Kontext der stochastischen Überlagerung der inkonsistente KQ-Schätzer auf Grundlage der anonymisierten Daten $\hat{\beta}_{ETV}$, der Instrumentvariablenschätzer $\hat{\beta}_{IV}$ mit Instrumenten, die analog zu den anonymisierten Regressoren erzeugt werden, sowie der korrigierte Fehler-in-den-Variablen Schätzer $\hat{\beta}_C$. Zur Berechnung des Korrekturterms verwenden wir die bekannte Varianz-Kovarianz-Matrix der Anonymisierungsstörgrößen Σ_{uu} sowie die empirische Momentenmatrix von X als Schätzung für Q.

Tabelle 1 fasst die Monte-Carlo-Ergebnisse für die Aggregationsschätzer zusammen. Aus Platzgründen werden nur die Resultate für den Koeffizienten vor der X_1 – Variablen wieder gegeben, zumal die Ergebnisse für β_0 und β_2 sich nicht substantiell von den Ergebnissen für β_1 unterscheiden. Neben der mittleren Schätzung und der mittleren Verzerrung berechnen wir die Wurzel des mittleren quadratischen Fehler, RMSE, als Maß für die Schätzunsicherheit in endlichen Stichproben. Der relative Standardfehler, RELSE, ist definiert als das Verhältnis von mittlerem Standardfehler der R Schätzungen zur Standardabweichung der Schätzungen. Da für $R \to \infty$ die Standardabweichungen der Schätzungen gegen den wahren Standardfehler der Schätzung für endliches N konvergieren, geben Abweichungen des relativen Standardfehlers von 1 Auskunft über die Genauigkeit der Schätzung der Standardabweichung des Schätzers aufgrund der asymptotischen Verteilung. Dieses Maß ist vor allem für Schätzer von Interesse, deren Standardfehler nicht für endliche Stichproben berechnet werden können, und bei denen deshalb asymptotische Approximationen verwendet werden müssen. Der relative Standardfehler wird in zwei Varianten ausgewiesen. Der unkorrigierte relative Standardfehler gibt Auskunft über das Ausmaß der fehlerhaften Inferenz, wenn das Aggregationsniveau bei der Inferenz durch eine entsprechende Korrektur der Freiheitsgrade unberücksichtigt bleibt.

3) Die etwas ungewöhnlichen Werte für den Beobachtungsumfang wurden gewählt, damit N ein Vielfaches des Aggregationsniveaus $A = 3$, 4 und 5 ist.

Der korrigierte relative Standardfehler verwendet im Zähler die korrekten Standardfehler basierend auf M statt auf N Beobachtungen.

Tabelle 1: Monte-Carlo Ergebnisse: Mikroaggregation im linearen Modell *)

$\beta = 1$	Mittelwert	Verzerrung	RMSE	RELSE korrigiert	RELSE unkorrigiert
		$N = 120$			
$\hat{\beta}$	.995	-.005	.145	–	.993
$\hat{\beta}_{A=3}$	.987	-.013	.277	.912	.513
$\hat{\beta}_{A=4}$	.991	-.009	.308	.957	.460
$\hat{\beta}_{A=5}$	.989	-.011	.365	.927	.393
$\hat{\beta}_{B=3}$	.992	-.008	.200	1.277	.718
		$N = 1200$			
$\hat{\beta}$	1.002	.002	.045	–	1.028
$\hat{\beta}_{A=3}$	1.001	.001	.077	1.000	.576
$\hat{\beta}_{A=4}$	.999	-.001	.089	.989	.493
$\hat{\beta}_{A=5}$	1.000	.000	.095	1.040	.463
$\hat{\beta}_{B=3}$	1.001	.001	.063	1.270	.732
		$N = 3600$			
$\hat{\beta}$	1.000	.000	.032	–	1.016
$\hat{\beta}_{A=3}$	1.000	.000	.045	.978	.564
$\hat{\beta}_{A=4}$	1.003	.003	.056	.978	.493
$\hat{\beta}_{A=5}$	1.001	.001	.056	1.021	.456
$\beta_{B=3}$	.999	-.001	.032	1.263	.729

*) Schätzung des Koeffizienten vor der ersten erklärenden Variablen, Anzahl der Replikationen = 1 000.

Wie bereits theoretisch gezeigt, führt die einfache listenweise Mikroaggregation zu keinerlei Verzerrung. Selbst für den kleinen Stichprobenumfang von $N = 120$ liegen die Schätzungen im Mittel für alle betrachteten Aggregationsniveaus recht nahe beim wahren Parameterwert, was offensichtlich eine Folge der gewählten Größe von $V[\varepsilon]$ ist. Allerdings führt die Aggregation zu Effizienzverlusten. Die Streuung der Schätzungen gemessen in termini des mittleren quadratischen Fehlers (RMSE) steigt erheblich mit dem Aggregationsniveau an. Nur für die größte Stichprobe mit 3 600 Beobachtungen sind Unterschiede im mittleren quadratischen Fehler nicht mehr auszumachen. Recht erfolgreich schneidet die Boostrap-Aggregation ab. Bei gleichem Aggregationsniveau ist der mittlere quadratische Fehler für $\hat{\beta}_{B=3}$ deutlich geringer als der mittlere quadratische Fehler von $\hat{\beta}_{A=3}$.

Die unkorrigierten relativen Standardfehler der Aggregationsschätzer liegen deutlich unter 1. Wie zu erwarten war, steigt die Verzerrung der Standardfehler mit dem Aggregationsniveau. Die Ignorierung der wahren Freiheitsgrade des Modells führt zu einer fehlerhaften Inferenz, die dem empirischen Wirtschaftsforscher niedrigere p – Werte (höhere t – Statistiken) vorgaukelt als tatsächlich vorhanden sind. Die Korrektur um die wahre Anzahl von Freiheitsgraden führt dagegen zu Standardabweichungen, die den empirischen Standardabweichungen recht nahe kommen. Eine Ausnahme bildet der Schätzer auf Grundlage der bootstrap-aggregierten Daten. Die unkorrigierten relativen Standardfehler liegen deutlich unter 1, während die korrigierten relativen Standardfehler auf eine Überkorrektur hinweisen. Aufgrund der heteroskedastischen Struktur des bootstrapagregierten Modells ist deshalb zu überprüfen, ob eine heteroskedastie-robuster Schätzer der Standardfehler genauere Schätzungen liefert.

Tabelle 2: Monte-Carlo Ergebnisse: Stochastische Überlagerung im linearen Modell*)

$\beta = 1$	Mittelwert	Verzerrung	RMSE	RELSE
		$N = 120$		
$\hat{\beta}$	.995	-.005	.145	.993
$\hat{\beta}_{EIV}$	.707	-.293	.324	1.018
$\hat{\beta}_{IV}$	1.001	.001	.195	.999
$\hat{\beta}_{CEIV}$	1.008	.008	.184	1.040
		$N = 1200$		
$\hat{\beta}$	1.002	.002	.045	1.028
$\hat{\beta}_{EIV}$	.707	-.293	.297	1.016
$\hat{\beta}_{IV}$	.998	-.002	.063	.985
$\hat{\beta}_{CEIV}$	1.002	-.002	.055	1.028
		$N = 3600$		
$\hat{\beta}$	1.000	.000	.032	1.016
$\hat{\beta}_{EIV}$	.706	-.294	.295	1.020
$\hat{\beta}_{IV}$	.999	-.001	.032	1.018
$\hat{\beta}_{CEIV}$	1.000	.000	.032	1.040

*) Schätzung des Koeffizienten vor der ersten erklärenden Variablen, Anzahl der Replikationen = 1 000.

Tabelle 2 enthält die Ergebnisse der Monte-Carlo-Studie für die Anonymisierung durch stochastische Überlagerung. Die Verzerrung des gewöhnlichen KQ-Schätzers erweist sich als erheblich und wird aufgrund der Inkonsistenz dieses Schätzers auch nicht mit zunehmendem Stichprobenumfang reduziert. Von einer naiven Verwendung von Datensätzen, die durch stochastische Überlagerung anonymisiert werden, ist deshalb abzuraten. Der Instrumentvariablenschätzer erweist sich als recht leistungsstark, obwohl hier sogar auch der Fehlerterm der abhängigen Variablen aufgrund der Anonymisierungsstörgröße von Y eine größere Varianz aufweist als im Originalmodell. Anders als bei der konventionellen IV-Schätzung auf Grundlage nicht experimenteller Instrumente sind hier die Instrumente per Konstruktion des Anonymisierungsverfahrens stark mit den anony-

misierten erklärenden Variablen korreliert. Der mittlere quadratische Fehler, der von der Korrelation zwischen (anonymisierten) Regressoren und den Instrumenten abhängt, ist für die gewählte Parameterkonstellation auch im Vergleich zum KQ-Schätzer auf Grundlage der Originaldaten recht klein. Ähnlich erfolgreich ist der korrigierte Fehler-in-den-Variablen Schätzer. Hier hängt die Präzision des Schätzers von der Schätzgenauigkeit von Q ab. Mit steigendem Beobachtungsumfang konvergiert die Schätzung für Q gegen den wahren Wert. Schon bei einem Stichprobenumfang von $N = 1200$ lassen sich keine wesentlichen Unterschiede zwischen $\hat{\beta}$ und $\hat{\beta}_{CEIV}$ ausmachen. Aus der Sicht des Ökonometrikers, für den bei gegebener faktischer Anonymisierung die Qualität der Schätzung im Vordergrund steht, stellt die stochastische Überlagerung im Kontext des linearen Regressionsmodells eine echte Alternative zur Mikroaggregation dar.

3 Anonymisierung und nichtlineare ökonometrische Modelle

Da sowohl Mikroaggregation als auch die stochastische Überlagerung, wie sie im vorherigen Abschnitt eingeführt wurden, lineare Transformationen der Originaldaten darstellen, sind die Auswirkungen dieser Anonymisierungsmethoden auf Schätzungen linearer Regressionsmodelle sehr viel einfacher zu analysieren als im Falle nichtlinearer Modelle. So führt die stochastische Überlagerung bei nichtlinearen Modellen zu einem nichtlinearen Fehler-in-den-Variablen-Modell. Der Umfang der Literatur zu Messfehlern in nichtlinearen Modellen muss als vergleichsweise gering bezeichnet werden. Spezielle Aspekte werden in den Arbeiten von Amemiya (1985), Hausman, Newey und Powell (1995), Lee und Sepanski (1995) sowie Hong und Tamer (2002) behandelt. Eine Übersicht über neuere Verfahren bietet die Monographie von Carroll, Ruppert und Stefanski (1995). Die Mikroaggegation nichtlinearer Modelle scheint nur unter Inkaufnahme von Approximationsfehlern ein gangbarer Weg zu sein (Lechner u. Pohlmeier 2003). Als Alternative bieten sich teilweise nichtlineare Modelle für gruppierte Daten an, die jedoch nicht unbedingt Rückschlüsse auf den datengenerierenden Prozess der Mikroebene zulassen.

Im Folgenden schlagen wir deshalb ein alternatives Anonymisierungsverfahren vor, das auch auf den Fall nichtlinearer Regressionsmodelle erweiterbar ist. Die Idee beruht darauf, dass Beobachtungen, die ein hohes Risiko der Reidentifikation aufweisen, zensiert werden, bzw. aus dem Datensatz gelöscht werden (Blanking). Geschätzt werden soll das nichtlineare Regressionsmodell mit additivem Fehlerterm

$$Y_i = f(X_i, \beta) + \varepsilon_i \tag{3.1}$$

Wir unterstellen, dass das Reidentifikationsrisiko nur bei den Beobachtungseinheiten des Datensatzes groß ist, die extreme Werte für irgendeine der Variablen aufweisen. Es sei W_i der Vektor von insgesamt L Variablen für Beobachtung i. Dieser Vektor enthält die erklärenden Variablen, die zu erklärende Variable Y_i sowie andere sicherheitsrelevante Variablen des Datensatzes, die nicht zwingend Regressoren in (3.1) sein müssen. Eine Beobachtung wird nicht anonymisiert übernommen, wenn alle Variablen von W_i innerhalb der Quantile θ_l und θ_u liegen. Der binäre Indikator für die nichtanonymisierte Übernahme der Variablen von i in den Datensatz ist demnach definiert als

$$S_i = \begin{cases} 1 \text{ wenn } & q_{\theta_l}(W_{1j},\ldots,W_{nj}) < W_{ij} < q_{\theta_u}(W_{1j},\ldots,W_{nj}), \quad \forall\ j = 1,\ldots,L \\ 0 \text{ sonst}, & \end{cases} \tag{3.2}$$

wobei $q_\theta(\cdot)$ das θ – Quantil der Variablen W_j bezeichnet mit $\theta_l < \theta_u$. In der Regel sollte das Reidentifikationsrisiko für besonders große Werte von W_{ij} hoch sein, so dass eine Selektion über das untere Quantil zu vernachlässigen ist. Alternativ können auch andere Anonymisierungsregeln unterstellt werden, die beispielsweise von einer hohen Reidentifikationswahrscheinlichkeit aufgrund von Kombinationen der Variablen ausgehen. Es sei nun unterstellt, dass die gewählte Selektionsregel durch eine semiparametrische „Single-Index"-Form approximiert werden kann:

$$S_i = 1\left|\left(\varphi\left(Z_i'\gamma\right) > \tau\right)\right. \tag{3.3}$$

Hierbei bezeichnet $\varphi(\cdot)$ eine zweifach differenzierbare bekannte Funktion bezüglich der Indexfunktion $I_i = Z_i'\gamma$ und τ einen unbekannten zusätzlichen Schwellenparameter. Gleichung (3.3) stellt eine Verallgemeinerung der üblichen linearen Selektionsregel $S_i = 1\left|\left(Z_i'\gamma + u_i > 0\right)\right.$ dar. Kontrollvariablen der Selektionsgleichung (3.3) sind die erklärenden Variablen der Strukturgleichung, da sie via Strukturgleichung die Größe der abhängigen Variablen bestimmen, sowie andere Varablen, die aus dem Datensatz für die Beobachtung i zu löschen sind und S_i über Kreuzkorrelationen beeinflussen. Da die erklärenden Variablen in der Selektionsgleichung ebenfalls der Anonymisierung unterliegen können, müssen sie anonymisierter in die Selektionsgleichung eingehen. Für eine zu anonymisierende Variable Z_{ij} der Selektionsgleichung schlagen wir folgende Transformation vor, wenn nur eine Anonymisierung großer Werte erfolgen soll:

$$Z_{ij}^* = 1\left|\left(Z_{ij} < q_{\theta u}\left(Z_j\right)\right)\right. Z_{ij} + \left[1 - 1\left|\left(Z_{ij} < q_{\theta u}\left(Z_j\right)\right)\right.\right] E\left[Z_{ij}\left|Z_{ij} \geq q_{\theta u}\left(Z_j\right)\right.\right] \tag{3.4}$$

Bei dieser Transformation bleiben die nicht zu anonymisierenden Werte in Originalform erhalten, während die zu anonymisierenden Beobachtungen durch den konditionalen Erwartungswert besetzt werden, der durch das bedingte arithmetische Mitel aus den Originaldaten geschätzt werden kann.

Letztlich kann die Selektionsgleichung auch Variablen enthalten, die über exogene, nicht auf der Quantilsregel (3.2) beruhende Selektionskriterien beruhen. Beispielsweise ist es üblich, Informationen über Branchen, die weniger als eine vorgegebene Anzahl von Unternehmen aufweisen, zu löschen. Die Anzahl der Firmen in einer Branche könnte in diesem Fall ein derartiger Regressor sein.

Im Falle einer linearen Strukturgleichung, $f(X_i,\beta) = X_i'\beta$, können die von einer hohen Reanonymisierungswahrscheinlichkeit betroffenen Beobachtungen zusätzlich auch als Mikroaggregate verwenden werden. Die abhängige Variable $\tilde{Y}_i$ ist in diesem Fall eine Originalbeobachtung oder ein anonymisierter Wert Y_i^*:

$$\tilde{Y}_i = S_i\, Y_i + (1 - S_i)\, Y_i^* \tag{3.5}$$

Damit wird der Informationsverlust durch Anonymisierung im Vergleich zur Mikroaggregation über sämtliche Beobachtungen reduziert.

Es sei $n < N$ die Anzahl der Beobachtungen, die nicht von der Anonymisierung (Blanking) betroffen sind. Für diese Bebachtungen gilt die konditionale Populationsregressionsfunktion:

$$\begin{aligned} E\left[Y_i|\varphi\left(Z_i'\gamma\right), S_i = 1\right] &= f\left(X_i, \beta\right) + E\left[\varepsilon_i|\varphi\left(Z_i', \gamma\right), S_i = 1\right] \\ &= f\left(X_i, \beta\right) + \lambda\left(Z_i'\gamma\right) + \zeta_i, \end{aligned} \tag{3.6}$$

wobei $\lambda(\cdot)$ eine allgemeine Selektionskontrollfunktion bezeichnet und $\zeta_i = \varepsilon_i - \lambda_i$ ein heteroskedastischer Fehlerterm mit $E\left[\zeta_i|\varphi\left(Z_i, \gamma\right), S_i = 1\right] = 0$ ist. Die Identifikationsbedingungen für dieses semiparametrische Modell mit linearer Regressionsfunktion werden ausführlich in Newey (1999) diskutiert. Das Problem adäquater Ausschlussrestriktionen im Fall der Anonymisierung ist deutlich geringer als bei typischen Anwendungen von Selektionskorrekturverfahren, die auf dem Prinzip der Selektion über unbeobachtbare Faktoren (selection on unobservables) beruhen. In unserem Fall beruht die Selektion in aller Regel auch auf Variablen, die nicht in der Strukturgleichung als erklärende Variablen enthalten sind. Diese Variablen liefern die notwendigen überidentifizierenden Restriktionen. Es ist wichtig darauf hinzuweisen, dass eine eventuelle Konstante in diesem Modell über den Korrekturterm λ_i aufgefangen wird und nicht ohne weitere Annahmen (vgl. Andrews u. Schafgans 1998) identifizierbar ist.

Das nichtlineare Modell mit semiparametrischer Selektionskontrollfunktion wird im Folgenden über ein zweistufiges Verfahren ähnlich dem Zwei-Stufen-Schätzer von Heckman geschätzt. In der ersten Stufe werden die Parameter der Selektionsgleichung mit Hilfe eines semiparametrischen Schätzers für binäre Auswahlmodelle geschätzt. Hierfür verwenden wir den von Klein und Spady (1993) vorgeschlagenen semiparametrisch effizienten Schätzer. Als Alternative sind andere semiparametrische $\sqrt{N}-$konsistente Schätzer denkbar, wie z.B. der semi-nichtparametrische Likelihood-Ansatz für binäre Auswahlmodelle von Gabler, Laisney, Lechner (1993) oder der semiparametrische Momentenschätzer von Ichimura (1993). Für die zweite Schätzstufe verwenden wir Neweys (1999) semiparametrischen Schätzer, bei dem die Selektionskontrollfunktion durch eine allgemeine Reihenapproximation ersetzt wird.[4)]

Das Klein-Spady Verfahren beruht auf einem parametrischen Likelihood-Ansatz, bei dem die binäre Auswahlwahrscheinlichkeit $P\left(S_i = 1|Z_i'\gamma\right)$ unspezifiziert bleibt:

$$ln\, L(\gamma) = \sum_{i=1}^{n} S_i\, ln\, P\left[S_i = 1|Z_i'\gamma\right] + (1 - S_i)\, ln\left[1 - P\left[S_i = 1|Z_i'\gamma\right]\right] \tag{3.7}$$

4) Für den linearen Fall bietet sich auch an, das Verfahren von Powell (1987) zu verwenden, das auf einer Kernschätzung der Kontrollfunktion beruht. Siehe Newey, Powell und Walker (1990) für eine vergleichende Studie.

Klein und Spady formulieren diese Wahrscheinlichkeit mittels des Bayes Theorems um als

$$P(S_i = 1|Z_i'\gamma) = \frac{P(S_i = 1)\, g_{I|S=1}(Z_i'\gamma|S_i = 1)}{g_I(Z_i'\gamma)}, \qquad (3.8)$$

wobei g_I die Dichte der Indexfunktion $I_i = Z_i'\gamma$ ist und $g_{I|S=1}$ die konditionale Dichte, gegeben $S_i = 1$. Die Auswahlwahrscheinlichkeit (3.8) wird geschätzt, indem sämtliche Terme dieser Wahrscheinlichkeit unabhängig voneinander nichtparametrisch geschätzt werden.[5] Durch Ersetzen der Auswahlwahrscheinlichkeit durch den Schätzer ergibt sich die Quasi-Likelihood-Funktion:[6]

$$\max_{\gamma} \ln Q(\gamma) = \sum_{i=1}^{n} S_i \ln\left(\left[\hat{P}\left[S_i = 1|Z_i'\gamma\right]\right]^2\right) + (1 - S_i)\ln\left(\left(1 - \hat{P}\left[S_i = 1|Z'\gamma_i\right]\right)^2\right) \quad (3.9)$$

Für die zweite Schätzstufe schlägt Newey vor, die unbekannte Kontrollfunktion $\lambda(\cdot)$ mit einer linearen Kombination von J Grundfunktionen ρ_j zu approximieren:

$$\lambda(\cdot) = \sum_{j=1}^{J} \eta_j \cdot \rho_j, \qquad (3.10)$$

wobei für $J \to \infty$ der Approximationsfehler verschwindet und η_j ein unbekannter zu schätzender Koeffizient ist. Diese Grundfunktionen hängen nur von der Indexfunktion ab. Ersetzen wir λ durch die Approximation (3.10) erhalten wir:

$$Y_i = f(X_i, \beta) + \sum_{j=1}^{J} \eta_j \rho_j(\tau - Z_i'\hat{\gamma}) + \hat{\xi}_i \qquad \hat{\xi}_i = \Sigma \eta_j (\rho_j - \hat{\rho}_j) + \xi_i \qquad (3.11)$$

Die Koeffizienten β und η_j können nun mit der nichtlinearen KQ-Methode geschätzt werden. Die optimale Ordnung von J wird durch ein Optimierungsverfahren bestimmt (siehe Appendix A II). Newey schlägt vor, die folgende polynomiale Approximation zu verwenden:

$$\rho_j(\tau - Z_i'\gamma) = [\Psi(\tau - Z_i'\gamma)]^j,$$

wobei Ψ eine monotone auf das Intervall $[-1;1]$ beschränkte Funktion ist. Weitere Details zum Newey-Verfahren findet der interessierte Leser in Appendix A II.

Monte-Carlo Evidenz

Mit Hilfe einer einfachen Monte-Carlo Studie soll im folgenden überprüft werden, ob die wahren Modellparameter möglichst akkurat mit Hilfe des vorgestellten zweistufigen semiparametrischen Selektionsverfahrens geschätzt werden können, wenn die Anonymi-

5) Die beiden Dichten lassen sich mit univariatem Kernschätzer schätzen. $P[S_i = 1]$ kann durch das arithmetische Mittel geschätzt werden.

6) Die geschätzte Wahrscheinlichkeit wird quadriert, weil deren Schätzung u. U. auch negativ sein kann.

sierung durch Blanking gemäß der Quantilsregel (3.2) erfolgt. Die zu anonymisierenden Regressoren der Selektionsgleichung werden gemäß (3.4) transformiert. Das gewählte Design der Simulationen ist im Wesentlichen das gleiche wie im vorherigen Abschnitt. Geschätzt werden soll wiederum ein lineares Modell mit den selben wahren Parameterwerten und einem t – verteilten Fehlertermprozess.

Die erklärenden Variablen und die weiteren Instrumente der Selektionsgleichung werden als multivariate normalverteilte Zufallsvariablen der Form

$$\begin{pmatrix} Z_1 \\ Z_2 \\ Z_3 \\ Z_4 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & .4 & .2 & .1 \\ .4 & 1 & .3 & .2 \\ .2 & .3 & 1.2 & .1 \\ .1 & .2 & .1 & .8 \end{pmatrix} \right)$$

gezogen, wobei $X_1 = Z_1$ und $X_2 = Z_2$ die erklärenden Variablen des Regressionsmodells bilden. Eine Beobachtung i wird aus dem Datensatz gelöscht $(S_i = 0)$, wenn irgendeine der Variablen von $(W_i = Y_{1i}, Z_{1i}, \ldots, Z_{4i})$ größer ist als das 90%-Quantil dieser Variablen.

$$S_i = \begin{cases} 1 \text{ wenn } 1\!\!1(Y_i < q_{.90}(Y)) \cdot \prod_{j=1}^{4} 1\!\!1(Z_{ij} < q_{.90}(Z_j)) = 1 \\ 0 \text{ sonst} \end{cases}$$

Tabelle 3 (siehe S. 131) gibt die Ergebnisse der Monte-Carlo Simulationen für die Stückprobenumfänge N = 120, 1 200 und 3 600 wieder. Durch das Löschen von Beobachtungen beruhen jedoch die Regressionen der zweiten Stufe auf geringeren Stichprobenumfängen. Für die Stichprobenverzerrung relevant ist ausschließlich die Unterdrückung von Beobachtungen mit Ausprägung der abhängigen Variablen oberhalb des 90 %-Quantils, während die Unterdrückung von Beobachtungen aufgrund extremer Werte anderer Variablen zu einem Verlust an Effizienz führt. Der Effizienzverlust durch die hier vorgegebene Form des Data Blanking speist sich aus zwei Quellen. Zum einen werden in der zweiten Stufe weniger Beobachtungen verwendet, zum anderen wird die Stichprobenvariation der erklärenden Variablen reduziert. In Tabelle 3 bezeichnet $\bar{n}$ die durchschnittliche Anzahl von Beobachtungen, die aufgrund der Anonymisierung in der zweiten Stufe verwendet wurde, während $\bar{n}_X$ den durchschnittlichen Beobachtungsumfang bezeichnet, wenn die Selektion ausschließlich über die Z -Variablen erfolgt. Für die drei Experimente reduziert sich der Stichprobenumfang in der zweiten Stufe um 37 bis 40 %.

Bei einem kleinen Stichprobenumfang von $N = 120$ (bzw. $\bar{n} = 72.81$) weist der zweistufige semiparametrische Schätzer eine mittlere Verzerrung auf, die über der Verzerrung der Aggregationsschätzer liegt. Allerdings reduziert sich diese Verzerrung deutlich mit steigendem Stichprobenumfang. Selbst bei kleinen Stichprobenumfängen ist die Schätzunsicherheit in termini des RMSE auf einem vergleichbaren Niveau wie die der IV-Schätzer und liegt deutlich niedriger als bei den Aggregationsschätzern.

Für kleine Stichprobenumfänge wird der Standardfehler des Selektionsschätzers deutlich zu hoch ausgewiesen. Aber schon bei einer Stichprobengröße von $N = 1200$ ($\overline{n} = 754.09$) scheint die asymptotische Approximation zu greifen, so dass sich Standardabweichung der Schätzungen und Mittelwert der geschätzten Standardabweichungen angleichen.

Tabelle 3: Monte-Carlo Ergebnis: Semiparametrisches Selektionskorrektur-Modell *)

$\beta = 1$	Mittelwert	Verzerrung	RMSE	RELSE
		$N = 120\ (\overline{n} = 72.81, \overline{n}_Z = 80.51)$		
$\hat{\beta}$	1.006	.006	.133	1.016
$\hat{\beta}_{NS}$	1.019	.019	.194	1.646
		$N = 1200\ (\overline{n} = 754.09, \overline{n}_Z = 827.22)$		
$\hat{\beta}$	1.001	.001	.042	1.003
$\hat{\beta}_{NS}$	1.004	-.004	.054	1.048
		$N = 3600\ (\overline{n} = 2269.54, \overline{n}_Z = 2487.20)$		
$\hat{\beta}$	1.000	.000	.024	.995
$\hat{\beta}_{NS}$	1.007	-.007	.034	.989

*) Schätzung des Koeffizienten vor der ersten erklärenden Variablen, Anzahl der Replikationen = 1 000.

4 Ein illustratives Beispiel

Da die praktische Relevanz von Monte-Carlo Ergebnissen von den Annahmen über den zugrunde gelegten stochastischen Prozess bzw. der Realitätsnähe dieser Annahmen abhängen, sollen anhand einer empirischen Anwendung die Auswirkungen von Aggregationsmethoden untersucht werden. Hierfür verwenden wir einen Querschnitt von 3 600 Firmen des Verarbeitenden Gewerbes der Kostenstrukturerhebung (KSE) des Jahres 1999. Erklärt werden soll der Anteil der gesetzlichen Sozialkosten einer Firma in Abhängigkeit von der Anzahl der vollzeitbeschäftigten Arbeitnehmer und der Anzahl der teilzeitbeschäftigten Arbeitnehmer. Das gewählte Anwendungsbeispiel soll eine mögliche, wenn auch stark vereinfachte Anwendung für anonymisierte Daten der KSE sein. In diesem Beispiel geben die Regressionskoeffizienten einen Hinweis darauf, inwieweit die gesetzliche Sozialkostenbelastung auf Unternehmensebene von der Beschäftigungsstruktur abhängt. Nicht uninteressant ist die Fragestellung, ob die Beschäftigung von Teilzeitbeschäftigten im Vergleich zu Vollzeitbeschäftigten kostenneutral erfolgt. Die

beiden erklärenden Variablen werden in standardisierter Form als Regressoren verwendet. Eine Standardisierung ist sinnvoll, um Regressoren von unterschiedlicher Dimension oder unterschiedlicher Skalierung mit Störgrößen mit gleicher Varianz zu überlagern. Wie in der Monte-Carlo-Studie zuvor, wählen wir normal verteilte Überlagerungsfehler mit einer Varianz von 0.25.

Tabelle 4 (siehe S. 133) gibt die Schätzergebnisse für die gewöhnliche KQ-Schätzung auf der Grundlage der Originaldaten sowie die Ergebnisse für anonymisierten Datensätze wieder. Deutlicher als in den beiden Monte-Carlo Experimenten zuvor zeigen sich erhebliche Unterschiede zwischen der „Originalschätzung" und den Schätzungen, die auf den weniger informativen anonymisierten Datensätzen beruhen.

Unsere Schätzergebnisse verdeutlichen recht anschaulich, dass die Wahl der Anonymisierungsmethode sowie die Wahl der entsprechenden Anonymisierungsparameter (z.B. Höhe des Aggregationsniveaus, Größenordnung der Überlagerung) die Schätzergebnisse substantiell beeinflussen. Die auf Grundlage der Originaldaten geschätzten Koeffizienten sind statistisch auf dem 1 % Signifikanzniveau abgesichert. Die Aggregationsschätzer und der Bootstrap-Aggregationsschätzer liefern ähnliche Parameterschätzungen. Allerdings ist der Koeffizient vor der Variablen Teilzeitbeschäftigte für den Bootstrap-Schätzer und den Aggregationsschätzer mit A = 5 nicht mehr statistisch abgesichert. Die Ergebnisse sind aber möglicherweise für die einfachen Aggregationsschätzer beschönigend, da durch die spezielle Sortierung des Originaldatensatzes homogene Firmen der gleichen Bereiche aggregiert wurden. Der Bootstrap-Aggregationsschätzer beruht auf einer einzigen Bootstrap-Aggregation für $B = 3$. Eine Schätzung auf einer anderen zufälligen Aggregation, die hier nicht wiedergegeben wird, führt zu einem positiven Koeffizienten vor der Teilzeitbeschäftigungsvariablen. Der Instrumentvariablen-Schätzer und der korrigierte Fehler-in-den-Variablen-Schätzer liefern ähnliche Ergebnisse wie der OLS-Schätzer, jedoch ist auch hier der letzte Regressionskoeffizient statistisch nicht abgesichert.

Tabelle 4: Auswirkungen der Anonymisierung: Ein Anwendungsbeispiel *)

	Konstante	Vollzeitbeschäftigte	Teilzeitbeschäftigte
OLS	0.120 (5.39)	0.598 (9.98)	-0.165 (-2.747)
$B = 3$	0.120 (9.15)	0.485 (2.14)	-0.029 (-0.16)
$A = 3$	0.120 (4.98)	0.627 (6.44)	-0.212 (-2.00)
$A = 4$	0.120 (4.83)	0.755 (6.48	-0.404 (-3.12)
$A = 5$	0.120 (4.71)	0.608 (4.73	-0.157 (-1.09)
EIV	0.130 (5.41)	0.786 (8.96)	0.083 (2.59)
IV	0.131 (5.45)	0.412 (3.33)	0.003 (0.02)
CEIV	0.134 (5.58)	0.660 (4.95)	0.0243 (-1.82)

*) Abhängige Variable: log Gesetzliche Sozialkosten, t -Werte in Klammern.

5 Schlussfolgerung

In dieser Arbeit werden verschiedene Anonymisierungsmethoden hinsichtlich ihrer Auswirkung auf die Qualität von ökonometrischen Schätzungen untersucht. Es wird gezeigt, dass standardmäßige Anonymisierungsverfahren wie Mikroaggregation und stochastische Überlagerung, sofern ihre Auswirkungen auf den generierenden Prozess für den Anwender bekannt sind, nicht unbedingt zu einer gravierenden Reduktion der Qualität der Schätzungen führen müssen. Hierzu muss jedoch die Struktur des Anonymisierungsverfahrens (z.B. Verlässlichkeitsquoten im Falle der stochastischen Überlagerung) dem Empiriker bekannt sein. Bei kleinen Stichproben kann Mikroaggregation zu einer deutlichen Reduktion der Schätzgenauigkeit führen. Wir zeigen, dass die stochastische Überlagerung als Anonymisierungsverfahren eine attraktive Alternative zur Mikroaggregation darstellt, sofern die datenerhebende Institution Informationen über die Kovarianzstruktur der Überlagerung dem Empiriker zu Händen gibt.

Die schöne heile Welt der Anonymisierung kann aber nur für einfache Anonymisierungsverfahren und Anwendungen des linearen Regressionsmodells aufrecht erhalten werden. Sobald die Aggregation gewichtet erfolgt und die Gewichtung auf einer potentiell endogenen Variablen beruht, haben wir es mit komplexen Selektionsmechanismen zu tun, die sich nur schwer modellieren lassen.

Die Analyse von Mikrodaten erfordert fast zwangsläufig die Verwendung von nichtlinearen Regressionsmodellen (qualitative Auswahlmodelle, Regressionsmodelle für begrenzt abhängige Variablen, Zähldatenmodelle etc.). Stochastische Überlagerung führt in diesem Fall zu komplexen nichtlinearen Fehler-in-den-Variablen-Modellen. Diese Modelle für eine allgemeine Struktur der Überlagerungsfehler (Zählvariablen-Fehler, Fehler für nominal skalierte Variablen, Fehler für stetige intervallskalierte Variablen etc.) und eine allgemeine nichtlineare Form zu schätzen, ist nicht unbedingt als trivial zu bezeichnen. In dieser Arbeit zeigen wir, wie ein allgemeines, möglicherweise nichtlineares Modell über einen semiparametrischen, zweistufigen Selektionskontrollschätzer geschätzt werden kann. Der Schätzer unterscheidet sich von Heckmans Zwei-Stufen-Schätzer für Selektionsmodelle dadurch, dass keine Verteilungsannahmen bezüglich der Fehlerterme der Selektionsgleichung und der Strukturgleichung getroffen werden und die Selektionswahrscheinlichkeit nur auf der Single-Index-Struktur beruht. Anhand von Monte-Carlo-Simulationen und eines empirischen Beispiels zeigen wir, dass dieser Ansatz zumindest bei größeren Stichproben ein gangbarer Weg ist, eine Selektionskorrektur infolge von „Data Blanking" in nichtlinearen Modellen durchzuführen. Obwohl der hier verwendete Blanking-Mechanismus nicht die Form eines schwellenüberschreitenden binären Auswahlmodells aufweist, scheint die semiparametrische Single-Index-Struktur durchaus geeignet zu sein, den Selektionsmechanismus abzubilden.

Die zukünftige Forschung sollte sich weiter darauf konzentrieren, adäquate nichtlineare Schätzer für anonymisierte Mikrodaten zu entwickeln, da anderenfalls der Wert wissenschaftlich ergiebiger, aber anonymisierter Individualdaten erheblich eingeschränkt wird. Mehrere Wege bieten sich für die zukünftige Forschung an. Im Kontext der Selektionsmodelle scheint der Versuch sinnvoll zu sein, die Anonymisierungswahrscheinlichkeit genauer abzubilden, um in der zweiten Stufe eine präziser geschätzte Kontrollfunktion zu erhalten. Für lineare Strukturgleichungen sollten andere Verfahren (z.B. der Schätzer von Powell 1987) mit den hier verwendeten Schätzern verglichen werden.

Das „Blanking" von Daten ist nur ein grobes Anonymisierungsverfahren. Selektionsmodelle könnten analog zu Lanot und Walker (1998) um eine weitere Gleichung für anonymisierte Beobachtungen erweitert werden, um sämtliche Beobachtungen des Originaldatensatzes für die Regressionsanalyse zu verwenden und somit den Informationsverlust zu reduzieren.

Literaturhinweise

Amemiya, T. (1985): Instrumental Variable Estimator for the Non-linear Errors in Variable Model, in: Journal of Econometrics, 28, S. 273 – 289.

Andrews, D. and Schafgans, M. (1998): Semiparametric Estimation of the Intercept of a Sample Selection Model, in: Review of Economic Studies, 65, S. 497 – 517.

Brand, R. (2000): Anonymität von Betriebsdaten, Beiträge zur Arbeitsmarkt- und Berufsforschung, BeitrAB 237, IAB, Nürnberg.

Carroll, R.; Ruppert, D. and Stefanski, L. F. (1995): Measurement Error in Nonlinear Models, Chapman and Hall.

Fuller, W. A. (1987): Measurement Error Models, Wiley.

Gabler, S.; Laisney, F. and Lechner, M. (1993): Semiparametric Estimation of Binary Choice Models with an Application to Labor Force Participation, in: Journal of Business and Economic Statistics, 11, S. 61 – 80.

Gottschalk, S. (2002): Anonymisierung von Unternehmensdaten: Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels, Discussion Paper No. 02 – 23, Mannheim, ZEW.

Hausman, J.; Newey, W. and Powell, J. (1995): Nonlinear Errors in Variables Models, in: Journal of Econometrics, 41, S. 159 – 185.

Hong, H. and Tamer, E. (2002): A Simple Estimator for Nonlinear Error in Variable Models, Princeton University, unpublished.

Ichimura, H. (1993): Semiparametric Least Squares (SLS) and weighted SLS Estimation of Single-Index Models, in: Journal of Econometrics, 58, S. 71 – 120.

Klein, R. W. and Spady, R. S. (1993): An Efficient Semiparametric Estimator of the Binary Response Model, in: Econometrica, 61, S. 387 – 421.

Lanot, G. and Walker, I. (1998): The Union/Non Union Wage Differential: An Application of Semi-Parametric Methods, in: Journal of Econometrics, 84, S. 327 – 349.

Lee, L. F. and Sepanski, J. H. (1995): Estimation of Linear and Nonlinear Error in Variables Models Using Validation Data, in: Journal of the American Statistical Association, 90, S. 130 – 140.

Lechner, S. and Pohlmeier, W. (2003): Microaggregation in Nonlinear Models: A Note, Center of Finance and Econometrics, University of Konstanz, unpublished working paper.

Newey, W. K.; Powell, J. L. and Walker, J. R. (1990): Semiparametric Estimation of Selection Models: Some Empirical Results, in: American Economic Review, Paper and Proceedings, 80, S. 324 – 328.

Newey, W. K. (1999): Two step Series Estimation of Sample selection Models, Department of Economics, Working Papers No–99–04, Massachusetts, Institute of Technology.

Paaß, G. und Wauschkuhn, U. (1984): Datenzugang, Datenschutz, und Anonymisierung, Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten, in: Berichte der Gesellschaft für Mathematik und Datenverarbeitung, Bericht 148, Oldenbourg Verlag.

Powell, J. L. (1987): Semiparametric Estimation of Bivariate Latent Variable Models, Working Paper No. 8704, SSRI, University of Wisconsin.

Anhang A I

Proposition 1:

$V\left[\hat{\beta}_A\right]-V\left[\hat{\beta}\right]$ ist positiv definit.

Beweis:

$V\left[\hat{\beta}_A\right]-V\left[\hat{\beta}\right]$ ist nur positiv definit, wenn und nur wenn die Differenz der Inversen der Varianz-Kovarianzmatrizen $V\left[\hat{\beta}\right]^{-1}-V\left[\hat{\beta}_A\right]^{-1}$, positiv definit ist.

Unter Vernachlässigung von σ^2 gilt hierfür

$$\begin{aligned} X'X'-X^{*'}X^{*} &= X'[I-D'D]X \\ &= X'[I-D]X \\ &= X'WX, \end{aligned}$$

wobei D und $W=I-D$ symmetrische, idempotente Formen sind. Da X und D vollen Spaltenrang besitzen, gilt für jeden Vektor $q\neq 0$:

$$q'X'WXq = q'\tilde{X}'\tilde{X}q = v'v > 0,$$

wobei $\tilde{X}=WX$ und $v=\tilde{X}q$.

Proposition 2:

$$E\left[\frac{e^{*'}e^{*}}{M-K}\right]=\sigma^2$$

Beweis:

$$\begin{aligned} E\left[e^{*'}e^{*}\right] &= E\left[\varepsilon^{*'}M^{*}\varepsilon^{*}\right] \\ &= E\left[tr\varepsilon^{*'}M^{*}\varepsilon^{*}\right] \\ &= trM^{*}E\left[\varepsilon^{*}\varepsilon^{*'}\right] \\ &= \sigma^2 trM^{*}D \\ &= \sigma^2\left(trD-trX^{*'}\left(X^{*'}X^{*}\right)^{-1}X^{*'}\right) \\ &= \sigma^2(M-K), \end{aligned}$$

so dass für die um die Freiheitsgrade $M-K$ korrigierte Fehlerquadratsumme die Proposition 2 hält.

Anhang A II

Varianz Matrix des Newey-Series Schätzer

Es sei $\hat{W}_i = \left(X_i', \hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{Ji}\right)'$,

$\hat{W} = \left(\hat{W}_1, \hat{W}_2 \ldots, \hat{W}_n\right)'$,

$\hat{\theta} = \left(\hat{\beta}_{NS}', \hat{\eta}_1, \hat{\eta}_2, \ldots, \hat{\eta}_J\right)'$,

Die optimale Anzahl der Grundfunktionen minimiert die folgende Funktion

$$J_{OPT} = arg\,min\; CV\,(J) = \sum_{i=1}^{n}\left[\frac{\left(1-2\,\hat{\delta}_i\right)\hat{e}_i}{1-\hat{\delta}_i}\right]^2,$$

wobei $\hat{\delta}_i = \hat{W}_i'\left(\hat{W}'\hat{W}\right)^{-1}\hat{W}_i$ und $\hat{e}_i = Y_i - \hat{W}_i'\hat{\theta}$.

$$\hat{V}_{NS} = [I_k, 0]\left(\frac{\hat{W}'\hat{W}}{n}\right)^{-1} \quad A\left(\frac{\hat{W}'W}{n}\right)^{-1}[I_k\ 0]'$$

$$A = \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i\hat{W}_i' \quad \left(Y_i - \hat{W}_i\,\hat{\theta}\right)^2 + \hat{H}\,\hat{V}\,(\hat{\gamma})\,\hat{H}'$$

$$\hat{H} = \frac{1}{n}\sum_{i=1}^{n}\hat{W}_i\left[\frac{\partial\left(\sum_{j=1}^{J}\hat{\eta}_j\cdot\rho_j\left(Z_i'\,\hat{\gamma}\right)\right)}{\partial\left(Z_i'\,\hat{\gamma}\right)}\right]\cdot\left[\frac{\partial\left(Z_i'\,\hat{\gamma}\right)}{\partial\,\gamma'}\right].$$

wobei I_k die Einheitsmatrix, deren Dimension gleich der Anzahl der erklärenden Variablen in der Strukturellgleichung ist. $\hat{V}\,(\hat{\gamma})$ ist eine konsistente Schätzung der Varianz des Schätzers der ersten Stufe.

Gerd Ronning *)

Ökonometrie und Anonymisierung von Mikrodaten

Koreferat zum Beitrag „Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten"

Der Beitrag von Lechner und Pohlmeier nimmt eine Frage auf, die bisher so gut wie gar nicht gestellt wurde, einfach weil es keinen Anlass dafür gab. Mikrodaten wurden entweder als Originaldaten zur Verfügung gestellt und mit ökonometrischen Methoden analysiert oder die Daten waren – vor allem aus Geheimhaltungsgründen – nicht verfügbar. Die jetzt erkundete Möglichkeit, Mikrodaten in anonymisierter Form den empirischen Wirtschafts- und Sozialforschern zur Verfügung zu stellen, provoziert unmittelbar die Frage, was dies für die Schätzung ökonometrischer Modelle bedeutet. Ich bin den beiden Autoren dankbar, dass sie sich im Rahmen unseres Forschungsprojektes dieser Frage widmen und mit dem vorliegenden Beitrag innovativ und befruchtend Neuland der ökonometrischen Methodik betreten. Und es wäre schön, wenn in dieser Richtung bald weitere Untersuchungen durchgeführt würden.

Die Autoren demonstrieren, dass sich bei bestimmten Anonymisierungsverfahren das Ausmaß der Verzerrung quantifizieren lässt und eine Schätzformel, die diese Verzerrung berücksichtigt, zufriedenstellende Ergebnisse bringt. Für den Fall der Zufallsüberlagerung, die eine direkte Umsetzung des „Fehler-in-den-Variablen"-Modells ist, ist dies nicht so überraschend und übrigens bereits in den achtziger Jahren von Wayne Fuller als Idee skizziert worden.

Die beiden Autoren untersuchen jedoch auch weitere Anonymisierungsverfahren, wobei anzumerken ist, dass diese in der betrachteten Form nicht im Projekt verwendet werden. Neben einer „deterministischen" Variante der Mikroaggregation, die rein zufällig jeweils Gruppen bestimmter Größe bildet und nicht die Ähnlichkeit berücksichtigt, wird auch eine „stochastische" Variante der Mikroaggregation betrachtet, in der jeder Einheit das arithmetische Mittel aus zufällig gezogenen Stichproben-Werten als anonymisierter Wert zugeordnet wird. Die Autoren nennen dieses Verfahren auch „Bootstrap-Aggregation". Zu beachten ist, dass die Autoren den Begriff Mikroaggregation in bestimmten Sinne wörtlich nehmen: Der anonymisierte Datensatz enthält – in der deterministischen Variante – nur noch Gruppen als Untersuchungseinheiten, während üblicherweise jedem Element der Gruppe dieser Mittelwert zugeordnet wird. Deshalb problematisieren die Autoren – in Abschnitt 2 – auch die Schätzung der Restvarianz bzw. die korrekte Anzahl Freiheitsgrade, die andernfalls entfallen würde. Man beachte, dass im Fall der stochastischen Mikroaggregation jede Untersuchungseinheit mit einem anonymisierten Wert versorgt wird. Dort stellt sich das Problem der angemessenen Schätzung der Restvarianz ohnehin in andere Weise, die von den Autoren beschrieben wird.

*) Prof. Dr. Gerd Ronning, Universität Tübingen und Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen.

Im Rahmen des linearen Modells lassen sich sowohl stochastische Überlagerung als auch beide Varianten der Mikroaggregation direkt in das Modell bzw. in die Koeffizienten-Schätzung integrieren. Für die deterministischen Variante der Mikroaggregation ergibt sich bei Verwendung der „naiven" Kleinstquadrate-Schätzung weiterhin eine unverzerrte Schätzung, aber ein Effizienzverlust, der umso größer ist, je weniger ähnlich die Gruppen sind, was sich auch analytisch belegen lässt. Auch für die Bootstrap-Aggregation ergeben sich zufriedenstellende Schätzwerte. Wegen des im Zweifel geringeren Re-Identifikationsrisikos sollte diese Variante der Mikroaggregation in zukünftigen Untersuchungen mit berücksichtigt werden.

Für die stochastische Überlagerung wird der Instrumentvariablen-Schätzer erfolgreich eingesetzt, wobei die spezielle Konstruktion der Instrumente in dieser Anwendung bemerkenswert ist: Die Instrumente werden durch stochastische Überlagerung der Originalwerte erzeugt und erfüllen damit perfekt die Anforderungen an ökonometrische Instrument-Variablen. Allerdings ist dies in der Praxis nicht operational, weil die Originaldaten nicht zur Verfügung stehen. Die Autoren stellen diese Anonymisierungsmethode als „echte Alternative" zur Mikroaggregation dar. Aus Sicht des Reidentifikations-Risikos dürfte sich diese Sichtweise jedoch nicht halten lassen.

Besonders interessant war für mich der Vorschlag, die Anonymisierung durch Zensierung, d.h. Aussonderung von Untersuchungseinheiten mit besonders großen Werte von exogenen und endogenen Variablen in einem nichtlinearen Modell zu modellieren. Diese Selektionsprozedur wird in Anlehnung an eine Idee von Klein und Spady durch einen semiparametrischen Ansatz formuliert. Grundsätzlich ließen sich aber auch alternative Vorgehensweisen, z.B. Ersatz des Wertes durch den Schwellenwert, spezifizieren. Im simulierten Datensatz und bei Vorgabe der Zensierung am 90 %-Quantil ergeben sich bei Verwendung des Klein-Spady-Schätzers erstaunlich gute Schätzresultate. Allerdings ist anzumerken, dass die Autoren – entgegen der Überschrift von Abschnitt 3 – ihre Methode an einem linearen Modell erproben.

Zusammenfassend bleibt festzustellen, dass die in diesem Beitrag untersuchten Anonymisierungsmethoden überwiegend nicht die ansonsten im Projekt betrachteten Methoden sind, dass gleichwohl diese Arbeit einen wesentlichen Schritt in Richtung der Erweiterung der ökonometrischen Methodik tut, die angesichts der Verfügbarkeit von Scientific-Use-Files bald ein wesentliches Forschungsfeld werden dürfte. Dabei zeigt der Beitrag von Lechner und Pohlmeier deutlich, dass für eine analytische Integration der Anonymisierungsverfahren in die Schätzroutinen zwei Vorbedingungen gegeben sein müssen: (1) Das Anonymisierungsverfahren muss dem Anwender bekannt gegeben werden. (2) Das Anonymisierungsverfahren muss so einfach strukturiert sein, dass es analytisch zu bewältigen ist.

Joachim Wagner *)

Arbeiten mit Einzeldaten der amtlichen Statistik am Beispiel des Monatsberichts im Verarbeitenden Gewerbe

1 Warum gibt es diesen Vortrag hier auf diesem Workshop?

Wenn (fast) am Ende eines Workshops zum Thema „Anonymisierung wirtschaftsstatistischer Einzeldaten" ein Vortrag steht, in dem über die Arbeit mit nicht-anonymisierten Einzeldaten der amtlichen Statistik am Beispiel des Monatsberichts im Verarbeitenden Gewerbe berichtet wird, dann geschieht dies aus einem nahe liegenden Grund: Es soll aufgezeigt werden, wofür diese Daten bisher – d.h. in nicht-anonymisierter Form – verwendet wurden, um damit Hinweise darauf zu geben, wofür die Daten in der Zukunft – und dann in einfacher zugänglicher Form – genutzt werden können, wenn (und nur wenn) eine Anonymisierung in einer für diese potentiellen Nutzungen geeigneten Form gelingt.

2 Was sind denn das für Daten und wer nutzt sie wie?

Ehe ich über die Arbeiten mit den Mikrodaten für Industriebetriebe aus Erhebungen der amtlichen Statistik berichte, möchte ich kurz daran erinnern, welche Informationen die Datensätze enthalten sowie wer bisher wie damit arbeiten konnte (vgl. ausführlicher hierzu Wagner 1999a, 2000):

Unter Vernachlässigung einiger Detailregelungen sind die Betriebe im Bergbau und Verarbeitenden Gewerbe, die eine Mindestgröße von 20 tätigen Personen überschreiten oder Teil eines entsprechend großen Mehrbetriebsunternehmens sind, die meldepflichtigen Betriebe zum so genannten Monatsbericht. Aus den Erhebungen der Statistischen Ämter der Länder liegen für diese Betriebe u.a. Informationen über Betriebsort, Wirtschaftszweig geleistete Arbeiterstunden, Bruttolohn- und -gehaltsumme, Anzahl der tätigen Personen insgesamt und der tätigen Arbeiter, Inlands- und Auslandsumsatz sowie den Produktionswert vor. Aus den monatlichen Meldungen lassen sich Jahressummen bzw. Durchschnittswerte für jeden Betrieb bilden. Diesen Daten können Informationen aus der jährlichen Investitionserhebung zugespielt werden. Anhand der unveränderlichen Betriebsnummer können die Angaben eines Betriebes über die Zeit hinweg zu einem Paneldatensatz verknüpft werden, dem so genannten Monatsmelder-Panel. Ergänzt man diesen um die entsprechend aufbereiteten Daten aus der jährlichen Erhebung in industriellen Kleinbetrieben, die unter dem o.a. Schwellenwert von 20 tätigen Personen liegen, so erhält man für den Bereich der Industrie das so genannte Totalerhebungs-Panel, das jedoch wegen des wesentlich weniger umfangreichen Fragenprogramms in der Kleinbetriebserhebung nur Informationen über Ort, Wirtschaftszweig, Anzahl tätige Personen und Umsatz enthält.

Paneldatensätze dieses eben beschriebenen Typs wurden in einer Reihe von Statistischen Ämtern der Länder aufbereitet, wobei das Niedersächsische Landesamt für Statistik (NLS) Anfang der neunziger Jahre des vorigen Jahrhunderts eine Pionierrolle spielte.

*) Prof. Joachim Wagner, Institut Volkswirtschaftslehre, Universität Lüneburg, HWWA Hamburgisches Welt-Wirtschaftsarchiv, Hamburg, IZA Forschungsinstitut zur Zukunft der Arbeit, Bonn.

Externe Wissenschaftler können mit diesen Daten arbeiten, wenn sie einen entsprechenden Vertag mit dem jeweiligen Statistischen Amt der Länder schließen. Hierbei ist zentral, dass diese geheimen Einzeldaten nur im Amt selbst genutzt werden können und sämtliche Ergebnisse vor einer Weitergabe auf Geheimhaltungsfreiheit zu prüfen sind. Für die Art des Datennutzungszugangs gibt es im Kern zwei Modelle, die sich mit den Schlagworten „Datenfernverarbeitung" und „Internalisierung von Externen" kennzeichnen lassen. Im ersten Fall senden die Wissenschaftler Auswertungsprogramme ins Amt; der Output wird dann geprüft und an die Wissenschaftler weitergeleitet, wenn er keine geheim zu haltenden Informationen enthält. Im zweiten Fall erhalten externe Wissenschaftler einen Status als Mitarbeiter im Amt (z.B. als Praktikant ohne Bezahlung), sind dann wie andere Mitarbeiter auch zur Geheimhaltung erhaltener Informationen verpflichtet, können vor Ort mit den Daten arbeiten, müssen aber die Ergebnisse vor einer Weitergabe ebenfalls auf Geheimhaltungsfreiheit prüfen lassen. Die Forschergruppen, die in verschiedenen Bundesländern mit diesen Datensätzen arbeiten, haben sich im Netzwerk FiDASt – Firmendaten aus der Amtlichen Statistik – zusammengeschlossen; dieses Netzwerk veranstaltet Workshops zum Erfahrungsaustausch und zur Diskussion von Forschungsergebnissen (vgl. Schasse und Wagner 1999, 2001).

3 Was kann man denn mit diesen Daten so machen?

Die zu Panels aufbereiteten Betriebsdaten aus dem Monatsbericht sowie aus der Verknüpfung von Monatsbericht und Kleinbetriebserhebung wurden in den vergangenen zehn Jahren für eine Vielzahl ganz unterschiedlicher empirischer Untersuchungen genutzt (vgl. Gerlach und Wagner 1997a sowie die oben genannten FiDASt-Workshop-Bände für Überblicke). Um einen Eindruck hiervon zu vermitteln, möchte ich zu fünf Arbeitsbereichen jeweils ein (eigenes) Beispiel kurz vorstellen:

Bisher unbekannte Fakten dokumentieren

Die Verknüpfung der Daten aus den Querschnittserhebungen zu einem Panel ermöglicht die Analyse von zeitlichen Entwicklungsprozessen auf einzelbetrieblicher Ebene. Dies eröffnet Auswertungsperspektiven, die über die von den statistischen Ämtern vorgenommenen Querschnittsanalysen weit hinausgehen. So können z.B. mit einem Totalerhebungspanel – wenn auch etwas unscharf – Betriebsgründungen identifiziert werden. Diese neuen Betriebe können dann über die Jahre nach der Gründung in dem Panel verfolgt werden. Damit wird sichtbar, was aus jedem dieser Betriebe geworden ist – wie lange hat er überlebt, ist er gewachsen oder geschrumpft? Mit dem niedersächsischen Totalerhebungspanel habe ich eine solche Studie durchgeführt (vgl. Wagner 1994a). Es zeigt sich, dass von einer Gründungskohorte (d.h. von allen Betrieben, die in einem bestimmten Jahr neu in den Markt eintraten) rund die Hälfte der Betriebe nach zehn Jahren bereits wieder geschlossen ist; allerdings sind die überlebenden Betriebe im Durchschnitt so stark gewachsen, dass der Beitrag einer Kohorte zur Gesamtbeschäftigung über die Zeit in etwa konstant bleibt. Die Anzahl neuer Arbeitsplätze in einer Kohorte im Jahr der Gründung ist damit ein recht guter Indikator für die Anzahl der Arbeitsplätze in dieser Kohorte in den Folgejahren – ein wichtiger Hinweis für eine Einschätzung der Beschäftigungseffekte von Gründungen.

Weitere neue Fakten dokumentierende Studien mit den Paneldaten untersuchen z.B. die Altersverteilung von Arbeitsplätzen (Wagner 1996), regionale (Gerlach/Wagner 1994) und sektorale Unterschiede im Gründungsgeschehen (Wagner 1994b), die Lebensgeschichte von geschlossenen Firmen (Wagner 1999b), Ein- und Austritte auf Exportmärkte (Bernard und Wagner 1997, 2001) und die Auswirkungen der Aufnahme einer Exporttätigkeit auf die Firmen (Wagner 2002).

Theoretische Hypothesen überprüfen

Die Paneldaten lassen sich verwenden, um theoretischen Hypothesen zu überprüfen, die Aussagen über Entwicklungen von Betrieben im Zeitablauf machen. Ein Beispiel hierfür ist die weit verbreitete These, dass die zu einem Zeitpunkt beobachtete Verteilung der Betriebsgröße das Ergebnis eines Zufallsprozesses ist, denn die Wachstumsrate eines Betriebes sei erstens von der Betriebsgröße im Vorjahr und zweitens von der Wachstumsrate in der Vorperiode unabhängig – es gelte demnach das Gibrat-Gesetz. Diese Hypothese habe ich mit den Daten aus dem niedersächsischen Totalerhebungspanel ökonometrisch überprüft – und verworfen (vgl. Wagner 1992): Zwar wachsen kleinere Firmen nicht systematisch schneller oder langsamer als größere, aber die Wachstumsraten sind von Periode zu Periode nicht unabhängig voneinander. Demnach wird die Hypothese von der Gültigkeit des Gibrat-Gesetzes hier abgelehnt.

Weitere Beispiele für die ökonometrische Überprüfung von theoretischen Hypothesen mit den Paneldaten betreffen Zusammenhänge zwischen Exportänderungen und Firmenwachstum (Wagner 1993, 1995a) und zwischen Betriebsgröße und Exporten (Wagner 2003). In diesen Studien wird der Panelcharakter der Daten nicht ausschließlich dafür genutzt, Veränderungsprozesse auf einzelbetrieblicher Ebene zu erfassen; hinzu kommt hier die Kontrolle der unbeobachteten Heterogenität durch die Verwendung von panelökonometrischen Schätzverfahren.

Eine fundamentale Annahme makroökonomischer Modelle kritisieren

Auch ein theoretischer Beitrag sollte, wie Wolfgang Franz (1995, S. 3) einmal bemerkt hat, nicht ohne Not an der Realität vorbeigehen. Wenn die fundamentalen Annahmen theoretischer Modelle im offensichtlichen Widerspruch zur uns umgebenden Realität stehen, so ist dies eine Herausforderung an die Modelltheoretiker – sie müssen begründen, warum sie trotzdem mit diesen Annahmen arbeiten. Ein Beispiel hierfür, dessen „Entdeckung" auf die Verwendung von Paneldaten des hier besprochenen Typs zurückgeht, folgt aus der in Analysen mit diesen Daten immer wieder aufgezeigten ausgeprägten Heterogenität der Betriebe: Gerlach und Wagner (1995) zeigen, dass sich auch bei sehr tiefer Disaggregation nach Industriezweigen innerhalb von Zwei-Jahres-Zeiträumen simultan nebeneinander wachsende und schrumpfende, neu gegründete und geschlossene Betriebe finden. Anders gesagt sind, um eines der Beispiele der genannten Studie zu nennen, die niedersächsischen Betriebe aus dem Industriezweig „Herstellung von Hütten- und Walzwerkeinrichtungen" nicht nur nicht alle im gleichen Maße gewachsen oder geschrumpft – sie entwickeln sich derart unterschiedlich, dass es angesichts dieser ausgeprägten Heterogenität keinen Sinn macht, von einem repräsentativen Durchschnittsbetrieb zu sprechen. Genau solche „repräsentative Firmen" sind aber ein Baustein der üblichen Makromodelle (soweit sie überhaupt eine Mikrofundierung verwenden). Modelltheoretiker müssen sich die Frage gefallen lassen, warum.

Eine solche ausgeprägte Heterogenität haben mikroökonometrische Studien mit Individual- und Firmendaten vielfach aufgezeigt. James Heckman (2001, p. 674 und p. 732) hat es in seiner Nobelpreisrede auf den Punkt gebracht: „The most important discovery [from microeconometric investigations, J.W.] was the evidence on the pervasiveness of heterogeneity and diversity in economic life. ... The evidence from microeconometric data has already had a substantial effect on the development of macroeconometric theory, which is slowly abandoning the representative agent paradigm."

Politiker beraten (vergeblich?)

Zu den hartnäckig von vielen (Wirtschafts-)Politikern vertretenen Thesen gehört die von „Jobmotor Mittelstand": Kleine Firmen schaffen Arbeitsplätze, große Firmen bauen welche ab. Gestützt wird diese These mit dem Hinweis auf Statistiken, in denen die Arbeitsplatzentwicklung in Betrieben verschiedener Größenklassen gegenüber gestellt wird. Hierbei werden die Betriebe nach ihrer Größe im jeweiligen Basisjahr klassifiziert. Es ist theoretisch bereits seit langer Zeit bekannt, dass in dem Maße, in dem die Arbeitsplatzentwicklung auf einzelbetrieblicher Ebene transitorischen Charakter hat und mit Größenklassenwechseln verbunden ist, diese Klassifikation dazu führt, dass kleinere Betriebe systematisch zu gut und grössere systematisch zu schlecht abschneiden. Das Ausmaß, in dem dies der Fall ist, und die Stärke der Verzerrung des Bildes von der Rolle kleiner und großer Betriebe bei der Arbeitsplatzschaffung und -vernichtung läßt sich mit den hier behandelten Paneldaten aufzeigen. Untersuchungen für die niedersächsische Industrie zeigen, dass bei einer theoretisch korrekten Klassifikation der Betriebe nach ihrer Durchschnittsgröße (statt nach ihrer Größe im Basisjahr) die These eines inversen Zusammenhangs zwischen Betriebsgröße und Arbeitsplatzschaffung nicht haltbar ist (vgl. Wagner 1995b).

Diese Schlussfolgerung wird gestützt durch die oben angesprochenen Ergebnisse der Überprüfung des Gibrat-Gesetzes – gezeigt wurde, dass es keinen Zusammenhang zwischen Betriebsgröße und Wachstumsrate gibt – und durch die ebenfalls erwähnten Befunde zur ausgeprägten Heterogenität der Betriebe – wachsende und schrumpfende, neue und geschlossene Betriebe finden sich auch innerhalb jeder Größenklasse in jeder Periode. (Wirtschafts-)Politiker nehmen diese Tatsachen (zumindest nach meiner Erfahrung) nicht zur Kenntnis – auch dann nicht, wenn sie in ihrem Auftrag erarbeiten bzw. ihnen vorgetragen werden. Aber das liegt an den Politikern; die Daten liefern jedenfalls eine gute Basis für fundierte politikberatende angewandte empirische Forschung zu „handfesten" Thesen und Themen.

Die Qualität der Ausbildung in empirischer Wirtschaftsforschung steigern

Schwimmen lernt man im See, Schlitten fahren im Schnee – und empirische Wirtschaftsforschung (auch) am PC, und zwar am besten mit echten Daten und all ihren echten Problemen und Tücken. Forderungen, hier vermehrte Anstrengungen zu unternehmen, um die deutsche wirtschaftswissenschaftliche Forschung an den internationalen Standard einen Schritt (oder auch mehrere) heran zu führen, finden sich in jüngerer Zeit z.B. von der Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (2001) oder vom Wissenschaftsrat (2002). Studierenden können die Paneldaten aus den Erhebungen der amtlichen Statistik bei Industriebetrieben in nicht-anonymisierter Form nur in Ausnahmefällen über die oben geschilderten Wege zu-

gänglich gemacht werden; hier bietet es sich an, die so genannten „Spieldatensätze" zu nutzen, die Originaldatensätze enthalten, bei den Identnummern, Wirtschaftszweig und Kreis durch aussagelose veränderte Angaben ersetzt wurden. Meine Erfahrungen hiermit sind gut. Doktoranden und Habilitanden steht die Nutzung der Daten offen wie allen anderen externen Wissenschaftlern auch – und auch hier sind bereits sehr gute Erfolge zu verzeichnen.

In Zukunft könnte die Bedeutung der Rolle der Betriebspanel aus Erhebungen der amtlichen Statistik bei der Ausbildung des wissenschaftlichen Nachwuchses noch steigen, wenn es im Zuge der flächendeckenden Einführung von Promotionsstudiengängen an mehr und mehr Universitäten verpflichtenden Module gibt, in denen ein empirisches Projekt mit Mikrodaten zu bearbeiten ist.

4 Braucht man dafür denn unbedingt diese geheimen Daten?

Ja. Zwar lassen sich mit den Betriebspaneldaten aus den Erhebungen der amtlichen Statistik aufgrund des recht eingeschränkten Fragenprogramms auch nur ausgewählte Probleme untersuchen – die aber wegen des umfassenden Berichtskreises und der Auskunftspflicht auf einer verlässlichen Grundlage und nicht nur auf der Basis einer Stichprobe mit oft fehlenden, fehlerhaften oder falschen Angaben. Zudem sind diese Daten zu geringen Kosten und für einen langen Zeitraum sofort verfügbar. Betriebspaneldaten aus eigenen Erhebungen der Wissenschaft wie das IAB-Betriebspanel (Kölling 2000) oder das Hannoveraner Firmenpanel (Gerlach, Hübler und Meyer 2003) sind hierfür kein Ersatz, sondern eine Ergänzung mit Stärken und Schwächen in anderen Bereichen (vgl. Gerlach und Wagner 1997b).

Die Betriebspaneldaten der amtlichen Statistik stellen damit einen unverzichtbaren Bestandteil des Datenbestandes dar, den die empirische Wirtschaftsforschung für ihre Arbeiten in der Forschung, der wissenschaftlichen Politikberatung und der Ausbildung des wissenschaftlichen Nachwuchses benötigt. Sie in umfassendem Ausmaß so einfach und unbürokratisch wie möglich der Wissenschaft zur Verfügung zu stellen ist eine wichtige Aufgabe. Ich hoffe, dieser Vortrag trägt ein gaaaanz klein wenig dazu bei, auf dem Weg zu diesem Ziel voran zu kommen.

Literaturhinweise

Bernard, Andrew B. und Wagner, Joachim (1997): Exports and Success in German Manufacturing, in: Weltwirtschaftliches Archiv 133, pp. 134 – 157.

Bernard, Andrew B. und Wagner, Joachim (2001): Export Entry and Exit by German Firms, *in*: Weltwirtschaftliches Archiv 137, pp. 105 – 123.

Franz, Wolfgang (1995): Theoretische Ansätze zur Erklärung der Arbeitslosigkeit: Wo stehen wir 1995?, in: Universität Konstanz, Fakultät für Wirtschaftswissenschaften und Statistik, Forschungsschwerpunkt „Internationale Arbeitsmarktforschung", Diskussionspapier 27.

Gerlach, Knut; Hübler, Olaf und Meyer, Wolfgang (2003): The Hannover Firm Panel (HFP), in: Schmollers Jahrbuch – Journal of Applied Social Science Studies (forthcoming).

Gerlach, Knut und Wagner, Joachim (1994): Regional Differences in Small Firm Entry in Manufacturing Industries: Lower Saxony, 1979 – 1991, in: Entrepreneurship & Regional Development 6, pp. 63 – 80.

Gerlach, Knut und Wagner, Joachim (1995): Die Heterogenität der Arbeitsplatzdynamik innerhalb der Industrie – Zum Verhältnis von Belegschafts- und Betriebsfluktuation im Verarbeitenden Gewerbe Niedersachsens (1978 – 1990), in: Klaus Semlinger und Bernd Frick (Hrsg.), Betriebliche Modernisierung in personeller Erneuerung, Berlin: Edition Sigma, S. 39 – 57.

Gerlach, Knut und Wagner, Joachim (1997a): Analysen zur Nachfrageseite des Arbeitsmarktes mit Betriebspaneldaten aus Erhebungen der amtlichen Industriestatistik - Ein Überblick über Ansätze und Ergebnisse für niedersächsische Industriebetriebe, in: Jürgen Kühl, Manfred Lahner und Joachim Wagner (Hrsg.), Die Nachfrageseite des Arbeitsmarktes – Ergebnisse aus Analysen mit deutschen Firmenpaneldaten (Beiträge zur Arbeitsmarkt- und Berufsforschung BeitrAB 204), Nürnberg: Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit 1997, S. 11 – 82.

Gerlach, Knut und Wagner, Joachim (1997b): Paneldaten für niedersächsische Industriebetriebe aus der amtlichen Statistik und aus dem Hannoveraner Firmenpanel, in: Reinhard Hujer, Ulrich Rendtel und Gert Wagner (Hrsg.), Wirtschafts- und Sozialwissenschaftliche Panel-Studien, Sonderheft zum Allgemeinen Statistischen Archiv, Heft 30, Göttingen: Vandenhoeck & Ruprecht, S. 211 – 227.

Heckman, James J. (2001): Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture, in: Journal of Political Economy 109, pp. 673 – 748.

Kölling, Arnd (2000): The IAB-Establishment Panel, in: Schmollers Jahrbuch – Journal of Applied Social Science Studies 120, pp. 291 – 300.

Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (2001): Wege zu einer besseren informationellen Infrastruktur, Baden-Baden: Nomos.

Schasse, Ulrich und Wagner, Joachim (Hrsg., 1999): Entwicklung von Arbeitsplätzen, Exporten und Produktivität im interregionalen Vergleich – Empirische Untersuchungen mit Betriebspaneldaten, in: Beiträge zum Workshop FiDASt '99, NIW-Vortragsreihe, Band 13, Hannover: Niedersächsisches Institut für Wirtschaftsforschung.

Schasse, Ulrich und Wagner, Joachim (Hrsg., 2001): Regionale Wirtschaftsanalysen mit Betriebspaneldaten – Ansätze und Ergebnisse, in: Beiträge zum Workshop FiDASt 2001, NIW-Vortragsreihe, Band 14, Hannover: Niedersächsisches Institut für Wirtschaftsforschung.

Wagner, Joachim (1992): Firm Size, Firm Growth, and Persistence of Chance: Testing GIBRAT's Law With Establishment Data From Lower Saxony, 1978 – 1989, in: Small Business Economics 4, pp. 125 – 131.

Wagner, Joachim (1993): Firm Size, Firm Growth, and Export Performance: Evidence From Longitudinal Data for German Establishments, in: Jahrbücher für Nationalökonomie und Statistik 211, p. 417 – 420.

Wagner, Joachim (1994a): The Post-Entry Performance of New Small Firms in German Manufacturing Industries, in: Journal of Industrial Economics XLII, pp. 141 – 152.

Wagner, Joachim (1994b): Small Firm Entry in Manufacturing Industries: Lower Saxony, 1979 – 1989, in: Small Business Economics 6, pp. 211 – 223.

Wagner, Joachim (1995a): Exports, Firm Size, and Firm Dynamics, in: Small Business Economics 7, pp. 29 – 39.

Wagner, Joachim (1995b): Firm Size and Job Creation in Germany, in: Small Business Economics 7, pp. 469 – 474.

Wagner, Joachim (1996): Firm Size, Firm Age, and Job Duration, in: Review of Industrial Organization, 11 (1996), pp. 201 – 210.

Wagner, Joachim (1999a): Nutzung von betrieblichen Einzeldaten aus der amtlichen Statistik durch externe Wissenschaftler – Modelle, Erfahrungen, Perspektiven, in: Statistisches Bundesamt (Hrsg.), Möglichkeiten einer wissenschaftlichen Nutzung von Unternehmensdaten aus der amtlichen Statistik, Spektrum Bundesstatistik, Band 14, Stuttgart: Metzler Poeschel, S. 9 – 17.

Wagner, Joachim (1999b): The Life History of Cohorts of Exits from German Manufacturing, in: Small Business Economics 13, pp. 71 – 79.

Wagner, Joachim (2000): Firm Panel Data from German Official Statistics, in: Schmollers Jahrbuch –Journal of Applied Social Science Studies 120, pp. 143 – 150.

Wagner, Joachim (2002): The causal effects of exports on firm size and labor productivity: First evidence from a matching approach, in: Economics Letters 77, pp. 287 – 292.

Wagner, Joachim (2003): Unobserved firm heterogeneity and the size – exports nexus: Evidence from German panel data, in: *Weltwirtschaftliches Archiv* (forthcoming).

Wissenschaftsrat (2002): Empfehlungen zur Stärkung wirtschaftswissenschaftlicher Forschung an den Hochschulen, in: Drucksache 5455–02, Saarbrücken, 15. November 2002.

Harald Strotmann *)

Zu den Erwartungen der Datennutzer an das Anonymisierungsprojekt

Koreferat zum Beitrag „Arbeiten mit Einzeldaten der amtlichen Statistik am Beispiel des Monatsberichts im Verarbeitenden Gewerbe"

Das Referat von Joachim Wagner hat in einem kurzweiligen Überblick aufgezeigt, für welch vielfältige Analysezwecke vor allem er selbst, aber auch einige andere Wissenschaftler die Originaleinzeldaten aus den Monatsberichten im Produzierenden Gewerbe in Kooperationsprojekten mit der amtlichen Statistik bisher bereits verwendet haben. Sowohl die von Joachim Wagner aufgezeigten inhaltlichen Fragestellungen und Ergebnisse als auch die Tatsache, dass viele dieser Arbeiten in referierten internationalen und nationalen Zeitschriften publiziert wurden, unterstreichen, dass die Mikrodaten aus den Monatsberichten für die Wissenschaft eine sehr wertvolle Datenbasis darstellen. Insofern ist zu erwarten, dass von Seiten möglicher Datennutzer Interesse an faktisch anonymisierten Daten aus der amtlichen Industriestatistik und gerade auch an Einzeldaten aus den Monatsberichten bestehen dürfte.

Im folgenden Koreferat möchte ich zunächst ebenfalls das Analysepotenzial des Datensatzes anhand eines Beispiels veranschaulichen. Auf der Grundlage dieses Beispiels, das die besondere Bedeutung der Bereitstellung der Daten als Paneldaten und die Möglichkeit einer Verknüpfung der Monatsberichtsdaten mit den Daten aus der industriellen Kleinbetriebserhebung betont, werde ich mich dann der Frage zuwenden, was die bisherigen Erfahrungen der Datennutzer für das Anonymisierungsprojekt bedeuten.

1 Vertiefendes Beispiel einer empirischen Analyse mit den Monatsberichten – Überleben und Sterben von Gründungen

Joachim Wagner hat bei seinem Überblick über verschiedene inhaltliche Fragestellungen bereits die Themen „Überleben und Sterben von Neugründungen" sowie „Beschäftigungsbeitrag von Gründungen" angeführt. Da man anhand dieses Themenkomplexes zwei grundlegende Aspekte, die mir für das Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" sehr wichtig erscheinen, besonders gut verdeutlichen kann, möchte ich dieses Beispiel für Baden-Württemberg einleitend kurz etwas vertiefen.

Die folgenden Ergebnisse basieren auf einem Kooperationsprojekt zwischen dem Statistischen Landesamt Baden-Württemberg und der Universität Hohenheim, in dessen Rahmen mir die Möglichkeit geboten wurde, unter Wahrung des Statistikgeheimnisses vor Ort im Landesamt mit den amtlichen Industriedaten zu arbeiten. Praktiziert wurde somit

*) Dr. Harald Strotmann, Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen.

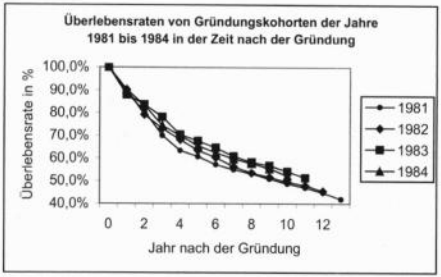
in Baden-Württemberg anders als in Niedersachsen nicht die Lösung über eine Schalterstelle mit Diskettentransfer, sondern die „Mitarbeiterlösung ohne Bezahlung".[1)]

Datengrundlage ist, wie bereits von Joachim Wagner für Niedersachsen beschrieben, ein Betriebspaneldatensatz, der auf den Monatsberichten und der industriellen Kleinbetriebserhebung basiert und somit alle Industriebetriebe in Baden-Württemberg erfasst. Durch die Panelstruktur des Datensatzes können neu im Datensatz auftretende Betriebe („Gründungen") und wegfallende Betriebe („Schließungen") von Jahr zu Jahr identifiziert werden. Um verschiedene Unschärfen bei der Identifikation von tatsächlichen Neugründungen und Schließungen zu reduzieren, wurden Korrekturen durchgeführt und Annahmen getroffen, die jedoch hier nicht weiter diskutiert werden sollen.

Für die folgenden Kohortenanalysen wurden die Gründungskohorten der Jahre 1981 bis 1984 betrachtet und deren Entwicklung bis einschließlich 1994 im Zeitablauf verfolgt. Die Jahre ab 1995 wurden nicht in die Analysen einbezogen, da es hier aufgrund der Änderung der Wirtschaftszweigsystematik von der SYPRO auf die WZ 93 zu erheblichen Änderungen in der Zusammensetzung des Berichtskreises der Erhebungen kam. Insgesamt standen 2 605 Gründungen für die Analysen der „post-entry-performance" zur Verfügung.

Die Ergebnisse der Kohortenanalysen, die erst durch die Panelstruktur der Daten möglich werden, verdeutlichen, dass die Sterblichkeit von Gründungen in den Jahren unmittelbar nach der Gründung sehr hoch ist (Abbildung 1): Nach einem Jahr sind bereits etwa 20% der Neugründungen wieder aus dem Markt ausgeschieden, nach 5 Jahren bereits rund 40 % und nach 10 Jahren nahezu jede zweite Gründung. Dieser „Drehtüreffekt" zeigt, dass es einem erheblichen Teil der Gründungen nicht gelingt, sich am Markt zu etablieren.

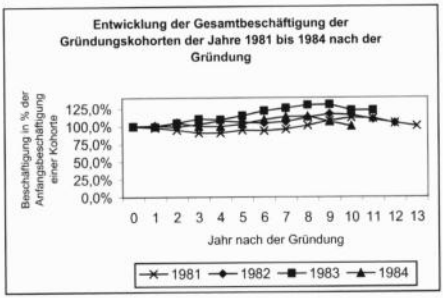
Abbildung 1



1) Die Idee für die Kooperation in Baden-Württemberg entstand während des Statistischen Kolloquiums im November 1996 im Statistischen Bundesamt, bei dem Joachim Wagner über sein „Pilotprojekt" in Niedersachsen berichtete. Für eine ausführliche Darstellung der Organisation der Kooperation in Baden-Württemberg vgl. Strotmann (1999).

Ein Blick auf die Entwicklung der Gesamtbeschäftigung der Gründungskohorten verdeutlicht jedoch, dass diese sich trotz der hohen Sterblichkeit der Gründungen über die Zeit hinweg relativ stabil entwickelt (Abbildung 2). Diese Stabilität der Gesamtbeschäftigung, die von weiteren Studien und auch in den Arbeiten von Joachim Wagner bestätigt wird, ist das Ergebnis eines erheblich überdurchschnittlichen Beschäftigungswachstums der überlebenden Gründungsbetriebe.

Abbildung 2



Analysen möglicher Bestimmungsfaktoren der „post-entry-performance" von Betrieben stellen vor dem Hintergrund dieser empirischen Fakten entweder das Überleben und Sterben von Gründungen oder aber Determinanten des Wachstums von Gründungen in der Folgezeit der Gründung in den Mittelpunkt ihrer Analysen. An dieser Stelle soll und kann nicht detailliert auf inhaltliche Ergebnisse derartiger Analysen eingegangen werden.[2] Hier sollte anhand des Beispiels zweierlei aufgezeigt werden: einerseits der besondere Wert der Verknüpfung jährlicher Monatsberichtsdaten zu Paneldaten, andererseits, dass bei vielen bisherigen Studien gerade auch das Zuspielen der Daten aus der industriellen Kleinbetriebserhebung von zentraler Bedeutung ist. Beides hat Auswirkungen auf das Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten", auf die in Kapitel 3 kurz eingegangen werden soll.

2) Für verweildaueranalytische Analysen der Determinanten des Überlebens und Sterbens von Neugründungen vgl. z.B. Strotmann (2002a, b); für eine Untersuchung der Einflussgrößen des Beschäftigungswachstum vgl. Strotmann (2002b) oder Wagner (1994).

3 Was bedeuten die Erfahrungen aus Sicht der Datennutzer für das Anonymisierungsprojekt?

3.1 Notwendigkeit der faktischen Anonymisierung von *Panel*daten

Das angeführte Beispiel unterstreicht zunächst die Tatsache, dass die Daten aus den Monatsberichten insbesondere dann von wissenschaftlichem Interesse sind, wenn sie zu einem Paneldatensatz verknüpft werden. Auch für fast alle der von Joachim Wagner angeführten Forschungsthemen ist die Existenz eines Panels Grundvoraussetzung für die inhaltlichen Analysen. Dies gilt nicht nur im Hinblick auf mögliche multivariate Analysen, sondern auch bereits für einfache deskriptive Auswertungen wie z.B. die Komponentenanalysen zur Arbeitsplatzdynamik (vgl. dazu z.B. Gerlach/Wagner 1997 oder Strotmann 2002a).

Sicherlich sind auch bloße Querschnittsinformationen aus einer einzigen Welle in anonymisierter Form für einen Teil der Nutzer interessant. Letztlich zeigt der Beitrag von Joachim Wagner jedoch sehr deutlich, dass es aus Sicht der Nutzer von wesentlicher Bedeutung ist, bei der Anonymisierung von amtlichen Industriedaten neben der Querschnittsdimension auch die Längsschnittdimension einzubeziehen. Für eine nutzerorientierte faktische Anonymisierung der Daten aus den Monatsberichten muss es daher das langfristige Ziel sein, die Daten aus den Monatsberichten als Paneldatensätze zu anonymisieren.

Für diesen Analyseschritt müssen dann alle Fragen, die gegenwärtig im Projekt für den Querschnitt behandelt werden und in verschiedenen Beiträgen auf dieser Tagung bereits diskutiert wurden, neu gestellt und beantwortet werden.

Bei der Frage, was die Anonymisierung eines Paneldatensatzes für einen möglichen Datenangriff und das Re-Identifikationsrisiko bedeuten kann, ist aus mehreren Gründen zu erwarten, dass das Re-Identifikationsrisiko steigt. Einerseits steigt es bereits dadurch, dass derselbe Betrieb mehrfach im Datensatz enthalten ist. Andererseits stellen Veränderungen von Variablen über die Zeit neue Zusatzinformationen dar, die eine Re-Identifikation erleichtern könnten. Darüber hinaus ermöglichen Ereignisse im Zeitablauf, wie z.B. eine Zusammenlegung von Betrieben, eine Abspaltung oder eine Produktionsstillegung, die im Querschnitt nicht identifiziert werden können, eventuell eine leichtere Re-Identifikation. Zu überprüfen ist, ob es im Gegenzug auch Argumente geben kann, die für einen Rückgang des Re-Identifikationsrisikos sprechen.

Zu bedenken ist auf jeden Fall, dass zumindest der Nutzen einer Re-Identifikation für Zeitpunkte in der weiteren Vergangenheit erheblich geringer sein dürfte als am aktuellen Rand, was bei einem Kosten-Nutzen-Kalkül zur Beurteilung der faktischen Anonymität berücksichtigt werden müsste.

Systematisch zu überprüfen ist zudem, welche Implikationen die Verwendung von Paneldaten für die Wahl der zu verwendenden Anonymisierungsverfahren einerseits und den Erhalt des Analysepotenzials andererseits hat.

Für den Erhalt des Analysepotenzials kommt zu dem Ziel, im Querschnitt die Momente und Produktmomente von Verteilungen möglichst gut abzubilden, das zusätzliche Ziel, einzelbetriebliche Veränderungen im Zeitablauf möglichst gut wiedergeben zu können.

Mit Blick auf die Anonymisierungsverfahren ist daher zu klären, ob und in welchem Maße sich die Ergebnisse einer Bewertung der Anonymisierungsverfahren im Querschnitt und die verwendeten Bewertungskriterien auch auf Panelanalysen übertragen lassen. Auch und gerade die Auswirkungen verschiedener Anonymisierungsverfahren auf die Eigenschaften von Schätzverfahren, wie sie zum Beispiel von Lechner und Pohlmeier (2003) und Rosemann (2003) auf dieser Tagung für Querschnittsdaten vorgestellt und diskutiert wurden, müssen dann auch für panelökonometrische Schätzverfahren und deren Eigenschaften soweit möglich theoretisch und empirisch evaluiert werden.

Zu begrüßen ist, dass im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" grundsätzlich geplant ist, Daten aus den Monatsberichten der Jahre 2000 bis 2002 als Paneldatensatz zu anonymisieren. Anhand von drei Wellen aus dem Panel lassen sich dann einige Fragestellungen bereits untersuchen; zahlreiche inhaltliche Studien, wie auch die oben angeführten Analysen des Überlebens und Sterbens von Gründungen, setzen jedoch die Verfügbarkeit längerer Zeiträume voraus. Ob und in welchem Maße sich die Ergebnisse zur Anonymisierung von drei Wellen auch auf mehrere Wellen übertragen lassen, bleibt dann auch noch zu klären.

Klar ist jedoch auch, dass es zunächst für die Anonymisierung von Querschnittsdaten noch eine ganze Menge offener Fragen zu beantworten gilt, bevor man sich der Anonymisierung von Paneldaten zuwenden kann.

3.2 Einbeziehung der industriellen Kleinbetriebserhebung?

Ein zweites Problem, das ebenfalls an obigem Beispiel veranschaulicht werden kann und aus Nutzersicht für die Entwicklung von Scientific-Use-Files mit Unternehmensdaten bedeutend sein könnte, betrifft die Einbeziehung der industriellen Kleinbetriebserhebung.

Die Mehrzahl der Forschungsarbeiten, die Joachim Wagner vorgestellt hat, basiert auf einer Verknüpfung der Monatsberichte im Produzierenden Gewerbe mit der industriellen Kleinbetriebserhebung. Für viele Fragestellungen werden sinnvolle Analysen auch erst durch die Verknüpfung dieser beiden Datengrundlagen möglich. So ist es gerade für Analysen des Gründungs- und Schließungsgeschehens mit Hilfe der amtlichen Industriedaten von wesentlicher Bedeutung, auch Informationen über Kleinbetriebe zu haben. Die Probleme bei der Identifikation originärer Neugründungen mit den Daten sind ohnehin bereits beträchtlich, aber durch bestimmte Korrekturen und Annahmen zumindest einigermaßen „heilbar". Solange man jedoch ausschließlich auf Betriebe mit 20 oder mehr Beschäftigten zurückgreifen kann, lassen sich Gründungsanalysen nicht mehr sinnvoll durchführen, da der kleinbetriebliche Bereich überhaupt nicht abgedeckt werden kann und originäre Gründungen in den meisten Fällen weniger als 20 Beschäftigte aufweisen.[3)]

Allerdings ist zu bedenken, dass die industrielle Kleinbetriebserhebung ohnehin mit dem Berichtsjahr 2002 eingestellt wird, so dass eine Verwendung anonymisierter Daten hier nur noch für die Vergangenheit möglich wäre. Über die Anonymisierung der industriellen Kleinbetriebserhebung und deren Verknüpfung mit den Monatsberichten für die

3) Die Kleinbetriebserhebung beschränkt sich auf die Erhebung der industriellen Kleinbetriebe, so dass kleine Handwerksbetriebe ohnehin nicht in der Kleinbetriebserhebung enthalten sind.

Vergangenheit bis 2002 sollte dennoch auf jeden Fall diskutiert werden. Panelanalysen gehen immer in die Vergangenheit zurück und für multivariate Zwecke ist es häufig nicht entscheidend, ob der Datensatz ganz bis an den aktuellen Rand reicht. Darüber hinaus dürfte die Anonymisierung der industriellen Kleinbetriebserhebung relativ einfach möglich sein, da es sich per Definition um kleine Betriebe handelt (vgl. dazu z.B. Brand 2000 oder Vorgrimler 2003).

Würde letztlich nur ein anonymisierter Paneldatensatz aus den Monatsberichten zur Verfügung gestellt, so wäre dieser aus Nutzersicht für einige Fragestellungen verwendbar, jedoch könnten andere Themenfelder, die bisher mit den Daten in verschiedenen Studien bearbeitet wurden, zukünftig nicht mehr Gegenstand von Forschungsarbeiten sein.[4)]

4 Schlussbemerkungen

Auch andere Datensätze aus dem Bereich der Wirtschaftsstatistik werden gerade dadurch für die Wissenschaft interessant, dass sie zu einem Panel verknüpft werden. Fritsch/Stephan (2003) arbeiten z.B. zurzeit an einem Projekt, bei dem die Kostenstrukturerhebung im Verarbeitenden Gewerbe zu einem Paneldatensatz verknüpft wurde. Hier wäre dann auch zu klären, ob und in welchem Maße sich die Ergebnisse bezüglich der Angriffsszenarien und der Bewertung der verwendeten Anonymisierungsverfahren vor dem Hintergrund des Ziels einer größtmöglichen Erhaltung des Analysepotenzials, die für die Monatsberichte gewonnen werden können, auch auf andere (Panel-)Datensätze übertragen lassen.

Da man jedoch den zweiten oder dritten Schritt nicht vor dem ersten machen kann, muss es in den weiteren Projektschritten zunächst weiter darum gehen, die Ergebnisse für die Anonymisierung von Querschnittsdaten zu verfeinern und abzusichern. Allerdings sollte man schon im Hinterkopf haben, dass die wissenschaftlichen Nutzer der Daten und somit die potenziellen Kunden der faktisch anonymisierten Scientific-Use-Files insbesondere auch an analysefähigen Paneldaten interessiert sind.

4) Abzuwarten bleibt, inwiefern in Zukunft Angaben aus dem kleinbetrieblichen Bereich eventuell aus anderen Quellen verfügbar sein könnten.

Literaturhinweise

Brand, Ruth (2000): Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos, in: Beiträge zur Arbeitsmarkt- und Berufsforschung, Nr. 237.

Fritsch, Michael und Stephan, Andreas (2003): The Distribution of Inefficiency Within Industries – an Empirical Analysis, Beitrag auf dem FiDASt-Workshop am 06./07. März 2003 in Berlin, mimeo.

Gerlach, Knut und Wagner, Joachim (1997): Analysen zur Nachfrageseite des Arbeitsmarktes mit Betriebspaneldaten aus der amtlichen Industriestatistik. Ein Überblick über Ansätze und Ergebnisse für niedersächsische Industriebetriebe, in: Kühl, J./Lahner, M./Wagner, J. (Hrsg.), Die Nachfrageseite des Arbeitsmarktes – Ergebnisse aus Analysen mit deutschen Firmenpaneldaten, Beiträge zur Arbeitsmarkt- und Berufsforschung des IAB, 204, S. 11 – 82.

Lechner, Sandra und Pohlmeier, Winfried (2003): Anonymisierungsmethoden und ökonometrische Modelle, in: Forum der Bundesstatistik (in diesem Band S. 115 ff.), Wiesbaden.

Rosemann, Martin (2003): Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik, in: Forum der Bundesstatistik (in diesem Band S. 154 ff.), Wiesbaden.

Strotmann, Harald (1999): Zur wissenschaftlichen Nutzung von Betriebsdaten aus der amtlichen Statistik, in: Spektrum Bundesstatistik, Band 14, Möglichkeiten einer wissenschaftlichen Nutzung von Unternehmensdaten aus der amtlichen Statistik, S. 29 – 44.

Strotmann, Harald (2002a): Arbeitsplatzdynamik in der baden-württembergischen Industrie – eine Analyse mit amtlichen Betriebspaneldaten, Hohenheimer Volkswirtschaftliche Schriften, Band 39, Peter Lang: Frankfurt.

Strotmann, Harald (2002b): Determinanten des Überlebens von Neugründungen und Schließungen in der baden-württembergischen Industrie – eine empirische Analyse mit amtlichen Betriebsdaten, in: IAW-Diskussionspapiere, Nr. 6, Tübingen.

Vorgrimler, Daniel (2003): Re-Identifikationsrisiken und Re-Identifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios, in: Forum der Bundesstatistik (in diesem Band S. 40 ff.), Wiesbaden.

Wagner, Joachim (1994): The Post-Entry Performance of New Small Firms in Manufacturing Industries, in: Journal of Industrial Economics, 42(2), S. 141 – 154.

Martin Rosemann *)

Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik

Einführung

Gegenstand des vorliegenden Beitrags ist es, erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten aus dem Bereich der Unternehmens- und Betriebsdaten vorzustellen. Mit Hilfe dieser Ergebnisse werden erste Schlussfolgerungen gezogen, inwiefern verschiedene im Rahmen des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten"[1] erprobte Verfahren zur Anonymisierung von Einzeldaten das Analysepotenzial der betrachteten Datensätze verändern. Darauf aufbauend kann entschieden werden, welche Verfahren oder Verfahrensgruppen nach dem Kriterium der geringsten Verminderung des Analysepotenzials für eine tiefergehende Betrachtung im weiteren Verlauf des Projekts in Frage kommen und künftig für die faktische Anonymisierung wirtschaftsstatistischer Einzeldaten angewendet werden können.

Zunächst wird in Kapitel 1 der Versuch unternommen, zu klären, was unter dem Begriff Analysepotenzial zu verstehen ist und wie die Verringerung des Analysepotenzials operationalisiert werden kann. In Kapitel 2 wird ein Überblick über die im Rahmen dieses Beitrags betrachteten Datensätze, die Kostenstrukturerhebung und die Umsatzsteuerstatistik, gegeben.[2] In Kapitel 3 werden die Auswirkungen traditioneller und datenverändernder Anonymisierungsverfahren auf die Umsatzsteuerstatistik einer näheren Untersuchung unterzogen. Dabei werden zunächst in Abschnitt 3.1 die angewendeten Anonymisierungsverfahren kurz erläutert. Abschnitt 3.2 widmet sich der Veränderung wichtiger Charakteristika der Verteilung. In Abschnitt 3.3 werden die Ergebnisse deskriptiver Analysen mit der Umsatzsteuerstatistik bei anonymisierten und Originaldaten verglichen. Abschnitt 3.4 zieht ein Zwischenfazit für die Umsatzsteuerstatistik. Die Untersuchung der Auswirkungen von datenverändernden Anonymisierungsverfahren auf das Analysepotenzial der Kostenstrukturerhebung wird in Kapitel 4 vorgenommen. Dabei werden in Abschnitt 4.1 die verwendeten Anonymisierungsverfahren vorgestellt. Abschnitt 4.2 untersucht die Veränderungen wichtiger Charakteristika der Verteilung. In Abschnitt 4.3 werden die Ergebnisse deskriptiver Analysen mit der Kostenstrukturerhebung bei anonymisierten und Originaldaten verglichen. In Abschnitt 4.4 wird untersucht, wie sich die Ergebnisse ökonometrischer Schätzungen bei Anwendung verschiedener Anonymisierungsverfahren verändern und inwieweit sich die Ergebnisse für verschiedene

*) Martin Rosemann, Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen.

1) Das Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" wird von den statistischen Ämtern getragen und vom Bundesministerium für Bildung und Forschung mitfinanziert. Das IAW Tübingen ist an diesem Projekt als Unterauftragnehmer des Statistischen Bundesamts beteiligt.

2) Die Gründe, warum diese Datensätze im Projekt zuerst bearbeitet werden, werden ebenfalls in Kapitel 2 erläutert.

Verfahren unterscheiden. Abschnitt 4.5 zieht ein Zwischenfazit für die Kostenstrukturerhebung. In Kapitel 5 werden die wesentlichen Ergebnisse der Untersuchungen der Veränderung des Analysepotenzials zusammengefasst, erste Schlussfolgerungen für die Verwendung von Anonymisierungsverfahren aus Sicht des Analysepotenzials gezogen und ein abschließender Ausblick auf weitere Forschungsarbeiten gegeben.

1 Operationalisierung des Analysepotenzials

Ziel ist es, Anonymisierungsverfahren danach zu beurteilen, wie stark sie das Analysepotenzial eines Datensatzes einschränken. Hierzu ist es notwendig, das Analysepotenzial zu operationalisieren und möglichst objektive Kriterien für seine Veränderung aufzustellen.

Als besonders problematisch erweist sich zunächst, dass auch das Analysepotenzial eines Originaldatensatzes nicht eindeutig und objektiv fassbar ist. Die Untersuchungsziele sind ebenso vielseitig wie die Methoden und Verfahren, mit denen sie erreicht werden sollen. Und letztlich hängt das Analysepotenzial vor allem von Art und Beschaffenheit der zugrundeliegenden Statistik ab.

Dennoch lässt sich leicht nachvollziehen, dass unabhängig von der konkreten Fragestellung bzw. dem angewendeten Verfahren bestimmte Verteilungseigenschaften für die betrachteten Merkmale auch nach der Anwendung von Anonymisierungsverfahren wenigstens annähernd erhalten bleiben sollten. Schließlich ist die gesamte in statistischen Auswertungen genutzte Information eines Datensatzes in der multivariaten Verteilung der erhobenen Variablen enthalten (vgl. Brand et al. 1999, S. 158). Erhaltenswerte Eigenschaften der multivariaten Verteilung sind insbesondere:

- Mittelwerte und Streuungsmaße der univariaten Verteilungen,
- Kovarianzen und Korrelationen zwischen den Variablen bzw. die Rangkorrelationen oder andere robuste Zusammenhangsmaße, insbesondere bei schiefen Verteilungen.

Genannt werden können auch die dritten und vierten Momente, wobei man davon ausgehen muss, dass bei der Einbeziehung von zu vielen Zielen und Kriterien Kompromisse zwischen unterschiedlichen Zielen gemacht werden müssen oder aber eine Prioritätensetzung erfolgen muss. So ist beispielsweise die Anwendung der meisten Standardverfahren, wie beispielsweise einer Regressionsschätzung, daran gebunden, dass die Mittelwerte und die Varianz-Kovarianz-Matrix im anonymisierten Datensatz näherungsweise erhalten bleiben (vgl. Brand, Bender, Kohaut 1999).

Als Kriterium für die Einschränkung des Analysepotenzials durch verschiedene Anonymisierungsverfahren dient dann die Abweichung der Kennzahl (des Charakteristikums) des anonymisierten Datensatzes von der Kennzahl des Originaldatensatzes. Angelehnt an Sebé et al. (2002) sowie Dandekar, Domingo-Ferrer und Sebé (2002) werden die mittleren Fehler

- der Mediane,
- der arithmetischen Mittel,
- der Varianzen,

- der Kovarianzen,
- der Korrelationskoeffizienten (Bravais-Pearson) und
- der Rang-Korrelationskoeffzienten (Spearman)

verwendet. Diese Kriterien werden unabhängig von der betrachteten Statistik, der interessierenden Fragestellung und der anzuwenden Analysemethode zur Beurteilung der Verringerung des Analysepotenzials herangezogen.[3] Im Unterschied zu Sebé et al. (2002) sowie Dandekar, Domingo-Ferrer und Sebé (2002) werden die Maße in diesem Beitrag allerdings nicht zu einer einzigen Kennzahl verdichtet. Vielmehr werden die Kriterien einzeln betrachtet.

Folgende Kriterien werden im Rahmen dieser Arbeit untersucht (vgl. Höhne 2002):

(Bei den mit * gekennzeichneten Größen handelt es sich jeweils um die anonymisierten Werte, bei den nicht gekennzeichneten Größen um die Originalwerte.)

a) $$\frac{\sum_{j=1}^{d} \frac{\left|\overline{x_j} - \overline{x_j^*}\right|}{\left|\overline{x_j}\right|}}{d}$$ Mittlerer relativer Fehler der arithmetischen Mittel

b) $$\frac{\sum_{j=1}^{d} \sum_{1 \le i < j} \frac{\left|\mathrm{cov}_{ij} - \mathrm{cov}_{ij}^*\right|}{\left|\mathrm{cov}_{ij}\right|}}{\frac{1}{2} d(d-1)}$$ Mittlerer relativer Fehler der Kovarianzen

c) $$\frac{\sum_{j=1}^{d} \frac{\left|\mathrm{var}_{jj} - \mathrm{var}_{jj}^*\right|}{\left|\mathrm{var}_{jj}\right|}}{d}$$ Mittlerer relativer Fehler der Varianzen

d) $$\frac{\sum_{j=1}^{d} \sum_{1 \le i < j} \left|r_{ij} - r_{ij}^*\right|}{\frac{1}{2} d(d-1)}$$ Mittlerer absoluter Fehler der Korrelationskoeffizienten

e) $$\frac{\sum_{j=1}^{d} \sum_{1 \le i < j} \left|s_{ij} - s_{ij}^*\right|}{\frac{1}{2} d(d-1)}$$ Mittlerer absoluter Fehler der Rangkorrelationen

3) Wenig überzeugend ist die Betrachtung der Abweichung der Einzelwerte, wie sie in Sebé et al. (2002) sowie Dandekar, Domingo-Ferrer und Sebé (2002) ebenfalls vorgenommen wird. Zum einen ist es gerade die Kernidee datenverändernder Anonymisierungsverfahren, dass die Werte voneinander abweichen, zum anderen ist mit der Abweichung der einzelnen Werte noch keine Aussage über die Verteilungseigenschaften und damit über das Analysepotenzial verbunden.

Eine abschließende Bewertung über die Einschränkung des Analysepotenzials kann dennoch erst im Zusammenhang mit den Auswirkungen von Anonymisierungsmaßnahmen auf unterschiedliche Arten von Analysen vorgenommen werden. Im Weiteren werden daher verschiedene Arten von Analysen durchgeführt, um die Effekte von Anonymisierungsverfahren auf das Anylsepotenzial zu untersuchen. Dabei werden zum einen beispielhafte deskriptive Auswertungen vorgenommen. Von besonderer Bedeutung ist die deskriptive Auswertung von Teilmassen, wie z.B. die Berechnung bestimmter Durchschnittsgrößen nach Wirtschaftszweigen und/oder Beschäftigtengrößenklassen sowie ihrer Rangzahlen. Zum anderen werden die Auswirkungen der Anonymisierungsverfahren auf die Koeffizienten verschiedener ökonometrischer Modelle sowie auf die entsprechenden Test-Statistiken untersucht. Vorgegangen wird also in drei Schritten:

1. Vergleich der arithmetischen Mittel, der Korrelationsstruktur, der Rangkorrelationen, der Kovarianzen und der Varianzen; Kriterium: Abweichungen dieser Größen.
2. Vergleich deskriptiver Kennzahlen von Teilmassen (z.B. FuE-Intensitäten nach Wirtschaftszweigen; FuE-Intensitäten nach Größenklassen, FuE-Intensitäten nach Wirtschaftszweigen und Größenklassen); Kriterien: Veränderung der Mittelwerte, Veränderung der Rangfolgen zwischen den Teilmassen.
3. Durchführung ökonometrischer Schätzungen (lineare und nichtlineare Modelle); Kriterium: Veränderung der Koeffizienten und der Teststatistiken.

Bei der Untersuchung der Veränderung von Koeffizienten im Rahmen ökonometrischer Modelle muss insbesondere analysiert werden,

- inwiefern sich die statistische Signifikanz eines Einflussfaktors verändert,
- inwiefern sich bei gegebener statistischer Signifikanz das Vorzeichen eines Koeffizienten verändert,
- ob die Veränderung der Werte von Koeffizienten statistisch signifikant ist.

Erst all dies gemeinsam liefert die Grundlage für die Entscheidung darüber, ob ein Anonymisierungsverfahren vor dem Hintergrund der Zielsetzung der weitgehenden Erhaltung des Analysepotenzials geeignet ist.

2 Zugrundeliegende Datensätze: Umsatzsteuerstatistik und Kostenstrukturerhebung

Im Rahmen des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" wird für sechs verschiedene Statistiken untersucht, ob und gegebenenfalls mit welchen Verfahren sie faktisch anonymisiert werden können.

Für die ersten Untersuchungen wurden zunächst die Umsatzsteuerstatistik und die Kostenstrukturerhebung (KSE) ausgewählt, weil

- die Umsatzsteuerstatistik aufgrund einer sehr hohen Zahl an Untersuchungseinheiten (2,9 Millionen Unternehmen) sowie einer geringen Zahl an Variablen tendenziell eher leichter zu anonymisieren sein dürfte und der Einsatz traditioneller Anonymisierungsverfahren hier zunächst am ehesten ausreichend erscheint (vgl. Vorgrimler 2002).

- die Kostenstrukturhebung auf der einen Seite aufgrund einer vergleichsweise geringen Zahl von Untersuchungseinheiten (16 918), einer großen Zahl von Variablen sowie eines großen vermuteten Zusatzwissens nur schwer zu anonymisieren ist und daher vermutlich in jedem Fall der Einsatz datenverändernder Verfahren notwendig sein dürfte. Auf der anderen Seite dürfte sich die Kostenstrukturerhebung eines recht großen Interesses von Seiten der Nutzer erfreuen.

Damit werden gleich zu Beginn des Projektes zwei Datensätze bearbeitet, die sich sowohl in ihren Nutzungsmöglichkeiten als auch in ihren Charakteristika und damit in den notwendigen Anonymisierungsverfahren deutlich unterscheiden. In den Tabellen 2.1 und 2.2 sind die zur Verfügung stehenden Variablen beider Statistiken aufgeführt.

Die Umsatzsteuerstatistik ist eine Sekundärstatistik, die auf die Daten zurückgreift, die bei der Finanzverwaltung im Rahmen des Umsatzsteuer-Voranmeldungs- und -Vorauszahlungsverfahrens anfallen. Erfasst werden alle Unternehmen mit einem Jahresumsatz (ohne Umsatzsteuer) von über 32 500 DM (16 617 Euro), die Umsatzsteuer-Voranmeldungen abgeben. Seit 1996 wird die Umsatzsteuerstatistik jährlich durchgeführt[4] (Statistisches Bundesamt 2002).

Tabelle 2.1: Variablen der Umsatzsteuerstatistik

1. Regionalbezug (BBR-Schlüssel, sog. „Neuner-Kategorie")
2. Wirtschaftszweig (WZ93)
3. Dauer der Steuerpflicht (typisiert)
4. Organschaft nach § 2 Abs. 2 Nr. 2 UstG (0= nein, 1 = ja)
5. Rechtsform
Jahreswerte in Euro
6. Lieferungen und Leistungen
7. Steuerpflichtige Lieferungen und Leistungen
8. Zu 16 %
9. Zu 7 %
10. Steuerfreie Lieferungen und Leistungen
11. Mit Vorsteuerabzug
12. Innergemeinschaftliche Lieferungen und Leistungen
13. Ohne Vorsteuerabzug
14. Umsatzsteuer vor Abzug der Vorsteuer
15. Für Lieferungen und Leistungen
16. Für innergemeinschaftliche Erwerbe
17. Abziehbare Vorsteuer
18. Für Lieferungen und Leistungen
19. Aus Rechnungen anderer Unternehmen
20. Einfuhrumsatzsteuer
21. Für innergemeinschaftliche Erwerbe
22. Vorauszahlungssoll
23. Nachrichtlich: innergemeinschaftliche Erwerbe
Vorjahreswerte in Euro
24. Lieferungen und Leistungen
25. Vorauszahlungssoll

4) Für die im Rahmen des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" durchgeführten Untersuchungen liegen Daten der Umsatzsteuerstatistik für das Jahr 2000 vor.

Die Kostenstrukturerhebung im Verarbeitenden Gewerbe erfasst als hochrechnungsfähige Stichprobe maximal 18 000 Unternehmen mit 20 Beschäftigten und mehr; die Befragung erfolgt zentral durch das Statistische Bundesamt im Wege der Selbstausfüllung durch die Unternehmen. Die in der Stichprobe gewonnenen Ergebnisse werden auf die Gesamtheit der Unternehmen mit 20 Beschäftigten und mehr hochgerechnet. Diese Stichprobe wird i.d.R. alle vier Jahre neu gezogen, so dass kleinere und mittlere Unternehmen durch Rotation entlastet werden können. Unternehmen mit 500 Beschäftigten und mehr, aber auch Unternehmen in Wirtschaftszweigen mit geringer Besetzungszahl, werden zur Sicherstellung der Qualität der Ergebnisse vollständig einbezogen. Den Ergebnissen für das Berichtsjahr 1999 liegt eine neue Stichprobenauswahl zugrunde. In dieser Stichprobe werden rd. 43 % der Unternehmen des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden mit 20 Beschäftigten und mehr erfasst. Die Konstruktion des Stichprobenplans garantiert, dass diese Unternehmen zu 76 % zur Gesamtzahl der tätigen Personen und zu 84 % zum Gesamtumsatz im Berichtskreis beitragen (Statistisches Bundesamt 2002).

Tabelle 2.2: Variablen der Kostenstrukturerhebung

1. Wirtschaftszweig (WZ 93)
2. Regionalbezug (BBR-Schlüssel, sog. „Neuner-Kategorie")
3. Beschäftigtengrößenklasse
4. Tätige Inhaber
5. Angestellte und Arbeiter
6. Teilzeitbeschäftigte
7. Teilzeitbeschäftigte umgerechnet in Vollzeiteinheiten
8. Tätige Personen insgesamt
9. Umsatz aus eigenen Erzeugnissen
10. Umsatz aus Handelsware
11. Gesamtumsatz (entspricht nicht der Summe aus 9. und 10.)
12. Anfangsbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen am Umsatz aus eigenen Erzeugnissen
13. Endbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen am Umsatz aus eigenen Erzeugnissen
14. Bestandveränderung an unfertigen/fertigen Erzeugnissen
15. Gesamtleistung/Bruttoproduktionswert
16. Anfangsbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und Betriebsstoffen gemessen am Umsatz aus eigenen Erzeugnissen
17. Endbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und Betriebsstoffen gemessen am Umsatz aus eigenen Erzeugnissen
18. Verbrauch an Rohstoffen
19. Energieverbrauch
20. Anfangsbestand an Handelsware gemessen am Umsatz aus Handelsware
21. Endbestand an Handelsware gemessen am Umsatz aus Handelsware
22. Einsatz an Handelsware
23. Bruttogehalts- und -lohnsumme
24. Gesetzliche Sozialkosten
25. Sonstige Sozialkosten
26. Kosten für Leiharbeitnehmer
27. Kosten für Lohnarbeiten
28. Kosten für Reparaturen
29. Mieten und Pachten
30. Sonstige Kosten
31. Fremdkapitalzinsen
32. Kosten insgesamt
33. Bruttowertschöpfung zu Faktorkosten
34. Nettowertschöpfung zu Faktorkosten
35. Gesamtaufwendungen für innerbetriebliche Forschung und Entwicklung
36. Anzahl der für Forschung und Entwicklung eingesetzten Lohn- und Gehaltsempfänger

3 Untersuchungen für die Umsatzsteuerstatistik

3.1 Angewendete Anonymisierungsverfahren

Für die Prüfung der Veränderung des Analysepotenzials durch Anonymisierungsmaßnahmen im Rahmen dieses Beitrags stehen sowohl mit so genannten traditionellen Verfahren (probe)anonymisierte Datensätze als auch mit datenverändernden Verfahren be-

arbeitete Datensätze zur Verfügung.[5] Bei traditionellen Verfahren werden im Unterschied zu den sogenannten datenverändernden Verfahren keine generellen Veränderungen der einzelnen Werte vorgenommen (vgl. Brand 2000; Höhne 2003 sowie Ronning et al. 2002).

3.1.1 Auf die Umsatzsteuerstatistik angewendete traditionelle Verfahren [6]

Anonymisiert werden nur die Überschneidungsmerkmale Umsatz (Lieferungen und Leistungen), Regionaltyp, Wirtschaftszweig und Rechtsform sowie solche Merkmale, die mit dem Umsatz hoch korreliert sind.[7]

Im Einzelnen werden folgende Maßnahmen vorgenommen:

- Die Wirtschaftszweige werden bis zu einem Umsatz von 500 Millionen Euro als Zweisteller (WZ 93) ausgewiesen, ab einem Umsatz von 500 Millionen Euro lediglich als Einsteller.
- Die Wirtschaftszweigklassifikationen werden so zusammengefasst, dass jeweils mindestens 3 500 Einheiten in einer Klasse vertreten sind.
- Die Rechtsform wird so umkodiert (vergröbert), dass nur noch vier verschiedene Kategorien ausgewiesen werden: Kategorie 1 (Personengesellschaften) umfasst Einzelunternehmen, OHG und KG. Kategorie 2 (Kapitalgesellschaften) umfasst AG und GmbH. Kategorie 3 umfasst Erwerbs- und Wirtschaftsgenossenschaften und Körperschaften des öffentlichen Rechts. Kategorie 4 umfasst die sonstigen Rechtsformen.
- Für das Merkmal Umsatz wird für Unternehmen mit einem Umsatz von mindestens 500 Millionen Euro das Replacement-Verfahren angewendet. D.h. alle Umsätze, die zwischen 500 Millionen und 1 Mrd. Euro liegen, werden durch das arithmetische Mittel dieser Umsatzwerte ersetzt. Das gleiche wird für Umsätze über 1 Mrd. Euro durchgeführt.
- Bei Unternehmen mit einem Umsatz von weniger als 500 Mill. Euro werden Rundungen vorgenommen. Umsatzwerte zwischen 50 und 500 Mill. Euro werden auf die ersten beiden Stellen gerundet, Umsätze von weniger als 50 Mill. Euro auf die erste Stelle.

Außerdem werden verschieden große Stichproben gezogen. Variante 1 stellt wie der Originaldatensatz die volle Erhebung dar (Trad1). Variante 2 ist eine 80 %-Stichprobe (Trad1 80 %), Variante 3 eine 50 %-Stichprobe (Trad1 50 %).

---

5) Die Anonymisierung wurde im Rahmen des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" durch das Statistische Bundesamt und das Statistische Landesamt Berlin vorgenommen. Das Vorgehen, zunächst mit so genannten traditionellen Verfahren zu beginnen, wurde vereinbart, da bei der Umsatzsteuerstatistik aufgrund einer hohen Zahl von Merkmalsträgern und einer vergleichsweise geringen Zahl von Variablen, am ehesten die Hoffnung bestand, mit traditionellen Verfahren einen ausreichenden Schutz vor Re-Identifikationen sicherzustellen.

6) Für einen Überblick über die traditionellen Anonymisierungsverfahren siehe insbesondere Müller et al. (1991) sowie Ronning et al. (2002). Einen Überblick über die datenverändernden Verfahren geben Höhne (2003), Ronning et al. (2002); Brand (2000) sowie Gottschalk (2002).

7) Vgl. hierzu Vorgrimler (2002).

3.1.2 Auf die Umsatzsteuerstatistik angewendete datenverändernde Verfahren

Bei der Anonymisierung der Umsatzsteuerstatistik wird das Verfahren SAFE des Statistischen Landesamts Berlin in zwei Varianten angewendet (vgl. Evers, Höhne 1999 und Höhne 2003). Dabei wird zuerst eine Lösung gesucht, die unter ausschließlicher Betrachtung der diskreten Fälle keine Einzel- oder Zweierfälle mehr aufweist. Dieser Lösung werden die originalen Sätze zugeordnet (mit größter Ähnlichkeit und Veränderungen bei möglichst kleinen stetigen Merkmalen). Anschließend werden die stetigen Merkmale nach zwei Varianten anonymisiert.

- SAFE1: Trippelbildung innerhalb der Gruppen der diskreten Merkmale (gleiche Ausprägungen bei den diskreten Merkmalen). Innerhalb der identischen Gruppen werden die Unternehmen absteigend nach einem ausgewählten dominierenden Merkmal sortiert und anschließend für jeweils drei benachbarte Werte alle stetigen Merkmalswerte durch das arithmetische Mittel ersetzt.
- SAFE2: Anonymisierung der stetigen Werte mit eindimensionaler Mikroaggregation (für jedes Merkmal getrennt). Die Gruppengröße weist eine Mindestgröße von drei Elementen sowie eine minimale Schwankungsbreite von 7 % auf.

3.2 Vergleich wesentlicher Charakteristika der Verteilungen

Die Ergebnisse in Tabelle 3.1. zeigen, dass die beiden Varianten des SAFE zu keiner Veränderung der arithmetischen Mittel führen. Die Anwendung der traditionellen Verfahren ohne zusätzliche Stichprobenziehung führt nur zu einer relativ geringen Veränderung der arithmetischen Mittel um 0,2 %. Durch die Stichprobenziehung ergibt sich dann eine zusätzliche größere Veränderung. Bei der Veränderung der Varianzen schneidet Verfahren SAFE2 mit einer Veränderung um 8,6 % mit Abstand am besten ab. Alle anderen Verfahren haben Veränderungen von mindestens 30 % zur Folge. Bei den Kovarianzen schneiden alle Verfahren mit eine Veränderung von ca. 60 % ungefähr gleich ab. Die geringsten Veränderungen der Korrelationskoeffizienten verursacht Verfahren SAFE1. Die anderen Verfahren liegen bei der Veränderung der Korrelationskoeffizienten etwa gleichauf. Bei den Rangkorrelationen schneidet Verfahren SAFE1 mit Abstand am schlechtesten ab. Alle anderen Verfahren haben durchschnittliche absolute Veränderungen von etwa 0,005. Bei den traditionellen Verfahren kann noch festgehalten werden, dass die Zusatzmaßnahme der Stichprobenziehung die arithmetischen Mittel deutlich stärker verändert als die Varianzen, Kovarianzen, Korrelationskoeffizienten und Rangkorrelationen.

Insgesamt lässt sich aus diesen Veränderungen der Verteilungscharakteristika noch keine eindeutige Antwort geben, wie stark welches Anonymisierungsverfahren das Analysepotenzial einschränkt.

Tabelle 3.1: Veränderung von Verteilungscharakteristika durch die Anonymisierung der Umsatzsteuerstatistik

Verfahren	Mittlerer relativer Fehler			Mittlerer absoluter Fehler	
	Arithmetische Mittel	Varianzen	Kovarianzen	Korrelationen	Rangkorrelationen
	in %			(x 100)	
SAFE1	0,0	56,4	57,4	3,5	8,5
SAFE2	0,0	8,6	56,6	10,8	0,5
Trad1	0,2	33,9	60,3	15,2	0,5
Trad1 80 %	2,5	31,5	60,1	15,0	0,5
Trad1 50 %	3,5	37,1	64,9	16,2	0,5

Quelle: IAW-Berechnungen

3.3 Vergleich deskriptiver Auswertungen

Für die Umsatzsteuerstatistik werden folgende deskriptive Auswertungen vorgenommen:

- Berechnung der Anteile der steuerpflichtigen Lieferungen und Leistungen zu 7 % und zu 16 % an allen Lieferungen und Leistungen sowie an allen steuerpflichtigen Lieferungen und Leistungen nach Wirtschaftszweigen. Damit kann analysiert werden, für welche Wirtschaftszweige der ermäßigte Mehrwertsteuersatz von 7 % von größerer und für welche er von geringerer Bedeutung ist.
- Berechnung der Anteile der innergemeinschaftlichen Lieferungen und Leistungen und der sonstigen steuerfreien Lieferungen und Leistungen mit Vorsteuerabzug (Exporte außerhalb der EU) an den Lieferungen und Leistungen. Damit kann analysiert werden, welche Wirtschaftszweige stärker und welche weniger stark exportabhängig sind.
- Berechnung der Anteile der abziehbaren Vorsteuer aus Einfuhrumsatzsteuer sowie für innergemeinschaftliche Lieferungen und Leistungen an der gesamten abziehbaren Vorsteuer. Damit kann analysiert werden, welche Wirtschaftszweige stärker und welche weniger stark importabhängig sind.

In Tabelle 3.2 sind beispielhaft für die Untersuchung der Bedeutung des ermäßigten Mehrwertsteuersatzes die Veränderungen der arithmetischen Mittel nach WZ- Zweistellern (auf Basis der im Rahmen der traditionellen Anonymisierung teilweise zusammengefassten Wirtschaftszweige) sowie der Ränge der Wirtschaftszweige nach arithmetischen Mitteln dargestellt, die sich durch die Anwendung der verschiedenen Anonymisierungsverfahren ergeben.

Tabelle 3.2: Veränderung der arithmetischen Mittel sowie der Ränge der Wirtschaftszweige nach arithmetischen Mitteln durch die Anonymisierung der Umsatzsteuerstatistik bei der Untersuchung der Bedeutung des ermäßigten Mehrwertsteuersatzes

Verfahren	Durchschnittliche relative Veränderung der arithmetischen Mittel in %		Durchschnittliche absolute Veränderung der Ränge	
	Umsatzanteile zu 7 % am Gesamtumsatz	Umsatzanteile zu 7 % am steuerpflichtigen Umsatz	Umsatzanteile zu 7 % am Gesamtumsatz	Umsatzanteile zu 7 % am steuerpflichtigen Umsatz
SAFE1	0,0	0,0	0,0	0,0
SAFE2	1,2	1,2	0,0	0,0
Trad1	14,3	10,9	0,6	0,7
Trad1, 80 %	19,8	15,2	1,0	1,0
Trad1, 50 %	30,4	26,7	1,6	1,4

Quelle: IAW-Berechnungen

Es ist zu erkennen, dass die beiden Varianten des SAFE-Verfahrens, insbesondere SAFE1, die Ergebnisse kaum verändern und von den getesteten Verfahren am besten abschneiden. Die traditionellen Verfahren schneiden hingegen deutlich schlechter ab. Es wird deutlich, dass sowohl die zunächst durchgeführten traditionellen Verfahren als auch die anschließende Stichprobenziehung eine erhebliche Einschränkung des Analysepotenzials bedeuten.

In Tabelle 3.3 sind die Ergebnisse für alle oben aufgeführten Auswertungen, die mit der Umsatzsteuerstatistik vorgenommen wurden, verdichtet dargestellt. Ausgewiesen ist jeweils die geringste und die größte Abweichung von den Originalergebnissen, die sich bei allen Auswertungen beobachten lässt. Dabei werden die für das in Tabelle 3.2 aufgeführte Beispiel festgehaltenen Ergebnisse bestätigt. Deutlich wird, dass mit SAFE1 bei den deskriptiven Auswertungen der Umsatzsteuerstatistik bessere Ergebnisse erzielt werden als mit SAFE2. Die traditionellen Verfahren schneiden zwar bei der Veränderung der arithmetischen Mittel selbst schlechter ab als das Verfahren SAFE2, nicht jedoch bei der Veränderung der Ränge.

Tabelle 3.3: Minimale und maximale Veränderung der arithmetischen Mittel sowie der Ränge der Wirtschaftszweige nach arithmetischen Mitteln durch die Anonymisierung der Umsatzsteuerstatistik für alle Auswertungen

Verfahren	Intervall der durchschnittlichen Veränderungen der arithmetischen Mittel nach WZ-Zweistellern in %	Intervall der durchschnittlichen absoluten Veränderung der Ränge der Zweisteller nach arithmetischen Mitteln
SAFE1	[0,0 ; 0,0]	[0,0 ; 0,0]
SAFE2	[0,4 ; 2,5]	[0,0 ; 3,5]
Trad1	[0,7 ; 16,2]	[0,2 ; 2,5]
Trad1, 80 %	[1,1 ; 20,9]	[0,7 ; 2,8]
Trad1, 50 %	[1,6 ; 30,4]	[1,4 ; 3,2]

Quelle: IAW-Berechnungen

3.4 Ein Zwischenfazit für die Umsatzsteuerstatistik

Das aussichtsreichste Verfahren bei der Anonymisierung der Umsatzsteuerstatistik scheint nach den ersten durchgeführten ersten Auswertungen das Verfahren SAFE des Statistischen Landesamts Berlin in der Variante SAFE1 zu sein. Die angewendeten traditionellen Verfahren führen zu einer größeren Einschränkung des Analysepotenzials als die SAFE-Verfahren. Dies gilt nicht für alle berechneten Charakteristika der Verteilung, allerdings werden Mittelwerte und Korrelationskoeffizienten deutlich stärker verändert. Bei der deskriptiven Analyse ist die Abweichung der Ergebnisse zwischen den beiden Verfahrenstypen bei den arithmetischen Mitteln innerhalb der Wirtschaftszweige stärker ausgeprägt als bei der Veränderung der Ränge. Dennoch ist das Ergebnis bei allen durchgeführten Auswertungen anzutreffen. Unklar bleibt, ob die traditionellen Verfahren vor der Stichprobenziehung eine größere Einschränkung des Analysepotenzials mit sich bringen als die anschließende Stichprobenziehung.

4 Untersuchungen für die Kostenstrukturerhebung

4.1 Angewendete Anonymisierungsverfahren

Bei der Probe-Anonymisierung der Kostenstrukturerhebung werden bisher nur datenverändernde Verfahren angewendet. Es werden folgende Verfahrensgruppen[8] zur Anonymisierung getestet und in Bezug auf die Verringerung des Analysepotenzials verglichen:

- Mikroaggregation (MA),
- SAFE (Verfahren des Statistischen Landesamtes Berlin),
- Rank-Swapping (RSWP) und
- Latin Hypercube Sampling (LHS).

8) Zu den Verfahren im Einzelnen vgl. Höhne (2003) sowie Ronning et al. (2002).

a) Zu den angewendeten Varianten der Mikroaggregation (MA)

Die Gruppengröße beträgt mindestens drei Einheiten. Die Originalwerte werden durch die arithmetischen Mittel der Gruppe ersetzt.

Folgende Varianten werden getestet:

- Alle stetigen Variablen werden gemeinsam betrachtet. D.h., die Gruppen werden nach der kleinsten euklidischen Distanz aus allen stetigen Merkmalen gebildet (MA1g).
- Es wird eine Blockung in Variablen für Handelstätigkeit einerseits und andere Variablen andererseits vorgenommen. Die Gruppenbildung im Rahmen der Mikroaggregation erfolgt für die beiden Blöcke getrennt (MA2g).
- Jede der insgesamt 33 stetigen Variablen wird getrennt mikroaggregiert (MA33g).

b) Zu den angewendeten Varianten des SAFE

Beim Verfahren SAFE wird zunächst eine Lösung gesucht, die unter ausschließlicher Betrachtung der diskreten Fälle keine Einzel- oder Zweierfälle mehr aufweist. Es wird gleichzeitig gewährleistet, dass die Fehler bei allen daraus aggregierbaren Häufigkeitsverteilungen einen minimalen Maximalfehler haben und die Anzahl aller Objekte erhalten bleibt. Dieser Lösung werden die originalen Sätze zugeordnet (mit größter Ähnlichkeit und Veränderungen bei möglichst kleinen stetigen Merkmalen). (vgl. Höhne 2003)

Anschließend werden die stetigen Merkmale in zwei Varianten anonymisiert:

- SAFE1: Trippelbildung innerhalb der Gruppen der diskreten Merkmale (gleiche Ausprägungen bei den diskreten Merkmalen): Innerhalb der identischen Gruppen werden die Unternehmen absteigend nach einem ausgewählten dominierenden Merkmal (hier: tätige Personen) sortiert. Anschließend werden für jeweils drei benachbarte Werte alle stetigen Merkmalswerte durch das arithmetische Mittel ersetzt.
- SAFE2: Anonymisierung der stetigen Werte mit eindimensionaler Mikroaggregation analog zu MA33g. Die Gruppengröße weist eine Mindestgröße von drei Elementen sowie eine minimale Schwankungsbreite von 7 % auf.

Bei den Mikroaggregationsverfahren sowie bei SAFE2 werden Felder, die sich als Summe aus anderen Feldern berechnen lassen bzw. unmittelbar und logisch aus anderen Feldern hervorgehen (z.B. Beschäftigtengrößenklasse) bei der Anonymisierung zunächst vernachlässigt und im Anschluss an die Anonymisierung der anderen Felder aus diesen neu berechnet.[9)]

c) Zu den angewendeten Varianten des Rank-Swapping (RSWP)

Alle Variablen (stetige und diskrete) werden getrennt voneinander bearbeitet. Bei jeder Variable wird ein Tausch von Merkmalswerten zwischen jeweils zwei Merkmalsträgern in einem vorab festgelegten Nachbarschaftsbereich vorgenommen. Die Nachbarschaftsbereiche betragen 10 % (RSWP 10p), 5 % (RSWP 5p) und 1 % (RSWP 1p) der Anzahl aller

9) So ergibt sich die neue Beschäftigtengrößenklasse aus dem anonymisierten Wert für die tätigen Personen insgesamt.

Untersuchungseinheiten. (Zum Beispiel würden bei n = 1 000 Beobachtungen und einem Nachbarschaftsbereich von 1 % innerhalb der nächstgelegenen 10 Beobachtungen getauscht).

d) Zur angewendeten Variante des Latin Hypercube Sampling (LHS)

Latin Hypercube Sampling (LHS) ist ein Simulationsverfahren, das von R. Dandekar für die Anonymisierung von Einzeldaten vorgeschlagen wurde (Dandekar 1993; Dandekar, Domingo-Ferrer, Sebé 2002; Dandekar, Cohen, Kirkendall 2002). Das Verfahren erlaubt es, (beliebige) univariate Verteilungen zu simulieren. Darüber hinaus werden die Rangkorrelationen erhalten (vgl. Höhne 2003 und Ronning et al. 2002). Für die hier vorgenommenen Untersuchungen werden die Unternehmen der KSE in vier Gruppen unterteilt. Es wird danach unterschieden, ob Handel[10] und/oder Forschung und Entwicklung betrieben wird[11]. Für diese vier Gruppen werden die stetigen Variablen mit LHS anonymisiert. Aus den anonymisierten Teildatensätzen werden anschließend mithilfe einer stark vereinfachten Variante der in Dandekar, Domingo-Ferrer, Sebé (2002) vorgeschlagenen Methode LHS Hybrid Daten gebildet, indem die anonymisierten Teildatensätze anhand der Ausprägung eines einzigen Merkmals, nämlich der Bruttogehalt- und Lohnsumme, den Ursprungsdatensätzen und damit den diskreten Variablen zugeordnet werden.[12] An dieser Stelle muss deutlich darauf hingewiesen werden, dass die nach diesem Verfahren (im folgenden als LHS1 bezeichnet) gebildeten LHS Hybrid Daten eigentlich keine Auswertungen nach Ausprägungen der diskreten Variablen (sog. Teilmassenbetrachtungen) zulassen. Dies machen die in Abschnitt 4.3 und 4.4 zusammengestellten Untersuchungsergebnisse deutlich, bei denen die Daten ausschließlich auf Teilmassenebene, bzw. unter Einbeziehung der Ausprägungen der diskreten Variablen ausgewertet werden.[13]

Anzumerken ist noch, dass bei der Anwendung von LHS1 die Zahl Merkmalsträger erhöht werden kann. Schließlich werden die Werte der stetigen Variablen synthetisch erzeugt. Bei der durchgeführten Variante des LHS wird die Zahl der Unternehmen von 16 918 auf 17 100 erhöht. Es kommt also vor, dass einer originalen Kombination von diskreten Merkmalen zwei unterschiedliche synthetische Sätze an stetigen Merkmalswerten zugeordnet werden.

---

10) Entscheidendes Kriterium hierfür ist, dass die Variable „Einsatz an Handelsware" größer Null ist.

11) Herangezogen wird die Variable „Anzahl der für Forschung und Entwicklung eingesetzten Lohn- und Gehaltsempfänger".

12) Dandekar, Domingo-Ferrer, Sebé (2002) schlagen vor, diese Paarbildung nicht anhand eines, sondern anhand aller im LHS Datensatz vorhandenen Merkmale vorzunehmen.

13) Falls es zu den Vorgaben für das Anonymisierungsverfahren gehört, dass solche Teilmassenbetrachtungen angestellt werden können sollen, wird in Dandekar, Cohen, Kirkendall (2002) vorgeschlagen, das LHS Verfahren getrennt auf diese Teilmassen anzuwenden und, wo dies wegen zu geringer Gruppenbesetzungen nicht möglich ist, LHS auf einen Datensatz anzuwenden, der die betreffenden kategorialen Merkmale bereits enthält. Allerdings ist es nur schwer vorstellbar, für unterschiedliche Teilmassenauswertungen unterschiedlich anonymisierte Scientific-Use-Files zur Verfügung zu stellen. Dennoch sind Verbesserungen gegenüber der Variante LHS1 in dieser Hinsicht möglich.

4.2 Vergleich wesentlicher Charakteristika der Verteilungen

Tabelle 4.1: Veränderung von Verteilungscharakteristika durch die Anonymisierung der Kostenstrukturerhebung

Verfahren	Mittlerer relativer Fehler			Mittlerer absoluter Fehler	
	Arithmetisches Mittel	Varianzen	Varianz-Kovarianzmatrix	Korrelationen	Rangkorrelationen
	in %		in %[1]	(x 100)	
MA1g	3,5	21,3	75,8	5,8	9,0
MA2g	2,5	23,4	62,0	4,8	6,8
MA33g	0,0	5,9	21,2	2,4	0,0
SAFE1	2,8	46,9	88,6	4,4	6,6
SAFE2	0,0	7,3	96,4	3,8	0,5
RSWP 10p	0,0	0,0	131,9	35,4	1,6
RSWP 5p	0,0	0,0	130,8	34,5	0,5
RSWP 1p	0,0	0,0	147,6	31,6	0,1
LHS1	1,0	0,6	219,6	36,2	0,8

1) Im Unterschied zu den Abschnitten 1.2 und 3.2 wird hier die Abweichung der gesamten Varianz-Kovarianzmatrix berechnet.

Quelle: Berechnungen des Statistischen Landesamts Berlin und des IAW

Tabelle 4.1 zeigt die Veränderung der wesentlichen Verteilungscharakteristika durch die Anonymisierungsverfahren. Es ist zu erkennen, dass durch das Rank-Swapping- Verfahren die ersten und zweiten Momente nicht verändert werden, dies ist verfahrensimmanent. Allerdings werden die Varianz-Kovarianzmatrix und die Korrelationskoeffizienten deutlich verändert. Ein ähnliches Bild ergibt sich für das Simulationsverfahren LHS1. Hier werden die Momente der univariaten Verteilung kaum verändert[14], die Kovarianzen und die Korrelationsstruktur hingegen deutlich. Die Mikroaggregationsverfahren MA1g und MA2g sowie das Verfahren SAFE1 führen zu den höchsten Fehlern bei den Varianzen. Dies liegt daran, dass durch die Durchschnittsbildung innerhalb der Gruppen bei Mikroaggregation und SAFE automatisch ein Teil der Variation verloren geht. Bei den Fehlern der Kovarianzen liegen sie im mittleren Bereich, die Korrelationskoeffizienten verringern sich nur geringfügig.[15] Die geringsten Abweichungen bei den Kovarianzen und Korrelationskoeffizienten weist das Mikroaggregationsverfahren MA33g auf. Auch bei den Varianzen und arithmetischen Mitteln ist die Abweichung gering. Das Verfahren SAFE2 schließlich weist ebenso wie MA33g geringe Abweichungen bei den Korrelationskoeffizienten, arithmetischen Mitteln und Varianzen auf, allerdings eine vergleichsweise hohe Abweichung der Kovarianzen.

14) Die Veränderungen entstehen lediglich durch die zusätzlich geschaffenen künstlichen Merkmalsträger.

15) Die Abweichungen der arithmetischen Mittel bei den Mikroaggregationsverfahren ergeben sich dadurch, dass die Lagerbestände im Ursprungsdatensatz als Anteile an Umsatzgrößen ausgewiesen sind. Vor der Anonymisierung wird auf die absoluten Werte zurückgerechnet. Diese werden anonymisiert und anschließend neu auf die ebenfalls anonymisierten Umsatzwerte bezogen.

Es zeigt sich, dass die Rank-Swapping-Verfahren und das Latin Hypercube Sampling zwar die univariaten Verteilungen erhalten, die Zusammenhänge zwischen den Variablen aber nur sehr unzureichend abbilden, während die Mikroaggregationsverfahren und SAFE zwar zu stärkeren Veränderungen bei den univariaten Verteilungen führen, die Zusammenhänge aber weniger zerstören.

Bei den Rangkorrelationen ergibt sich nochmals ein völlig verändertes Bild. Hier weist nach dem Verfahren MA33g das Verfahren RSWP 1p die geringste Veränderung auf. Es folgen RSWP 5p, SAFE2 und LHS1. Die größten Abweichungen bei den Rangkorrelationen weisen MA1g und MA2g auf. Ebenso wie bei der Umsatzsteuerstatistik verursacht auch bei der KSE das Verfahren SAFE1 eine vergleichsweise hohe Abweichung bei den Rangkorrelationen.

Diese unterschiedlichen Ergebnisse machen nochmals deutlich, dass es sinnvoll ist, die Maßzahlen, die sich aus der Veränderung von Verteilungscharakteristika ergeben, nicht zu einer einzigen Kennzahl zu verdichten, sondern die Auswirkung der Verfahren auf konkrete Analysen zu untersuchen, um ein geschlossenes Gesamtbild zu erhalten.

4.3 Vergleich deskriptiver Auswertungen

Im Rahmen deskriptiver Auswertungen mit der Kostenstrukturerhebung werden die FuE-Beschäftigungs- und Ausgabenintensitäten nach Wirtschaftszweigen und Beschäftigtengrößenklassen untersucht.[16] Betrachtet werden jeweils die Veränderung der Kennzahlen selbst sowie die Veränderungen in der Rangstruktur.[17]

Die FuE-Beschäftigungsintensitäten werden als Anteil der für Forschung und Entwicklung eingesetzten Beschäftigten an der Anzahl der insgesamt im Unternehmen tätigen Personen bestimmt. Die FuE-Ausgabenintensitäten werden als Anteil der Ausgaben für Forschung und Entwicklung am Gesamtumsatz berechnet.

Die durchschnittlichen Intensitäten werden jeweils für WZ-Viersteller, WZ-Zweisteller und Beschäftigtengrößenklassen sowie für die möglichen Kombinationen aus WZ-Vier- und Zweistellern und den verschiedenen Beschäftigtengrößenklassen berechnet.

Beispielhaft ist in den Tabellen 4.2 bis 4.4 zunächst dargestellt, wie sich die arithmetischen Mittel bzw. die Ränge durch die Anonymisierungsverfahren bei einer Untersuchung der FuE-Intensitäten aller Unternehmen nach WZ-Zweistellern, WZ-Vierstellern sowie WZ-Zweistellern und Beschäftigtengrößenklassen verändern.

Es ist zu erkennen, dass das Verfahren der Mikroaggregation, bei dem alle Variablen getrennt voneinander bearbeitet werden (MA33g), mit Abstand am besten abschneidet. Gefolgt wird es von den beiden Varianten des SAFE, wobei SAFE2 in der Regel besser abschneidet als SAFE1. Die einzige Ausnahme stellt die Untersuchung nach WZ-Vierstellern dar. Die schlechtesten Ergebnisse liefern die Rank-Swapping-Verfahren, insbesondere RSWP 10p und RSWP 5p. MA2g, MA1g und die Hybrid Daten (LHS1) liegen im Mittelfeld.

16) Vergleichbare Untersuchungen wurden auch für die Kostenstruktur (Anteile verschiedener Kostenpositionen am Gesamtumsatz) der Unternehmen durchgeführt. Die Auswirkungen der Anonymisierungsmaßnahmen sind allerdings ähnlich.

17) Die Berechnungen wurden am Institut für Angewandte Wirtschaftsforschung in Tübingen durchgeführt. Vorarbeiten und wertvolle Hinweise stammen von Professor Dr. Joachim Wagner (Universität Lüneburg).

Es zeigt sich also, dass die Tatsache, dass die Rankswapping-Verfahren die univariaten Verteilungen der Ursprungsvariablen erhalten, selbst bei einfachen deskriptiven Auswertungen nicht ausreicht, um ähnliche Ergebnisse wie mit den Originaldaten zu erreichen. Dies liegt daran, dass bei der Berechnung der Intensitäten der Quotient aus zwei Variablen gebildet wird, die bei der Anwendung des Rank-Swappings völlig unabhängig voneinander getauscht werden. Außerdem werden beim Rank-Swapping zusätzlich auch die Ausprägungen der diskreten Variablen getauscht, damit auch die Wirtschaftszweige.

Tabelle 4.2: Veränderung der arithmetischen Mittel und der Ränge bei der Untersuchung der FuE-Beschäftigungs- und FuE-Ausgabenintensitäten nach WZ-Zweistellern

Verfahren	Durchschnittliche relative Veränderung der arithmetischen Mittel in %		Durchschnittliche absolute Veränderung der Ränge	
	FuE-Beschäftigungsintensitäten	FuE-Ausgabenintensitäten	FuE-Beschäftigungsintensitäten	FuE-Ausgabenintensitäten
MA1g	117,4	82,3	1,7	1,9
MA2g	129,9	107,6	0,6	1,1
MA33g	0,2	0,2	0,0	0,0
SAFE1	18,9	21,0	0,8	0,9
SAFE2	12,3	15,6	0,9	0,5
RSWP 10p	747,6	974,4	5,2	4,6
RSWP 5p	310,7	478,6	3,1	4,2
RSWP 1p	97,1	96,4	2,2	2,5
LHS1	70,0	156,6	2,2	2,8

Quelle: IAW-Berechnungen

Tabelle 4.3: Veränderung der arithmetischen Mittel und der Ränge bei der Untersuchung der FuE-Beschäftigungs- und FuE-Ausgabenintensitäten nach WZ-Vierstellern

Verfahren	Durchschnittliche relative Veränderung der arithmetischen Mittel in %		Durchschnittliche absolute Veränderung der Ränge	
	FuE-Beschäftigungsintensitäten	FuE-Ausgabenintensitäten	FuE-Beschäftigungsintensitäten	FuE-Ausgabenintensitäten
MA1g	203,4	222,2	32,7	30,0
MA2g	269,6	282,6	21,4	24,0
MA33g	0,2	0,2	0,5	0,3
SAFE1	56,0	87,0	21,2	21,8
SAFE2	119,4	138,8	24,2	25,8
RSWP 10p	7037,2	6348,8	69,1	66,7
RSWP 5p	953,8	1100,9	56,2	56,2
RSWP 1p	285,6	354,6	46,9	43,2
LHS1	222,9	597,6	32,9	39,0

Quelle: IAW-Berechnungen

Tabelle 4.4: Veränderung der arithmetischen Mittel und der Ränge bei der Untersuchung der FuE-Beschäftigungs- und Ausgabenintensitäten nach WZ-Zweistellern und Beschäftigtengrößenklassen

Verfahren	Durchschnittliche relative Veränderung der arithmetischen Mittel in %		Durchschnittliche absolute Veränderung der Ränge	
	FuE-Beschäftigungsintensitäten	FuE-Ausgabenintensitäten	FuE-Beschäftigungsintensitäten	FuE-Ausgabenintensitäten
MA1g	89,7	87,7	13,5	13,9
MA2g	95,7	143,0	12,8	12,8
MA33g	0,2	0,2	0,7	0,7
SAFE1	58,1	82,7	8,8	8,7
SAFE2	10,4	9,3	8,5	8,0
RSWP 10p	934,7	929,1	35,1	32,4
RSWP 5p	583,0	678,7	25,9	23,7
RSWP 1p	69,7	103,8	13,1	12,8
LHS1	225,6	407,4	17,4	18,9

Quelle: IAW-Berechnungen

Tabelle 4.5: Bandbreiten der Veränderungen der arithmetischen Mittel sowie der Mediane der FuE-Beschäftigungs- und Ausgabenintensitäten und ihrer Ränge bei zweidimensionalen Analysen (Wirtschaftszweige oder Beschäftigtengrößenklassen)

Verfahren	Intervall der durchschnittlichen Veränderungen der arithm. Mittel		Intervall der durchschnittlichen Veränderungen der Mediane	
	Veränderung der Werte in %	Veränderung der Ränge (normiert)	Veränderung der Werte in %	Veränderung der Ränge (normiert)
MA1g	[3,6 ; 41,9]	[0,0 ; 0,1]	[3,1 ; 61,1]	[0,0 ; 0,2]
MA2g	[1,9 ; 25,2]	[0,0 ; 0,1]	[1,0 ; 36,0]	[0,0 ; 0,1]
MA33g	[0,0 ; 0,2]	[0,0 ; 0,1]	[0,0 ; 0,4]	[0,0 ; 0,1]
RSWP 10p	[34,3 ; 123,0]	[0,3 ; 0,5]	[3,2 ; 86,1]	[0,0 ; 0,4]
RSWP 5p	[10,7 ; 92,5]	[0,0 ; 0,5]	[2,1 ; 83,4]	[0,1 ; 0,3]
RSWP 1p	[5,9 ; 96,9]	[0,1 ; 0,3]	[0,9 ; 73,5]	[0,1 ; 0,3]
SAFE1	[3,9 ; 126,3]	[0,0 ; 0,1]	[3,9 ; 42,5]	[0,0 ; 0,2]
SAFE2	[0,1 ; 21,5]	[0,0 ; 0,1]	[0,1 ; 42,2]	[0,0 ; 0,1]
LHS1	[10,4 ; 126,3]	[0,0 ; 0,4]	[4,5 ; 75,0]	[0,0 ; 0,4]

Quelle: IAW-Berechnungen

Tabelle 4.6: Bandbreiten der Veränderungen der arithmetischen Mittel sowie der Mediane der FuE-Beschäftigungs- und Ausgabenintensitäten und ihrer Ränge bei dreidimensionalen Analysen (Wirtschaftszweige und Beschäftigtengrößenklassen)

Verfahren	Intervall der durchschnittlichen Veränderungen der arithm. Mittel		Intervall der durchschnittlichen Veränderungen der Mediane	
	Veränderung der Werte (in %)	Veränderung der Ränge (normiert)	Veränderung der Werte (in %)	Veränderung der Ränge (normiert)
MA1g	[6,8 ; 570,6]	[0,1 ; 0,1]	[5,9 ; 798,7]	[0,1 ; 0,2]
MA2g	[5,5 ; 305,1]	[0,1 ; 0,1]	[4,9 ; 401,6]	[0,1 ; 0,1]
MA33g	[0,1 ; 0,3]	[0,0 ; 0,0]	[0,1 ; 0,5]	[0,0 ; 0,0]
RSWP 10p	[91,7 ; 747,3]	[0,3 ; 0,3]	[15,8 ; 803,5]	[0,2 ; 0,3]
RSWP 5p	[42,8 ; 1184,9]	[0,2 ; 0,3]	[15,1 ; 1368,1]	[0,2 ; 0,3]
RSWP 1p	[20,1 ; 606,0]	[0,1 ; 0,2]	[10,1 ; 451,6]	[0,1 ; 0,2]
SAFE1	[7,0 ; 309,8]	[0,1 ; 0,1]	[6,9 ; 423,3]	[0,1 ; 0,2]
SAFE2	[2,4 ; 94,2]	[0,0 ; 0,1]	[2,5 ; 162,9]	[0,0 ; 0,1]
LHS1	[19,4 ; 2225,2]	[0,2 ; 0,4]	[14,3 ; 2292,5]	[0,3 ; 0,3]

Quelle: IAW-Berechnungen

In den Tabellen 4.5 und 4.6 wird jeweils die Bandbreite der Abweichungen, die sich durch die einzelnen Anonymisierungsverfahren ergeben, ausgewiesen. Dabei wird danach differenziert, ob es sich um zwei- oder dreidimensionale Auswertungen[18] handelt. Die Abweichungen der Ränge werden dadurch normiert, dass durch die Gesamtzahl der Ränge geteilt wird.

Im Wesentlichen bestätigt auch die Betrachtung der Bandbreiten die obigen Ergebnisse der obigen Beispiele. MA33g zeigt sich als das überlegene Verfahren gefolgt von SAFE2. Im Durchschnitt führen LHS1 und die Rank-Swapping-Verfahren zu den größten Veränderungen. Zu beachten ist allerdings, dass dies für die Veränderung der Mediane und seiner Ränge nicht uneingeschränkt gilt.

4.4 Vergleich ökonometrischer Schätzungen

Um die Auswirkungen der verschiedenen Anonymisierungsverfahren auf ökonometrische Modelle zu untersuchen, gibt es grundsätzlich zwei Vorgehensweisen. Zum einen können theoretische Überlegungen darüber angestellt werden, wie sich Anonymisierungsmaßnahmen auf die Eigenschaften (Erwartungstreue, Konsistenz, Effizienz) von Schätzern auswirken (vgl. hierzu Lechner und Pohlmeier 2003), zum anderen können die Auswirkungen empirisch überprüft werden, indem verschiedene Modelle sowohl für die Originaldaten als auch für die anonymisierten Daten geschätzt werden und die Veränderung

18) Zweidimensional: Nach WZ-Zweistellern, WZ-Vierstellern oder Beschäftigtengrößenklassen. Dreidimensional: Nach WZ-Zweistellern oder WZ-Vierstellern in Verbindung mit Beschäftigtengrößenklassen.

der Koeffizienten und der statistischen Signifikanz von Einflussfaktoren untersucht wird. Solche Simulationen sind insbesondere deshalb notwendig, weil es sich bei den Anonymisierungsverfahren teilweise um sehr komplizierte Algorithmen handelt, deren Auswirkungen nicht ohne weiteres theoretisch abgeleitet werden können[19]. Sinvoll ist dabei insbesondere, sowohl lineare als auch nichtlineare Regressionsmodelle (insbesondere Probit-, Logit- und Tobitmodelle) mit in die Untersuchungen einzubeziehen. Dies kann im Rahmen dieses Beitrags nur für eine sehr eingeschränkte Fragestellung vorgenommen werden.

4.4.1 Die untersuchten Modelle: OLS-, Probit- und Tobitmodelle zur Erklärung der Forschungs- und Entwicklungsintensitäten

Erklärt werden soll die Höhe der FuE-Beschäftigungs- und Ausgabenintensitäten wie sie in Abschnitt 4.3 definiert wurden. Den Regressionsmodellen liegt kein theoretisches Modell zugrunde, vielmehr orientieren sie sich am vorhandenen Datenbestand. Als Einflussgrößen werden die Wirtschaftszweige auf Zweistellerebene (WZ 93), die Beschäftigten (in Tausend), die quadrierten Beschäftigten (in Millionen), die Nettowertschöpfung, der Energieverbrauch, sowie die Anteile der Vorleistungen, Personalausgaben und Fremdkapitalzinsen am Gesamtumsatz[20] verwendet.

Für die Erklärung der Höhe der beiden FuE-Intensitäten werden zunächst normale OLS-Regressionen geschätzt. Aufgrund der extrem linkssteilen Verteilung der Intensitäten werden alternativ auch die logarithmierten Intensitäten als abhängige Variable verwendet. Allerdings gehen durch die Logarithmierung im Originaldatensatz etwa dreiviertel aller Beobachtungen verloren, da sie eine FuE-Intensität von Null aufweisen. Zur Erklärung des qualitativen Unterschieds zwischen vorhandener FuE-Tätigkeit und nicht vorhandener FuE-Tätigkeit werden Probit-Modelle geschätzt.[21] Um den qualitativen Unterschied zwischen einer vorhandenen FuE-Tätigkeit und einer nicht vorhandenen FuE-Tätigkeit einerseits und die Höhe der FuE-Intensitäten in Abhängigkeit von den verschiedenen Einflussvariablen andererseits erklären zu können, werden ergänzend noch Tobit-Modelle angewendet.[22]

4.4.2 Veränderung der Ergebnisse der Schätzungen durch die angewendeten Anonymisierungsverfahren

Beispielhaft sind zunächst in Tabelle 4.9 die Ergebnisse der OLS-Regression der logarithmierten FuE-Beschäftigungsintensitäten auf die verschiedenen Einflussfaktoren abgebildet.

---

19) Die theoretische Ableitung der Auswirkungen der Anonymisierungsmaßnahmen auf die Eigenschaften der Schätzer könnte aber auch ein Kriterium zur Auswahl von Verfahren bei der Erstellung von Scientific-use-files sein (Lechner und Pohlmeier 2003).

20) Die letzten vier Variablen sollen insbesondere die Struktur des Unternehmens abbilden.

21) Hierfür wird eine abhängige Variable gebildet, die bei positiver FuE-Intensität den Wert 1 annimmt, bei einer FuE-Intensität von Null den Wert Null.

22) Zu den mikroökonometrischen Probit- und Tobit-Modellen vgl. insbesondere Ronning (1991).

Tabelle 4.9 a: Ergebnisse von OLS-Schätzungen für die logarithmierten FuE-Beschäftigungsintensitäten (Mikroaggregation und SAFE), P-Werte in Klammern

	(0) Original	(1) MA1g	(2) MA2g	(3) MA33g	(4) SAFE1	(5) SAFE2
Energieverbrauch (Mio)	-0.001	-0.003	-0.004	-0.002	-0.003	-0.002
	(0.107)	(0.000)***	(0.000)***	(0.012)**	(0.001)***	(0.011)**
Vorleistungsquote	0.000	0.001	0.000	0.000	-0.001	0.000
	(0.862)	(0.356)	(0.926)	(0.850)	(0.262)	(0.641)
Personalausgabenquote	-0.001	-0.005	0.004	-0.001	-0.009	-0.001
	(0.282)	(0.002)***	(0.016)**	(0.311)	(0.000)***	(0.347)
Zinsaufwandsquote	0.053	0.086	0.075	0.053	0.059	0.052
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 11 [1]	1.625	-0.369	-0.449	1.785	4.044	2.571
	(0.087)*	(0.469)	(0.383)	(0.058)*	(0.000)***	(0.004)***
WZ 93 Nr. 14	1.929	-0.312	0.015	2.149	3.995	2.323
	(0.015)**	(0.399)	(0.965)	(0.006)***	(0.000)***	(0.003)***
WZ 93 Nr. 15	1.628	-0.274	-0.007	1.843	3.407	1.835
	(0.032)**	(0.374)	(0.981)	(0.014)**	(0.000)***	(0.015)**
WZ 93 Nr. 16	1.989	-0.673	0.077	2.151	5.085	2.149
	(0.028)**	(0.113)	(0.869)	(0.016)**	(0.000)***	(0.013)**
WZ 93 Nr. 17	2.301	0.429	0.272	2.519	4.144	2.485
	(0.002)***	(0.168)	(0.385)	(0.001)***	(0.000)***	(0.001)***
WZ 93 Nr. 18	3.164	0.597	0.265	3.382	4.744	3.240
	(0.000)***	(0.062)*	(0.408)	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 19	2.123	0.227	0.086	2.341	3.532	2.309
	(0.007)***	(0.501)	(0.799)	(0.003)***	(0.000)***	(0.003)***
WZ 93 Nr. 20	1.794	-0.066	0.017	2.013	3.711	1.979
	(0.019)**	(0.837)	(0.958)	(0.008)***	(0.000)***	(0.009)***
WZ 93 Nr. 21	1.598	-0.083	-0.103	1.816	3.542	1.789
	(0.036)**	(0.792)	(0.743)	(0.016)**	(0.000)***	(0.018)**
WZ 93 Nr. 22	2.002	0.151	0.011	2.215	3.792	2.161
	(0.010)**	(0.631)	(0.971)	(0.004)***	(0.000)***	(0.005)***
WZ 93 Nr. 23	2.764	1.111	0.594	2.997	4.125	3.005
	(0.000)***	(0.001)***	(0.097)*	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 24	2.949	1.078	0.835	3.163	5.065	3.128
	(0.000)***	(0.000)***	(0.007)***	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 25	2.111	0.608	0.215	2.325	4.262	2.290
	(0.005)***	(0.048)**	(0.488)	(0.002)***	(0.000)***	(0.002)***
WZ 93 Nr. 26	2.042	0.344	0.178	2.259	4.043	2.225
	(0.007)***	(0.264)	(0.567)	(0.002)***	(0.000)***	(0.003)***
WZ 93 Nr. 27	1.541	0.217	0.067	1.764	3.524	1.735
	(0.042)**	(0.484)	(0.829)	(0.019)**	(0.000)***	(0.021)**
WZ 93 Nr. 28	2.067	0.629	0.243	2.281	4.132	2.250
	(0.006)***	(0.041)**	(0.431)	(0.002)***	(0.000)***	(0.003)***

* = signifikant zu 10 %; ** = signifikant zu 5 %; *** = signifikant zu 1 %.

Quelle: IAW-Berechnungen

1) Statistisches Bundesamt, Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ 93).

Tabelle 4.9 a: Ergebnisse von OLS-Schätzungen für die logarithmierten FuE-Beschäftigungsintensitäten (Mikroaggregation und SAFE), P-Werte in Klammern

	(0) Original	(1) MA1g	(2) MA2g	(3) MA33g	(4) SAFE1	(5) SAFE2
WZ 93 Nr. 29	2.644	1.071	0.727	2.856	4.917	2.824
	(0.000)***	(0.000)***	(0.018)**	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 30	3.784	1.713	1.527	3.989	5.988	3.931
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 31	2.847	1.313	0.878	3.060	5.017	3.031
	(0.000)***	(0.000)***	(0.005)***	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 32	3.322	1.625	1.255	3.530	5.593	3.512
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 33	3.215	1.576	1.243	3.429	5.535	3.385
	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 34	2.527	1.007	0.594	2.735	4.621	2.709
	(0.001)***	(0.001)***	(0.057)*	(0.000)***	(0.000)***	(0.000)***
WZ 93 Nr. 35	2.400	1.030	0.696	2.595	4.850	2.550
	(0.002)***	(0.001)***	(0.030)**	(0.001)***	(0.000)***	(0.001)***
WZ 93 Nr. 36	2.287	0.425	0.295	2.504	4.144	2.473
	(0.002)***	(0.169)	(0.342)	(0.001)***	(0.000)***	(0.001)***
WZ 93 Nr. 37	2.613	0.454	0.435	2.832	3.950	2.803
	(0.011)**	(0.223)	(0.283)	(0.006)***	(0.000)***	(0.006)***
Beschäftigte (Tausend)	-0.002	-0.105	-0.093	-0.010	0.279	-0.010
	(0.916)	(0.000)***	(0.002)***	(0.614)	(0.000)***	(0.605)
Quadratbesch. (Mill.)	-0.000	-0.001	-0.001	-0.000	-0.002	-0.000
	(0.041)**	(0.000)***	(0.000)***	(0.001)***	(0.000)***	(0.001)***
Nettowertsch. (Mill.)	0.000	0.002	0.002	0.001	-0.001	0.001
	(0.026)**	(0.000)***	(0.000)***	(0.002)***	(0.006)***	(0.002)***
Konstante	-1.259	-0.186	-0.178	-1.482	-3.693	-1.459
	(0.096)*	(0.574)	(0.594)	(0.048)**	(0.000)***	(0.052)*
Beobachtungen	4518	8572	8082	4518	8893	4518
Bestimmtheitsmaß	0.213	0.219	0.149	0.214	0.287	0.210
Korr. Bestimmtheitsmaß	0.207	0.217	0.145	0.209	0.284	0.204
F-Wert	37.952	75.038	43.941	38.208	111.248	37.274

* = signifikant zu 10 %; ** = signifikant zu 5 %; *** = signifikant zu 1 %.

Quelle: IAW-Berechnungen

Tabelle 4.9 b: Ergebnisse von OLS-Schätzungen für die logarithmierten FuE-Beschäftigungsintensitäten (Rank-Swapping und LHS1), P-Werte in Klammern

	(0) Original	(6) RSWP 10p	(7) RSWP 5p	(8) RSWP 1 p	(9) LHS1
Energieverbrauch (Mio)	-0.001	-0.003	-0.002	-0.001	-0.004
	(0.107)	(0.003)***	(0.157)	(0.170)	(0.000)***
Vorleistungsquote	0.000	-0.000	0.000	0.000	0.000
	(0.862)	(0.416)	(0.312)	(0.352)	(0.743)
Personalausgabenquote	-0.001	0.000	0.000	-0.000	-0.000
	(0.282)	(0.831)	(0.602)	(0.719)	(0.923)
Zinsaufwandsquote	0.053	-0.001	0.002	0.026	0.012
	(0.000)***	(0.409)	(0.277)	(0.000)***	(0.019)**
WZ 93 Nr. 11	1.625	-1.589	0.000	-0.799	-0.131
	(0.087)*	(0.278)	(.)	(0.513)	(0.905)
WZ 93 Nr. 14	1.929	0.710	-0.218	-0.444	1.440
	(0.015)**	(0.520)	(0.773)	(0.512)	(0.111)
WZ 93 Nr. 15	1.628	0.350	-0.150	-0.707	1.172
	(0.032)**	(0.737)	(0.812)	(0.250)	(0.172)
WZ 93 Nr. 16	1.989	-0.061	-0.469	-0.391	0.899
	(0.028)**	(0.958)	(0.578)	(0.749)	(0.389)
WZ 93 Nr. 17	2.301	0.162	0.143	-0.182	1.175
	(0.002)***	(0.877)	(0.822)	(0.768)	(0.171)
WZ 93 Nr. 18	3.164	0.302	0.205	0.526	1.346
	(0.000)***	(0.775)	(0.752)	(0.407)	(0.128)
WZ 93 Nr. 19	2.123	0.134	0.478	-0.538	0.939
	(0.007)***	(0.902)	(0.496)	(0.419)	(0.295)
WZ 93 Nr. 20	1.794	0.406	-0.444	-0.637	0.984
	(0.019)**	(0.700)	(0.496)	(0.310)	(0.259)
WZ 93 Nr. 21	1.598	0.510	-0.215	-0.770	0.896
	(0.036)**	(0.628)	(0.744)	(0.220)	(0.301)
WZ 93 Nr. 22	2.002	0.740	0.377	0.067	1.078
	(0.010)**	(0.477)	(0.553)	(0.916)	(0.225)
WZ 93 Nr. 23	2.764	-0.046	0.393	0.470	1.376
	(0.000)***	(0.968)	(0.597)	(0.478)	(0.122)
WZ 93 Nr. 24	2.949	0.700	0.689	0.540	1.132
	(0.000)***	(0.501)	(0.275)	(0.378)	(0.185)
WZ 93 Nr. 25	2.111	0.599	0.363	-0.123	1.223
	(0.005)***	(0.565)	(0.566)	(0.841)	(0.154)
WZ 93 Nr. 26	2.042	0.459	0.002	-0.375	1.163
	(0.007)***	(0.659)	(0.998)	(0.541)	(0.174)
WZ 93 Nr. 27	1.541	0.427	0.112	-0.755	1.154
	(0.042)**	(0.683)	(0.861)	(0.221)	(0.179)
WZ 93 Nr. 28	2.067	0.535	0.197	-0.272	1.105
	(0.006)***	(0.606)	(0.755)	(0.658)	(0.197)

* = signifikant zu 10 %; ** = signifikant zu 5 %; *** = signifikant zu 1 %.

Quelle: IAW-Berechnungen

Tabelle 4.9 b: Ergebnisse von OLS-Schätzungen für die logarithmierten FuE-Beschäftigungsintensitäten (Rank-Swapping und LHS1), P-Werte in Klammern

	(0) Original	(6) RSWP 10p	(7) RSWP 5p	(8) RSWP 1 p	(9) LHS1
WZ 93 Nr. 29	2.644	0.801	0.659	0.318	1.207
	(0.000)***	(0.440)	(0.295)	(0.603)	(0.158)
WZ 93 Nr. 30	3.784	1.195	0.697	0.744	1.289
	(0.000)***	(0.257)	(0.281)	(0.232)	(0.135)
WZ 93 Nr. 31	2.847	1.034	1.024	0.526	1.257
	(0.000)***	(0.320)	(0.105)	(0.391)	(0.142)
WZ 93 Nr. 32	3.322	1.010	1.135	0.991	1.223
	(0.000)***	(0.334)	(0.075)*	(0.108)	(0.154)
WZ 93 Nr. 33	3.215	0.989	1.093	0.862	1.195
	(0.000)***	(0.342)	(0.084)*	(0.160)	(0.163)
WZ 93 Nr. 34	2.527	0.841	0.681	0.146	1.075
	(0.001)***	(0.420)	(0.284)	(0.812)	(0.210)
WZ 93 Nr. 35	2.400	1.150	0.528	0.194	1.368
	(0.002)***	(0.274)	(0.415)	(0.756)	(0.113)
WZ 93 Nr. 36	2.287	0.743	0.320	-0.076	1.172
	(0.002)***	(0.475)	(0.614)	(0.902)	(0.172)
WZ 93 Nr. 37	2.613	0.890	-0.300	0.296	1.170
	(0.011)**	(0.402)	(0.676)	(0.667)	(0.331)
Beschäftigte (Tausend)	-0.002	-0.285	-0.204	-0.050	-0.111
	(0.916)	(0.000)***	(0.000)***	(0.000)***	(0.000)***
Quadratbesch. (Mill.)	-0.000	0.002	0.001	0.000	0.001
	(0.041)**	(0.000)***	(0.000)***	(0.025)**	(0.000)***
Nettowertsch. (Mill.)	0.000	-0.000	0.000	0.000	0.000
	(0.026)**	(0.651)	(0.379)	(0.299)	(0.000)***
Konstante	-1.259	0.802	0.917	1.147	0.179
	(0.096)*	(0.439)	(0.145)	(0.061)*	(0.834)
Beobachtungen	4518	4518	4518	4518	4600
Bestimmtheitsmaß	0.213	0.092	0.137	0.180	0.060
Korr. Bestimmtheitsmaß	0.207	0.085	0.131	0.174	0.053
F-Wert	37.952	14.117	22.948	30.716	9.093

* = signifikant zu 10 %; ** = signifikant zu 5 %; *** = signifikant zu 1 %.

Quelle: IAW-Berechnungen

Es ist zu erkennen, dass insbesondere die mit den Verfahren MA33g und SAFE2 bearbeiteten Daten recht ähnliche Ergebnisse wie die Originaldaten erzeugen. Die stärksten Abweichungen ergeben sich bei den mit Rank-Swapping bearbeiteten Daten und bei den mit LHS1 erzeugten Hybrid Daten. Große Abweichungen sind aber auch bei den Verfahren MA1g und MA2g zu beobachten. Das Verfahren SAFE1 verursacht zwar weniger Veränderungen bei der statistischen Signifikanz der Einflussfaktoren, führt aber zu recht großen Veränderungen bei den Werten der Regressionskoeffizienten.

Auffällig ist noch, dass die Verfahren MA1g und SAFE1 zu einer Erhöhung des korrigierten Bestimmtheitsmaßes und des F-Wertes führen, der Aussagen über den gemeinsamen Einfluss aller Einflussfaktoren des Modells macht. Dies ist voraussichtlich auf die bei diesen Anonymisierungsverfahren vergleichsweise starke „Glättung" durch die Durchschnittsbildung zurückzuführen.

Zu erkennen ist auch, dass sich bei den Verfahren MA1g, MA2g und SAFE1 für die vorgenommenen Schätzungen die Zahl der Beobachtungen gegenüber den Originaldaten deutlich erhöht. Dies liegt daran, dass aufgrund der vorgenommenen Durchschnittsbildung im Rahmen der Anonymisierung die Zahl der Unternehmen mit einer FuE-Beschäftigungsintensität von Null systematisch abnimmt. Dieses Phänomen ist insbesondere auch bei den vorgenommenen Probit- und Tobit-Schätzungen von Bedeutung.

Systematische Unterschiede zwischen den unterschiedlichen Schätzmodellen können dennoch bei den hier geschätzten Modellen nicht beobachtet werden. Deshalb ist in Tabelle 4.10 zusammenfassend dargestellt, wie sich die verschiedenen Anonymisierungsverfahren auf die Ergebnisse der Modellschätzungen auswirken.

Als Kriterien für die Veränderung der Ergebnisse werden dabei folgende Maße verwendet:

- Anteil der Einflussfaktoren, bei denen eine Veränderung der statistischen Signifikanz beobachtet werden kann, an allen Einflussfaktoren. Als Grenzen für eine Veränderung werden Signifikanzniveaus von 1 %, 5 % und 10 % herangezogen.
- Anteil derjenigen Einflussfaktoren, die im *Originaldatensatz statistisch signifikant* sind, im *anonymisierten Datensatz hingegen nicht* (mindestens 10 % Signifikanzniveau).
- Anteil derjenigen Einflussfaktoren, die im *Originaldatensatz nicht statistisch signifikant*, im *anonymisierten Datensatz hingegen statistisch signifikant* sind (mindestens 10 % Signifikanzniveau).
- Anteil derjenigen Koeffizienten, die aufgrund der Anonymisierung ihr Vorzeichen verändern.
- Anteil derjenigen Koeffizienten, *die bei gegebener statistischer Signifikanz im Originaldatensatz* (mindestens 10 % Signifikanzniveau) aufgrund der vorgenommenen Anonymisierung ihr Vorzeichen verändern.
- Anteil derjenigen Koeffizienten, *die bei gegebener statistischer Signifikanz im anonymisierten Datensatz* (mindestens 10 % Signifikanzniveau) aufgrund der Anonymisierung ihr Vorzeichen verändern.
- Anteil derjenigen Koeffizienten, deren Werte durch die Anonymisierung statistisch signifikant verändert werden. Eine statistisch signifikante Veränderung der Werte wird dann angenommen, wenn die Koeffizientenwerte bei einem anonymisierten Datensatz außerhalb des 95 % Konfidenzintervalls des originalen Koeffizienten liegen.
- Anteil derjenigen Koeffizienten, deren Werte durch die Anonymisierung statistisch signifikant verändert werden – *bei gegebener statistischer Signifikanz im Originaldatensatz.*

- Anteil derjenigen Koeffizienten, deren Werte durch die Anonymisierung statistisch signifikant verändert werden – *bei gegebener statistischer Signifikanz im anonymisierten Datensatz.*

Auch die komprimierte Darstellung bestätigt die oben bereits erläuterten Resultate. Die besten Ergebnisse nach fast allen Kriterien zeigen die Verfahren MA33g und SAFE2. Hier ergeben sich die geringsten Veränderungen der Ergebnisse. Alle anderen Verfahren verursachen zumindest in der hier erprobten Variante recht eindeutige Veränderungen sowohl bei der statistischen Signifikanz von Einflussfaktoren als auch bei den Werten der Koeffizienten.

4.5 Zwischenfazit für die Kostenstrukturerhebung

Diejenigen Verfahren, welche die univariaten Verteilungen ganz oder weitgehend erhalten (LHS1, Rank-Swapping) führen auf der anderen Seite zu einer weitgehenden Zerstörung der Zusammenhänge zwischen den Variablen. Dies wird insbesondere an der Veränderung der Korrelationskoeffizienten deutlich, ist aber bei den Rangkorrelationen deutlich weniger ausgeprägt. Dennoch führt dies sowohl bei den hier durchgeführten deskriptiven Analysen als auch bei den ökonometrischen Schätzungen zu deutlichen Veränderungen der Ergebnisse.

Insgesamt schneidet das Verfahren am besten ab, bei dem eine getrennte Mikroaggregation für alle 33 stetigen Variablen durchgeführt wird und die diskreten Variablen unbehandelt bleiben. Die Behandlung der diskreten Variablen in Verbindung mit dieser Form der Mikroaggregation (SAFE2) liefert die zweitbesten Ergebnisse. Insbesondere bei den deskriptiven Auswertungen wird aber deutlich, dass die Behandlung der diskreten Merkmale zu einer zusätzlichen Einschränkung des Analysepotenzials führt.[23] [24]

23) Der geringe Anteil derjenigen Einflussfaktoren bei den Rank-Swapping-Verfahren, der im anonymisierten Datensatz statistisch signifikant ist, im Originaldatensatz hingegen nicht, ist darauf zurückzuführen, dass hier insgesamt deutlich weniger Einflussgrößen nach der Anonymisierung überhaupt noch statistisch signifikant sind.

24) Die Modelle sind noch verbesserungsfähig. Eine optimierte Modellspezifikation wird sich vermutlich auch positiv auf die Wirkung der Anonymisierungsverfahren auswirken (vgl. Licht 2003).

Tabelle 4.10: Auswirkungen von Anonymisierungsverfahren auf die Ergebnisse aller durchgeführten Modellschätzungen (OLS, Probit, Tobit)

Verfahren	(1) MA1g	(2) MA2g	(3) MA33g	(4) SAFE1	(5) SAFE2	(6) RSWP 10p	(7) RSWP 5p	(8) RSWP 1p	(9) LHS 1
Anteile jeweils in % an allen Einflussfaktoren									
Veränderung der Signifikanz von Einflussfaktoren	47	47	17	49	31	70	69	62	60
Im Original signifikant, bei Anonymisierung nicht	25	28	0	2	1	51	55	48	35
Bei Anonymisierung signifikant, im Original nicht	3	3	3	14	5	3	1	2	3
Veränderung der Vorzeichen von Koeffizienten	9	14	0	3	0	13	15	22	7
Veränderung von Vorzeichen bei gegebener Signifikanz im Original	6	9	0	1	0	9	12	19	5
Veränderung von Vorzeichen bei gegebener Signifikanz bei Anonymisierung	0	1	0	3	0	3	2	3	2
Signifikante Veränderung der Werte von Koeffizienten	65	67	6	48	11	71	75	66	31
Signifikante Veränderung der Werte von Koeffizienten bei gegebener Signifikanz im Original	62	67	3	41	10	65	72	64	29
Signifikante Veränderung der Werte von Koeffizienten bei gegebener Signifikanz bei Anonymisierung	42	47	6	11	10	30	29	25	20

(Grau unterlegt sind die nach dem jeweiligen Kriterium besten beiden Verfahren.)

Quelle: IAW-Berechnungen

5 Fazit und Ausblick

Mit diesem Beitrag wird ein erstes Zwischenfazit gezogen, wie sich verschiedene Anonymisierungsverfahren auf die Veränderung des Analysepotenzials auswirken. Beispielhaft wird dies für verschiedene Anonymisierungsverfahren mit den Daten der Umsatzsteuerstatistik und der Kostenstrukturerhebung durchgeführt. In einem ersten Schritt wird versucht, das Analysepotenzial zu operationalisieren. Es zeigt sich, dass dies nur bedingt möglich ist. Zur Bewertung der Veränderung des Analysepotenzials durch eine Anonymisierungsmaßnahme muss vielmehr sowohl die Veränderung wesentlicher Charakteristika der Verteilungen als auch die Veränderung der Ergebnisse ganz konkreter deskriptiver und ökonometrischer Analysen herangezogen werden. Neben den in diesem Beitrag untersuchten Veränderungen der arithmetischen Mittel, der Varianzen, der Kovarianzen, der Korrelationskoeffizienten und der Rangkorrelationen sollten insbesondere die Veränderungen der Mediane Gegenstand weiterer Untersuchungen sein. Sinnvoll erscheint allerdings auch die Analyse der Veränderung verschiedener Konzentrationsmaße. Bei der Untersuchung der Veränderung der Korrelationsstruktur sollte neben einer Untersuchung der Korrelationskoeffizienten auch untersucht werden, inwiefern die Zusammenhänge vor und nach der Anonymisierung statistisch signifikant sind.

Auch die vorgenommenen deskriptiven und ökonometrischen Untersuchungen mit beiden Datensätzen sind nicht ausreichend und nur beispielhaft zu verstehen. Deshalb sind weitere Untersuchungen notwendig, um die Stabilität der gefundenen Ergebnisse inhaltlich abzusichern. Für die Kostenstrukturerhebung ist daran gedacht, in einem nächsten Schritt verschiedene Produktionsfunktionen zu schätzen. Daneben sollen die hier vorgestellten ökonometrischen Modelle weiter optimiert werden.

Dennoch lassen sich aus den ersten Untersuchungen bereits Schlussfolgerungen ziehen, welche Anonymisierungsverfahren in welcher Weise das Analysepotenzial verringern. So ist deutlich geworden, dass die Verfahrensgruppe SAFE für die Umsatzsteuerstatistik die besten Ergebnisse erzielt. Dennoch sind hier weitere Verfahren zu untersuchen und zusätzliche Auswertungen vorzunehmen.

Für die Kostenstrukturerhebung kann gezeigt werden, dass diejenigen Verfahren, welche die univariaten Verteilungen annähernd erhalten, wie Rank-Swapping und Latin Hypercube Sampling zu einer teilweise sehr starken Zerstörung der Zusammenhänge zwischen den Variablen führen. Dies bedeutet insbesondere für ökonometrische Schätzungen, aber auch für die deskriptiven Auswertungen eine große Verringerung des Analysepotenzials. Die geringste Verringerung des Analysepotenzials ergibt sich durch die Mikroaggregation für jedes Merkmal getrennt. Die zusätzliche Behandlung der diskreten Variablen im Rahmen des Verfahrens SAFE2 führt zu einer zusätzlichen Einschränkung des Analysepotenzials, dennoch schneidet auch dieses Verfahren recht gut ab.

Die Verfahren unterscheiden sich ferner in ihrer Flexibilität für die Nutzer. Während bei der Anonymisierung mit Latin Hypercube Sampling (LHS) mögliche Teilmassenuntersuchungen bereits bei der Anoymisierung berücksichtigt werden müssen, ist dies bei den anderen Verfahren nicht notwendig. Teilmassenuntersuchungen werden aber insbesondere für die deskriptiven Auswertungen eine wesentliche Rolle spielen.

Die Untersuchung der beiden sehr verschiedenen Datensätze Umsatzsteuerstatistik und Kostenstrukturerhebung zeigt aber auch, dass die Wirkungsweise der verschiedenen

Anonymisierungsverfahren von der Struktur und Beschaffenheit der Daten abhängig ist. Besonders schön beobachten kann man dies daran, dass das Verfahren SAFE2 bei der Umsatzsteuerstatistik zu einer stärkeren Verringerung des Analysepotenzials führt als SAFE1, während es sich bei der Kostenstrukturerhebung genaus andersherum verhält. Der Grund hierfür besteht darin, dass die Umsatzsteuerstatistik aufgrund ihrer großen Zahl an Unternehmen sehr viel dichter besetzt ist und daher die bei SAFE1 durchgeführte Mikroaggregation für alle stetigen Variablen gemeinsam zu geringeren Veränderungen führt als bei der KSE.

Mit diesen ersten Untersuchungen wird deutlich, dass weitere Varianten der in diesem Beitrag betrachteten Verfahren getestet werden müssen. Auch andere Anonymisierungsverfahren, wie beispielsweise die Überlagerung mit Zufallszahlen, sind in die Analysen einzubeziehen. Gleichzeitig wird eine zentrale Aufgabe darin liegen, die Verringerung des Analysepotenzials weiter zu systematisieren und die Zusammenhänge zwischen der Veränderung wesentlicher Verteilungscharakteristika und den Veränderungen der Ergebnisse konkreter Analysen weiter herauszuarbeiten. Das Ziel einer möglichst weitgehenden Erhaltung des Analysepotenzials muss mit der Sicherstellung der faktischen Anonymität in Einklang gebracht werden. Vor dem Hintergrund dieser beiden Ziele müssen die Anonymisierungsverfahren dann verfeinert und für die verschiedenen Erhebungen optimiert werden.

Literaturhinweise

Brand, R. (2000): Anonymität von Betriebsdaten – Verfahren zur Erfassung und Maßnahmen zur Verringerung des Reidentifikationsrisikos. In: Beiträge zur Arbeitsmarkt- und Berufsforschung 237.

Brand, R.; Bender, S. und Kohaut, S. (1999): Möglichkeiten der Erstellung eines Scientific-Use-Files aus dem IAB-Betriebspanel. In: Spektrum Bundesstatistik, Band 14, Statistisches Bundesamt.

Dandekar, R. A. (1993): Performace improvement of Restricted Pairing Algorithm for Latin Hypercube Sampling, ASA Summer Conference, unpublished manuscript.

Dandekar, R. A.; Cohen, M. und Kirkendall, N. (2002): Sensitive Micro Data Protection Using Latin Hypercube Sampling Technique, 2001. In: Domingo-Ferrer, Josep (Ed): Inference Control in Statistical Data Bases – From Theory to Practice. Springer.

Dandekar, R. A.; Domingo-Ferrer, J. und Sebé, F. (2002): LHS-Based Hybrid Microdata vs. Rank Swapping and Microaggregation for Numeric Microdata Protection, 2001. In: Domingo-Ferrer, Josep (Ed): Inference Control in Statistical Data Bases – From Theory to Practice. Springer.

Evers, K. und Höhne, J. (1999): SAFE – Ein Verfahren zur Anonymisierung und statistischen Geheimhaltung wirtschaftsstatistischer Einzeldaten. In: Spektrum Bundesstatistik, Band 14, S. 136 – 147, Wiesbaden.

Gottschalk, S. (2002): Anonymisierung von Unternehmensdaten. Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels. ZEW-Discussion-Paper No. 02 – 23.

Höhne, J. (2002): Messung der Qualität einer anonymen Datei. Arbeitspapier der Projektgruppe „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten".

Höhne, J. (2003): Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten. In: Gnoss, R. und G. Ronning: Anonymisierung wirtschaftsstatistischer Einzeldaten, Forum der Bundesstatistik (in diesem Band S. 69 ff.), Wiesbaden.

Lechner, S. und Pohlmeier, W. (2003): Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten, in: Gnoss, R. und G. Ronning : Anonymisierung wirtschaftsstatistischer Einzeldaten, Forum der Bundesstatistik (in diesem Band S. 115 ff.), Wiesbaden.

Licht, G. (2003): Koreferat, in: Gnoss, R. und G. Ronning: Anonymisierung wirtschaftsstatistischer Einzeldaten, Forum der Bundesstatistik (in diesem Band S. 184 ff.), Wiesbaden.

Müller, W.; Blien, U.; Knoche, P.; Wirth, H. u.a. (1991): Die faktische Anonymität von Mikrodaten. In: Statistisches Bundesamt (Hrsg.): Forum der Bundesstatistik, Band 19.

Ronning, G. (1991): Mikroökonometrie, Berlin: Springer.

Ronning, G.; Brand, R.; Höhne, J.; Rosemann, M. und Wiegert, R. (2002): Anonymisierungsverfahren – Überblick und erste Bewertung. Arbeitspapier der Projektgruppe „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten".

Sebé, F.; Domingo-Ferrer, J.; Mateo-Sanz, J. M. und Torra, V. (2001): Post-Masking Optimization of the Tradeoff between Information Loss and Disclosure Risk in masked Microdata Sets. In: Domingo-Ferrer, Josep (Ed): Inference Control in Statistical Data Bases – From Theory to Practice. Springer, 2002.

Statistisches Bundesamt (2002): Kurzbeschreibungen der Projektdatensätze, Wiesbaden.

Vorgrimler, D. (2002): Probe-Anonymisierung der Umsatzsteuerstatistik. Arbeitspapier der Projektgruppe „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten".

Georg Licht *)

Koreferat zum Beitrag „Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik"

Die Verwendung von Individualdaten hat in den letzten zwanzig Jahren in der ökonomischen und sozialwissenschaftlichen Forschung erheblich zugenommen. Dabei dominierte der Einsatz von personen- und haushaltsbezogenen Individualdaten, was unmittelbar mit der Verfügbarkeit entsprechender Datensätze (z.B. Sozio-ökonomisches Panel, Mikrozensus) zusammenhängt. Das Angebot an Unternehmens- oder Betriebsdaten ist dagegen vergleichsweise gering. Dies liegt jedoch nicht daran, dass kaum Daten auf der Unternehmens- oder Betriebsebene erhoben werden. Vielmehr besteht auf Seiten der datenproduzierenden Institutionen die Sorge, dass einzelne im Datensatz enthaltene Unternehmen oder Betriebe deutlich leichter erkannt werden können als Individuen in personenbezogenen Daten. Daher werden Mikrodaten für Unternehmen/Betriebe entweder überhaupt nicht für Forschungszwecke zur Verfügung gestellt oder nur in einer Form, bei der nicht die ganze im Datensatz enthaltene Information weitergegeben wird. Die Informationsreduktion dient dabei dem Zweck, die Identifikation einzelner Unternehmen zu verhindern. In den letzten Jahren wurde eine Reihe von Vorschlägen entwickelt, wie bei der Anonymisierung von Unternehmensdaten vorgegangen werden kann. Trotz großer Unterschiede im einzelnen ist allen Verfahren gemein, dass die in den Daten enthaltene Information reduziert wird und damit natürlich auch die Analysemöglichkeiten, die ein Datensatz bietet, eingeschränkt wird. Bislang wurde nur in seltenen Fällen explizit der Frage nachgegangen, wie stark die Informationsreduktion ist und welche Auswirkungen die Informationsreduktion auf statistische Auswertungen hat. Das Papier von Rosemann stellt sich genau diese Aufgabe. Für die Auswahl eines konkreten Verfahrens zur Anonymisierung von Daten kommt es nämlich nicht nur darauf an, ob und wie leicht die Identität der Unternehmen erkannt werden kann. Vielmehr wird das Angebot an anonymisierten Unternehmensdaten aus der amtlichen Statistik nur dann von Wissenschaftlern angenommen werden, wenn in den Daten hinreichend Information verbleibt, die anspruchsvolle Auswertungen ermöglicht.

Rosemann nähert sich der Analyse der Informationszerstörung durch Anonymisierung in zwei unterschiedlichen Ansätzen: Zum einen betrachtet er die Abweichungen der ersten und zweiten Momente stetiger Variablen zwischen anonymisierten und Originaldaten; zum anderen demonstriert er die Auswirkungen der Anonymisierung am Beispiel von typischen deskriptiven Auswertungen und von linearen und nicht-linearen Regressionsmodellen, die sowohl für anonymisierte Daten als auch für nicht-anonymisierte Daten geschätzt werden.

*) Dr. Georg Licht, Zentrum für Europäischen Wirtschaftsforschung, Mannheim.

Die meisten getesteten Verfahren werden bei dieser Betrachtung dem Ziel der Erhaltung des Analysepotenzials nicht gerecht und sind damit für die Anonymisierung von Einzeldaten nicht geeignet. Eine eindeutige Hierarchie aller Verfahren im Hinblick auf die Informationserhaltung lässt sich auf der Basis des vorliegenden Beitrags nicht ermitteln. Allerdings sind interessante Unterschiede in der Wirkungsweise der verschiedenen Verfahren erkennbar. So zeigen die Berechnungen, dass die Relevanz der Informationszerstörung durch Anonymisierung auch von der Art der statistischen Analysen abhängt, die mit dem anonymisierten Datenmaterial durchgeführt werden sollen. Dies wird insbesondere deutlich, wenn man die von Rosemann berechneten Veränderungsmaße für die verschiedenen Verteilungschrakteristika betrachtet. Es fällt auf, dass manche Verfahren sowohl die ersten (und zweiten) Momente als auch die Zusammenhänge zwischen den Variablen ganz gut erhalten (MA33g und SAFE2). Andere Verfahren erhalten lediglich die univariaten Verteilungseigenschaften, während sie die Zusammenhänge sehr stark zerstören (LHS1, Rank-Swapping-Verfahren). Diese Ergebnisse sind in den Abbildungen 1 und 2 visualisiert.

Allerdings sollte man klarer herausstellen, dass einige der analysierten Anonymisierungsverfahren zu so hohen Abweichungen bei den univariaten und/oder multivariaten Statistiken führen, dass auf der Basis solchermaßen anonymisierter Daten kaum verlässliche Analysen durchgeführt werden können. Entsprechend wird sich kaum ein Wissenschaftler finden lassen, der solchermaßen anonymisierte Daten auch tatsächlich als Analysegrundlage akzeptieren wird.

Abbildung 1: Statistischer Vergleich von Anonymisierungsmethoden hinsichtlich des Erhalts der 1. Momente und der Korrelationen

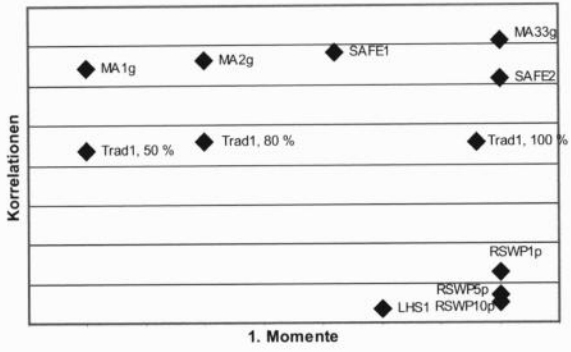
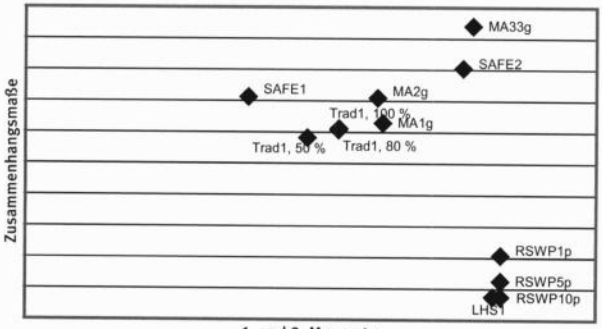


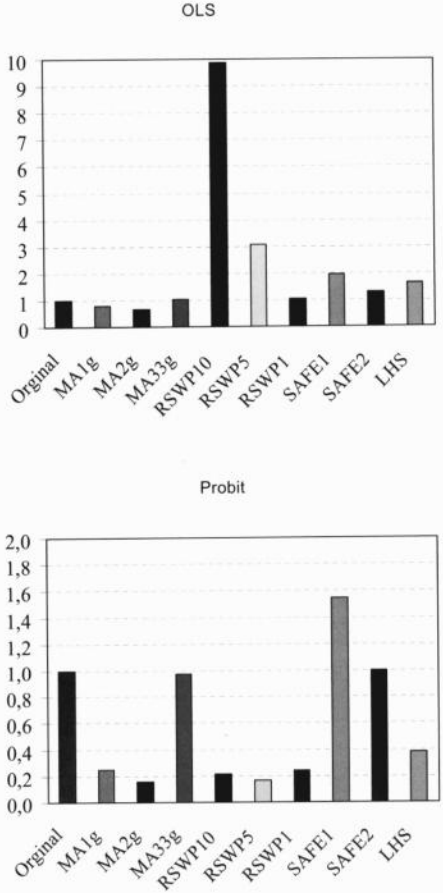
Abbildung 2: Statistischer Vergleich von Anonymisierungsmethoden hinsichtlich des Erhalts der 1. und 2. Momente und der Zusammenhangsmaße (Korrelationen und Rangkorrelationen)



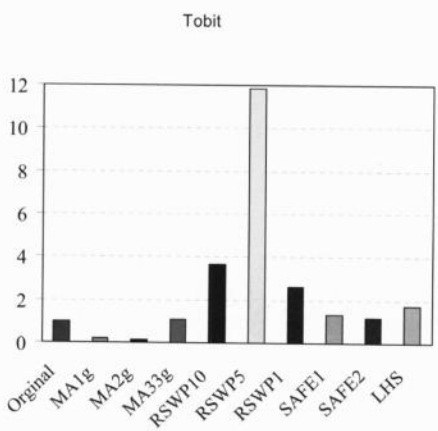
Rosemann zeigt, dass das Ausmaß der Informationserhaltung nicht nur von der gewählten Anonymisierungsmethode sondern auch von den Eigenschaften des ursprünglichen Datenmaterials abhängig ist. So ist es auffällig, dass die Wirkungsweise mancher Verfahren bei den beiden von Rosemann untersuchten Statistiken unterschiedlich ausfällt. Insbesondere das Verfahren SAFE1 schneidet bei der Umsatzsteuerstatistik deutlich besser ab als bei der Kostenstrukturerhebung, was Rosemann auf die unterschiedliche Dichte der Datenpunkte zurückführt.

Entscheidender ist jedoch der Unterschied zwischen typischen deskriptiven Auswertungen und multivariaten Verfahren wie beispielsweise Regressionsanalysen. Allerdings lässt sich bei den im Beitrag von Rosemann vorgenommenen Untersuchungen kein systematischer Unterschied zwischen beiden Arten der Auswertung herauslesen – die Auswertungen stehen hierzu bisher auf keiner ausreichend breiten Grundlage. Da multivariate Analysen bisher nur mit der KSE durchgeführt wurden, stehen Ergebnisse für multivariate Analysen nicht für alle Verfahren zur Verfügung. Insgesamt schneidet ohnehin aus Sicht des Analysepotenzials lediglich das Verfahren der Mikroaggregation für alle Variablen getrennt (MA33g) gut ab. Mit Einschränkungen sind noch die Verfahren SAFE2 für die Kostenstrukturerhebung und SAFE1 für die Umsatzsteuerstatistik als hoffnungsvoll zu bezeichnen. In Abbildung 3 ist beispielhaft visualisiert, wie sich der geschätzte Koeffizient für den Wirtschaftszweig 29 (Maschinenbau) durch die verschiedenen Anonymisierungsverfahren in den von Rosemann mit der KSE geschätzten OLS-, Probit- und Tobitmodellen verändert.

Abbildung 3: Vergleich der geschätzten Koeffizienten für WZ 29 (Maschinenbau) Beispiel: FuE-Personalintensität



noch Abbildung 3: Vergleich der geschätzten Koeffizienten für WZ 29 (Maschinenbau) Beispiel FuE-Personalintensität



Während ein Teil der Wissenschaftler für ihre Arbeiten primär deskriptive Statistiken benötigt, wird ein zweiter Teil sich intensiv auf der Basis multivariater Verfahren sozial- und wirtschaftswissenschaftlichen Problemstellungen widmen. Sollte sich im Verlauf der weiteren Untersuchungen herausstellen, dass gerade unterschiedliche Anonymisierungsverfahren geeignet sind, die Ergebnisse deskriptiver Auswertungen einerseits und die Ergebnisse multivariater Analysen andererseits zu erhalten, würde dies bedeuten, dass die beiden genannten Gruppen von Wissenschaftlern unterschiedliche Anonymisierungsverfahren präferieren. Als eine – im Papier nicht angesprochene – Frage ergibt sich dann, ob nicht für unterschiedliche Zielgruppen unterschiedliche Anonymisierungsverfahren eingesetzt werden sollten. Daraus ergäbe sich aber unmittelbar die Frage, inwieweit nicht die Gleichzeitigkeit eines solchen Mehrfachangebots selbst wiederum eine zusätzliche Gefährdung der Anonymität von Einzeldaten darstellt.

Für weitere Untersuchungen der Auswirkungen von Anonymisierungsmaßnahmen auf das Analysepontenzial von Einzeldaten können folgende Anregungen von Interesse sein:

- Hilfreich ist die Verwendung eines simulierten Datensatzes und/oder unterschiedlich großer Stichproben, um zu untersuchen, wie sensitiv einzelne Verfahren auf die „Dichte" der Merkmalsausprägungen oder auf Ausreißer reagieren. Dies würde es erleichtern, Rückschlüsse hinsichtlich der Auswirkungen der Anonymisierung bei anderen Datensätzen zu ziehen.

- Sinnvoll erscheint es, die Wirkungsweise der so genannten traditionellen Verfahren auch im Hinblick auf multivariate Modelle zu untersuchen.
- Untersucht werden sollte die Sensitivität der Ergebnisse gegenüber Fehlspezifikationen ökonometrischer Modelle bei unterschiedlichen Anonymisierungsverfahren.
- Die Betrachtung einzelner Verfahren scheint für die Lösung des Problems nicht ausreichend zu sein. Deshalb sollten die Möglichkeiten und Auswirkungen eines Methoden-Mix untersucht werden. Beispielsweise könnte man die Auswirkung einer Kombination von traditionellen Verfahren in Verbindung mit der Mikroaggregation für alle Variablen getrennt (MA33g) studieren.

Die von Rosemann geschätzten Modelle zur Erklärung der FuE-Intensitäten sind noch nicht optimal spezifiziert. Eine optimale Spezifikation könnte sich auch vorteilhaft für die Ergebnisse mit den anonymisierten Daten auswirken. Es ist üblich, die FuE-Ausgabenintensität als Anteil der FuE-Gesamtausgaben am Umsatz aus eigenen Erzeugnissen zu definieren. Rosemann weicht davon ab, indem er die FuE-Ausgaben auf den Gesamtumsatz bezieht. Weiterhin sind als Einlfussgrößen zur Erklärung der FuE-Intensitäten folgende Variablen zu empfehlen:

Größe, Größe^2, Größe^3, Personalkosten/Nettowertschöpfung, Fremdkapitalaufwand/Nettowertschöpfung, Wirtschaftszweige-Dummies, Unternehmenskonzentration, Spillover.

Schließlich sollte angemerkt werden, dass die Entscheidung für oder gegen eine Anonymisierungsmethode auch den Grad des intendierten Schutzes gegen „Datenangreifer" berücksichtigen muss. Das heißt aber auch, dass die hier vorgelegten Untersuchungen um Indikatoren zu ergänzen sind, mit Hilfe derer die Schutzwirkung beurteilt werden kann. Allerdings sei hier daran erinnert, dass es immer um einen Kompromiss zwischen Schutzwirkung und Informationserhaltung geht. Werden aus Gründen des Datenschutzes „zu viel" Informationen aus dem Datensatz entfernt, wird die Bereitschaft von Wissenschaftlern, mit anonymisierten Daten zu arbeiten, im Extremfall auf Null reduziert. Damit verliert aber die Diskussion um Anonymisierungsmethoden ihre Berechtigung. Die hier vorgelegten Ergebnisse unterstreichen vor diesem Hintergrund auch die Bedeutung der im Aufbau befindlichen Forschungsdatenzentren, die ein Arbeiten mit Originaldaten ermöglichen. Aber selbst wenn sich herausstellt, dass der angesprochene Kompromiss zwischen Datenschutz und Informationserhaltung kaum erreichbar ist, ist die Diskussion um die Anonymisierung wichtig. Denn möglicherweise könnte aus dem Zusammenspiel des Angebots der Forschungsdatenzentren und dem Angebot an anonymisierten Datensätzen der Königsweg darin bestehen, dass Wissenschaftler auf der Basis von anonymisierten Daten erste Ergebnisse produzieren, Schätzprogramme etc. entwickeln und anschließend mit Originaldaten in den Forschungsdatenzentren die letztendlichen empirischen Ergebnisse ermitteln. Auch unter dieser Perspektive ist das Papier von Rosemann ein wichtiger Beitrag zur laufenden Diskussion, wie das Angebot an Forschungsdaten in Deutschland verbessert werden kann.

Gerd Ronning *)

Was hat die Veranstaltung gebracht? Ein Resumee der Tagung

Die im vorliegenden Tagungsband enthaltenen Beiträge berichten über den Stand eines Forschungsprojektes, dessen Ziel es ist, die methodischen Grundlagen für die Entwicklung von Scientific-use-files für Einzeldaten über Unternehmen und Betriebe zu legen, nachdem bereits vor etwa 10 Jahren unter Walter Müllers Leitung die Möglichkeit untersucht wurde, Public-Use-Files aus dem Bereich der personenbezogenen Daten zur Verfügung zu stellen. Das wurde bereits im Vorwort zu diesem Band angesprochen. Hier soll versucht werden, eine kurze Bewertung der präsentierten Ergebnisse vorzunehmen, soweit dies für jemand, der aktiv als Leiter des Projektes, aber auch bei der Erarbeitung von Teilergebnissen mitarbeitet, überhaupt objektiv möglich ist.

Wesentliche Herausforderung ist nach wie vor, eine angemessene operationale sprich datenbezogene Formulierung der Faktischen Anonymität festzulegen. Die komplexen Aspekte des Risikos der Re-Identifikation werden sich nur dann befriedigend mit den Forderungen der Datennutzer nach einem ausreichendenden Analyse-Potenzial vereinigen lassen, wenn das wissenschaftliche Interesse der Datennutzer weit höher gewichtet wird als die – nie auszuschließende – Möglichkeit, dass einzelne Personen die Geheimhaltung brechen wollen. Ich verweise dazu auf dezidierte Äußerungen im Beitrag von Heike Wirth. Die im Beitrag von Daniel Vorgrimler betrachteten Angriffs-Szenarien und die für konkrete Datensätze identifizierten Einheiten machen zudem deutlich, dass das Ausmaß der Re-Identifikation durchaus unterschiedlich gewertet werden kann und vermutlich auch von Datennutzern und Datenanbietern unterschiedlich gesehen wird. Wir sind zuversichtlich, für konkrete Datensätze einen machbaren Mittelweg zu finden.

Die Datennutzer ihrerseits müssen sich darüber klar werden, dass jede Form der Anonymisierung den Aussagegehalt von Daten einschränkt. Inwieweit rein statistisch-technisch orientierte Maße dabei helfen können, das verbleibende Analysepotenzial zu messen, ist meines Erachtens noch völlig offen. Auf jeden Fall sollten auch andere Aspekte bei der Quantifizierung des Verlustes an Analyse-Potenzial berücksichtigt werden. Martin Rosemanns Beitrag gibt dazu erste Anregungen.

Die Beiträge von Jörg Höhne sowie von Rolf Wiegert zeigen, dass die Entwicklung der Methoden und Verfahren im Bereich der Anonymisierung einerseits und der Re-Identifikation andererseits im letzten Jahrzehnt stark fortgeschritten ist. Ferner zeigt die Arbeit von Sandra Lechner und Winfried Pohlmeier einen neuen Weg auf, um die Anonymisierung auf analytischem Wege in die statistische und ökonometrische Analyse einzuarbeiten. Vor allem diese drei zuletzt genannten Beiträge offenbaren aber auch, dass das Thema „(Faktische) Anonymisierung" aus der Sicht der Datenanalyse noch ganz am Anfang steht. Die Resonanz auf die Tagung macht andererseits deutlich, dass in verschiedenen wissenschaftlichen Bereichen großes Interesse an diesem Projekt besteht und dass die Tagungsbeiträge wichtige Denkanstöße für die weitere Arbeit zur Erstellung von Scientific-Use-Files aus dem Unternehmensbereich gegeben haben.

---

*) Prof. Dr. Gerd Ronning, Universität Tübingen und Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen.

Teilnehmerverzeichnis

Grußwort

Friedrich, Reinhold; *Bundesministerium für Bildung und Forschung*, Bonn
Ronning, Prof. Dr. Gerd; *Institut für Angewandte Wirtschaftsforschung*, Tübingen
Schaich, Prof. Dr. Eberhard; *Universität Tübingen*

Referentinnen/Referenten – Ko-Referentinnen/-Referenten

Blien, Dr. Uwe; *Institut für Arbeitsmarkt- und Berufsforschung (IAB)*, Nürnberg
Brand, Dr. Ruth; *Statistisches Bundesamt*, Wiesbaden
Domingo-Ferrer, Dr. Josep; *Universität Rovira y Virgili*, Tarragona/Spanien
Höhne, Jörg; *Statistisches Landesamt Berlin*
Giessing, Sarah; *Statistisches Bundesamt*, Wiesbaden
Gnoss, Dr. Roland; *Statistisches Bundesam*, Wiesbaden
Licht, Dr. Georg; *Zentrum für Europäische Wirtschaftsforschung (ZEW)*, Mannheim
Pohlmeier, Prof. Dr. Winfried; *Universität Konstanz*
Rosemann, Martin; *Institut für Angewandte Wirtschaftsforschung (IAW)*, Tübingen
Strotmann, Dr. Harald; *Institut für Angewandte Wirtschaftsforschung (IAW)*, Tübingen
Vorgrimler, Dr. Daniel; *Statistisches Bundesamt*, Wiesbaden
Wagner, Prof. Dr. Joachim; *Universtität Lüneburg*
Wiegert, Dr. Rolf; *Institut für Angewandte Wirtschaftsforschung (IAW)*, Tübingen
Wirth, Dr. Heike; *Zentrum für Umfragen, Methoden und Analysen (ZUMA)*, Mannheim

Weitere Teilnehmerinnen und Teilnehmer

A

Arndt, Christian; *Institut für Angewandte Wirtschaftsforschung (IAW)*, Tübingen

B

Bajaja, Dr. Vladislav; *Statistisches Bundesamt*, Wiesbaden
Bender, Stefan; *Institut für Arbeitsmarkt- und Berufsforschung (IAB)*, Nürnberg
Buschle, Dr. Nicole; *Statistisches Bundesamt*, Wiesbaden

D

Deutschmann, Prof. Dr. Christoph; *Universtität Tübingen*

F

Frank, Eberhard; *Statistisches Amt Stuttgart*
Fitzenberger, Prof. Bernd; *Universität Mannheim*

Frietsch, Rainer; *Fraunhofer-Institut für Systemtechnik und Innovationsforschung (ISI)*, Karlsruhe

G

Gottschalk, Sandra; *Zentrum für Europäische Wirtschaftsforschung (ZEW)*, Mannheim
Gräb, Christopher; *Statistisches Bundesamt*, Wiesbaden
Gruber, Winfried; *Statistisches Landesamt Baden-Württemberg*, Stuttgart
Grupp, Prof. Dr. Hariolf; *Fraunhofer-Institut für Systemtechnik und Innovationsforschung (ISI)*, Karlsruhe
Günterberg, Brigitte; *Institut für Mittelstandsforschung*, Bonn

H

Hellmich, Eva; *Statistisches Landesamt Sachsen-Anhalt*, Halle (Saale)

J

Jung, Dr. Robert; *Universität Tübingen*

K

Klee, Günther; *Institut für Angewandte Wirtschaftsforschung (IAW)*, Tübingen
Koch, Iris; *Bundesanstalt für Arbeit*, Nürnberg
Krüger, Dr. Antje; *Landesamt für Datenverarbeitung und Statistik NRW*, Düsseldorf
Kurtschanowa, Marija; *Institut für Angewandte Wirtschaftsforschung (IAW)*, Tübingen

L

Lechner, Sandra; *Universität Konstanz*
Lenz, Dr. Rainer; *Statistisches Bundesamt*, Wiesbaden
Loose, Dr. Brigitte; *Institut für Wirtschaftsforschung*, Halle (Saale)
Loreth, Dr. Hans; *Statistisches Landesamt Baden-Württemberg* , Stuttgart
Luckert, Hilmar; *Verband Deutscher Rentenversicherungsträger (VDR)*, Frankfurt/Main

M

Mai, Joachim; *Niedersächsisches Landesamt für Statistik*, Hannover
Meinken, Holger; *Bundesanstalt für Arbeit*, Nürnberg
Merz, Prof. Dr. Joachim; *Universität Lüneburg*
Metschke, Rainer; *Datenschutzbeauftragter Berlin*

O

Opfermann, Rainer; *Statistisches Bundesamt*, Wiesbaden

P

Padberg, Dr. Frank; *Hessisches Statistisches Landesamt*, Wiesbaden
Pohl, Ramona; *FHTW Berlin*

R

Rehfeld, Uwe; *Verband Deutscher Rentenversicherungsträger (VDR)*, Frankfurt/Main
Revermann, Christa; *Wissenschaftsstatistik im Stifterverband für die Deutsche Wissenschaft*, Essen
Rose, Petra; *Landesamt für Datenverarbeitung und Statistik NRW*, Düsseldorf

S

Scharnhorst, Sebastian; *Landesamt für Datenverarbeitung und Statistik NRW*, Düsseldorf
Schimpl-Neimanns, Bernhard; *Zentrum für Umfragen, Methoden und Analysen* (ZUMA), Mannheim
Schneider, Dr. Patrick; *Landesamt für Datenverarbeitung und Statistik NRW*, Düsseldorf
Schrödter, Dietmar; *Statistisches Landesamt Schleswig-Holstein*, Kiel
Schürle, Josef; *Universität Tübingen*
Schuster, Michael; *Deutsche Forschungsgemeinschaft (DFG)*, Bonn
Stegenwaller, Lars; *Landesamt für Datenverarbeitung und Statistik NRW*, Düsseldorf
Stephan, Andreas; *Deutsches Institut für Wirtschaftsforschung (DIW)*, Berlin
Stock, Dr. Gerhard; *Statistisches Bundesamt*, Wiesbaden
Sturm, Roland; *Statistisches Bundesamt*, Wiesbaden

V

v. der Heyde, Christian A.; *Infratest Sozialforschung GmbH*, München

W

Wahl, Dr. Anke; *Universtität Tübingen*
Walter, Mario; *Bayerisches Landesamt für Statistik und Datenverarbeitung*, München
Wende, Thomas; *Forschungsdatenzentrum des Statistischen Bundesamtes*, Wiesbaden
Wingerter, Christian; *Institut für Angewandte Wirtschaftsforschung (IAW)*, Tübingen
Wohlers, Dr. Eckhardt; *Hamburgisches Welt-Wirtschafts-Archiv (HWWA)*, Hamburg

Z

Zühlke, Dr. Sylvia; *Forschungsdatenzentrum der Statistischen Ämter der Länder*, Düsseldorf
Zwick, Markus; *Forschungsdatenzentrum des Statistischen Bundesamtes*, Wiesbaden