Gottschalk, Sandra

Working Paper

Anonymisierung von Unternehmensdaten: Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels

ZEW Discussion Papers, No. 02-23

Provided in Cooperation with:
ZEW - Leibniz Centre for European Economic Research

Suggested Citation: Gottschalk, Sandra (2002) : Anonymisierung von Unternehmensdaten: Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels, ZEW Discussion Papers, No. 02-23, Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim

This Version is available at:
https://hdl.handle.net/10419/24783

Discussion Paper No. 02-23

Anonymisierung von Unternehmensdaten

Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels

Sandra Gottschalk

ZEW

Zentrum für Europäische Wirtschaftsforschung GmbH

Centre for European Economic Research

Discussion Paper No. 02-23

Anonymisierung von Unternehmensdaten

Ein Überblick und beispielhafte Darstellung anhand des Mannheimer Innovationspanels

Sandra Gottschalk

Das Wichtigste in Kürze

Empirische Forschung in den Sozial- und Wirtschaftswissenschaften erfordert einen regelmäßigen Zugang zu Einzelangaben über Personen, Haushalte und Unternehmen. Die vom Bundesministerium für Bildung und Forschung eingesetzte Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik empfiehlt daher unter Anderem die Entwicklung von faktisch anonymisierten Datensätzen (Scientific-Use-Files) und deren Weitergabe an wissenschaftliche Datennutzer.

Im Gegensatz zu personenbezogenen Daten existieren für Firmendaten keine einheitlichen Anonymisierungsregeln. Das Risiko einer Reidentifikation eines Merkmalsträgers ist wesentlich höher bei Unternehmens- bzw. Betriebsdaten als bei Personendaten. Daher sind weiterreichende Schutzmaßnahmen erforderlich. Die Anonymisierung steht dabei im Konflikt zu dem Nutzen, den Daten für wissenschaftliche Zwecke stiften sollen. Durch die Anwendung geeigneter ökonometrischer Methoden können die eingebrachten Fehler jedoch bis zu einem gewissen Grad korrigiert werden.

Ziel dieses Papiers ist es, einen Überblick über gängige Anonymisierungsmaßnahmen zu geben und beispielhaft eine Korrekturmöglichkeit bei ökonometrischen Schätzungen vorzustellen: Anhand von Simulationen und Schätzungen mit dem Mannheimer Innovationspanel wird gezeigt, wie Verzerrungen von Regressionskoeffizienten, die durch eine multiplikative Fehlerüberlagerung entstehen, behoben werden können.

Anonymisierung von Unternehmensdaten

Ein Überblick und Beispielhafte Darstellung anhand des Mannheimer Innovationspanels

von
Sandra Gottschalk

Zentrum für Europäische Wirtschaftsforschung (ZEW)

3. April 2002

Zusammenfassung: Für Unternehmensdaten existieren im Gegensatz zu Personendaten keine einheitlichen Regeln zur Erstellung von Scientific-Use-Files, d.h. eines anonymisierten Datenfiles zur wissenschaftlichen Nutzung. Verschiedene Anonymisierungsmethoden werden hier vorgestellt. Um Mikrodaten einer breiten wissenschaftlichen Forschung zugänglich zu machen, soll deren Analysefähigkeit weitgehend erhalten bleiben. Es muss demnach ein Kompromiss zwischen den Anforderungen, der durch die den Merkmalsträgern - hier Unternehmen - gegebenen Zusage des Vertrauensschutzes entsteht, und dem Erhalt von wesentlichen Informationen für wissenschaftliche Untersuchungen gefunden werden. Ferner kann der Datennutzer selbst die Qualität seiner empirischen Analysen mit anonymisierten Mikrodaten verbessern. Zur Demonstration wird hier beispielhaft eine Korrekturmöglichkeit durch die Anononymisierung verzerrter Daten bei ökonometrischen Schätzungen vorgestellt.

Keywords: Unternehmensdaten, Anonymisierungsmaßnahmen, Analysefähigkeit, Korrekturmöglichkeit verzerrter Daten

JEL Klassifikation: C13, C15, C81

1 Einleitung

Empirische Forschung in den Sozial- und Wirtschaftswissenschaften beruht vielfach auf Informationen über Haushalte und Unternehmen. Diese Informationen liegen als so genannte Mikrodaten, d.h. Einzelangaben von Personen, Haushalten und Unternehmen, vor, die durch Befragungen, durch systematische Beobachtung oder durch die Aufzeichnung der Geschäftstätigkeit von Behörden oder privaten Einrichtungen gewonnen werden. In den letzten Jahren hat die empirische Forschung und damit auch der Wunsch nach Freigabe von Mikrodaten stark zugenommen, was zum einen auf die gestiegene Verfügbarkeit leistungsstarker Hard- und Softwaresysteme zurückzuführen ist. Die Untersuchung wirtschaftlichen und sozialen Wandels erfordert zum anderen die Verfügbarkeit von Informationen auf der Mikroebene von Beobachtungseinheiten zu verschiedenen Zeitpunkten. Die Anwendung statistischer Methoden zur Messung von Verhaltensregelmäßigkeiten und -zusammenhängen erfordert den Zugang zu Mikrodaten.

Um die Zusammenarbeit zwischen Wissenschaft und Statistik zu verbessern und damit Daten einem größeren Kreis wissenschaftlicher Nutzer verfügbar zu machen, hat das Bundesministerium für Bildung und Forschung (bmb+f) eine Kommission zu Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik eingesetzt. Die Aufgaben der Kommission bestanden darin zu überprüfen, ob die informationelle Infrastruktur in Deutschland den Ansprüchen der modernen Informationsgesellschaft noch gerecht werden kann und Vorschläge zu deren Verbesserung zu erarbeiten. Es wurden alle in Deutschland vorhandenen Datenquellen miteinbezogen und internationale Erfahrungen im Umgang mit Daten genutzt. Die Kommission empfiehlt die Koordination zwischen Wissenschaft und Statistik durch institutionelle Regelungen zu verbessern. Die Wissenschaft solle systematisch an der Aufstellung von Erhebungs- und Aufbereitungsprogrammen der amtlichen Statistik mitwirken. Ferner sei die Entwicklung von faktisch anonymisierten Datensätzen (Scientific-Use-Files) als wichtigstes Instrument des Mikrodatenzugangs weiter voranzutreiben. Um auch die Nutzung nicht anonymisierter Daten für Wissenschaftler ermöglichen zu können, wird die Einrichtung von Forschungsdatenzentren empfohlen.

Im Allgemeinen sind Mikrodaten nicht frei zugänglich. Amtliche Statistiken unterliegen den Regelungen des Bundesdatenschutzgesetzes (BDSG), des Bundesstatistikgesetzes (BStatG) oder des Sozialgesetzbuches (SGB). Die gesetzlichen Bestimmungen beziehen zwar Befragungen von Forschungseinrichtungen und privaten Meinungsforschungsinstituten nicht mit ein. Jedoch dürfen auch derartig gewonnene Informationen nicht frei zugänglich gemacht werden, wenn den Befragten eine Vertraulichkeitszusage gegeben wurde. Bei Firmenbefragungen werden Vertraulichkeitszusagen erteilt, um die Bereitschaft der Befragten zu erhöhen, potenziell sensible Auskünfte zu erteilen. Nur so kann die Qualität einer Datenerhebung sichergestellt werden. Diese Zusicherung impliziert, dass die erhobenen Informationen in Form des Mikrodatenfiles nur an Dritte weitergegeben dürfen, wenn keine Rückschlüsse auf die Angaben der Befragten möglich sind. Daher müssen die Daten vor der Weitergabe vom Daten erhebenden Institut anonymisiert werden.

Für Firmendaten existieren keine einheitlichen Regeln zur Anonymisierung von Mikrodaten. Die Methoden, die dabei verwendet werden, werden hier kurz vorgestellt. Es kann zwischen Maßnahmen unterschieden werden, die keine direkte Datenmodifikation vorsehen, wie die Unterdrückung von Angaben, einer Vergröberung von Informationen und zuletzt einer Verfälschung des Datenbestandes. Die Art des Eingriffs bzw. die Wahl

einer dieser Methoden zur Anonymisierung hängt dabei in erster Linie von der Gefahr einer Reidentifikation eines Mikrodatensatzes ab. Eine Reidentifikation hat stattgefunden, wenn eine Beobachtung des anonymisierten Mikrodatenfiles eindeutig einem Datensatz aus dem Bestand des Zusatzwissens, einem so genannten Identifikationsfile, zugeordnet werden kann.

Die Auswahl einer Anonymisierungsmaßnahme sollte in erster Linie den an die Befragten erteilten Vertrauensschutz berücksichtigen. Um die Mikrodaten einer breiten wissenschaftlichen Forschung zugänglich zu machen, sollte die Analysefähigkeit der Mikrodaten allerdings weitgehend erhalten bleiben. Es muss demnach ein Kompromiss zwischen den Anforderungen, der durch die Zusage des Vertrauensschutzes entsteht, und dem Erhalt von wesentlichen Informationen für die wissenschaftliche Untersuchungen gefunden werden. Ein sinnvoller Einsatz von Anonymisierungsmethoden, auch Maskierungsmethoden genannt, nimmt auf diese Maßgaben Bezug.

Ziel dieses Papers ist es, einen Überblick über gängige Anonymisierungsmaßnahmen zu geben und beispielhaft Korrekturmöglichkeiten bei ökonometrischen Schätzungen vorzustellen.

In dieser Arbeit werden im folgenden Kapitel zunächst die datenschutzrechtlichen Bedingungen einer Weitergabe von Mikrodaten erläutert. In Abschnitt 3 wird auf die Gefahr einer Reidentifikation von Merkmalsträgern hingewiesen und zusammenfassend Reidentifikationstechniken vorgestellt. Ein Überblick der gängigen Anonymisierungsmaßnahmen wird im anschließenden Kapitel gegeben. Die Methoden, die zur Anonymisierung der Daten des Mannheimer Innovationspanels verwendet werden, werden darauffolgend genannt. Mit Hilfe einer Simulation wird am Beispiel einer Überlagerung mit einem Zufallsfehler eine Korrekturmöglichkeit des OLS-Schätzers demonstriert.

2 Anonymisierung von Mikrodaten als Beitrag zum Datenschutz

Für Betriebsdaten besteht im Vergleich zu Personendaten ein wesentlich höheres Reidentifikationsrisiko (vgl. z.B. Brand, 2000). Eine Reidentifikation ist dann erfolgt, wenn einem Element eines anonymisierten Mikrodatenfiles genau eine Beobachtung aus anderen zugänglichen Datensätzen, so genanntes Zusatzwissen, zugeordnet werden kann. Zum Abgleich der beiden vorliegenden Datensätze ist das Vorhandensein von Überschneidungsmerkmalen (Schlüsselvariablen) zwingend erforderlich, d.h. Informationen, die sowohl im Zusatzwissen als auch in dem anzugreifenden Mikrodatenfile zu finden sind. Das Reidentifikationsrisiko steigt, wenn dem Datenangreifer mit Sicherheit bekannt ist, dass eine Person bzw. ein Unternehmen/Betrieb an der Erhebung, die dem Mirkrodatenfile zu Grunde liegt, teilgenommen hat.

Das Vorliegen von Zusatzwissen mit Überschneidungsmerkmalen ist im Fall von Firmendaten im Gegensatz zu Personendaten wahrscheinlicher. Informationen über Umsätze, Beschäftigte, Firmensitz etc. (z.B. durch CREDITREFORM[1], Hoppenstedt-Datenbank)

[1]CREDITREFORM ist die größte deutsche Kreditauskunftei, die über eine umfassende Datenbank zu deutschen Unternehmen verfügt.

sind im Allgemeinen öffentlich zugänglich. Daher besteht eine erhöhte Gefahr der Reidentifikation bei Unternehmensdaten. Maßnahmen, die den Schutzanforderungen bei Personendaten genügen, sind hier unzureichend. Die für personenbezogene Umfragedaten allgemein anerkannten Richtlinien bei der faktischen Anonymisierung können demnach nicht direkt auf das Maskierungsproblem bei Unternehmensdaten übertragen werden. Weiterreichende Schutzmaßnahmen sind erforderlich.

Der Vertrauensschutz, der den Unternehmen zugesichert wird, steht offensichtlich im Konflikt zu dem Nutzen, den die Daten für wissenschaftliche Zwecke stiften sollen. Bei der Datenanonymisierung wird ein Datensatz zwangsläufig verfälscht. Diese Verzerrungen führen dabei zu einer mehr oder weniger starken Beeeinflussung des Analysepotenzials. So muss die Daten erhebende Einrichtung bzw. Anonymisierungsstelle einen Weg finden, sowohl die notwendigen Schutzvorkehrungen vor potenziellen Datenangreifern zu gewährleisten, als auch die wissenschaftliche Verwertbarkeit der Daten zu bewahren.

Dabei hängt der Grad der Beeinträchtigung der Analysemöglichkeiten eines Datensatzes erheblich von der Art der vorgesehenen Untersuchungen und den Themengebieten ab. Um ein Beispiel zu nennen, sei hier der Vergleich von Analysen genannt, die zum einen auf die Betrachtung gesamtwirtschaftlicher Entwicklungen von Aggregaten abzielen und zum anderen betriebliche Prozesse abbilden wollen. Es wurde vielfach gezeigt (z.B. Brand, 2000), dass für kleine und mittlere Betriebe eine sinnvolle Anonymisierung möglich ist, d.h., die Maskierung der Daten beeinflusst nur unwesentlich das Analysepotenzial. Damit ist eine Darstellung von Prozessen von Betrieben mit anonymisierten Daten sehr gut vorstellbar, da hier „extreme“ Gruppen, beispielsweise Großunternehmen, keine Rolle spielen. Bei der Betrachtung von gesamtwirtschaftlichen Aggregaten aber muss eine repräsentative Stichprobe vorliegen.

Die Nutzungsmöglichkeiten von maskierten Datensätzen hängt folglich sehr stark von dem Forschungsthema und den Analysetechniken ab. Daher konnten bisher keine allgemeinen Regeln für ein optimales Anonymisierungsverfahren aufgestellt werden. Es ist dem Daten erfassenden und -weitergebenden Institut im Weiteren aus Kostengründen i.d.R. nicht möglich, Datensätze speziell für bestimmte Forschungsvorhaben aufzubereiten. Es muss einen Kompromiss eingehen, der dem Großteil der Datennutzer ein bestmögliches Analysepotenzial gewährleistet, ohne dabei den zugesicherten Vertrauensschutz für die Unternehmen zu gefährden.

2.1 Rechtliche Vorschriften

Einzelangaben (Mikrodaten), die durch Befragungen im Rahmen der amtlichen Statistik, d.h. vom Statistischen Bundesamt und den statistischen Landesämtern, erhoben werden, sind in der Regel nicht frei zugänglich. Die Erhebungen unterliegen dem Bundesstatistikgesetz. Dieses schreibt die Nichtweitergabe von Mikrodaten vor, wenn dadurch eine Identifizierung eines Befragten möglich wird. Wenn der Datensatz an jedermann weitergegeben werden soll, müssen die Angaben „absolut anonym“ sein (Public-Use-File). Soll der Datensatz ausschließlich für wissenschaftliche Zwecke genutzt werden, genügt die „faktische Anonymität“ (§16(6) BStatG) (Scientific-Use-File). Im deutschen Recht existieren drei verschiedene Formen von Anonymität: die formale, die faktische und die absolute Anonymität. Enthält ein Datensatz weder Name noch Adresse der Befragten, spricht man von formaler Anonymität. Faktische Anonymität liegt vor, wenn ein Datensatz derart

verändert wurde, dass eine Reidentifizierung von Befragten auf Grund der herausgegebenen Informationen nur mit einem unverhältnismäßig großem Aufwand an Zeit, Kosten und Arbeitskraft möglich wird (§3(7) BDSG). Eine Enthüllung der Identität kann also mit diesem Konzept nicht mit absoluter Sicherheit ausgeschlossen werden. Wenn absolut ausgeschlossen werden kann, dass Einzelangaben auf bestimmte Personen bzw. Unternehmen/Betriebe zurückgeführt werden können, wird von „absoluter Anonymität“ gesprochen.

Von privaten Forschungsinstituten durchgeführte Befragungen unterliegen nicht den Regeln des Bundesstatistikgesetzes. Bei der Weitergabe von Mikrodaten ist allerdings das Bundesdatenschutzgesetz (BDSG) zu beachten. In §3 BDSG wird die vertrauliche Behandlung von personengebundenen Daten festgelegt. Unternehmensbefragungen unterliegen diesen Regeln nicht direkt. Wenn den befragten Firmen Vertraulichkeitszusagen gegeben werden, sind die Daten aber ebenfalls nicht frei zugänglich. Denn die Zusicherung einer vertraulichen Behandlung der Angaben stellt eine Selbstverpflichtung dar und eine Weitergabe der Einzelangaben darf lediglich in anonymisierter Form erfolgen.

2.2 Vertrauensschutz

Um die Bereitschaft des befragten Unternehmens unter Umständen sensible Informationen weiterzugeben zu erhöhen, ist eine Vertraulichkeitszusicherung dringend erforderlich. Nur so ist eine hohe Qualität der Erhebung zu gewährleisten. Am ZEW werden die Angaben von Unternehmen im Rahmen von Unternehmensbefragungen nur für wissenschaftliche Zwecke an externe Datennutzer weitergeleitet. Die Wissenschaftler verpflichten sich im Rahmen eines Datenüberlassungsvertrages, die anonymisierten Daten ausschließlich im Rahmen von wissenschaftlichen, nicht gewerblichen und sonstigen wirtschaftlichen Zwecken zu nutzen.

Es wird als ethische Verpflichtung der Statistik betrachtet, dass Daten vertraulich zu behandeln seien (American Statistical Association, 1983). Aus diesem Grund wurde im ICC/ESOMAR-Internationaler Kodex für die Praxis der Markt- und Sozialforschung[2] verbindlich vorgeschrieben, dass Daten jeglicher Informanten, d.h. auch juristischer Personen, dem Vertrauensschutz unterliegen. In dieser Regelung wird eine Anonymisierung von Einzelangaben vor der Weitergabe an Dritte verbindlich vorgesehen: „Informationen über natürliche und juristische Personen sowie Personengruppen dürfen in der Markt- und Sozialforschung nur in anonymisierter Form übermittelt werden. Dementsprechend dürfen insbesondere Markt- und Sozialforschungsinstitute Daten nur in anonymisierter Form an ihre Auftraggeber (Forschungsinstitute eingeschlossen) weitergeben.“(Koschnick, 1995, S. 452).

3 Reidentifikationsversuche

Zur Reidentifikation anonymisierter Datensätze eines Mikrodatenfiles benötigt ein so genannter Datenangreifer ein Identifikationsfile als Zusatzwissen, wobei die Identität der

[2]Ein in enger Zusammenarbeit zwischen der European Society for Marketing and Opinion Research (ESOMAR) und der Internationalen Handelskammer (ICC) erarbeiteter Code für ethische Grundsätze und Normen im Bereich der Marketing- und Sozialforschung (Koschnick, 1995, S. 451 ff.).

Merkmalsträger dieses Identifikationsfiles dem Datenangreifer bekannt sein muss. Der Mikrodatensatz und das Identifikationsfile müssen gemeinsame Merkmale, sogenannte Überschneidungsmerkmale, besitzen. Ein Reidentifikationsversuch besteht nun darin, die Ausprägungen der einzelnen Datensätze beider Files zu vergleichen und auf Grund der Übereinstimmungen der Werte der Überschneidungsmerkmale die Datensätze einander zuzuordnen. Ein Datensatz gilt als identifiziert, wenn eine eindeutige Zuordnung gefunden werden kann. Das Motiv eines Reidentifikationsversuchs könnte darin bestehen, das resultierende Zusatzwissen für wissenschaftliche oder andere Zwecke zu nutzen.

Für Personendaten wurden Verfahren zur Schätzung des Reidentifikationsrisikos, d.h. der Zuordnungswahrscheinlichkeit, von Datensätzen in Mikrodatenfiles entwickelt. Zu unterscheiden sind der bayesianische Ansatz (z.B. Paaß und Wauschkuhn, 1984, Paaß, 1986, 1987, 1988) und ein vereinfachtes Konzept zur Evaluation von Reidentifikationsrisiken, der Uniqueness-Ansatz (z.B. Willenborg und de Waal, 1996). Auf der Grundlage der Bestimmung von Reidentifikationsrisiken wurden Anonymisierungsregeln für Personendaten entwickelt. Für Firmendaten liegen nur wenige Studien zur Bestimmung des Reidentifikationsrisikos und zur Entwicklung von Anonymisierungsregeln vor (z.B. Brand, 2000).

Auch ohne eine eindeutige Zuordnung kann eine Enthüllung von Informationen vorliegen. Wenn z.B. alle Befragten in einer Umfrage für eine bestimmte sensible Information die gleichen Angaben machen, ist allein aus dem Wissen, dass ein Merkmalsträger an der Erhebung teilgenommen hat, die sensible Information ableitbar. Eine ähnliche Situation liegt vor, wenn einem Datenangreifer bekannt ist, dass eine Beobachtung zu einer Gruppe gehört, die durch bestimmte Variablenausprägungen gekennzeichnet ist und alle Befragten in dieser Gruppe dieselbe sensible Information aufweisen.

4 Anonymisierungsmaßnahmen

Bei der Auswahl von Anonymisierungsmaßnahmen ist insbesondere darauf zu achten, dass einfache Abgleichtechniken durch die gewählte Methode gestört werden (Müller et al., 1991, S. 388-390). Hierbei handelt es sich um Techniken, die unterhalb der Schwelle des Kriteriums der Unverhältnismäßigkeit für die faktische Anonymität liegen. Unter Voraussetzung dieser Bedingung soll der statistische Gehalt der Daten durch die Maßnahme so wenig wie möglich beeinträchtigt werden. Dies bedeutet zum einen, den Verlust an Individualinformationen möglichst gering zu halten und zum anderen, die Beziehungen zwischen den Variablen in Form von Korrelationen, Kovarianzen u.a., unverändert zu erhalten. Anonymisierungsmaßnahmen lassen sich grob in folgende Gruppen unterteilen:

- Methoden, bei denen die Daten verändert aber nicht verfälscht werden. Die Merkmalsausprägungen werden vergröbert und dadurch entsteht ein mehr oder weniger hoher Informationsverlust.
- Maßnahmen, die keine direkte Modifikation der Daten erfordern, wie die Weitergabe von „veralteten“ Daten und Ziehung von zufälligen Substichproben.
- Methoden, die auf dem bewussten Einbringen von falschen Angaben beruhen.

Bei der Auswahl einer dieser Methoden sollten die anzuwendenden Analysemethoden bzw. Forschungsinhalte der Datennutzer möglichst berücksichtigt werden, da die Maßnahmen in unterschiedlicher Art und Weise den Informationsgehalt der Daten beeinflussen.

4.1 Vergröberung von Merkmalen

Die Vergröberung von Merkmalsausprägungen kann wesentlich zur Verringerung des Reidentifikationsrisikos beitragen, ohne die Daten zu verfälschen. Bei regionaler und/oder sachlicher Tiefengliederung (z.B. Brancheneinteilung) eines Mikrodatenfiles oder schwach besetzten Ausprägungen eines Merkmals ist das Risiko einer Reidentifikation eines Merkmalsträgers besonders groß. Durch schwach besetzte Zellen ist es möglich Einzelfälle auszugrenzen und gezielt Zusatzwissen, z.B. bei starker regionaler Untergliederung, zu beschaffen, so dass ein einfacher Abgleich der Informationen möglich wird. Zur Verringerung des Reidentifikationsrisikos kann eine Vergröberung der Merkmalsausprägungen dienen, die gleichzeitig die Daten nicht verfälscht (Müller et al., 1991, S. 394-397). Differenziert ausgewiesene Merkmale können in Kategorien bzw. Intervallen zusammengefasst werden, so dass sich eine gleichbleibende Anzahl von Fällen auf weniger Ausprägungen beziehen. Die Wahrscheinlichkeit von einzigartigen Ausprägungskombinationen sinkt, während die Zahl von statistischen Doppelgängern steigt.

Der Verlust an Informationen ist umso größer, je gröber die Einteilung der Kategorien vorgenommen wird. Die Vergröberung der erhobenen Merkmale durch die Zusammenfassung zu Intervallen bzw. Gruppen beeinflusst deskriptive Statistiken erheblich. Mittelwerte und Varianzen sowie Kovarianzen bleiben i.d.R. nicht erhalten. Im Grenzfall kann die geringere Differenzierung der Merkmalsausprägungen zu einer essentiellen Veränderung der Beziehung zweier Variablen führen und damit die Analysefähigkeit des Datensatzes stark beeinflussen. Bei kausalanalytischen Ansätzen können aber durch die Verwendung von Modellen zur Analyse von gruppierten Daten konsistente Schätzergebnisse der Modellparameter erzielt werden (Grouped data-Modelle) (z.B. Ronning, 1983).

Eine zu grobe Klassenbildung wirkt sich besonders negativ auf Längsschnittsanalysen aus, wenn minimale aber bedeutsame Veränderungen nicht zu einem Wechsel der Klassen führen (Krupp und Preißl, 1989, S. 125).

Die Beeinträchtigung des Analysepotenzials von Datensätzen lässt sich begrenzen, wenn die Klassenbildung unter Berücksichtigung von forschungsrelevanten Aspekten, falls diese bekannt sind, vorgenommen wird.

Statt der Bildung von Ausprägungskategorien reicht oft auch eine Stutzung der Extremwerte eines Merkmals (Censoring-Verfahren), anhand derer ein Merkmalsträger u.U. identifiziert werden kann. Die Merkmalsausprägungen werden bei Überschreitung einer vorher festgesetzten Schwelle, die eine Gefahr der Reidentifikation darstellt, zu diesem Schwellenwert zusammengefasst, gestutzt. Zu denken ist hier z.B. an extrem hohe FuE-Intensitäten von Unternehmen in Unternehmensdatensätzen wie dem Mannheimer Innovationspanel[3], anhand derer Firmen von Konkurrenzunternehmen erkannt werden.

Auch bei der Verwendung von Censoring-Verfahren können Mittelwerte und Varianzen nicht mehr erwartungstreu mit den anonymisierten Daten geschätzt werden (z.B. Stange, 1970, S. 81 f.). Da sich das wissenschaftliche Interesse in der Regel aber nicht auf seltene Fälle richtet, wirkt sich diese Maßnahme nicht besonders schädlich aus. Bei der Angabe der Ober- bzw. Untergrenze einer Stutzung kann der Informationsverlust gemindert werden. Im Weiteren eignen sich zur konsistenten Schätzung von Modellparametern Modelle für zensierte Daten (Tobit-Modelle) (z.B. Maddala, 1983). Die Censoring-Technik beinhaltet

[3] Vgl. Anhang A.1.

einen geringeren Eingriff in die Aussagekraft eines Datensatzes als die Klassenbildung von Informationen.

4.2 Maßnahmen, die keine direkte Datenmodifikation vorsehen

Bei diesen Maßnahmen handelt es sich um eine Unterdrückung von Informationen, mit denen ein besonders hohes Deanonymisierungsrisiko verbunden ist (Müller et al., 1991, S. 398-402). Eine einfache Methode ist das Entfernen von Variablen. Hierunter fallen vor allem personen- bzw. unternehmensbezogene Angaben, wie Name und Anschrift. Aber auch Merkmale mit stark differenzierten Ausprägungen oder extremer Verteilung der Häufigkeiten oder Merkmale, für die Zusatzwissen besonders leicht zu beschaffen ist, kommen in Frage.

Der Informationsverlust kann durch die Unterdrückung von Merkmalen unter Umständen erheblich sein. Bei der Parameterschätzung von Regressionsmodellen tritt eine Fehlspezifikation auf, wenn erklärende Variablen fehlen (vgl z.B. Frohn, 1980, S. 71 ff.) . Diese Methode stellt demnach eine gravierende Einschränkung des Analysepotentials dar und sollte deshalb auf wenige Variablen beschränkt bleiben.

Eine Abwandlung der Entfernung von Merkmalen stellt die Neukonstruktion von Variablen dar. Statt der Originalvariablen tauchen deren Linearkombinationen im anonymisierten Datensatz auf. Ist die Verwendung der neuen Variablen für ökonomische Analysen hinreichend, ergeben sich durch diesen Eingriff keine Probleme. Wenn die disaggregierten Informationen notwendig sind für die angestrebte Untersuchung, sind die Regressionsmodelle wie bei der Entfernung von Informationen fehlspezifiziert.

Eine weitere Methode zur Unterdrückung von Informationen besteht in der Entfernung einzelner diskreter Werte (local suppression) aus dem Datensatz. Dabei werden bei Beobachtungen mit in der Stichprobe sehr seltenen oder einzigartigen Ausprägungskombinationen der Variablen die Ausprägungen einer oder mehrerer Variablen durch einen fehlenden Wert ersetzt. Bei diesem Verfahren ist ebenfalls mit Veränderungen der inferenzstatistischen Eigenschaften und verzerrten Koeffizientenschätzungen in Regressionen zu rechnen (Willenborg und de Waal, 1998).

Eine andere Schutzmaßnahme besteht in der Weitergabe von veralteten Daten. Dies ist zwar keine Anonymisierungsmaßnahme im eigentlichen Sinne, schützt aber die Merkmalsträger insofern, dass potenzielle Angreifer nur an aktuellen Informationen interessiert sind bzw. die Reidentifizierung mit alten Angaben nicht mehr möglich ist. Wenn der Abstand zwischen Erhebung und Freigabe der Daten groß genug ist, ist das aktuelle Zusatzwissen u.U. nicht mehr kompatibel mit dem Datensatz (Südfeld, 1987). Dies stellt allerdings auch eine entscheidenden Nachteil für die Wissenschaft dar, wenn wissenschaftliche Analysen den aktuellen gesellschaftlichen bzw. wirtschaftlichen Wandel abbilden wollen. Eine Politikberatung kann auf der Basis von veralteten Informationen nicht stattfinden, geschweige denn Prognosen angestellt werden.

Eine weitere Methode zur Verringerung des Deanonymisierungsrisikos besteht in der Ziehung und Weitergabe von Substichproben des vorliegenden Mikrodatenfiles. Für eine Reidentifikation eines Merkmalsträgers in einer Substichprobe muss dieser sowohl im Mikrodatenfile als auch im Identifikationsfile enthalten sein. Die Erfolgswahrscheinlichkeit entspricht dem Auswahlsatz der Stichprobe. Diese Schutzwirkung besteht nicht mehr,

wenn dem Angreifer bekannt ist, dass eine Person an der Stichprobenerhebung teilgenommen hat bzw. wenn er weiß, dass sie in der Substichprobe enthalten ist. Im Allgemeinen ist davon auszugehen, dass derartige Informationen nicht zur Verfügung stehen. Dann besteht aber noch die Möglichkeit bei Vorlage der Grundgesamtheit statistische Doppelgänger zu identifizieren und somit die Wahrscheinlichkeit eines Treffers in der Stichprobe abschätzen zu können.

Der sich aus dem potenziellen Wissen eines Angreifers über das Antwortverhalten ergebende Risikofaktor kann durch das zufällige Ziehen von Stichproben eingeschränkt werden. Diese Maßnahme bildet einen wichtigen Schutz vor Reidentifikationsversuchen. Das Analysepotenzial der Daten bei Stichprobenziehungen bleibt insofern gewahrt, als dass keine Datenmodifikation vorgenommen wird. Duch die Reduzierung der Stichprobe erhöhen sich allerdings die Schätzfehler (z.B. Greene, 1997). Beim Vorhandensein ohnehin schwach besetzter Zellen kann eine Stichprobenziehung zu deutlichen Verzerrungen führen.

Problematisch wird es, wenn zur empirischen Beurteilung der ökonomischen Fragestellung die ausgeschlossenen Datenmengen bei nicht zufälliger Stichprobenziehung einen wesentlichen Beitrag liefern. Bei der Anwendung auf Betriebsdaten spielt dieses Verfahren eine entscheidende Rolle, da Großbetriebe einem hohen Identifikationsrisiko unterliegen. Werden Informationen über große Unternehmen vor der Weitergabe der Daten gestrichen, führt das zur Veränderung der Ergebnisse inferenzstatistischer Analysen (Pursey, 1999, Barnett und Lewis, 1994). Werden jedoch bei Untersuchungen Extremwerte, sogenannte „Ausreißer“, durch den Datennutzer aus statistischen Erwägungen ohnehin entfernt, spielt eine derartige Manipulation praktisch keine Rolle für die Analysefähigkeit der Daten.

4.3 Einbringen von falschen Angaben in das Datenmaterial

Eine Möglichkeit die Kompatibilität zwischen Mikrodatenfile und potenziellem Zusatzwissen zu reduzieren besteht darin, das originale Datenmaterial des Mikrodatenfiles mit „falschen“ Angaben zu überlagern. Die Einführung von Zufallsfehlern (Zufallsrauschen) in das Datenmaterial stellt eine Variante dieser auch als „perturbation“ bzw. „contamination“ bezeichneten Anonymisierungstechniken dar (Blien et al., 1991, Skinner et al., 1990, Marsh et al., 1991). Folgende Ansätze sind zu nennen:

- Das Einschätzen einzelner Werte (Blanking und Imputation).
- Das zufällige Vertauschen von benachbarten Merkmalsausprägungen bei kontinuierlichen Merkmalen (Data-Swapping).
- Die Addition von oder Multiplikation mit Zufallsvariablen bei kontinuierlichen Merkmalen.[4]
- Das Ersetzen von Variablenausprägungen durch vorher ermittelte Mittelwerte von jeweils ähnlichen Datensätzen (Mikroaggregation).
- Die Erzeugung von synthetischen Datensätzen, deren Elemente nicht mehr auf die Ausprägungen der Merkmale des Ausgangsdatensatzes zurückgeführt werden können (Resampling).

[4]Bei der Anonymisierung der Daten des Mannheimer Innovationspanels wird eine gleichverteilte Zufallsvariable verwendet.

Eine Reidentifikation eines Merkmalsträgers durch einfache Abgleichtechniken ist bei derartig modifizierten Daten kaum mehr möglich.

Imputationsverfahren beruhen auf dem Austausch von Originalangaben durch geschätzte Werte (z.B. Rubin, 1993). Sie werden auch im Rahmen von Nonresponse-Analysen zur Ersetzung von fehlenden Werten verwendet (z.B. Rubin, 1987). Die Güte dieser Verfahren ist abhängig von den Korrelationen zwischen den Variablen (Kovar, Withridge, 1995). Bei schwachen Abhängigkeiten zwischen den Merkmalen ist die Anwendung nicht sinnvoll.

Bei der Durchführung von Data-Swaps werden eine Reihe von zufälligen Vertauschungen von Werten vorgenommen. Dies geschieht in der Regel innerhalb von Gruppen mit bestimmten Ausprägungskombinationen für einige besonders bedeutsame Variablen (Reiss, 1980). Die univariaten Verteilungen bleiben erhalten, während multivariate Verteilungen und Korrelationen zwischen den Variablen nicht konsistent mit dem anonymisierten Datensatz geschätzt werden können. Die Beeinträchtigung der Analysemöglichkeiten lässt sich verringern, wenn die Vertauschungen unter Nebenbedingungen stattfinden, die den Erhalt von Varianz-Kovarianzmatrizen zwischen Variablen sicherstellen (Kim und Winkler, 1995, 1997). Ein solches Vorgehen ist allerdings sehr rechenintensiv.

Zur Überlagerung von zu maskierenden Variablen mit Zufallsfehlern sind verschiedene Methoden zu unterscheiden. Die einfache oder naive Überlagerung beruht auf der Addition einer normalverteilten Störgröße mit einem Erwartungswert von null zu den Werten stetiger Variablen (z.B. Spruill, 1983). Die Varianzen der Störgrößen sollen proportional zu den Varianzen der Merkmale sein. Im Weiteren sollten die Überlagerungen der unterschiedlichen Variablen unabhängig voneinander sein. Für Betriebsdaten kann schon mit „kleinen“ Überlagerungen eine recht gute Schutzwirkung erzielt werden, so dass ökonomische Analysen weiterhin möglich sind. Wenn die Anzahl der Schlüsselvariablen allerdings eine kritische Zahl überschreitet (mehr als vier bis sechs), müssen größere Überlagerungen vorgenommen werden. Dadurch wird die Analysefähigkeit unter Umständen erheblich eingeschränkt (Spruill, 1983).

Diese Methode erhält zwar die Erwartungswerte der maskierten Variablen nicht aber deren Varianzen und die Korrelationen mit anderen Merkmalen. Dies hat natürlich auch Auswirkungen auf die Schätzung von Parametern in Regressionsmodellen. Ist das Ausmaß der Verzerrung der Varianzen allerdings bekannt, können Schätzungen von Korrelationen und Modellparametern entsprechend korrigiert werden. Nicht korrigierbar ist allerdings, dass die Verteilungen der maskierten Variablen nicht mit denen der Originalvariablen übereinstimmen, sofern es sich nicht um eine normalverteilte Größe handelt (Brand, 2000, 183 ff.).

Beim Ansatz von Kim (1986) kann im Gegensatz zur naiven Überlagerung durch zusätzliche Transformation erreicht werden, dass die Stichproben-Kovarianzmatrix der Originalvariablen konsistent mit den anonymisierten Daten geschätzt werden kann. Die univariaten Verteilungen der Originalvariablen bleiben aber im Allgemeinen (s.o.) auch bei dieser Methode nicht erhalten. Aufgrund der Struktur der Transformation können nur stetige Variablen mit diesem Ansatz maskiert werden. Die Schutzwirkung nimmt mit zunehmendem Überlagerungsfaktor bis zu einem kritischen Wert zu. Danach kann keine zusätzliche Schutzwirkung mehr erzielt werden (Moore, 1996).

Ausgehend von dem Ansatz von Kim schlägt Sullivan (1989) ein Maskierungsverfahren vor, das zusätzlich auf diskrete Variablen angewendet werden kann. Diese Methode lässt

zu, dass auch die univariaten Verteilungen der zu maskierenden Variablen erhalten bleiben. Der Ansatz besteht aus einem mehrstufigen, aufwendigen Verfahren, einer Kombination von Transformationen mit einer additiven Überlagerung. Während der Durchführung wird überprüft, ob die Schutzwirkung der Anonymisierung ausreichend ist.

Hwang (1986) betrachtet den Fall, dass Variablen mit multiplikativen Zufallsfehlern überlagert werden. Dabei wird angenommen, dass die Störgrößen mit einem Erwartungswert von eins unabhängig identisch verteilt sind. Die maskierten Variablen sind somit weiterhin erwartungstreu, aber deren Varianzen können nicht mehr konsistent geschätzt werden. Parameterschätzungen von Regressionsmodellen sind verzerrt. Hwang (1986) zeigt, wie eine konsistente Schätzung von Koeffizienten eines linearen Regressionsmodells mit anonymisierten Daten durch eine Korrektur erzielt werden kann.

Die Methode der Mikroaggregation beruht auf der Ermittlung von Mittelwerten von jeweils ähnlichen Datensätzen und die Ersetzung der Variablen durch eben diese Werte (Cox et al., 1985, Paaß, Wauschkuhn, 1985). Die Gesamtmittelwerte des Original- und des anonymisierten Datensatzes sind weitestgehend identisch, während die Standardabweichungen der Merkmalsausprägungen verringert werden. Auch die Korrelationen und Schätzungen von Modellparametern werden systematisch verzerrt. Das Ausmaß des Bias hängt dabei davon ab, wie hoch die Korrelationen der Variablen des Ausgangsdatensatzes sind. Bei Betriebsdaten sind zwischen den Variablen in vielen Fällen geringe Korrelationen zu beobachten. Daher sollte die Mikroaggregation zur Anonymisierung von Firmendaten nur eingeschränkt verwendet werden.

Beim Resampling wird ein synthetischer Datensatz generiert, so dass dessen Elemente nicht mehr auf die Ausprägungen der Merkmale des Ausgangsdatensatzes zurückgeführt werden können. Dazu wird die multivariate Dichte des gesamten Datensatzes mittels nichtparametrischer Verfahren konsistent geschätzt (z.B. Fienberg et al., 1998). Daran anschließend wird eine Stichprobe generiert, die der geschätzten Dichte folgt. Ergebnisse inferenzstatistischer Analysen und Koeffizientenschätzungen von Regressionsmodellen mit den anonymisierten Daten weichen daher nicht systematisch von gleichen Analysen mit dem Originaldatensatz ab. Nur die Standardabweichungen geschätzter Regressionskoeffizienten sollten korregiert werden[5]. Eine praktische Anwendung ist nur dann möglich, wenn eine ausreichende Zahl an Beobachtungen vorliegt, um die multivariate Dichte hinreichend genau schätzen zu können.

5 ZEW-Methoden

5.1 Überblick

Zur Erhaltung des den Unternehmen zugesagten Vertrauensschutzes bezüglich ihrer Angaben, werden zur externen Nutzung der Daten der Mannheimer Innovationspanels im verarbeitenden Gewerbe und Bergbau sowie im Dienstleistungssektor für wissenschaftliche Zwecke die Mikrodaten anonymisiert. Dazu werden abhängig vom Variablentyp unterschiedliche Verfahren eingesetzt. Diese Verfahren verhindern, dass ein Unternehmen anhand seiner Angaben identifiziert werden kann. Dies impliziert eine „Vernichtung“ von Informationen im Datensatz. Diese Verfahren wurden nach den Kriterien ausgewählt, dass

[5] Geeignet erweisen sich Bootstrap-Verfahren (z.B. Fienberg, 1997).

im Datensatz ein Maximum an Informationen verbleibt, gleichzeitig aber eine eindeutige Zuordnung eines bestimmten Wertes zur Originalangabe der Unternehmen nicht mehr durchgeführt werden kann.

Streichung von Beobachtungen

Aus dem Datensatz werden alle Unternehmen entfernt, die nicht für das Unternehmen sondern für ihre gesamte Unternehmensgruppe, den Konzern, geantwortet haben. Informationen von leicht zu identifizierenden Unternehmen werden dadurch zurückgehalten.

Streichung von Merkmalen

Einige Angaben sind im Datensatz nicht enthalten. Dazu gehören z.B. Angaben über Anzahl der Teilzeit- und Auslandsbeschäftigten sowie Ortsangaben.

Konstruktion neuer Variablen

In einigen Erhebungen der Mannheimer Innovationspanels wurden Informationen zusammengefasst. In der Erhebung 1993 wurden z.B. bei den Fragen nach dem Technologieerwerb und Technologietransfer in der Fragebogenversion der Produktionsunternehmen mit mindestens 50 Beschäftigten nach der geographischen Zuordnung des Erwerbs/Transfers entsprechend der Einteilung - BRD, EG, Nicht EG, USA, Japan und sonstige Länder außerhalb Europas - gefragt. Im anonymisierten Datensatz wird jedoch lediglich die Einteilung - BRD und Ausland - ausgewiesen. Ähnlich wurde bei der geographischen Zuordnung der FuE-Kooperationspartner vorgegangen: Im anonymisierten Datensatz wird ausgewiesen, ob der Kooperationspartner aus der Region, aus der BRD oder aus dem Ausland stammt.

Überlagerung mit einem multiplikativen Zufallsfehler

Dieses Verfahren besteht darin, dass der von einem Unternehmen angegebene Wert mit einer gleichverteilten Zufallszahl multipliziert wird, die in einem Intervall zwischen 0,5 und 1,5 liegt. Ihr Erwartungswert beträgt somit 1. Diese Zufallszahl stellt eine unternehmensspezifische Konstante dar, d.h., jede so randomisierte Variable wird mit der gleichen Zahl multipliziert. Dies garantiert, dass das Unternehmen nicht mehr an den von ihm angegebenen absoluten Zahlenwerten erkannt werden kann. Die Methode wird für die Umsatzangaben und die Anzahl der Beschäftigten verwendet. Der Quotient dieser beiden Variablen (Umsatz pro Vollzeitbeschäftigten) bleibt aber unverändert.

Ausweis von Intensitäten und Quoten

Anstatt alle absoluten Größen der Originaldaten mit einer konstanten Zufallszahl zu multiplizieren, werden die Größen in Relation zum Umsatz oder den Beschäftigten ausgewiesen. Diese Intensitäten oder Quoten sind dann im anonymisierten Datensatz aufgeführt. So werden beispielsweise die gesamten Innovationsaufwendungen, die FuE-Aufwendungen,

der Auslandsumsatz, die Personalkosten und der Materialaufwand in Relation zum Umsatz, die Anzahl der FuE-Beschäftigten und der Beschäftigten insgesamt nach Qualifikationsstruktur in Relation zu den Beschäftigten ausgewiesen. Die Nutzer können bei Bedarf durch Umrechnung wieder (randomisierte) absolute Größen bzw. weitere Intensitäten wie z.B. FuE-Aufwendungen pro Beschäftigten berechnen.

Stutzung von Intensitäten

In Einzelfällen kann es vorkommen, dass Unternehmen „extreme“ Intensitäten aufweisen, z.B. eine FuE-Intensität von 25%. Um zu verhindern, dass Firmen an diesen Intensitäten erkannt werden können, werden diese Extremfälle, die in der Population und in der Stichprobe nur selten anzutreffen sind, gestutzt. Abhängig von der Verteilung der jeweiligen Intensitäten wurde dabei mit unterschiedlichen Obergrenzen gearbeitet. Zum Beispiel liegt die Obergrenze bei der FuE-Intensität (FuE-Aufwendungen/Umsatz) bei 0,15. Weist ein Unternehmen eine FuE-Intensität von 0,25 auf, so wird die FuE-Intensität auf 0,15 gestutzt. Um den Benutzer erkennen zu lassen, ob die jeweilige Variable gestutzt ist, wird eine zusätzliche Variable in den Datensatz aufgenommen, die eine Stutzung anzeigt. Mit Hilfe entsprechender ökonometrischer Verfahren kann damit der Benutzer die Auswirkungen dieser Zensierung in Grenzen halten. Eine Stutzung erfolgt z.B. bei der FuE-Intensität, der Innovationsintensität (Innovationsaufwendungen/Umsatz), der Investitionsintensität (Investitionen/Umsatz) und der Exportquote (Exporte/Umsatz).

Gruppierung

Bei einigen Variablen, z.B. bei Umsatzanteilen mit neuen Produkten, wird in den anonymisierten Datensätzen lediglich angegeben, in welchem Intervall die Merkmalsausprägungen liegen.

5.2 Simulation einer Regressionskorrektur

Im folgenden Abschnitt soll anhand einer Simulation demonstriert werden, wie die Überlagerung mit einem multiplikativen Zufallsfehler innnerhalb eines Regressionsmodells korrigiert werden kann. Die Zufallszahl ist genau wie bei der Anonymisierung der Daten des MIP eine gleichverteilte Zufallsvariable (ZV) in dem Intervall von 0,5 bis 1,5 und wird einem Set von vier generierten Variablen (x_1, x_2, y, u) von jeweils 100 Beobachtungen hinzugefügt.

Die Variablen sind folgendermaßen definiert:

$$x_1 \sim N(1,1), \quad x_2 \sim \chi^2(1), \quad u \sim N(0,1), \quad ZV \sim Uni[0,5;1,5]$$

$$x_1^* = x_1 * ZV$$

$$y = 2 + 2x_1 + 0,5x_2 + u$$

x_1^* ist die durch den Zufallsfehler ZV überlagerte Variable x_1, d.h. die ursprüngliche Information x_1 ist maskiert. Die Verteilung von x_1 bleibt nicht erhalten, obgleich x_1^* weiterhin erwartungstreu ist. Um die Wirkung einer Regressionskorrektur zu veranschaulichen, werden jeweils drei lineare Modelle konstruiert:

1. OLS-Schätzung mit den unverfälschten Variablen:

   $$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

   $$\hat{\beta} = (X'X)^{-1}X'y, \quad X' = (1, x_1, x_2)$$

   Der OLS-Schätzer $\hat{\beta}$ ist konsistent.

2. OLS-Schätzung mit der verfälschten Variablen x_1^*:

   $$y = \beta_0 + \beta_1 x_1^* + \beta_2 x_2$$

   $$\beta^* = (X^{*'}X^*)^{-1}X^{*'}y, \quad X^{*'} = (1, x_1^*, x_2)$$

   Der OLS-Schätzer β^* ist inkonsistent.

3. OLS-Schätzung des Modells mit Korrekturterm M (Hwang, 1986):

   $$\tilde{\beta} = [(X^{*'}X^*) \div M]^{-1}X^{*'}y$$

   $A \div B$ bezeichnet den Hadamard Quotienten (jedes Element (i, j) der Matrix A wird durch das entsprechende Element (i, j) der Matrix B dividiert). Der Korrekturterm M kann leicht berechnet werden, wenn die Verteilung des Zufallsfehlers bekannt ist. In diesem Beispiel ist er gleichverteilt mit $E(ZV) = 1$, $Var(ZV) = \frac{1}{12}$ und $E(ZV'ZV) = 1\frac{1}{12}$. Damit ergibt sich:

   $$M = \begin{pmatrix} 1 & 1 & 1 \\ 1 & E(ZV'ZV) & 1 \\ 1 & 1 & 1 \end{pmatrix}$$

   Der Schätzer $\tilde{\beta}$ ist konsistent.

Um die Wirkungsweise der Korrektur erfassen zu können, werden Monte-Carlo Simulationen (100 Replikationen) der drei Schätzgleichungen durchgeführt. Tabelle 1 stellt die Mittelwerte der OLS-Schätzer, deren Varianzen und die Bestimmtheitsmaße der drei verschiedenen Schätzungen dar. Die mit * gekennzeichneten Werte zeigen signifikante Unterschiede zu den Ergebnissen mit den Originaldaten an.

Beim Vergleich der Regressionskoeffizienten zeigt sich deutlich die Verzerrung der Schätzung mit der maskierten Variablen x_1^*, während t-Tests auf Gleichheit der Mittelwerte der Koeffizienten der Regressionen mit den Originalwerten und der korrigierten Schätzung keine signifikanten Unterschiede feststellen. Für die geschätzten Varianzen lassen sich jedoch keine eindeutigen Schlüsse aus den Testergebnissen ableiten.

Tabelle 1: Monte-Carlo Simulation

Schätzer	Mittelwert	Std.Abw.	Min	Max
β_0 Original	2,002	0,158	1,540	2,469
β_0 Verzerrt*	2,276	0,187	1,744	2,827
β_0 Korrigiert*	1,964	0,196	1,455	2,584
$Var(\beta_0)$ Original	0,026	0,005	0,016	0,037
$Var(\beta_0)$ Verzerrt*	0,039	0,008	0,022	0,058
$Var(\beta_0)$ Korrigiert*	0,044	0,009	0,024	0,069
β_1 Original	1,998	0,109	1,659	2,264
β_1 Verzerrt*	1,733	0,141	1,292	2,110
β_1 Korrigiert*	2,037	0,163	1,570	2,486
$Var(\beta_1)$ Original	0,010	0,002	0,006	0,016
$Var(\beta_1)$ Verzerrt*	0,014	0,003	0,009	0,021
$Var(\beta_1)$ Korrigiert*	0,017	0,004	0,011	0,027
β_2 Original	0,492	0,069	0,338	0,656
β_2 Verzerrt	0,490	0,088	0,253	0,662
β_2 Korrigiert	0,491	0,094	0,216	0,672
$Var(\beta_2)$ Original	0,006	0,002	0,002	0,014
$Var(\beta_2)$ Verzerrt*	0,009	0,003	0,004	0,020
$Var(\beta_2)$ Korrigiert*	0,010	0,004	0,004	0,021
R^2 Original	0,819	0,035	0,707	0,888
R^2 Verzerrt*	0,718	0,047	0,582	0,804
R^2 Korrigiert*	0,829	0,050	0,690	0,930

5.3 Simulation einer Regressionskorrektur am Beispiel des MIP

Die Korrektur der Parameterschätzer im OLS-Modell soll hier anhand eines realen Datensatzes, des MIP des verarbeitenden Gewerbes 1997, demonstriert werden. Zur beispielhaften Darstellung wurde eine Produktionsfunktion sowohl mit den Originaldaten als auch mit einer verzerrten Variablen mit und ohne Korrektur des multiplikativen Zufallsfehlers (siehe Abschnitt 5.2) geschätzt:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6$$

mit:

y - Arbeitsproduktivität

x_1 - FuE-Intensität

x_2 - Beschäftigtengrößenklassen

x_3 - Investitionsintensität

x_4 - Indikator: Chemieindustrie

x_5 - Indikator: Medizin-, Mess-, Steuer- und Regelungstechnik (MMSR)

x_6 - Indikator: Metallverarbeitende Industrie

Die Parameter des Modells werden im ersten Schritt mit den Originaldaten durch die OLS-Methode geschätzt. Im zweiten Schritt wird x_1, die FuE-Intensität, mit der Zufallsvariablen multipliziert und eine erneute Schätzung durchgeführt. Zuletzt wird die duch die Fehlerüberlagerung verzerrte Schätzung durch die Methode von Hwang (1986) korrigiert. In der Monte-Carlo Simulation wird der Zufallsfehler 100 Mal variiert.

In Tabelle 2 sind die durchschnittlichen Parameterschätzer der Simulation zusammengefasst. Die Ergebnisse der Schätzung mit den verfälschten Daten weichen bei einigen Variablen signifikant von den Resultaten der Schätzung mit den ursprünglichen Daten ab. Die korrigierten Werte stimmen dagegen im Mittel signifikant mit den Koeffizienten der Schätzung mit den unverzerrten Daten überein.

Durch die Simulation wird deutlich, wie ein Datennutzer selbst das Analysepotenzial anonymisierter Daten verbesseren kann. Wenn ihm das Ausmaß der Verzerrung durch einen multiplikativen Fehler bekannt ist, kann er mit Hilfe dieser Korrekturmethode konsistente Regressionsergebnisse erzeugen.

Tabelle 2: Ergebnis der Monte-Carlo Simulation mit dem MIP

Schätzer	Mittelwert	Std.Abw.	Min	Max
FuE-Intensität				
β_1 Original	-0,743	-	-	-
β_1 Verzerrt*	-0,685	0,071	-0,864	-0,536
β_1 Korrigiert	-0,756	0,079	-0,956	-0,591
$Var(\beta_1)$ Original	0,053	-	-	-
$Var(\beta_1)$ Verzerrt*	0,046	0,005	0,037	0,060
$Var(\beta_1)$ Korrigiert	0,051	0,006	0,041	0,067
Größenklassen				
β_2 Original	0,044	-	-	-
β_2 Verzerrt	0,043	0,000	0,043	0,044
β_2 Korrigiert	0,044	0,000	0,043	0,044
$Var(\beta_2)$ Original	0,000	-	-	-
$Var(\beta_2)$ Verzerrt*	0,000	0,000	0,000	0,000
$Var(\beta_2)$ Korrigiert*	0,000	0,000	0,000	0,000
Investitionsintensität				
β_3 Original	-0,086	-	-	-
β_3 Verzerrt	-0,087	0,001	-0,089	-0,084
β_3 Korrigiert	-0,086	0,001	-0,089	-0,083
$Var(\beta_3)$ Original	0,002	-	-	-
$Var(\beta_3)$ Verzerrt*	0,002	0,000	0,002	0,002
$Var(\beta_3)$ Korrigiert*	0,002	0,000	0,002	0,002

Schätzer	Mittelwert	Std.Abw.	Min	Max
Indikator: Chemieindustrie				
β_4 Original	0,239	-	-	-
β_4 Verzerrt*	0,238	0,001	0,235	0,241
β_4 Korrigiert	0,239	0,001	0,236	0,242
$Var(\beta_4)$ Original	0,001	-	-	-
$Var(\beta_4)$ Verzerrt	0,001	0,000	0,001	0,001
$Var(\beta_4)$ Korrigiert*	0,001	0,000	0,001	0,001
Indikator: MMSR				
β_5 Original	-0,051	-	-	-
β_5 Verzerrt*	-0,053	0,002	-0,059	-0,047
β_5 Korrigiert	-0,050	0,003	-0,057	-0,044
$Var(\beta_5)$ Original	0,001	-	-	-
$Var(\beta_5)$ Verzerrt*	0,001	0,000	0,001	0,001
$Var(\beta_5)$ Korrigiert	0,001	0,000	0,001	0,001
Indikator: Metallindustrie				
β_6 Original	-0,075	-	-	-
β_6 Verzerrt*	-0,075	0,000	-0,076	-0,074
β_6 Korrigiert	-0,076	0,000	-0,077	-0,074
$Var(\beta_6)$ Original	0,001	-	-	-
$Var(\beta_6)$ Verzerrt*	0,001	0,000	0,001	0,001
$Var(\beta_6)$ Korrigiert*	0,001	0,000	0,001	0,001

Schätzer	Mittelwert	Std.Abw.	Min	Max
Konstante				
β_0 Original	0,217	-	-	-
β_0 Verzerrt*	0,216	0,001	0,214	0,218
β_0 Korrigiert	0,217	0,001	0,215	0,219
$Var(\beta_0)$ Original	0,001	-	-	-
$Var(\beta_0)$ Verzerrt*	0,001	0,000	0,001	0,001
$Var(\beta_0)$ Korrigiert*	0,001	0,000	0,001	0,001
R^2 Original	0,054	-	-	-
R^2 Verzerrt*	0,053	0,001	0,052	0,055
R^2 Korrigiert*	0,059	0,001	0,057	0,060

6 Zusammenfassung

Dieser Aufsatz stellt in einem kurzen Überblick die gängigen Anonymierungsmaßnahmen für Mikrodaten dar und beleuchtet sie im Hinblick auf ihren Beitrag zum Datenschutz und den Erhalt der Analysefähigkeit der Daten. Anschließend werden die Maskierungsmethoden beschrieben, die das ZEW vor der Weitergabe der Daten des Mannheimer Innovationspanels an externe wissenschaftliche Datennutzer anwendet.

Die Problematik, der sich ein Daten erhebendes Institut gegenüber sieht, besteht darin, zum einen den Vertrauensschutz gegenüber den befragten Unternehmen zu wahren und zum anderen die Analysefähigkeit der Daten zu erhalten. Für Betriebsdaten wurden im Gegensatz zu Personendaten keine allgemeinen Anonymisierungsregeln aufgestellt. Dies erscheint auch nicht sinnvoll, da an einen Datensatz in Abhängigkeit vom Forschungsauftrag jeweils unterschiedliche Anforderungen gestellt werden. Die Herleitung eines optimalen Anonymisierungsansatzes und der daraus resultierenden Maskierungsregeln wäre nur unter Nebenbedingungen denkbar und setzt daher die Kenntnis des Untersuchungsgegenstandes für das anonymisierende Institut voraus. Auf Grund des zeitlichen Aufwandes ist ein derartiges Vorgehen nicht praktikabel. Für die Weitergabe von Mikrodatenfiles bedeutet dies, einen Kompromiss zu finden, der die Möglichkeiten zur Datennutzung nicht zu sehr einschränkt.

Durch die Anwendung geeigneter ökonometrischer Methoden können die eingebrachten Fehler, wie oben beispielhaft gezeigt, bis zu einem gewissen Grad korrigiert werden. Die Analysefähigkeit der anonymisierten Daten lässt sich also durch den Datennutzer selbst optimieren. Die Korrektur von Informationsverlusten und Verfälschungen wird in der Regel durch komplexere Analysemethoden erschwert. Ziel zukünftiger Forschung ist demnach, eine Vielzahl an Korrekturmöglichkeiten zur Verfügung zu stellen.

A Anhang

A.1 Die Mannheimer Innovationspanels MIP und MIP-DL

Seit 1993 bzw. 1995 führt das Zentrum für Europäische Wirtschaftsforschung (ZEW) jährlich systematische Erhebungen zum Innovationsverhalten im verarbeitenden Gewerbe und Bergbau und im Dienstleistungssektor durch.

Das MIP und MIP-DL umfassen folgende Themengebiete:

- Allgemeine Angaben,
- Entwicklung und Verbreitung von Innovationsaktivitäten,
- Entwicklung des Innovationserfolgs (Innovationsoutput),
- Bedeutung und Struktur von Innovationshemmnissen,
- Verbreitung und Ergebnisse öffentlicher Förderung,
- Kostenstruktur und Sachanlagen und
- Qualifikation und Weiterbildung.

Im MIP-DL gibt es zusätzlich Angaben zu Informationstechnologien und Kundenbeziehungen.

Das Konzept der Erhebung besteht dabei zum einen aus jährlich wiederkehrenden Fragen, z.B. nach der Durchführung von Produkt- und Prozessinnovationen, den wirtschaftlichen Effekten von Innovationen sowie der Höhe von FuE- und Innovationsaufwendungen. Zum anderen werden im zweijährigen Turnus zusätzliche Schwerpunktthemen untersucht, wie z.B. Formen des Technologietransfers, Informationsquellen, Kooperationen und Hemmnisfaktoren.

A.2 Übersichtstabelle

Kategorie	Methode	Auswirkung				
		Mittelwerte	Varianzen	Regressions-koeffizienten	univariate Verteilungen	multivariate Verteilungen
Vergröberung von Merkmalen	Gruppierung	verzerrt	verzerrt	verzerrt	verzerrt	verzerrt
	Stutzung	verzerrt	verzerrt	verzerrt	verzerrt	verzerrt
Maßnahmen, die keine direkte Datenmodifikation vorsehen	Variablenunterdrückung	-	-	fehlspez.	-	-
	Neukonstruktion von Variablen	-	-	fehlspez.	-	-
	Local Suppression	verzerrt	verzerrt	verzerrt	verzerrt	verzerrt
	Weitergabe veralteter Daten	unverzerrt	unverzerrt	unverzerrt	unverzerrt	unverzerrt
	Substichprobenziehung - zufällig	unverzerrt	unverzerrt	un-/verzerrt	unverzerrt	unverzerrt
	Substichprobenziehung - systematisch	verzerrt	verzerrt	un-/verzerrt	verzerrt	verzerrt

Kategorie	Methode	Auswirkung				
		Mittelwerte	Varianzen	Regressions-koeffizienten	univariate Verteilungen	multivariate Verteilungen
Einbringen von falschen Angaben in das Datenmaterial	Data-Swapping	unverzerrt	unverzerrt	un-/verzerrt	unverzerrt	un-/verzerrt
	Imputation	unverzerrt	geringer	verzerrt	verzerrt	verzerrt
	Resampling	unverzerrt	unverzerrt	unverzerrt	unverzerrt	unverzerrt
	Mikroaggregation	unverzerrt	geringer	verzerrt	verzerrt	verzerrt
	Naive Überlagerung	unverzerrt	verzerrt	verzerrt	verzerrt	verzerrt
	Ansatz von Kim	unverzerrt	unverzerrt	verzerrt	verzerrt	verzerrt
	Ansatz von Sullivan	unverzerrt	unverzerrt	verzerrt	unverzerrt	verzerrt
	Überlagerung mit multiplikativem Zufallsfehler	unverzerrt	verzerrt	verzerrt	verzerrt	verzerrt

B Literatur

Barnett, V. und T. Lewis (1994), Outliers in Statistical Data, 3. Auflage, John Wiley & Sons, Chichester u.a.O.

BMBF (2001), KVI Gutachten, Wege zu einer besseren informationellen Infrastruktur, Gutachten der vom Bundesministerium fr Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik.

Brand, R. (2000), Anonymität von Betriebsdaten, Beiträge zur Arbeitsmarkt- und Berufsforschung, BeitrAB 237, Nürnberg.

Cox, L.H., Johnson, B., McDonald, S.-K., Nelson, D. und V. Vazquez (1985), Confidentiality Issues at the Census Bureau, First Annual Research Conference, Proceedings 1985, U.S. Department of Commerce, Bureau of the Census, 199 ff.

Fienberg, S.E. (1997), Confidentiality and Disclosure Limitation Methodology: Challenges for National Statistics and Statistical Research, Technical Report No. 161, Carnegie Mellon University, Pittsburgh.

Fienberg, S.E., Makov, U.E. und R.J. Steele (1998), Disclosure Limitation Using Perturbation and Related Methods for Categorial Data, Journal of Official Statistics, 14, 485-502.

Frohn, J. (1980), Grundausbildung in Ökonometrie, de Gruyter, Berlin, New York.

Fuller, W.A. (1993), Masking Procedures for Microdata Disclosure Limitation, Journal of Official Statistics, 9, 2, 383-406.

Greene, W.H. (1997), Econometric Analysis, 3. Auflage, Prentice Hall, Englewood Cliffs.

Hwang, J.T. (1986), Multiplicative Errors-in-Variables Models with Application to Recent Data Released by U.S. Department of Energy, Journal of the American Statistical Association, 81, 395, 680-688.

Kim, J. (1986), A Method for Limiting Disclosure in Microdata Based on Random Noise and Transformation, American Statistical Association: Proceedings of the Section on Survey Research Methods, 370-374.

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

Kim, J.J. und W.E. Winkler (1997), Masking Microdata Files, Statistical Research Division RR97/03, U.S. Bureau of the Census, Washington, D.C.

Koschnick, W.J. (1995), Standard-Lexikon für Markt- und Konsumforschung, Band 1: A-K, Saur, München u.a.O., 451-467.

Kovar, J.G. und P.J. Withridge (1985), Imputation of Business Survey Data, in: Cox, B.G., Binder, D.A., Chinnappa, B.N., Christianson, A., Colledge, M.J. und P.S. Kott (Hrsg.), Business Survey Methods, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, New York u.a.O., 403-423.

Krupp, H.-J. und B. Preil (1989), Die Neufassung des BDSG und die wissenschaftliche Forschung, Computer und Recht 5/2, 121 ff.

Maddala, G.S. (1983), Limited-Dependent and Qualitative Variables in Econometrics, Cambridge University Press, Cambridge.

Marsh et al. (1991), The Case for Samples of Anonymized Records from the 1991 Census, Journal of the Royal Statistical Society, 154.

Moore, R.A. (1996), Analysis of the Kim-Winkler Algorithm for Masking Microdata Files - How Much Masking is Necessary and Sufficient? Conjectures for the Development of a Controllable Algorithm, Statistical Research Division RR96/05, U.S. Bureau of the Census, Washington, D.C.

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

Paaß, G. (1986), Identifizierbarkeit und Anonymisierung von Mikrodaten, Allgemeines Statistisches Archiv, 70, 344-367.

Paaß, G. (1987), Reidentifikationsrisiko von Einzelangaben, Statistisches Bundesamt (Hrsg.), Nutzung von anonymisierten Einzelangaben aus der amtlichen Statistik, Bedingungen und Möglichkeiten, Schriftenreihe Forum der Bundesstatistik, Band 5, Kohlhammer, Stuttgart, Mainz, 89-100.

Paaß, G. (1988), Disclosure Risk and Disclosure Avoidance for Microdata, Journal of Business & Economic Statistics, 6, 487-500.

Paaß, G. und U. Wauschkuhn (1984), Datenzugang, Datenschutz und Anonymisierung, Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten, Berichte der Gesellschaft für Mathematik und Datenverarbeitung, Bericht Nr. 148, R. Oldenbourg Verlag, München, Wien.

Pursey, S. (1999), Disclosure Control Methods in the Public Release of a Microdata File of Small Businesses, Beitrag zur: Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, 08.-10. März, Thessaloniki.

Reiss, S.P. (1980), Practical Data-Swapping: The First Steps, Proceedings of the IEEE Symposium on Security and Privacy, 38-43.

Ronning, G. (1991), Mikrökonometrie, Springer, Berlin.

Rubin, D. (1987), Multiple Imputation for Nonresponse in Surveys, Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, USA.

Rubin, D. (1993), Satisfying Confidentiality Constraints Through the Use of Synthetic Multiply-Imputed Microdata, Journal of Official Statistics, 9, 641-468.

Skinner, C.J., Marsh, C., Openshaw, S. und C. Wymer (1990), Disclosure Avoidance for Census Microdata in Great Britain, Annual Research Conference, Proceedings, 1990, U.S. Department of Commerce, Bureau of the Census, 131 ff.

Spruill, N.L. (1983), The Confidentiality and Analytic Usefulness of Masked Business Microdata, American Statistical Association, Proceedings of the Section on Survey Research Methods 1983, Washington, D.C., 602-610.

Spruill, N.L. (1984), Protecting Confidentiality of Business Microdata by Masking, The Public Research Institute, Alexandria, Virginia.

Stange, K. (1970), Angewandte Statistik, Erster Teil, Eindimensionale Probleme, Springer, Berlin, Heidelberg, New York.

Südfeld, E. (1987), Anonymisierungsstandards und generelle Abwicklungsregelungen für Anforderungen nach anonymisierten Einzelangaben im Statistischen Bundesamt, in: Statistisches Bundesamt (Hrsg.), Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik, 1987, Kohlhammer, Stuttgart, Mainz.

Sullivan, G.R. (1989), The Use of Added Error to Avoid Disclosure in Microdata Releases, unveröffentlichte Dissertation, Iowa State University.

Willenborg, L. und T. de Waal (1996), Statistical Disclosure Control in Practice, Springer, New York, Berlin, Heidelberg.