DESTATIS
wissen.nutzen.

STATISTIK UND WISSENSCHAFT

Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten

Gerd Ronning, Roland Sturm, Jörg Höhne, Rainer Lenz,
Martin Rosemann, Michael Scheffler und Daniel Vorgrimler

Unter Mitarbeit von:
Stefan Dittrich, Sandra Gottschalk, Harald Strotmann und Rolf Wiegert

Band 4

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 B – FDZ
Tel.: +49 (0) 611 / 75 26 36
Fax: +49 (0) 611 / 75 39 50
forschungsdatenzentrum@destatis.de

Allgemeine Informationen
zum Datenangebot:

Informationsservice,
Tel.: +49 (0) 611 / 75 24 05
Fax: +49 (0) 611 / 75 33 30
www.destatis.de/kontakt

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

Erschienen im September 2005

Bestellnummer: 1030804059004

Vertriebspartner: SFG Servicecenter Fachverlage
Part of the Elsevier Group
Postfach 43 43
72774 Reutlingen
Tel.: +49 (0) 70 71 / 93 53 50
Fax: +49 (0) 70 71 / 93 53 35
destatis@s-f-g.com
www.destatis.de/shop

Vorwort

Empirisch arbeitende Sozial- und Wirtschaftswissenschaftler brauchen Mikrodaten. Um ihre Datenbestände insoweit für die Wissenschaft zu erschließen, haben die Statistischen Ämter des Bundes und der Länder - gefördert vom Bundesministerium für Bildung und Forschung (BMBF) - mit Vertretern der Wissenschaft das Forschungsprojekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ durchgeführt. Von der Wissenschaft ist das Vorhaben einhellig begrüßt worden, die Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und amtlicher Statistik (KVI) unterstützte seine Durchführung in ihrem Abschlussgutachten ausdrücklich. Bei der Projektbearbeitung wurde immer wieder deutlich, wie stark die Wissenschaft anonymisierte Mikrodaten nachfragt.

Mit dem Projekt sollten die Möglichkeiten der faktischen Anonymisierung von Mikrodaten zu Unternehmen und Betrieben untersucht, beschrieben und so aufbereitet werden, dass die so entwickelten Anonymisierungsverfahren künftig auf Datensätze der amtlichen Statistik angewendet werden können. Es handelte sich also um Grundlagenforschung zur Verbesserung der Arbeitsbedingungen der empirischen Wirtschaftsforschung.

Während die Statistischen Ämter von Bund und Ländern seit langem faktisch anonymisierte Personendaten bereitstellen, man hier von einer bewährten Praxis sprechen kann, weisen wirtschaftsstatistische Mikrodaten einige Besonderheiten auf, die bisher eine faktische Anonymisierung als ausgesprochen schwierig erscheinen ließen. Skeptiker hielten solches sogar für aussichtslos. Die wissenschaftliche Herausforderung des Projektes bestand deshalb darin, ob auch für Daten über Unternehmen und Betriebe der klassische Zielkonflikt lösbar ist: Sicherstellung der faktischen Anonymität bei gleichzeitigem Erhalt eines Potenzials für wissenschaftliche Analysen. Die Untersuchungen haben erfreulicherweise gezeigt, dass dies gelingen kann. Schon während der Laufzeit des Projektes konnten so erste anonymisierte Daten aus der amtlichen Statistik angeboten werden. Seit dem Frühjahr 2005 stehen der Wirtschaftswissenschaft faktisch anonymisierte Daten aus der Kostenstrukturerhebung im Verarbeitenden Gewerbe, aus der Einzelhandelsstatistik und aus der Umsatzsteuerstatistik zur Verfügung.

Mit dem hier vorgelegten „Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten“ stellt das Projektteam die Ergebnisse seiner Arbeiten allen Interessierten zur Verfügung: Die Entwicklung und Anwendung von Methoden der statistischen Geheimhaltung,

die Operationalisierung des Begriffes der faktischen Anonymität und der Beurteilung des Analysepotenzials wirtschaftsstatistischer Mikrodaten und – aufbauend auf diesen Grundlagen – Empfehlungen für die Vorgehensweise bei der Anonymisierung von Wirtschaftsdaten unterschiedlicher Struktur. Die statistischen Ämter und andere Datenhalter können nun mit vertretbarem Aufwand der Wissenschaft eine Forschung mit Mikrodaten auch für Unternehmens- und Betriebsdaten ermöglichen.

Mein herzlicher Dank gilt dem wissenschaftlichen Leiter des Projektes, Herrn Prof. Dr. Gerd Ronning, Inhaber des Lehrstuhls für Statistik und Ökonometrie an der wirtschaftswissenschaftlichen Fakultät der Universität Tübingen und Direktor des Instituts für Angewandte Wirtschaftsforschung (IAW), Tübingen, sowie den beiden Projektleitern im Statistischen Bundesamt, den Herren Dr. Roland Gnoss und Roland Sturm. Ebenso danke ich den wissenschaftlichen Bearbeitern aus den am Projekt beteiligten Institutionen: Herrn Martin Rosemann, Herrn Dr. Harald Strotmann und Herrn Dr. Rolf Wiegert (Institut für Angewandte Wirtschaftsforschung), Herrn Jörg Höhne (Statistisches Landesamt Berlin), Frau Dr. Ruth Brand, Herrn Dr. Rainer Lenz, Herrn Michael Scheffler und Herrn Dr. Daniel Vorgrimler (Statistisches Bundesamt) sowie Herrn Thorsten Doherr und Frau Dr. Sandra Gottschalk (Zentrum für Europäische Wirtschaftsforschung, Mannheim). Für viele wertvolle Impulse während des Projektes danke ich dem wissenschaftlichen Begleitkreis des Projektes, dem die Herren PD Dr. Uwe Blien, Prof. Dr. Reinhard Hujer, Dr. Georg Licht, Prof. Dr. Winfried Pohlmeier, Prof. Dr. Gerhard Wagenhals, Prof. Dr. Joachim Wagner und Frau Dr. Heike Wirth angehörten.

Schließlich gilt mein besonderer Dank dem Bundesministerium für Bildung und Forschung (BMBF) für die finanzielle Förderung, ohne die das durchaus aufwendige Forschungsprojekt nicht durchgeführt hätte werden können.

Wiesbaden, im September 2005

Der Präsident des Statistischen Bundesamtes

Johann Hahlen

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Vorbemerkung

Die statistischen Ämter haben, gefördert vom Bundesministerium für Bildung und Forschung (BMBF), gemeinsam mit der Wissenschaft das Projekt „Faktische Anonymisierung wirtschaftstatistischer Einzeldaten“ durchgeführt. Mit diesem Handbuch stellt die Projektgruppe ihre Ergebnisse zusammengefasst der Öffentlichkeit zur Verfügung.

Das Projekt hatte zur Aufgabe, die Möglichkeiten der faktischen Anonymisierung von Mikrodaten über Unternehmen und Betriebe zu untersuchen, zu beschreiben und so aufzubereiten, dass sie künftig auf verschiedene Datensätze angewendet werden können. Damit wurde Grundlagenforschung zur Verbesserung der Arbeitsbedingungen der empirischen Wirtschaftsforschung geleistet. Ausgehend von den Projektergebnissen werden die statistischen Ämter – und auch andere Datenhalter – der Wissenschaft künftig den bevorzugten Weg der Forschung mit Mikrodaten der amtlichen Statistik mit vertretbarem Aufwand auch für Unternehmens- und Betriebsdaten eröffnen. Der Erfolg dieser Grundlagenarbeit dokumentiert sich augenfällig darin, dass erste Datenangebote bereits im Rahmen des Projekts geschaffen wurden und seit dem Frühjahr 2005 für die Forschung bereitgestellt werden. Faktisch anonymisierte Unternehmensdaten stehen nun für die interessierte Forschung zur Verfügung.

Projektnehmer gegenüber dem BMBF war das Statistische Bundesamt. An den Projektarbeiten beteiligten sich vier weitere statistische Ämter: das Statistische Landesamt Berlin, das Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, das Statistische Landesamt Schleswig-Holstein und das Bayerische Landesamt für Statistik und Datenverarbeitung. Die übrigen statistischen Ämter der Länder unterstützten die Arbeiten durch die Bereitstellung von Projektdaten. Professor Dr. Gerd Ronning, Inhaber des Lehrstuhls für Statistik und Ökonometrie an der Wirtschaftswissenschaftlichen Fakultät der Universität Tübingen und Direktor des Instituts für Angewandte Wirtschaftsforschung (IAW), Tübingen, war Wissenschaftlicher Leiter des Projekts. Das IAW übernahm wesentliche Projektarbeiten. Als weiterer Datenhalter, der an der Anonymisierbarkeit seiner Mikrodaten arbeitet, hat sich das Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (IAB), Nürnberg mit Daten seines Betriebspanels in die Projektarbeiten eingebracht. Mit dem Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim gab es eine Kooperation.

Dieses Handbuch umfasst zwölf Teile und einen Anhang. Im ersten Teil werden die Ziele, die wesentlichen Projektergebnisse und der weitere Forschungsbedarf zusammengefasst. Der Teil II bietet einen Überblick über die verfügbaren Anonymisierungsverfahren und erläutert die Auswahl aussichtsreicher Verfahren für die Projektanonymisierungen. Teil III stellt die im Projekt genutzten echten Datenbestände aus statistischen Erhebungen vor. In den Teilen IV bis VI wird die Operationalisierung der beiden Randbedingungen für sinnvolle anonymisierte Datensätze vorgestellt – der ausreichende Schutz („faktische Anonymität") und der ausreichende Erhalt des Analysepotenzials der Daten – und das Vorgehen des Verfahrensvergleiches in der Praxis erläutert. Es folgt die Beschreibung der Wirkung von datenverändernden Verfahren auf das Analysepotenzial (Teile VII – IX). Teil X befasst sich mit dem Vorschlag, die Anonymisierung auf die Überschneidungsmerkmale von Zieldaten und externem Zusatzwissen zu beschränken. Im Teil XI werden die Arbeiten zur Anonymisierung von konkreten Erhebungen vorgestellt. Hier finden sich die präzisen Handlungsempfehlungen für Datenhalter/-anbieter, wie verschiedene Grundtypen von Datenbeständen aussichtsreich anonymisiert werden können. Drei der behandelten Erhebungen werden von den statistischen Ämtern inzwischen als Scientific-Use-Files angeboten. Teil XII enthält die Zusammenfassung der wichtigsten Ergebnisse und daraus abgeleitete allgemeine Handlungsempfehlungen für die Anonymisierung wirtschaftsstatistischer Mikrodaten. Aus dem Anhang kann man die praxisnahe Gestaltung der Projektarbeiten und den stattgefundenen Austausch mit der Fachöffentlichkeit ersehen.[1]

1) Von Martin Rosemann geschriebene Teile dieses Handbuchs sind zugleich in eine an der wirtschaftswissenschaftlichen Fakultät der Universität Tübingen eingereichte Dissertationsschrift eingegangen. Es handelt sich dabei um Abschnitt 1.3, die Kapitel 4, 5 und 6, Teil V, die Kapitel 20, 21, 22, 23 und 26 sowie wesentliche Bestandteile der Kapitel 19, 27, 29, 30 und 35.

Teil I

Kurzfassung - Überblick über die Leistungen des Projekts

Im ersten Teil des Handbuchs erfolgt eine Zusammenfassung der wichtigsten Leistungen und Ergebnisse des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“. Kapitel 1 beschreibt die mit dem Projekt verbundenen Ziele, die an den Projektarbeiten beteiligten Institutionen und die für das Vorhaben der Anonymisierung von Betriebs- und Unternehmensdaten wesentlichen Herausforderungen. Kapitel 2 gibt einen Überblick über die einzelnen Projektarbeiten und die wesentlichen Projektergebnisse. In Kapitel 3 erfolgt eine Zusammenstellung der durch die Projektarbeiten aufgeworfenen offenen Fragen und des weiteren Forschungsbedarfs.

Kapitel 1

Projektziele, Projekthintergrund und Projektbeteiligte

1.1 Projektziele

Aufgabe des Projekts war es, die Möglichkeiten der faktischen Anonymisierung von Mikrodaten über Unternehmen und Betriebe zu untersuchen, zu beschreiben und so aufzubereiten, dass sie künftig auf verschiedene Datensätze angewendet werden können. Damit wurde Grundlagenforschung zur Verbesserung der Arbeitsbedingungen der empirischen Wirtschaftsforschung geleistet: Die aus Literatur und der Anwendung bekannten Anonymisierungsverfahren mussten gesammelt, in ihrer Wirkungsweise beschrieben und systematisiert werden. Die Methodenkenntnisse über Anonymisierungsverfahren sollten damit so aufbereitet werden, dass eine gezielte Auswahl geeigneter Verfahren möglich wird. Die Verfügbarkeit von aussichtsreichen und direkt einsetzbaren Verfahren sollte dabei nach ursprünglicher Projektplanung als Randbedingung der Arbeiten gesehen werden. Im Projektverlauf wurde aber die Notwendigkeit erkannt, auch zur Weiterentwicklung von Verfahren beizutragen, insbesondere im Falle der als aussichtsreich erachteten stochastischen Überlagerung. Dadurch erhielt das Projekt in seinem Verlauf auch einen Auftrag zur Methodenentwicklung.

Anonymisierte Einzeldaten müssen zwei gleichrangigen Ansprüchen genügen: Einen ausreichenden Schutz der Einzelangaben gewährleisten (§ 16 Absatz 6 Bundesstatistikgesetzes (BStatG)) und ihre analytische Aussagekraft weiterhin behalten. Beide Ziele sind sowohl für die Nutzer als auch für die statistischen Ämter von großer Bedeutung.

Die Sicherung der Vertraulichkeit von Einzelangaben ist zunächst eine den statistischen Ämtern vom Gesetzgeber vorgegebene Aufgabe. Den Nutzern faktisch anonymisierter Daten wird von Seiten der statistischen Ämter keine Missbrauchsabsicht unterstellt. Unbesehen davon kann der amtlichen Statistik – anders als den Auskunftsgebenden – nicht erst durch tatsächlichen Missbrauch der Daten Schaden entstehen. Vielmehr birgt bereits der Verdacht des leichtfertigen Umgangs der Ämter mit Einzelangaben erhebliche Gefahren für die Auskunftsbereitschaft der Befragten. Diesem Aspekt des von den Auskunftsgebenden

empfundenen Schutzes ihrer Angaben wird national wie international Relevanz zugemessen. Der erkennbar verantwortungsvolle Umgang mit der statistischen Geheimhaltung ist daher wichtig für die Qualität der Statistik insgesamt. Dies ist neben der gesetzlichen Verpflichtung des § 16 BStatG das zweite bestimmende Moment, weswegen die statistischen Ämter dem Schutz der ihnen anvertrauten Daten große Bedeutung beimessen. Eine Erschütterung des Vertrauens in die Vertraulichkeit der von ihr verwalteten Daten hätte für die amtliche Statistik fatale Folgen und würde auch der Wissenschaft als einem ihrer wichtigsten Nutzer schaden. Das Projekt sollte die Umsetzbarkeit dieser Regelung in die Praxis umfassend untersuchen und beschreiben. Die Herausforderung bestand darin, eine Operationalisierung von „faktischer Anonymität" zu schaffen, die Aussagen darüber erlaubt, ob und wann Einzeldaten von Unternehmen und Betrieben als ausreichend geschützt gelten können.

Die primäre Anforderung der Nutzerseite besteht im möglichst weit gehenden Erhalt des Analysepotenzials anonymisierter Daten. Dies wird von den Ämtern in gleicher Weise für wichtig erachtet, denn die statistischen Ämter haben selbst ein großes Interesse daran, dass Analysen auf der Grundlage ihrer faktisch anonymisierten Daten valide Ergebnisse liefern. Die Beteiligung der künftigen Nutzer an den Arbeiten war deshalb eine wesentliche Rahmenvorgabe für die Projektarbeiten. Augenfällig wird diese Einbindung der Nutzer durch den Projektpartner Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen und die wissenschaftliche Leitung des Projekts durch den Direktor des IAW, Professor Dr. Gerd Ronning. Das IAW sollte die Analysemöglichkeiten von probeanonymisierten Daten umfänglich untersuchen und bewerten. Die stärkere Nutzerbeteiligung wurde durch die Einrichtung eines Wissenschaftlichen Begleitkreises beim IAW verankert, der die gesamten Projektarbeiten begleitet hat. Ergänzend wurden eine schriftliche Nutzerumfrage und zwei wissenschaftliche Kolloquien zum Analysepotenzial durchgeführt.

Ein wesentliches Charakteristikum der Projektkonzeption bestand darin, für die Arbeiten echte Mikrodaten der statistischen Ämter zur Verfügung zu stellen. Die Untersuchungen sollten alle Eigenschaften echter Mikrodatenbestände einbeziehen können, um die Realitätsnähe und die Anwendbarkeit der Forschungsergebnisse sicherzustellen. Bei den Prüfungen zur Schutzwirkung von Anonymisierungsmaßnahmen sollten damit die realen Möglichkeiten von Datenangreifern zugrunde gelegt werden können. Bei den Bewertungen des Analysepotenzials wurden Beispielrechnungen als Ergänzung der theoretischen Ableitungen erschlossen. Ein sehr positiver Begleiteffekt dieses Vorgehens sollte sein, möglichst schon projektbegleitend Datenangebote zu erstellen.

Der Anspruch, den die Initiatoren dieses Projekts an ihre Arbeit hatten, ging jedoch über die Schaffung eines konkreten Datenangebotes für bestimmte Erhebungen der statistischen Ämter hinaus. Wesentliche Leistung der durchgeführten Grundlagenforschung ist es, den Haltern von Mikrodaten eine ressourcenschonende Anonymisierung zu ermöglichen, die es ihnen erlaubt, ihre Datenbestände Forschern mit vertretbarem Aufwand zugänglich zu machen. Für die statistischen Ämter als wesentliche Datenhalter bedeutet dies konkret, dass sie die vom Bundesstatistikgesetz geforderte faktische Anonymität von Mik-

rodaten erzeugen können. Dazu wird hier ein Kompendium der Methoden der Anonymisierung und Hilfen für deren sinnvolle und effiziente Anwendung bereitgestellt. Kontakte zu anderen Datenhaltern im In- und Ausland haben bereits gezeigt, dass hieran ein erhebliches Interesse besteht. Das vorliegende Handbuch zur Anonymisierung enthält auch den Vergleich der Eignung alternativer Anonymisierungsstrategien: Anonymisierungen, die sich auf Überschneidungsmerkmale beschränken versus Anonymisierungen, die sich auch auf Nicht-Überschneidungsmerkmale erstrecken. Im Ergebnis wurden vom Projekt differenzierten Aussagen über die Möglichkeiten faktischer Anonymisierbarkeit erwartet. Das Kompendium, das mit diesem Band vorgelegt wird, liefert den Ämtern und allen weiteren interessierten Datenanbietern Hilfen für die sinnvolle und effiziente Anonymisierung. Es bietet die Grundlage für die Methodenauswahl, die Anwendung der Methoden und die Prüfung ihrer Wirkungsweise künftiger Anonymisierungen.

1.2 Projektbeteiligte

Zur Erreichung der beschriebenen Ziele wurde ein Projektteam aus Datenanbietern und Datennutzern gebildet. Projektnehmer gegenüber dem BMBF war das Statistische Bundesamt. An den Projektarbeiten beteiligten sich vier weitere statistische Ämter: das Statistische Landesamt Berlin, das Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, das Statistische Landesamt Schleswig-Holstein und das Bayerische Landesamt für Statistik und Datenverarbeitung. Die übrigen statistischen Ämter der Länder unterstützten die Arbeiten durch die Bereitstellung von Projektdaten. Damit standen den Projektbearbeitern deutschlandweite Daten aus sechs statistischen Erhebungen zur Verfügung. Alle statistischen Ämter wurden über den Fortgang der Projektarbeiten regelmäßig informiert. Das im Projekt entwickelte Schutzwirkungskonzept wurde mit den Statistischen Ämtern im Rahmen eines Workshops diskutiert.

Die künftigen Nutzer von anonymisierten Einzeldaten waren umfangreich in die Projektarbeiten eingebunden. Professor Dr. Gerd Ronning, Inhaber des Lehrstuhls für Statistik und Ökonometrie an der Wirtschaftswissenschaftlichen Fakultät der Universität Tübingen und Direktor des Instituts für Angewandte Wirtschaftsforschung (IAW), Tübingen, war Wissenschaftlicher Leiter des Projekts. Das IAW übernahm wesentliche Projektarbeiten. Dem IAW oblag nicht nur die Bewertung des Analysepotenzials, sondern es hat sich auch wesentlich in der Entwicklung geeigneter Anonymisierungsstrategien engagiert.

Zur Verbreiterung der Basis der Nutzereinbindung hat das IAW in Abstimmung mit den Auftraggebern einen Wissenschaftlichen Begleitkreis (WBK) eingerichtet, der sich aus empirisch arbeitenden Wirtschaftswissenschafltern und Sozialwissenschaftlern zusammensetzte. Mitglieder des Wissenschaftlichen Begleitkreises waren PD Dr. Uwe Blien (Institut für Arbeitsmarkt- und Berufsforschung, Nürnberg), Dr. Georg Licht (Zentrum für Europäische Wirtschaftsforschung, Mannheim), Prof. Dr. Winfried Pohlmeier (Universität Konstanz), Prof. Dr. Gerhard Wagenhals (Universität Hohenheim, bis November 2002), Prof. Dr.

Joachim Wagner (Universität Lüneburg), Dr. Heike Wirth (Zentrum für Umfragen Methoden und Analysen, Mannheim) und Prof. Dr. Reinhard Hujer (Universität Frankfurt, ab November 2002). Weitere Vertreter der empirischen Wirtschafts- und Sozialforschung wurden durch eine Nutzerumfrage und zwei Nutzerkonferenzen angesprochen.

An den Projektabeiten beteiligten sich auch weitere Forschungseinrichtungen: Als weiterer Datenhalter, der an der Anonymisierbarkeit seiner Mikrodaten arbeitet, hat sich das Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit (IAB), Nürnberg mit Daten seines Betriebspanels in die Projektarbeiten eingebracht. Mit dem Zentrum für Europäische Wirtschaftsforschung (ZEW), Mannheim gab es eine zweifache Kooperation: Zum einen wurden gemeinsame Arbeiten zum Record Linkage durchgeführt und zum anderen die Wirkungsweise von Resampling-Verfahren untersucht.

1.3 Die besondere Problemlage bei der Anonymisierung wirtschaftsstatistischer Einzeldaten

Die faktische Anonymisierung wirtschaftsstatistischer Einzeldaten wurde bisher allgemein als schwierig eingeschätzt. Diese Einschätzung hat verschiedene Gründe. Zunächst sind bei Unternehmens- und Betriebsdaten im Allgemeinen die Grundgesamtheiten kleiner als bei Haushalts- und Personendaten. Dies hat zur Folge, dass die Besetzungszahlen innerhalb einzelner Gruppen häufig sehr klein sind. Bei Unternehmens- und Betriebsdaten existieren dadurch mehr einzigartige, häufig auch sehr leicht zu identifizierende Fälle (Brand 2000; Rosemann und Vorgrimler 2004). Letzteres ist vor allem darauf zurückzuführen, dass die Verteilungen der quantitativen Variablen wesentlich heterogener sind (KVI 2001). Eng damit verbunden ist die Tatsache, dass bei Unternehmens- und Betriebsdaten im Gegensatz zu Haushalts- und Personendaten häufig Dominanzen auftreten, beispielsweise auf ein oder wenige Unternehmen ein Großteil des Umsatzes einer Branche entfällt (KVI 2001, S.166). Ein weiteres Problem besteht darin, dass bei Unternehmens- und Betriebsdaten in der Regel größere Stichprobenauswahlsätze anzutreffen sind. Es treten zumindest in bestimmten Größenklassen sogar Vollerhebungen auf. Im Gegensatz dazu ist beispielsweise der Mikrozensus als eine bedeutende Erhebung im Bereich der Personen- und Haushaltsdaten gemessen an der Anzahl der erhobenen Fälle zwar eine große Erhebung der amtlichen Statistik, jedoch mit einem sehr geringen Auswahlsatz (von einem Prozent). Zuletzt unterscheiden sich die Daten von Unternehmen und Betrieben sehr stark in ihrer Größe. Insbesondere gibt es nur sehr wenige große Einheiten (KVI 2001, S.166).

All diese Faktoren machen es potenziellen Datenangreifern leichter, Unternehmen voneinander zu unterscheiden und damit auch in formal anonymisierten Datenbeständen zu erkennen, als dies bei Personen- und Haushaltsdaten der Fall ist. Will ein Angreifer jedoch ein, mehrere oder viele Unternehmen erkennen, so benötigt er entsprechendes Zusatzwissen, das ihm eine Zuordnung ermöglicht. Beispielsweise benötigt er Kenntnisse über die Branchenzugehörigkeit, die Rechtsform, den Standort, den Umsatz und die Beschäftigten-

zahl der gesuchten Unternehmen. Dabei ist leicht einsichtig, dass solcherlei Zusatzwissen für Unternehmen und Betriebe in weitaus größerem Umfang, deutlich leichter zugänglich und besser aufbereitet zur Verfügung steht als für Haushalte und Personen. Dies ergibt sich insbesondere aus Publizitätspflichten von Unternehmen, der Existenz von allgemeinen Unternehmensdatenbanken und Bilanzdatenbanken (KVI 2001, S.166), (Vorgrimler 2002). Allerdings dürfen Informationen von den statistischen Ämtern selbst dann nicht preisgegeben werden, wenn sie gleichzeitig aus den genannten Quellen bezogen werden können. Es kommt hinzu, dass Unternehmen ab einer gewissen Größe oft zu mehreren Erhebungen meldepflichtig sind (Sturm 2002a, S.472). Damit ergeben sich Probleme insbesondere für die Anonymisierung größerer Unternehmen, die häufiger innerhalb bestimmter Merkmalskombinationen einzigartig sind und für die aufgrund von Publizitätspflichten vergleichsweise viel Zusatzwissen zur Verfügung steht (Rosemann und Vorgrimler 2004, S.4). Allerdings dürften Datenfehler und insbesondere Dateninkompatiblitäten zwischen den verschiedenen Erhebungen (beispielsweise durch eine unterschiedliche Abgrenzung des Unternehmens oder des Umsatzes eines Unternehmens) den Abgleich zwischen Zusatzwissen und den Daten der amtlichen Statistik erschweren und damit bereits eine „natürliche Schutzwirkung“ der Daten erzeugen.

Diese Überlegungen lassen vermuten, dass es angebracht ist, bei wirtschaftsstatistischen Einzeldaten „stärkere“ Anonymisierungsvarianten – sowohl im Bereich der Verfahren zur Informationsreduktion als auch im Bereich der datenverändernden Verfahren – zu ergreifen als bei Haushalts- und Personendaten. Dies gilt umso mehr, wenn man bedenkt, dass auch der mögliche Nutzen aus der Enthüllung vertraulicher Informationen über Unternehmen und Betriebe wesentlich höher eingestuft werden muss als bei Personen- und Haushaltserhebungen (Sturm 2002a, S.472). Die Verwendung datenverändernder Verfahren zur Ergänzung des Repertoires von Anonymisierungsmaßnahmen bedeutet per se nicht, dass die Anonymisierung aus Sicht des Analysepotenzials gravierender in die Daten eingreift als bei einer Beschränkung auf die traditionell eingesetzten Verfahren zur Informationsreduktion. Der diskursive Abwägungsprozess über die jeweils adäquaten Verfahren wird in diesem Handbuch thematisiert. Wichtig an dieser Stelle ist deshalb nur der grundlegende Hinweis, dass die Entscheidung für den Einsatz datenverändernder Verfahren aus Gesichtspunkten des Analysepotenzials auch die schonendere Maßnahme bedeuten kann.

Kapitel 2

Überblick über die Projektarbeiten und wesentliche Projektergebnisse

Die Leistungen des Projekts lassen sich wie folgt zusammenfassen:

1. Sichtung, Weiterentwicklung und Anwendung von Anonymisierungsverfahren auf Daten der statistischen Ämter und des IAB.

   Im Rahmen des Projekts wurde ein breiter Kanon von verfügbaren Anonymisierungsverfahren untersucht. Teil II dieses Handbuchs beschreibt umfassend die Anonymisierungsmethoden, die im Rahmen dieses Projekts zusammengestellt und deren Eigenschaften analysiert wurden. Dem Leser bietet sich hier die Möglichkeit, einen schnellen und anschaulichen Überblick über die in der Anonymisierungsforschung eingesetzten Verfahren zu gewinnen. Darüber hinaus werden die Kriterien beschrieben, nach denen die Auswahl aussichtsreicher Verfahren für die Projektanonymisierungen vorgenommen wurde.

   Es wurden zur Anonymisierung auch Verfahren eingesetzt, die sich noch in der Phase der Entwicklung befinden. Einige dieser Methoden wurden in diesem Projekt einem ersten Test anhand von echten Daten unterzogen und die Forscher, die diese Methoden entwickeln, erhielten aus dem Projekt Anregungen zur Weiterentwicklung und Modifizierung.
   Nach ursprünglicher Projektplanung sollte sich die Auswahl und der Einsatz von Verfahren auf das bereits vorfindbare Methodenspektrum beschränken. Im Projektverlauf wurde jedoch deutlich, dass bei den als aussichtreich eingestuften stochastischen Überlagerungen eine Weiterentwicklung der vorgefundenen Methoden dringend angeraten schien. Durch eine Aufstockung der Projektförderung des BMBF wurde diese Methodenarbeit ermöglicht.

2. Entwicklung einer Konzeption zur Bewertung der Vertraulichkeit und zur Gewährleistung der faktischen Anonymität (Schutzwirkungskonzept).

Ein neuartiger Ansatz zur Operationalisierung der faktischen Anonymität ergänzt die Betrachtung der Kostenseite von De-Anonymisierungsversuchen um eine Nutzenbetrachtung. Dies bietet die Möglichkeit, anhand von Kriterien, die auch den Nutzen einer möglichen Reidentifikation von Unternehmen berücksichtigen, zu beurteilen, ob faktische Anonymität erreicht ist. Mit diesem Konzept hebt sich das Projekt deutlich von seinem Vorgänger (Müller et al. 1991) und von anderen aktuellen Anonymisierungsprojekten ab. Zur Implementierung des neuen Ansatzes werden Motivation und Möglichkeiten von Datenangreifern beleuchtet und mögliche Angriffsszenarien beschrieben, ferner Quellen des Zusatzwissens von Angreifern und technische Möglichkeiten eines Datenangriffs aufgezeigt.

3. Entwicklung von Strategien zur Bewertung des Analysepotenzials anonymisierter Daten.

   Niemandem ist damit gedient, anonymisierte Daten anzubieten, die die Nutzer nicht einsetzen können oder wollen. Daher wurde dieses Projekt von den Datenanbietern und Datennutzern gemeinsam durchgeführt. Der Projektpartner IAW hatte die Aufgabe, die Nutzeranforderungen im Projekt zu vertreten. Um die Bedürfnisse der Anwender besser kennen zu lernen, wurden eine Nutzerbefragung und zwei Nutzerworkshops durchgeführt. Die Ergebnisse geben einen guten Einblick in die Bedürfnisse der zukünftigen Nutzer. Diese Kenntnisse wurden im Verlauf der Projektarbeiten aufgegriffen und berücksichtigt. Als Novum in der Anonymisierungsforschung wird der Begriff des Analysepotenzials bzw. seine Veränderung aufgrund von Anonymisierungsmaßnahmen systematisiert und dann umfassend operationalisiert. Auf Basis dieser Arbeiten wird im Projekt das Analysepotenzial der anonymisierten Daten beurteilt und ein Raster an Empfehlungen für die Bewertung künftiger Anonymisierungen bereitgestellt. Darüber hinaus gibt das Projekt Anregungen an die Datenanbieter, wie sie auf Unsicherheiten und Vorbehalte der Datennutzer eingehen und diesen Unterstützung beim sachgerechten Einsatz von anonymisierten Daten geben können.

4. Systematisierung der Auswirkungen datenverändernder Anonymisierungsverfahren auf Analysen.

   Im Projekt wurden die Auswirkungen unterschiedlicher datenverändernder Anonymisierungsverfahren auf deskriptive Verteilungsmaße sowie lineare und nichtlineare Modelle theoretisch untersucht. Außerdem wurden Korrekturansätze bei datenverändernden Anonymisierungsverfahren in linearen und nichtlinearen Modellen überprüft.

5. Untersuchung der Auswirkungen unterschiedlicher Anonymisierungsmaßnahmen auf das Analysepotenzial konkreter Projektdatensätze (Kostenstrukturerhebung im Verarbeitenden Gewerbe, Umsatzsteuerstatistik, Einzelhandelsstatistik, Querschnitt aus dem IAB-Betriebspanel).

   Für die konkreten Anonymisierungen der Projektstatistiken wurden die Auswirkungen auf das Analysepotenzial – wie es unter Punkt 3 skizziert wurde – untersucht und beschrieben. Für die Anonymisierungsvarianten, die für die erstellten Scientific-Use-Files gewählt wurden, sind die jeweils vorgegebenen Abweichungstoleranzen weitgehend

eingehalten. Problematische Bereiche in den Daten werden zudem im Metadatenmaterial der jeweiligen Datenangebote beschrieben.

6. Untersuchung der Auswirkungen unterschiedlicher Anonymisierungsmaßnahmen auf die Vertraulichkeit anhand der Simulation von Angriffsszenarien.

   Das Projekt entwickelte zum Test der Schutzwirkung modernste Methoden und stand im Dialog mit der aktuellen Anonymisierungsforschung im In- und Ausland. Weitere Datenanbieter in Deutschland, aber auch im Ausland zeigten großes Interesse an unseren Arbeiten und möchten auf unseren Ergebnissen aufbauen.

   Zu den Möglichkeiten faktischer Anonymisierbarkeit, die nicht mit einem generellen „Ja“ oder „Nein“ beantwortet werden können, hat das Projekt differenzierte Aussagen ermöglicht, die zum einen die Betrachtung der gängigen Methoden und Analyseformen der empirischen Forschung einbeziehen und zum anderen auch die organisatorische und juristische Ausgestaltung von Datennutzungen betreffen. Der Ansatz, die Untersuchungen anhand von echten Datenbeständen aus Unternehmens- und Betriebserhebungen durchzuführen, hat sich bewährt, da es dadurch möglich wird, die Praxistauglichkeit der einzelnen Verfahren realitätsbezogen zu testen. Insbesondere die Vielfalt der im Projekt verwendeten Daten ermöglicht es, die Auswirkungen der Anonymisierung auf unterschiedlich strukturierte Daten (z.B. unterschiedlicher Umfang, unterschiedliche Dichte) zu untersuchen.

7. Erstellung von anonymisierten Daten: Public-Use-File kleiner und mittlerer Unternehmen im Verarbeitenden Gewerbe; faktisch anonymisierte Scientific-Use-Files für Querschnittsdaten der Kostenstrukturerhebung im Verarbeitenden Gewerbe, der Umsatzsteuerstatistik und der Einzelhandelsstatistik.

   Das für die Nutzer wichtigste Ziel ist es, Datensätze verschiedener, für die Forschung interessanter, Erhebungen möglichst komfortabel nutzen zu können. Um die in Deutschland in den statistischen Ämtern neu eingerichteten Forschungsdatenzentren inhaltlich mit Leben zu füllen, sollen faktisch anonymisierte Einzeldaten ein wesentlicher Baustein sein. Ein wesentlicher Erfolg des Projekts zeigt sich darin, dass bereits während der Laufzeit des Grundlagenprojekts mehrere anonymisierte Datensätze erstellt werden konnten, die von interessierten Forschern bereits über die Forschungsdatenzentren der Statistischen Ämter bezogen werden können.

   Das Projekt hat vier Scientific-Use-Files für drei verschiedene Projektstatistiken realisiert und bereits in der Projektlaufzeit etappenweise bereitgestellt. Den Einstieg stellte eine Datei über 500 kleine und mittlere Unternehmen der Kostenstrukturerhebung im Verarbeitenden Gewerbe dar, der als Public-Use-File jedermann zugänglich ist und ein neues Datenangebot der Forschungsdatenzentren der statistischen Ämter mitbegründete. Es folgten faktisch anonymisierte Daten über alle 2,9 Millionen Unternehmen der Umsatzsteuerstatistik, über 13.000 Klein- und Mittelunternehmen des Verarbeitenden Gewerbes und 14.500 Unternehmen des Einzelhandels. Schließlich wurden zu Projektende die großen Unternehmen der beiden letztgenannten Bereiche anonymisiert und angeboten. Das Projekt hat sich dabei auf die Anonymisierung von

Einzeldaten im Querschnitt beschränkt. Die Anonymisierung von Paneldaten konnte wegen der zunächst erforderlichen Grundlagenforschung – insbesondere über die Wirkungsweise verschiedener Anonymisierungsverfahren – sowie wegen der zusätzlichen Probleme und Anforderungen bei der Anonymisierung von Paneldaten nicht im abgeschlossenen Projekt erreicht werden. Hierzu haben die Projektpartner ein weiteres Forschungsprojekt initiiert.

8. Formulierung von Handlungsempfehlungen für die Anonymisierung wirtschaftsstatistischer Einzeldaten.

   Auf der Grundlage dieser theoretischen und empirischen Ergebnisse, die mit Hilfe realer Daten gewonnen wurden, können in diesem Handbuch Schlussfolgerungen zur Anonymisierbarkeit von Unternehmens- und Betriebsdaten gezogen werden. Die Ergebnisse machen deutlich, welche Wege der Anonymisierung Erfolg versprechend sind und daher auch bei weiteren Daten verfolgt werden sollten. Das vorliegende Kompendium will eine aussagekräftige Grundlage für die Methodenauswahl, die Anwendung der Methoden und die Prüfung ihrer Wirkungsweise bieten. Dadurch wird die Anwendung auf alle weiteren Datenbestände ermöglicht, an denen ein Nutzerinteresse besteht. Damit unterstützt das BMBF die Bereitstellung von anonymisierten Daten an die Wissenschaft weit über die Datenbestände der statistischen Ämter hinaus.

Kapitel 3

Offene Fragen und weiterer Forschungsbedarf

Das Projekt hat seine Ziele (siehe dazu Abschnitt 1.1) im Wesentlichen erreicht. Zur konkreten Abgrenzung des Projektrahmens gehörte dabei insbesondere (a) die Konzentration auf Anonymisierungsverfahren, für die theoretische Ergebnisse und möglichst auch geeignete Software sowie erste Erfahrungen bezüglich ihrer Anwendung verfügbar waren und (b) eine Konzentration auf einige, allerdings sehr wichtige und für die Datennutzer attraktive Datensätze aus dem Bereich der wirtschaftsstatistischen Daten, die die amtliche Statistik erhebt. Außerhalb dieser für das Projekt notwendigen Abgrenzung wurde im Laufe der Untersuchungen der Forschungsbedarf deutlich, auf den sich aus Sicht des Projektteams weitere Untersuchungen erstrecken sollten.

3.1 Anonymisierungsverfahren und ökonometrisch-statistische Methoden

Von den datenverändernden Verfahren, die in Kapitel 6 beschrieben werden, wurden vor allem die Mikroaggregation und die stochastische Überlagerung, die beide ausschließlich auf metrische Merkmale anzuwenden sind, im Projekt sehr ausführlich sowohl theoretisch als auch anhand von Datensätzen überprüft. Daneben wurden auch ausgewählte Simulationsmethoden eingesetzt: Zum einen das Latin Hypercube Sampling (=LHS)-Verfahren von Dandekar (Dandekar et al. 2001), zum anderen das von Sandra Gottschalk vorgeschlagene Verfahren des Resamplings (Gottschalk 2005), das allerdings durchaus auch als Variante der additiven stochastischen Überlagerung angesehen werden kann. Allerdings ist damit keineswegs die Klasse der Simulations- und stichprobenbezogenen Verfahren erschöpfend behandelt. Insbesondere spielen in der neueren Literatur Imputationsverfahren eine wichtige Rolle, wobei allerdings hinzugefügt werden muss, dass bisher kaum Anwendungen für die Anonymisierung bekannt sind. Alle bisher genannten Methoden werden nur auf metrische Merkmale angewendet. Dagegen kann, wie im Einzelnen in Kapitel 6 beschrieben wird, für kategoriale Merkmale das Verfahren der Post-Randomisierung (PRAM) eingesetzt werden.

Im Rahmen des Projekts wurden diese Verfahren sowohl theoretisch als auch anhand echter Datensätze untersucht. Die theoretischen Untersuchungen dienten dazu, allgemein gültige Aussagen über den Effekt bestimmter Anonymisierungsverfahren auf die Schätzung ökonometrischer Modelle zu gewinnen. Teilweise wurden diese Ergebnisse durch Simulationsstudien untermauert. (Siehe dazu die Resultate in den Teilen VIII und IX). Gleichzeitig wurden die im Projekt behandelten Datensätze zu diesen Untersuchungen herangezogen. Dabei ging es vor allem darum, einen Einblick in die Auswirkung von bestimmten Anonymisierungsmethoden auf die Schätzung konkreter ökonometrischer Modelle zu gewinnen. Auch dies ist in den genannten Teilen dokumentiert.

Weiteren Forschungsbedarf gibt es vor allem in folgenden Bereichen:

- Ergänzende Untersuchungen, die das Gebiet der Mikroaggregation mit ihren verschiedenen Varianten bezüglich der Effekte auf die ökonometrisch statistische Analyse untersuchen, wären wünschenswert. Beispielsweise betrachtet die Arbeit von Schmid et al. (2005) eine spezielle Variante der Mikroaggregation (Aggregation nach der endogenen Variablen) und zeigt, dass der Effekt auf die Schätzung stark von der Korrelation zwischen endogener und exogener Variablen abhängt. Insbesondere sollten die Auswirkungen der Mikroaggregation auf die Teststatistiken bei verschiedenen Formen der Mikroaggregation einer vertieften Anlayse unterzogen werden.

- Die multiplikative Überlagerung ist bisher nur teilweise bezüglich geeigneter Schätzverfahren, vor allem unter Korrektur der Verzerrung, die durch die Anonymisierung entsteht, bearbeitet worden. Insbesondere sollte noch systematischer untersucht werden, wie der SIMEX-Schätzer, der ursprünglich für additive Überlagerung entwickelt wurde, in diesem Fall arbeitet. Wünschenswert wären auch entsprechende Realisierungen in den statistischen Programmpaketen wie beispielsweise STATA. Außerdem sollten mögliche alternative Korrekturverfahren (z.B. GMM und SIMEX) systematisch miteinander verglichen werden.

- Für das im Projekt betrachtete Resamplingverfahren sollten geeignete Korrekturverfahren entwickelt werden. Wegen der Verwandtschaft mit der additiven Überlagerung sind entsprechende Ergebnisse dort vermutlich zumindest teilweise übertragbar.

- Auch die Anonymisierung kategorialer Merkmale mittels Post-Randomisierung (PRAM) ist bisher nicht umfassend genug untersucht worden. Neben dem Problem, wie bei Merkmalen mit mehr als zwei Kategorien und vor allem bei ordinalen Merkmalen die Übergangswahrscheinlichkeiten zu bestimmen sind, ist bisher nicht untersucht worden, wie sich PRAM auf mehrere kategoriale Merkmale gemeinsam auswirkt. Weiterer Forschungsbedarf besteht auch hinsichtlich des Zusammenspiels der Post-Randomisierung einer diskreten abhängigen Variablen mit einer multiplikativen stochastischen Überlagerung der Regressoren im Probit-Modell.

- Wenn eine mit PRAM anonymisierte binäre Variable als Einflussgröße (Dummy-Variable) verwendet wird, ergibt sich ein spezielles Fehler-in-den-Variablen-Modell.

Es sollte untersucht werden, inwieweit auch in diesem Fall ein Korrekturverfahren analog dem SIMEX-Verfahren für den metrischen Fall konsistente Schätzungen ergibt.

- Sowohl stochastische Überlagerungen als auch die Post-Randomisierung führen in Schätzungen in der Regel zu einem Effizienzverlust und damit zu einer Verringerung der Werte von Teststatistiken. Damit werden die Schlussfolgerungen durch die Anonymisierung unsicherer. Es sollte geprüft werden, ob sich für die genannten Verfahrensarten „Faustregeln“ aufstellen lassen, um den durch die Anonymisierung hervorgerufenen Effizienzverlust zu quantifizieren und damit die statistische Signifikanz in den Originaldaten abzuschätzen.

- Die Interaktion zwischen verschiedenen Anonymisierungsverfahren, die gemeinsam auf einen Datensatz angewendet werden (z.B. Mikroaggregation bei den metrischen und PRAM bei den kategorialen Merkmalen) sollte ebenfalls systematischer analysiert werden. Beispielsweise sollte die Auswirkung der Festlegung von Parametern in den einzelnen Verfahren näher untersucht werden („Kalibrierungsproblem“). Dies sollte allerdings stets im Rahmen von konkreten Anonymisierungsvorhaben betrachtet werden.

- Der bereits oben erwähnte Ansatz der (multiplen) Imputation, der ebenfalls als Anonymisierungsmethode vorgeschlagen wurde, sollte in zukünftige Untersuchungen einbezogen werden.

3.2 Anonymisierung von Paneldaten

Die Auswahl der im vorliegenden Projekt betrachteten Datensätze erfolgte nach dem Gesichtspunkt einer möglichst großen Vielfalt, wobei die ausgewählten Datensätze durchaus als repräsentativ für das Datenangebot der amtlichen Statistik im Bereich der Wirtschaftsstatistik (Unternehmensdaten) angesehen werden können. Andererseits konnte die Ausweitung der Untersuchungen auf Paneldaten nicht geleistet werden.

Paneldaten stehen immer stärker im Zentrum des wirtschaftswissenschaftlichen Interesses, weil sie vor allem folgende zwei bedeutsame Erweiterungen der Analysemöglichkeiten gegenüber der Verwendung von Makro- bzw. Mikrodaten im Querschnitt zulassen:

- Nur mit Paneldaten kann die individuelle Dynamik der Befragungseinheiten untersucht und damit Evidenz für die Mikrofundierung in der Wirtschaftstheorie geliefert werden. Beispielsweise ist zu erwarten, dass eine im Zeitverlauf vergleichsweise konstante makroökonomische Kennzahl, wie z.B. das Wirtschaftswachstum als Veränderungsrate des BSP, durchaus durch erhebliche Schwankungen auf der Mikroebene gekennzeichnet ist. Somit können insbesondere die Determinanten des Unternehmenswachstums

zum Untersuchungsgegenstand werden und die resultierenden Erkenntnisse für gezielte wirtschaftspolitische Maßnahmen genutzt werden.

- Bei vielen Untersuchungen ist zu erwarten, dass in den unterstellten Wirkungszusammenhängen beobachtbare oder auch unbeobachtbare Eigenschaften der Mikroeinheiten eine Rolle spielen. Beispielsweise möchte man untersuchen, welche Charakteristika einer Person die Höhe der Entlohnung im Beruf beeinflussen, oder ob eine bestimmte Branchenzugehörigkeit auf das Verhalten eines Unternehmens Einfluss hat. Wesentlich ist, dass diese Eigenschaften (z.B. das Geschlecht bei Personen oder die Branchenzugehörigkeit bei Unternehmen) über die Zeit hin konstant sind. Im Fall unbeobachtbarer Einflüsse greift man zu speziellen Techniken, die unter dem Begriff der unbeobachteten Heterogenität bekannt sind.

Neuere Untersuchungen vor allem aus dem Bereich der Industrieökonomik und der Innovationsökonomik sind nicht mehr an Panels mit konstanter Menge von Untersuchungseinheiten interessiert, sondern an den Ein- und Austritten aus einer Population und damit an entsprechenden Beobachtungen im Paneldatensatz. Typische Beispiele sind das Mannheimer Innovations Panel (MIP) (Gottschalk 2004), das auf der Basis der von Creditreform erhobenen Angaben das Gründungs- und Innovationsverhalten von Unternehmen untersucht, oder die Monatsberichte der amtlichen Statistik in Verbindung mit der industriellen Kleinbetriebserhebung mit ähnlichen Forschungsfragen (siehe unter anderem Wagner (1992), Wagner (1994b), Wagner (1994a) und Strotmann (2001)). Die Anwendung klassischer panel-ökonometrischer Methoden (Baltagi 2001) ist für solche Untersuchungen nicht mehr angemessen, weil das Stichproben-Design fundamental anders ist: Die Stichprobenziehung ist endogen (siehe Ronning (2003) und die dort zitierte Literatur). Auf der anderen Seite sind solche Datensätze, in denen Gründe für Zu- und Abgänge eigentliches Untersuchungsziel sind, in der modernen Wirtschafts- und Sozialforschung ein wichtiges Forschungsgebiet. Problematisch ist, dass Zu- und Abgänge auch durch andere Gründe bedingt sein können. So können z.B. Unternehmen aus den Monatsberichten im Verarbeitenden Gewerbe herausfallen, weil sie eine neue wirtschaftszweigsystematische Zuordnung erhalten.

Die Anonymisierung von Paneldaten stellt zusätzliche Anforderungen an Anonymisierungsverfahren. Während nicht verknüpfbare Querschnittsdaten für verschiedene Zeitpunkte unabhängig voneinander anonymisiert werden können, ist dies für Paneldaten nicht der Fall. Die zeitliche Dimension im Paneldatensatz erfordert auch die Berücksichtigung der zeitlichen Korrelation (neben der Korrelation im Querschnitt zwischen den Merkmalen), wobei noch zwischen Korrelation innerhalb einer Variablen (auto correlation) und der Korrelation zwischen Merkmalen (auto cross correlation) zu unterscheiden ist. Querschnittsdaten beziehen sich hingegen stets nur auf einen Zeitpunkt.

Die geschilderten Aspekte zeigen, dass die Anonymisierung von Paneldaten eine zusätzliche Herausforderung darstellt, die in einem zur Zeit beantragten Anschlussprojekt gesondert in Angriff genommen werden soll.

Teil II

Anonymisierungsansätze und Anonymisierungsverfahren

Der zweite Teil des Handbuchs bietet einen umfangreichen Überblick über die unterschiedlichen Ansätze zur Anonymisierung von Einzeldaten und die wichtigsten Gruppen von Anonymisierungsverfahren und ihren Repräsentanten. Kapitel 4 beschreibt zunächst die Ansätze zur Systematisierung von Anonymisierungsverfahren. In Kapitel 5 werden die Verfahren zur Informationsreduktion, in Kapitel 6 die datenverändernden Anonymisierungsverfahren vorgestellt. Kapitel 7 beschreibt abschließend die Kriterien zur Auswahl der im Projekt näher untersuchten Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten.

Kapitel 4

Abgrenzung der verschiedenen Anonymisierungsverfahren

Wie in Abschnitt 1.3 beschrieben, existiert bereits eine natürliche Schutzwirkung gegenüber Angriffen bei wirtschaftsstatistischen Einzeldaten. Ziel jeder Anonymisierungsmaßnahme ist es, diese natürliche Schutzwirkung zu erhöhen und die Unsicherheiten für einen Datenangreifer zu erhöhen.

In Anlehnung an die in Höhne et al. (2003) vorgenommene Operationalisierung der faktischen Anonymität, die in Teil IV ausführlich beschrieben wird, werden grundsätzlich zwei unterschiedliche Effekte, die zur faktischen Anonymität beitragen können, unterschieden:

- Verhinderung der eindeutigen Zuordnung von Merkmalsträgern
- Verhinderung eines Informationsgewinns (Entfernen der Information) beziehungsweise Reduzierung des Nutzens des Informationsgewinns (Veränderung der Information) bei erfolgter Zuordnung.

Alle Anonymisierungsverfahren führen Veränderungen an den Einzeldaten durch. Dabei existieren verschiedene Ansätze der Veränderung der Einzeldaten (Höhne 2003a, S.2):

- Die Bearbeitung ausgewählter Merkmale reduziert für alle Merkmalsträger die Möglichkeit der eindeutigen Zuordnung (bei Bearbeitung der Überschneidungsmerkmale), aber auch den potenziellen Informationsgewinn (bei Bearbeitung der sensiblen Merkmale).
- Die Bearbeitung aller Merkmale einzelner Merkmalsträger reduziert für diese sowohl die Zuordnungsmöglichkeit als auch den Informationsgewinn.
- Die Bearbeitung einzelner auffälliger Merkmalswerte von bestimmten Merkmalsträgern kann für diese sowohl das Zuordnungsrisiko (bei Bearbeitung der Überschneidungsmerkmale) als auch den Informationsgewinn (bei Bearbeitung von sensiblen Merkmalen) reduzieren.

- Die Bearbeitung aller Einzelwerte reduziert in jedem Fall sowohl die Zuordnungswahrscheinlichkeit als auch den Informationsgewinn.

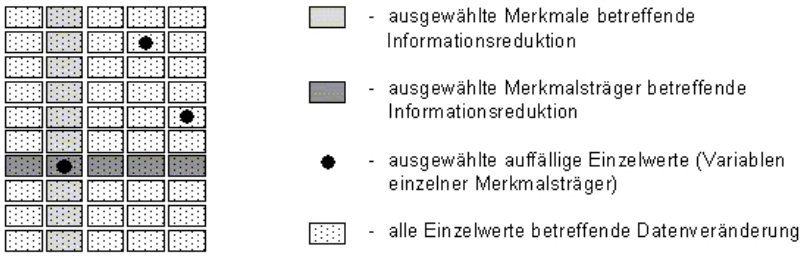

Abbildung 4.1: Ansätze der Veränderung von Mikrodaten

Im Laufe der Jahre und Jahrzehnte hat sich eine Vielzahl verschiedener Maßnahmen zur Anonymisierung von Daten entwickelt. Diese werden von der Wissenschaft bislang in die traditionellen und die datenverändernden Verfahren eingeteilt (Brand 2000; Statistische Ämter des Bundes und der Länder und IAW 2003). Diese Unterscheidung ist im Wesentlichen historisch gewachsen, da die traditionellen Verfahren bereits seit längerem in den statistischen Ämtern zum Einsatz kommen. Jedoch führt auch manch traditionelles Verfahren zu einer Veränderung von Einzelwerten, wie beispielsweise das Top-Coding. Allerdings basieren die aufgrund dieser Einteilung als datenverändernde Verfahren bezeichneten Maßnahmen meist auf einem systematischen Algorithmus, während die Datenveränderungen bei den traditionellen Verfahren eher inhaltlich motiviert sind.

Eindeutiger ist eine Einteilung der Verfahren, die darauf basiert, dass die Anonymisierung grundsätzlich in der Einschränkung oder in der Veränderung von Informationen besteht. Höhne (2003a) unterscheidet deshalb zwischen Informationsreduktion („Verschweigen“) und Informationsveränderung („Notlüge“). Eine weitere Systematisierung der Verfahren kann danach erfolgen, ob sie auf metrische oder auf kategoriale Variablen angewendet werden können. Diese Unterscheidung ist deshalb von besonderer Bedeutung, weil bei Unternehmens- und Betriebsdaten metrische Variablen vorherrschen, während bei Personen- und Haushaltsdaten vor allem kategoriale Variablen eine Rolle spielen.

In den Kapiteln 5 und 6 werden Verfahren und Verfahrensgruppen zur Anonymisierung von Einzeldaten vorgestellt. Dabei wird zunächst grundsätzlich danach unterschieden, ob Informationsreduktion (Kapitel 5) oder Datenveränderung (Kapitel 6) angewendet wird. Innerhalb der Verfahren, bei denen eine Informationsreduktion zur Anwendung kommt, wird danach unterschieden, ob diese merkmalsträger-, merkmals- oder ausprägungsbezogen vorgenommen wird. Die datenverändernden Verfahren werden im Wesentlichen danach unterschieden, ob sie auf kategoriale oder metrische Variablen angewendet werden können.

Kapitel 5

Verfahren zur Informationsreduktion

Die Verfahren in diesem Kapitel werden in der Regel bei der Anonymisierung von Haushalts- und Personendaten verwendet. Sie sind daher ausnahmslos auch den traditionellen Verfahren zuzuordnen (vgl. hierzu auch Brand (2000); Müller et al. (1991); Ronning et al. (2002)). Informationsreduktionen werden realisiert, indem Informationen unterdrückt oder vergröbert werden. Sie können grundsätzlich an einzelnen oder Gruppen von Merkmalsträgern (Abschnitt 5.1), an einzelnen oder mehreren Merkmalen (Abschnitt 5.2) oder an einzelnen Ausprägungen (Abschnitt 5.3) ansetzen.

5.1 Merkmalsträgerbezogene Verfahren zur Informationsreduktion

Merkmalsträgerbezogene Anonymisierungsverfahren verfolgen in der Regel das Ziel, besonders sensible Merkmalsträger – in diesem Fall Unternehmen oder Betriebe – zu schützen oder besonders auffällige Merkmalsträger vor einer Enthüllung zu bewahren.

Entfernen auffälliger Merkmalsträger: Ausreißer, d.h. besonders auffällige und daher reidentifikationsgefährdete Merkmalsträger werden entfernt. Dies sind Merkmalsträger, die einzigartige oder seltene Merkmalskombinationen aufweisen. Die entfernten Merkmalsträger können aus dem veröffentlichten Datensatz nicht mehr reidentifiziert werden.

Die Probleme bei der Analyse können vermieden werden, wenn auch der Nutzer die entsprechenden Beobachtungen vor Beginn der eigentlichen Analyse aus substanziellen oder statistischen Erwägungen entfernen würde. Schwierigkeiten ergeben sich jedoch, wenn sich die Kriterien zur Bestimmung von Ausreißern zwischen dem Nutzer und der veröffentlichenden Institution unterscheiden. Die einmal ausgeschlossenen Beobachtungen können auch nicht mehr durch spezielle Verfahren in die Analyse einbezogen werden. Damit werden möglicherweise besonders einflussreiche Beobachtungen bei Schätzungen unterdrückt.

Vor diesem Hintergrund führt die Entfernung der seltenen Beobachtungen aus dem Datensatz zu systematischen Veränderungen der Ergebnisse inferenzstatistischer Analysen.

Das Verfahren hat sich für die Anonymisierung von Personen- und Haushaltsdaten bewährt. Es muss jedoch bei der Anonymisierung von Wirtschaftsdaten beachtet werden, dass Ausreißer nicht nur häufiger vorkommen, sondern in der Regel auch einen größeren Einfluss auf das Verhalten der Aggregate haben (Ronning et al. 2002).

Allerdings können auch systematisch abgrenzbare Teilgesamtheiten eines Mikrodatenbestandes einem besonders hohen Reidentifikationsrisiko ausgesetzt sein. In diesem Fall ist auch eine systematische Einschränkung der Grundgesamtheit durch das Entfernen einer kompletten Teilgesamtheit als Anonymisierungsmaßnahme vorstellbar.

Systematische Einschränkung der Grundgesamtheit: Beispielsweise werden alle publizitätspflichtigen Unternehmen, eine komplette Branche oder, wie bei der Kostenstrukturerhebung und der Einzelhandelsstatistik im Projekt praktiziert, die Großunternehmen ab einer bestimmten Beschäftigtenzahl oder einem bestimmten Umsatz aus dem Datenbestand entfernt. Die entfernten Teilgesamtheiten können nicht mehr aus dem veröffentlichten Mikrodatenfile reidentifiziert werden. Allerdings sind für sie dann auch keine Analysen 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. Das Verfahren wird bei Personendaten manchmal zum Schutz von bestimmten Personengruppen, z.B. Abgeordneten, eingesetzt. Wichtiger erscheint die Anwendung des Verfahrens bei Wirtschaftsdaten, weil insbesondere Großbetriebe und Großunternehmen einem besonders hohen Reidentifikationsrisiko unterliegen. Hier könnte die Anwendung des Verfahrens dazu führen, dass wenigstens für Teile des ursprünglichen Datenbestandes ein Scientific-Use-File erstellt werden kann (Ronning et al. 2002).

Während das systematische Entfernen einzelner Merkmalsträger oder ganzer Teilgesamtheiten zwar deren absoluten Schutz gewährleistet, die Reidentifikationsgefahr der im Datenbestand verbliebenen Merkmalsträger aber nicht reduziert, verfolgt das zufällige Entfernen von Merkmalsträgern durch eine Stichprobenziehung das Ziel, den Schutz des gesamten Datenbestandes zu erhöhen.

(Sub-)Stichprobenziehung: Durch die Ziehung einer (Sub-)Stichprobe wird die Teilnahmewahrscheinlichkeit jedes Merkmalsträgers verringert.

Eine (Sub-)Stichprobe kann auch mit ungleichen Auswahlwahrscheinlichkeiten (beispielsweise in unterschiedlichen Beschäftigtengrößenklassen) gezogen wer-

den. Des Weiteren sind Ziehungen mit Zurücklegen denkbar. Damit besteht die Möglichkeit, dass auch in der Grundgesamtheit einzigartige Elemente mehrmals in die Stichprobe gelangen.
Bei einer Vollerhebung soll mit Hilfe einer Stichprobenziehung gewährleistet werden, dass die Wahrscheinlichkeit einer Reidentifikation dadurch vermindert wird, dass der Angreifer nicht weiß, ob sein Ziel überhaupt im veröffentlichten Datensatz enthalten ist. Er weiß lediglich, mit welcher Wahrscheinlichkeit sich der gesuchte Merkmalsträger in der Stichprobe befindet.

Für den Fall einer Sub-Stichprobenziehung wird gewährleistet, dass sich die Wahrscheinlichkeit dafür, dass der gesuchte Merkmalsträger Teil des veröffentlichten Datenfiles ist, zusätzlich verringert. Selbst wenn einem Angreifer bekannt ist, dass ein Merkmalsträger an einer Erhebung teilnimmt, weiß er bei einer Stichprobenziehung nicht, ob der gesuchte Merkmalsträger Teil des veröffentlichten Datenfiles ist. Zwar kann die (Sub-)Stichprobenziehung zu einer Erhöhung der (eindeutigen) Zuordnungen führen, diese sind jedoch mit größerer Wahrscheinlichkeit falsch. Durch eine Ziehung mit Zurücklegen kann die Schutzwirkung weiter verbessert werden.

Sub-Stichprobenziehungen haben sich für die Anonymisierung von Personen- und Haushaltsdaten bewährt (z.B. Mikrozensus). Ihr Nutzen für die Anonymisierung von Unternehmensdaten muss als deutlich geringer eingeschätzt werden, weil bei Unternehmensdaten die Grundgesamtheiten kleiner und die Stichprobenauswahlsätze größer sind (Brand 2000; Ronning et al. 2002). Außerdem dürften Stichprobenziehungen bei Unternehmens- und Betriebsdaten aufgrund der schieferen Verteilungen erheblich größere Auswirkungen auf das Analysepotenzial nach sich ziehen.

5.2 Merkmalsbezogene Verfahren zur Informationsreduktion

Merkmalsbezogene Anonymisierungsmaßnahmen behandeln im Gegensatz zu merkmalsträgerbezogenen Maßnahmen einzelne oder mehrere Merkmale bzw. Variablen. Sie werden in der Regel auf Überschneidungsmerkmale angewendet, um eine Zuordnung zu verhindern, oder auf besonders sensible Merkmale, um die wahren und exakten Werte vor Enthüllungen zu schützen. Dabei können Merkmale komplett entfernt, ersetzt oder geeignet vergröbert werden.

Verfahren, bei denen Merkmale beseitigt, ersetzt oder zusammengefasst werden: Die Merkmale werden vollständig beseitigt (Variablenunterdrückung) oder durch adäquate Linearkombinationen, Kennziffern oder Indizes ersetzt. Für die Ersetzung von Merkmalen bestehen die folgenden Möglichkeiten:

- Konstruktion von neuen Merkmalen aus mehreren ursprünglichen Merk-

malen beispielsweise durch die Bildung von Linearkombinationen (z.B. Bildung der Summe aus Inlands- und Auslandsumsatz),

- Bildung von statistisch interpretierbaren Beziehungs- bzw. Verhältniszahlen als Kennziffern (z.B. Bestand an Handelsware am Jahresanfang bezogen auf den Jahresumsatz),
- Indexbildung auf einer plausiblen Basis, insbesondere bei Zeitreihen- und Paneldaten (z.B. Jahresumsatz des Jahres 1999 bezogen auf den Jahresumsatz des Jahres 1980).

Während die Variablenunterdrückung gleichermaßen für metrische wie kategoriale Variablen anwendbar ist, lässt sich die Ersetzung von Variablen durch Linearkombinationen, Beziehungs- und Verhältniszahlen sowie Indizes nur für metrische Variablen realisieren.

Die Schutzwirkung dieser Verfahren beruht allein auf der Verringerung der Informationen im Datensatz. Werden durch die Anwendung dieser Verfahren Überschneidungsmerkmale entfernt, so sinkt die Zuordnungswahrscheinlichkeit. Werden hingegen sensible Variablen aus dem Datensatz entfernt, so werden die Anreize verringert, eine Enthüllung vorzunehmen, da der Nutzen einer Reidentifikation für den potenziellen Angreifer sinkt.
Bedingung für die Durchführung einer Ersetzung von Merkmalen ist, dass die ökonomische Interpretierbarkeit der neu gebildeten Variablen und ihre sinnvolle Verwendbarkeit für Schätzungen sichergestellt werden kann. Die Unterdrückung von Variablen, d.h. ihre Entfernung aus dem Datensatz, ist für die rein statistische Analyse des restlichen Datensatzes ohne Auswirkungen. Nachteilig an diesem Verfahren ist jedoch der für einzelne empirische Fragestellungen erhebliche Informationsverlust, der gerade in Regressionsanalysen auch zu Fehlspezifikationen aufgrund unterdrückter Variablen führen kann. Bei der Konstruktion neuer Variablen ergeben sich keine Abweichungen der Ergebnisse, sofern zur ökonomischen Analyse ausschließlich die konstruierte Variable benötigt wird. Sind zur Überprüfung des ökonomischen Modells jedoch die ursprünglichen Informationen notwendig, so sind die formulierten Schätzansätze aufgrund unterdrückter Variablen bzw. falscher Parameterrestriktionen fehlspezifiziert (Ronning et al. 2002).

Vergröberung von Merkmalsausprägungen (Gruppierung, Zusammenfassung): Bei der Vergröberung von Merkmalsausprägungen existieren in Abhängigkeit von der Skalierung der Merkmale unterschiedliche Ansätze.

- Gruppierung von metrischen Merkmalen (z.B. Bildung von Beschäftigtengrößenklassen oder Umsatzgrößenklassen),
- Vergröberung durch Rundung der Werte metrischer Variablen (z.B. Rundung von Umsatzwerten auf ganze Tausenderbeträge),

- Zusammenfassung bereits existierender Kategorien (z.B. Zusammenfassung von zwei benachbarten Beschäftigtengrößenklassen).

Durch die Vergröberung der Merkmalsausprägungen beziehungsweise durch die Zusammenfassung von Ausprägungskategorien bei Überschneidungsmerkmalen verringert sich die Anzahl der möglichen Schlüsselvariablenkombinationen. Eine Verringerung der Ausprägungsmöglichkeiten bei einer gleichbleibenden Zahl von Fällen führt dazu, dass die Wahrscheinlichkeit des Auftretens von einzigartigen Ausprägungskombinationen sinkt. Zugleich steigt die Zahl der identischen Beobachtungen. Durch die verringerte Zahl von Merkmalsausprägungen erhöht sich zugleich auch die Wahrscheinlichkeit von übereinstimmenden Ausprägungskombinationen zwischen Mikrodaten- und Identifikationsfile. Hierdurch kann sich die Zahl der auf der Basis von identischen Ausprägungskombinationen vorgenommenen Zuordnungen – und zwar sowohl der ein- als auch der mehrdeutigen – erhöhen. Für einen Datenangreifer ergibt sich hieraus jedoch eine erhöhte Unsicherheit, da die Wahrscheinlichkeit von Falschzuordnungen steigt. Außerdem sinkt der Nutzen durch eine Enthüllung, weil mit der Vergröberung ein Informationsverlust verbunden ist. Somit sinken auch die Anreize, eine Enthüllung zu versuchen.
Auf der anderen Seite wird der Informationsgehalt der solcherart bearbeiteten Merkmale erheblich verringert. Insbesondere die Umwandlung metrischer in kategoriale Variablen kann komplette Analysen ausschließen, vor allem wenn zur Modellspezifikation die Variable in metrischer Form erforderlich ist. Zudem werden auch die deskriptiven Statistiken des Datensatzes erheblich beeinflusst. Der Informationsverlust hängt dabei entscheidend vom Grad der Vergröberung ab.

5.3 Ausprägungsbezogene Verfahren zur Informationsreduktion

Beim ausprägungsbezogenen Vorgehen zur Informationsreduktion handelt es sich in der Regel um die Unterdrückung einzelner Werte (local suppression). Dies geschieht meist bei Beobachtungen mit Ausprägungen oder Ausprägungskombinationen, die in der Stichprobe sehr selten oder einzigartig sind. Durch die Unterdrückung entstehen „Missing Values“. Damit ist keine Neukodierung der gesamten Variablen erforderlich, vielmehr bleibt die Variablendefinition des Ausgangsdatensatzes erhalten. Dies kann insbesondere dann sinnvoll sein, wenn Probleme bei der Anonymisierung von Werten für einzelne Großbetriebe und Großunternehmen bestehen. Variablenwerte können sowohl bei metrischen als auch bei kategorialen Variablen unterdrückt werden.
Die Schutzfunktion dieses Verfahrens besteht zum einen in der direkten Verminderung der Zahl der Schlüsselvariablenkombinationen, insbesondere bei diskreten bzw. kategorialen Variablen. Durch die Unterdrückung sind vorher seltene oder einmalige Schlüsselvariablenkombinationen nicht mehr aufdeckbar. Zum anderen können sensible Informationen

für einzelne Beobachtungen unterdrückt werden, sofern es sich um seltene Ausprägungen handelt (Ronning et al. 2002).

Eine Alternative zur Unterdrückung einzelner Werte und der Erzeugung von „Missing Values“ besteht in der Ersetzung einzelner Werte durch anonymisierte Werte. Die Ersetzung folgt dann in der Regel auf Basis eines merkmalsbezogenen Anonymisierungsverfahrens, sei es durch die Vergröberung von Merkmalen oder durch die Anwendung datenverändernder Verfahren, die jedoch ausschließlich für bestimmte Merkmalswerte zur Anwendung kommen. Hierfür kommen insbesondere Imputationsverfahren in Frage (vgl. Unterabschnitt 6.2.2).

Kapitel 6

Datenverändernde Anonymisierungsverfahren

6.1 Datenverändernde Anonymisierungsverfahren für kategoriale Variablen

6.1.1 Vertauschungsverfahren (Swapping) bei kategorialen Variablen

Die Swapping-Verfahren basieren auf der Vertauschung von Werten. Sie sind sowohl für metrische als auch für kategoriale Variablen durchführbar (siehe deshalb auch Unterabschnitt 6.2.1). Die Vertauschungen zwischen den Merkmalsträgern erfolgen für alle Variablen getrennt. Die Vertauschungen können auf einen bestimmten Vertauschungsbereich beschränkt sein. Außerdem können Vertauschungen auch so ausgestaltet werden, dass sie nur innerhalb bestimmter Ausprägungskombinationen für die nicht getauschten Merkmale vorgenommen werden. Durch die Anwendung des Verfahrens wird sowohl die Zuordnung erschwert als auch die enthüllte Information unbrauchbar.

Da sich bei Vertauschungsverfahren die Merkmalswerte in jedem Fall ändern, bedeutet die Anwendung der Verfahren bei kategorialen Variablen eine sehr starke Informationsveränderung. Es bietet sich daher eher an, die Merkmalsausprägungen nur mit einer festgelegten Wahrscheinlichkeit zu verändern. Dies ist bei der Post-Randomisierung der Fall.

6.1.2 Post-Randomisierung

Beim Verfahren der Post-Randomisierung (PRAM) werden diskrete Merkmale durch die Definition von Übergangswahrscheinlichkeiten randomisiert (Kooiman et al. 1997; Willenborg und de Waal 2001).

Das Verfahren entspricht der bei Erhebungen verwendeten Randomisierung von Antworten, die durchgeführt werden, um zu erreichen, dass die Befragten auch auf sensible Fragen

antworten, beispielsweise nach dem Drogenkonsum oder einer Aids-Erkrankung (Warner 1965). Särndal et al. (1992, S.573) schlugen vor, diese Methode zu verwenden, um die Anonymität von Individuen zu schützen. Während im Fall randomisierter Antworten das stochastische Modell allerdings definiert werden muss, bevor die Daten erhoben werden, wird bei der Post-Randomisierung die Methode auf die bereits erhobenen Daten angewendet (Van den Hout und van der Heijden 2002).

Bei der Post-Randomisierung werden die Merkmalswerte mit bei der Anwendung festzulegenden Übergangswahrscheinlichkeiten in andere Ausprägungen transformiert. Für den Fall einer dichotomen Variablen ist die Matrix der Übergangswahrscheinlichkeiten in Gleichung (6.1) dargestellt (Ronning et al. 2002).

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \tag{6.1}$$

Im Folgenden wird die randomisierte dichotome Variable mit Y^a bezeichnet, ihre Realisierung mit y^a. Für die Übergangswahrscheinlichkeiten gilt dann Gleichung (6.2):

$$p_{jk} \equiv P(Y^a = j \,|\, Y = k) \text{ mit } j,k \,\varepsilon\, \{0,1\} \text{ und } p_{j0} + p_{j1} = 1 \text{ für } j = 0,1 \tag{6.2}$$

Bei einer symmetrischen Übergangsmatrix gilt Gleichung (6.3). Für den Fall mehrerer Kategorien bei ordinalen Merkmalen kann es sinnvoll sein, für den Tausch mit benachbarten Kategorien höhere Übergangswahrscheinlichkeiten vorzusehen.

$$P = \begin{pmatrix} \pi & 1-\pi \\ 1-\pi & \pi \end{pmatrix} \tag{6.3}$$

Eine Modifikation der Post-Randomisierung – das invariante PRAM – wurde von Höhne (2003a) vorgeschlagen und von Ronning und Rosemann (2004) sowie Ronning et al. (2005) formal dargestellt. Beim invarianten PRAM werden die Randverteilungen erhalten. Der Erwartungswert der anonymisierten Variablen entspricht dem Erwartungswert der Originalvariablen.

Damit gilt für den Fall von zwei Kategorien.

$$E[Y^a] = E[Y]. \tag{6.4}$$

und für den Fall mehrerer Kategorien

$$E[Y_j^a] \quad = \quad E[Y_j] \quad = \quad \theta_j \ , \ j = 1, \ldots, r. \tag{6.5}$$

Dies erfordert folgenden Zusammenhang:

$$\theta_k^a \ \equiv \ P(Y^a = k) \ = \ \sum_{j=1}^{r} \ P(Y^a = k | Y = j) \ P(Y = j) \ = \ \sum_j p_{jk} \ \theta_j, \quad k \ = \ 1, 2, \ldots, r \tag{6.6}$$

oder kompakter geschrieben

$$\boldsymbol{\theta}^a \quad = \quad \mathbf{P} \, \boldsymbol{\theta} \quad . \tag{6.7}$$

Dabei sind $\boldsymbol{\theta}$ und $\boldsymbol{\theta}^a$ Vektoren. Im Fall des invarianten PRAMS ist erforderlich, dass $\boldsymbol{\theta} \ = \ \boldsymbol{\theta}^a$ gilt. Somit muss die Matrix $\mathbf{P}$ so gewählt werden, dass

$$\boldsymbol{\theta} \quad = \quad \mathbf{P} \, \boldsymbol{\theta} \tag{6.8}$$

gilt.

Äquivalent hierzu ist:

$$P(Y^a \ = \ k) \quad = \quad P(Y \ = \ k) \tag{6.9}$$

Da die Matrix $\mathbf{P}$ $r \, (r - 1)$ unabhängige Parameter besitzt, aber das Gleichungssystem aus Gleichung (6.8) lediglich aus r Gleichungen besteht, existiert keine eindeutige Lösung für das invariante PRAM. Eine Möglichkeit besteht darin, zunächst die Wahrscheinlichkeit λ dafür festzulegen, dass eine Einheit überhaupt für einen Tausch in Frage kommt $(0 < \lambda < 1)$, und dann die tatsächlichen Tauschwahrscheinlichkeiten wie folgt zu definieren:

$$\begin{aligned} p_{ii} \quad &= \quad \lambda \ + \ (1 \ - \ \lambda) \, \theta_i, i = 1, \ldots, r \\ p_{ij} \quad &= \quad (1 \ - \ \lambda) \, \theta_j, j = 1, \ldots, r \ ; \ i \neq j \ . \end{aligned} \tag{6.10}$$

Für eine beliebige Anzahl an Kategorien k folgt aus dieser Definition:

$$\theta_k^a \quad = \quad \sum_j p_{jk} \ \theta_j \quad = \quad [\lambda + (1 - \lambda)\theta_k] \ \theta_k \ + \ \sum_{j \neq k} [(1 - \lambda) \, \theta_k] \ \ \theta_j \quad = \quad \theta_k \ . \tag{6.11}$$

In dem Spezialfall von zwei Kategorien $(r = 2)$ gilt damit für die Übergangsmatrix $\boldsymbol{P}$

$$\begin{aligned} \boldsymbol{P} &= \begin{pmatrix} \lambda + (1-\lambda)\theta_1 & (1-\lambda)\theta_2 \\ (1-\lambda)\theta_1 & \lambda + (1-\lambda)\theta_2 \end{pmatrix} \\ &= \begin{pmatrix} \lambda + (1-\lambda)\theta & (1-\lambda)(1-\theta) \\ (1-\lambda)\theta & \lambda + (1-\lambda)(1-\theta) \end{pmatrix} \end{aligned} \tag{6.12}$$

In der Praxis werden die unbekannten Wahrscheinlichkeiten θ_j durch die beobachtbaren relativen Häufigkeiten ersetzt.

Post-Randomisierung führt dazu, dass die veröffentlichten Werte nur noch mit einer durch das Verfahren festgelegten Wahrscheinlichkeit den Werten im Originaldatensatz entsprechen. Da kategoriale Variablen häufig als Zusatzwissen verwendet werden können[2], senkt das Verfahren vorrangig die Zuordnungswahrscheinlichkeit bei Datenangriffen.

Die univariaten Verteilungen und die Zellwerte (Häufigkeiten) in beliebigen Kreuztabellen können konsistent geschätzt werden, sofern dem Nutzer die Matrix der Übergangswahrscheinlichkeiten bekannt ist. Bei der Anwendung multivariater Verfahren muss beachtet werden, dass im veränderten Datensatz die Originalvariablen nur noch als latente Variablen analysierbar sind. Grundsätzlich können Schätzungen mit Statistikpaketen durchgeführt werden, die es erlauben, das Vorhandensein randomisierter Antworten oder latenter Variablen explizit einzubeziehen. Die Anwendung von nicht modifizierten Standardverfahren führt dagegen zu fehlerhaften Schätzungen.

Wesentlich ist, dass das Verfahren der Post-Randomisierung die einzige Anonymisierungsmethode für diskrete bzw. kategoriale Variablen darstellt, der ein stochastisches Modell zugrunde liegt.

6.1.3 Risikoorientierte Veränderung kategorialer Variablen durch das SAFE-Verfahren

Die Idee des Verfahrens SAFE besteht darin, einen Datenbestand zu erzeugen, in dem jeder Merkmalsträger bezüglich aller betrachteten Merkmale mit mindestens zwei weiteren Merkmalsträgern identisch ist (Evers und Höhne 1999; Höhne 2003a,b,c). Bei den diskreten beziehungsweise kategorialen Variablen wird dies durch das gezielte Verändern von Merkmalswerten realisiert. Dabei wird nach dem Kriterium minimaler Fehler in den Randsummen

2) Insbesondere eignet sich diskretes Zusatzwissen zur Blockung des Datensatzes bei Datenangriffen (Lenz 2003a). Dabei wird der Datensatz nach den Ausprägungen der diskreten Merkmalen aufgeteilt und das Matching-Verfahren, also die Zuordnung nach den metrischen Merkmalen, auf die einzelnen Blöcke getrennt durchgeführt.

eine diskrete Basisdatei erstellt, in der alle Ausprägungskombinationen diskreter Variablen mit mindestens drei Einheiten besetzt sind. Anschließend werden die Merkmalswerte der metrischen Variablen neu den Merkmalskombinationen der diskreten Variablen zugeordnet. Dies geschieht unter den Gesichtspunkten der Erhaltung der größten Ähnlichkeit in den diskreten Variablen und dem Verschieben der kleinsten Merkmalswerte der metrischen Variablen. Bei den metrischen Variablen wird eine Form der Mikroaggregation durchgeführt (vgl. Unterabschnitt 6.2.4).

Der Algorithmus wurde ursprünglich für die Tabellengeheimhaltung entwickelt. SAFE reduziert das Risiko der eindeutigen Zuordnung von Einheiten, da immer mehrere gleiche Einheiten vorhanden sind (Ronning et al. 2002; Höhne 2003a; Statistische Ämter des Bundes und der Länder und IAW 2003).

6.2 Datenverändernde Anonymisierungsverfahren für metrische Variablen

6.2.1 Vertauschungsverfahren (Swapping) bei metrischen Variablen

Wie bereits in Unterabschnitt 6.1.1 erwähnt, können die Vertauschungsverfahren grundsätzlich gleichermaßen für kategoriale wie für metrische Variablen verwendet werden. Auch bei den metrischen Variablen können die Vertauschungen so beschränkt werden, dass sie nur innerhalb bestimmter Ausprägungskombinationen vorgenommen werden. Sortiert man die Merkmalswerte für jede einzelne Variable nach ihrer Größe und definiert gleichzeitig Nachbarschaftsbereiche, auf die der Tausch beschränkt wird, so kann der annähernde Erhalt der Rangstatistiken garantiert werden. Diese Verfahrensvariante wird daher auch – in Abgrenzung zum allgemeinen Data Swapping – als Rank Swapping bezeichnet.

Werden metrische Überschneidungsmerkmale mit Swapping-Verfahren bearbeitet, so wird in erster Linie das Zuordnungsrisiko reduziert. Werden auch andere metrische Variablen mit den Verfahren anonymisiert, so geht damit auch die Reduzierung des Nutzens einer Enthüllung für den potenziellen Datenangreifer einher.

Die univariaten Verteilungen werden durch das Verfahren erhalten, während die gemeinsame Verteilung deutlich verändert wird. Die Veränderung ist dabei umso stärker, je größer die Bandbreite ist, innerhalb derer die Vertauschung stattfindet (Ronning et al. 2002; Rosemann 2003).

Modifikationen des Verfahrens können die Auswirkungen auf das Analysepotenzial verbessern (Gottschalk 2005, S.48). So schlagen Dalenius und Reiss (1982) ein Verfahren vor, bei dem die Randverteilungen der Häufigkeitsverteilung von jeweils zwei bearbeiteten Variablen erhalten bleiben. Damit lassen sich die bivariaten Verteilungen erhalten, allerdings nicht die höherdimensionalen. Außerdem können die Vertauschungen unter Nebenbedingungen statt-

finden, die den Erhalt der Varianz-Kovarianzmatrix sicherstellen (Kim und Winkler 1995, 1997).

6.2.2 Imputationsverfahren

Imputationsverfahren bestehen in einem Austausch von Angaben durch eingeschätzte Werte. Diese Idee wurde zuerst von Rubin (1993) vorgeschlagen und baut auf den Imputationsverfahren im Falle von fehlenden Antworten („Missing Values“) auf, die im Rahmen der Nonresponseforschung entwickelt wurden (Fienberg 1997). Im Unterschied zur klassischen Anwendung der Imputation bei „Missing Values“ werden hier nicht fehlende Angaben, sondern besonders sensible Merkmalswerte oder Merkmalswerte von Schlüsselvariablen durch die eingeschätzten Werte ersetzt. Dabei können einzelne Merkmalswerte, die Merkmalswerte besonders gefährdeter Merkmalsträger oder alle Merkmalswerte einzelner Variablen anonymisiert werden.

Es können Imputationen auf der Basis von parametrischen und nicht parametrischen Regressionsmodellen gewählt werden (Pollettini et al. 2002). Dabei muss stets ein bestimmter (Regressions-)Zusammenhang unterstellt werden.

Es wird zwischen einfacher Imputation (single Imputation) und multipler Imputation (multiple Imputation) unterschieden. Während bei der einfachen Imputation die Einschätzung auf Basis eines einmal unter Einbeziehung aller vorhandenen Beobachtungen geschätzten Regressionsmodells vorgenommen wird, werden bei der multiplen Imputation Bootstrap-Schätzer ermittelt, indem die Regressionsschätzung mit k Bootstrap-Stichproben durchgeführt wird (Rubin und Schenker 1991; Little 1993; Raghunathan et al. 2003). In diesem Fall bekommen die Datennutzer mehrere anonymisierte Datensätze zur Verfügung gestellt. Dies kann sich allerdings negativ auf das Reidentifikationsrisiko auswirken.

Die Folgen für die Schutzwirkung sind schwer abschätzbar. Sie hängen insbesondere davon ab, welches Verfahren konkret zur Anwendung kommt und wie stark die anonymisierten Werte den Originalwerten ähneln. Da, wie bereits oben erwähnt wurde, die Einschätzung der Werte üblicherweise auf Basis eines unterstellten Regressionszusammenhangs erfolgt, und diese häufig den anschließend im Rahmen von Analysen mit den eingeschätzten Daten geschätzten Regressionsmodellen ähneln, kann es bei Anwendung solcher Verfahren zu einer systematischen und massiven Überschätzung des Bestimmtheitsmaßes und damit des Erklärungsgehalts des geschätzten Modells kommen (Ronning et al. 2002).

Imputationsverfahren wurden beispielsweise von Abowd und Woodcock (2002) zur Anonymisierung eines französischen Linked Employer-Employee-Paneldatensatzes des INSEE (Institut National de la Statistique et des Etudes Economiques) verwendet.

6.2.3 Stochastische Überlagerung

Stochastische Überlagerungen beziehungsweise Überlagerungen mit Zufallsfehlern stellen eine umfangreiche Verfahrensgruppe zur Anonymisierung von Einzeldaten dar. Die Grundidee besteht darin, dass zu den metrischen Variablen eines Datensatzes Zufallszahlen addiert oder die Merkmalswerte mit Zufallszahlen multipliziert werden, so dass die Originalwerte durch die überlagerten Werte ersetzt werden.

Bei einer stochastischen Überlagerung entsprechen die veröffentlichten Werte mit der Wahrscheinlichkeit Null den Originalwerten. Für jeden Originalwert können jedoch Intervalle um den überlagerten Wert ermittelt werden, in denen dieser mit vorgegebener Wahrscheinlichkeit liegt.

Die stochastischen Überlagerungen unterteilen sich grundsätzlich in additive und multiplikative Überlagerungen. Variiert werden kann auch die Verteilung des Zufallsfehlers. Additive Zufallsfehler sind in der Regel normalverteilt mit einem Erwartungswert von Null. Neben der Überlagerung mit einer einfachen Normalverteilung ist jedoch auch die Überlagerung mit einem Zufallsfehler möglich, der aus einer Mischungsverteilung aus mehreren Normalverteilungen gezogen wird. Daneben ist es denkbar, die Varianz der Zufallsfehler in Abhängigkeit von den zu anonymisierenden Merkmalswerten zu variieren (heteroskedastische additive Überlagerung). Die gleichen Verteilungen sind auch bei multiplikativen Überlagerungen verwendbar. Allerdings muss dann eine nicht-negative Verteilung mit Erwartungswert Eins gewählt werden. Alternativ kann der multiplikative Zufallsfehler auch gleichverteilt sein, einer Lognormalverteilung oder einer gestutzten Normalverteilung entstammen.

a) Additive stochastische Überlagerung

a1) Additive stochastische Überlagerung mit einer Normalverteilung

Bei einer additiven stochastischen Überlagerung mit einer Normalverteilung werden die einzelnen Merkmalswerte mit einem Zufallsfehler überlagert, dessen Erwartungswert den Wert Null aufweist und dessen Varianz beziehungsweise Varianz-Kovarianzmatrix konstant ist. Die naive additive Überlagerung lässt sich damit für die gesamte Datenmatrix $\mathbf{X}$ wie folgt darstellen (vgl. u.a. Höhne (2004a)):

$$\boldsymbol{X}^a = \boldsymbol{X} + \boldsymbol{W} \tag{6.13}$$

mit

$$\boldsymbol{W} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{ww}), \tag{6.14}$$

wobei $\mathbf{X}^a$ und $\mathbf{W}$ dieselbe Dimension aufweisen wie $\mathbf{X}$.

Wird die Kovarianzmatrix der Überlagerungen proportional zur Kovarianzmatrix der Originalwerte gewählt, gilt also Gleichung (6.15), so wird auch die Kovarianzmatrix der anonymisierten Daten proportional zur Kovarianzmatrix der Originalwerte (Gleichung (6.16)).

$$\mathbf{\Sigma}_{ww} = d\mathbf{\Sigma}_{xx} \tag{6.15}$$

$$\mathbf{\Sigma}_{x^a x^a} = (d+1)\mathbf{\Sigma}_{xx} \tag{6.16}$$

Damit wird die Korrelationsmatrix erwartungstreu geschätzt. In linearen Regressionsmodellen werden die Regressionsschätzer somit asymptotisch erwartungstreu, sofern alle Variablen (einschließlich der abhängigen) additiv überlagert werden. Die Varianz der Störgrößen wird jedoch um den Faktor $(1 + d)$ überschätzt (Brand 2000).

Für die Matrix der Störterme W gilt damit:

$$\boldsymbol{W} = \boldsymbol{V}\sqrt{d}\ \mathbf{\Sigma}_{xx}^{1/2} \tag{6.17}$$

Dabei ist $\boldsymbol{V}$ eine Matrix aus untereinander unkorrelierten standardnormalverteilten Zufallsvariablen.

Daraus folgt für die Matrix der anonymisierten Werte:

$$\boldsymbol{X}_1^a = \boldsymbol{X} + \boldsymbol{W} \tag{6.18}$$

oder

$$\boldsymbol{X}_1^a = \boldsymbol{X} + \boldsymbol{V}\sqrt{d}\ \mathbf{\Sigma}_{xx}^{1/2}. \tag{6.19}$$

Kim (1986) schlägt vor, anschließend die Varianz-Kovarianzmatrix der Originaldaten durch die folgende Operation wieder herzustellen:

$$\boldsymbol{X}_2^a = \frac{1}{\sqrt{1+d}}\boldsymbol{X}_1^a + \left(1 - \frac{1}{\sqrt{1+d}}\right)\underline{1}\boldsymbol{\mu}_x \tag{6.20}$$

mit $\boldsymbol{\mu}_x$ als Vektor der Erwartungswerte von $\boldsymbol{X}$.

a2) Additive Überlagerung mit einer Mischungsverteilung

Ein Problem der additiven stochastischen Überlagerung mit einer einfachen Normalverteilung besteht darin, dass der Zufallsfehler mit einer hohen Wahrscheinlichkeit Werte nahe Null annimmt, die überlagerten Werte folglich auch mit einer hohen Wahrscheinlichkeit nahe bei den Originalwerten liegen. Will man das ändern, so kann man eine höhere Varianz verwenden. Dies birgt allerdings das Risiko, dass einzelne Werte sehr stark von den entsprechenden Originalwerten abweichen. Deshalb hat Roque (2000) vorgeschlagen, bei der Überlagerung anstatt einer einfachen Normalverteilung eine Mischung aus normalverteilten Zufallswerten zu nutzen. Damit kann bei gleicher Varianz erreicht werden, dass ein größerer Anteil der überlagerten Werte weiter von den Originalwerten entfernt ist.

Es seien V_1 und V_2 zwei stetige Zufallsvariablen mit Dichtefunktionen f_1 und f_2 sowie Erwartungswerten μ_i und Varianzen σ_i^2 für $i = 1, 2$. Eine Zufallsvariable W entstammt dann einer Mischung der Verteilungen von V_1 und V_2, falls ihre Dichtefunktion durch

$$g(w) = \alpha f_1(w) + (1 - \alpha) f_2(w) \tag{6.21}$$

gegeben ist ($0 < \alpha < 1$).

Damit kann man sich eine Mischungsverteilung gedanklich auch wie folgt vorstellen: „Es gibt zwei (bzw. allgemeiner k) Zustände, die jeweils mit Wahrscheinlichkeit α_i auftreten und die sich gegenseitig ausschließende Ereignisse darstellen. Für jeden Zustand gibt es eine Verteilung der Zufallsvariablen W. Je nachdem welcher Zustand eintritt, wird der Wert der Zufallsvariablen W aus der betreffenden Verteilung generiert. Diese Modellvorstellung nutzt man aus, um Zufallszahlen aus Mischungsverteilungen zu erzeugen“ (Ronning 2004b).[3]

Für den Erwartungswert von W ergibt sich damit bei einer Mischungsverteilung aus zwei Normalverteilungen:

$$E[W] = \alpha\mu_1 + (1 - \alpha)\mu_2. \tag{6.22}$$

Weiterhin gilt für den Erwartungswert der quadrierten Zufallsvariablen:

$$E\left[W^2\right] = \alpha E\left[V_1^2\right] + (1 - \alpha) E\left[V_2^2\right]. \tag{6.23}$$

3) vgl. auch z.B. McLachlan und Peel (2000).

Wegen $var[W] = E\left[W^2\right] - E[W]^2$ folgt für die Varianz von W:

$$var[W] = \alpha E\left[V_1^2\right] + (1-\alpha)E\left[V_2^2\right] - (\alpha\mu_1 + (1-\alpha)\mu_1)^2 \tag{6.24}$$

und damit

$$var[W] = \alpha\sigma_1^2 + (1-\alpha)\sigma_2^2 + 2\alpha(1-\alpha)\mu_1\mu_2. \tag{6.25}$$

Für den allgemeinen Fall einer Mischungsverteilung aus k Normalverteilungen ergibt sich für Erwartungswertvektor und Varianz-Kovarianzmatrix (Ronning 2004b):

$$E[\boldsymbol{W}] = \sum_{i=1}^{k} \alpha_i \boldsymbol{\mu}_i, \tag{6.26}$$

$$\boldsymbol{\Sigma}_{ww} = \sum_{i=1}^{k} \alpha_i \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i\boldsymbol{\mu}_i'\right) - \left(\sum_{i=1}^{k} \alpha_i\boldsymbol{\mu}_i\right)\left(\sum_{i=1}^{k} \alpha_i\boldsymbol{\mu}_i\right)'. \tag{6.27}$$

Bei der additiven stochastischen Überlagerung mit einer Mischungsverteilung wird die Verteilung der Störterme aus Gleichung (6.14) durch die folgende allgemeine Annahme über die Verteilung der Störterme ersetzt:

$$\boldsymbol{W} \sim \sum_{i=1}^{k} \alpha_i N\left(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\right). \tag{6.28}$$

Aus der Bedingung, dass der Erwartungsvektor der Mischungsverteilung gleich dem Nullvektor sein soll, ergeben sich die folgenden Restriktionen (Roque 2000; Ronning 2004b):

$$\sum_{i=1}^{k} \alpha_i\boldsymbol{\mu}_i = \mathbf{0} \tag{6.29}$$

oder auch

$$\sum_{i=1}^{k} \alpha_i\mu_{ji} = 0. \tag{6.30}$$

Aus Gleichung (6.27) in Verbindung mit Gleichung (6.29) folgt außerdem:

$$\boldsymbol{\Sigma}_{ww} = \sum_{i=1}^{k} \alpha_i \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i'\right). \tag{6.31}$$

Verwendet man nach wie vor die Annahme, dass die Varianz-Kovarianzmatrix der Störterme proportional zur Varianz-Kovarianzmatrix der Originalvariablen ist (Gleichung (6.15)), so gilt:

$$d\,\boldsymbol{\Sigma}_{xx} = \sum_{i=1}^{k} \alpha_i \left(\boldsymbol{\Sigma}_i + \boldsymbol{\mu}_i \boldsymbol{\mu}_i'\right). \tag{6.32}$$

Roque (2000) nimmt nun an, dass die Kovarianzmatrizen aller Mischungskomponenten proportional zur Kovarianzmatrix $\boldsymbol{\Sigma}_{xx}$ der Originalmerkmale sind, d.h.

$$cov\left[\boldsymbol{V}_i\right] = \boldsymbol{\Sigma}_i = d_i \boldsymbol{\Sigma}_{xx}. \tag{6.33}$$

Daraus ergibt sich für die Gleichung (6.32):

$$\sum_{i=1}^{k} \alpha_i \boldsymbol{\mu}_i \boldsymbol{\mu}_i' = \left(d \; - \; \sum_{i=1}^{k} \alpha_i d_i\right) \boldsymbol{\Sigma}_{xx}. \tag{6.34}$$

Um sicherzustellen, dass beide Seiten der Gleichung (6.33) positiv definit sind, muss auch der skalare Faktor auf der rechten Seite der Gleichung positiv sein. Dies ergibt die folgende Restriktion (vgl. hierzu auch Ronning (2004b)):

$$d \; > \; \sum_{i=1}^{k} \alpha_i d_i. \tag{6.35}$$

Roque (2000) leitet aus der Gleichung (6.33) zudem die wesentliche Forderung ab, dass die Anzahl der Mischungskomponenten (k) mindestens so groß sein muss wie die Anzahl der zu anonymisierenden Merkmale (r), weil die rechte Seite der Gleichung in jedem Fall Rang r besitzt, die linke Seite hingegen maximal Rang k erreicht.[4]

$$k \; \geq \; r \tag{6.36}$$

4) Ronning (2004b) stellt darüber hinaus fest, dass ein unangenehmes Nebenprodukt darin besteht, dass alle Erwartungswertvektoren $\boldsymbol{\mu}_i$ unterschiedlich sein müssen.

Höhne (2004a) schlägt ein alternatives Vorgehen vor, um diese Bedingung zu umgehen und lediglich zwei Mischungskomponenten verwenden zu können. Hintergrund dieser Überlegungen ist die Tatsache, dass eine über zwei hinausgehende Anzahl von Mischungskomponenten keinen entscheidenden Einfluss auf die Erhöhung der Sicherheit hat. Eine Erhöhung der Sicherheit wird in erster Linie durch die Mittelwerte und Standardabweichungen der Mischungskomponenten bestimmt.

Höhne (2004a) ersetzt deshalb die Vereinfachungsannahme von Roque (2000) in Gleichung (6.33), dass die Kovarianzmatrizen aller Mischungskomponenten proportional zur Kovarianzmatrix der Originaldaten sein sollen, durch die Annahme, dass alle Mischungskomponenten die gleiche Kovarianzmatrix $\mathbf{\Sigma}_{mm}$ aufweisen sollen.

$$\mathbf{\Sigma}_i = \mathbf{\Sigma}_{mm} \tag{6.37}$$

Damit kann man die Gleichung (6.32) wie folgt umschreiben:

$$d\,\mathbf{\Sigma}_{xx} - \sum_{i=1}^{k} \alpha_i \left(\mu_i \mu_i'\right) = \sum_{i=1}^{k} \alpha_i \mathbf{\Sigma}_{mm}. \tag{6.38}$$

Dann kann die Anzahl der Mischungskomponenten k auf zwei reduziert werden, und es folgt:

$$\mathbf{\Sigma}_{mm} = d\,\mathbf{\Sigma}_{xx} - \sum_{i=1}^{2} \alpha_i \left(\boldsymbol{\mu}_i \boldsymbol{\mu}_i'\right). \tag{6.39}$$

Für eine symmetrische Mischung ($\mu_1 = -\mu_2$ und $\alpha_1 = \alpha_2 = 0,5$) kann diese Gleichung wie folgt weiter vereinfacht werden (Höhne 2004a):

$$\mathbf{\Sigma}_{mm} = d\,\mathbf{\Sigma}_{xx} - \boldsymbol{\mu}_i \boldsymbol{\mu}_i'. \tag{6.40}$$

Dabei ergibt sich die Restriktion, dass die Verschiebung der beiden Mischungsverteilungen nur so groß gewählt werden darf, dass die rechte Seite der Gleichung (6.40) positiv definit bleibt. Will man stärkere Verschiebungen realisieren, so ist auch eine größere Varianz der Fehlerüberlagerungen (größeres d) erforderlich (Höhne 2004a). Deshalb wird der Parameter f eingeführt, der den Grad der Abweichung der Mittelwerte der beiden Mischungskomponenten als Anteil an der Standardabweichung der Originalvariablen festlegt.

$$\boldsymbol{\mu}_1 = f\ \boldsymbol{s}(\boldsymbol{X}) \tag{6.41}$$

und

$$\boldsymbol{\mu}_2 = -\boldsymbol{\mu}_1 = -f\ \boldsymbol{s}(\boldsymbol{X}). \tag{6.42}$$

Dabei ist $\boldsymbol{s}(\boldsymbol{X})$ der Vektor der Standardabweichungen der Originalvariablen.

Außerdem gilt für die Parameter f und d:

$$0 \ < \ f \ < \ d. \tag{6.43}$$

Eine dem Vorschlag von Höhne (2004a) ähnliche Modifikation des Verfahrens von Roque (2000) wird von Yancey et al. (2002) vorgeschlagen.[5] Auch diese Modifikation geht von einer Mischung von Normalverteilungen aus: Der Erwartungswert der Mischungsverteilung ist wiederum Null, so dass die Gleichungen (6.27) bis (6.31) weiterhin gelten. Allerdings sollen die einzelnen Komponenten der Mischungsverteilung $\boldsymbol{V}_i$ im Unterschied zu den Vorschlägen von Roque (2000) und Höhne (2004a) bereits bei der Generierung der Varianz-Kovarianzmatrix der Originalvariablen $\boldsymbol{\Sigma}_{xx}$ genügen (Ronning 2004b).

$\boldsymbol{\epsilon}$ sei ein Zufallsvektor, für den gilt:

$$E\left[\boldsymbol{\epsilon}\right] = \boldsymbol{0} \tag{6.44}$$

und

$$cov\left[\boldsymbol{\epsilon}\right] = \boldsymbol{I}. \tag{6.45}$$

Allerdings wird angenommen, dass der Vektor $\boldsymbol{\epsilon}$ aus einer Mischungsverteilung stammt, deren Komponenten $\boldsymbol{\eta}_i$ normalverteilt sind mit:

$$\boldsymbol{\eta}_i \sim N\left(\boldsymbol{\mu}_i\, \tau^2 \boldsymbol{I}\right). \tag{6.46}$$

Die Mischung besitzt dann die folgende Verteilung:

$$\boldsymbol{\epsilon} \sim \sum_{i=1}^{k} \boldsymbol{\alpha}_i N\left(\boldsymbol{\mu}_i\, \tau^2 \boldsymbol{I}\right). \tag{6.47}$$

Der Vorteil der Methode von Yancey et al. (2002) liegt darin, dass die Anzahl der Mischungskomponenten k beliebig ist. Allerdings ergeben sich durch die unterschiedlichen

5) Diese Methode wurde von Kim und Winkler (1995) verwendet.

Restriktionen auf die Mischungsverteilung Widersprüche, die nur gelöst werden können, indem nicht eine mehrdimensionale Mischungsverteilung erzeugt wird, sondern k eindimensionale identische Mischungsverteilungen (Ronning 2004b).

Mit Hilfe der Choleski-Zerlegung der Varianz-Kovarianzmatrix der Originalvariablen wird dann für jeden Beobachtungspunkt der r-dimensionale Vektor

$$\boldsymbol{w}^* = \boldsymbol{\Sigma}_{xx}^{1/2}\boldsymbol{\epsilon} \tag{6.48}$$

generiert.

Die stochastische Überlagerung erfolgt dann analog zu Gleichung (6.19) bei der Überlagerung mit einer einfachen Normalverteilung durch:

$$\boldsymbol{X}^a = \boldsymbol{X} + \sqrt{d}\boldsymbol{w}^*. \tag{6.49}$$

Für die Verfahren von Höhne (2004a) und Yancey et al. (2002) kann ebenfalls wieder die Transformation von Kim (1986) aus Gleichung (6.20) durchgeführt werden, damit die Varianz-Kovarianzmatrix der anonymisierten Daten der Varianz-Kovarianzmatrix der Originaldaten entspricht.

a3) Heteroskedastische additive Überlagerung

Ein Grundproblem der additiven stochastischen Überlagerung besteht darin, dass bei konstanter Varianz der Zufallsfehler die kleinen Werte sehr stark, die großen Werte hingegen kaum verfremdet werden, und dies, obwohl gerade Großunternehmen und Großbetriebe, die bei den meisten quantitativen Merkmalen auch große Merkmalswerte aufweisen, einer besonders hohen Reidentifikationsgefahr ausgesetzt sind (Brand 2000; Vorgrimler 2003c; Statistische Ämter des Bundes und der Länder und IAW 2003; Rosemann et al. 2004). Eine Möglichkeit, diesem Problem zu begegnen, besteht darin, mit einer größenabhängigen Varianz zu arbeiten und somit einen heteroskedastischen Fehler zu verwenden. Die Zufallsfehler werden damit in funktionaler Abhängigkeit von den Originalvariablen erzeugt (Höhne 2004a).

Für die Störvariablen gilt damit:

$$W = f(X). \tag{6.50}$$

Am einfachsten erreicht man das Ziel einer größenabhängigen Streuung der Fehler durch die folgende funktionale Verknüpfung:

$$W = VX \tag{6.51}$$

mit

$$E[V] = 0. \tag{6.52}$$

Daraus ergibt sich aber

$$X^a = X + W \;=\; X + VX \;=\; (1+V)X \;=\; W^*X \tag{6.53}$$

mit

$$E[W^*] = 1. \tag{6.54}$$

Damit ist die heteroskedastische additive Überlagerung in dieser Form identisch mit der multiplikativen Überlagerung (Höhne 2004a). Allerdings gilt dies nur für den Fall eines linearen Zusammenhangs zwischen der Varianz der Störgröße und den Originalwerten.

Alternativ zu dieser Form der Heteroskedastie kommt noch eine additive stochastische Überlagerung mit verschiedenen Stufen fester Varianz – abhängig von der Größe der Merkmalswerte beziehungsweise des Merkmalsträgers – in Betracht (Höhne 2004a). Dabei könnte bei den kleinsten Unternehmen möglicherweise ganz auf eine Überlagerung verzichtet werden, wenn ein ausreichender Schutz bereits besteht. Ein solches Vorgehen würde es den Datennutzern erlauben, durch gezieltes Entfernen der großen Merkmalsträger einen fehlerfreien Datensatz zu verwenden. Soll jedoch bei ökonometrischen Analysen der gesamte Datenbestand verwendet werden, so müssten heteroskedastie-konsistente Verfahren verwendet werden.

a4) Weitere Varianten der additiven stochastischen Überlagerung

Das oben beschriebene Problem, dass bei einer additiven Überlagerung kleinere Unternehmen zu stark und größere Unternehmen zu schwach anonymisiert werden, könnte auch gelöst werden, indem die erzeugten Zufallsfehler vor der Überlagerung nach ihrem Betrag (Absolutwert) sortiert werden. Anschließend wird der größte Originalwert mit dem größten absoluten Fehler, der zweitgrößte mit dem zweitgrößten absoluten Fehler usw. überlagert.

Allerdings kann bei additiven stochastischen Überlagerungen das zusätzliche Problem auftreten, dass die Zufallsfehler zu unerwünschten Vorzeichenwechseln führen. Vorzeichenwechsel sind insbesondere dann unerwünscht, wenn die betroffene Variable per Definition

nur nichtnegative Werte aufweist, wie die „Beschäftigung“ oder der „Umsatz“. Dieses Problem kann insbesondere durch den Einsatz multiplikativer anstelle additiver Überlagerungen begegnet werden, sofern die Überlagerungsfaktoren ausschließlich im positiven Bereich liegen. Allerdings gibt es auch Möglichkeiten, das Problem beim Einsatz additiver Überlagerungen zu vermeiden: Beispielsweise kann eine additive Überlagerung verwendet werden, die statt eines Mittelwerts von Null jeweils den Mittelwert der zu anonymisierenden Originalvariablen aufweist und deren Realisationen ausschließlich im positiven Bereich liegen. Dann bleiben die originalen Mittelwerte zwar nicht erhalten, können jedoch einfach berechnet werden, indem der durch die Anonymisierung entstandene Mittelwert durch den Faktor 2 dividiert wird.

Anstelle einer additiven Überlagerung wird von einer Reihe von Autoren über den Vorschlag von Kim (1986) hinaus eine Kombination von Überlagerungen mit Transformationen vorgeschlagen. Kim und Winkler (2001) schlagen die Kombination nichtlinearer Transformationen mit dem Schema von Kim (1986) vor. Sullivan (1989) kombiniert die Überlagerung mit Zufallsfehlern mit nichtlinearen Transformationen und stellt eine Erweiterung vor, die die Anwendung auf diskrete Variablen erlaubt (Brand 2002; Ronning et al. 2002).

b) Multiplikative stochastische Überlagerung

Wie bereits in den vorangegangenen Ausführungen erwähnt, weist die multiplikative gegenüber der additiven stochastischen Überlagerung (ohne heteroskedastische Varianz) den Vorteil auf, dass der stärkeren Reidentifikationsgefahr größerer Unternehmen Rechnung getragen wird. Zudem erhält sie die Nullen und, sofern ausschließlich positive Überlagerungsfaktoren verwendet werden, auch die Vorzeichen.

Allgemein gilt für die anonymisierten Daten im Fall der multiplikativen stochastischen Überlagerung (Ronning 2004c):

$$X^a = WX. \tag{6.55}$$

Dabei ist W eine stetige Zufallsvariable mit Erwartungswert 1 und Varianz $\sigma_w^2 > 0$, die unabhängig von X ist.

Im multivariaten Fall gilt:

$$\mathbf{X}^a = \mathbf{W} \odot \mathbf{X}. \tag{6.56}$$

Dabei sind $\mathbf{X}^a$, $\mathbf{W}$ und $\mathbf{X}$ Zufallsmatrizen. $\odot$ ist das so genannte Hadamard-Produkt (vgl. z.B. Lütkepohl (1997)), bei dem eine elementweise Multiplikation durchgeführt wird.

Bei der multiplikativen Überlagerung besteht die Möglichkeit, die Daten eines Merkmalsträgers entweder mit einem konstanten Zufallsfaktor zu überlagern oder für jedes Merkmal einen neuen Zufallsfaktor zu erzeugen (Höhne 2004a). Die erste Vorgehensweise hat aus Nutzersicht den Vorteil, dass die relativen Beziehungen zwischen den Variablen eines Merkmalsträgers erhalten bleiben. Dies kann allerdings auch zu einem höheren Reidentifikationsrisiko führen.

Wie bereits erwähnt, sollte es sich bei der Zufallsvariable W in Gleichung (6.53) um eine positive Variable handeln, damit die Vorzeichen der Merkmalsträger nicht systematisch verändert werden. Im multivariaten Fall bedeutet dies, dass es sich bei der Matrix $\boldsymbol{W}$ um eine positive Matrix handeln muss. Aus diesem Grund werden bei multiplikativen Überlagerungen für die Störgrößen in der Regel Verteilungen verwendet, die lediglich positive Ausprägungen der Merkmalswerte zulassen. Gottschalk (2005) verwendet daher eine Gleichverteilung im Intervall von 0,5 bis 1,5. Das größte denkbare Intervall reicht von 0 bis 2, weil nur so alle Elemente positiv sind und der Erwartungswert 1 beträgt.

Kim und Winkler (2001) schlagen eine alternative Methode unter Verwendung einer Lognormalverteilung vor. Dabei werden die Originalvariablen zunächst logarithmiert (Gottschalk 2005, S.58).

$$V = ln(X). \tag{6.57}$$

Allerdings kann dieses Verfahren grundsätzlich nur für positive Werte der Originalvariablen durchgeführt werden. Bei Originalwerten von Null kann eine kleine Zahl hinzuaddiert werden, damit der Logarithmus berechnet werden kann (Gottschalk 2005).

Die Matrix der Zufallsfehler entstammt einer Normalverteilung $N(0, \boldsymbol{\Sigma}_{ww})$ mit

$$\boldsymbol{\Sigma}_{ww} = d\boldsymbol{\Sigma}_{vv} \tag{6.58}$$

mit $0 < d < 1$.

Für die anonymisierten Variablen gilt dann:

$$X^a = exp(V + W), \tag{6.59}$$

$$X^a = exp(ln(X) + W) \tag{6.60}$$

und damit

$$X^a = X exp(W). \tag{6.61}$$

In linearen Regressionsmodellen wirkt die Anwendung des Verfahrens wie ein additiver Zufallsfehler, wenn X^a logarithmiert wird (Gottschalk 2005, S.58). Allerdings werden die Mittelwerte der Variablen systematisch verzerrt. Es gilt nämlich (Gottschalk 2005, S.58):

$$E[X^a] = E[X exp(W)] \tag{6.62}$$

und aufgrund der Unabhängigkeit von X und W:

$$\begin{aligned} E[X^a] &= E[X]\, E[exp(W)] \\ E[X^a] &= E[X]\, exp(\sigma_w^2/2). \end{aligned} \tag{6.63}$$

Um einen erwartungstreuen Schätzer für den Mittelwert zu erhalten, muss daher die Stichprobenvarianz des Zufallsfehlers σ_w^2 aus den anonymisierten Daten berechnet werden. Dies wird von Gottschalk (2005) wie folgt hergeleitet:

$$\begin{aligned} var[ln(X^a)] &= var[V + W] \\ var[ln(X^a)] &= (1 + d)\sigma_v^2, \end{aligned} \tag{6.64}$$

$$\sigma_v^2 = \frac{var[ln(X^a)]}{1 + d} \tag{6.65}$$

und damit

$$\sigma_w^2 = var[ln(X^a)] \frac{d}{1 + d}. \tag{6.66}$$

Außerdem können die Varianzen der Originalvariablen und die Kovarianzen zwischen zwei Originalvariablen geschätzt werden (Gottschalk 2005):

$$var[X] = \frac{var[X^a]}{exp(2\sigma_w^2)} + \frac{E[X]^2}{exp(\sigma_w^2)} - E[X]^2 \tag{6.67}$$

und

$$cov\left[X_k X_s\right] = \left[\frac{\sum_i X_{ik}^a X_{is}^a}{exp((\sigma_{w_k}^2 + 2\rho_{ks}\sigma_{w_k}\sigma_{w_s} + \sigma_{w_s}^2)/2)} - \frac{n\bar{X}_k^a \bar{X}_s^a}{exp(\sigma_{w_k}^2 + \sigma_{w_s}^2)}\right]/(n-1) \quad (6.68)$$

mit $\bar{X}_k^a = \sum_i X_{ik}^a/n$ und ρ_{ks} Korrelationskoeffizient zwischen X_k^a und X_s^a.

Um lediglich positive Werte für die Zufallsfehler zu erhalten, kann auch, wie von Kim und Winkler (2001) ebenfalls vorgeschlagen, bei der direkten multiplikativen Überlagerung der Originalwerte eine gestutzte Normalverteilung verwendet werden. Außerdem können die Originalvariablen auch multiplikativ mit lognormalverteilten Störgrößen oder Störgrößen aus anderen nichtnegativen Verteilungen überlagert werden. Höhne (2004a) schlägt hingegen vor, die Varianzen der Zufallsfehler so gering zu wählen, dass Werte ≤ 0 bei einem Erwartungswert von Eins auch bei einer einfachen Normalverteilung nur mit sehr kleiner Wahrscheinlichkeit auftreten. Treten sie in der Praxis dennoch auf, werden die Zufallsfehler erneut erzeugt. Um dennoch einen ausreichenden Schutz der Daten zu gewährleisten, soll wiederum eine zweigipflige Mischungsverteilung für den Zufallsfehler verwendet werden. Die Verteilung der Zufallszahlen ist in diesem Fall durch

$$W \sim 0,5N\left(1-\mu,\sigma\right) + 0,5N\left(1+\mu,\sigma\right) \quad (6.69)$$

gegeben.

Eine spezielle Form einer unechten Mischungsverteilung wird von (Höhne 2004a) zur Anonymisierung vorgeschlagen: Bei diesem im Folgenden auch als Höhne-Verfahren bezeichneten Vorgehen wird zunächst mit Wahrscheinlichkeit von 0,5 festgelegt, ob die Merkmalswerte eines Merkmalsträgers verkleinert oder vergrößert werden. Hierzu werden die Grundüberlagerungsfaktoren $1-f$ und $1+f$ festgelegt. Jeder Merkmalsträger erhält einen dieser Grundüberlagerungsfaktoren zugewiesen. Die Grundüberlagerungsfaktoren werden anschließend für jeden Merkmalswert unabhängig additiv mit einer Normalverteilung mit Erwartungswert Null und Standardabweichung s ($s < f/2$) überlagert. Somit wird jeder Merkmalsträger in die gleiche Richtung verzerrt, dennoch erhält jeder Merkmalswert einen unterschiedlichen Überlagerungsfaktor.

Auch nach einer multiplikativen Überlagerung kann eine Korrektur analog zur Transformation von Kim (1986) vorgenommen werden, um die ersten und zweiten Momente wieder herzustellen. Diese lautet wie folgt (Höhne 2004a):

$$x_i^{a^R} = \frac{\sigma_x}{\sigma_{x^a}}\left(x_i^a - \mu_{x^a}\right) + \mu_x \quad (6.70)$$

mit μ_x dem Durchschnitt der Originalvariablen, μ_{x^a} dem Durchschnitt der multiplikativ überlagerten Variablen, σ_x der Standardabweichung der Originalvariablen und σ_{x^a} der Standardabweichung der multiplikativ überlagerten Variablen.

Insbesondere bei sehr schief verteilten Originalvariablen hängt die Stärke der Abweichungen durch Überlagerung von der Konstellation der Zufallszahlen bei wenigen großen Merkmalsträgern ab. Dadurch kann es passieren, dass trotz der asymptotischen Erwartungstreue und qualitativ hochwertig generierten Zufallszahlen die Mittelwerte und Summen nur sehr schlecht reproduziert werden (Höhne 2004a). Höhne (2004a) entwickelte deshalb einen Algorithmus für eine „kontrollierte" multiplikative Überlagerung. Dabei werden zunächst normalverteilte Zufallszahlen W_i mit Erwartungswert größer Null erzeugt. Diese sollen lediglich positive Werte annehmen. Anschließend wird der Datenbestand für das zu bearbeitende Merkmal nach den Merkmalswerten der Größe absteigend sortiert. Der hinsichtlich des zu bearbeitenden Merkmals größte Merkmalsträger wird durch

$$X_1^a = (1 - W_1)X_1 \tag{6.71}$$

in jedem Fall verkleinert.

Die folgenden Merkmalsträger $i = 2, ..., m-1$ werden wie folgt bearbeitet:

$$X_i^a = (1 - W_i)X_i \text{ , wenn} \sum_{k=1}^{i-1} X_k^a > \sum_{k=1}^{i-1} X_k \tag{6.72}$$

$$X_i^a = (1 + W_i)X_i \text{ , wenn} \sum_{k=1}^{i-1} X_k^a \leq \sum_{k=1}^{i-1} X_k \tag{6.73}$$

Damit wird sichergestellt, dass sich die Verkleinerungen und Vergrößerungen der einzelnen Merkmalswerte durch die stochastische Überlagerung auch innerhalb der einzelnen Größenbereiche gegenseitig aufheben und Summen und Mittelwerte erhalten bleiben.

Für die Anonymisierung des letzten Merkmalsträgers gilt dann:

$$X_m^a = X_m - \left(\sum_{k=1}^{m-1} X_k^a - \sum_{k=1}^{m-1} X_k \right), \tag{6.74}$$

womit sichergestellt wird, dass auch die Gesamtsumme erhalten bleibt.

Dieses Vorgehen kann auch mit der oben beschriebenen Überlagerung mit einer unechten „Mischungsverteilung" nach dem Verfahren von Höhne kombiniert werden. Dann erhält der größte Merkmalsträger den Grundfaktor $1 - f$, der zweitgrößte den Wert $1 + f$ zugewiesen.

6.2.4 Mikroaggregationsverfahren

Die Grundidee der Mikroaggregationsverfahren besteht darin, ähnliche Objekte zu Gruppen zusammenzufassen und die Ursprungswerte durch die arithmetischen Mittel der Merkmalswerte aller Merkmalsträger innerhalb der Gruppen zu ersetzen (Mateo-Sanz und Domingo-Ferrer 1998; Ronning et al. 2002; Höhne 2003a).

Alle Gruppierungsverfahren gehen von Gruppengrößen von mindestens drei Werten aus, denn bei nur zwei Merkmalsträgern können die Werte des einen Merkmalsträgers bei Kenntnis der Werte des anderen Merkmalsträgers in jedem Fall enthüllt werden.

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 (Höhne 2003a).

Grundsätzlich kann zwischen zwei Arten der Mikroaggregation unterschieden werden (Rosemann 2004):

- Die deterministische bzw. abstandsorientierte Mikroaggregation, bei der möglichst ähnliche Einheiten zusammengefasst werden.
- Die stochastische Mikroaggregation, bei der die Gruppenbildung zufällig erfolgt.

Zudem erfolgt eine Unterscheidung danach, ob die Mikroaggregation für alle Variablen gemeinsam erfolgt – für die Durchschnittsbildung bei den verschiedenen Variablen folglich die gleichen Gruppen gebildet werden – oder die Gruppenbildung für jede Variable getrennt erfolgt (Rosemann 2004).

a) Deterministische beziehungsweise abstandsorientierte Mikroaggregation

Die Idee der deterministischen beziehungsweise abstandsorientierten Mikroaggregation besteht darin, möglichst ähnliche Merkmalsträger zu Gruppen zusammenzufassen und deren Originalwerte durch die arithmetischen Mittel innerhalb der Gruppen zu ersetzen. Die einzelnen Verfahrensvarianten unterscheiden sich zum einen danach, ob die Gruppenbildung für alle metrischen Variablen – oder auch Gruppen von Variablen – gemeinsam erfolgt (mehrdimensionale Mikroaggregation) oder die Variablen getrennt mikroaggregiert (eindimensionale Mikroaggregation) werden. Zum anderen unterscheiden sich die mehrdimensionalen Mikroaggregationsverfahren hinsichtlich der Bestimmung des Abstandes zwischen den einzelnen Objekten.

a1) Abstandsorientierte Mikroaggregation für alle Variablen gemeinsam (Gemeinsame Mikroaggregation)

- **Mikroaggregation nach einer Variablen**: Es wird eine dominierende Variable herausgesucht und der Datenbestand danach sortiert. Danach werden absteigend immer drei benachbarte Merkmalsträger in einer Gruppe zusammengefasst 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 einer Hilfsvariablen**: Die Sortierung erfolgt anhand 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).
- **Mikroaggregation nach allen metrischen Variablen:** Die Gruppenbildung erfolgt nach der euklidischen Distanz zwischen den Merkmalsträgern. Dabei werden die beiden Merkmalsträger herausgesucht, die den größten Abstand untereinander haben. Danach werden diesen beiden jeweils die zwei dichtesten Merkmalsträger hinzu gruppiert. Die verbleibenden, noch nicht gruppierten Merkmalsträger werden wieder analog behandelt (Mateo-Sanz und Domingo-Ferrer 1998).

a2) Abstandsorientierte Mikroaggregation für alle Variablen getrennt (Getrennte Mikroaggregation)

Der Datenbestand wird jeweils nach der zu anonymisierenden Variable sortiert. Danach werden absteigend immer drei bis fünf benachbarte Merkmalsträger in einer Gruppe zusammengefasst und alle ihre stetigen Merkmalswerte durch den Durchschnitt der Werte ersetzt. Anschließend wird der Vorgang für die anderen metrischen Variablen wiederholt.

Die unabhängige eindimensionale Mikroaggregation hat gegenüber den anderen Mikroaggregationsverfahren ein zusätzliches Sicherheitsrisiko, weil für jeden Merkmalswert eine Ober- und Untergrenze des originalen Wertes durch die Durchschnitte der benachbarten Gruppen verfügbar ist. Diese Werte können den Bereich, in dem ein potenzieller Datenangreifer den Originalwert vermuten kann, stark einschränken. Aus diesem Grund schlägt Höhne (2004b) vor, eine Mindestabweichung zwischen den Durchschnittswerten der einzelnen Gruppen einzuführen und die Gruppengröße entsprechend dieser Mindestbreite variabel zu gestalten. Allerdings stellt sich die Frage, ob dieses Problem in der Praxis tatsächlich eine große Relevanz besitzt. Ein vergleichsweise geringer Abstand zwischen den Durchschnitten benachbarter Gruppen tritt in der Regel in Bereichen mit einer hohen Dichte an Datenpunkten auf. Die Dichte der Datenpunkte ist aber im Bereich der kleineren Unternehmen am größten, dort ist zugleich das Reidentifikationsrisiko am geringsten (Statistische Ämter des Bundes und der Länder und IAW 2003; Vorgrimler 2003c; Rosemann et al. 2004).

a3) Abstandsorientierte Mikroaggregation für Gruppen von Variablen (Gruppierte Mikroaggregation / Teilweise gemeinsame Mikroaggregation)

Bei dieser Verfahrensvariante werden die Variablen zunächst gruppiert und anschließend innerhalb der gebildeten Gruppen gemeinsam mikroaggregiert. Diese Variante der Mikroaggregation wurde von Domingo-Ferrer und Mateo-Sanz (2001) entwickelt. Die Gruppenbildung der Variablen erfolgt nach den Korrelationen zwischen den Variablen (Statistische Ämter des Bundes und der Länder und IAW 2003; Rosemann 2004). Für die einzelnen Variablengruppen kann die Mikroaggregation analog der in a1) beschriebenen Varianten für die gemeinsame Mikroaggregation vorgenommen werden. Die abstandsorientierte Gruppenbildung der Objekte kann somit nach einer Variablen (aus der Gruppe), aus diesen gebildeten Hilfsvariablen oder nach allen metrischen Variablen innerhalb der Gruppe erfolgen. Mit dem Verfahren sollen die Vorteile der gemeinsamen Mikroaggregation mit den Vorteilen der getrennten Mikroaggregation verbunden werden.

a4) Modifikation von Mikroaggregationsverfahren

Eine Modifikation der abstandsorientierten Mikroaggregationsverfahren besteht in der Einführung einer variablen Gruppengröße, um eine noch stärker datenorientierte Gruppenbildung zu ermöglichen. Dabei werden Gruppen bis zu einer Größe von fünf angestrebt. Gruppengrößen von mehr als fünf sind nicht sinnvoll, da eine Teilung in diesem Fall zu besseren Ergebnissen führen würde. Bei diesen Verfahren werden die einzelnen Objekte nach dem Kriterium der größten Ähnlichkeit gegebenenfalls auch Gruppen zugeordnet, die bereits drei oder mehr Elemente enthalten. Erreichen Gruppen eine Größe von mehr als fünf Elementen, werden sie durch hierarchische Anwendung des Verfahrens wieder geteilt (Mateo-Sanz und Domingo-Ferrer 1998).

b) Stochastische Mikroaggregation

Stochastische Mikroaggregationsverfahren wurden erstmals von Lechner und Pohlmeier (2003) vorgeschlagen. Lechner und Pohlmeier (2003) beschreiben zwei Möglichkeiten der stochastischen Mikroaggregation, die im Folgenden als zufällige Mikroaggregation und Bootstrap-Mikroaggregation bezeichnet werden.

b1) Zufällige Mikroaggregation

Das Vorgehen bei der zufälligen Mikroaggregation entspricht grundsätzlich dem Vorgehen bei der deterministischen Mikroaggregation, allerdings erfolgt die Gruppenbildung der Objekte nicht abstandsorientiert, sondern zufällig. Damit spielt die Ähnlichkeit der Objekte bei der Gruppenbildung keine Rolle. Die zufällige Gruppenbildung kann analog zur deterministischen Mikroaggregation für alle Variablen – beziehungsweise für Gruppen von

Variablen – gemeinsam oder für alle Variablen getrennt erfolgen.

b2) Bootstrap-Mikroaggregation

Für jedes Objekt werden zufällig zwei weitere gezogen. Die Ziehung erfolgt mit Zurücklegen – auch das erste Unternehmen selbst kann nochmals gezogen werden –, so dass es sich um eine Art Bootstrap-Verfahren handelt. Diese drei Objekte bilden eine Gruppe, deren durchschnittliche Merkmalswerte an die Stelle der Werte für das erste Objekt treten. Die Bootstrap-Mikroaggregation hat den Vorteil, dass im Vergleich mit den anderen Mikroaggregationsvarianten die Varianz weniger stark reduziert wird.

c) Formale Darstellung der Mikroaggregation

Wird ein Vektor aus n Variablenwerten mikroaggregiert, so gilt für den anonymisierten Variablenvektor

$$\boldsymbol{x}^a = \boldsymbol{D}\boldsymbol{x}. \tag{6.75}$$

Wird eine Matrix aus K Variablen (Spalten) und n Beobachtungen (Zeilen) gemeinsam mikroaggregiert, so gilt für die anonymisierte Datenmatrix:

$$\boldsymbol{X}^a = \boldsymbol{D}\boldsymbol{X}. \tag{6.76}$$

Dabei ist $\boldsymbol{D}$ die $n \times n$-blockdiagonale Matrix

$$\boldsymbol{D} = \boldsymbol{I}_M \otimes \frac{1}{A}\boldsymbol{\iota}\boldsymbol{\iota}'. \tag{6.77}$$

Bei der Mikroaggregation werden A Beobachtungen zu einer Gruppe zusammengefasst. Die Anzahl der Gruppen beziehungsweise die Anzahl der unterschiedlichen Werte einer Variablen beträgt somit $M = n/A$.

Die Mikroaggregationsmatrix $\boldsymbol{D}$ ist symmetrisch idempotent, d.h. es gilt:

$$\mathbf{DD} = \mathbf{D} \tag{6.78}$$

und

$$\mathbf{D}' = \mathbf{D}. \tag{6.79}$$

Dieses Schema gilt sowohl für die abstandsorientierte Mikroaggregation als auch für die zufällige Mikroaggregation. Allerdings ist im Fall einer deterministischen Mikroaggregation aufgrund der abstandsorientierten Gruppenbildung die Aggregationsmatrix von der beziehungsweise den zu anonymisierenden Variable/n abhängig. Es gilt dann: $\boldsymbol{D} = \boldsymbol{D}(\boldsymbol{X})$ (Lechner und Pohlmeier 2003).

Wird hingegen eine Bootstrap-Mikroaggregation vorgenommen, so ergibt sich die $n \times n$-Aggregationsmatrix durch (Lechner und Pohlmeier 2003):

$$\boldsymbol{D}_{BS} = \frac{1}{B}\left(\boldsymbol{I}_n + \boldsymbol{S}_1 + \boldsymbol{S}_2 + \cdots + \boldsymbol{S}_{B-1}\right). \tag{6.80}$$

Dabei ist $\boldsymbol{S}_b$ eine $n \times n$-Selektionsmatrix, die jeweils in einer Zeile an einer zufällig ausgewählten Position eine Eins und sonst Nullen enthält. Somit ist die Aggregationsmatrix $\boldsymbol{D}_{BS}$ eine Zufallsmatrix. Diese ist jedoch nicht symmetrisch idempotent.

d) Bewertung der Mikroaggregation

Der Vorteil der Mikroaggregationsverfahren besteht darin, dass die arithmetischen Mittel erhalten bleiben, Standardabweichungen und Varianzen werden systematisch reduziert. Auch die Korrelationsmatrix und die Kovarianzen werden je nach konkretem Verfahren und Beschaffenheit der Daten verzerrt. Die Gesamtvarianz im Datenbestand lässt sich bei deterministischer und zufälliger Mikroaggregation additiv in die Varianz zwischen den Gruppen und die Varianz innerhalb der Gruppen zerlegen. Nach der Mikroaggregation wird durch die Durchschnittsbildung die Varianz innerhalb der Gruppen entfernt. Die verbleibende Varianz liegt somit immer unterhalb der Originalvarianz. Die methodischen Untersuchungen von Baeyens und Defays (1999) zeigen jedoch, dass mit zunehmender Größe des Datenbestandes der Varianzverlust asymptotisch gegen Null konvergiert. Die Eigenschaft von systematisch unterschätzten Varianzen führt auch zu verzerrten Ergebnissen von Parametertests in ökonometrischen Modellen (Brand 2000; Lechner und Pohlmeier 2003).

Höhne (2004b) schlägt deshalb eine Modifikation der Mikroaggregation vor, mit dem Ziel, die Varianz in den Daten zu erhalten. Konkret sollen Gruppen mit jeweils vier Merkmalsträgern gebildet werden, von denen zwei den um die Standardabweichung innerhalb der Gruppe verminderten Gruppendurchschnitt als neuen Merkmalswert zugewiesen bekommen, die beiden anderen den um die Standardabweichung erhöhten Gruppendurchschnitt. Innerhalb der Gruppe aus vier Merkmalsträgern bleiben so Mittelwert und Varianz erhalten. Für größere Gruppen mit einer geraden Anzahl an Merkmalsträgern kann die Bearbeitung analog erfolgen. Bei ungerader Anzahl muss die Transformationsregel angepasst werden.

Grundsätzlich stellt eine Gruppe mit lediglich zwei identischen Werten ein besonderes Risiko dar. Der für die beiden Merkmalsträger ausgewiesene Wert entspricht jedoch nun nicht mehr dem Durchschnitt aus den beiden Originalwerten. Für die gesamte Aggregationsgruppe bleiben die Durchschnitte zwar erhalten, diese besteht jedoch zum einen aus mindestens vier Merkmalsträgern. Zum anderen sind die beiden Teilgruppen nicht mehr als zum gleichen Mikroaggregat gehörig erkennbar, was selbst Distanzprobleme innerhalb von Gruppen nicht mehr erkennen lässt. Die Entscheidung darüber, welche Merkmalsträger den um die Standardabweichung reduzierten Wert zugewiesen bekommen und welche den um die Standardabweichung erhöhten, sollte so gewählt werden, dass die Korrelationsmatrix möglichst gut erhalten bleibt (Höhne 2004b).

6.2.5 Simulationsverfahren

Das Grundprinzip von Simulationsverfahren ist die Erzeugung von synthetischen Merkmalsträgern, die die empirische Verteilung approximieren. 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 weitaus kleinere oder größere Testdatensätze erzeugen.

Im Idealfall genügen synthetische Datensätze sowohl den Vorgaben des Datenschutzes als auch den Anforderungen der Datennutzer (Gottschalk 2005, S.97f.). Die Testdaten lassen sich daher im Idealfall nicht mehr auf die Originaldaten zurückführen (Statistische Ämter des Bundes und der Länder und IAW 2003). Auf der anderen Seite werden den Datennutzern für die Untersuchung von kausalen Zusammenhängen und die Schätzung von Populationsmerkmalen optimale anonymisierte Daten zur Verfügung gestellt (Gottschalk 2005, S.97).

Die optimale Umsetzung dieser Idee bestünde darin, die Kerndichte des gesamten Datenbestandes zu schätzen (Fienberg 1997). Mit Hilfe dieser Dichte würde dann die gewünschte Anzahl der synthetischen Datensätze erzeugt. Allerdings ist die Schätzung mehrdimensionaler empirischer Verteilungen bisher nur für niedrig-dimensionale Daten gelungen. Bereits zwei- bis dreidimensionale Kerndichteschätzungen scheitern daran, dass zu wenig Beobachtungen vorliegen. Folglich können nicht alle Informationen des Originaldatensatzes vollständig im simulierten Datensatz erhalten bleiben (Gottschalk 2005, S.97). Dennoch existieren eine Reihe von Simulationsansätzen mit dem Ziel, dem Optimum möglichst nahe zu kommen. Zwei davon, das Latin Hypercube Sampling und das Resampling, werden im Folgenden vorgestellt.

a) Latin Hypercube Sampling (LHS)

Latin Hypercube Sampling (LHS) wurde von Dandekar et al. (2001) zur Anonymisierung von Daten mit großem Stichprobenumfang vorgeschlagen. Das Verfahren basiert auf dem LHS-Verfahren von McKay et al. (1979), das synthetische Datensätze unkorrelierter Variablen so erzeugen kann, dass die univariaten Verteilungen annähernd abgebildet werden. Dieses wurde von Iman und Conover (1982) dahingehend weiterentwickelt, dass auch Rangkorrelationen erhalten bleiben (Gottschalk 2005).

Beim Latin Hypercube Sampling 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 Werten erzeugt. Diese synthetischen Merkmalswerte werden in einem zweiten Schritt durch ein Swapping-Verfahren so umgeordnet, dass die Rangkorrelationen optimiert werden (Höhne 2003a).

Der iterative Algorithmus zur Erzeugung des künstlichen Datensatzes lautet wie folgt (Dandekar et al. 2001):

1. Berechne die Matrix der Rangkorrelationen $\boldsymbol{R}$ der Merkmale $d = 1, 2, ..., D$ der zu anonymisierenden Daten, die sowohl ordinale als auch kardinale Variablen enthalten können.

2. Ausgangspunkt der Anonymisierung ist ein Datensatz $\boldsymbol{S}$ mit n gewünschten Beobachtungen ($i = 1, 2, ..., n$). Dieser wird, für jede Variable getrennt, mit Hilfe der geglätteten empirischen Verteilungsfunktion $F(X)$ (ggf. auch eine theoretische Verteilungsfunktion) simuliert. Dazu wird für jedes Element der Menge $\left\{\frac{i}{n+1}\right\}$ $x_i = F^{-1}\left(\frac{i}{n+1}\right)$ berechnet.

3. Die Merkmalswerte der Datenmatrix $\boldsymbol{S}$ werden durch zufälliges Vertauschen so umsortiert, dass die Spalten nicht mehr perfekt korreliert sind (Rangkorrelationen). Es entsteht die transformierte Matrix $\boldsymbol{S}^*$.

4. Man berechnet die Rangkorrelationsmatrix $\boldsymbol{R}^*$ der Matrix $\boldsymbol{S}^*$, d.h. $\boldsymbol{R}^*$ hat die Dimension $D \times D$.

5. Man berechnet $\mathbf{A}$, so dass $\mathbf{AA}' = \mathbf{R}^*$.

6. Man berechnet $\mathbf{B}$, so dass $\mathbf{BB}' = \mathbf{R}$.

7. Man berechnet die Korrekturmatrix $\mathbf{R}^{**} = \mathbf{BA}^{-1}$.

8. Man berechnet die $(n \times D)$-Matrix $\mathbf{H} = \mathbf{S}^*\mathbf{R}^{**'}$, so dass die Rangkorrelationen von $\mathbf{H}$ die Rangkorrelationsmatrix $\boldsymbol{R}$ der Ausgangsdaten approximiert.

9. Man ersetzt die n Merkmalsausprägungen der D Spalten von $\mathbf{H}$ durch ihre Ränge 1 bis n.
10. Man ordnet jeweils die n Werte der D Spalten von $\mathbf{S}^*$ in derselben Reihenfolge wie die Ränge der Matrix $\mathbf{H}$. Die umgeordnete Matrix sei $\mathbf{S}^{**}$.
11. Man ersetzt $\mathbf{S}^* = \mathbf{S}^{**}$ und wiederholt die Schritte 4 bis 10 so lange, bis die mittlere absolute Abweichung der Nebendiagonalelemente von $\mathbf{R}$ und $\mathbf{R}^{**}$ nicht mehr signifikant verringert werden kann.
12. Man berechnet die univariaten Verteilungsfunktionen der Variablen $d = 1, 2, ..., D$ der Ausgangsdaten.
13. Man ersetzt jeden Rang in den Spalten $d = 1, 2, ..., D$ der Matrix $\mathbf{H}$ mit dem entsprechenden Wert der inversen empirischen Verteilungsfunktion an der Stelle $i/(n + 1)$ der Variablen $d = 1, 2, ..., D$ aus Schritt 12.

Das Verfahren kann für einzelne Teilgesamtheiten getrennt angewendet werden. Dies setzt allerdings jeweils eine ausreichende Anzahl an Beobachtungen voraus, um die Korrelationen zwischen den Variablen weiterhin zuverlässig schätzen zu können. Kategoriale Variablen können zur Abgrenzung dieser Teilgesamtheiten verwendet werden. Alternativ können sie auch im Rahmen des Anonymisierungsalgorithmus berücksichtigt werden. Dabei werden die entsprechenden empirischen univariaten diskreten Verteilungen abgebildet (Gottschalk 2005).

Eine Modifikation des Verfahrens – die so genannten Hybrid Masking Methode – besteht darin, die synthetischen Daten und die Originaldaten additiv oder multiplikativ miteinander zu verknüpfen (Dandekar et al. 2002), um zu verhindern, dass die anonymisierten Daten zu stark von den Originaldaten abweichen. Die Kombination erfolgt durch einen Match der synthetischen und der originalen Beobachtungen über ein Distanzmaß (Gottschalk 2005, S.100).

Echte Reidentifikationen sind nach Durchführung des LHS nicht mehr möglich, da die einzelnen Einheiten nicht mehr in ihrer ursprünglichen Form im Datenbestand auftauchen. Damit sind in der Regel auch keine Rückschlüsse auf die Originalangaben mehr möglich. Allerdings kann keine absolute Anonymität garantiert werden, insbesondere wenn weiterhin extreme Beobachtungen in den Daten enthalten sind. Dies ist insbesondere deshalb der Fall, weil LHS die Besonderheit aufweist, dass alle originalen Merkmalswerte auch im synthetischen Datensatz wieder auftauchen, allerdings in einer veränderten Zuordnung mit möglichst optimalem Erhalt der Rangkorrelationen.

b) Resampling

Beim Resampling wird zunächst die multivariate Dichte der Ausgangsdaten mittels semi- oder nichtparametrischer Verfahren geschätzt. Anschließend wird eine Stichprobe gezogen,

die der geschätzten Dichte folgt (Gottschalk 2005). Erfolgen die Ziehungen mit Zurücklegen, so können die künstlichen Mikrodatensätze als Bootstrap-Stichproben betrachtet werden.

Wenn $\hat{F}$, die geschätzte empirische Verteilungsfunktion, ein konsistenter Schätzer für die wahre Verteilungsfunktion F ist, und für wiederholte Ziehungen mit Zurücklegen $R_1, R_2, ..., R_B$, mit $B \rightharpoonup \infty$

$\frac{1}{B} \sum_{j=1}^{B} \hat{F}_{R_j} \rightharpoonup \hat{F}$ gilt,

können Ergebnisse von Datenanalysen mit den Originaldaten unter Verwendung mehrerer Resamples reproduziert werden. Es ist daher für die Nutzer besser, mehrere Resamples zu nutzen. Die Bootstrap-Schätzer berechnen sich dann als Mittelwerte der empirischen Verteilungsparameter der einzelnen Bootstrap-Stichproben (Gottschalk 2005, S.102).

Auch beim Resampling stellt sich das oben bereits beschriebene Problem, dass die Schätzung einer mehrdimensionalen Kerndichte bisher nicht in dem erforderlichen Umfang möglich ist. Gottschalk (2005) beschreibt drei mögliche Lösungswege. Zum Ersten kann eine Verteilungsannahme getroffen werden. Anschließend werden mit Hilfe der Daten die Parameter dieser theoretischen Verteilung geschätzt. Allerdings ist es sehr fraglich, ob eine empirische Verteilung tatsächlich einer theoretischen folgt. Zum Zweiten wird daher alternativ ein bayesianischer Ansatz diskutiert (Fienberg 1997; Fienberg et al. 1996). Grundlage des bayesianischen Ansatzes ist eine bestimmte Verteilungsannahme eines Modellparameters, die aufgrund theoretischer Überlegungen angenommen wird. Diese à priori Verteilung wird mit Hilfe der Datenbasis $\boldsymbol{X}$ evaluiert. Dabei werden die zu schätzenden Parameter so gewählt, dass die erwarteten Kosten einer falschen à priori Annahme über die Parameter minimiert werden, wenn die Kosten quadratisch mit dem Fehler wachsen. Zum Dritten wird ein nicht-parametrisches Resampling zur Anonymisierung von Einzeldaten vorgeschlagen. Dabei werden die simulierten Beobachtungen direkt aus den Originaldaten erzeugt. Die Dichtefunktion wird nicht explizit bestimmt.

Ausgangspunkt des Verfahrens sind nicht-parametrische uni- oder multivariate Kerndichteschätzungen (Devroye und Györfi 1985; Silverman 1986). Die Kernfunktion K schätzt die Dichte an der Stelle x („lokale" Dichte). Sie ist eine Gewichtungsfunktion, die nahe bei x liegende Beobachtungen hoch und weit von x entfernte Beobachtungen niedrig gewichtet (Gottschalk 2005). Eine Kernfunktion ist symmetrisch um Null verteilt, ihr Integral beträgt 1. Mögliche Kernfunktionen sind Epanechnikov-, Rechtecks- und Gausskerne. Die Summe der einzelnen Kerne ergibt die Dichteschätzung für die gesamte Verteilung. Mit der Wahl der Bandbreite h wird festgelegt, wie stark die benachbarten Beobachtungen bei der Schätzung der „lokalen" Dichte gewichtet werden. Bei kleinen Werten von h werden weiter entfernt liegende Beobachtungen weniger stark gewichtet. Je größer die Bandbreite h gewählt wird, desto stärker ist die Dichteschätzung geglättet. Eine zu geringe Glättung

führt zu einer stark schwankenden Funktion, womit die Verteilung von X nicht eindeutig zu erkennen ist. Eine zu starke Glättung führt dazu, dass vorhandene systematische Strukturen möglicherweise nicht mehr erkannt werden können. Die Wahl der optimalen Bandbreite ist daher für eine gute Dichteschätzung von besonderer Bedeutung (Gottschalk 2005, S.106f.). In Parzen (1962) und Silverman (1986) finden sich Vorschläge für die Wahl der optimalen Bandbreite.

Stichprobenziehungen aus der geglätteten Kerndichtefunktion können nach den folgenden Schritten erfolgen (Gottschalk 2005, S.107f.):

1. Ziehe eine Zufallsstichprobe X_Z mit Zurücklegen.
2. Berechne k aus der Kernfunktion K.
3. Bilde $X^a = X_Z + hk$.

Zur Erhaltung der ersten beiden Momente von X kann X^a nach dem Vorschlag von Silverman (1986) umskaliert werden:

$$X^a = \bar{X} + (X_Z - \bar{X} + hk)/(1 + h^2\sigma_k^2/\sigma_x^2). \tag{6.81}$$

Dabei ist $\bar{X}$der Mittelwert von X sowie σ_x^2 und σ_k^2 die Varianzen von X beziehungsweise des Kerns k.

Im mehrdimensionalen Fall ist für jede Variable eine Bandbreite h zu wählen. Außerdem müssen für alle Variablen Kernfunktionen gebildet werden. Werden die Kernfunktionen unabhängig voneinander gebildet, so sollten die Kerne unter Berücksichtigung der mehrdimensionalen Verteilung gewichtet werden, um auch die multivariaten Verteilungseigenschaften zu erhalten. Hierzu werden von Gottschalk (2005) verschiedene Varianten vorgeschlagen:

- Die Kerne werden mit der Varianz-Kovarianzmatrix der Originalvariablen gewichtet.
- Die mehrdimensionale Kernmatrix wird so umskaliert, dass sie dieselbe Kovarianzstruktur aufweist wie die originale Datenmatrix.
- Die mehrdimensionale Kernmatrix wird so umskaliert, dass sie dieselbe Korrelationsmatrix aufweist wie die originale Datenmatrix.

Eine wesentliche Stellschraube des Verfahrens besteht in der Wahl der Bandbreite der Kerne. Dabei gilt grundsätzlich die Devise: Maximiere die Bandbreiten und erhalte dabei die Analysefähigkeit der Daten. An diesem kritischen Punkt führen kleine Erweiterungen der Bandbreiten zu einer deutlichen Verschlechterung der Datenqualität. Das bedeutet, dass die Verzerrung der Ursprungsverteilungen bis zu dem kritischen Punkt zwar iterativ erhöht

wird, aber dennoch sichergestellt ist, dass valide wissenschaftliche Auswertungen mit den anonymisierten Daten möglich sind. Eine weitere Anonymisierung beziehungsweise Glättung der Daten würde die Datensätze unbrauchbar werden lassen. Die Restriktion dieses Optimierungsproblems – die Erhaltung der Analysefähigkeit – kann ebenfalls nicht eindeutig definiert werden. Erste Experimente haben gezeigt, dass eindimensionales Resampling univariate Verteilungsparameter gut reproduziert, aber mehrdimensionale Strukturen zum Teil verzerrt. Multivariates Resampling bildet dagegen mehrdimensionale Beziehungen besser ab als univariate Parameter (Gottschalk 2005). Je nach Forschungsaufgabe oder -ziel und der anzuwendenden Analysemethodik ist ein Datennutzer an ein- oder mehrdimensionalen Statistiken interessiert.

Für den Ökonometriker sind insbesondere die Abhängigkeiten zwischen den Merkmalen aufschlussreich. Daher stehen in Gottschalk (2005) die Erfordernisse multivariater Analysen und ökonometrischen Arbeitens im Vordergrund. Im Rahmen einer Monte-Carlo-Simulation wird eine kritische Bandbreite gesucht. Ausgangspunkt dieser Simulation bildet ein beispielhaftes lineares Eingleichungsmodell. Die exogenen und endogenen Variablen dieses Modells werden aus theoretischen Verteilungsfunktionen gezogen. Eine kritische Bandbreite ist hier erreicht, wenn die Koeffizienten dieses linearen Modells durch multivariates Resampling mit einer Kernmatrix, die dieselbe Korrelationsmatrix aufweist wie die Datengrundlage der Monte-Carlo-Simulation (3. Variante des multivariaten Resamplings, siehe oben), gerade noch nicht signifikant von den Schätzwerten mit den Originaldaten abweichen. Das heißt, bei einer marginalen Erhöhung der Bandbreiten an diesem kritischen Punkt wären Parameterschätzungen deutlich verzerrt.

Ausgangspunkt der Monte-Carlo-Simulation bei Gottschalk (2005) sind die quasi-optimalen Bandbreiten für univariate Verteilungen, die Silverman (1986, S.47-48) vorgeschlagen hat.[6] Während der Monte-Carlo-Simulation wird die Bandbreite iterativ erhöht, solange die Restriktion erfüllt ist. Das Maximierungsproblem endet bei einem Bandbreitenfaktor von 2,1. Die Bandbreiten können demzufolge bis um etwa das Doppelte der quasi-optimalen Bandbreiten vergrößert werden, ohne die Aussagekraft der durchgeführten Modellschätzung zu gefährden. Gottschalk (2005) übernimmt diese kritische Bandbreite für Anwendungen mit dem Mannheimer Innovationspanel (MIP) (vgl. Janz et al. (2001)).

Da beim Resampling keine echten Daten weitergegeben werden, ist der Vertrauensschutz grundsätzlich sichergestellt. Dennoch kann keine absolute Anonymität gewährleistet werden, da nicht vollständig ausgeschlossen werden kann, dass die künstlichen Beobachtungen den Merkmalsträgern des Originaldatensatzes zugeordnet werden können (Gottschalk 2005, S.110). Das Risiko ist besonders groß, wenn die künstlichen Beobachtungen den Originalwerten sehr ähnlich sind und wenige auffällige Merkmalsträger existieren. Daher sind auch beim Resampling Untersuchungen hinsichtlich der Sicherstellung der faktischen

6) Sie stellt eine gute Approximation der optimalen Bandbreite (hinsichtlich der Güte der Schätzung univariater Verteilungen) für alle t-Verteilungen und Lognormalverteilungen mit einem Skewnessfaktor bis etwa 1,8 dar.

Anonymität notwendig.

Wenn die Speicherkapazität ausreicht, kann es sinnvoll sein, einem Nutzer statt der Originaldaten mehrere geglättete Resamples zu überlassen, so dass dieser damit den Bootstrap-Schätzer berechnen kann (Gottschalk 2005). Allerdings steigt durch die Verfügbarkeit mehrerer Resamples auch das Reidentifikationsrisiko der Einzelangaben.

Gottschalk (2005) untersucht die unterschiedlichen Varianten des Resamplings hinsichtlich Schutzwirkung und Analysepotenzial bei der Anonymisierung von Einzeldaten in Simulationsexperimenten und mit dem Mannheimer Innovationspanel (MIP). Hinsichtlich des Analysepotenzials werden insbesondere bei Anpassung der Kernmatrix an die Korrelationsstruktur gute Ergebnisse ermittelt.

Das hier beschriebene nicht-parametrische Resamplingverfahren kann als eine Verknüpfung des Resamplingansatzes von Fienberg (1997) und stochastischer Fehlerüberlagerung (vgl. Abschnitt 6.2.3) interpretiert werden. Einerseits ist die Methode eine spezielle praktische Anwendung der Resamplingtheorie, weil die Wahrscheinlichkeitsdichte des Resamples identisch ist mit der geglätteten Kerndichte der Originaldaten und die wahren Beobachtungen nicht eindeutig einem Element des Resamples zugeordnet werden können. Andererseits sind die anonymisierten Daten faktisch eine Stichprobe der Originaldaten und Schritt 3 der Resamplingprozedur ist eine spezielle Form der additiven stochastischen Fehlerüberlagerung.

Die ersten beiden Momente der Variablen können konsistent geschätzt werden. Ferner werden durch multivariates Resampling mehrdimensionale Strukturen abgebildet. Dennoch kann insbesondere bei sehr starker Glättung, wenn aufgrund des hohen Reidentifikationsrisikos große Bandbreitenfaktoren gewählt wurden, nicht sicher gestellt werden, dass im Allgemeinen konsistente Analyseergebnisse erzielt werden. Dem Datennutzer der anonymisierten Resamples können aber Verteilungsparameter der Originalvariablen mitgeteilt werden, da dadurch das Reidentifikationsrisiko der Einzelangaben nicht steigt. Mit dieser Information wird dem Datennutzer ermöglicht, bei starker Glättung der Verteilungen mit Hilfe eines Fehlerkorrekturverfahrens die Qualität und Aussagekraft ökonometrischer Schätzungen zu verbessern.

In den Kapitel 22 werden Korrekturverfahren für lineare und nichtlineare ökonometrische Modelle beschrieben. Beispielsweise ist die Simulations-Extrapolationsmethode (SIMEX) (vgl. Cook und Stefanski (1994) und Abschnitt 22.1.2) eine (relativ zu alternativen anwendbaren Methoden) einfach ausführbare Möglichkeit, Schätzungen zu verbessern. Der SIMEX-Algorithmus setzt keine Annahmen über die Verteilungen der nicht anonymisierten Variablen voraus. Der Prozess der Fehlerüberlagerung wird mithilfe der vorliegenden anonymisierten Daten und mit dem Wissen über die Art der Fehlerüberlagerung simuliert und „negativ“ fortgeschrieben, bis der Fehler verschwindet. So kann der SIMEX-Algorithmus mindestens approximativ konsistente Schätzergebnisse für lineare und nichtlineare Regressionsmodelle erzielen.

6.3 Datenverändernde Verfahren zum Schutz besonders gefährdeter Merkmalsträger

Bisher wurden datenverändernde Verfahren vorgestellt, die auf ganze Datenbestände oder zumindest für einzelne Variablen anwendbar sind. In diesem Abschnitt werden hingegen solche datenverändernden Verfahren betrachtet, die lediglich auf besonders gefährdete Merkmalsträger angewendet werden. Dabei wird danach unterschieden, ob die Verfahren auf einzelne, besonders auffällige, Merkmalsträger oder auf systematisch abgrenzbare Gruppen auffälliger Merkmalsträger beschränkt werden.

6.3.1 Auf einzelne Merkmalsträger beschränkte datenverändernde Verfahren

- **Klonen von Merkmalsträgern**: Einzelne Merkmalsträger, die wegen ihrer seltenen diskreten Merkmalskombinationen auffällig sind, werden anonymisiert, indem gleichartige künstliche Merkmalsträger erzeugt werden. Die künstlichen Merkmalsträger haben die gleichen Ausprägungen bei den diskreten Merkmalen und ähnliche stetige Merkmalswerte (Statistische Ämter des Bundes und der Länder und IAW 2003). Das Verfahren führt zu einer Reduzierung der eindeutigen Zuordnungen und damit zu einer steigenden Unsicherheit beim Datenangreifer. Allerdings werden die uni- und multivariaten Verteilungscharakteristika systematisch verzerrt. Die Verzerrung ist umso größer, je geringer die Anzahl der Merkmalsträger im Datensatz ist.
- **Zerlegung von Merkmalswerten:** Einzelne Merkmalsträger, die wegen der Größe ihrer stetigen Merkmalswerte auffällig sind, werden anonymisiert, indem ihre metrischen Merkmalswerte auf mehrere künstliche Merkmalsträger nach einem geheimen Verteilungsschlüssel verteilt werden (Statistische Ämter des Bundes und der Länder und IAW 2003). Das Verfahren führt sowohl zu einer Reduzierung der Zuordnungswahrscheinlichkeit als auch zu einer Verringerung der Brauchbarkeit der Werte. Allerdings werden auch bei diesem Verfahren die uni- und multivariaten Verteilungscharakteristika systematisch verzerrt.

Klonen bietet sich bei kleineren Unternehmen an, um die Einzigartigkeit von Fällen zu verschleiern, Zerlegung kann der Anonymisierung von Großunternehmen dienen.

6.3.2 Gruppenspezifische datenverändernde Verfahren

Diese Verfahrensgruppe wird zumindest teilweise bereits in der Praxis der statistischen Ämter angewandt und kann daher ebenso wie die in Kapitel 5 beschriebenen Verfahren zu den traditionellen gerechnet werden.

- **Censoring-Verfahren**: Beim Censoring werden Merkmalswerte, die oberhalb oder unterhalb einer gewissen Grenze liegen, auf eben diesen Wert festgesetzt (topcoding, bottomcoding). Zum Beispiel werden besonders hohe Jahresumsätze auf eine obere Grenze gesetzt, um Ausreißer zu vermeiden. Damit wird sowohl das Zuordnungsrisiko als auch die Brauchbarkeit der in dieser Weise veränderten Werte für einen potenziellen Datenangreifer reduziert. Allerdings werden auch die uni- und multivariaten Verteilungscharakteristika für die betroffene Teilgesamtheit und damit auch für den Gesamtdatensatz verzerrt. Im Fall des Censoring sind allerdings zensierte Tobit-Modelle anwendbar, sofern es sich bei der betreffenden Variablen um die abhängige handelt (Statistische Ämter des Bundes und der Länder und IAW 2003).
- **Replacement-Verfahren**: Beim Replacement werden Merkmalswerte, die oberhalb oder unterhalb einer gewissen Grenze liegen, durch das arithmetische Mittel aller Werte oberhalb beziehungsweise unterhalb der entsprechenden Grenze ersetzt. So werden beispielsweise besonders hohe Jahresumsätze durch den Durchschnitt dieser Umsätze ersetzt. Die Schutzwirkung dürfte wegen der im Durchschnitt geringeren Abweichung der Einzelwerte kleiner sein als beim Censoring. Das Verfahren erhält im Gegensatz zum Censoring die arithmetischen Mittel. Allerdings werden die Varianzen und die multivariaten Verteilungscharakteristika verändert (Vorgrimler 2003b; Statistische Ämter des Bundes und der Länder und IAW 2003).

Neben diesen speziell auf besonders gefährdete Gruppen ausgerichteten Verfahren können auch alle anderen datenverändernden Verfahren, wie die Mikroaggregationsverfahren oder die stochastischen Überlagerungen, auf gefährdete Teilgesamtheiten beschränkt werden.

6.4 Kombination unterschiedlicher datenverändernder Verfahren

Eine besondere Bedeutung kommt auch der Kombination unterschiedlicher Verfahren und Verfahrensgruppen zu, insbesondere wenn sowohl metrische als auch kategoriale Variablen anonymisiert werden sollen. In der Praxis sind die verschiedensten Verfahrenskombinationen denkbar. Bisher haben im Wesentlichen zwei Verfahrenskombinationen in der Literatur Beachtung gefunden, das Verfahren von Winkler und SAFE, das Verfahren des Statistischen Landesamts Berlin.

6.4.1 Das Verfahren von Winkler

Das Verfahren von Winkler (Kim und Winkler 1995) stellt eine Kombination aus stochastischer Überlagerung und Zufallsvertauschung dar. Merkmalsträger, die durch stochastische Überlagerung 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 Dataswapping erreicht wird, mit der

höheren Datenqualität der stochastische Überlagerung verbunden wird, da nur bei den Problemfällen Dataswapping eingesetzt wird.

6.4.2 SAFE – Das Verfahren des Statistischen Landesamts Berlin

Die Idee des Verfahrens SAFE besteht darin, einen Datenbestand zu erzeugen, in dem jeder Datensatz mit mindestens zwei weiteren identisch ist (Evers und Höhne 1999; Höhne 2003a,b,c). Das Verfahren SAFE stellt eine Kombination aus dem Vertauschen beziehungsweise Verändern diskreter Merkmalswerte mit einer Mikroaggregation bei den metrischen Variablen dar. 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 Mikroaggregationsverfahren. Das Verfahren ermöglicht auch die Vereinheitlichung von diskreten Merkmalen. Es werden folgende Schritte vorgenommen:

1. 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.
2. Zuordnung der Merkmalswerte der stetigen Variablen: Dabei werden die originalen Sätze zu den Merkmalskombinationen der diskreten Variablen zugeordnet. Ziele sind die Erhaltung der größten Ähnlichkeit in den diskreten Variablen und das Verschieben der kleinsten Merkmalswerte der stetigen Variablen.
3. Bearbeitung der Dominanzen und des Problems der merkmalsbezogenen Fallzahlen unter zwei.
4. Optimierung und Qualitätssicherung der Ergebnisse.
5. Mikroaggregation der stetigen Merkmale innerhalb der gebildeten Gruppen.

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 faktisch anonymisierter Einzeldaten gegebenenfalls verzichtet werden (Ronning et al. 2002; Höhne 2003a; Statistische Ämter des Bundes und der Länder und IAW 2003).

Kapitel 7

Kriterien für die Auswahl der untersuchten Verfahren

Im Rahmen des Projekts wurde eine Reihe der im vorigen Kapitel beschriebenen 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 und gleichzeitig möglichst viele unterschiedliche Verfahrensgruppen zu testen. Eine endgültige qualitative Bewertung der Verfahren war damit noch nicht verbunden. Die Qualitätsbewertung erfolgte im Wesentlichen im Rahmen der im Projekt durchgeführten Verfahrenstests, da auch die konkreten Beurteilungskriterien erst im Projektverlauf entwickelt wurden.
Kriterien zur Verfahrensauswahl waren:

- Leichte Handhabbarkeit des Verfahrens:
  Da die Verfahren später in den statistischen Ämtern und bei den anderen Anbietern von anonymisierten Einzeldaten 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. Es muss möglich sein, nachdem für eine spezielle Statistik die Anonymisierungsanforderungen (Sicherheitsanforderungen und Qualitätskriterien für die Analysefähigkeit) definiert wurden, unabhängig von der jeweiligen Welle der Datenlieferung die Regeln für die Verfahrensanwendung zu formulieren. Verfahren, die z.B. erst eine langwierige Analyse der Abhängigkeiten und Gefährdungsrisiken bei jeder neuen Datenwelle erfordern, erweisen sich als wenig praktikabel für eine regelmäßige Anwendung auf immer wieder neu bereitzustellende Datenbestände.
- Erfolgsaussichten der Verfahren:
  Hier wurde auf Bewertungen in der Literatur zurückgegriffen. Bei manchen Verfahren sind bereits in der Literatur einige Nachteile bekannt, die durch andere Verfahren im Sinne einer Weiterentwicklung bereits gelöst sind. Andere Verfahren befinden sich

noch in einer theoretischen Entwicklungsstufe, so dass sie noch nicht praktikabel einsetzbar sind (z.B. einige Simulationsverfahren). Deshalb wurde versucht, die erfolgversprechenderen Varianten der Verfahren zu nutzen. Wurden beim Anwenden der Verfahren einzelne kleine Mängel festgestellt, so wurden im Projekt ggf. auch Weiterentwicklungen vorgenommen. Diese wurden bei der Beschreibung der Verfahren mit dargestellt.

- Repräsentative Vertretung der Verfahrensgruppen:
  Die Auswahl sollte Verfahren aus allen Verfahrensgruppen enthalten. Da Maße zur Datenqualität im Hinblick auf die Analysefähigkeit und Datensicherheit erst im Laufe des Projekts definiert wurden, sollte keine Verfahrensgruppe generell ausgeschlossen werden, weil jede Verfahrensgruppe einen anderen Ansatz der Veränderung der Einzeldaten repräsentiert. Jede Verfahrensgruppe hat deshalb andere Schwächen und Stärken, die erst am Ende des Projekts abgewogen und beurteilt werden können.

- Methodenmix von Verfahren:
  Oftmals ist eine wirkungsvolle Anonymisierung nur durch das Verwenden von mehreren Verfahren gleichzeitig zu erzielen. Hier müssen natürlich alle Verfahren berücksichtigt werden, die zu einem solchen Methodenmix beitragen können. So lassen sich z.B. die metrischen Angaben großer Einheiten selten durch alleinige Behandlung der kategorialen Variablen schützen. Umgekehrt bewirken datenverändernde Anonymisierungsverfahren für metrische Variablen ohne eine zusätzliche Anonymisierung der kategorialen Variablen meist keine ausreichende Schutzwirkung.

Von den Anonymisierungsverfahren (siehe Abb. 7.1) wurde eine Reihe von Verfahren auf Basis der beschriebenen Auswahlkriterien für intensive Untersuchungen ausgewählt. Von den datenverändernden Verfahren für metrische Variablen wurden stochastische Überlagerungen und Mikroaggregationsverfahren am intensivsten getestet. Weniger intensiv wurde eine Auswahl von Simulationsverfahren (Resampling und Latin Hypercube Sampling) untersucht. Von den datenverändernden Verfahren für kategoriale Variablen wurde die Post-Randomisierung getestet. Informationsreduzierende Verfahren wurden im Einzelfall problemadäquat eingesetzt.

Bei folgenden Verfahren beziehungsweise Verfahrensgruppen wurde von vornherein auf einen Test verzichtet:

- *local suppression (auch mit Imputationen), Klonen, Zerlegung:*
  Diese Verfahren sind nur dann wirkungsvoll, wenn im Datenbestand sehr wenige geheimhaltungsbedürftige Merkmalsträger enthalten sind. Dabei bedürfen diese Verfahren eines großen manuellen Aufwandes, um diese Merkmalsträger zu lokalisieren und in den Datenbestand einzugreifen. Da die Verfahren somit schlecht automatisierbar sind, wurden sie als wenig praktikabel eingestuft.
  Bei local suppression mit Imputationen ergibt sich außerdem dass Problem, dass für

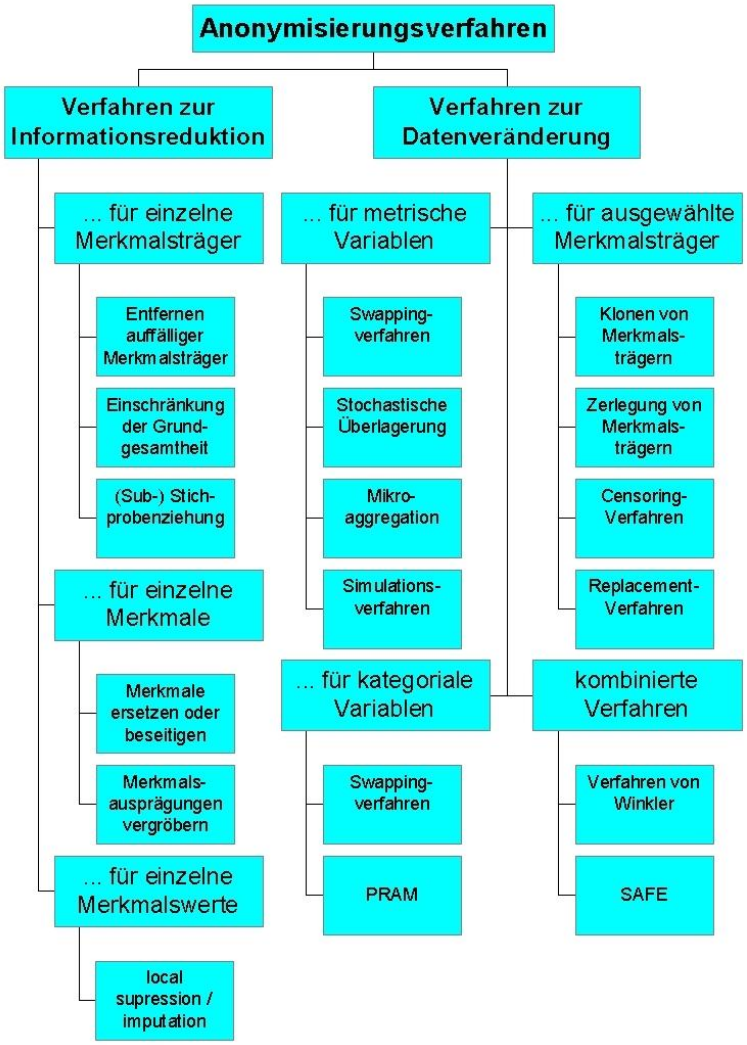

Abbildung 7.1: Übersicht der Anonymisierungsverfahren

Imputationen Modelleigenschaften zugrunde gelegt werden müssen, die dann analoge Modellanalysen der Daten systematisch verzerren.

- *Stochastische Überlagerung nach dem Verfahren von Sullivan:*
  Zum Zeitpunkt des Projektbeginns stellte das Verfahren von Sullivan eine sehr moderne Erweiterung der stochastischen Überlagerung dar. Mit diesem Verfahren schien sowohl die Anonymisierung kategorialer Variablen als auch die Berücksichtigung der

schiefen Verteilungen, die bei Unternehmensdaten vorwiegend anzutreffen sind, möglich. Leider konnte das Verfahren den hohen Erwartungen nicht gerecht werden. Das lag in erster Linie darin begründet, dass die Anwendung des Verfahrens ein sehr hohes Spezialwissen voraussetzt. Dieses kann in den statistischen Ämtern aber nicht als Standardfertigkeit für die Anwendung der Anonymisierungsverfahren erwartet werden. Die verfügbaren Programmversionen des Verfahrens ermöglichen es insbesondere nicht, die Projektdatensätze in einem Stück zu anonymisieren. Damit ergeben sich für die Anonymisierung immer sehr umfangreiche Datenmanipulationsarbeiten. Die Anwendung des Verfahrens auf kategoriale Variablen basiert auf einer Transformation von kategorialen in stetige Variablen (siehe Brand (2000)). Diese ist jedoch nur unter großen Restriktionen einsetzbar, so dass sie für die Projektdaten nicht nutzbar war.

- *Swapping für kategoriale Variablen:*
  Auf das Swapping für kategoriale Variablen wurde verzichtet, weil es mit den Verfahren der Postrandomisierung (PRAM) verwandt ist. Auch hier lassen sich auf Basis der Häufigkeit des Auftretens von Kategorien im Datenbestand „Übergangswahrscheinlichkeiten“ bestimmen, mit denen festgelegt wird, mit welcher Wahrscheinlichkeit eine Ausprägung mit einer anderen „geswappt“ wird. Die Eigenschaften sind somit gleichartig modellierbar. PRAM wurde einer intensiven methodischen Betrachtung unterzogen.

- *Verfahren von Winkler:*
  Das Verfahren von Winkler als einer Mischung aus Zufallsüberlagerung und punktuellem Dataswapping wurde nicht untersucht, weil die schlechten Erfahrungen im Projektverlauf mit der additiven Zufallsüberlagerung und den Swappingverfahren (Rankswapping) dem Verfahren keine Erfolgsaussichten beimaßen. Außerdem ist das Verfahren mit einem hohen manuellen Aufwand bei der Bestimmung der „Restrisikobereiche“ verbunden, der es schwer praktikabel einsetzbar macht.

Weitere Verfahren wurden zwar geprüft und in Ansätzen empirisch getestet, aufgrund der im Projektverlauf gemachten Erfahrungen (insbesondere bis zum Zwischenbericht), jedoch vor allem wegen ihrer einschneidenden Auswirkungen auf das Analysepotenzial nicht weiter untersucht (Statistische Ämter des Bundes und der Länder und IAW 2003, S. 64 ff.). Dabei handelt es sich um:

- *Entfernen auffälliger Merkmalsträger:*
  Das Entfernen auffälliger Merkmalsträger erweist sich in wirtschaftsstatischen Datenbeständen als ein sehr problematisches Verfahren. Auffällige Einheiten, d.h. Einheiten, für die viel Zusatzwissen verfügbar gemacht werden kann, sind gerade die großen Einheiten im Datenbestand. Diese tragen – wie dies bei Personendaten so nicht denkbar wäre – einen erheblichen Anteil zur Gesamtmasse bei den stetigen Merkmalen bei. Bei wirtschaftsstatistischen Mikrodaten sind Merkmalsträger oftmals deshalb auffällig, weil sie eine Monopolstellung haben (z.B. Bahn, Post u.a.). Ein Verzicht auf

diese Einheiten würde sowohl die Abbildung ganzer Branchen verhindern, als auch die gesamtwirtchaftliche Darstellung stark verzerren, da die Objekte sehr groß sind. (siehe auch Bemerkungen bei Stichprobenziehung)

- *Stichprobenziehung/Substichprobenziehung:*
  Die Stichprobenziehung wurde keiner intensiveren Untersuchung unterzogen, weil es sich bei Unternehmensdaten in der Regel um bedeutend kleinere Grundgesamtheiten als bei Personenstatistiken handelt.
  Hier würde durch das Ziehen einer Stichprobe die Unsicherheit in den Daten nicht bedeutend vergrößert, wie bei sehr großen Grundgesamtheiten. Besitzt ein Datenangreifer nämlich qualitativ hochwertiges Angriffswissen, so ist er trotzdem in der Lage für die im Bestand verbliebenen Merkmalsträger eine sichere Zuordnung vorzunehmen bzw. eine Falschzuordnung zu erkennen. Viele Unternehmensstatistiken stellen bereits Stichprobenerhebungen dar. Gerade bei großen Objekten sind dort aber sehr hohe Auswahlsätze vorhanden (oft auch Vollerhebung). Eine höhere Sicherheit kann hier nur durch sehr kleine Auswahlsätze für die Substichprobe bei großen Einheiten erreicht werden. Auf Grund des hohen Beitrages der großen Merkmalsträger zum Gesamtvolumen der Merkmale können diese geringen Auswahlsätze jedoch nicht angestrebt werden, da die Stichprobenfehler bei Analysen mit den Daten erheblich zunehmen würden. (Statistische Ämter des Bundes und der Länder und IAW 2003, S. 65)
  Außerdem ist die Methode der Stichprobenziehung mit Zurücklegen für Unternehmensdaten nicht brauchbar, weil der hohe Anteil stetiger Merkmale es einem Datenangreifer leicht ermöglicht, Doppelziehungen zu erkennen. Die völlige Identität der sehr fein ausgewiesenen stetigen Merkmale kann hier als sicheres Indiz für eine Doppelziehung angesehen werden. Nur bei gleichzeitiger Anwendung von starker Rundung der Werte wäre eine Stichprobenziehung mit Zurücklegen anwendbar.
  Stichprobenziehung bewirkt somit nur für die nicht gezogenen Einheiten eine Schutzwirkung, während die Schutzwirkung für die gezogenen Einheiten vernachlässigt werden kann, wenn nicht sehr niedrige Auswahlsätze aus homogenen Datenbeständen gezogen werden. Homogene Datenbestände sind aber bei wirtschaftsstatistischen Datenbeständen nicht üblich.
  Stichprobenziehung erfolgreich anzuwenden ist deshalb nur möglich, wenn sie in Verbindung mit anderen Verfahren erfolgt, die die sehr großen Einheiten (Ausreißer) im Datenbestand schützen. Dafür bieten sich z.B. die Maßnahmen „Einschränkung der Grundgesamtheit“ oder Verfahren wie „Censoring“ oder „Replacement“ an. Beides wurde im Projektverlauf getestet und im Zwischenbericht dokumentiert (Statistische Ämter des Bundes und der Länder und IAW 2003, S. 64 ff.). Die Tests bei der Anonymisierung der Umsatzsteuerstatistik zeigten, dass die Stichprobenziehung nicht mit anderen bis zum Zwischenbericht getesteten Verfahren mithalten konnte. Sowohl bei den deskriptiven Analysen als auch bei ökonometrischen Schätzungen konnte die Zufallsauswahl nicht überzeugen.
  Nichtsdestotrotz hat eine bei der Erhebung bereits durchgeführte Stichprobenziehung eine Anonymisierungswirkung auf den Datenbestand, die besonders bei kleinen Unter-

nehmen dazu führen kann, dass auf starke zusätzliche Anonymisierungsmaßnahmen verzichtet werden kann (siehe z.B. Anonymisierung der Kostenstrukturerhebung und der Einzelhandelsstatistik).
Aus der Gruppe der Verfahren der Informationsreduktion für einzelne Merkmalsträger stellt somit die Einschränkung der Grundgesamtheit ein viel wertvolleres Verfahren dar, weil die Regeln der Einschränkung dokumentiert und bei Analysen berücksichtigt werden können. Hier kann sich die Einschränkung genau auf den Bereich der datenschutzkritischen Merkmalsträger konzentrieren.

- *Rankswapping:*
  Rankswapping wurde aus der Gruppe der Swappingverfahren in der ersten Projekthälfte getestet. Auch dieses Verfahren hatte wie z.B. Stichprobenziehung bezüglich der Analysequalität der anonymen Daten erhebliche Mängel, so dass es nicht weiter betrachtet wurde. Der Erhalt der ersten und zweiten Momente beim Rankswapping ist leider mit starken Fehlern bei den Korrelationskoeffizenten und den mehrdimensionalen Häufigkeitsverteilungen verbunden. Diese starken Fehler führen auch zu sehr schlechten Ergebnissen in ökonometrischen Analysen. Das Verfahren wurde deshalb bereits im Zwischenbericht als nicht nutzbar für die Erstellung von Scientific-Use-Files eingeschätzt.

- *SAFE:*
  Das Verfahren SAFE basiert auf der völligen Vereinheitlichung von mindesten drei Merkmalsträgern. Das wird durch eine Veränderung der kategorialen Merkmale derart erreicht, das jeder Merkmalsträger mit mindestens zwei weiteren Merkmalsträgern identisch ist. Dass die Entscheidung über die Veränderung der kategorialen Merkmale auf der Grundlage der Minimierung der Abweichungen in den Häufigkeitsverteilungen erfolgt, bewirkt einerseits eine Verschiebung von möglichst wenigen Einheiten. Diese ist jedoch in Einzelfällen mit sehr starken Verschiebungen der metrischen Merkmalswerte verbunden, wenn große Unternehmen in einer Branche eine Monopolstellung inne haben. In diesem Fall ist es mit dem Verfahren nur über die Zusammenlegung mit einer anderen Branche möglich, die Anonymität zu gewährleisten. Da diese Zusammenlegungen aus Sicherheitsgründen nur eingeschränkt dokumentiert werden können, ist die Brauchbarkeit der Angaben stark gefährdet, weil die Mittelwerte der Gruppen durch einzelne Ausreißer stark verzerrt werden können (Statistische Ämter des Bundes und der Länder und IAW 2003, S. 62 f.). Diese Ausreißerproblematik führte dazu, dass die Akzeptanz für das Verfahren im Bereich der Wissenschaft stark eingeschränkt war. Deshalb wurde das Verfahren nicht weiter verfolgt.

Teil III

Die Projektdatensätze

Die Untersuchung der Wirkung der unterschiedlichen Anonymisierungsverfahren auf Analysepotenzial und Datensicherheit wurde im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ vorrangig an konkreten Datengrundlagen der statistischen Ämter und des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) vorgenommen. In diesem Teil werden die zu Testrechnungen herangezogenen Projektstatistiken näher beschrieben. In Kapitel 8 werden zunächst die Kriterien für die Auswahl der Projektdatensätze dargestellt. Kapitel 9 enthält die Beschreibung der einzelnen Projektstatistiken (Kostenstrukturerhebung im Verarbeitenden Gewerbe und Bergbau, Umsatzsteuerstatistik, Einzelhandelsstatistik und IAB-Betriebspanel).

Kapitel 8

Kriterien für die Auswahl der Projektdatensätze

Die statistischen Ämter erheben in regelmäßigen Abständen – im Wesentlichen monatlich, quartalsweise, jährlich – Daten über Unternehmen und Betriebe. Diese Erhebungen sind gekennzeichnet durch eine Reihe von Eigenschaften, die sie für die wissenschaftliche Forschung interessant machen:

- Art der Berichtskreise: Die Daten werden bei systematisch ausgewählten Berichtskreisen erhoben. Soweit Stichproben gezogen werden, werden dazu anhand der Struktur der Grundgesamtheiten spezifische Auswahlpläne erstellt. Das ermöglicht qualifizierte Hochrechnungen.
- Größe der Berichtskreise: Die Erhebungen der statistischen Ämter umfassen Berichtskreise, die in diesem Umfang für einzelne wissenschaftliche Forschungsprojekte nicht befragt werden können. Manche Erhebungen beziehen mehrere zehntausend Befragte ein.
- Merkmale: Erfragt werden monatlich und quartalsweise die wichtigsten konjunkturellen Indikatoren, jährlich oder mehrjährlich ein breiter Kranz von ökonomischen Strukturdaten. Die konkreten Erhebungsinhalte sind dabei auf die Gegebenheiten der Wirtschaftsbereiche abgestimmt.
- Wiederholungsbefragungen mit stabilem Merkmalskanon: Die Erhebungen sind gesetzlich angeordnet und werden regelmäßig durchgeführt. Aufgrund der eingerichteten Erhebungsorganisation sind alle Arbeitsschritte vom Versand der Fragebögen über die Kontrolle des Rücklaufs, der Erfassung und Plausibilisierung der Meldungen durch EDV-Verfahren (und der nachfolgenden Arbeiten zur Erstellung statistischer Daten und ihrer Veröffentlichung) aufeinander abgestimmt und werden nach einem festen Zeitplan durchgeführt.
- Auskunftspflicht: Die Erhebungen der statistischen Ämter werden in aller Regel mit Auskunftspflicht durchgeführt. Dadurch wird ein sehr hoher Rücklauf innerhalb bekannter Zeithorizonte erreicht.

Das Einzeldatenmaterial bundesstatistischer Erhebungen ist aus diesen Gründen sehr aussagekräftig und zuverlässig und damit für wissenschaftliche Forschung wertvoll. Ziel der Projektdatenauswahl war es, die Auswirkungen von Anonymisierungsverfahren auf unterschiedlich strukturierte Daten zu untersuchen. Grundlegende Unterschiede zweier Datensätze können im Umfang (Anzahl der im Datensatz enthaltenen Merkmalsträger) oder in der Dichte (Verteilung der Merkmalsträger bezüglich der im Datensatz enthaltenen qualitativen Merkmale) bestehen. Feinere Unterschiede werden bei der Betrachtung der Merkmalstypen (metrisch oder kategorial) der in den Datensätzen enthaltenen Merkmale sichtbar. Um aussagekräftige Bewertungen von Anonymisierungsmethoden und deren Auswirkungen zu erhalten, wurde zumeist realen Daten der Vorzug gegenüber synthetischen Daten gegeben.

In den folgenden Abschnitten werden vier Datensätze beschrieben, die im Projekt besondere Aufmerksamkeit erfahren haben: Die amtlichen Datensätze der Kostenstrukturerhebung im Verarbeitenden Gewerbe des Jahres 1999 (Abschnitt 9.1), der Umsatzsteuerstatistik des Jahres 2000 (Abschnitt 9.2), der Einzelhandelsstatistik des Jahres 1999 (Abschnitt 9.3) sowie das Betriebspanel des Instituts für Arbeitsmarkt- und Berufsforschung (Abschnitt 9.4).

Bei der Umsatzsteuerstatistik handelt es sich um eine Erhebung mit niedriger Abschneidegrenze (nahezu alle Unternehmen mit einem Jahresumsatz von über 16.617 Euro werden erfasst), die alle Wirtschaftszweige einbezieht. Von den hier betrachteten Erhebungen weist sie mit über 2,9 Millionen Merkmalsträgern den größten Umfang auf. Gleichzeitig besitzt sie eine etwas geringere Anzahl an Merkmalen (27), deren metrische sich auf zwei Gruppen von Aussagetypen, nämlich Umsätze und die darauf anfallenden Steuern verteilen. Aufgrund dieser Struktur wurden die Anonymisierungsmöglichkeiten für diese Erhebung zu Beginn des Projekts positiv beurteilt.

Im Gegensatz zur Umsatzsteuerstatistik ist die Kostenstrukturerhebung eine Stichprobenerhebung mit Abschneidegrenze (es werden Unternehmen mit wenigstens 20 Beschäftigten erhoben), die sich auf einen Wirtschaftsbereich – das Verarbeitende Gewerbe und den Bergbau – beschränkt. Der für die Projektarbeiten ausgewählte Auszug weist 36 (davon 33 metrische) Merkmale auf. Drei metrische Merkmale wurden im Laufe der Projektarbeiten entfernt. Sie hat eine vergleichsweise geringe Anzahl an Merkmalsträgern (rund 17.000), die aber einen hohen Anteil der Wertschöpfung in ihrem Wirtschaftszweig abdecken. Aus diesem Grunde wurden die Anonymisierungsmöglichkeiten der Kostenstrukturerhebung zu Projektbeginn als besonders schwierig eingeschätzt.

Ähnlich schwierig erschienen die Möglichkeiten einer Anonymisierung der Einzelhandelsstatistik, die eine Stichprobenerhebung ohne Abschneidegrenze mit rund 23.500 Merkmalsträgern darstellt. Während sich die Merkmalsträger der Umsatzsteuerstatistik auf sämtliche Abteilungen und die der Kostenstrukturerhebung auf 28 Abteilungen (so genannte Zweisteller) verteilen, wird der Einzelhandel in eine einzige Abteilung (der Zweisteller 52) klassifiziert, die sich in der nächst tieferen Gliederungsebene in sieben Gruppen (Dreisteller)

zerlegt. Mit 36 (davon 34 metrischen) Merkmalen liegt sie im Bereich der Kostenstrukturerhebung.

Das IAB-Betriebspanel zeichnet sich im Vergleich mit den anderen Projektstatistiken durch eine große Zahl an kategorialen (auch binären) Merkmalen aus. Es eignet sich daher besonders zur Untersuchung datenverändernder Verfahren bei kategorialen Merkmalen. Beim IAB-Betriebspanel handelt es sich, wie bei der Kostenstrukturerhebung und der Einzelhandelsstatistik, um eine nach Betriebsgröße und Wirtschaftszweig geschichtete Stichprobe. Im Unterschied zu den Erhebungen der statistischen Ämter handelt es sich bei den betrachteten Einheiten nicht um Unternehmen, sondern um Betriebe. Die Anzahl der Betriebe im Bundesdatensatz für das im Projekt betrachtete Jahr 2002 beträgt etwa 14.800 und liegt damit in einer ähnlichen Größenordnung wie die Anzahl der Unternehmen in der Kostenstrukturerhebung im Verarbeitenden Gewerbe. Die Anzahl der Betriebe für Baden-Württemberg, mit denen ebenfalls gearbeitet wurde, liegt bei etwa 1.200. Dies stellt im Projektkontext eine kleine Erhebung dar. Ebenso wie die Umsatzsteuerstatistik ist auch das IAB-Betriebspanel ein branchenübergreifender Datensatz. Das IAB-Betriebspanel ergänzt das inhaltliche Angebot der Projektstatistiken insbesondere um Arbeitsmarktinformationen. Wie von Seiten der statistischen Ämter und ihren Forschungsdatenzentren, besteht auch von Seiten des IAB Interesse daran, seine Betriebsdaten der Wissenschaft in faktisch anonymisierter Form zur Verfügung zu stellen.[7] Allerdings wurde im Rahmen des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ keine faktische Anonymisierung des IAB-Betriebspanels und somit auch nicht die Erstellung eines Scientific-Use-Files mit den Betriebsdaten des IAB angestrebt. Vielmehr diente mit der Welle des Jahres 2002 ein Querschnitt aus dem IAB-Betriebspanel als Grundlage für Tests mit verschiedenen datenverändernden Anonymisierungsverfahren. [8]

7) Zu ersten Versuchen der Anonymisierung des IAB-Betriebspanels vgl. Brand et al. (1999).

8) Die Anonymisierung des IAB-Betriebspanels ist u.a. Gegenstand eines Folgeprojekts der Forschungsdatenzentren von Bund und Ländern, des IAW und des IAB.

Kapitel 9

Die eingesetzten Projektdaten

9.1 Die Kostenstrukturerhebung im Verarbeitenden Gewerbe und Bergbau

9.1.1 Ziele der Kostenstrukturerhebung

Inhaltlich liefert die Kostenstrukturerhebung (KSE) die umfassendsten Informationen zu den Unternehmen im Bereich der Statistik im Produzierenden Gewerbe. Sie dienen als Ausgangspunkt für vielfältige Strukturuntersuchungen nicht nur in Politik und Verwaltung, sondern auch in der Wirtschaft und ihren Verbänden sowie in der Wissenschaft und vielen anderen gesellschaftlichen Gruppierungen. Die Informationen der KSE bilden darüber hinaus eine unentbehrliche Datengrundlage für die Volkswirtschaftlichen Gesamtrechnungen. Hier werden die Ergebnisse vor allem für die Berechnung der Wertschöpfung und ihrer Komponenten nach Wirtschaftsbereichen im Rahmen der Entstehungsrechnung herangezogen; schließlich liefern sie auch wichtige Informationen für die Input-Output-Rechnungen. Im Rahmen der Statistiken im Produzierenden Gewerbe bilden die Kostenstrukturstatistiken u.a. eine Grundlage für die Gewichtung von Produktionsindizes.

9.1.2 Methode der Erhebung

Die Kostenstrukturerhebung im Verarbeitenden Gewerbe erfasst als hochrechnungsfähige Stichprobe maximal 18.000 Unternehmen mit 20 und mehr Beschäftigten. Die Befragung erfolgt zentral durch das Statistische Bundesamt im Wege der Selbstausfüllung durch die Unternehmen. Die aus 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 und mehr Beschäftigten, aber auch Unternehmen in Wirtschaftszweigen mit geringer Besetzungszahl werden zur Sicherstellung der Qualität der Ergebnisse vollständig einbezogen. Den Ergebnissen ab dem Berichtsjahr 1999 liegt eine neue Stichprobenauswahl zugrunde. In dieser Stichprobe werden rund 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 führt dazu, dass diese Unternehmen 76% zur Gesamtzahl der tätigen Personen und 84% zum Gesamtumsatz im Berichtskreis beitragen.

Anteil der Stichprobenunternehmen an allen Unternehmen des Berichtskreises in den Größenklassen:

20-249 Beschäftigte: ca. 38%
250-499 Beschäftigte: ca. 73%
500-999 Beschäftigte: 100%
1.000 und mehr Beschäftigte: 100%

9.1.3 Inhalte der Erhebung

Im Projekt wurden Daten der Kostenstrukturerhebung des Jahres 1999 betrachtet. Tabelle 9.1 liefert Auskunft darüber, wie viele von den 16.918 in die Auswertung einbezogenen Unternehmen Angaben zu unterschiedlichen Merkmalen machten.

Tabelle 9.1: Fallzahlen nach ausgewählten Merkmalen (KSE)

Teilzeitbeschäftigte	13.776
Umsatz aus Handelsware	8.027
Einsatz an Handelsware	8.027
Sonstige Sozialkosten	13.848
Kosten für Leiharbeitnehmer	6.735
Kosten für Lohnarbeiten	8.685
Kosten für Reparaturen	16.138
Mieten und Pachten	16.014
Nachrichtlich: Anzahl der Datensätze insgesamt	16.918

In einem zweistufigen Hochrechnungsverfahren werden die Stichprobenergebnisse auf die Gesamtheit der Unternehmen mit 20 Beschäftigten und mehr hochgerechnet. Dabei werden nach einer freien Hochrechnung in einem zweiten Schritt die frei hochgerechneten Werte an die Ergebnisse der Investitionserhebung für Unternehmen, in der die Grundgesamtheit aller Unternehmen mit 20 Beschäftigten und mehr erfasst wird, mittels Korrekturfaktoren angeglichen. Die Korrekturfaktoren werden für drei Bezugsmerkmale (Anzahl der Unternehmen,

Umsatz, Beschäftigte) durch Abgleich zwischen den frei hochgerechneten Ergebnissen und den erhobenen Ergebnissen der Hochrechnungsgrundlage (Investitionserhebung) ermittelt.

Die hochgerechneten Ergebnisse liefern veröffentlichungsfähige absolute Werte, die von Jahr zu Jahr miteinander verglichen und deren zwischenzeitliche Veränderungen mit ausreichender Sicherheit festgestellt werden können. Die Ergebnisse werden gegliedert nach der Klassifikation der Wirtschaftszweige (WZ 93) und Beschäftigtengrößenklassen dargestellt, wobei die Zuordnung der Unternehmen nach ihrem wirtschaftlichen Schwerpunkt erfolgt.

Tabelle 9.2: Merkmale des Datensatzes der Kostenstrukturerhebung

1.	Wirtschaftszweig (WZ93)
2.	Regionalbezug (Ost-West)
3.	Beschäftigtengrößenklasse 07=20-49 08=50-99 11=100-249 14=250-499 17=500-999 22=1000 und mehr
4.	Teilzeitbeschäftigte
5.	Teilzeitbeschäftigte umgerechnet in Vollzeiteinheiten
6.	Tätige Personen insgesamt
7.	Umsatz aus eigenen Erzeugnissen
8.	Umsatz aus Handelsware
9.	Gesamtumsatz (entspricht nicht der Summe aus 7. und 8.)
10.	Anfangsbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen am Umsatz aus eigenen Erzeugnissen
11.	Endbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen am Umsatz aus eigenen Erzeugnissen
12.	Gesamtleistung/Bruttoproduktionswert
13.	Anfangsbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und Betriebsstoffen gemessen am Umsatz aus eigenen Erzeugnissen
14.	Endbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und Betriebsstoffen gemessen am Umsatz aus eigenen Erzeugnissen
15.	Verbrauch an Rohstoffen
16.	Energieverbrauch
17.	Anfangsbestand an Handelsware gemessen am Umsatz aus Handelsware
18.	Endbestand an Handelsware gemessen am Umsatz aus Handelsware
19.	Einsatz an Handelsware
20.	Bruttogehalts- und -lohnsumme
21.	Gesetzliche Sozialkosten
22.	Sonstige Sozialkosten
23.	Kosten für Leiharbeitnehmer
24.	Kosten für Lohnarbeiten
25.	Kosten für Reparaturen
26.	Mieten und Pachten
27.	Sonstige Kosten
28.	Fremdkapitalzinsen
29.	Kosten insgesamt
30.	Bruttowertschöpfung zu Faktorkosten
31.	Nettowertschöpfung zu Faktorkosten
32.	Gesamtaufwendungen für innerbetriebliche Forschung und Entwicklung
33.	Anzahl der für Forschung und Entwicklung eingesetzten Lohn- und Gehaltsempfänger

9.2 Die Umsatzsteuerstatistik

9.2.1 Datengrundlage

Im Rahmen der Umsatzbesteuerung werden von den Unternehmen Umsatzsteuer-Voranmeldungen bei den Finanzbehörden abgegeben. Durch die Auswertung der monatlichen und vierteljährlichen Umsatzsteuer-Voranmeldungen gewinnt die amtliche Statistik Informationen über die Entstehung der Umsatzsteuer, über steuerpflichtige Unternehmen und deren Umsätze sowie über die innergemeinschaftlichen Erwerbe (Einfuhren aus anderen Mitgliedstaaten der Europäischen Union). Abweichungen zwischen den Angaben der Voranmeldungen und den tatsächlichen Umsätzen (beispielsweise durch Korrekturen bei Rückgaben) sind möglich, jedoch insgesamt nicht von größerer Bedeutung. Die Durchführung der Umsatzsteuerstatistik als Bundesstatistik ist im Gesetz über Steuerstatistiken (StStatG) vom Oktober 1995 geregelt.

Umsatzsteuerpflichtig und damit in der Umsatzsteuerstatistik abgebildet sind grundsätzlich alle Unternehmer, die Lieferungen und Leistungen im Inland gegen Entgelt im Rahmen ihres Unternehmens erbringen. Nicht erfasst sind in der Umsatzsteuerstatistik Unternehmen, die aufgrund ihrer Umsatzhöhe oder durch die Erbringung vorwiegend umsatzsteuerfreier Leistungen von der unterjährigen Abgabe von Umsatzsteuer-Voranmeldungen befreit sind (z.B. niedergelassene Ärzte, Behörden, Theater und Museen). Die Umsatzsteuerstatistik liefert somit für fast alle Wirtschaftsbereiche Daten. Es ist zu beachten, dass bei finanziell, wirtschaftlich und organisatorisch zusammen gehörenden Unternehmen (z.B. Filialen, Zweigbetrieben oder Tochterunternehmen), die Unternehmen als Einheit mit dem gesamten Jahresumsatz von dem für den Sitz der Geschäftsleitung zuständigen Finanzamt zentral erfasst werden.

9.2.2 Aussagekraft der Umsatzsteuerstatistik

In der Umsatzsteuerstatistik werden u.a. folgende kategorialen Merkmale abgebildet: Gewerbekennzahl, amtlicher Gemeindeschlüssel, letztgültige Zahlungsweise, Dauer der Steuerpflicht, Organschaft, Rechtsform. Die metrischen Merkmale umfassen u.a.: steuerbarer Umsatz (ohne die der Einfuhrumsatzsteuer unterliegenden Umsätze), Umsatzsteuer vor Abzug der Vorsteuerbeträge, abziehbare Vorsteuerbeträge, Umsatzsteuer-Vorauszahlung. Für wirtschaftsstatistische Untersuchungen ist dabei der Umsatz der Unternehmen von besonderem Interesse. Zum steuerbaren Umsatz der Unternehmen zählen im Umsatzsteuerrecht neben den als „Lieferungen und Leistungen“ bezeichneten Umsätzen der Unternehmen auch deren aus EU-Ländern importierte Güter (innergemeinschaftliche Erwerbe). Als Umsatz werden im wirtschaftsstatistischen Sinn aber lediglich die Lieferungen und Leistungen betrachtet, da die innergemeinschaftlichen Erwerbe den Vorleistungen zuzurechnen sind.

Im Vergleich zu Primärerhebungen ist neben der erhebungstechnischen Abgrenzung des Umsatzes auch die inhaltliche Abgrenzung des steuerlichen Umsatzbegriffs zu beachten. Während in Primärerhebungen allein der Umsatz aus der laufenden Produktionstätigkeit betrachtet wird, beinhaltet der steuerliche Umsatz zusätzlich außerordentliche Erträge, z.B. aus einem nicht dem Betriebszweck dienenden Verkauf von Gebäuden. Das am häufigsten genutzte kategoriale Merkmal der Umsatzsteuerstatistik ist die Wirtschaftszweigzuordnung (Klassifikation der Wirtschaftszweige, WZ93). Maßgebend für die Zuordnung zu einem Wirtschaftszweig ist der Schwerpunkt der wirtschaftlichen Tätigkeit eines Unternehmens. Ergebnisse zu allen 830 nachgewiesenen Wirtschaftszweigen können kostenlos über den Statistik-Shop des Statistischen Bundesamtes bezogen werden (www.ec-destatis.de http://www.ec-destatis.de, vgl. Dittrich (2004)).

9.2.3 Wirtschaftsstatistische Analysemöglichkeiten

Neben den unmittelbar aus den Daten der Finanzverwaltung ableitbaren Erkenntnissen lässt sich die Umsatzsteuerstatistik auch als Basis für weitere wirtschaftsstatistische Analysen nutzen.

Eine nahe liegende Untersuchungsmöglichkeit, welche die Umsatzsteuerstatistik bietet, ist die Analyse der Unternehmenskonzentration. Die zugrunde gelegten Daten der Umsatzsteuerstatistik enthalten die Unternehmensumsätze der Jahre 2000 und 1999, womit die Konzentration für diese beiden Zeitpunkte messbar wird. Der größte Vorteil der Umsatzsteuerstatistik für Konzentrationsanalysen liegt in ihrer fast vollständigen Erfassung der Unternehmen. Nachteilig ist, dass die gesamten Umsätze von Unternehmen, die in mehreren Branchen tätig sind, einer Branche zugeschlagen werden. Des Weiteren können Unternehmensverflechtungen nicht abgebildet werden. Trotz dieser Einschränkungen kann die Umsatzsteuerstatistik erste sichere Hinweise darauf geben, welche Wirtschaftsbereiche stärker durch Unternehmenskonzentration gekennzeichnet sind.

Nach § 15 Abs. 2 Umsatzsteuerstatistikgesetz (UStG) sind steuerfreie Lieferungen und Leistungen vom Vorsteuerabzug ausgeschlossen. Dieser grundsätzliche Ausschluss gilt jedoch nicht bei Lieferungen in Mitgliedstaaten der EU sowie Ausfuhren in Drittländer. Eine Steuerfreiheit mit Vorsteuerabzug ist darüber hinaus bspw. für Umsätze der Seeschifffahrt und der Luftfahrt, Goldlieferungen an Zentralbanken und Lieferungen an Vertragsparteien des Nordatlantikvertrages vorgesehen. Diese Lieferungen sind nicht quantifizierbar, dürften jedoch insgesamt nicht von größerer Bedeutung sein, so dass anhand des Merkmals der steuerfreien Lieferungen und Leistungen mit Vorsteuerabzug die Exporttätigkeit der Unternehmen beschrieben werden kann. Die Exportquoten können mit allen kategorialen Merkmalen des Mikrodatenfiles kombiniert werden (Wirtschaftsklassifikation, Rechtsform oder Regionalgliederung). Zudem werden die steuerfreien Umsätze mit Vorsteuerabzug für innergemeinschaftliche Lieferungen an Abnehmer mit Umsatzsteuer-Identifikationsnummer (§ 4 Nr. 1b UStG) gesondert nachgewiesen, sodass sich auch die EU-Exporte gezielt un-

tersuchen lassen.

Umfassende Importquoten lassen sich mit der Umsatzsteuerstatistik hingegen nicht bilden, da bei Importen aus Drittländern die so genannte Einfuhrumsatzsteuer direkt bei der Zollverwaltung erhoben wird. Dagegen werden die Importe aus EU-Mitgliedstaaten als innergemeinschaftliche Erwerbe im Rahmen der Umsatzsteuer-Voranmeldungen angegeben. Es lassen sich daher anhand der Daten immerhin die EU-Importe näher untersuchen.

Ein weiteres Forschungsfeld ist die Untersuchung der Auswirkungen von Steuerrechtsänderungen im Bereich der Umsatzsteuer. Wie aktuelle Untersuchungen zeigen, spielen steuerpolitische Themen in der öffentlichen Diskussion immer wieder eine große Rolle. So kann beispielsweise die Frage, welche Branchen am meisten von dem ermäßigten Umsatzsteuersatz profitieren, mit den Daten der Umsatzsteuerstatistik beantwortet werden.

Durch das Merkmal Beginn der Steuerpflicht können Analysen der Unternehmensdemographie durchgeführt werden. Die Möglichkeiten und Grenzen der Umsatzsteuerstatistik in diesem Zusammenhang werden in Gräb und Zwick (2002) beschrieben.

Die Datensatzbeschreibung der Umsatzsteuerstatistik, wie sie als Projektstatistik in die Projektarbeiten eingegangen ist, ist in Tabelle 9.3 enthalten.

In Tabelle 9.4 sind die Ausprägungen für die Dauer der Steuerpflicht, in Tabelle 9.5 für die Rechtsform dargestellt.

Tabelle 9.3: Datensatzbeschreibung der Projektstatistik Umsatzsteuerstatistik 2000

Nr.	Merkmal	Untermerkmal	Untermerkmal
1	zufällig vergebene Nummer		
2	Amtlicher Gemeindeschlüssel		
3	Wirtschaftszweig		
4	Dauer der Steuerpflicht		
5	Organschaft 0=nein; 1=ja		
6	Rechtsform		
7	Lieferungen und Leistungen (LuL)		
8	Steuerpflichtige LuL		
9		zu 16%	
10		zu 7%	
11	Steuerfreie LuL		
12		mit Vorsteuerabzug	
13			innergemeinschaftl. LuL
14			weitere steuerfreie LuL
15		ohne Vorsteuerabzug	
16	Umsatzsteuer vor Abzug der Vorsteuer		
17		für LuL	
18		für innergemeinschaftl. Erwerbe	
19	Abziehbare Vorsteuer		
20		für LuL	
21			aus Rechnungen anderer Unternehmen
22			Einfuhrumsatzsteuer
23		für innergemeinschaftl. Erwerbe	
24	Vorauszahlungssoll		
25	Nachricht.: innergemeinschaftl. Erwerbe		
26	LuL 1999		
27	Vorauszahlungssoll 1999		

Tabelle 9.4: Ausprägungen

1	Beginn vor dem 1.1 des Vorjahres; Ende nicht vor dem 1.1. des Folgejahres
2	Beginn im Vorjahr; Ende nicht vom dem 1.1. des Folgejahres
3	Beginn im Berichtsjahr Ende nach dem 31.12 des Berichtsjahres
4	Beginn vor dem 1.1. des Berichtsjahres Ende im Berichtsjahr
7	Beginn im Berichtsjahr Ende im Berichtsjahr
8	wie 1 ohne Vorjahresumsatz

Tabelle 9.5: Rechtsformen

1	Einzelunternehmen
2	OHG
3	KG
4	AG
5	GmbH
6	Erwerbs- und Wirtschaftsgenossenschaften
7	Betriebe gewerblicher Art von Körperschaften des öffentlichen Rechts
8	sonstige Rechtsformen

9.3 Die Einzelhandelsstatistik

9.3.1 Ziele der Einzelhandelsstatistik

Ziel der Jahreserhebung im Handel ist die Beurteilung der Struktur und Entwicklung dieses Wirtschaftsbereichs und seiner wirtschaftlichen Bedeutung. Der Darstellung der Entwicklung des privaten Verbrauchs kommt in den Volkswirtschaftlichen Gesamtrechnungen bei der Verwendungsrechnung des Bruttoinlandsprodukts eine herausragende Bedeutung zu. Ergebnisse aller Bereiche des Handels und Gastgewerbes fließen in die Berechnung des Privaten Verbrauchs ein, vor allem die Ergebnisse der Einzelhandels- und Gastgewerbestatistik. Die Aufgliederung der Umsätze nach Produktarten erlaubt die Beobachtung der Entwicklung der Sortimentsstruktur im Fachhandel und in den nicht spezialisierten Unternehmen (z.B. Kaufhäuser, Warenhäuser, Versandhandelsunternehmen, Filialunternehmen mit breitem Sortiment). Die Erfassung des Warensortiments ist auch die einzige und verlässliche Grundlage für die Aktualisierung der Preisindizes im Handel. Im Rahmen der VGR, insbesondere für die Darstellung der gütermäßigen Verflechtung im Rahmen von Input-Output-Rechnungen und für die Darstellung von Marktverflechtungen ist eine möglichst tiefe Gliederung der Sortimente und der Wirtschaftszweige ebenfalls erforderlich. Entsprechendes gilt auch für die Zwecke der Entstehungsrechnung des Sozialprodukts, der Berechnung des Privaten Verbrauchs und der Ausrüstungsinvestitionen sowie für Untersuchungen des Verbraucherverhaltens. Die detaillierte Darstellung der Zusammensetzung des Warensortiments bietet auch den Unternehmen wichtige Anhaltspunkte für Zwecke der Marktanalyse und unternehmerische Dispositionen. Mit den in der EG-Struktur-Verordnung (EG, Euratom Nr. 58/97) geforderten Angaben über die Aufwendungen für Handelsware, Material und Dienstleistungen als wichtigstem Teil der Vorleistungen und den entsprechenden Bestandsveränderungen lässt sich die Nettoleistung berechnen. Ihre Erfassung ist für die Darstellung im Rahmen der VGR unentbehrlich. Die Bestandsveränderungen stellen ferner eine wesentliche Grundinformation für die Schätzung der Vorratsveränderung in der gesamten Volkswirtschaft dar. Darüber hinaus kommt ihnen unter betriebswirtschaftlichen Aspekten erhebliche Bedeutung zu. Sie werden zur Berechnung von Roherträgen, Handelsspannen und Warenumschlagskoeffizienten benötigt. Zu den Vorleistungen gehören auch Aufwendungen für gepachtete, gemietete und geleaste Sachanlagen.

9.3.2 Methode der Erhebung

Die jährliche Einzelhandelsstatistik erfasst als Stichprobe für das Jahr 1999 etwa 23500 Unternehmen. Die Befragung erfolgt überwiegend dezentral durch die Statistischen Landesämter. Die in der Stichprobe gewonnenen Ergebnisse werden auf die Gesamtheit der Unternehmen hochgerechnet.

Die Stichprobe ist dreifach geschichtet:

1.) Unterteilung der Grundgesamtheit nach Bundesländern,

2.) innerhalb jedes Bundeslandes Schichtung nach Branchengruppen,

3.) Schichtung innerhalb jeder so gebildeten Schicht nach Umsatzgrößenklassen.

Auf diese Weise ergeben sich für den gesamten Handel in Deutschland einige tausend Schichten. Aus jeder Schicht wird nach einem vorab festgelegten Auswahlplan eine Stichprobe gezogen. Die Schichtung ist so angelegt, dass sowohl für Bundesländer, als auch für Branchen, als auch für Umsatzgrößenklassen repräsentative Ergebnisse mit in etwa gleicher Genauigkeit ermittelt werden können.

Durch das komplizierte Schichtungsverfahren soll sichergestellt werden, dass die Stichprobe die Marktstruktur, die regionale Verteilung und die Konzentration in der Grundgesamtheit und deren Änderungen im Zeitablauf möglichst gut (d.h. unter Optimierung der resultierenden Standardfehler) abbildet.

Bei der dritten Schichtung ist die höchste Umsatzgrößenklasse so angelegt, dass sie die Totalschicht umfasst. Die Unternehmen dieser Schicht werden alle in die Stichprobe aufgenommen. Sie erhalten den Hochrechnungsfaktor $1,0$. Es handelt sich um die Großunternehmen der jeweiligen Branche, auf die in der Stichprobe wegen ihrer Bedeutung nicht verzichtet werden kann.

Die Unternehmen mit einem Umsatz unter der Totalschwelle werden mit einem Auswahlsatz gezogen (sog. Repräsentativschichten), wobei dieser Auswahlsatz um so kleiner wird, je mehr Unternehmen in einer Umsatzschicht vertreten sind; das ist der Fall in den Umsatzgrößenklassen mit geringem Umsatz. Der Hochrechnungsfaktor ist dann der Kehrwert des Auswahlsatzes. Beispiel: Bei einem Auswahlsatz von $0,25$ (d.h. es wird aus dieser Schicht jedes vierte Unternehmen in die Stichprobe aufgenommen) ergibt sich ein Hochrechnungsfaktor von 4 für alle Unternehmen dieser Schicht. In der untersten Umsatzgrößenklasse kann der Auswahlsatz auf ca. $0,02$ sinken und der Hochrechnungsfaktor somit auf ca. 50 steigen.

9.3.3 Inhalte der Erhebung

In der Jahreserhebung werden im Wesentlichen Angaben zu den in der EG-Struktur-Verordnung festgelegten Sachverhalten erhoben. Zu ihnen gehören auch so genannte zusammengefasste Merkmale wie Produktionswert, Bruttogewinnspanne bei Handelswaren, Bruttowertschöpfung zu Faktorkosten, Bruttobetriebsüberschuss. Diese werden jedoch nicht erhoben, sondern aus den anderen Merkmalen berechnet.

Zu den erfragten Merkmalen gehören auch Angaben zu den tätigen Personen und zu dem Personalaufwand, also Angaben für arbeitsmarkt- und beschäftigungspolitische Fragen. Tätige Personen schließen dabei die Selbständigen, Arbeitnehmer und Auszubildenden ein.

Angaben über die tätigen Personen werden zur Beurteilung der Personalkosten und für die Bildung wichtiger Beziehungszahlen (Produktivitätszahlen) benötigt. Die Unterteilungen nach der Stellung im Beruf und nach dem Geschlecht dienen der Beurteilung der Beschäftigungssituation und -entwicklung in sozioökonomischer Hinsicht, die Unterscheidung in Voll- und Teilzeitbeschäftigte der Beurteilung des Arbeitsvolumens. Die Frage nach den Bruttolöhnen und -gehältern gibt einerseits Aufschluss über die Höhe der Einkommen aus unselbständiger Tätigkeit, andererseits sind die Bruttolöhne und -gehälter im Handel und Gastgewerbe im Allgemeinen die wichtigsten Kostengrößen. Zusammen mit den gesetzlichen und übrigen Sozialaufwendungen der Arbeitgeber ergeben sie den Aufwand für den Personaleinsatz.

Aus den Angaben zu Umsatz und Vorleistungen lässt sich die Bruttowertschöpfung für die einzelnen Wirtschaftszweige und ihr Beitrag zum Bruttoinlandsprodukt ermitteln.

Mit dem Nachweis der „Umsätze nach Art der Tätigkeiten“ wird das Ziel verfolgt, die Differenzierung der Umsätze in funktionaler Gliederung (im Handel sind dies im Wesentlichen Umsätze aus Großhandel, aus Handelsvermittlung, aus Einzelhandel, aus Herstellung, aus Be- und Verarbeitung und aus anderen Tätigkeiten) zu zeigen. Die Darstellung dieses Sachverhaltes ist als Maßstab für die Betriebsleistung unentbehrlich und dient darüber hinaus auch der Beobachtung der Spezialisierungs- und Diversifikationstendenzen der Unternehmen. Nicht zuletzt liefert diese Gliederung auch Unterlagen zur Beurteilung der statistischen Zuordnung der Befragten zum Handel oder Gastgewerbe. Die Angaben dazu dienen zur Feststellung des Schwerpunktes des Unternehmens entsprechend seiner Wertschöpfung.

Bei den „sonstigen betrieblichen Erträgen“ handelt es sich um eine Restgröße für betriebsbedingte Erträge, die keinen anderen Erträgen zugeordnet werden, z.B. Honorare für Patente, Warenzeichen und Lizenzen. Sie werden zur Berechnung des Produktionswertes benötigt.

Schließlich werden als Merkmal auch die Investitionen (in vier Ausprägungen, siehe Variablen 26-29 in Tabelle 9.6) erfragt. Investitionen sind gesamtwirtschaftlich wichtige Aggregate, die das Wirtschaftswachstum und die Beschäftigungsentwicklung stark beeinflussen.

Viele der auskunftspflichtigen Unternehmen unterliegen einer Publizitätspflicht (z.B. nach §§ 325 ff. HGB). Aus diesem Grunde muss generell davon ausgegangen werden, dass einzelne Tatbestände aus anderen Quellen (z.B. Geschäftsberichten) bereits bekannt sind und daher auch zur Reidentifikation ganzer Datensätze verwendet werden können. Der endgültige Datensatz enthält die in Tabelle 9.6 aufgeführten Merkmale.

Tabelle 9.6: Merkmale des Datensatzes der Einzelhandelsstatistik

1.	Wirtschaftszweig (WZ93)
2.	Regionalbezug (Ost-West Merkmal)
3.	Gesamtumsatz
	Umsatzanteile in % aus
4.	Großhandel
5.	Einzelhandel, Reparatur von Gebrauchsgütern
6.	Sonstigen Dienstleistungstätigkeiten
7.	Herstellung, Verarbeitung, anderen industriellen Tätigkeiten oder aus Land- und Forstwirtschaft und Fischerei
8.	Sonstige betriebliche Erträge
	Einzelhandelsumsatz in % nach Absatzformen
9.	In Verkaufsräumen
10.	Aus Versandhandel
11.	An Verkaufsständen und auf Märkten
12.	Aus sonstigem Einzelhandel
13.	Anfangsbestand an Handelsware
14.	Endbestand an Handelsware
15.	Anfangsbestand an Roh-, Hilfs- und Betriebsstoffen
16.	Endbestand an Roh-, Hilfs- und Betriebsstoffen
17.	Anfangsbestand an selbsthergestellten und bearbeiteten Halb- und Fertigerzeugnissen
18.	Endbestand an selbsthergestellten und bearbeiteten Halb- und Fertigerzeugnissen
19.	Bezüge von Handelsware
20.	Bezüge von Roh-, Hilfs- und Betriebsstoffen
21.	Löhne und Gehälter
22.	Sozialabgaben
23.	Mieten und Pachten einschl. Kosten für Operate Leasing
24.	Betriebliche Steuern und Abgaben
25.	Bezogene Leistungen und andere betriebliche Aufwendungen
	Bruttoinvestitionen in
26.	Grundstücke
27.	Bestehende Gebäude
28.	Errichtung, Umbau und Erweiterung von Gebäuden
29.	Maschinen, Einrichtungen und Fahrzeuge
30.	Verkäufe von Sachanlagen
31.	Wert der im Geschäftsjahr über Finanzierungsleasing erworbenen Sachanlagen
32.	Zahl der rechtlich unselbständigen örtlichen Einheiten des Unternehmens am 31.12.
	Zahl der Beschäftigten am 30.9.
33.	Beschäftigte insgesamt
34.	Darunter Lohn- und Gehaltsempfänger
35.	Darunter Teilzeitbeschäftigte
36.	Hochrechnungsfaktor

9.4 Das IAB-Betriebspanel

Grundlage der Befragung des IAB-Betriebspanels sind die über die Betriebsnummer zum 30.6. eines Jahres aggregierten Angaben aus der Beschäftigtenstatistik. Somit sind in dem Panel nur Betriebe enthalten, die mindestens eine sozialversicherungspflichtige Person beschäftigen. Nicht berücksichtigt werden also beispielsweise Ein-Personen-Unternehmen und Unternehmen ohne sozialversicherungspflichtig Beschäftigte (z.B. mithelfende Familienangehörige) sowie Scheingründungen. Die Stichprobe wird nach dem Prinzip der optimalen Schichtung nach den Schichtungszellen der Betriebsgrößenklasse (10 Kategorien) und des Wirtschaftszweigs (16 Kategorien[9]) gezogen. Diese Schichtungszellen dienen auch der Gewichtung und Hochrechnung der Stichprobe. Die Befragung erfolgt durch Interviewer von TNS Infratest Sozialforschung. Für die erste Welle wurden im 3. Quartal 1993 in den alten Bundesländern 4.265 Betriebe befragt. Das Betriebspanel wird seitdem jährlich – seit 1996 zudem mit über 4.700 Betrieben in Ostdeutschland – durchgeführt. Die Antwortquote von wiederholt befragten Einheiten beträgt über 80%. Das Panel wird in jedem Jahr durch Ergänzungs- und Nachbearbeitungsstichproben flankiert, um neue oder wieder auflebende Betriebsnummern zu befragen und Ausfälle durch Verweigerungen oder nicht mehr vorhandene Betriebsnummern zu kompensieren. Das Fragenprogramm beinhaltet detaillierte Informationen über die Personalstruktur, -entwicklung und -politik der Betriebe. Aufgrund der Zusammenarbeit mit verschiedenen Länderministerien ist es seit dem Jahr 2001 möglich, eine Nettostichprobe mit mehr als 15.000 Betrieben zu erzielen, die auch regionalisierte Auswertungen auf Bundeslandsebene ermöglicht.[10]

An den Ziehungswahrscheinlichkeiten und Rücklaufquoten ändert sich im Verlauf der Jahre wenig. Da es aber wenig große Betriebe gibt, kann eine dort auftretende Panelmortalität kaum durch Ersatzbetriebe ausgeglichen werden. Sie werden in den späteren Wellen durch mittelgroße Betriebe ersetzt. Die Zahl der Betriebe ist daher im Verlauf des Panels angestiegen. Die Zahl der in den Betrieben beschäftigten Personen aber bleibt nahezu unverändert[11].

Im IAB-Betriebspanel stehen u.a. folgende Betriebsinformationen zur Verfügung:

- Beschäftigtenzahl (getrennt nach Qualifikationsgruppen)
- Anzahl befristet Beschäftigter und Leiharbeiter
- Wochenarbeitszeit für Vollzeitbeschäftigte und Überstunden

9) Ab dem Jahr 2000 erfolgt die Schichtung nach 20 Wirtschaftszweigen.

10) In Ostdeutschland ist das IAB-Betriebspanel bereits seit dem Jahr 1996 auf der Bundeslandebene repräsentativ.

11) Die Gewichtungsfaktoren des IAB-Betriebspanels passen sich über die Zeit den Veränderungen der Zahl der Betriebe und der Zahl der sozialversicherungspflichtig Beschäftigten an.

- Tarifbindung des Betriebs, Existenz eines Betriebsrats
- Umsatz, Vorleistungen und Exportanteil
- Investitionssumme
- Gesamtlohnsumme im Juni des Befragungsjahres
- technischer Stand der Betriebsanlagen
- Betriebsalter, Rechtsform und Unternehmensstellung
- Beurteilung der betrieblich-wirtschaftlichen Gesamtsituation
- Reorganisationsmaßnahmen und betriebliche Weiterbildungsaktivitäten (in mehrjährigen Abständen)
- Betriebsgröße und Wirtschaftszweig

Diese Informationen werden mit wenigen Ausnahmen jährlich erhoben. Ein detaillierter Überblick über das Fragenprogramm des IAB-Betriebspanels und über die Verfügbarkeit von Variablen in einzelnen Befragungsjahren kann der Online-Dokumentation zum IAB-Betriebspanel entnommen werden.[12]

Finanziert vom Bundesministerium für Arbeit stieg die Anzahl der Betriebe in Baden-Württemberg seit dem Jahr 2000 auf mehr als 1.200. Auf diese Weise wurde seitdem eine repräsentative Analyse für Baden-Württemberg ermöglicht.

Für die empirischen Auswertungen wird der Querschnitt für das Jahr 2002 verwendet. Auswertungen werden sowohl mit dem Bundesdatensatz als auch mit den Daten für Baden-Württemberg vorgenommen.[13] Aus der Vielzahl der Variablen wurde zunächst zur besseren Überschaubarkeit und weil beim IAB-Betriebspanel ausschließlich das Analysepotenzial untersucht werden sollte, eine Auswahl vorgenommen, die auf die inhaltlichen Auswertungen zugeschnitten ist.

12) Es ist auch ein Codebuch verfügbar, das detaillierte jährliche Häufigkeitsauszählungen für alle erhobenen Variablen beinhaltet.

13) Die Auswertungen für Baden-Württemberg wurden vorgeschaltet, weil die entsprechenden Daten im IAW selbst vorliegen.

Teil IV

Die Operationalisierung der Faktischen Anonymität

Eine der wesentlichen Aufgaben des Projekts bestand darin, den Begriff der faktischen Anonymität für Daten aus dem Bereich der Wirtschaftsstatistiken zu operationalisieren. Das hierbei gewählte Vorgehen wird in diesem Teil dargestellt. Zunächst wird in Kapitel 10 der Begriff der faktischen Anonymität erläutert. In Kapitel 11 werden anschließend die für die Messung der Datensicherheit relevanten Szenarien für mögliche Datenangriffe und das bei wirtschaftsstatistischen Einzeldaten zu beachtende Zusatzwissen dargestellt. Kapitel 12 beschreibt das auf diesen Grundlagen im Projekt entwickelte Konzept für die Beurteilung der Anonymität einer Mikrodatendatei. In Kapitel 13 werden die methodischen Grundlagen für die Simulation von Massenfischzügen zur Überprüfung der Datensicherheit dargestellt.

Kapitel 10

Der Begriff der faktischen Anonymität

Grundsätzlich dürfen die Statistischen Ämter Einzeldaten an Dritte nur weitergeben, wenn es für den Empfänger keine Möglichkeit gibt, die Identität eines Merkmalsträgers zu ermitteln (d.h. zu „deanonymisieren"). Diese strenge Formulierung, die sich im § 16 Abs. 1 des Bundesstatistikgesetzes (BStatG) findet, bedeutete für lange Zeit ein faktisches Verbot der Übermittlung von Unternehmensdaten, da es eigentlich nie auszuschließen ist, dass ein Merkmalsträger (z.B. ein Unternehmen) identifiziert wird. Da aber neben dem Schutz der „informationellen Selbstbestimmung" des Einzelnen auch die „Wissenschaftsfreiheit" einen gesellschaftlich hohen Stellenwert genießt, hat der Gesetzgeber bei der Überarbeitung des BStatG im Jahre 1987 durch den § 16 Abs. 6 (BStatG) eine Möglichkeit geschaffen, der Wissenschaft Einzeldaten zur Verfügung zu stellen, die nicht die beschriebenen hohen Ansprüche des § 16 Abs. 1 BStatG erfüllen müssen. Die Daten, die den geringeren Ansprüchen des § 16 Abs. 6 BStatG genügen, werden zur besseren Abgrenzung als „faktisch anonyme" Daten bezeichnet und dürfen ausschließlich wissenschaftlichen Einrichtungen mit der Aufgabe unabhängiger Forschung zu Analysezwecken übermittelt werden.

Ausgehend vom Wortlaut des zitierten § 16 Abs. 6 (BStatG) gilt eine Datei als faktisch anonym, wenn der potenzielle Datenangreifer aus rationalem Kalkül die Kosten der Deanonymisierung höher einschätzt als den Nutzen, den er aus einem erfolgreichen „Angriff" erwartet (Unverhältnismäßigkeitsgebot). Demnach entscheidet über Anonymität in diesem Sinne nicht eine technische Größe, sondern es wird auf Basis einer ökonomischen Kosten-Nutzen-Analyse entschieden, ob eine Datei als faktisch anonym gelten kann. Diese Vorgehensweise wurde bereits grundsätzlich im Projekt zur Anonymisierung von Personendaten verwendet (Müller et al. (1991) und Helmcke und Knoche (1992)).

Eine Datei kann ebenfalls als faktisch anonym gelten, wenn der potenzielle Angreifer die gewünschten Informationen aus alternativen Quellen kostengünstiger als durch eine Deanonymisierung der vertraulichen Daten beschaffen kann. Er wird auch in diesem Fall aus rationalem Kalkül heraus auf einen Deanonymisierungsversuch verzichten, da der Aufwand für ihn unverhältnismäßig hoch sein wird (vgl. Sturm (2002b), S. 107). Es könnte demnach

die Konstellation entstehen, dass für einen Angreifer nach Abwägung des Nutzens und der Kosten der Deanonymisierung zwar ein Nutzengewinn verbleibt, er aber den „Angriff“ unterlässt, da er auf einem alternativen Wege einen höheren Nutzengewinn generieren kann.

Nach dem Willen des Gesetzgebers kann und muss also die Enthüllung von faktisch anonymisierten Merkmalsträgern z.B. durch einen „Datenschutzidealisten“ nicht mit absoluter Sicherheit ausgeschlossen werden (dies gilt im Prinzip für alle Arten von anonymisierten Daten, vgl. z.B. Duncan und Lambert (1989), S. 207). Vielmehr kann dem Fall des „Datenschutzidealisten“ im Konzept der faktischen Anonymität keine Relevanz zugemessen werden. Er stellt vielmehr einen Sonderfall dar, bei dem davon ausgegangen wird, dass es dem Angreifer nicht um den Wert der von ihm erfolgreich enthüllten Information geht, sondern darum zu zeigen, dass die Deanonymisierung prinzipiell möglich ist. Ein Datenschutzidealist wird sehr viel höhere Kosten akzeptieren, weil ihm die Enthüllung als solche wichtig ist. Diese Kosten-Nutzen-Relation liegt außerhalb eines für das Konzept der „faktischen Anonymität“ sinnvoll anzunehmenden Rahmens und kann nicht betrachtet werden. Im Übrigen muss auch beachtet werden, dass die mit Strafandrohung bewehrten Regelungen des § 16 Abs. 6 BStatG auch für einen Datenschutzidealisten eine abschreckende Wirkung entfalten.

Aus der Begriffsklärung ergeben sich zwei im Projekt behandelte Fragestellungen bezüglich der Schutzwirkungen von Anonymisierungsmethoden:

- Inwieweit wird die Kosten-Nutzen-Relation des potenziellen Angreifers durch die Anonymisierung beeinflusst? Sei es, weil sich der Nutzen der gewonnenen Informationen durch die Anonymisierung verringert oder weil sich die Kosten der Deanonymisierung erhöhen (natürlich ist auch eine Kombination beider von Interesse). Beides führt für den Datenangreifer zu einer negativen Beeinflussung seiner Kosten-Nutzen-Relation, wodurch ein „Angriff“ unwahrscheinlicher wird.
- Welchen Einfluss haben alternative Quellen der Informationsbeschaffung (anstelle eines Deanonymisierungsversuchs) auf das „notwendige Anonymisierungsniveau“? Je kostengünstiger diese Quellen für einen Datenangreifer zu erschließen sind, desto unwahrscheinlicher ist ein Datenangriff, und desto geringer ist das notwendige Anonymisierungsniveau.

Kapitel 11

Angriffsszenarien und Zusatzwissen

11.1 Definitionen und Bezeichnungen

Im Zusammenhang mit Deanonymisierungsversuchen werden Informationen, die einem Datenangreifer zusätzlich zum anonymisierten Datensatz vorliegen oder beschaffbar sind und die zur Deanonymisierung eingesetzt werden können, als Zusatzwissen bezeichnet. Das Zusatzwissen zeichnet sich dadurch aus, dass darin die direkten Identifikatoren der Merkmalsträger enthalten sind. Diejenigen Merkmale, die sowohl im Zusatzwissen als auch in den Zieldaten (bzw. den geheimhaltungspflichtigen Daten) enthalten sind, werden als Schlüssel- oder Überschneidungsmerkmale bezeichnet (siehe Abbildung 11.1 und vgl. z.B. Höhne et al. (2003)). Unter einer Reidentifikation/Deanonymisierung wird schließlich verstanden, ein bestimmtes Zusatzwissen mit Hilfe der Überschneidungsmerkmale mit dem Zieldatensatz eindeutig und richtig zu verknüpfen. Ein Zuordnungsversuch kann i.d.R. nur mittels der Überschneidungsmerkmale vorgenommen werden. Die Reidentifikation (richtige und eindeutige Zuordnung eines gesuchten Merkmalsträgers) eröffnet einem Datenangreifer dabei Kenntnisse über sensible Sachverhalte von ihn interessierenden Merkmalsträgern (Zielmerkmale).

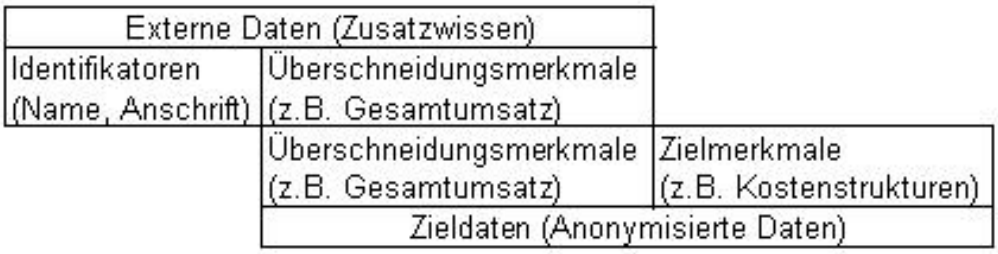

Abbildung 11.1: Überschneidungsmerkmale

11.2 Angriffsszenarien

Zu einer Einschätzung des Reidentifikationsrisikos von Merkmalsträgern bzw. des Enthüllungsrisikos für einzelne Werte in den Zieldaten können Simulationen verschiedener Datenangriffsszenarien dienen (vgl. Elliot und Dale (1999), ausführlich zu Angriffsszenarien bei wirtschaftsstatistischen Einzeldaten vgl. Wirth (2003)). Die relevantesten Szenarien sind der so genannte Einzelangriff und der Massenfischzug. Um das Reidentifikationsrisiko für einen Merkmalsträger u in den Zieldaten vernünftig beurteilen zu können, müssen beide Szenarien ihre Anwendung finden. Der Schätzer $R(u)$ für das Reidentifikationsrisiko ergibt sich dann als Maximum der mit dem Einzelangriff und Massenfischzug verbundenen Risiken $R_E(u)$ und $R_M(u)$, d.h.

$$R(u) := \max\{R_E(u), R_M(u)\}.$$

Bei einem Einzelangriff versucht der Datenangreifer, eine oder mehrere Informationen über einen bestimmten Merkmalsträger zu enthüllen. Dem Datenangreifer wird bei einem Einzelangriff ein spezielles Zusatzwissen und die Kenntnis der Teilnahme des interessierenden Unternehmens an der Erhebung (vertrauliche Zieldaten) unterstellt. Weitere Unternehmensinformationen kann er über kommerzielle Datenbanken und allgemein verfügbare Quellen wie z.B. die Geschäftsberichte der Unternehmen sammeln.

Bei einem Massenfischzug hingegen versucht der Datenangreifer, möglichst viele Merkmalsträger einer externen Datenbank (z.B. einer kommerziell erhältlichen Unternehmensdatenbank) den Zieldaten korrekt zuzuordnen, um seiner externen Datenbank weitere Infomationen zuzuspielen.

Das Risiko der Reidentifikation eines Merkmalsträgers bzw. der Enthüllung eines Einzelwertes dieses Merkmalsträgers hängt sehr stark von der Besetzung relevanter Teilmassen der Gesamtdatei, die diesen Merkmalsträger enthalten, ab. Befindet sich ein Unternehmen etwa in einem sehr dünn besetzten Wirtschaftszweig und/oder in einer oberen Beschäftigtengrößenklasse, so ist eine Reidentifikation bzw. Enthüllung durch den Datenangreifer wahrscheinlicher als im allgemeinen Falle. Hier ist die Wahrscheinlichkeit einer Reidentifikation mittels eines Einzelangriffes höher als mittels eines Massenfischzuges einzustufen, da ein potenzieller Einzelangreifer die möglichen Kandidaten für eine richtige Zuordnung leichter überblicken und weitere individuelle Kenntnisse über das gesuchte Unternehmen zum Vergleich einbringen kann. Der Aufwand für einen (einzigen) Einzelangriff ist aus Sicht des Datenangreifers in der Regel überschaubar. Allerdings kann von Merkmalsträger zu Merkmalsträger die Menge der Überschneidungsmerkmale und die damit verbundene Recherchearbeit stark variieren. Auf der anderen Seite ist in sehr dicht besetzten Teilmassen der Massenfischzug dem Einzelangriff überlegen, da bei einer Vielzahl von ähnlichen Kandidaten durch kompliziertere Strukturvergleiche und Distanzmaße auch feinere, mit dem bloßen Auge kaum sichtbare Unterschiede transparent werden können.

Der Aufwand für einen Massenfischzug ist aus Sicht des Datenangreifers sehr hoch einzustufen, da ihm neben den Anschaffungskosten für einen leistungsfähigen Rechner und eine

qualitativ hochwertige kommerzielle Unternehmensdatenbank vor allem die Kosten für die Entwicklung einer Simulationssoftware entstehen.

Es empfiehlt sich in der Praxis für den Datenanbieter, zur Abschätzung des mit den vertraulichen Zieldaten verbundenen Reidentifikations- bzw. Enthüllungsrisikos zunächst die Ergebnisse von Massenfischzugsimulationen heranzuziehen. Auf diese Weise wird die globale Wirkung der Anonymisierung auf die gesamte Zieldatei sichtbar. Der Aufwand ist seitens des Datenanbieters bei vorhandenem Zusatzwissen und vorhandener Simulationssoftware im Wesentlichen durch die i.d.R. überschaubare Rechenzeit begrenzt. In einem zweiten Schritt sollten für besonders reidentifikationsgefährdete Bereiche in den Zieldaten (wie z.B. die Klasse der Großunternehmen oder dünn besetzte Wirtschaftszweige), welche den Fachleuten oftmals bereits vor der Anwendung von Anonymisierungsmaßnahmen bekannt sind, Einzelangriffe durchgeführt werden. Dies gilt auch für Bereiche, in denen der zuvor simulierte Massenfischzug möglicherweise bereits hohe Reidentifikationsrisiken aufgedeckt hat, diese aber nur eine Untergrenze für das tatsächliche, mit einem Einzelangriff besser abschätzbare Risiko darstellen. Hierbei ist zu beachten, dass aus Sicht des Datenanbieters die Simulation zahlreicher Einzelangriffe sehr zeitaufwändig werden kann.

11.3 Struktur des Zusatzwissens

Das Zusatzwissen, welches einem potenziellen Datenangreifer zur Verfügung stehen kann, ist ständigen Veränderungen durch die Umwelt unterworfen. So hat z.B. das Internet zu völlig neuartigen Möglichkeiten geführt, Wissen über einen Merkmalsträger zu generieren, das dann als Zusatzwissen verwendbar ist. Eine Analyse des Zusatzwissens ist daher immer eine Momentaufnahme, so dass ein Datenanbieter bei der Entwicklung eines Scientific-Use-Files eine solche Analyse für seinen speziellen Fall der Risikoabschätzung jedesmal vorschalten muss. Bei der zeitlichen Variabilität des Zusatzwissens kann dieses insoweit abgeschätzt werden, dass zwar die konkreten Inhalte sehr großen Veränderungen unterworfen sind, sich die grundlegende Struktur aber nur geringfügig verändert. Daher versucht die folgende Analyse, solche zeitlich relativ stabilen Strukturmerkmale aufzuzeigen. Dies kann bei einer aktuellen Erstellung eines Scientific-Use-Files als Hilfestellung dafür dienen, das jeweilige Zusatzwissen zu analysieren. Die folgenden konkreten Beispiele von Zusatzwissen sind aus genannten Gründen als Momentaufnahme der Analyse innerhalb des Projekts zu sehen. Sie illustrieren, wie künftig die Abschätzung des jeweils relevanten Zusatzwissens erfolgen kann.

Zusatzwissen kann danach unterschieden werden, ob es ein Datenangreifer aus kommerziellen Datenbanken, nichtkommerziellen (öffentlichen) Informationsquellen (Telefonbüchern, Handelsregistern, Internet usw.) oder aus persönlichen Quellen bezieht (vgl. Elliot und Dale (1999, S. 7)).

11.3.1 Kommerzielle Unternehmensdatenbanken

Unter den kommerziellen Unternehmensdatenbanken werden diejenigen eingruppiert, die gegen Entgelt einem Nutzer Informationen über ein bestimmtes oder eine Auswahl von Unternehmen bereitstellen. Als Beispiele sind hier die Datenbanken von Hoppenstedt oder die MARKUS-Datenbank zu nennen. Ein Nutzer kann zwischen Datenbanken für bestimmte Branchen und allgemeinen Unternehmensdatenbanken wählen. So kann er z.B. mit Hilfe der M&M-Handelsdatenbank Zusatzwissen über Handelsunternehmen und mit der MARKUS-Datenbank Informationen über Unternehmen aller Branchen generieren.

Wichtigste Überschneidungsmerkmale, die in fast allen kommerziellen Datenbanken und in den meisten zu schützenden Wirtschaftsstatistiken zu finden sind, sind die Anzahl der Beschäftigten, der Umsatz, die Branchenzugehörigkeit und die regionale Eingliederung eines Unternehmens. Ist, wie z.B. in der Umsatzsteuerstatistik, die Rechtsform des Unternehmens in den Zieldaten enthalten, so muss sie als Überschneidungsmerkmal eingestuft werden, da sie ebenfalls in kommerziellen Datenbanken zu finden ist.

Die Vorteile kommerzieller Unternehmensdatenbanken liegen in der hohen Anzahl von erfassten Unternehmen und der klaren Struktur der enthaltenen Informationen. Daher ermöglichen diese Datenbanken prinzipiell, wie in späteren Abschnitten gezeigt wird, einen Massenfischzug. Allerdings muss für eine komplette Datei mit erheblichen Kosten gerechnet werden, wohingegen die Informationen einzelner Unternehmen relativ kostengünstig zu erhalten sind. Letzteres gilt im besonderem Maße, da man bei einer Unternehmensrecherche auch die Recherchezeit und das Risiko, keine Informationen über das gesuchte Unternehmen zu erhalten, berücksichtigen muss. Diese Kosten sind bei den kommerziellen Datenbanken im Verhältnis zu anderen Wegen des Zusatzwissens relativ gering.

Als kommerzielle Anbieter von Daten sind die Datenbankanbieter auf eine hohe Datenqualität angewiesen. Daher werden von ihnen Maßnahmen zur Kontrolle und zur Verbesserung der Datenqualität ergriffen, was sich positiv auf die Eignung der Datenbank als Zusatzwissen auswirkt. Die in den Datenbanken verfügbaren Überschneidungsmerkmale können qualitativ in zwei Gruppen eingeteilt werden (zur Qualität von Überschneidungsmerkmalen vgl. Elliot und Dale (1999, S. 9-10)):

- Merkmale mit vielen Ausprägungen, die einen Datenbestand sehr stark differenzieren, die aber im Zeitablauf eine hohe Variabilität aufweisen. Dies sind z.B. Umsatz und Beschäftigte.
- Merkmale mit wenigen Ausprägungen, die einen Datenbestand weniger stark differenzieren, die jedoch im Zeitablauf nur eine geringe Variabilität aufweisen. Hierzu gehören Regionalkennung, Branchenzugehörigkeit und Rechtsform.

Da der Grundsatz gilt, dass, je mehr Ausprägungen ein Überschneidungsmerkmal annehmen kann, desto mehr es zur Identifikation beitragen kann (vgl. Fürnrohr et al. (2002,

S. 310)), sind für die Reidentifikation zunächst Merkmale wie Beschäftigte und Umsatz interessant. Allerdings können aufgrund der hohen zeitlichen Variabilität die Ausprägungen im Zusatzwissen schnell von denen in den Zieldaten abweichen. Dadurch steigt die Gefahr von fehlerhaften Zuordnungen (vgl. Fürnrohr et al. (2002)). Um diese Gefahr zu minimieren, sollte die kommerzielle Datenbank den gleichen Zeitraum abbilden wie die zu deanonymisierende Datei. Die Datenbankanbieter legen sehr großen Wert auf Aktualität und weniger auf eine historische Führung ihrer Daten. Dies senkt die Qualität der Merkmale wie Umsatz und Beschäftigte als Überschneidungsmerkmale, zumindest dann, wenn der Erhebungszeitraum des veröffentlichten Zieldatensatzes einige Jahre zurückliegt.

Merkmale wie Regionalkennung, Branche und Rechtsform eines Unternehmens sind im Zeitablauf zwar relativ stabil und nur gering fehlerbehaftet, dafür werden aber die Unternehmen weniger stark differenziert als bei Merkmalen wie Umsatz und Beschäftigte. Eindeutige Zuordnungen aufgrund dieser Überschneidungsmerkmale werden die seltene Ausnahme sein, für einen Datenangreifer bieten sie aber eine sichere Möglichkeit, die Unternehmen in einem anonymisierten Datensatz zu kategorisieren und auf dieser Basis den Datenangriff weiterzuführen.

Abschließend zu diesem Unterabschnitt sind in Tabelle 11.1 einige bekannte kommerzielle Unternehmensdatenbanken mit ihren Eigenschaften aufgelistet. Die Tabelle erhebt dabei nicht den Anspruch auf Vollständigkeit. Als Merkmale, die je nach Zieldaten als Überschneidungsmerkmale in Frage kommen, sind die Anzahl der Beschäftigten, die Höhe des Umsatzes, die Rechtsform und der Regionalbezug zu nennen.

Tabelle 11.1: Auswahl an Unternehmensdatenbanken

Name	Schwerpunkt	Anzahl Unternehmen	Preis
Hoppenstedt Top 7000	7.000 größten Unternehmen	7.000	13.500 €+ MWSt Teildatensätze zu niedrigeren Preisen erhältlich
Hoppenstedt Firmen-datenbank	allgemeine Unternehmens-datenbank	ca. 152.000	1.500 €+ MWSt Prepaid möglich z.B. 20 Profile für 155 €
M&M deutsche Handels-datenbank	Handels-unternehmen (Top-Firmen)	ca. 300	275 €
BvD MARKUS	allgemeine Unternehmens-datenbank	ca. 850.000 aus Deutschland und Öster-reich	VhB Pay per view Mindestumsatz: 5.000 Dollar
Zollner Unternehmens-profile	allgemeine Unternehmens-datenbank	ca. 70.000	k.A.

11.3.2 Nichtkommerzielle Informationsquellen

Mit nichtkommerziellen öffentlichen Informationsquellen sind Quellen gemeint, die allgemein zugänglich sind und deren Zweck nicht die kommerzielle Übermittlung von Informationen über Unternehmen ist. Die Übermittlung der Information dient bei diesen Informationsquellen einem Nebenzweck. Bei einem Telefonbuch verfolgt die Telekom nicht das Ziel, eine Telefonnummer zu verkaufen, sondern die Möglichkeit zu bieten, jemanden anzurufen. Die IHK-Datenbanken haben, ebenso wie die Unternehmenspräsentationen im Internet, nicht zum Ziel, Informationen von Unternehmen zu verkaufen, sondern dienen zur Anbahnung von Geschäftsbeziehungen. So weist die IHK Osnabrück-Emsland auf Ihrer Internetseite darauf hin, „dass die (...) veröffentlichten Daten (...) nur zur Förderung von Geschäftsabschlüssen und zu anderen dem Wirtschaftsverkehr dienenden Zwecken benutzt werden dürfen“. Abgeleitet hiervon sind nichtkommerzielle Informationsquellen als solche definiert, für deren Abruf der Nutzer kein Entgeld bezahlen muss. Zu nennen sind hierbei in erster Linie Datenbanken von Verbänden und Industrie- und Handelskammern (IHK), sowie die Internetauftritte der Unternehmen. Am Beispiel der IHK-Datenbanken sollen im Folgenden diese Informationsquellen bewertet werden.[14]

Neben frei zugänglichen bieten die IHKs auch kostenpflichtige Datenbanken an. Diese werden zu den kommerziellen Datenbanken gezählt und daher in diesem Kontext nicht betrachtet. Der wichtigste Vorteil ist die (kosten)freie Verfügbarkeit der Datenbanken über das Internet. Daneben ist weiterhin positiv, dass sie auch Kleinstunternehmen enthalten. Diesen Vorteilen stehen aber schwerwiegende Nachteile gegenüber. So werden nicht von allen IHKs solche Datenbanken angeboten. Diejenigen, die eine anbieten, bieten i.d.R. lediglich eine Suchmöglichkeit innerhalb des eigenen Kammerbezirks. Eine Ausnahme bilden die IHKs in Baden Württemberg, die eine Suche innerhalb des eigenen Bundeslandes ermöglichen. Durch diese Einschränkung wird eine bundesweit kostenfreie Suche über die IHK-Datenbanken erschwert. Der Nachteil der regional beschränkten Suche wirkt allerdings nur, wenn ein Datenangreifer nicht gezielt nach einem Unternehmen sucht. Sucht er dagegen nach einem bestimmen Unternehmen, kann er in der dazugehörigen Kammer gezielt nach Informationen über das Unternehmen suchen.

Ein weiterer Nachteil für einen potenziellen Datenangreifer ist, dass die Datenbanken der einzelnen Kammern unterschiedlich aufgebaut sind. Dies erschwert die Suche in verschiedenen Datenbanken. Allerdings sind Harmonisierungsbestrebungen zu beobachten, durch welche die einzelnen Datenbanken in ein einheitliches Format überführt werden sollen. Hauptproblem der IHK-Datenbanken ist, dass die Einträge freiwilliger Natur sind. Dadurch ist völlig unklar, wie hoch der Anteil der Unternehmen ist, die Einträge in die Datenbanken tätigen und wie gesichert deren Datenqualität ist (d.h. ob und mit welcher Genauigkeit

14) Bei den IHK-Datenbanken handelt es sich nicht um eine einzige, zentral vorliegende Unternehmensdatenbank. Vielmehr bieten die einzelnen Kammern jeweils eigene Datenbanken für ihren Kammerbezirk an. Daher wird im Folgenden von den IHK-Datenbanken gesprochen, womit die Gesamheit der einzelnen regionalen Datenbanken verstanden wird.

die Unternehmen die angeforderten Merkmale angeben). Die Angabe der IHKs, wie viele Unternehmenseinträge eine Datenbank aufweist, ist ein Indiz, mit dem die erste Frage näherungsweise beantwortet werden kann, nicht jedoch die zweite.

Die Überschneidungsmerkmale der IHK-Datenbanken zu den amtlichen Erhebungen sind der Regionalbezug und die Branchenzugehörigkeit des Unternehmens. Das Merkmal „Anzahl der Beschäftigten" ist in vielen IHK-Datenbanken vorhanden, die Höhe des Umsatzes dagegen nur in wenigen. Durch die Freiwilligkeit der Einträge ist nicht gesichert, dass alle Merkmale von den Unternehmen ausgefüllt werden, auch wenn diese abgefragt werden. Ebenso ist die Qualität der Daten fragwürdig; dies gilt selbst bei dem als stabil eingeschätzten Merkmal Branchenzugehörigkeit.

Aufgrund der erwähnten Nachteile erscheinen diese Datenbanken nicht für einen Massenfischzug geeignet. Die Informationen sind zwar kostenlos, jedoch nicht für alle Unternehmen erhältlich und zudem auf die einzelnen Kammern verstreut und freiwillig. Die Datenqualität ist eher fraglich. Dies alles erschwert für einen Datenangreifer den Aufbau einer größeren Abgleichdatei. Eignen könnten sich diese Datenbanken für einen Datenangreifer im Rahmen eines Einzelangriffsszenarios, um sich erste Informationen über das gesuchte Unternehmen zu beschaffen. Da die Abrufe kostenlos sind, muss er nur die Zeit der Suche investieren. Im Rahmen eines Einzelangriffsszenarios wurde in den IHK-Datenbanken nach Zusatzwissen recherchiert (zu den Ergebnissen vgl. Vorgrimler (2002)). Dabei ergaben sich die erwarteten Probleme. In den meisten Fällen war das gesuchte Unternehmen nicht in der Datenbank auffindbar. Bei denjenigen Unternehmen, die in einer der Datenbanken enthalten waren, war die Datenqualität nur selten befriedigend. Aus diesen Gründen waren die IHK-Datenbanken nur in wenigen Fällen zur Reidentifikation eines Unternehmens geeignet.

Das Internet, das bisher als Medium zur Verwendung der IHK-Datenbanken bereits in die Analyse mit eingeflossen ist, kann auch ganz allgemein als nichtkommerzielle Informationsquelle aufgefasst werden. Zu den relevanten Informationsquellen, die über diesen Weg erschlossen werden können, zählen neben den bereits beschriebenen IHK-Datenbanken vor allem die Internetauftritte der gesuchten Unternehmen, Nachrichtenportale und vieles weitere mehr. Mit Hilfe des Internets könnnen Datenangreifer Zusatzwissen generieren, welches sie zu einem Reidentifikationsversuch verwenden können. Neben dem Internet stehen hierzu noch Fachzeitschriften, Tagespresse, Fernsehen und weitere denkbare Medien zur Verfügung.

Das Beispiel der IHK-Datenbanken zeigt die Schwierigkeiten auf, die mit der Generierung des Zusatzwissens mittels nichtkommerzieller Informationsquellen verbunden sind. Da diese Probleme (besonders die der Datenqualität aber auch der Quantität bezüglich verschiedener Unternehmen) auch bei anderen nichtkommerziellen Quellen auftreten, sind solche Quellen höchstens im Rahmen von Einzelangriffen verwendbar. Um einen Einblick in die Möglichkeiten der eigenen Recherche mittels Internet zu erhalten, wurde im Rahmen eines Einzelangriffsszenarios versucht, über eine bestimmte Anzahl von Unternehmen soviel Zusatzwissen wie möglich über eigene Recherchen zu generieren. Die Ergebnisse verschie-

dener Reidentifikationsversuche haben gezeigt, dass eine Reidentifikation allein mit dem Internet als Quelle für Zusatzwissen prinzipiell möglich ist. Sie haben aber auch gezeigt, dass dies in erster Linie nur für größere Unternehmen gilt. Für kleinere Unternehmen war das zu findende Zusatzwissen zu ungenau, als dass eine erfolgreiche Reidentifikation möglich gewesen wäre. Darüber hinaus entstehen auch bei einer Generierung des Zusatzwissens in nichtkommerziellen Informationsquellen für einen Angreifer Kosten. Diese drücken sich in erster Linie im Zeitaufwand der Suche und dem Risiko, keine Informationen über das gesuchte Unternehmen zu erhalten, aus. Die Kosten sind dabei sehr variabel und kaum abschätzbar. Sie hängen zum einen vom gesuchten Merkmalsträger und zum anderen vom Datenangreifer selbst ab. Für ein sehr großes Unternehmen kann die Suche nach geeignetem Zusatzwissen bereits nach wenigen Minuten mit sehr geringen Kosten erfolgreich beendet sein. Für kleine Unternehmen kann die Suche dagegen sehr zeit- und kostenintesiv sein. Diese Unterschiede sind in der Anonymisierungsstrategie zu berücksichtigen.

11.3.3 Persönliche Informationsquellen

Die persönlichen Informationsquellen sind am schwierigsten abgrenzbar. Dazu zählen sowohl persönliche Erfahrungen als auch zufälliges Wissen. Beides kann auch als Expertenwissen bezeichnet werden.

Das persönliche Wissen steht einem Datenangreifer quasi zum Nulltarif zur Verfügung. Er hat es bereits in der Vergangeheit zu einem anderen Zweck generiert und kann es nun wieder für den Nebenzweck als Zusatzwissen verwenden. Hierzu können private und berufliche Erfahrungen zählen. Ein Experte der Automobilbranche hat beruflich über diesen Markt und seinen Unternehmen einen hohen Wissensstand und kann diesen eventuell als Zusatzwissen bei einen Datenangriff einsetzen.

Da nicht davon ausgegangen werden kann, dass eine oder mehrere Personen Informationen über eine Vielzahl von Unternehmen besitzen, die ausreichend sind, um einen Massenfischzug durchzuführen, scheidet dieser auf dieser Basis aus. Anders ist jedoch die Möglichkeit für Einzelangriffe zu beurteilen. Hier können die persönlichen Informationsquellen bei einem Datenangriff sogar die entscheidende Rolle spielen. Ein Angreifer hat vielleicht bereits vorab viele Informationen über ein Unternehmen und kann diese durch gezielte Recherchen relativ kostengünstig vergrößern und so genügend Wissen sammeln, um ein Reidentifikationsversuch erfolgreich durchführen zu können.

Die verschiedenen denkbaren Datenangreifer bringen jeweils unterschiedliches Expertenwissen mit. So hat, wie bereits erwähnt, ein Marktexperte in seinem speziellen Bereich bereits vorab eine hohe Kenntnis über die betreffenden Unternehmen, während andere wesentlich weniger Vorabwissen mitbringen. Diese Problematik ist nicht mit einer geeigneten Anonymisierungsstrategie zu lösen. In diesem Zusammenhang muss auf die Rahmenbedingungen der Weitergabe faktisch anonymer Daten und auf die gesetzlichen Bestimmungen nach §16 Abs. 6 BStatG hingewiesen werden. Der Kundenkreis ist gesetzlich auf wissenschaftli-

che Forschungsinstitute und Universitäten beschränkt und zwar auf die Einrichtungen und nicht auf Personen. Die Personen, die letztlich die Daten erhalten, beschränken sich auf Mitarbeiter dieser Einrichtungen. Sie sind den statistischen Ämtern über die vertraglichen Verpflichtungen, welche die Institutionen eingehen, bekannt. Die statistischen Ämter können daher einerseits bereits vorab prüfen, inwieweit es eventuell zu Interessenkonflikten bei denjenigen kommen kann, die einen Scientific-Use-File erhalten. Die Wissenschaftler, die die Einzeldaten in faktisch anonymer Form erhalten, müssen sich andererseits selbst vertraglich verpflichten, dass ein eventueller Interessenkonflikt nicht im Widerspruch zur faktischen Anonymität der Daten steht. Die rechtlichen Rahmenbedingungen und vertraglichen Ausgestaltungen müssen das Problem des Expertenwissens lösen. Nur auf diese Weise kann eine sinnvolle Anonymisierung der Daten gewährleistet werden. Würde man auch das Expertenwissen mit in die Anonymisierungsüberlegungen einbeziehen, würde die daraus folgende Anonymisierung zu einem sehr starken Eingriff in die Daten führen müssen.

Zusammenfassend zeigt sich, dass sich kommerzielle Datenbanken sowohl für Einzelangriffe als auch für Massenfischzüge eignen. Ein großer Anteil der Reidentifikationsgefahr ist im Rahmen von Einzelangriffen auch auf das individuelle Zusatzwissen des Datenangreifers zurück zu führen.

Kapitel 12

Das Konzept der Schutzwirkung

12.1 Korrektheit von Zuordnungsversuchen

Ein Schutzeffekt von Anonymisierungsmaßnahmen besteht darin, die Verwendbarkeit der Überschneidungsmerkmale im Zieldatensatz für eine Zuordnung zu stören. Ein Zuordnungsversuch verursacht für einen Datenangreifer einen Aufwand und damit Kosten. Es ist wesentlich, sich vor Augen zu halten, dass ein Datenangreifer vor dem Problem steht, korrekte von falschen Zuordnungen zu unterscheiden. Da er die Richtigkeit seiner vorgenommenen Zuordnungen nicht prüfen kann, wird ein rationaler Datenangreifer die Unsicherheit, ob eine von ihm angenommene Zuordnung zutrifft, bei der Bewertung seines Nutzens berücksichtigen.

Aufgrund seines Rationalkalküls wird ein Datenangreifer von dem Versuch einer Reidentifikation absehen, wenn er davon ausgehen muss, dass der Anteil der falsch zuordenbaren an allen zuordenbaren Datensätzen über einer kritischen Schwelle liegt, also sein Zuordnungsrisiko zu hoch ist.

12.2 Qualität enthüllter Informationen

Eine erfolgreiche Zuordnung führt nicht unbedingt zu einem Nutzen für den Datenangreifer und damit zu einer Verletzung der faktischen Anonymität. Vielmehr ergibt sich sein Nutzen erst aus den „brauchbaren“ Informationen, die er bei einer erfolgreichen Reidentifikation gewinnen kann. Im Gegensatz zu Wissenschaftlern, für deren Analyse nicht unbedingt die Richtigkeit der Einzelwerte einer Datei, sondern die aus den Einzelwerten errechneten Größen entscheidend sind, hängt der Nutzen für Datenangreifer von der Qualität konkreter Einzelwerte ab.

Im Folgenden werden Unsicherheiten untersucht, die auf der Ebene der einzelnen Merkmale

einer Datei bestehen. Wird durch einen Datenangriff ein Datensatz erfolgreich zugeordnet (identifiziert), so gewinnt (enthüllt) der Datenangreifer auf der Merkmalsebene Informationen über alle Zielmerkmale im Datensatz. Daneben liefert eine Reidentifikation auch neue Informationen über die Ausprägungen der Überschneidungsmerkmale, die vor allem bei metrischen Größen von den jeweiligen Ausprägungen im Zusatzwissen abweichen können. Der Datenangreifer muss für diese Merkmale selbst entscheiden, welche Werte er für die zuverlässigeren hält. Wird im Folgenden daher von richtig zugeordneten Informationen gesprochen, so sind alle Merkmale im jeweils reidentifizierten Datensatz gemeint.

Die Brauchbarkeit einer Information ist im Falle einer richtigen Zuordnung nicht von vornherein gesichert, da die Merkmalsausprägungen beim Einsatz datenverändernder Verfahren vom Originalwert abweichen können. Dies spielt für die Wahrung der faktischen Anonymität eine wichtige Rolle. Brauchbar ist eine Information nur dann, wenn der gefundene Wert dem „wahren“ Wert[15] entspricht oder diesem in einem bestimmten Maß ähnelt. Ab einer gewissen Abweichung wird ein Datenangreifer keinen Nutzen mehr aus einer richtig zugeordneten Information erzielen können. Die Abweichungsgrenze, ab der ein Wert als unbrauchbar gilt, wird im Folgenden „Nützlichkeitsschwelle“ genannt. Die jeweiligen Datenanbieter (wie z.B. statistische Ämter) müssen vernünftige Werte für diese Schwellen festlegen, mit deren Hilfe dann das Risiko eines Datenangreifers geschätzt werden kann. Im Rahmen der Anonymisierung von Tabellen wurden vergleichbare Schwellen bereits festgelegt.

Auch an dieser Stelle ist wieder wesentlich zu beachten: Ob eine korrekte Reidentifikation brauchbare Informationen liefert, also die Abweichung des gefundenen Wertes zu dem entsprechenden Originalwert innerhalb einer bestimmten Umgebung liegt, kann von einem Datenangreifer nicht geprüft werden. Er kann allenfalls eine Wahrscheinlichkeit für die Brauchbarkeit der gefundenen Informationen abschätzen. Liegt diese Wahrscheinlichkeit unterhalb einer gewissen Schwelle, wird das Risiko, eine unbrauchbare Information mit einer brauchbaren Information zu verwechseln, als zu hoch eingeschätzt werden. Daraus folgt, dass für einen Datenangreifer eine Information wertlos werden kann, unabhängig davon, ob die gefundene Information innerhalb oder außerhalb einer bestimmten Abweichungsumgebung liegt. Datensätze können daher auch als faktisch anonym gelten, selbst wenn vereinzelt durch Reidentifikation vereinzelt brauchbare Informationen zugeordnet werden können.

12.3 Zusammenführung zu einem Maß für faktische Anonymität

In die nachfolgende Definition eines Maßes für die faktische Anonymität fließen nun die beiden in den vorherigen Abschnitten 12.1 und 12.2 besprochenen Elemente ein. Nach einer durchgeführten Simulation eines Datenangriffes wird zum einen der Anteil richtiger

15) Als „wahrer“ Wert wird hier die Ausprägung in den Originaldaten verstanden.

Zuordnungen und zum anderen der Anteil nützlicher Informationen innerhalb der richtigen Zuordnungen berechnet. Auch eine richtige Zuordnung eines Merkmalsträgers M kann einen für den potenziellen Datenangreifer erfolglosen Angriffsversuch darstellen, nämlich wenn der ihn interessierende Einzelwert (bzw. die ihn interessierende Information) relativ um wenigstens γ von dem tatsächlichen Originalwert abweicht (in den Experimenten in Teil XI wurden $\gamma = 0,05$ und $\gamma = 0,1$ gewählt). Nun werden beide Elemente zu einem Gesamtmaß zusammengeführt: Die Wahrscheinlichkeit, dass ein Datenangreifer einen Einzelwert w, welcher um weniger als γ von seinem Originalwert w relativ abweicht, richtig zuordnet, werde mit $P_\gamma(w \text{ enthüllt})$ bezeichnet und im Folgenden Enthüllungsrisiko genannt. Nach diesem von Höhne et al. (2003) vorgestellten und in Lenz et al. (2004b) formalisierten Schutzwirkungskonzept ergibt sich der Nutzen einer erfolgreichen Zuordnung aus den brauchbaren Informationen, die ein Datenangreifer bei einer erfolgreichen Reidentifikation enthüllen kann. Auf Merkmalsebene wird untersucht, in welchem Umfang ein durchgeführter Datenangriff auf einen Einzelwert (bzw. eine Information) als erfolgreich einzustufen ist, also die beiden Ereignisse „erfolgreiche Zuordnung der Merkmalsträger“ und „Zuordnung einer brauchbaren Information“ gleichzeitig auftreten. Als Formalisierung dieser Abschätzung erhalten wir daher die Wahrscheinlichkeit einer Enthüllung:

$$\begin{aligned} P_\gamma(w \text{ enthüllt}) \quad := \quad & P(m \text{ erfolgreich zugeordnet}) * P_\gamma(w' \text{ weicht relativ} \\ & \text{weniger als } \gamma \text{ von } w \text{ ab} \,|\, m \text{ erfolgreich zugeordnet}), \end{aligned}$$

wobei der erste Faktor durch den Anteil der erfolgreichen Zuordnungen geschätzt werden kann. Der zweite Faktor liest sich als bedingte Wahrscheinlichkeit, „eine brauchbare Information (auf Merkmalsebene) zu finden, gegeben eine erfolgreiche Zuordnung (auf Merkmalsträgerebene)“. Insgesamt erhält man demnach einen Schätzer $\hat{P}_\gamma(w \,\text{enthüllt})$ für das Enthüllungsrisiko $P_\gamma(w \,\text{enthüllt})$. Ein risikoaverser Datenangreifer wird ab einer bestimmten Wahrscheinlichkeit, aus einem Enthüllungsversuch für ihn unbrauchbare Werte zu erhalten, von einem Datenangriff absehen. Da ihm die Informationen fehlen, um reidentifizierte von nicht reidentifizierten Merkmalsträgern und brauchbare von unbrauchbaren Werten zu unterscheiden, bedeutet das, dass eine Datei als faktisch anonym eingestuft werden kann, auch wenn Merkmalsträger dieser Datei richtig zuordenbar sind *und* dabei brauchbare Einzelwerte enthüllt werden können. Die formale Bedingung für faktische Anonymität lautet demnach, dass eine vorgegebene obere Schwelle τ für das Enthüllungsrisiko nicht überschritten werden darf, d.h. es muss

$$\hat{P}_\gamma(w \,\text{enthüllt}) < \tau$$

gelten. Wird diese von den Datenhaltern gewissenhaft a priori festgelegte obere Schwelle nicht überschritten, so kann die zugrundeliegende Datei als faktisch anonym eingestuft werden. Je nach betrachteter Statistik kann diese Schwelle mit unterschiedlichen Werten angesetzt werden. In Teil XI wird das Maß mit Vorschlägen für eine geeignete Schwellenwahl bei den Projektstatistiken Kostenstrukturerhebung im Verarbeitenden Gewerbe, Umsatzsteuerstatistik und Einzelhandelsstatistik vorgestellt.

Kapitel 13

Simulation von Massenfischzügen

Wie bereits in Abschnitt 11.1 geschildert, sind zur Gegenüberstellung zweier Datenquellen so genannte Überschneidungsmerkmale nötig, die in beiden Quellen enthalten sind. Offenbar ist die Qualität solcher Merkmale wesentlich für den Erfolg eines Reidentifikationsprozesses. Wir unterscheiden zwischen zwei Typen von Überschneidungsmerkmalen, den metrischen und kategorialen Merkmalen, die nachfolgend beschrieben werden.

13.1 Merkmals- und Distanztypen

13.1.1 Metrische und kategoriale Merkmale

Metrische Merkmale sind definiert als diskrete oder stetige Merkmale, für welche der Differenz zwischen den einzelnen Ausprägungen Bedeutung zukommt, wie z.B. *Größe, Gewicht* einer Person oder *Anzahl der Beschäftigten, Gesamtumsatz* eines Unternehmens. Als Distanzmaß wird bei metrischen Merkmalen oftmals die quadratische Abweichung gewählt. Sind $a = (a_1, \ldots, a_n)$ und $b = (b_1, \ldots, b_n)$ zwei Merkmalsträger und a_i bzw. b_i die Ausprägungen im metrischen Merkmal v_i, so bestimmt $d_i(a, b) = (a_i - b_i)^2$ die Distanz zwischen a und b in diesem Merkmal.

Bei den kategorialen Merkmalen werden die Merkmalsausprägungen als Kategorien (oder Klassen) interpretiert, wobei jeder Merkmalsträger in eine bestimmte Kategorie fällt. Wir unterscheiden hier zwischen **nominalen Merkmalen** (es gibt keine Ordnung auf den Kategorien) und **ordinalen Merkmalen** (es gibt eine lineare Ordnung auf den Kategorien, wobei Differenzen zwischen den Kategorien keine Bedeutung zukommen muss). Die Ausprägungen nominaler Merkmale v_i können nur auf Gleichheit untersucht werden. D.h., mit

$$d_i(a, b) \quad = \quad \begin{cases} 0, & \text{wenn } a_i = b_i, \\ 1 & \text{sonst} \end{cases} \tag{13.1}$$

wird ein Distanzmaß für nominale Merkmale definiert. Sei nun v_i ein ordinales Merkmal und $c_1 <_i c_2 <_i \ldots <_i c_r$ der zugehörige geordnete Wertebereich, wobei $<_i$ (lies „kleiner als“) die Ordnungsrelation auf dem Wertebereich beschreibt. Die erweiterte Relation $\leq_i$ (lies „kleiner oder gleich“) schließt den Fall der Gleichheit mit ein. Wir definieren

$$d_i(a, b) = \frac{|\{c_j \mid \min(a_i, b_i) \leq_i c_j <_i \max(a_i, b_i)\}|}{r} \tag{13.2}$$

als Distanzmaß. Da die Differenz zwischen zwei Kategorien bei ordinalen Merkmalen in der Regel bedeutungslos ist, werden in obigem Maß die dazwischen liegenden Kategorien gezählt. In der Praxis kann es vorkommen, dass der Wertebereich eines Merkmals keine lineare Struktur besitzt. Im Falle solcher hierarchischer Merkmale wird folgendes Distanzmaß vorgeschlagen:

$$d_i(a, b) = \max\{f(c_j) \mid c_j < a_i \text{ und } c_j < b_i\}, \tag{13.3}$$

wobei $f(c_k) = 0$ für die ordnungserhaltende (monotone) Abbildung f gilt, wenn es kein c_l mit $c_k <_i c_l$ gibt. In unserem Kontext können hierarchische Merkmale u.a. dort auftauchen, wo ein kategoriales Merkmal infolge einer traditionellen Anonymisierungsmaßnahme in verschiedene Gliederungstiefen vergröbert wurde. Zum Beispiel kann in einer Wirtschaftsstatistik das Merkmal *Wirtschaftszweigklassifikation* (NACE Code) für einige Merkmalsträger auf Dreistellerebene (etwa mit der Ausprägung „101“) und für andere auf Zweistellerebene (etwa mit der Ausprägung „10“) ausgewiesen sein (siehe Abbildung 13.1).

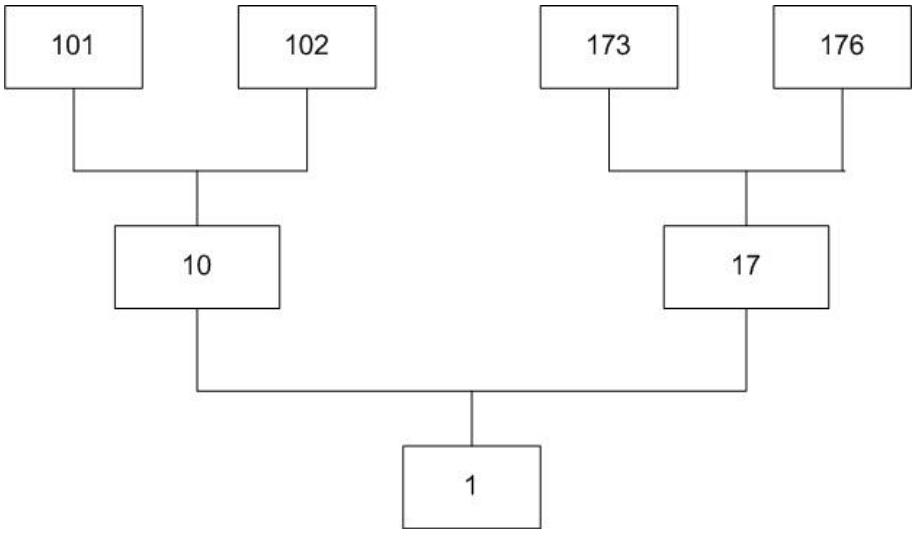

Abbildung 13.1: Vergröberung von Wirtschaftsklassen

In diesem Beispiel ist die via Gleichung (13.3) berechnete Distanz der einfachen $(0-1)$-Distanz in Gleichung (13.1) vorzuziehen, da letztere hier zu stark separierend wirkte.

13.1.2 Blockmerkmale

Bei großen Datenmengen ist es aus Sicht eines potenziellen Datenangreifers empfehlenswert, die Daten zunächst in geeignete Blöcke zu zerlegen. Die Blockung von Daten ist ein Verfahren zur Vorauswahl von möglichen Paaren von Merkmalsträgern für eine spätere Zuordnung. Paare von Merkmalsträgern werden a priori von der Zuordnung ausgeschlossen, wenn sie sich in einigen ausgezeichneten Merkmalen unterscheiden. Diese Merkmale werden **Blockmerkmale** genannt.

Bei geschickter Wahl der Blockmerkmale können Fehlzuordnungen vermieden, Speicherplatz gespart und Rechenaufwand reduziert werden. Obwohl die Anzahl möglicher Fehlzuordnungen mit der Anzahl falsch klassifizierter Merkmalsträger wächst (d.h., zwei Merkmalsträger a and b, die zu derselben zugrundeliegenden Einheit gehören, landen möglicherweise nicht in demselben Block), sind Fehlzuordnungen besonders in großen Blöcken mit vielen ähnlichen Merkmalsträgern zu erwarten.[16] Wie gut es einem Datenangreifer gelingen kann, hier einen vernünftigen Kompromiss zu finden, hängt in erster Linie von der Zuverlässigkeit der zur Blockung verwendeten Merkmale ab. Inkompatibilitäten zwischen den beiden Datenquellen in den Blockmerkmalen sind für den Datenangreifer schwierig auszumachen. Im (für den Datenangreifer) schlimmsten Falle sind die in den beiden Quellen gebildeten zueinander gehörigen Blöcke disjunkt (durchschnittsleer), im besten Falle dienen die Blockmerkmale als direkte Identifikatoren, so dass genau zwei zueinander gehörige Merkmalsträger einen gemeinsamen Block bilden. Der Datenangreifer wird daher nach Möglichkeit solche Merkmale als Blockmerkmale auswählen, die seines Wissens nach in den Zieldaten nicht oder nur geringfügig mit datenverändernden Verfahren behandelt wurden. Wenn allein informationsreduzierende Verfahren auf ein Merkmal angewendet wurden, dann kann das Merkmal mit den jeweiligen Einschränkungen zur Blockung verwendet werden. In den späteren Anwendungen werden daher kategoriale Merkmale (wie z.B. Wirtschaftszweigklassifikation oder Rechtsform) als Blockmerkmale gewählt. Sollte ein metrisches Merkmal zur Blockung vorgesehen sein, empfiehlt es sich, zuvor den Wertebereich in disjunkte Intervalle zu aggregieren.

13.2 Lineares Zuordnungsproblem

Um die verschiedenen Typen der komponentenweise definierten Distanzen d_r angemessen zusammenzubringen, müssen diese zunächst standardisiert werden, z.B. durch die $max - min$ Standardisierung

$$\widetilde{d}_r(a,b) := \frac{d_r(a,b) - \min_{(\alpha,\beta)\in A\times B} d_r(\alpha,\beta)}{\max_{(\alpha,\beta)\in A\times B} d_r(\alpha,\beta) - \min_{(\alpha,\beta)\in A\times B} d_r(\alpha,\beta)}. \tag{13.4}$$

16) Eine empirische Untersuchung findet sich in (Lenz und Vorgrimler 2005).

Da der Erfolg der späteren Zuordnung im Wesentlichen von der Wahl der Distanzmaße und der Qualität der einzelnen Überschneidungsmerkmale abhängt, wird der Datenangreifer einige Merkmale den anderen vorziehen. Wir definieren daher für jedes $(a, b) \in A \times B$ die Distanz

$$d(a, b) := \sum_{i=1}^{k} \lambda_i \cdot d_i(a, b). \tag{13.5}$$

Dieser Ausdruck ist eine gewichtete Summe aller komponentenweise definierten Distanzen. Eine Formalisierung dieses Ansatzes findet sich in Lenz (2003c).

Auf Basis der Distanzen besteht nun die Aufgabe darin, möglichst viele Merkmalsträger der externen Datei den zugehörigen Merkmalsträgern der Zieldatei korrekt zuzuordnen. Da es unmöglich erscheint, sämtliche gesuchte Einheiten korrekt zuzuordnen, wird im Folgenden versucht, die Anzahl der Fehlzuordnungen zu minimieren. Sei hierzu $n = |A| = |B| = m$. Andernfalls betrachten wir ohne Beschränkung der Allgemeinheit den Fall $m < n$ und definieren neue Merkmalsträger $b_{m+1}, \ldots, b_n$ und damit neue mögliche Paarungen (a_i, b_j) für $i = 1, \ldots, n$ und $j = m + 1, \ldots, n$, deren zugehörige Distanzen auf

$$d(a_i, b_j) := \max_{(a,b) \in A \times B} d(a, b) \tag{13.6}$$

gesetzt werden.

Wir kürzen die zuvor berechneten Distanzen mit $d_{ij} := d(a_i, b_j)$ ab und formulieren folgendes Zuordnungsproblem.[17]

$$\text{Minimiere} \quad \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} x_{ij} \tag{13.7}$$

$$\begin{aligned} \text{unter} \quad & x_{ij} \in \{0, 1\} \quad \text{for} \quad i, j = 1, \ldots, n, \\ & \sum_{j=1}^{n} x_{ij} = 1 \quad \text{für} \quad i = 1, \ldots, n \quad \text{und} \\ & \sum_{i=1}^{n} x_{ij} = 1 \quad \text{für} \quad j = 1, \ldots, n. \end{aligned}$$

Die Nebenbedingungen stellen sicher, dass jedes a_i genau einem b_j zugeordnet wird und umgekehrt. Es gilt $x_{ij} = 1$ genau dann, wenn a_i und b_j einander zugeordnet werden.

Die Aufgabe besteht also mit anderen Worten darin, eine Permutation π über der Menge $\{1, \ldots, n\}$ zu finden, welche die Summe $\sum_{i=1}^{n} d_{i,\pi(i)}$ (i.e. die Gesamtdistanz aller Zuordnungen) minimiert. Der naivste und rechentechnisch aufwändigste Weg zu einer Lösung

17) Ein allgemeinerer Zugang über multikriterielle Optimierung ist in Lenz (2003b) dargestellt

wäre sicher, alle $n!$ möglichen Permutationen durchzuprobieren und diejenige mit der geringsten Gesamtdistanz auszuwählen. Obgleich es klassische Verfahren gibt, wie z.B. die bekannte Simplex-Methode, die sich – trotz exponentieller Worst-Case Laufzeit – in vielen praktischen Anwendungen als effizient erwiesen hat, traten in dem vorliegenden Falle bereits Probleme bei der Arbeit mit (für Wirtschaftsstatistiken) verhältnismäßig kleinen Datenmengen der Größenordnung von 10.000 Einheiten auf. Die in Gleichung (13.7) formulierten Nebenbedingungen stellen ein lineares Gleichungssystem mit folgender Matrix dar:

$$\left(\begin{array}{cccc|cccc|cccc|ccc|c}
1 & 1 & \cdots & 1 & 0 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
0 & 0 & \cdots & 0 & 1 & 1 & \cdots & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \\
0 & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 1 & \cdots & 1 \\
\hline
 & \ddots & & & & \ddots & & & & \ddots & & & & \ddots & & \\
 & & I_n & & & & I_n & & & & I_n & & & & I_n & \\
 & & & \ddots & & & & \ddots & & & & \ddots & & & & \ddots
\end{array}\right)$$

wobei $I_n = diag(1, 1, \ldots, 1)$ die Einheitsmatrix der Dimension $n \times n$ definiere. Allein diese Koeffizientenmatrix, bestehend aus $2n$ Zeilen und n^2 Spalten, kann zum Überschreiten des Arbeitsspeichers eines handelsüblichen PC führen. Eine leichte Verbesserung ist durch die so genannte Ungarische Methode (siehe Kuhn (1955)) gegeben, deren Aufwand von der Ordnung $O(n^3)$ (d.h., man kann den Rechenaufwand durch ein Polynom dritten Grades in der Länge des Eingabevektors abschätzen) bei größeren Datenmengen ebenfalls beachtlich ist.

13.3 Greedy-Heuristiken

Um den Rechenaufwand zu reduzieren, wurden im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ zwei Heuristiken auf die Projektdaten angewendet und ihr Ergebnis mit der optimalen Lösung des linearen Programmes Gleichung (13.7) verglichen (für eine ausführliche Beschreibung siehe Lenz (2003a)).

Obwohl die beiden unten aufgeführten Greedy-Heuristiken nur Näherungslösungen für Gleichung (13.7) liefern, wurden sie aufgrund ihres zweifellosen Vorteils der guten Effizienz herangezogen. In der Tat weisen die beiden Heuristiken eine Komplexität von $O(knm)$ auf, wobei k die Anzahl der Überschneidungsmerkmale ist. Da in der Regel $k << n$ und $k << m$ gilt (d.h., es gibt weitaus mehr Merkmalsträger als Merkmale), können wir den Faktor k bei der Komplexitätsanalyse vernachlässigen.

Prozedur I: `begin` {PROC I}

$\mathcal{M} := \emptyset$

$i := 1$

`While` $(i \leq n$ `and` $B \neq \emptyset)$ `do`

$b' := arg\,min_{b \in B} d(a_i, b)$

$\mathcal{M} := \mathcal{M} \cup \{(a_i, b')\}$

$B := B \setminus \{b'\}$

$i := i + 1$

`end` {PROC I}

Die Ausgabe der Prozedur ist eine eindeutige Zuordnung zwischen A und B. Offensichtlich hängt die Ausgabe von der Anfangsnummerierung der Merkmalsträger $a_1, \ldots, a_n$ ab. Seien nun ohne Beschränkung der Allgemeinheit $a_1, \ldots, a_r$ und $b_{\pi(1)}, \ldots, b_{\pi(r)}$ paarweise einander zugeordnet. In Schritt $r+1$ wird a_{r+1} dem Merkmalsträger b mit minimalem Abstand zu a_{r+1} zugeordnet. Dabei ist b einer der verbliebenen Merkmalsträger in B, die zu diesem Zeitpunkt noch nicht zugeordnet wurden. An dieser Stelle sei angemerkt, dass eine Streichung der siebten Zeile in Prozedur I bewirken würde, dass ein a_i verschiedenen $b \in B$ zugeordnet werden könnte und damit die resultierende Zuordnung nicht mehr eindeutig wäre. Eine solche Heuristik kann als Schätzer für einen Einzelangriff interpretiert werden, allerdings mit der Einschränkung, dass dem Zusatzwissen einige, allein dem Einzelangreifer individuell verfügbare Überschneidungsmerkmale fehlen.

Das gegenüber der optimalen Lösung des linearen Programmes in Gleichung 13.7 schlechte Abschneiden von Prozedur I war aus oben genannten Gründen zu erwarten. Sie wurde aber dennoch Tests unterzogen, weil sie in der Literatur oft verwendet wird. Eine wesentliche Verbesserung wird durch untenstehende Prozedur II erreicht, bei der die Anfangsnummerierung der Merkmalsträger weitaus weniger Einfluss auf die Ausgabe hat. Die Idee besteht in einer sukzessiven Auswahl von Paaren mit kleinstmöglicher Distanz. Die Prozedur endet, wenn eine der beiden Datenquellen abgearbeitet ist.

Prozedur II: `begin` {PROC II}

Sortiere die Distanzen in einer aufsteigenden Liste $\mathcal{L}$

`While` L ist nichtleer `do`

Betrachte das erste Element d_{ij} von $\mathcal{L}$ und ordne (a_i, b_j) zu.

Entferne alle Elemente d_{rs} von $\mathcal{L}$, für die $r = i$ oder $s = j$ gilt.

`end` {PROC II}

Die Erfahrung hat gezeigt, dass Prozedur II Ergebnisse nahe bei der (optimalen) Lösung liefert. Darüber hinaus ist es derzeit nahezu unmöglich, für große Dateien (wie z.B. die in Abschnitt 9.2 beschriebene Umsatzsteuerstatistik) in angemessener Zeit die optimale Lösung zu bestimmen.

13.4 Algorithmus für den Massenfischzug

Basierend auf den vorhergehenden Überlegungen wird folgender Algorithmus zur Durchführung eines Massenfischzuges vorgeschlagen:

1) Input: Mengen $A = \{a_1, \ldots, a_n\}, B = \{b_1, \ldots, b_m\}$ von Merkmalsträgern und $V = \{v_1, \ldots, v_k\}$ von Überschneidungsmerkmalen.

2) Partitionierung der Dateien in Blöcke via Blockmerkmale $BV \subseteq V$.

3) Berechnung der Komponentendistanzen $d_r(a_i, b_j), \quad r = 1, \ldots, k$.

4) Optionales Setzen individueller Gewichte $\Lambda = (\lambda_1, \ldots, \lambda_k)$.

5) Anwendung von Prozedur II

6) Output: $(1-1)$-Zuordnung zwischen A und B.

Zur Illustration des Algorithmus betrachten wir ein kleines Beispiel. Wir versuchen, vier Merkmalsträger in $A = \{a_1, \ldots, a_4\}$ mit vier Merkmalsträgern in $B = \{b_1, \ldots, b_4\}$ zu verknüpfen. Als Überschneidungsmerkmale stehen fünf metrische Merkmale $v_1, \ldots, v_5$ zur Verfügung (siehe Tabelle 13.1).

Eine Aufzählung aller 24 möglichen eindeutigen Zuordnungen zwischen A und B liefert Abbildung 13.2.

Tabelle 13.1: Beispiel

M-träger/ Merkmale	v_1	v_2	v_3	v_4	v_5
a_1	14008906	755187	907264	6582133	4794809
a_2	14309437	673189	1179713	8111720	5407676
a_3	14330083	567300	920065	4871720	1667078
a_4	14780637	567553	1026861	5313029	3654241
b_1	14825332	563928	913631	4978410	1711353
b_2	14045802	724071	1040229	7064023	5078378
b_3	13945802	682110	973631	7378984	508494
b_4	14996199	563928	1050673	5252164	3871084

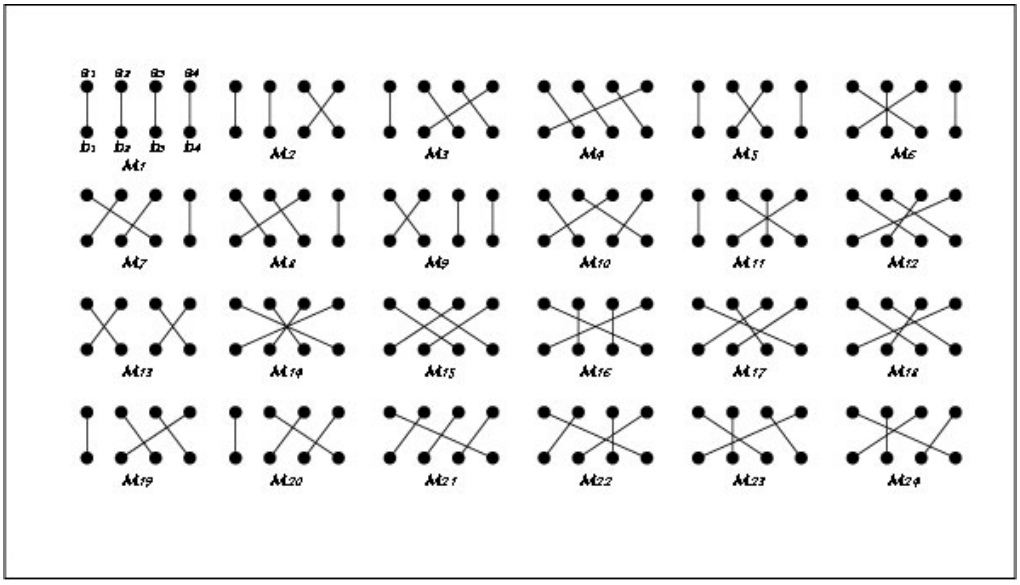

Abbildung 13.2: Eindeutige Zuordnungen

In diesem Beispiel führt der Algorithmus zu der Zuordnung $\mathcal{M}_6$, die gleichzeitig die optimale Lösung bestimmt.

Der hier beschriebene Algorithmus wird später zur Bewertung der Schutzwirkung von Anonymisierungsverfahren auf die Projektdaten angewendet.

Teil V

Die Operationalisierung des Analysepotenzials

Neben der Sicherstellung der faktischen Anonymität besteht die zweite große Herausforderung bei der Anonymisierung in der Erhaltung des Analysepotenzials von Daten. Im Rahmen des Projekts wurde der Versuch unternommen, die Aspekte der Beeinflussung des Analysepotenzials durch Anonymisierungsmaßnahmen möglichst weitgehend zu systematisieren. In diesem Teil wird das Vorgehen bei der Operationalisierung des Analysepotenzials beschrieben. In Kapitel 14 werden zunächst die Bestimmungsfaktoren des Analysepotenzials von anonymisierten Daten systematisch dargestellt. In Kapitel 15 erfolgt anschließend eine Beschreibung der Ansätze zur Messung der Einschränkungen des Analysepotenzials durch datenverändernde Anonymisierungsverfahren.

Kapitel 14

Bestimmungsfaktoren des Analysepotenzials von Daten

14.1 Zur generellen Definition des Analysepotenzials

Unter dem Analysepotenzial eines Datenbestandes wird grundsätzlich die Gesamtheit der Nutzungsmöglichkeiten verstanden, also das Potenzial der Daten in Bezug auf die Untersuchung inhaltlicher Fragestellungen mit Hilfe deskriptiver und inferenzstatistischer Methoden. Im Hinblick auf die Anonymisierung von Einzeldaten, insbesondere mit Hilfe datenverändernder Anonymisierungsverfahren, wie sie in Kapitel 6 beschrieben wurden, umfasst der Begriff des Analysepotenzials jedoch auch die Sicherheit für einen Datennutzer, mit den anonymisierten Daten inhaltlich gleichgerichtete und gleichwertige Ergebnisse zu erzielen wie mit den Originaldaten. Dies bedeutet, dass das Analysepotenzial im Wesentlichen durch drei Aspekte determiniert wird:

- die inhaltliche Fragestellung,
- die methodische Herangehensweise,
- die Analyseergebnisse.

Damit kann sich eine Verringerung des Analysepotenzials grundsätzlich auch in dreierlei Hinsicht ergeben:

- Die inhaltliche Fragestellung kann nicht mehr untersucht werden.
- Die Methode kann nicht mehr angewendet werden.
- Die Analyse kommt zu abweichenden Ergebnissen.

Die möglichen zu untersuchenden inhaltlichen Fragestellungen sind ebenso wie die Analysemethoden sehr vielseitig. Auch die Frage, ob eine Analyse zu anderen Ergebnissen führt,

kann nur in Abhängigkeit vom inhaltlichen Ziel der Untersuchung beantwortet werden. Beispielsweise interessieren in manchen Fällen bei einer Regressionsschätzung nur die Vorzeichen der Koeffizienten und die Signifikanz der Einflussfaktoren, in anderen sind hingegen die Größenordnungen der geschätzten Koeffizientenwerte von Interesse.

Damit ist festzustellen, dass das Analysepotenzial von Originaldaten und damit seine Einschränkung durch Anonymisierungsverfahren nicht eindeutig und objektiv erfassbar ist. Im Folgenden wird versucht, wesentliche Aspekte der Veränderung des Analysepotenzials durch Anonymisierungsverfahren herauszuarbeiten.

14.2 Informationseinschränkungen versus Ergebnisveränderung

Legt man die im vorangegangenen Abschnitt benannten Determinanten des Analysepotenzials von Daten zugrunde, so ergeben sich Einschränkungen des Analysepotenzials zum einen dadurch, dass bestimmte Auswertungen aufgrund von Anonymisierungsmaßnahmen von vornherein ausgeschlossen sind, weil entweder die inhaltliche Fragestellung nicht mehr untersucht werden kann oder die anzuwendende Methode inklusive gleichwertiger anderer Methoden nicht mehr durchführbar ist. Zum anderen entstehen sie durch die Veränderung von Analyseergebnissen bei anonymisierten Daten gegenüber den Originaldaten. Daraus folgt, dass die Wahl der Anonymisierungsmethode unmittelbare Folgen für die Art der Einschränkung des Analysepotenzials hat.

So geht die Anwendung von Verfahren zur Informationseinschränkung, wie sie in Kapitel 5 dieses Bandes dargestellt wurden, in der Regel mit einer Reduzierung der Bandbreite an inhaltlichen Auswertungsmöglichkeiten einher. Werden besonders reidentifikationsgefährdete Teilgesamtheiten, z.B. Großunternehmen, entfernt, so sind für diese auch keine Auswertungen mehr möglich. Werden Variablen zur Regionalangabe aus dem Datensatz entfernt, so sind auch keine Regionalanalysen beziehungsweise regional differenzierende Auswertungen mehr möglich.

Werden hingegen datenverändernde Verfahren (vgl. Kapitel 6) angewendet, so wird die Bandbreite der Analysemöglichkeiten aus inhaltlichen Gründen in der Regel nicht eingeschränkt. Allerdings können datenverändernde Anonymisierungsverfahren zu einer Veränderung von Analyseergebnissen führen. Außerdem kann die Anwendung von datenverändernden Anonymisierungsmethoden auch zu einer Einschränkung der methodischen Möglichkeiten führen oder den Einsatz anderer Methoden erforderlich machen, beispielsweise wenn der Einsatz einer bestimmten Methode an bestimmte Annahmen gebunden ist und diese aufgrund der durchgeführten Anonymisierungsmaßnahmen nicht mehr zutreffen. Als Beispiel kann hier die Anwendung der stochastischen Überlagerung angeführt werden. Der Einsatz dieser Methode macht in den meisten Fällen den Einsatz von korrigierten Schätzern für Fehler-in-den-Variablen-Modelle oder Instrumentvariablen-Schätzern erforderlich (vgl. hierzu auch die Unterabschnitte 22.1.1 und 22.1.2).

Allerdings kann auch die Anwendung datenverändernder Anonymisierungsverfahren in Einzelfällen bestimmte Auswertungen unmöglich machen. Ein Beispiel ist die Auswertung von Teilgesamtheiten mit simulierten Daten. Werden Simulationsverfahren auf den gesamten Datenbestand angewandt und werden kategoriale Merkmale dabei nicht beachtet, so sind Teilmassenauswertungen grundsätzlich nicht mehr möglich. Werden hingegen die diskreten Variablen bei der Simulation berücksichtigt und wird die Simulation für bestimmte Teilgesamtheiten getrennt durchgeführt, so sind zumindest für diese Teilgesamtheiten weiterhin getrennte Auswertungen möglich.

Zusammengefasst ergibt sich aus diesen Überlegungen die folgende Konsequenz für den Zusammenhang zwischen der Art der Anonymisierungsmaßnahme und der Wirkung auf das Analysepotenzial:

- Verfahren zur Informationseinschränkung (Kapitel 5) und in Einzelfällen auch datenverändernde Verfahren (Kapitel 6) führen zu einer Reduzierung der Informationen im Datensatz und damit zu einer Einschränkung der inhaltlichen Bandbreite an Analysemöglichkeiten. Dies kann dazu führen, dass Analysen nicht mehr oder nicht mehr mit dem gleichen Informationsgehalt durchgeführt werden können.

- Datenverändernde Anonymisierungsverfahren (Kapitel 6) führen zu Veränderungen der Einzelwerte. Dies kann dazu führen, dass Methoden nicht mehr oder nicht mehr in der gleichen Weise zur Anwendung gebracht werden können beziehungsweise die Ergebnisse bei Anwendung dieser Methoden auf die anonymisierten Daten verzerrt sind. In diesem Fall sind entweder Analysemethoden zu verwenden, die den Veränderungen in den Einzeldaten Rechnung tragen (Korrekturverfahren) oder die Ergebnisveränderungen müssen sich in einem Rahmen bewegen, der für den Nutzer noch akzeptabel ist.

Die Beurteilung, welche inhaltlichen Möglichkeiten ein Scientific-Use-File vorrangig erfüllen muss, entzieht sich einer objektiven wissenschaftlichen Bewertung. Diese Aufgabe muss vielmehr in enger Abstimmung zwischen den potenziellen Datennutzern und den Datenlieferanten für jedes einzelne Datenangebot geklärt werden. Letztlich müssen die Nutzungswünsche der Wissenschaft das Datenangebot und damit auch das Vorgehen beim Einsatz von Verfahren zur Informationsreduktion entscheidend determinieren. Zu empfehlen ist daher, für zukünftige Vorhaben zur Bereitstellung von Scientific-Use-Files die interessierende Gesamtheit, die erforderliche Variablenliste sowie die Tiefe der Ausprägungen bei kategorialen Variablen in einem Arbeitskreis unter Beteiligung der datenerhebenden Stelle und der potenziellen Nutzer festzulegen. Die Forschungsdatenzentren sowie der Rat für Wirtschafts- und Sozialdaten können hierbei eine koordinierende Funktion wahrnehmen.

Wissenschaftlich objektiv beantwortbar ist hingegen die Frage, ob und inwieweit Analysemethoden nach dem Einsatz datenverändernder Anonymisierungsverfahren noch anwendbar sind, welche Auswirkungen die Anwendung dieser Verfahren auf die anonymisierten Daten

hat und welche Korrekturverfahren gegebenfalls zur Anwendung kommen können oder müssen. Diese Fragen sind im Wesentlichen Gegenstand des theorieorientierten Ansatzes zur Beurteilung der Auswirkungen von datenverändernden Anonymisierungsverfahren auf das Analysepotenzial. In Abschnitt 15.3 wird dieser Ansatz vorgestellt. In diesem Handbuch sollen diese Fragen vor allem für stochastische Überlagerungen und Mikroaggregationsverfahren und für verschiedene deskriptive Maße und ökonometrische Modelle beantwortet werden.

Lässt sich die Veränderung von Analyseergebnissen nicht verhindern, so stellt sich weiterhin die Frage nach Grenzwerten, innerhalb derer die Ergebnisveränderungen aus Nutzersicht akzeptabel sind. Wie bereits im vorangegangenen Abschnitt angedeutet wurde, hängt auch dies entscheidend von der inhaltlichen Fragestellung und den Präferenzen der Nutzer ab und lässt sich daher nicht objektiv klären. Die Beurteilung einer Ergebnisveränderung hängt zudem von der Beschaffenheit und dem Aussagegehalt der Ergebnisse mit den Orginaldaten ab. Beispielsweise ist zu vermuten, dass bei einem starken statistisch gesicherten Zusammenhang zwischen abhängiger und erklärender Variablen in einem Regressionsmodell die Anwendung datenverändernder Anonymisierungsverfahren zu einer geringeren Beeinträchtigung dieses Zusammenhangs führt als bei einem nur schwach ausgeprägten Einfluss der erklärenden Variablen.

Anhaltspunkte für die Größenordnung von Veränderungen erhält man bei der Durchführung beispielhafter Auswertungen im Rahmen des in Abschnitt 15.2 vorgestellten anwendungsorientierten Ansatzes zur Beurteilung der Einschränkungen des Analysepotenzials durch datenverändernde Anonymisierungsverfahren. Auf dieser Grundlage kann durch die Festlegung und Veröffentlichung von bei der Anonymisierung zu beachtenden Abweichungsgrenzen für bestimmte Maße eine zusätzliche Information für den Nutzer hinsichtlich der Abweichung von Analyseergebnissen bereitgestellt werden. Hierbei kann auch der in Abschnitt 15.1 vorgestellte maßzahlorientierte Ansatz zur Beurteilung der Auswirkungen datenverändernder Anonymisierungsverfahren auf das Analysepotenzial hilfreich sein.

Da aus der Sicht der Datensicherheit die Anwendung von Informationseinschränkungen beispielsweise durch das Entfernen von Schlüsselmerkmalen einen ähnlichen Schutz erzeugen kann wie der entsprechend dosierte Einsatz datenverändernder Verfahren, ist im Einzelfall abzuwägen, welches Vorgehen aus Nutzersicht zu befürworten ist.

14.3 Relevanz unterschiedlicher Datengrundlagen

Es ist sehr wichtig zu betonen, dass auch die Beschaffenheit der jeweiligen Datengrundlage den Umfang der Verringerung des Analysepotenzials durch Anonymisierungsverfahren erheblich beeinflusst. Dies gilt sowohl hinsichtlich einer Einschränkung der inhaltlichen Bandbreite an Analysemöglichkeiten durch Informationsreduzierung als auch hinsichtlich der Veränderung von Analyseergebnissen durch Datenveränderungen.

Bei der Interpretation einer Veränderung von Analyseergebnissen durch den Einsatz datenverändernder Verfahren ist zu berücksichtigen, dass eine stärkere Ergebnisveränderung auch dadurch zustande kommen kann, dass bereits die mit den Originaldaten erzielten Ergebnisse nicht in ausreichendem Maße robust sind. Dies kann beispielsweise durch eine Fehlspezifikation des geschätzten Modells aufgrund mangelnder Datenverfügbarkeit, auf Datenfehler in den Originalangaben oder auf das Zugrundeliegen einer zu kleinen Stichprobe zurückzuführen sein. Die Lagerbestände in der Kostenstrukturerhebung für das Verarbeitende Gewerbe und den Bergbau werden beispielsweise von den Unternehmen selbst nur geschätzt. Die entsprechenden Merkmale weisen somit eine geringere Qualität auf als andere Merkmale dieser Erhebung. In Rosemann (2004) wird gezeigt, dass die Höhe der Korrelation zwischen erklärenden und abhängiger Variablen und damit auch der Erklärungsgehalt des Modells beziehungsweise die Güte der Modellspezifikation einen Einfluss auf die Wirkungsweise der getrennten Mikroaggregation in linearen Schätzmodellen hat. Wird für einen Datenbestand nur eine Stichprobe erfasst und ist die Stichprobe so beschaffen, dass der Einfluss eines Koeffizienten in einem ökonometrischen Modell nur schwach signifikant ist, so kann der Einsatz datenverändernder Verfahren eher dazu führen, dass die Signifikanz verloren geht.

Zuletzt ist von besonderer Bedeutung, dass die Anforderungen aus Sicht des Analysepotenzials an Paneldaten andere sind als an Querschnittsdaten. Die Anonymisierung von Paneldaten konnte insbesondere wegen dieser zusätzlichen Anforderungen innerhalb des Projektzeitraums nicht geleistet werden. Mit dieser Frage wird sich ein Folgeprojekt beschäftigen.

14.4 Berücksichtigung unterschiedlicher Analysemethoden

Wie bereits in Abschnitt 14.1 erwähnt wurde, sind die Methoden, die bei der Beurteilung der Auswirkungen von Anonymisierungsverfahren auf das Analysepotenzial eines Datenbestandes berücksichtigt werden müssten, unterschiedlich. Die Verfahren, die auf wirtschaftsstatistische Einzeldaten angewendet werden können, können grundsätzlich in deskriptive statistische Analysen (inklusive deterministischer Modelle, z.B. Steuermodelle) und ökonometrische Analysen stochastischer Modelle unterschieden werden.

Zu den wichtigsten deskriptiven Analysen gehören neben ein- und mehrdimensionalen Häufigkeitsauswertungen, wie beispielsweise die Häufigkeitsverteilung der Betriebe nach Wirtschaftszweigen, die Berechnung von Maßzahlen für den gesamten Datenbestand oder abgrenzbare Teilgesamtheiten. Dabei können die Teilgesamtheiten ein- und mehrdimensional sowohl nach kategorialen Merkmalen (beispielsweise Wirtschaftszweig oder Regionalkennung) als auch nach metrischen Merkmalen (Beschäftigung oder Umsatz) abgegrenzt werden. Die Auswertungen sind gleichermaßen für die ursprünglich im Datensatz enthaltenen Merkmale sowie für lineare und nichtlineare Transformationen dieser Merkmale durchführbar. Im Wesentlichen sind folgende Maßzahlen von Bedeutung:

- Mittelwerte (arithmetische Mittel und Mediane),
- Streuungsmaße, wie Varianzen und Variationskoeffizienten sowie Konzentrationsmaße,
- Korrelationsmaße (Bravais-Pearson Korrelationskoeffizienten und Spearman-Rangkorrelationskoeffizienten).

Weiterhin gehören deskriptive multivariate Verfahren wie Multiple Regressionsanalyse, Clusteranalyse, Hauptkomponentenanalyse und multidimensionale Skalierung zum Auswertungsprogramm wirtschaftsstatistischer Einzeldaten. Bei Paneldaten ist zudem die Analyse von Veränderungsraten und Übergängen relevant.

Die ökonometrischen Analysen[18] können grundsätzlich in lineare und nichtlineare stochastische Modelle unterteilt werden. Daneben wird die Modellstruktur im Wesentlichen durch die zugrundeliegenden Annahmen, beispielsweise in Bezug auf die Existenz heteroskedastischer Störterme, determiniert. Weitere inferenzstatistische multivariate Verfahren sind die Faktorenanalyse, LISREL-Verfahren sowie Latente Strukturmodelle.

Dabei ist offensichtlich, dass unterschiedliche Analysen auch unterschiedliche Anforderungen an die Anonymisierung von Einzeldaten stellen. Ökonometriker sind beispielsweise in erster Linie an der Korrelationsstruktur interessiert. Ausreißerbeobachtungen sind für sie nicht relevant, während diese bei der Berechnung univariater Verteilungsparameter oder bei Hochrechnungen quantitativer Angaben essentiell sein können (Gottschalk 2005, S. 41).

18) Es wird hier von ökonometrischen Analysen gesprochen, weil von ökonomischen Fragestellungen ausgegangen wird. Bei soziologischen Fragestellungen wird von soziometrischen Modellen gesprochen, in der Biologie von biometrischen.

Kapitel 15

Ansätze zur Messung des Analysepotenzials beim Einsatz datenverändernder Anonymisierungsverfahren

Wie bereits in Abschnitt 14.2 erwähnt wurde, sind die Einschränkungen des Analysepotenzials durch informationsreduzierende Maßnahmen nicht objektiv bewertbar. Vielmehr müssen die Auswirkungen solcher Verfahren auf das Analysepotenzial und die Rückschlüsse für ihren Einsatz in einem diskursiven Prozess zwischen Datenanbietern und Datennutzern geklärt werden. Die bei Projektbeginn vorgefundenen Ansätze zur Beurteilung von Anonymisierungsmaßnahmen vor dem Hintergrund einer möglichst weitgehenden Erhaltung des Analysepotenzials beschränken sich somit auf datenverändernde Anonymisierungsverfahren. Sie lassen sich wie folgt systematisieren (Ronning und Rosemann 2003):

- Maßzahlorientierter Ansatz
- Anwendungsorientierter Ansatz
- Theorieorientierter Ansatz

Während die beiden ersten Ansätze auf einer empirischen Herangehensweise an die Bewertung der Veränderung des Analysepotenzials beruhen, orientiert sich der dritte Ansatz an der theoretischen Ableitung der Auswirkungen von Anonymisierungsverfahren auf Analyseergebnisse. Im Folgenden werden die drei Ansätze im Einzelnen dargestellt.

15.1 Der maßzahlorientierte Ansatz

Die Idee des maßzahlorientierten Ansatzes besteht darin, die Veränderung des Analysepotenzials durch die Abweichung von Maßen zu bestimmen, mit der üblicherweise empirische Verteilungen charakterisiert werden. In der Literatur werden eine Reihe für die Bewertung

von Anonymisierungsmaßen in Frage kommender Abweichungsmaße von Verteilungseigenschaften dargestellt (Domingo-Ferrer et al. 2003; Höhne 2002; Ronning und Rosemann 2003). Dabei können für metrische Variablen folgende Arten von Maßen unterschieden werden:

- Abweichungen der einzelnen Werte
- Abweichungen der Mittelwerte und Varianzen (univariate Verteilungen)
- Abweichungen von Konzentrationsmaßen (univariate Verteilungen)
- Abweichungen von Zusammenhangsmaßen (multivariate Verteilungen)
- Abweichungen höherer Momente

In der Literatur werden bisher die Abweichungen der Einzelwerte, die Abweichungen der ersten und zweiten Momente sowie die Abweichungen von Zusammenhangsmaßen als Kriterien zur Beurteilung der Verringerung des Analysepotenzials durch Anonymisierungsmaßnahmen im Rahmen des maßzahlorientierten Ansatzes verwendet.

Grundsätzlich bietet es sich für die genannten Größen an, die mittleren relativen Abweichungen heranzuziehen, um zu verhindern, dass einzelne Variablen stärker gewichtet werden als andere. Allerdings gilt dies nur, sofern es sich – was bei den Projektstatistiken in der Regel der Fall ist – um nichtnegative Merkmale handelt und die Größen im Original nicht Null sind.

Gegen die Verwendung der mittleren relativen Abweichung bei den Medianen spricht, dass ein Großteil der Mediane bei den im Projekt untersuchten Datensätzen den Wert Null annimmt. Dies gilt umso stärker, je häufiger strukturelle Nullen auftreten. Dies bedeutet, dass bei bestimmten Variablen bestimmte Unternehmen von vornherein eine Null aufweisen müssen, weil sie beispielsweise keinen Handel betreiben oder keine Forschung und Entwicklung durchführen. Auf der anderen Seite ist der Median bei vielen Analysen bedeutsamer als das arithmetische Mittel, da er weniger empfindlich gegenüber Ausreißern ist.

Bei den Korrelationskoeffizienten scheint jedoch die Betrachtung der mittleren absoluten Abweichungen sinnvoller zu sein, weil diese zum einen bereits normiert sind und zum anderen bei sehr kleinen Werten (nahe bei Null) bereits marginale absolute Abweichungen zu sehr großen relativen Abweichungen führen und das Problem sehr kleiner Werte vor allem bei den Korrelationskoeffizienten auftritt.

Die mittlere Veränderung der einzelnen Werte ist als Maß zur Beurteilung der Verringerung des Analysepotenzials kritisch zu bewerten. 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 Veränderung von

Analyseergebnissen verbunden. Darüber hinaus ist die Berechnung dieses Maßes an die direkte Zuordnung der Untersuchungseinheiten gebunden. Dies ist jedoch für synthetische Datensätze in der Regel nicht möglich (vgl. Unterabschnitt 6.2.5).

Neben der Veränderung der Varianz selbst ist insbesondere die Veränderung einer normierten Varianz in Form eines Variationskoeffizienten von Interesse. Der Variationskoeffizient beziehungsweise die relative Varianz ergibt sich grundsätzlich als Verhältnis von Streuungsmaß zu Lagemaß. Besonders bekannt ist das Verhältnis von Standardabweichung und arithmetischem Mittel. Ein robusteres Maß wird durch das Verhältnis von Interquartilsabstand zu Median gegeben. Ein sehr oft verwendetes relevantes Streuungsmaß ist auch der Gini-Koeffizient, dessen Bezug zur Interpretation als Variationskoeffizient allerdings weniger geläufig ist. Man kann zeigen, dass der Gini-Koeffizient durch das Verhältnis von durchschnittlicher absoluter Differenz zum zweifachen arithmetischen Mittel gegeben ist (Ronning und Rosemann 2003). Der Gini-Koeffizient stellt ein Maß für die relative Konzentration dar. Ein Maß für die absolute Konzentration sind Konzentrationsraten sowie der Herfindahl-Index. Allerdings sind diese Maße nur interpretierbar, sofern die Variablen nur nichtnegative Werte annehmen können.

Ein weitergehender Vorschlag des maßzahlorientierten Ansatzes besteht darin, die aufgelisteten Maße zu einem Score beziehungsweise einer Maßzahl zu verdichten und anhand der Größe der Scorewerte eine Bewertung der Verringerung des Analysepotenzials durch die Anonymisierungsverfahren vorzunehmen (Domingo-Ferrer et al. 2003; Sebé et al. 2002; Dandekar et al. 2002).

In Domingo-Ferrer et al. (2003) wird folgender Score zur Messung der Verringerung des Analysepotenzials vorgeschlagen:

$$
\begin{aligned}
Score\ I \;=\; 100 \Bigg(& \frac{\sum_{j=1}^{d}\sum_{i=1}^{m}\frac{|x_{ij}-x_{ij}^{a}|}{|x_{ij}|}}{md} + \frac{\sum_{j=1}^{d}\frac{|\bar{x}_j-\bar{x}_j^{a}|}{|\bar{x}_j|}}{d} + \frac{\sum_{j=1}^{d}\sum_{1\le i\le j}\frac{|s_{ij}-s_{ij}^{a}|}{|s_{ij}|}}{\frac{1}{2}d(d+1)} \\
& + \frac{\sum_{j=1}^{d}\frac{|s_{jj}-s_{jj}^{a}|}{|s_{jj}|}}{d} + \frac{\sum_{j=1}^{d}\sum_{1\le i<j}\left|r_{ij}-r_{ij}^{a}\right|}{\frac{1}{2}d(d-1)} \Bigg)
\end{aligned}
\tag{15.1}
$$

Dabei ist d die Anzahl der Variablen, m die Anzahl der Untersuchungseinheiten. x_{ij} sind die Merkmalswerte der Einheit i für Merkmal j, $\bar{x}_j$ der Mittelwert des Merkmals j über alle Einheiten, s_{ij} die Kovarianz zwischen den Merkmalen i und j, s_{jj} die Varianz des Merkmals j und r_{ij} der Korrelationskoeffizient der Merkmale i und j.

Alternativ wird auch der folgende Score vorgeschlagen:

$$Score\ II = 100 \left(\frac{\sum_{j=1}^{d} \sum_{i=1}^{m} \frac{|x_{ij} - x_{ij}^a|}{|x_{ij}|}}{md} + 0,5 \left(\frac{\sum_{j=1}^{d} \frac{|\bar{x}_j - \bar{x}_j^a|}{|\bar{x}_j|}}{d} + \frac{\sum_{j=1}^{d} \frac{|s_{jj} - s_{jj}^a|}{|s_{jj}|}}{d} \right) \right.$$
$$\left. +0,5 \left(\frac{\sum_{j=1}^{d} \sum_{1 \le i \le j} \frac{|s_{ij} - s_{ij}^a|}{|s_{ij}|}}{\frac{1}{2} d\,(d+1)} + \frac{\sum_{j=1}^{d} \sum_{1 \le i < j} |r_{ij} - r_{ij}^a|}{\frac{1}{2} d\,(d-1)} \right) \right) \qquad (15.2)$$

Dieser alternative Vorschlag wird von Domingo-Ferrer et al. (2003); Sebé et al. (2002) und Dandekar et al. (2002) damit begründet, dass der Veränderung der einzelnen Werte, der Veränderung der univariaten Verteilungsmaße und der Veränderung der multivariaten Verteilungsmaße jeweils das gleiche Gewicht zukommen soll. Allerdings zeigen diese beiden alternativen Vorschläge, dass die Gewichtung der einzelnen Abweichungsmaße recht willkürlich erfolgt und hierfür keine objektiv begründbaren Kriterien vorliegen. Dies dürfte auch deshalb schwierig sein, weil für unterschiedliche Nutzungen auch unterschiedliche Abweichungsmaße eine höhere oder geringere Bedeutung haben. Man könnte allenfalls festhalten, dass der Erhalt von Mittelwerten ein eigenständiges Ziel bei der Auswertung von Daten darstellt, während die anderen Maße eher als Hilfsindikatoren für einen möglichst weitgehenden Erhalt der Ergebnisse möglichst vieler anderer Analysen dienen. Dabei könnte man zu der Hypothese gelangen, dass insbesondere die weitgehende Reproduktion der originalen Korrelationsstruktur in den anonymisierten Daten wesentlich für die Reproduktion von Schätzergebnissen in ökonometrischen Modellen ist.

Auch für diskrete Variablen lassen sich Maße in analoger Form berechnen. So kann auch hier die Abweichung der Einzelwerte bestimmt beziehungsweise gezählt werden, wie häufig die Ausprägungen übereinstimmen und wie häufig nicht. Außerdem können die Abweichungen der Zellwerte in Häufigkeitstabellen bestimmt werden. In Kooiman et al. (1997) wird die Verwendung von Shannons Entropie zur Messung des Informationsverlustes diskutiert.

Weiterhin wird vorgeschlagen, einen Score für die Verringerung des Analysepotenzials beziehungsweise den Informationsverlust mit dem über Matching-Verfahren bestimmten Reidentifikationsrisiko zu einem Score zur Bewertung von Anonymisierungsverfahren zusammenzufassen (Domingo-Ferrer et al. 2003; Sebé et al. 2002). Dieser weitergehende Vorschlag ist jedoch nicht praktikabel, weil das Konzept der faktischen Anonymität ein Enthüllungsrisiko brauchbarer Informationen bis zu einer bestimmten Risikoschwelle gestattet (Höhne et al. 2003). Diese Risikoschwelle darf aber in keinem Fall überschritten werden. Andererseits muss auch ein angemessenes Analysepotenzial in jedem Fall sichergestellt sein. Auch wenn im Zweifelsfall Kompromisse zwischen beiden Zielen gemacht werden müssen, stellt ein gegenseitiges Aufrechnen wie im Falle eines gemeinsamen Scores doch eine unerlaubte Vereinfachung des Abwägungsproblems dar. Vielmehr ist die Sicherstellung der faktischen

Anonymität eine Nebenbedingung für die Optimierung des Analysepotenzials bei der Anonymisierung.

Der maßzahlorientierte Ansatz verfolgt das Ziel, die Verringerung des Analysepotenzials durch Anonymisierungsmaßnahmen in Form einer oder mehrerer Maßzahlen quantitativ darzustellen, um so eine operationale Grundlage für die Entscheidung zu schaffen, ob mit einem Anonymisierungsverfahren bearbeitete Daten ein größeres oder ein kleineres Analysepotenzial aufweisen als Daten, die mit einem anderen Verfahren bearbeitet wurden. Der Ansatz ist aus mehreren Gründen problematisch. Zum Ersten existieren keine objektivierbaren Kriterien dafür, welche Maße mit welchem Gewicht bei diesem Ansatz berücksichtigt werden sollen. Zum Zweiten ignoriert der Ansatz die Tatsache, dass nicht nur quantitative Abweichungen, sondern auch die Systematik von Abweichungen und damit möglicherweise ihre Korrigierbarkeit ein Kriterium für die Bewertung der Eingriffe von datenverändernden Verfahren in das Analysepotenzial von Einzeldaten darstellt. Und zum Dritten bleibt auch unklar, ab welchem Wert für eine wie auch immer beschaffene Abweichungsmaßzahl die Anonymisierung aus Sicht des Analysepotenzials eigentlich als zu stark eingeschätzt werden soll. Insofern liefert der Ansatz allenfalls Ansatzpunkte für einen Vergleich verschiedener Anonymisierungsmaßnahmen hinsichtlich ihrer Auswirkungen auf das Analysepotenzial.

15.2 Der anwendungsorientierte Ansatz

Der anwendungsorientierte Ansatz verfolgt die Idee, die Auswirkungen verschiedener Anonymisierungsverfahren auf das Analysepotenzial von Einzeldaten danach zu bewerten, ob und inwiefern sich die Ergebnisse verschiedener Analysen verändern. Dies kann in der Praxis dadurch umgesetzt werden, dass ein repräsentativer Querschnitt an verschiedenen Fragestellungen mit unterschiedlichen statistischen Methoden untersucht wird. Dabei können sowohl die inhaltlichen Fragestellungen als auch die verwendeten Methoden stets nur beispielhaft sein. Der anwendungsorientierte Ansatz stellt damit gewissermaßen eine Weiterentwicklung des maßzahlorientierten Ansatzes dar, indem er die Zahl der relevanten Entscheidungskriterien auf die Ergebnisse statistischer und ökonometrischer Analysen ausdehnt. Die Entscheidungskriterien können dann umfassender sein, gleichzeitig jedoch niemals vollständig. Schließlich kann vom aktuellen Standpunkt aus nicht vollständig überblickt werden, welche möglichen Fragestellungen in der Zukunft mit einem Datenbestand noch untersucht werden können.

Einbezogen werden sollten möglichst unterschiedliche Analysemethoden, insbesondere sowohl deskriptive Auswertungen als auch ökonometrische Schätzungen. Die inhaltlichen Fragestellungen hängen dabei von den Informationen im Datensatz – also von den vorhandenen Variablen – ab.

Der anwendungsorientierte Ansatz bietet damit den Vorteil, dass Fragen, die sich den im vorangegangenen Abschnitt beschriebenen Maßzahlen beziehungsweise einem Score entzie-

hen, mit Hilfe dieses Ansatzes beantwortet werden können. Im Unterschied zum maßzahlorientierten Ansatz können mit Hilfe dieses Ansatzes sowohl die inhaltlichen Einschränkungen an Analysemöglichkeiten durch die Verringerung von Informationen im Datensatz als auch die sich durch datenverändernde Verfahren ergebenden Veränderungen von Analyseergebnissen beurteilt werden.

Allerdings weist auch dieser Ansatz gravierende Nachteile auf. Neben dem bereits beschriebenen Problem, dass alle getesteten Auswertungen immer nur beispielhaft sind und bei weitem nicht alle möglichen Auswertungen berücksichtigt werden können, entbindet auch der anwendungsorientierte Ansatz nicht von der Aufgabe, die Abweichungen der beispielhaften Auswertungen zu bewerten. Zuletzt werden auch in diesem Ansatz die Möglichkeiten der Korrektur oder der Verwendung alternativer Verfahren vernachlässigt.

15.3 Der theorieorientierte Ansatz

Während die beiden empirischen Ansätze davon ausgehen, dass eine größtmögliche Erhaltung des Analysepotenzials auch gleichzeitig eine größtmögliche Gleichheit der empirischen Ergebnisse von Analysen beziehungsweise die weitgehende Erhaltung von Verteilungsmaßen bedeutet, besteht der theorieorientierte Ansatz darin, Anonymisierungsverfahren danach zu bewerten, inwieweit ihre Auswirkungen auf die Eigenschaften der Schätzer in ökonometrischen Modellen oder auch deskriptiver statistischer Maße theoretisch unverändert beziehungsweise erwartungstreu oder korrigierbar sind. Hierzu werden die Auswirkungen der datenverändernden Verfahren für verschiedene Modelle oder Maße theoretisch abgeleitet (vgl. z.B. Lechner und Pohlmeier (2003, 2004) und Ronning (2005)).

Grundvoraussetzung für die Anwendung des Ansatzes ist, dass sich das datenverändernde Verfahren formal so darstellen lässt, dass sich seine Auswirkungen auch theoretisch ableiten lassen. Der Ansatz baut auf den Fehler-in-den-Variablen-Modellen auf (siehe hierzu beispielsweise Fuller (1980, 1984); Hwang (1986); Lin (1989)) oder auch auf Modellen für Messfehler in nichtlinearen Modellen (Carroll et al. 1995). Diese Arbeiten lassen sich für das Verfahren der stochastischen Überlagerung anwenden (Lechner und Pohlmeier 2003, 2004).

Lechner und Pohlmeier (2003) untersuchen die Auswirkungen von stochastischen Überlagerungen und stochastischen Mikroaggregationsverfahren im linearen Regressionsmodell. In Rosemann (2004) erfolgt eine theoretische Betrachtung der Auswirkungen von Mikroaggregationsverfahren im linearen Regressionsmodell. Schmid et al. (2005) untersuchen die Wirkung einer abstandsorientierten Mikroaggregation nach der abhängigen Variablen ebenfalls im linearen Regressionsmodell. Lechner und Pohlmeier (2004) analysieren die Möglichkeiten der Korrektur von Schätzungen bei stochastischen Überlagerungen in nichtlinearen Modellen, Lechner und Pohlmeier (2005) in nicht parametrischen Modellen. In Ronning (2005), Ronning und Rosemann (2004) sowie Ronning et al. (2005) werden die Auswirkungen der

Post-Randomisierung der binären abhängigen Variablen auf das Probit-Modell theoretisch abgeleitet. Ronning (2004a) schließlich analysiert die Wirkung der Post-Randomisierung einer Dummy-Variablen auf die Varianz- und Kovarianzanalyse.

Der Vorteil des theorieorientierten Ansatzes besteht darin, dass eine objektive Bewertung der datenverändernden Anonymisierungsverfahren dahingehend erfolgen kann, ob die Verfahren theoretisch zu keiner Veränderung bestimmter Analyseergebnisse führen beziehungsweise erwartungstreu sind oder sich die Erwartungstreue durch Korrekturverfahren wieder herstellen lässt. Allerdings ist eine theoretische Untersuchung einer Vielzahl von Modellen und deskriptver Maße erforderlich, um eine umfassende Bewertung auf Basis dieses Ansatzes abgeben zu können. Zudem können selbst korrigierte Verfahren, die in der Theorie zu erwartungstreuen Schätzern führen, in der Praxis bei endlichen Stichproben einen verzerrten Schätzer erzeugen. Zuletzt ist dieser Ansatz auch nicht auf alle datenverändernden Methoden anwendbar. So sollten die Schätzergebnisse bei Anwendung von Simulationsverfahren grundsätzlich erhalten bleiben. Eine Modellierung von Swapping-Verfahren in ökonometrischen Modellen ist beispielsweise nicht möglich. Auch dies könnte ein Grund dafür sein, ein solches Anonymisierungsverfahren nicht einzusetzen.

Teil VI

Grundsätzliches Vorgehen beim Verfahrensvergleich und bei der Erstellung von Scientific-Use-Files

Auf Basis des entwickelten Schutzwirkungskonzepts und der vorgenommenen Operationalisierung des Analysepotenzials wird in diesem Teil das Vorgehen bei der Beurteilung von Anonymisierungsverfahren hinsichtlich der beiden Kriterien Schutz und Analysepotenzial beschrieben. Dabei wird zunächst in Kapitel 16 das theoretisch optimale Vorgehen zur Erreichung beider Ziele sowie die dabei in der Praxis auftretenden Probleme dargestellt. Anschließend werden in den Kapiteln 17 und 18 die konkreten Vorgehensweisen bei der Untersuchung der Schutzwirkung beziehungsweise des Analysepotenzials anonymisierter Daten dargestellt, die im Projekt eingeschlagen wurden.

Kapitel 16

Das theoretisch optimale Vorgehen und Probleme in der Praxis

Die Erstellung eines Scientific-Use-Files soll dazu dienen, die Datennutzer mit „möglichst guten“ Daten zu versorgen. Allerdings sind dabei folgende Nebenbedingungen zu beachten: Einerseits muss die faktische Anonymität gewährleistet sein. Andererseits muss ein möglichst großes Analysepotenzial der Daten für möglichst viele Datennutzer erhalten bleiben. Theoretisch optimal ist es daher, die Anonymisierungsmaßnahmen so zu wählen, dass das Analysepotenzial maximiert wird unter der Nebenbedingung, dass die faktische Anonymität gerade noch erreicht wird. Während jedoch das Enthüllungsrisiko eines Datensatzes als eindimensionales Zielkriterium darstellbar ist (vgl. hierzu Teil IV dieses Handbuchs), ist das Analysepotenzial aufgrund der Vielfältigkeit der zu beachtenden Aspekte (vgl. Teil V) und der unterschiedlichen Anforderungen durch die verschiedenen Nutzergruppen ein mehrdimensionales Zielkriterium. Dabei kann wegen der unterschiedlichen Bewertung der Nutzer keine verbindliche Rangfolge der Ziele festgelegt werden. Zudem sind die Effekte von Anonymisierungsmaßnahmen auf das Analysepotenzial häufig nicht quantifizierbar oder es bestehen gar widersprüchliche Bewertungen aus der Sicht unterschiedlicher Gruppen von potenziellen Datennutzern.

Um dies zu illustrieren, sei folgendes Alternativ-Szenario angedacht: Für eine bestimmte Erhebung sei die faktische Anonymisierung alternativ entweder durch Elimination jeglicher regionalspezifischer Identifikatoren oder durch Elimination jeglicher branchenspezifischer Identifikatoren zu erreichen. Natürlich wird der Regionalwissenschaftler eher die zweite Variante und beispielsweise der industrieökonomisch orientierte Forscher eher die erste Variante präferieren. Damit ist bereits angedeutet, dass es in der Praxis „den“ optimalen Weg zur Erzeugung eines Scientific-Use-Files selbst dann nicht gibt, wenn man sich auf eine bestimmte Erhebung beschränkt.

Ferner soll das erzeugte Datenfile in unterschiedlichen Fragestellungen und vor allem bei Anwendung unterschiedlicher Methoden einsetzbar sein. Wenn ein Forscher eher statistisch-deskriptive Analysen für einzelne Merkmale durchführen möchte, dann wird es ihn über-

haupt nicht beschäftigen, wenn die Daten beispielsweise mittels Rank Swapping anonymisiert werden, da die (Rand-)Verteilung der Beobachtungswerte dadurch überhaupt nicht berührt wird, zumindest dann nicht, wenn die Analyse für die Gesamtheit und nicht für Teilgesamtheiten angestellt wird. Sobald jedoch der Zusammenhang zwischen verschiedenen Merkmalen, beispielsweise bei der Schätzung von stochastischen Modellen wie dem Regressionsmodell betrachtet wird, ist diese Methode aus Nutzersicht nicht mehr akzeptabel, weil die Zusammenhänge zwischen den Merkmalen stark verändert werden. Die genannte Methode wurde deshalb auch aus der Reihe der sinnvollen Anonymisierungsmethoden ausgeschlossen (siehe dazu Kapitel 7).

Da bei Erzeugung des Scientific-Use-Files nicht klar ist, ob ein bestimmtes Merkmal beispielsweise in einem Regressionsmodell als abhängige d.h. zu erklärende Variable oder als unabhängiger Regressor auftritt, müssen die Anonymisierungsverfahren so gewählt werden, dass sich in beiden Fällen zufriedenstellende Schätzungen ergeben, die nicht zu weit von den Schätzungen auf Basis der Originaldaten abweichen. Wie im Einzelnen beispielsweise in Teil VIII dargelegt wird, lässt sich für das Regressionsmodell analytisch zeigen, dass die Verfremdung der Regressoren mit einem formalen Anonymisierungsverfahren wie der stochastischen Überlagerung deutlich problematischer ist als die entsprechende Manipulation der abhängigen Variablen. Auf der anderen Seite hat im Fall der Mikroaggregation die Verfremdung der abhängigen Variablen deutlich stärkere negative Effekte auf die Schätzergebnisse.

Es kommt hinzu, dass durch die Möglichkeit, verschiedene Anonymisierungsverfahren miteinander zu kombinieren bzw. die Notwendigkeit, bei gemeinsamer Anonymisierung von metrischen und kategorialen Merkmalen einen Methodenmix einzusetzen, die Festlegung eines optimalen Verfahrens ein hochdimensionales Entscheidungsproblem ist. Somit ist auch die Erstellung eines effizienten Algorithmus zur systematischen Suche nach dem optimalen Anonymisierungsverfahren in der Praxis nicht möglich.

Die bisherigen Ausführungen sollen vor allem einen ersten Eindruck von der Komplexität des Prozesses zur Erzeugung eines Scientific-Use-Files, d.h. eines aus Nutzersicht akzeptablen bzw. forschungsrelevanten und gleichzeitig faktisch anonymen Files, geben. Wie in den beiden folgenden Kapiteln gezeigt wird, ist die Bewertung aus Sicht der Datenanbieter anders als aus Sicht der Datennutzer. Gleichwohl müssen beide in einem diskursiven, interaktiven Prozeß versuchen, ein Datenfile zur Verfügung zu stellen, das die Schutzwirkungsanforderungen – gerade – erfüllt und gleichzeitig noch ein möglichst hohes Analysepotenzial aufweist. Das Vorgehen bei einem solchen Prozess ist in Abbildung 16.1 dargestellt.

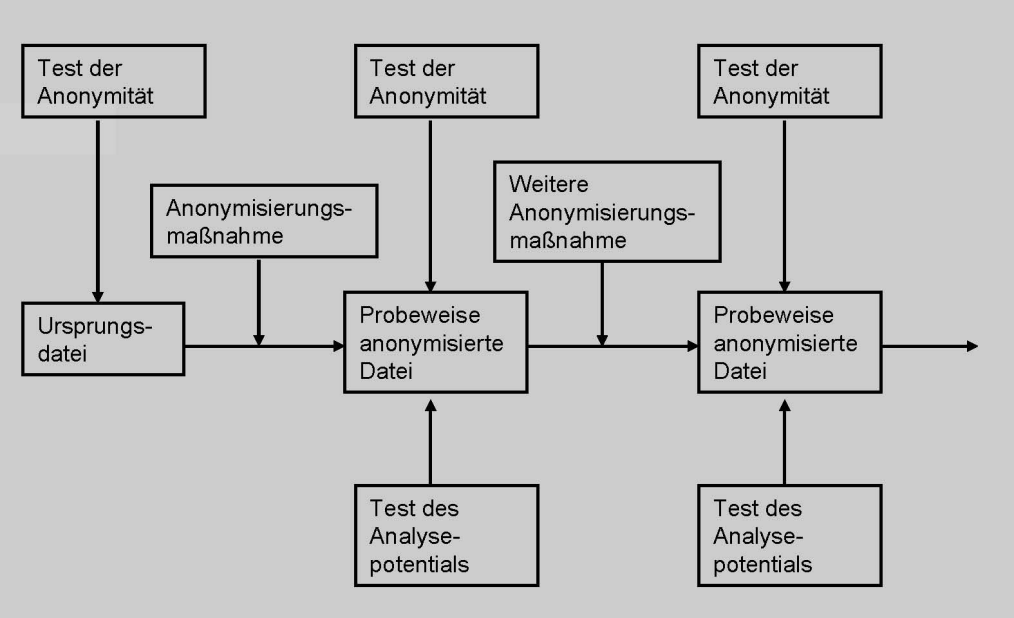

Abbildung 16.1: Prozess zur Erstellung eines Scientific-Use-Files

Kapitel 17

Untersuchung der Schutzwirkung anonymisierter Daten

Zu Beginn der Untersuchung sollte eine ausführliche Recherche über das mögliche Zusatzwissen eines potenziellen Datenangreifers durchgeführt werden (zum Vergleich siehe Abschnitt 11.3). Bereits hier können kritische, bei Datenangriffen besonders gefährdete Bereiche in den Daten aufgedeckt werden. Diese sind zum Teil den Fachleuten schon vor der Zusatzwissenrecherche bekannt (z.B. dünne Besetzungszahlen in Tabellen einer Fachserie). Des Weiteren sollten sich die Fachleute der Erhebung und der Anonymisierung zusammensetzen und sich auf Schwellen für die Brauchbarkeit von Einzelangaben (so genannte Nützlichkeitsschwellen) sowie eine obere Risikoschwelle verständigen.

In einem nächsten Schritt sollte versucht werden, eine Datenbank als Zusatzwissen für Massenfischzugsimulationen aufzubauen. Da im Allgemeinen keine gemeinsamen Identifikatoren zwischen Daten verschiedener Erhebungen vorliegen, ist hier auch seitens der Datenanbieter mit viel Aufwand zu rechnen (siehe zum Beispiel Lenz et al. (2004a)). In den meisten Fällen liegen in beiden Dateien Merkmale über Namen und Adressen vor, die aber mitunter stark in den beiden Quellen differieren können.

Ist Zusatzwissen in Form einer Datenbank vorhanden, kann ein Massenfischzug wie in Abschnitt 11.2 beschrieben simuliert werden. Ein Programm hierzu kann über das Forschungsdatenzentrum des Statistischen Bundesamtes bezogen werden. Dieses Programm ist selbstverständlich für einen potenziellen Datenangreifer unbrauchbar, da dieser keine Möglichkeit hat, die durch das Programm getroffenen Zuordnungen auf Korrektheit zu überprüfen.
Es ist sinnvoll, den Massenfischzug zunächst mit formal anonymisierten Daten (die aus den Originaldaten durch Entfernen direkter Identifikatoren wie Name, Adresse und Unternehmensnummer entstehen) durchzuführen. Auf diese Weise werden der natürliche Schutz in den Daten sowie weitere gefährdete Bereiche sichtbar, auf welche nun Einzelangriffe dürchgeführt werden sollten. Dies kann aus Sicht des Datenanbieters sehr zeitaufwändig sein (vgl. Abschnitt 11.2).

Bei der Entwicklung eines Anonymisierungskonzeptes – mit besonderem Fokus auf die zuvor

aufgedeckten gefährdeten Bereiche – erscheint oftmals eine Mischung aus informationsreduzierenden und datenverändernden Methoden als beste Lösung. Informationsreduzierende Methoden wie die Entfernung von Merkmalen, die für den Datennutzer nicht von Belang sind oder z.B. eine (weitere) Vergröberung kategorialer Überschneidungsmerkmale in der Form, dass die den Wissenschaftler interessierenden Teilmassenauswertungen weiterhin möglich sind, sollten bevorzugt angewendet werden, da hier Informationen nicht verfremdet, sondern lediglich unterdrückt werden. Auf diese Weise können mögliche, für den Datenangreifer nahezu unverzichtbare Blockmerkmale (zur Definition siehe Unterabschnitt 13.1.2) entschärft werden, da diese nach der Vergröberung die Daten weniger fein partitionieren. In Datenbereichen mit verhältnismäßig dichter Besetzung kann eine Anonymisierung allein mit informationsreduzierenden Methoden (wie z.B. bei Unternehmen der Einzelhandelsstatistik mit maximal 49 Beschäftigten, siehe Kapitel 37) oder mit geringfügiger Datenveränderung (z.B. bei Unternehmen der Kostenstrukturerhebung im Verarbeitenden Gewerbe mit maximal 49 Beschäftigten, siehe Kapitel 35) gelingen. In Bereichen mit dünner Besetzung sind dagegen datenverändernde Maßnahmen in der Regel unvermeidlich, wie z.B. bei marktführenden Unternehmen der Umsatzsteuerstatistik in bestimmten Branchen (siehe Kapitel 36).

Hat man sich bei den datenverändernden Anonymisierungsverfahren auf eine Verfahrensgruppe verständigt, so müssen im Folgenden die Parameter des Verfahrens ausbalanciert und Datenangriffe so lange simuliert werden, bis der gewünschte Anonymisierungsgrad erreicht und die vorgegebene obere Risikoschwelle (siehe Abschnitt 12.3) unterschritten wird. Wie im vorherigen Kapitel 16 beschrieben wurde, sollten diese Untersuchungen nicht isoliert von der Bewertung des Analysepotenzials durchgeführt werden. In die Wahl der Parameter des ausgewählten Anonymisierungsverfahrens sollten simultan die Untersuchungen hinsichtlich des Analysepotenzials, wie im nachfolgenden Kapitel 18 ausgeführt wird, eingehen.

Abbildung 17.1 beinhaltet zusammengefasst das Vorgehen bei der Schutzwirkung anonymisierter Daten.

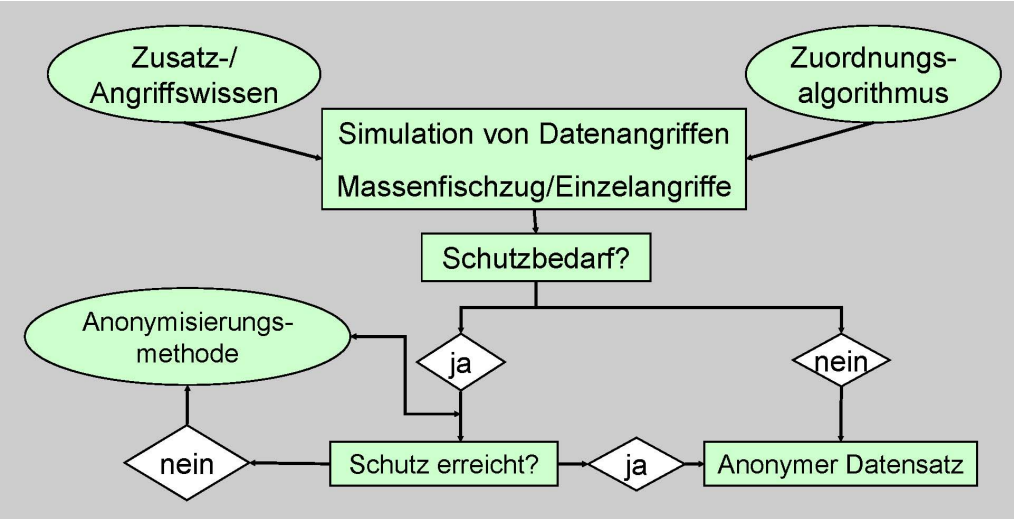

Abbildung 17.1: Vorgehen bei der Untersuchung der Schutzwirkung anonymisierter Daten

Kapitel 18

Untersuchung des Analysepotenzials anonymisierter Daten

Aus den in Kapitel 15 vorgestellten Ansätzen zur Beurteilung der Veränderung des Analysepotenzials von Einzeldaten durch Anonymisierungsverfahren ergibt sich kein eindeutig zwingendes Vorgehen für die Beurteilung von Anonymisierungsverfahren aus der Sicht des Analysepotenzials. In Abschnitt 14.2 wurde jedoch herausgearbeitet, dass sich die Wirkung von Anonymisierungsverfahren in Einschränkungen hinsichtlich der inhaltlichen Analysemöglichkeiten und potenziellen Ergebnisveränderungen systematisieren lässt. Daraus folgt, dass auch die Beurteilung von Anonymisierungsverfahren hinsichtlich dieser beiden Fragestellungen vorgenommen werden kann.

Dabei muss zunächst in einem diskursiven Prozess zwischen der datenbereitstellenden Institution und dem Kreis der potenziellen Datennutzer Einverständnis über das Analysespektrum und die daraus folgenden Anforderungen an die Informationen in der Datei hergestellt werden, da sich diese Fragen einer Beurteilung durch objektive Kriterien entziehen. Beispielsweise ist also zu klären, in welcher Tiefengliederung Wirtschaftszweige und Regionalinformation in einem Scientific-Use-File enthalten sein müssen. Solche Diskussionen müssen für jede als Scientific-Use-File zur Verfügung zu stellende Datei neu geführt werden.

Im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ wurden diese Fragen in einem Diskussionsprozess zwischen den Statistischen Ämtern des Bundes und der Länder, dem IAW und dem Wissenschaftlichen Begleitkreis für die einzelnen Projektstatistiken geklärt. Dabei wurde für alle Projektstatistiken vereinbart, dass eine regionale Unterscheidung nach Ost- und Westdeutschland nichtadministrativen Gebietsgliederungen des Bundesamts für Bauwesen und Raumordnung (BBR) vorzuziehen ist, insbesondere wenn aus Sicherheitsgründen lediglich die Regionsgrundtypen (drei siedlungsstrukturelle Typen auf der Ebene der Regionen)[19] zur Verfügung gestellt werden können. Hinsichtlich

19) Das Bundesamt für Bauwesen und Raumordnung (BBR) erstellt siedlungsstrukturelle Kreistypen nach dem Grad der Verdichtung der Stadt- und Landkreise. Hierfür werden zunächst die Regionen in drei siedlungsstrukturelle Regionstypen, die so genannten Grundtypen (BBR3), eingeteilt. Die Kreise innerhalb dieser drei Grundtypen werden weiter unterteilt, so dass sich insgesamt neun siedlungsstrukturelle

der Tiefengliederung der Wirtschaftszweige wurden unterschiedliche Vorgehensweisen für die einzelnen Projektstatistiken vereinbart. Für das Verarbeitende Gewerbe ist eine Differenzierung nach den Zweistellern der Wirtschaftszweigklassifikation WZ 93 ausreichend, während für den Dienstleistungsbereich und damit für Teile der Umsatzsteuerstatistik eine tiefere Gliederung notwendig ist. Beispielsweise gehören zum Wirtschaftszweig 74 nach der Gliederung WZ 93 sowohl Rechts-, Steuer- und Unternehmensberatungen (74.1) als auch Architektur- und Ingenieurbüros (74.2), Detekteien und Schutzdienste (74.6) oder Reinigung von Gebäuden, Inventar und Verkehrsmitteln (74.7). Der Zweisteller 93 umfasst sowohl Wäschereien und chemische Reinigungen (93.01) als auch Friseurgewerbe und Kosmetiksalons (93.02), Bestattungswesen (93.03) und Bäder, Saunas und Solarien (93.04). Diese Auflistung macht deutlich, dass für diesen Zweisteller aus Analysesicht eine tiefere Gliederung erforderlich ist. Da die Einzelhandelsstatistik nur den Zweisteller 52 erfasst, ist hier ohnehin eine größere Gliederungstiefe nötig. (Die Einzelheiten des Vorgehens sind in Teil XI für die einzelnen Projektstatistiken beschrieben).

Wenn man mit der Einschränkung von Informationen allein keine faktische Anonymität sicherstellen kann oder die hierfür notwendigen Maßnahmen so weitgehend sind, dass sie von den Nutzern nicht toleriert würden, dann wird zusätzlich die Anwendung datenverändernder Verfahren notwendig, die sich je nach Sicherheitslage und Verfahren auf alle Variablen erstrecken oder ausschließlich auf die Überschneidungsmerkmale beschränken können (vgl. hierzu Teil X).

Werden solche datenverändernden Verfahren angewendet, so ist eine Bewertung in zweierlei Hinsicht notwendig:

- Relative Bewertung: Ist ein datenveränderndes Verfahren besser als ein anderes?
- Absolute Bewertung: Sind die Auswirkungen auf die Analysen, die durch ein datenveränderndes Verfahren hervorgerufen werden, für die Nutzer akzeptabel oder nicht?

Als Entscheidungsgrundlage dienen sowohl theoretische Erkenntnisse über die Wirkungsweise verschiedener datenverändernder Anonymisierungsverfahren in unterschiedlichen Analysen, ergänzt um die Ergebnisse von Simulationsexperimenten, als auch Ergebnisse empirischer Berechnungen.

Die theoretischen Erkenntnisse gelten unabhängig von der jeweils zu anonymisierenden Datengrundlage. Aus empirischen Beispielrechnungen können zwar ebenfalls allgemein gültige Schlüsse über die Wirkungsweise von datenverändernden Anonymisierungsverfahren gezogen werden, dennoch sind die Wirkungsweisen von Verfahren besonders von der jeweiligen Datengrundlage abhängig. Daraus ergibt sich ein zweistufiges Verfahren zur Beurteilung der Eignung von datenverändernden Anonymisierungsverfahren.

Kreistypen ergeben (BBR9). Davon unabhängig existiert eine weitere Unterteilung der Kreise in sieben siedlungsstrukturelle Kreistypen (BBR7).

A) Beurteilung, welche datenverändernden Verfahren grundsätzlich, beziehungsweise für bestimmte Anwendungen, akzeptabel sind.

B) Beurteilung, ob ein konkreter faktisch anonymisierter Datensatz aus Analysesicht als Scientific-Use-File freigegeben werden kann.

Dabei wird für deskriptive Auswertungen wie auch für ökonometrische Modelle geprüft, ob

1. Ergebnisveränderungen von vornherein ausgeschlossen beziehungsweise die Ergebnisse erwartungstreu sind,
2. Ergebnisveränderungen bestimmter Maßnahmen im Rahmen der Analyse korrigiert werden können,
3. Ergebnisveränderungen sich in einem für den Nutzer akzeptablen Rahmen bewegen.

Die Fragen 1. und 2. können mit Hilfe theoretischer Überlegungen beantwortet werden. Allerdings beziehen sich die theoretischen Ergebnisse häufig (insbesondere bei stochastischen Überlagerungen) auf den asymptotischen Fall. Somit können – insbesondere bei endlichen Stichproben – auch theoretisch unverzerrte Schätzer Verzerrungen aufweisen, die sich in einer Größenordnung bewegen, die auch von anderen – nicht theoretisch optimalen – datenverändernden Verfahren erzeugt werden. Es kommt hinzu, dass kein Verfahren für alle denkbaren Analysemethoden zu theoretisch optimalen Ergebnissen führt. Da jedoch die Bereitstellung von Scientific-Use-Files mit einer möglichst großen Bandbreite an Nutzungsmöglichkeiten angestrebt wird, sollte am Ende ein Verfahren gewählt werden, das für möglichst viele Auswertungsmethoden nur zu geringen Ergebnisabweichungen führt.

Deshalb werden für die empirischen Auswertungen Kriterien entwickelt, nach denen die Verfahren ergänzend zu den theoretischen Erkenntnissen beurteilt und nach denen Verfahren miteinander verglichen werden können. Da die Anzahl der möglichen Analysemethoden groß ist, müssen sich die empirischen Tests und damit auch die zur Beurteilung herangezogenen Kriterien auf eine Auswahl der wichtigsten deskriptiven und inferenzstatistischen Auswertungen beschränken.

Für die Beispielschätzungen linearer und nichtlinearer Modelle sind die folgenden Kriterien wesentlich:

- Der annähernde Erhalt der Koeffizientenwerte (als Richtgröße wird eine maximale Abweichung von 10 Prozent angenommen. Zur Begründung siehe unten.)
- Der Ausschluss von Vorzeichenveränderungen bei den Koeffizienten.
- Die vergleichbare Signifikanz der Einflüsse (signifikant zum gleichen Signifikanzniveau) und damit eine vergleichbare Größenordnung bzw. Aussagequalität der Teststatistiken.

Während die Einhaltung dieser Kriterien stets nur für eine bestimmte Modellspezifikation geprüft werden kann, können Kriterien für deskriptive Maße die Veränderungen im Datensatz als Ganzes abbilden. Dabei werden folgende Kriterien betrachtet:

- Der Erhalt struktureller Nullen und von Vorzeichen,
- annähernder Erhalt der Häufigkeitsauszählungen für die nach kategorialen Variablen gebildeten Teilgesamtheiten,
- der Erhalt der Mittelwerte (arithmetische Mittel und Mediane) im Gesamtdatensatz sowie für abgrenzbare Teilgesamtheiten,
- die Abweichung der Standardabweichungen im Gesamtdatensatz sowie für abgrenzbare Teilgesamtheiten,
- die Abweichung der Korrelationskoeffizienten im Gesamtdatensatz.

Die Veröffentlichung dieser Kriterien soll den Nutzern damit zugleich einen fundierten Einblick in die möglichen Abweichungen von Analyseergebnissen geben. Um eine Einschätzung über die Brauchbarkeit anonymisierter Daten zu erhalten, werden Richtgrößen (Abweichungsschwellen) definiert, ab denen ein anonymisierter Datensatz hinsichtlich der jeweiligen Auswertung als problematisch eingeschätzt wird. Eigentlich müssten solche Richtgrößen abhängig vom konkreten Forschungsinteresse sowie der Skalierung und Verteilung der interessierenden Variable gewählt werden. Wenn man jedoch allgemeingültige Handlungsempfehlungen für die Anonymisierung von wirtschaftsstatistischen Einzeldaten entwickeln will, so ist eine Verständigung auf einheitliche Richtgrößen beziehungsweise Abweichungsschwellen notwendig. Dies scheint in Anbetracht der Beschaffenheit der Projektstatistiken und der Skalierung der dort erhobenen Variablen auch möglich zu sein. Ein solches Vorgehen entbindet aber nicht davon, bei zukünftigen Anonymisierungsvorhaben zu prüfen, ob die gewählten Kriterien auch auf die dann zu anonymisierenden Variablen anwendbar sind oder gegebenenfalls die hier gewählte Konvention verlassen werden sollte.

Mittelwerte:

- Es wird vorgeschlagen, für die Beurteilung der Brauchbarkeit von Mittelwerten aus der Sicht des Analysepotenzials die gleichen Abweichungsschwellen zu wählen, die für die Brauchbarkeit von Einzelwerten beim Datenangriff definiert wurden. Im Rahmen der Anwendung des Schutzkonzepts auf die Projektstatistiken wurde festgelegt, dass ein Einzelwert für einen Datenangreifer dann einen wertvollen Nutzen darstellt, wenn er nicht um mehr als 10 Prozent vom Originalwert abweicht. Es wird daher auch für die Auswertungen von arithmetischen Mitteln und Medianen eine maximale Abweichung der Mittelwerte von 10 Prozent angestrebt.
- Die Gleichheit der Mittelwerte von Teilgesamtheiten wird ferner mit t-Tests auf Mittelwertgleichheit für ein Signifikanzniveau von 10 Prozent überprüft.

Standardabweichungen:

- Die meisten datenverändernden Anonymisierungsverfahren führen zu einer systematischen Verzerrung der Streuungsmaße. Der weitestgehende Erhalt der Streuung ist jedoch von Bedeutung insbesondere für die Durchführung von inferenzstatistischen Analysen. Darüber hinaus sind auch die Standardabweichungen von Teilgesamtheiten für Streuungsvergleiche von Interesse. Es wird daher angestrebt, die durch datenverändernde Verfahren hervorgerufene Abweichung der Standardabweichungen sowohl im Gesamtdatensatz als auch in den durch kategoriale Merkmale abgrenzbaren Teilgesamtheiten ebenfalls auf 10 Prozent zu beschränken.

Korrelationen:

- Die Zusammenhänge zwischen den Variablen spielen insbesondere eine Rolle für die Ergebnisse ökonometrischer Schätzungen. Häufig wird der Modellspezifikation zudem eine Untersuchung der Zusammenhangsmaße in Form von Korrelationskoeffizienten vorgeschaltet. Die Berechnung von Korrelationskoeffizienten geht auch in andere statistische Analysemethoden, wie z.B. die Faktorenanalyse mit ein. Daneben kann auch ein eigenständiges Forschungsinteresse an deskriptiven Zusammenhangsuntersuchungen bestehen.
- Da die Korrelationskoeffizienten bereits normiert sind und hier häufig sehr kleine Werte beobachtet werden, wird die absolute Abweichung betrachtet und nicht die relative.
- Bei den Korrelationskoeffizienten wird angestrebt, eine absolute Abweichung von 0,10 nicht zu überschreiten.
- Bei den Rangkorrelationen sollten die absoluten Abweichungen den Wert von 0,05 nicht überschreiten.
- Wesentlich ist, dass es weder bei Korrelationskoeffizienten noch Rangkorrelationen zu einem Vorzeichenwechsel kommt.

Da ein Überschreiten der Richtgrößen/Abweichungsschwellen in Einzelfällen nicht zu verhindern sein dürfte, wird angestrebt, den Anteil der Fälle, bei denen eine Überschreitung der Richtgrößen beobachtet wird, auf 10 Prozent zu begrenzen. Wird dieser Anteil überschritten und ist keine alternative Anonymisierung möglich, so müssen die Datennutzer im Rahmen der Datenbeschreibung (Metadaten) über diese Überschreitung und auf damit verbundene mögliche Probleme bei der Datennutzung hingewiesen werden.

Für die verschiedenen Konzentrationsmaße wurden keine Abweichungsschwellen festgelegt. Konzentrationsuntersuchungen wurden jedoch als beispielhafte Auswertungen durchgeführt.

Sind die unterschiedlichen Anforderungen, die deskriptive und ökonometrische Auswertungen an einen anonymisierten Datensatz nach den benannten Kriterien stellen, nicht zu vereinbaren, so sollten gegebenenfalls mehrere Scientific-Use-Files für unterschiedliche Datennutzungen bereitgestellt werden. Auf grundsätzliche Einschränkungen bei der Datennutzung ist ebenfalls in der Datensatzbeschreibung hinzuweisen.

Teil VII

Wirkung datenverändernder Anonymisierungsmethoden auf deskriptive Auswertungen

Dieser Teil beschäftigt sich mit den Auswirkungen datenverändernder Anonymisierungsverfahren auf verschiedene deskriptive Verteilungsmaße. Dargestellt werden sowohl theoretische Herleitungen als auch Praxisbeispiele unter Verwendung der Projektdaten. Zunächst beschäftigt sich Kapitel 19 mit den Auswirkungen von stochastischen Überlagerungen und Mikroaggregationsverfahren auf verschiedene deskriptive Verteilungsmaße. Anschließend behandelt Kapitel 20 die Auswirkungen der Post-Randomisierung auf deskriptive Auswertungen.

Kapitel 19

Auswirkungen von Mikroaggregationsverfahren und stochastischen Überlagerungen auf deskriptive Auswertungen

19.1 Theoretische Untersuchungen

19.1.1 Einleitung

In diesem Abschnitt werden die Auswirkungen von Mikroaggregationsverfahren sowie der stochastischen Überlagerung auf verschiedene deskriptive Maße theoretisch untersucht. Dabei werden die Anonymisierungsverfahren

- Deterministische Migroaggregationsverfahren
- Zufällige Migroaggregation
- Additive stochastische Überlagerungen
- Multiplikative stochastische Überlagerungen

dargestellt sowie ihre Auswirkungen auf die deskriptiven statistischen Maße

- Mittelwerte und Quantile von Variablen
- Streuungsmaße
- Korrelationskoeffizienten
- Konzentrationsmaße.

Folgende theoretische Annahmen werden generell zugrunde gelegt: Endliche Grundgesamtheiten von N Einheiten z.B. von Unternehmen, Betrieben, Haushalten sowie Personen etc., jeweils mit ihren diversen, metrisch skalierten m Merkmalen und deren beobachtbaren Werten, liegen vor. Stichproben des Umfanges n wurden für die m Merkmale aus diesen endlichen Grundgesamtheiten gezogen. Diese beobachteten Werte der Stichproben werden als m Daten-Vektoren der Länge n in einer Datenmatrix $\mathbf{X} = \{x_{i,j} | \, i = 1, ..., n \quad ; \quad j = 1, ..., m\}$ zusammengefasst.

19.1.2 Auswirkungen der Mikroaggregation auf das arithmetische Mittel

Bei den deterministischen Mikroaggregationsverfahren wird in der Regel ein dreigliedriger Durchschnitt auf die nach Abstandskriterien zueinander sortierten Originaldaten angewendet (zum Vergleich siehe Unterabschnitt 6.2.4). Anstelle eines 3-er Durchschnitts kann bei Bedarf auch ein k-gliedriger Durchschnitt ($k \geq 3$) verwendet werden. Im Folgenden wird, wenn nichts anderes vermerkt, auf diesen Standardfall ($k = 3$) rekurriert. Es wird vereinfacht angenommen, dass $n \equiv 0 \; mod \; 3$ gilt.

Der sortierte Spaltenvektor $\mathbf{x}$ wird durch Mikroaggregation in $\mathbf{x}^a$ transformiert, d.h. für die Komponenten von $\mathbf{x}^a$ gilt dann:

$$x_i^a = x_{i+1}^a = x_{i+2}^a = \frac{1}{3}(x_i + x_{i+1} + x_{i+2}) \qquad (i = 3k+1; \; k = 0, 1, ...). \tag{19.1}$$

Es findet folgende Fehlerüberlagerung der größensortierten Originalwerte statt:

$$x_{i+k}^a = x_{i+k} + e'_{i+k} \qquad (i = 3k+1; \; k = 0, 1, ...) \qquad (j = 0, 1, 2) \tag{19.2}$$

mit daraus folgenden Überlagerungen für $e_i \quad (i = 3k+1; \; k = 0, 1, ...)$. Die Indizierung für j ist aufgelöst, $e'_{(i+j)}$ geht in die e_i über:

$$\begin{aligned} e_i &= \frac{1}{3}(-2x_i + x_{i+1} + x_{i+2}) \\ e_{i+1} &= \frac{1}{3}(x_i - 2x_{i+1} + x_{i+2}) \\ e_{i+2} &= \frac{1}{3}(x_i + x_{i+1} - 2x_{i+2}) \end{aligned} \tag{19.3}$$

Daraus folgt für die Fehlersumme:

$$\sum_{i=1}^{n} e_i = 0$$

und damit für die arithmetischen Mittelwertvektoren von $\mathbf{x}$, $\mathbf{x}^a$:

$$\bar{\mathbf{x}} = \bar{\mathbf{x}}^a.$$

Die Mikroaggregation ist somit mittelwerterhaltend für jeden Datenvektor bei Verwendung des arithmetischen Mittels. Das ist unabhängig von der zugrunde gelegten Sortierung beziehungsweise der konkreten Variante der Mikroaggregation.

19.1.3 Auswirkungen der Mikroaggregation auf den Median und andere Quantile

Der Median ist nicht invariant gegenüber der Mikroaggregation. In welchem Maße er verändert wird, hängt, wie oben ausgeführt, von der konkreten Konstellation der dem Median der Originaldaten benachbarten Originalwerte ab. Da ein potenzieller Nutzer anonymisierter Mikrodaten die Originalwerte nicht kennt, ist er auf das Intervall $[x^a_{(n-4)/2}; x^a_{(n+4)/2}]$für gerades n beziehungsweise $[x^a_{(n-3)/2}; x^a_{(n+5)/2}]$für ungerades n zur approximativen Abschätzung angewiesen.

Für weitere deskriptive Maße, wie Quantile und Perzentile, sind die entsprechenden Intervalle innerhalb der empirischen Häufigkeitsdarstellung der anonymisierten Werte für eine Abschätzung heranzuziehen. Die eingetretenen Verschiebungen dieser Größen durch Mikroaggregation sind analog zu denen für den Median zu beurteilen.

Betrachtet man die Auswirkungen der gemeinsamen oder gruppierten Mikroaggregation auf den Median, so hängt die Übertragbarkeit dieser Ergebnisse von den Korrelationen zwischen den Merkmalen ab. Wenn eine hohe Korrelation der Variablen besteht, so überträgt sich eine Sortierung nach einer Variablen annähernd auf die der anderen. Die obigen Überlegungen zur Medianverschiebung bei der getrennten Mikroaggregation gelten dann näherungsweise weiter. Wenn allerdings keine ausreichend hohe Korrelation vorhanden ist, verändert sich der Median in nicht kontrollierbarer Weise.

Die Resultate für Median und Quantile können nicht auf die Varianten der stochastischen Mikroaggregation übertragen werden, weil die Größenordnungen bei einer randomisierten Auswahl der zu aggregierenden k Werte völlig verschieden sein können.

19.1.4 Auswirkungen stochastischer Überlagerungen auf das arithmetische Mittel

Eine Menge von Vektoren aus Originalwerten (Variable) werden mit $\mathbf{x}_j$ $(j = 1, ..., m)$ bezeichnet. Zur Maskierung von Originalwerten werden diese in den einzelnen Variablen mit Zufallsfehlern (Störtermen), die aus bestimmten Verteilungen als Pseudo-Zufallszahlen (Stichprobe für jede Variable) generiert werden, additiv oder multiplikativ überlagert.

X sei eine beliebige Originalvariable, die mit einer Zufallsvariablen (Störterm) verschiedenen Verteilungstyps zur Anonymisierung ihrer Werte überlagert werden soll. Es werden die fol-

genden Verteilungen zur Generierung von Zufallszahlen für die Überlagerungen untersucht:

- Additive Überlagerung mit einer Variablen aus einer Normalverteilung $N(0, \sigma^2)$ bzw. einer Mischungsverteilung aus Normalverteilungen mit Mittel 0 und Varianz σ^2:

  $$X^a = X + W \quad bzw. \quad x_i^a = x_i + w_i \quad (i = 1, ..., n)$$

  Untersucht wird die Auswirkung der Überlagerung auf das arithmetische Mittel von x_i^a.

  $$\frac{1}{n}\sum_{i=1}^{n} x_i^a = \frac{1}{n}\sum_{i=1}^{n}(x_i + w_i) = \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} w_i \tag{19.4}$$

  Da $E(W) = 0$ ist der zweite Summand für großes n annähernd Null: Das arithmetische Mittel der überlagerten Variablen x_i^a stimmt also bis auf unwesentliche Abweichungen mit dem arithmetischen Mittel der originalen Ausprägungen x_i überein.

  Verwendet man innerhalb der naiven Überlagerung eine Mischungsverteilung mit den oben angegebenen Spezifikationen, so gelten die Ergebnisse analog, wobei die tatsächlich wirksame Varianz aus einer Mischungsverteilung stammt (siehe Unterabschnitt 6.2.3).

- Multiplikative stochastische Überlagerung

  W sei eine stetige Zufallsvariable mit $W > 0 \wedge E(W) = 1 \wedge var(W) = \sigma_w^2 > 0$.

  $$X^a = X * W \quad bzw. \quad x_i^a = x_i * w_i \quad (i = 1, ..., n) \tag{19.5}$$

  $$E(\bar{x}^a) = E(\frac{1}{n}\sum_{i=1}^{n} x_i^a) = \frac{1}{n}\sum_{i=1}^{n} x_i * E(w_i) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{19.6}$$

  Der Mittelwert von x_i^a ist approximativ gleich $\bar{x}$. Je nach Spezifikation der Zufallszahlen u_i und der Größe von n können auch größere Abweichungen auftreten. Im Mittel gilt:

  $$\bar{x}^a \approx \bar{x}.$$

  Siehe dazu auch Höhne (2004a) für den Fall von sehr schief verteilten Originaldaten.

19.1.5 Auswirkungen stochastischer Überlagerungen auf den Median und andere Quantile

Über die Auswirkungen der stochastischen Überlagerung auf den Median lässt sich sowohl im additiven als auch im multiplikativen Fall keine generelle Aussage machen, weil dies von der jeweiligen Situation im mittleren Bereich der Ausgangsverteilung abhängt. Entsprechendes gilt für andere Quantile.

19.1.6 Auswirkungen von Mikroaggregation und stochastischen Überlagerungen auf Streuungsmaße

In diesem Unterabschnitt sind die Auswirkungen von Mikroaggregation und Zufallsüberlagerung auf Varianzen bzw. Standardabweichungen zu analysieren. Die genannten Größen sind die in der deskriptiven Analyse standardmäßig verwendeten Dispersionsmaße, die darüber hinaus auch, bei weitergehenden Voraussetzungen, in der schätzenden Statistik von besonderer Bedeutung sind.

a) Auswirkungen der Mikroaggregation auf die Varianz

Gegeben seien wieder metrisch skalierte Variablen des originalen Datenmaterials. Die Daten werden mit der k-gliedrigen Mikroaggregation transformiert; wiederum sei $k = 3$, für $k > 3$ gilt eine analoge Argumentation. Der Mittelwert der Variablen ist, wie weiter oben ausgeführt, invariant gegenüber dieser Transformation. Damit gilt Folgendes:

Die Varianz einer originalen beziehungsweise mikroaggregierten Variablen X, X^a (n Werte, $n \equiv 0 \mod k$; k-gliedriger Mittelwert für die Aggregation) lässt sich mit Hilfe des Varianz-Zerlegungssatzes für klassierte Daten darstellen und dadurch in seiner Größenordnung bezogen auf die Varianz der Originalwerte beurteilen.

$x_1, \ldots, x_n$ wird in $r = n/k$ Klassen mit Anzahl von k Werten in jeder Klasse partitioniert. Eine Vorsortierung der Werte nach Größe kann bei dieser Betrachtung entfallen. Daten werden in originaler oder sonstwie beliebiger Anordnung behandelt. Infolge dessen lässt sich auch die Varianzänderung durch stochastische Mikroaggregation einbeziehen.

Die Varianz der Originalwerte lässt sich generell darstellen:

$$s_x^2 = \underbrace{\frac{1}{n}\sum_{j=1}^{r} n_j s_j^2}_{\text{Interne Varianz}} + \underbrace{\frac{1}{n}\sum_{j=1}^{r} n_j (\bar{x}_j - \bar{x})^2}_{\text{Externe Varianz}} \tag{19.7}$$

In der Formel bedeuten:
n_j Elemente der j-ten Partition; $\sum n_j = n$
s_j^2 Varianz innerhalb der j-ten Partition
$\bar{x}_j$ Mittelwert in j-ter Partition
$\bar{x}$ Gesamtmittelwert aller x_i

Bei der Mikroaggregation mit k-gliedrigem Durchschnitt werden die Originalwerte innerhalb jeder Klasse durch das zugehörige k-gliedrige arithmetische Mittel ersetzt. Die so

entstehende, veränderte Variable X wird mit X^a bezeichnet. Die Mittelwerte $\bar{x}_j, \bar{x}$ bleiben nach Konstruktion unverändert, ebenso die $n_j = k$ für $j = 1, ..., r$. Damit folgt:

- Die interne Varianz insgesamt wird Null für die anonymisierten Werte, weil innerhalb jeder Partition nur gleiche Werte auftreten; klasseninterne Varianz ist Null.
- $s^2_{x^a}$ ist damit gleich der externen Varianz der Originalwerte, welche infolge der Konstanz von Gesamtmittel und aller Mittelwerte der Partitionen konstant geblieben ist, d.h.:

$$s^2_{x^a} = \frac{1}{n}\sum_{j=1}^{r} n_j(\bar{x}_j - \bar{x})^2 \;\leq\; s^2_x \tag{19.8}$$

  Insbesondere folgt mit $n_j = k$:

$$var(x^a) = \frac{1}{r}\sum_{j=1}^{r}(\bar{x}_j - \bar{x})^2 \leq var(x) \tag{19.9}$$

Diese Abschätzung der Varianzen gilt somit für alle Formen der Mikroaggregation.

Die Ergebnisse zur Varianzabschätzung (Standardabweichungsabschätzung) bei einer mikroaggregierten Variablen lassen sich auf die gemeinsame und die gruppierte Mikroaggregation übertragen. Dabei ist die Varianzverringerung umso kleiner, je ähnlicher sich die Werte sind, die in einer Gruppe zusammengefasst werden. Somit ist bei der getrennten abstandsorientierten Mikroaggregation der Varianzverlust am geringsten. Bei der gemeinsamen abstandsorientierten Mikroaggregation ist der Varianzverlust dann geringer, wenn die Variablen stark miteinander korreliert sind. Somit fällt der Varianzverlsut am höchsten aus, wenn die Gruppenbildung zufällig erfolgt.

b) Auswirkungen stochastischer Überlagerung auf Varianzen

Die stochastische Überlagerung wird wiederum mit additiven Überlagerungen (z.B. naive Überlagerung mit $N(0, \sigma)$, Mischungsverteilung mehrerer Normalverteilungen) sowie mit multiplikativer Überlagerung vorgenommen.

- Additive stochastische Überlagerung

$$X^a = X + W \quad bzw. \quad x_i^a = x_i + w_i$$

  Für die Varianz ergibt sich:

$$\begin{aligned} var(X^a) = E((X^a - E(X^a))^2) &= E((X - E(X)) + (W - E(W))^2 \\ &= var(X) + {\sigma_w}^2 \end{aligned} \tag{19.10}$$

Die Varianz der überlagerten Merkmalswerte ist also größer als die Varianz der originalen Merkmalswerte.

- Multiplikative Überlagerung

$$X^a = X \cdot W \quad bzw. \quad x_i^a = x_i \cdot w_i \quad (i = 1, ..., n)$$

Für die Varianz ergibt sich:

$$\begin{aligned} var(X^a) &= var(XW) = E((XW)^2) - E(X)^2 E(W)^2 \\ &= var(X) + \sigma_w^2([E(X)]^2 + var(X)) \\ &> var(X) \end{aligned} \tag{19.11}$$

Die Varianz der multiplikativ überlagerten x-Werte ist also stets größer als die empirische Varianz der Originalwerte, wobei die Vergrößerung der Varianz auch von Erwartungswert und Varianz der Originalvariablen abhängt.

19.1.7 Auswirkungen von Mikroaggregation und stochastischen Überlagerungen auf die empirische Korrelation

Bei diesen weiteren Untersuchungen über die Auswirkungen bestimmter Anonymisierungsmethoden wird analysiert, wie sich die empirische, korrelative Verbindung zweier metrisch skalierter Variablen beziehungsweise die empirische Interkorrelation in Gruppen von Variablen unter dem Einfluss von Mikroaggregation bzw. stochastischer Überlagerung verändert.

a) Auswirkungen der Mikroaggregation auf empirische Korrelationen

Der Bravais-Pearson-Korrelationskoeffizient für die Realisationen der metrisch skalierten Variablen X_1 und X_2 ist durch die folgende Formel definiert:

$$r_{x_1 x_2} = \frac{s_{x_1 x_2}}{s_{x_1} s_{x_2}}. \tag{19.12}$$

Die Ergebnisse für die Varianzen (Standardabweichungen), die im Nenner der Formel für den empirischen Korrelationskoeffizienten auftreten, wurden bereits abgeleitet.

In analoger Art und Weise wie im Unterabschnitt 19.1.6 über die Wirkung von Mikroaggregation auf Varianzen lässt sich eine Zerlegungsformel für die Kovarianz bei klassierten Daten herleiten und beim vorliegenden Problem anwenden, allerdings nur, wenn die beiden Variablen, deren Kovarianz betrachtet wird, gemeinsam mikroaggregiert werden.

Die Kovarianz wird wiederum in eine interne und eine externe Komponente zerlegt, d.h. die Kovarianz ist eine Summe (lineare Kombination) der beiden Komponenten:

$$cov(X,Y) = \underbrace{\frac{1}{r}\sum_{i=1}^{r} cov(x_{i,j}, y_{i,j})}_{interne\ Kovarianz,\ (Term1)} + \underbrace{\frac{1}{r}\sum_{i=1}^{r}(\bar{x}_i - \bar{x})(\bar{y}_i - \bar{y})}_{externe\ Kovarianz,\ (Term2)} \qquad (19.13)$$

Die gesamte interne Kovarianz der Variablen X^a, Y^a wird Null; die restliche, externe Kovarianz gibt somit den Wert der Kovarianz zwischen den zu X,Y gehörigen mikroaggregierten Variablen X^a, Y^a an; sie wird als $cov(X^a, Y^a)$ bezeichnet.

Die externe Kovarianz ist bei Originaldaten und zugehörigen mikroaggregierten Daten identisch, weil die Mikroaggregation alle in der Formel auftretenden Mittelwerte invariant lässt. Sind die Vorzeichen von interner und externer Kovarianz beide positiv, so folgt, dass die Kovarianz der Originalwerte größer ist als die Kovarianz der zugehörigen mikroaggregierten Variablen; entsprechend gilt, wenn beide negativ sind, die Originalkovarianz kleiner ist als die der mikroaggregierten Variablen. Treten unterschiedliche Vorzeichen bei interner und externer Kovarianz auf, so entscheiden die Vorzeichenmuster der internen und externen Kovarianzen gemäß der Zerlegungsformel über die Größenrelation zwischen $cov(X,Y)$ und $cov(X^a, Y^a)$.

Die Aussagen zur Kovarianz lassen sich jedoch nur unter bestimmten Bedingungen auf die Korrelationen übertragen, nämlich dann, wenn sich die Kovarianzen bzw. ihre Beträge durch die Anonymisierung vergrößern. Dann gilt folgendes:

$$cov(X,Y) - cov(X^a, Y^a) < 0$$

$$\Downarrow$$

$$\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} - \frac{cov(X^a, Y^a)}{\sqrt{var(X)var(Y)}} < 0$$

$$\Downarrow$$

$$\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} - \frac{cov(X^a, Y^a)}{\sqrt{var(X^a)var(Y^a)}} < 0$$

$$\Downarrow$$

$$corr(X,Y) \quad < \quad corr(X^a,Y^a)$$

analog hierzu gilt:

$$|cov(X,Y)| - |cov(X^a,Y^a)| \quad < \quad 0$$

$$\Downarrow$$

$$\frac{|cov(X,Y)|}{\sqrt{var(X)var(Y)}} - \frac{|cov(X^a,Y^a)|}{\sqrt{var(X)var(Y)}} \quad < \quad 0$$

$$\Downarrow$$

$$\frac{|cov(X,Y)|}{\sqrt{var(X)var(Y)}} - \frac{|cov(X^a,Y^a)|}{\sqrt{var(X^a)var(Y^a)}} \quad < \quad 0$$

$$\Downarrow$$

$$|corr(X,Y)| \quad < \quad |corr(X^a,Y^a)|$$

Aussagen sind auch möglich, sofern sich die Kovarianzen durch die Mikroaggregation gar nicht verändern, weil die internen Kovarianzen Null sind. Im diesem Fall gilt:

$$cov(X,Y) - cov(X^a,Y^a) \quad = \quad 0$$

$$\Downarrow$$

$$\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} - \frac{cov(X^a,Y^a)}{\sqrt{var(X)var(Y)}} \quad = \quad 0$$

$$\Downarrow$$

$$\frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} - \frac{cov(X^a,Y^a)}{\sqrt{var(X^a)var(Y^a)}} \quad < \quad 0$$

$$\Downarrow$$

$$corr(X,Y) \quad < \quad corr(X^a,Y^a)$$

und analog hierzu:

$$|cov(X,Y)| - |cov(X^a,Y^a)| \quad = \quad 0$$

$$\Downarrow$$

$$\frac{|cov(X,Y)|}{\sqrt{var(X)var(Y)}} - \frac{|cov(X^a,Y^a)|}{\sqrt{var(X)var(Y)}} \quad = \quad 0$$

$$\Downarrow$$

$$\frac{|cov(X,Y)|}{\sqrt{var(X)var(Y)}} - \frac{|cov(X^a,Y^a)|}{\sqrt{var(X^a)var(Y^a)}} \quad < \quad 0$$

$$\Downarrow$$

$$|corr(X,Y)| \quad < \quad |corr(X^a,Y^a)|$$

Für andere Fälle sind keine Aussagen möglich.

Diese Herleitungen gelten für die gemeinsame Mikroaggregation und die gruppierte Mikroaggregation, sofern die betrachteten Merkmale gemeinsam mikroaggregiert werden. Sie können nicht auf die getrennte Mikroaggregation übertragen werden, weil für beide Merkmale, deren Kovarianz betrachtet wird, die Gruppenbildung unterschiedlich sein kann. Somit geht auch die „interne“ Kovarianz durch die Mikroaggregation nicht verloren.

Zu erwähnen ist noch, dass der Wert des Rangkorrelationskoeffizienten nach Spearman bei einer abstandsorientierten getrennten Mikroaggregation grundsätzlich erhalten wird. Lediglich durch eine höhere Anzahl von Bindungen können sich geringfügige Veränderungen ergeben.

b) Auswirkungen stochastischer Überlagerungen auf empirische Korrelationen

- Additive stochastische Überlagerungen mit Mittelwert Null und Varianz σ^2

  Im folgenden seien zwei Zufallsvariablen X und Y gegeben, die jeweils additiv stochastisch überlagert werden:

$$X^a = X + W \quad \text{und} \quad Y^a = Y + V$$

Zunächst wird gezeigt, dass die Kovarianz von X^a und Y^a mit der von X und Y identisch ist, sofern die Überlagerungen W und V unkorreliert sind:

$$\begin{aligned} cov(X^a, Y^a) &= E(X^a Y^a) - E(X^a)E(Y^a) \\ &= E(XY) + E(XV) + E(YW) + E(WV) - (E(X) + E(W))(E(Y) + E(V) \\ &= E(XY) - E(X)E(Y) \\ &= cov(X, Y). \end{aligned} \tag{19.14}$$

Somit ergibt sich eine Verringerung der Korrelation:

$$corr(X^a\,,\, Y^a) \;=\; \frac{cov(X\,,\,Y)}{\sqrt{(\sigma_x^2 + \sigma_w^2)\,(\sigma_y^2 + \sigma_v^2)}} \;<\; \frac{cov(X\,,\,Y)}{\sqrt{\sigma_x^2\,\sigma_y^2}} \;=\; corr(X\,,\,Y) \tag{19.15}$$

- Multiplikative Überlagerung

  Es wird vorausgesetzt: $W, V > 0$, $E(W) = E(V) = 1$. W und V sind dabei unabhängig von X und Y.

$$X^a = XW; \qquad Y^a = YV$$

Dann gilt für die Kovarianz zwischen den überlagerten Variablen:

$$\begin{aligned} cov(X^a, Y^a) &= E(XW \cdot YV) - E(XW)E(YV) \\ &= E(XY \cdot WV) - E(X)E(Y) \\ &= E(XY)E(WV) - E(X)E(Y) \\ &= cov(X, Y). \end{aligned} \tag{19.16}$$

falls W und V stochastisch unabhängig sind. Falls die Überlagerungsfaktoren der beiden Variablen hingegen korreliert sind, so bleiben die Kovarianzen nicht erhalten und es können keine generellen Aussagen zu den Korrelationen gemacht werden.

Damit gilt die gleiche Argumentation hinsichtlich der Korrelation und der zugehörigen empirischen Korrelation wie im Fall der additiven Überlagerung weiter oben. Die Korrelationen verkleinern sich, weil die Kovarianzen erhalten bleiben, aber die Varianzen durch die Überlagerung größer werden.

19.1.8 Auswirkungen von Mikroaggregation und stochastischen Überlagerungen auf ein Disparitätsmaß

Im Folgenden wird der Gini-Koeffizient unter Einfluss verschiedener Anonymisierungsverfahren für die Originalvariablen für einen endlichen Stichprobenumfang untersucht. Es liegen Originalwerte **x** einer metrisch skalierten Variablen X als $x_i | i = 1, ..., n$ mit $x_i > 0$ vor. Die x_i seien in aufsteigender Reihenfolge sortiert.

Der Gini-Koeffizient ist ein Maß für die Disparität der auf Teilmengen von Einheiten entfallenden kumulierten Merkmalsbeträge. Ihr enger Zusammenhang mit der Lorenzkurve ist bekannt. Die Definition des Gini-Koeffizienten lautet:

$$G := \frac{2 \cdot \sum_{1}^{n} i \cdot x_i - (n+1) \cdot \sum_{i=1}^{n} x_i}{n \cdot \sum_{i=1}^{n} x_i} = \frac{2 \cdot \sum_{i=1}^{n} i \cdot p_i - (n+1)}{n} \qquad (19.17)$$

Die p_i sind die relativen Häufigkeiten der x_i.

a) Auswirkungen der Mikroaggregation auf den Gini-Koeffizienten

Bei der Mikroaggregation gelte wieder für den Stichprobenumfang: $n = 3 \cdot \acute{n}$. Die aus x durch Mikroaggregation erzeugten Werte seien x^a benannt. Der Mittelwert einer Variablen ist invariant gegenüber der Mikroaggregation, d.h. es gilt: $\bar{x} = \bar{x}^a$.

Der Ausdruck $\sum_{i=1}^{n} i \cdot x_i^a$ in seiner Relation zu $\sum_{i=1}^{n} i \cdot x_i$ ist entscheidend für die Änderung von G bei Mikroaggregation. Es gilt:

$$\sum_{i=1}^{n} i \cdot x_i^a = (1 \cdot x_1^a + 2 \cdot x_1^a + 3 \cdot x_1^a) + (4 \cdot x_4^a + 5 \cdot x_4^a + 6 \cdot x_4^a) + \ldots + (n-2) \cdot x_{\acute{n}}^a + (n-1) \cdot x_{\acute{n}}^a + n \cdot x_{\acute{n}}^a \qquad (19.18)$$

Dieser Ausdruck lässt sich in folgende Form umschreiben:

$$x_1 + (1 \cdot x_1 + 2 \cdot x_2 + 3 \cdot x_3) - x_3 + x_4 + (4 \cdot x_4 + 5 \cdot x_5 + 6 \cdot x_6) - x_6 + \ldots + n \cdot x_n - x_n \qquad (19.19)$$

Durch weitere Vereinfachung ergibt sich:

$$\sum_{i=1}^{n} i \cdot x_i^a = \sum_{i=1}^{n} i \cdot x_i - \Delta_n \quad mit \quad \Delta_n := \sum_{i=1,4,7,..,(n-2)} (x_{i+2} - x_i) \tag{19.20}$$

$\Delta_n \geq 0$ nach Definition und Größensortierung. Damit erhält man folgende Abschätzung:

$$\sum_{i=1}^{n} i \cdot x_i^a \leq \sum_{i=1}^{n} i \cdot x_i \tag{19.21}$$

Also folgt: Der Gini-Koeffizient wird durch Mikroaggregation der Variablenwerte verkleinert. Sollten die x-Werte bereits in einer Form vorliegen, die einer mikroaggregierten Variablen entspricht, so ändert sich der Gini-Koeffizient nicht, weil dann $\Delta = 0$ gilt. Nur für diesen Fall gilt das Gleichheitszeichen in der Ungleichung.

Die Größenordnung der Verkleinerung von G wird durch den Wert von Δ bestimmt, entsprechend der folgenden Differenz, die nach Ausrechnung anhand der Originalwerte einen numerischen Größenvergleich zwischen G_x und G_{x^a} ermöglicht:

$$\sum_{i=1}^{n} i \cdot x_i - \sum_{i=1}^{n} i \cdot x_i^a = \Delta_n \tag{19.22}$$

bzw. direkt für G_x und G_{x^a}:

$$(G_x - G_{x^a}) = \frac{2 \cdot \Delta_n}{n \cdot \sum_{i=1}^{n} x_i} \tag{19.23}$$

Da $\Delta_n << \sum i = 1^n x_i$ gilt, konvergiert der Ausdruck auf der rechten Seite gegen Null und damit der Gini-Koeffizient für das anonymisierte Merkmal gegen den für die Originalwerte.

b) Auswirkungen von stochastischen Überlagerungen auf den Gini-Koeffizienten

- Die Modalitäten einer additiven stochastischen Überlagerung seien analog zu Unterabschnitt 19.1.4 spezifiziert. Die Originalwerte seien mit x, die anonymisierten Werte mit x^a bezeichnet. n sei der Stichprobenumfang der Variablen. Damit folgt:

$$x_i^a := x_i + w_i \qquad | \; i = 1, ..., n \tag{19.24}$$

  Für den Gini-Koeffizienten der Variablen X^a ergibt sich dann:

$$G_{x^a} := \frac{2 \cdot \sum_{i=1}^{n} i \cdot (x_i + w_i) - (n+1) \cdot \sum_{i=1}^{n} (x_i + w_i)}{n \cdot \sum_{i=1}^{n} (x_i + w_i)} \tag{19.25}$$

Umformungen führen zu:

$$\begin{aligned} G_{x^a} &= \frac{(2 \cdot \sum_{i=1}^{n} i \cdot x_i - (n+1) \cdot \sum_{i=1}^{n} x_i) + (2 \cdot \sum_{i=1}^{n} i \cdot w_i - (n+1) \cdot \sum_{i=1}^{n} w_i)}{n \left(\cdot \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} w_i \right)} \\ &= G_x \cdot \frac{1}{1 + \frac{\sum_{i=1}^{n} w_i}{\sum_{i=1}^{n} x_i}} + G_u \cdot \frac{1}{1 + \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} w_i}} \end{aligned} \tag{19.26}$$

Da $E(\sum w_i) = 0$ und demnach $\sum w_i \approx 0$, folgt, dass der Gini-Koeffizient sich nur geringfügig verändert (in der Größenordnung abhängig von n und dem zugrundeliegenden σ). Über die Richtung der Veränderung lässt sich keine Aussage treffen.

Sollten in der Variablen X^a infolge der stochastischen Überlagerung negative Werte auftreten, so ist der Gini-Koeffizient für diese Daten nicht definiert.

- Eine multiplikative Überlagerung wird wie in Unterabschnitt 19.1.4 spezifiziert.

  Es gilt entsprechend: $X^a = X \cdot W$, damit für die Komponenten der Variablen: $X_i^a = x_i \cdot w_i$ für i=1,....,n.

  Für den Gini-Koeffizienten von X^a folgt:

$$G_{x^a} = \frac{2 \sum_{1}^{n} i \cdot x_i \cdot w_i - (n+1) \sum_{1}^{n} x_i \cdot w_i}{n \sum_{1}^{n} x_i \cdot w_i} \tag{19.27}$$

  Aufgrund der Voraussetzungen gilt:

$$E(\sum x_i \cdot w_i) = \sum x_i; \qquad E(\sum i \cdot x_i \cdot w_i) = \sum i \cdot x_i, \tag{19.28}$$

  woraus approximativ (in der Größenordnung abhängig von n und σ) folgt:

$$\sum x_i \cdot w_i \approx \sum x_i; \qquad \sum i \cdot x_i \cdot w_i \approx \sum i \cdot x_i \tag{19.29}$$

Für die Relation G_{x^a} zu G_x ergibt sich damit:

$$G_{x^a} \approx G_x \tag{19.30}$$

Der Gini-Koeffizient G_{x^a} rangiert auch bei der multiplikativen Überlagerung mit geringfügigen Änderungen in der Größenordnung des G_x. Über die Richtung der Unterschiedlichkeit läßt sich keine Aussage treffen.

19.2 Praxisbeispiele

Am Beispiel der Kostenstrukturerhebung im Verarbeitenden Gewerbe wird in den Tabellen 19.1 und 19.2 gezeigt, wie stark einzelne Varianten der Mikroaggregation und der stochastischen Überlagerung die Verteilungsmaße beeinflussen. Ein direkter Vergleich zwischen beiden Tabellen ist nur ungefähr möglich, weil für die Berechnungen der Maße in Tabelle 19.1 die gesamte KSE für 1999 herangezogen wurde, während für die Berechnungen in Tabelle 19.2 Wirtschaftszweig 37 (Recycling) entfernt wurde. Außerdem umfassen die Berechnungsergebnisse in Tabelle 19.1 33 metrische Merkmale, die in Tabelle 19.2 hingegen lediglich 30, weil die Merkmale „Tätige Inhaber", „Angestellte und Arbeiter" sowie „Bestandsveränderungen an fertigen und unfertigen Erzeugnissen aus eigener Produktion" für die späteren Untersuchungen und die Erstellung eines Scientific-Use-Files entfernt wurden. Allerdings ist der Einfluss dieser beiden Veränderungen auf die Höhe der Abweichungen offenbar gering, was man für die getrennte Mikroaggregation sehen kann, für die die Abweichungen in beiden Fällen berechnet wurden.

Im Einzelnen werden folgende Mikroaggregationsverfahren betrachtet:

1. **Abstandsorientierte Mikroaggregationsverfahren**

   - Getrennte abstandsorientierte Mikroaggregation: MA33G/MA30G
   - Gemeinsame abstandsorientierte Mikroaggregation: MA1G
   - Teilweise gemeinsame abstandsorientierte Mikroaggregation mit elf Gruppen: Zunächst werden die Bravais-Pearson-Korrelationen zwischen den 33 stetigen Variablen berechnet. Anschließend werden die drei Variablen mit den höchsten Korrelationen zusammengefasst, dann die nächsten drei, bis schließlich noch drei stetige Variablen übrig bleiben. Auch diese werden zu einer Gruppe zusammengefasst. Für diese so entstandenen elf Gruppen von Variablen werden jeweils die euklidischen Distanzen zwischen den Unternehmen bestimmt. Auf dieser Basis werden für die Gruppen von Variablen getrennt die Unternehmen zu Gruppen mit einer Besetzungsgröße von mindestens drei zusammengefasst und die Einzelwerte der Variablen durch den Durchschnitt der (mindestens) drei Unternehmen ersetzt: MA11G.

- Teilweise gemeinsame abstandsorientierte Mikroaggregation mit acht Gruppen: Wiederum werden die Korrelationskoeffizienten zwischen den Variablen bestimmt. Allerdings werden die Variablen nun so zusammengefasst, dass Gruppen mit unterschiedlicher Größe entstehen, wobei die Variablen in einer Gruppe möglichst stark miteinander korreliert sind. Als problematisch erweist sich eine „Restgruppe" aus den Variablen „Tätige Inhaber" und „Bestandsveränderungen", da diese Variablen sowohl miteinander als auch mit den anderen Variablen nur sehr gering korreliert sind: MA8G.

2. **Stochastische Mikroaggregationsverfahren**

   - Zufällige gemeinsame Mikroaggregation mit einer Gruppengröße von drei bis fünf: MA1G_stoch,
   - Gemeinsame Bootstrap-Mikroaggregation mit einer Gruppengröße von drei: MA1G_BS

Folgende Varianten der stochastischen Überlagerung werden untersucht:

1. Additive Überlagerung mit Korrektur der ersten und zweiten Momente, dabei variiert der Überlagerungsfaktor d (1 Prozent, 5 Prozent, 10 Prozent): Kim_d1p, Kim_d5p, Kim_d10p
2. Multiplikative Überlagerung mit einem konstanten Faktor aus einer zweigipfligen Mischungsverteilung aus zwei Normalverteilungen, dabei variieren die Abstände zwischen den beiden Gipfeln ($2f$) und die Standardabweichungen (s) der Verteilungen. Im Einzelnen werden folgende Varianten getestet:
   - Abstand der Mittelwerte von 1: f=0,04; Standardabweichung: s=0,02: Mult_f04_s02
   - Abstand der Mittelwerte von 1: f=0,08; Standardabweichung: s=0,018: Mult_f08_s018
   - Abstand der Mittelwerte von 1: f=0,11; Standardabweichung: s=0,03: Mult_f11_s03
3. Multiplikative Überlagerung mit einem zunächst konstanten Faktor aus einer Mischungsverteilung aus zwei Normalverteilungen und anschließende Korrektur der ersten und zweiten Momente (Kim-Korrektur). Dabei variieren die Abstände zwischen den beiden Gipfeln (f) und die Standardabweichungen (s) der Verteilungen. Im Einzelnen werden folgende Varianten getestet:
   - Abstand der Mittelwerte von 1: f=0,08; Standardabweichung: s=0,018: Mult_f08_s018_trans
   - Abstand der Mittelwerte von 1: f=0,11; Standardabweichung: s=0,03: Mult_f11_s03_trans

4. Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne mit $f = 0,11$ und $s = 0,03$: Hoe_f11_s03

5. Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne mit $f = 0,11$ und $s = 0,03$ und anschließender Kim-Korrektur: Hoe_f11_s03_trans

6. Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne mit $f = 0,11$ und $s = 0,03$ und anschließende gruppenweise Kim-Korrektur. Dabei wurde der Datenbestand für jede Variable absteigend sortiert und die Korrektur in Blöcken von 100 Sätzen vorgenommen: Hoe_f11_s03_trans_grupp

7. Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne mit $f = 0,11$ und $s = 0,02$ und anschließende gruppenweise Kim-Korrektur. Dabei wurde der Datenbestand für jede Variable absteigend sortiert und die Korrektur in Blöcken von 100 Sätzen vorgenommen: Hoe_f11_s02_trans_grupp

Tabelle 19.1: Veränderung von Verteilungsmaßen der KSE durch verschiedene Mikroaggregationsverfahren, Datengrundlage: gesamte KSE 1999

	Mittlerer relativer Fehler		Mittlerer absoluter Fehler	
	Arithmetische Mittel	Varianzen	Korrelationen	Rangkorrelationen
	in %	in %	(x 100)	(x 100)
MA1G	3,5	21,3	5,8	9,0
MA8G	1,8	16,7	3,5	2,4
MA11G	3,9	29,5	2,5	1,5
MA33G	0,0	5,9	2,4	0,0
MA1G_stoch	0,0	66,6	0,3	6,3
MA1G_BS	81,6	67,3	2,0	5,9

Die Mikroaggregationsverfahren erhalten grundsätzlich die arithmetischen Mittel, allerdings wurden bei der gemeinsamen Mikroaggregation, der getrennten Mikroaggregation und den beiden Varianten der gruppierten Mikroaggregation die Anteile der Lagerbestände an den Umsatzgrößen nicht direkt mikroaggregiert, sondern Lagerbestände und Umsatzgrößen getrennt, anschließend wurden die Quotienten neu berechnet. Dies ruft die beobachtete Verzerrung der arithmetischen Mittel hervor.

Auffällig ist jedoch, dass sich die arithmetischen Mittel bei der Bootstrap-Mikroaggregation stark verändern. Möglicherweise sind mehrere mit Bootstrap-Mikroaggregation bearbeitete Datensätze notwendig, um gute Schätzer für die orginalen arithmetischen Mittel zu erhalten.

Gruppierte und gemeinsame Mikroaggregation sowie insbesondere die beiden Varianten der stochastischen Mikroaggregation führen zu sehr starken Verzerrungen der Varianzen. Dies

ist auch zu erwarten, da bei der gemeinsamen abstandsorientierten Mikroaggregation, und noch stärker ausgeprägt bei einer stochastischen Mikroaggregation, auch stark unterschiedliche Einzelwerte in einer Gruppe zusammengefasst werden. Damit geht mehr Variation im Datensatz verloren als wenn – wie bei der getrennten Mikroaggregation – lediglich sehr ähnlich Einzelwerte in einer Gruppe zusammengefasst werden.

Insgesamt werden die Korrelationen durch die Mikroaggregationsverfahren recht gut erhalten, wobei die gemeinsame abstandsorientierte Mikroaggregation sowohl bei den Korrelationskoeffizienten nach Bravais-Pearson als auch bei den Rangkorrelationen nach Spearman die größten Abweichungen hervorruft. Die stochastischen Mikroaggregationsvarianten ziehen stärkere Veränderungen bei den Rangkorrelationen als bei den Bravais-Pearson-Korrelationen nach sich.

Tabelle 19.2: Veränderung von Verteilungsmaßen der KSE durch verschiedene stochastische Überlagerungen und getrennte Mikroaggregation, Datengrundlage: KSE 1999 ohne Wirtschaftszweig 37 (Recycling)

	Mittlerer relativer Fehler		Mittlerer absoluter Fehler	
	Arithmetische Mittel	Varianzen	Korrelationen	Rang-korrelationen
	in %	in %	(x 100)	(x 100)
Kim_d1p	0,0	0,0	0,0	13,7
Kim_d5p	0,0	0,0	0,1	15,9
Kim_d10p	0,0	0,0	0,1	16,6
Mult_f04_s02	0,7	6,7	0,2	0,0
Mult_f08_s018	1,2	12,1	0,3	0,1
Mult_f08_s018_trans	0,0	0,0	0,3	0,1
Mult_f11_s03	1,7	17,4	0,5	0,2
Mult_f11_s03_trans	0,0	0,1	0,5	0,2
Hoe_f11_s03_trans	0,1	0,2	0,8	0,3
Hoe_f11_s03	0,0	5,2	0,8	0,2
Hoe_f11_s03_trans_grupp	0,1	0,4	0,6	0,3
Hoe_f11_s02_trans_grupp	0,1	0,5	0,6	0,3
MA30G	0,0	6,2	2,4	0,0

Die Mittelwerte werden sowohl durch die additiven stochastischen Überlagerungen mit Erwartungswert Null als auch durch die multiplikativen stochastischen Überlagerungen mit Erwartungswert Eins theoretisch erhalten. Man erkennt, dass bei endlichen Stichproben leichte Abweichungen auftreten, die mit der Varianz der Überlagerungen zunimmt. Durch die Kim-Transformation können dann sowohl bei additiven als auch bei multiplikativen Überlagerungen die arithmetischen Mittel im Original exakt erhalten werden.

Erwartungsgemäß führen sowohl additive als auch multiplikative Überlagerungen zu erhöhten Varianzen, wobei die Erhöhung mit der Varianz der Überlagerungen zunimmt. Dabei ist die Varianzerhöhung beim Verfahren von Höhne geringer als bei der multiplikativen

Überlagerung mit einem konstanten Faktor aus einer echten Mischungsverteilung.

Auch die Varianzen können durch die Kim-Transformation für beide Arten der stochastischen Überlagerung korrigiert werden; jedoch sprechen zwei Argumente gegen die Transformation: Zum einen werden vor allem die Merkmalswerte kleinerer Unternehmen zusätzlich verzerrt, die jedoch keinen stärkeren Schutz benötigen. Zum zweiten werden durch die Transformation bei der multiplikativen Überlagerung deren Vorteile gegenüber der additiven Überlagerung teilweise zunichte gemacht. Die multiplikative Überlagerung erhält Vorzeichen und Nullen. Erfolgt jedoch eine zusätzliche Transformation zum Erhalt der ersten und zweiten Momente, so ist dies nicht mehr automatisch der Fall. Dieses Problem tritt auch auf, wenn die Kim-Korrektur lediglich gruppenweise vorgenommen wird, wenn auch in geringerem Ausmaß.

Vergleicht man die Varianten der multiplikativen stochastischen Überlagerung mit der getrennten abstandsorientierten Mikroaggregation, so fällt auf, dass die Überlagerungen zu einer geringeren Beeinträchtigung der Korrelationen führen als die Mikroaggregation, während ohne Kim-Korrektur die multiplikativen Überlagerungen mit Ausnahme des Höhne-Verfahrens zu stärkeren Verzerrungen der Varianzen führen als die getrennte Mikroaggregation.

Einschränkend muss darauf hingewiesen werden, dass die Wirkung der stochastischen Überlagerung und der stochastischen Mikroaggregationsverfahren im Gegensatz zu den deterministischen Mikroaggregationsverfahren auch vom Zufallsgenerator (vom Startwert) abhängt. Insbesondere die Varianzverzerrung ist hiervon betroffen.

Häufig interessieren sich Datennutzer nicht nur für die deskriptiven Kennzahlen eines Datenbestandes, sondern für spezielle Teilgesamtheiten beziehungsweise Subpopulationen. Diesem Aspekt wird bei der Anonymisierung der drei Projektstatistiken in Teil XI dieses Handbuchs ausführlich Rechnung getragen.

19.3 Zusammenfassende Bewertung

Eindeutige Schlussfolgerungen für die Bewertung datenverändernder Anonymisierungsverfahren sind aus der Sicht deskriptiver Auswertungen nur bedingt möglich, weil die Verfahren sich sehr unterschiedlich auf einzelne Verteilungsmaße auswirken. Beispielsweise werden durch das im Rahmen dieser Arbeit nicht näher betrachteten Swapping-Verfahren alle univariaten Verteilungsmaße erhalten, während die Korrelationen zerstört werden.

Die Mikroaggregationsverfahren sind mit Blick auf die theoretischen Ableitungen und die beispielhaft durchgeführten deskriptiven Auswertungen wie folgt zu bewerten:

- Mikroaggregationsverfahren erhalten generell die arithmetischen Mittel. Eine Ausnah-

me stellt die Bootstrap-Mikroaggregation dar. Bei dieser können die wahren arithmetischen Mittel als Mittelwert der arithmetischen Mittel mehrerer Samples geschätzt werden. Nicht erhalten werden auch die arithmetischen Mittel von Teilgesamtheiten, wenn die Mikroaggregation nicht auf die einzelnen Teilgesamtheiten getrennt angewendet wird. Bei arithmetischen Mitteln von Teilgesamtheiten sind die Abweichungen der arithmetischen Mittel dann geringer, wenn die Einzelwerte weniger stark verändert werden. Die besten Ergebnisse lassen sich somit mit der abstandsorientierten getrennten Mikroaggregation erzielen.

- Die Auswirkungen von Mikroaggregationsverfahren auf Quantile lassen sich nicht exakt vorhersagen. Lediglich für die abstandsorientierte getrennte Mikroaggregation kann ein Bereich für die Veränderung von Quantilen vorgegeben werden.

- Varianzen beziehungsweise Standardabweichungen werden durch die Mikroaggregation generell verringert. Die Verzerrung ist dabei umso geringer, je ähnlicher sich die Einzelwerte sind, die durch den selben Durchschnittswert ersetzt werden. Somit ist die Varianzverzerrung bei der abstandsorientierten getrennten Mikroaggregation am geringsten. Für die Bootstrap-Mikroaggregation gilt dieses Ergebnis nicht. Hier wird die Streuung nicht systematisch reduziert. Im Praxisbeispiel ergeben sich starke Veränderungen der Varianzen.

- Allgemeingültige Aussagen zur Veränderung der Korrelationskoeffizienten durch die Mikroaggregation sind nicht möglich. Es zeigt sich aber, dass die abstandsorientierte getrennte Mikroaggregation unter den abstandsorientierten Verfahren zu den geringsten absoluten Abweichungen der Korrelationskoeffizienten führt. Die Rangkorrelationen bleiben bei dieser Variante der Mikroaggregation ohnehin erhalten. Leichte Veränderungen ergeben sich lediglich durch die höhere Zahl an Bindungen.

Für die Auswirkungen stochastischer Überlagerungen auf deskriptive Auswertungen kann folgendes festgehalten werden:

- Additive stochastische Überlagerungen mit Erwartungswert Null und multiplikative Überlagerungen mit Erwartungswert Eins führen bei einer hohen Anzahl an Beobachtungswerten zu unveränderten arithmetischen Mitteln. Bei endlichen Stichproben kann es zu Abweichungen kommen, die jedoch durch die so genannte Kim-Transformation korrigiert werden können. Die Auswirkungen von stochastischen Überlagerungen auf Quantile sind nicht einfach abzuleiten.

- Stochastische Überlagerungen führen im Gegensatz zu Mikroaggregationsverfahren zu einer Erhöhung der Varianz. Diese nimmt mit der Varianz der Überlagerungen zu. Bei multiplikativen stochastischen Überlagerungen hängt die Zunahme der Varianz durch die Überlagerung zudem positiv von der Originalvarianz und dem originalen Mittelwert ab. Auch die Verzerrung der Varianz kann durch die Kim-Transformation korrigiert werden.

- Bei stochastischen Überlagerungen werden die Kovarianzen erhalten, sofern die Überlagerungen miteinander unkorreliert sind. Damit werden die Korrelationskoeffizienten in diesem Fall verringert. Bei proportionaler Varianz-Kovarianzmatrix werden die Korrelationen dagegen erhalten. Ansonsten lassen sich keine generellen Aussagen machen. Allerdings zeigen die Beispielrechnungen, dass die absoluten Abweichungen der Korrelationskoeffizienten für alle getesteten Varianten der stochastischen Überlagerung unterhalb derer für die getrennte Mikroaggregation liegen.
- Multiplikative Überlagerungen lassen sich im Unterschied zu additiven Überlagerungen so konstruieren, dass sie erwartungstreu sind und trotzdem keine Vorzeichenwechsel hervorrufen (Erwartungswert Eins und nur positive Werte). Diese Vorteile können durch die Anwendung der Kim-Transformation jedoch verloren gehen.

Kapitel 20

Auswirkungen der Post-Randomisierung auf deskriptive Auswertungen

20.1 Theoretische Eigenschaften

Das in Unterabschnitt 6.1.2 beschriebene Verfahren der Post-Randomisierung zur Anonymisierung kategorialer Merkmale hat den Vorteil, dass die Originalverteilung der kategorialen Variablen X aus der Verteilung der maskierten Variablen X^a geschätzt werden kann, wenn die bei der Post-Randomisierung angewendete Übergangsmatrix bekannt ist (Kooiman et al. 1997; Ronning 2005).

Für den Fall einer dichotomen Variable X sei θ der Anteil der Merkmalsträger in der Grundgesamtheit, die den Wert Eins annehmen. Für den Erwartungswert der anonymisierten Variablen X^a gilt:

$$E(X^a) \quad = \quad \pi\theta \quad + \quad (1-\pi)(1-\theta). \tag{20.1}$$

Damit gilt für den Erwartungswert des arithmetischen Mittels $\hat{\theta}^a = \frac{1}{n}\sum_{i=1}^{n} x_i^a$, das sich als erwartungstreuer Schätzer für den Anteil der Einsen in den anonymisierten Daten interpretieren lässt:

$$E\left(\hat{\theta}^a\right) \quad = \quad \pi\theta \quad + \quad (1-\pi)(1-\theta). \tag{20.2}$$

Daraus ergibt sich ein unverzerrter Schätzer für θ durch

$$\hat{\theta} \quad = \quad \frac{T^a - (1-\pi)}{2\pi - 1} \tag{20.3}$$

mit T^a dem realisierten Anteil der Beobachtungen mit Wert 1 nach der Anonymisierung. Eine Lösung existiert für alle $\pi \neq 1/2$.

Für die Varianz des Schätzers gilt:

$$var\left(\hat{\theta}\right) = \frac{var\left(T^a\right)}{(2\pi - 1)^2} \tag{20.4}$$

Für die Varianz von T^a gilt (Ronning 2005):

$$var(T^a) = var(\theta^a) = \theta\pi(1-\pi) + (1-\theta)(1-\pi)\pi = \pi(1-\pi) \tag{20.5}$$

Setzt man dieses Ergebnis in Gleichung (20.4) ein, so ergibt sich für die Varianz des Schätzers $\hat{\theta}$:

$$var\left(\hat{\theta}\right) = \frac{\pi(1-\pi)}{(2\pi - 1)^2} \tag{20.6}$$

Für den Fall einer nicht symmetrischen Übergangsmatrix ergibt sich für den unverzerrten Schätzer $\hat{\theta}$:

$$\hat{\theta} = \frac{T^a - (1 - \pi_0)}{\pi_1 + \pi_0 - 1} \tag{20.7}$$

Und für dessen Varianz gilt:

$$var\left(\hat{\theta}\right) = \frac{\theta(\pi_1(1-\pi_1) - \pi_0(1-\pi_0)) + \pi_0(1-\pi_0)}{\pi_1 + \pi_0 - 1} \tag{20.8}$$

Damit ist im nicht symmetrischen Fall die Varianz des Schätzers im Gegensatz zum symmetrischen Fall nicht von θ unabhängig.

Für den allgemeinen Fall von r Kategorien gilt für die erwarteten Anteile in der Kategorie j nach der Post-Randomisierung:

$$E\left(\hat{\theta}_j^a | X = x\right) = \pi_{jj}\theta_j + \sum_{k=1, k\neq j}^{r} \pi_{kj}\theta_k, \tag{20.9}$$

mit π_{jj} der Bleibewahrscheinlichkeit in Kategorie j und π_{kj} der Übergangswahrscheinlichkeit von Kategorie k in Kategorie j.

Daraus ergibt sich allgemein für den Schätzer des Originalanteils in Kategorie j:

$$\hat{\theta}_j = \frac{T_j^a - \sum_{k=1,k\neq j}^{r} \pi_{kj}\hat{\theta}_k}{\pi_{jj}} \tag{20.10}$$

Außerdem gilt:

$$\sum_{j=1}^{r} \hat{\theta}_j = 1 \tag{20.11}$$

Somit können die originalen Randhäufigkeiten, also r Unbekannte, aus einem Gleichungssystem mit $r+1$ Gleichungen auch im Fall mehrerer Kategorien erwartungstreu geschätzt werden.

Beim invarianten PRAM werden die Randhäufigkeiten und damit im Fall von zwei Kategorien die Anteile der Beobachtungen mit Null oder Eins sogar immer erhalten. Dies ist gerade der Gegenstand der Definition des invarianten PRAM in Unterabschnitt 6.1.2.

Probleme ergeben sich durch die Post-Randomisierung insbesondere bei mehrdimensionalen Auswertungen. Werden zwei oder mehr kategoriale Variablen mit Post-Randomisierung bearbeitet, so muss die Vertauschung zwischen Ausprägungskombinationen erfolgen, damit die Variablen anschließend gemeinsam ausgewertet werden können. Erfolgt die Post-Randomisierung hingegen für die einzelnen Variablen unabhängig voneinander, so sind gemeinsame Auswertungen nur möglich, wenn die Variablen stochastisch unabhängig sind. Der Grund hierfür besteht darin, dass die Randverteilungen der Originalvariablen in jedem Fall wie oben beschrieben aus den Randverteilungen der maskierten Variablen ermittelt werden können. Bei unabhängiger Post-Randomisierung der einzelnen Variablen kann jedoch die gemeinsame Wahrscheinlichkeitsverteilung der Originalvariablen nicht direkt aus der gemeinsamen Wahrscheinlichkeitsverteilung der maskierten Variablen abgeleitet werden. Die gemeinsame Wahrscheinlichkeitsverteilung lässt sich aber aus den Randwahrscheinlichkeiten nur im Falle der stochastischen Unabhängigkeit berechnen.

Dahingegen können bei einer Post-Randomisierung einer mehrdimensionalen Häufigkeitsverteilung durch die Definition von Übergangswahrscheinlichkeiten für die einzelnen Ausprägungskombinationen oder Zellen die originalen gemeinsamen Wahrscheinlichkeiten beziehungsweise Häufigkeiten für die einzelnen Zellen nach dem oben beschriebenen Vorgehen unverzerrt geschätzt werden. Daraus lassen sich wiederum Schätzer für die Randhäufigkeiten für die einzelnen Variablen berechnen.

Noch problematischer gestaltet sich die Auswertung metrischer Variablen in Abhängigkeit von durch Post-Randomisierung anonymisierten kategorialen Variablen, sofern beide Variablen nicht stochastisch unabhängig sind. Beispielsweise ist leicht einsichtig, dass die Tarifbindung von Unternehmen mit der Unternehmensgröße zunimmt (Strotmann 2004).

Wird nun die dichotome Variable Tarifbindung mit den Ausprägungen 1 für „ja“ und 0 für „nein“ durch Post-Randomisierung unabhängig von der metrischen Variablen „Betriebsgröße“ anonymisiert, so führt die getrennte Analyse der Betriebsgröße für tarifgebundene und nicht tarifgebundene Unternehmen zu verzerrten Ergebnissen.

20.2 Praxisbeispiele

Mit Hilfe des IAB-Betriebspanels 2002 für Baden-Württemberg werden die obigen Überlegungen veranschaulicht. Hierzu wurde die Variable „Tarifbindung“ mit den Ausprägungen „ja“ und „nein“ durch Post-Randomisierung in vier Varianten anonymisiert:

- Einfache Post-Randomisierung mit einer Wechselwahrscheinlichkeit von 5 Prozent (Bleibewahrscheinlichkeit π = 0,95)
- Einfache Post-Randomisierung mit einer Wechselwahrscheinlichkeit von 10 Prozent (Bleibewahrscheinlichkeit π = 0,90)
- Invariante Post-Randomisierung mit $\lambda = 0,80$
- Invariante Post-Randomisierung mit $\lambda = 0,90$

Vor und nach der Anonymisierung wird in Tabelle 20.1 zum einen der Anteil der Betriebe mit Tarifbindung, zum anderen die durchschnittliche Beschäftigtenzahl der Unternehmen mit und ohne Tarifbindung ausgewiesen.

Tabelle 20.1: Auswirkungen der Post-Randomisierung auf deskriptive Auswertungen

	Original	Einfaches PRAM		Invariantes PRAM	
		$\pi = 0,95$	$\pi = 0,90$	$\lambda = 0,8$	$\lambda = 0,9$
Anteil der tarifgebundenen Betriebe	0,6195	0,6003	0,5920	0,6220	0,6128
Durchschnittliche Beschäftigung der tarifgebundenen Betriebe	341,76	344,06	321,11	327,03	330,94
Durchschnittliche Beschäftigung der nicht tarifgebundenen Betriebe	79,40	88,51	127,02	101,90	101,03

Beim invarianten PRAM muss der Anteil der Unternehmen mit Tarifbindung konstant bleiben, was auch der Fall ist. Allerdings bleibt der Anteil auch im Fall der einfachen Post-Randomisierung im Fallbeispiel weitgehend konstant, weil der Anteil der tarifgebundenen

Betriebe im Original ca. 62 Prozent beträgt, die Verteilung somit ziemlich symmetrisch ist und die Wechselwahrscheinlichkeiten gering sind.

Allerdings ergibt sich bereits bei den vergleichsweise geringen Wechselwahrscheinlichkeiten eine beachtliche Verzerrung der durchschnittlichen Beschäftigtenzahl der tarifgebundenen und insbesondere der nicht tarifgebundenen Betriebe. Dies liegt daran, dass die Tarifbindung von der Beschäftigtenzahl abhängt (vgl. hierzu Abschnitt 21.2), dieser Zusammenhang aber bei der Post-Randomisierung nicht berücksichtigt wurde.

Die Verzerrung der durchschnittlichen Betriebsgröße steigt mit der Wechselwahrscheinlichkeit deutlich an. Legt man eine maximal zu akzeptierende Abweichung der durchschnittlichen Beschäftigtenzahl von zehn Prozent zugrunde, so kann eine Wechselwahrscheinlichkeit von fünf Prozent in diesem Fall gerade noch vertreten werden.

20.3 Ein Fazit für den Einsatz der Post-Randomisierung in deskriptiven Auswertungen

Wird Post-Randomisierung zur Anonymisierung von kategorialen Merkmalen eingesetzt, so sollte in jedem Fall das invariante PRAM angewendet werden, damit die originale Randverteilungen der anonymisierten Variablen erhalten bleiben. Allerdings müssen zudem die Abhängigkeiten der randomisierten Variablen mit anderen Variablen berücksichtigt werden. Deshalb empfiehlt es sich, die Post-Randomisierung mehrerer kategorialer Variablen abgestimmt so vorzunehmen, dass die mehrdimensionale Häufigkeitsverteilung erhalten bleibt.

Doch auch die Abhängigkeiten zwischen kategorialen und metrischen Variablen – wie zwischen Betriebsgröße und Tarifbindung – dürfen nicht zerstört werden. Deshalb sollten auch diese Zusammenhänge bei der Anwendung des Verfahrens berücksichtigt werden, beispielsweise, indem nur innerhalb ähnlicher Größenordnungen die Merkmalsausprägung getauscht (wie auch beim Rank Swapping) oder die Wechselwahrscheinlichkeit so gering gewählt wird, dass die Verzerrungen mehrdimensionaler Auswertungen unter Berücksichtigung der randomisierten Variablen die in Kapitel 18 definierten Abweichungsschwellen nicht überschreiten. Nach den durchgeführten Berechnungen sollte eine Wechselwahrscheinlichkeit von fünf Prozent deshalb nicht überschritten werden.

Teil VIII

Wirkung datenverändernder Anonymisierungsmethoden bei metrischen Variablen auf lineare und nichtlineare Modelle

Die folgenden Kapitel behandeln die Auswirkungen unterschiedlicher datenverändernder Anonymisierungsverfahren auf die Schätzung linearer und nichtlinearer Modelle. Zur Beurteilung werden theoretische Überlegungen, Monte-Carlo-Simulationen und Praxisbeispiele mit den Projektdaten herangezogen. Stochastische Überlagerungen und Mikroaggregationsverfahren werden sehr ausführlich untersucht. Von den Simulationsverfahren werden das Resampling-Verfahren und Latin Hypercube Sampling (LHS) weniger ausführlich betrachtet.

Zunächst werden in Kapitel 21 die als Praxisbeispiele herangezogenen Modelle vorgestellt. Kapitel 22 behandelt die Auswirkungen der stochastischen Überlagerung, Kapitel 23 die Auswirkungen der Mikroaggregation auf lineare und nichtlineare Modelle. In Kapitel 24 werden die Auswirkungen des Resampling-Verfahrens in linearen Modellen beschrieben. Kapitel 25 enthält Ergebnisse zur Wirkung des LHS in linearen Modellen. Eine Zusammenfassung der Ergebnisse dieses Teils findet sich in Kapitel 26.

Kapitel 21

Anwendungsbeispiele für lineare und nichtlineare Modelle

In den folgenden Kapiteln werden die Auswirkungen von Mikroaggregationsverfahren, stochastischen Überlagerungen und Post-Randomisierung auf lineare und nichtlineare Modelle anhand theoretischer Überlegungen, von Monte-Carlo-Simulationen und von Praxisbeispielen untersucht. Als Praxisbeispiele werden die im Folgenden beschriebenen Modelle genutzt.

21.1 Linearisierte Cobb-Douglas-Produktionsfunktion mit den Daten der Kostenstrukturerhebung

In Anlehnung an das Vorgehen von Fritsch und Stephan (2003) werden die Produktionselastizitäten einer Cobb-Douglas-Produktionsfunktion für das Verarbeitende Gewerbe inklusive Bergbau sowohl mit den Originaldaten der Kostenstrukturerhebung als auch mit anonymisierten KSE-Daten geschätzt. Im Unterschied zu Fritsch und Stephan (2003) stehen lediglich Querschnittsdaten für das Jahr 1999 zur Verfügung. Die Cobb-Douglas-Produktionsfunktion hat folgende Gestalt:

$$Y = A \prod_{k=1}^{K} X_k^{\beta_k}. \tag{21.1}$$

Dabei ist A der konstante Technologieparameter, β_k sind die Produktionselastizitäten, X_k die Inputfaktoren und Y der Output. Der Output wird wie von Fritsch und Stephan (2003) als Bruttoproduktionswert vermindert um den Umsatz aus sonstiger Tätigkeit definiert. Es gilt konkret:

Y = Bruttoproduktionswert - (Gesamtumsatz - Umsatz aus eigenen Erzeugnissen - Umsatz aus Handelsware).

Die Inputfaktoren werden in Anlehnung an Fritsch und Stephan (2003) wie folgt definiert:

- Materialeinsatz
- Personalkosten
- Externe Dienstleistungen
- Sonstige Kosten
- Kapitalkosten

Der *Materialeinsatz* beinhaltet neben dem *Verbrauch an Rohstoffen* und dem *Einsatz an Handelsware* im Unterschied zu Fritsch und Stephan (2003) auch den *Energieverbrauch*.[20] Die *Personalkosten* bestehen aus der *Bruttolohn- und Gehaltssumme*, den *Gesetzlichen Sozialkosten*, den *Sonstigen Sozialkosten* und den *Kosten für Leiharbeit*. Die *externen Dienstleistungen* setzen sich aus den *Kosten für Reparaturen* und den *Kosten für Lohnarbeiten* zusammen. In die *Kapitalkosten* gehen die internen Kapitalkosten in Form von *Abschreibungen* und die externen Kapitalkosten in Form von *Mieten und Pachten* ein (Fritsch und Stephan 2003).[21]

Durch Bildung des Logarithmus auf beiden Seiten der Gleichung (21.1) kann die Funktion linearisiert werden. Man schätzt daher das folgende lineare Modell:

$$log(Y) = C + \sum_{k=1}^{K} \beta_k ln(X_k). \tag{21.2}$$

Die Schätzung des Modells wird zunächst mit den Originalwerten durchgeführt[22]. Grundsätzlich werden Unternehmen ausgeschlossen, bei denen ein Inputfaktor oder der Output den Wert Null aufweist.

Um zu vermeiden, dass Unternehmen mit extremen Merkmalsausprägungen einen zu großen Einfluss auf die Schätzergebnisse haben, wird alternativ zur Berücksichtigung aller Unternehmen im Datensatz ein Szenario geschätzt, bei dem der Datensatz vorher wie folgt „bereinigt“ wird: Unternehmen, bei denen der Produktionsanteil eines Inputfaktors weniger

20) Wird der *Energieverbrauch* als separater Inputfaktor betrachtet, so ergibt sich bei der Schätzung des Modells eine negative Produktionselastizität für den Faktor Energie. Der Grund hierfür besteht möglicherweise darin, dass in die geschätzte Produktionsfunktion Wertgrößen und keine Mengengrößen eingehen. Beim Faktor Energie kann sich dies als Problem erweisen, da gleichzeitig nur ein Querschnittsdatensatz zur Verfügung steht und die Preise für den Faktor Energie im Zeitablauf starken Schwankungen ausgesetzt sind.

21) Im Unterschied zu Fritsch und Stephan (2003) wird der Wert der Abschreibungen für das Jahr 1999 verwendet, da ein Durchschnittswert aus mehreren Jahren nicht verfügbar ist.

22) Hocherechnungsfaktoren werden nicht berücksichtigt, da Vergleichsrechnungen gezeigt haben, dass hierdurch nur marginale Ergebnisveränderungen hervorgerufen werden.

als das 1-Prozent-Quantil bzw. mehr als das 99-Prozent-Quantil der Verteilung der Produktionsanteile über alle Unternehmen beträgt, werden gemäß dem Vorgehen von Fritsch und Stephan (2003) von den Berechnungen ausgeschlossen. Die wichtigsten Kenngrößen der Inputfaktoren bezogen auf den Output für die in diesem Szenario berücksichtigten Unternehmen sind in den Tabellen 21.1 und 21.2 zusammengestellt. Dabei zeigt Tabelle 21.1 die entsprechenden Werte für den gesamten Datensatz der KSE für 1999, Tabelle 21.2 die Werte für die KSE ohne den Wirtschaftszweig 37 (Recycling), der im Laufe des Projekts sowohl aus Sicherheitsgründen als auch aus inhaltlichen Gründen entfernt wurde. Die Ergebnisse ähneln denen von Fritsch und Stephan (2003).

Tabelle 21.1: Kenngrößen der Anteile der Inputfaktoren am Output (Datensatz bereinigt)

Anteil am Output	Arithmetisches Mittel	Median	Standard-abweichung	Minimum	Maximum	Variations-koeffizient
Material-einsatz	0,443	0,439	0,163	0,051	0,877	36,720
Personal-kosten	0,327	0,316	0,133	0,052	0,765	40,590
Externe Dienst-leistungen	0,048	0,027	0,055	0,001	0,335	114,730
Sonstige Kosten	0,094	0,078	0,060	0,010	0,348	64,100
Kapital-kosten	0,066	0,055	0,043	0,006	0,257	64,730

Tabelle 21.2: Kenngrößen der Anteile der Inputfaktoren am Output (ohne Wirtschaftszweig Recycling), Datensatz bereinigt

Anteil am Output	Arithmetisches Mittel	Median	Standard-abweichung	Minimum	Maximum	Variations-koeffizient
Material-einsatz	0,443	0,439	0,163	0,051	0,878	36,680
Personal-kosten	0,328	0,316	0,133	0,052	0,765	40,500
Externe Dienst-leistungen	0,048	0,027	0,055	0,001	0,331	114,660
Sonstige Kosten	0,094	0,079	0,060	0,010	0,346	64,060
Kapital-kosten	0,066	0,054	0,043	0,006	0,253	64,470

Die Standardfehler werden alternativ für jedes Szenario zusätzlich heteroskedastiekonsistent geschätzt. Die Schätzergebnisse sind in den Tabellen 21.3 und 21.4 dargestellt. Es

ist zu erkennen, dass die geschätzten Produktionselastizitäten bei Ausreißerbereinigung ungefähr den durchschnittlichen Produktionsanteilen aus Tabelle 21.1 entsprechen. Die Summe der Produktionselastizitäten bei der OLS-Schätzung mit den „bereinigten Daten" des Gesamtdatensatzes beträgt 0,984. Auch dieses Ergebnis stimmt fast mit den Resultaten von Fritsch und Stephan (2003) überein und deutet außerdem auf konstante Skalenerträge hin.

Tabelle 21.3: Cobb-Douglas-Produktionsfunktion: Schätzergebnisse für die Originalwerte der KSE 1999

Inputfaktoren	OLS-Regression	OLS-Regression mit robusten Standardfehlern	OLS-Regression	OLS-Regression mit robusten Standardfehlern
	(Daten nicht bereinigt)		(Daten bereinigt)	
Materialeinsatz	0,414	0,414	0,435	0,435
(t-Werte)	(192,24)	(79,02)	(230,41)	(149,95)
Personalkosten	0,339	0,339	0,322	0,322
(t-Werte)	(100,11)	(58,87)	(116,99)	(100,69)
Externe Dienstleistungen	0,058	0,058	0,052	0,052
(t-Werte)	(40,13)	(29,27)	(44,39)	(39,31)
Sonstige Kosten	0,114	0,114	0,105	0,105
(t-Werte)	(50,52)	(35,01)	(56,95)	(46,91)
Kapitalkosten	0,055	0,055	0,07	0,07
(t-Werte)	(22,29)	(16,14)	(33,85)	(32,44)
Konstante	1,805	1,805	1,717	1,717
(t-Werte)	(78,37)	(66,42)	(99,49)	(97,88)
Anzahl Beobachtungen	16.343	16.343	15.017	15.017
R^2	0,977	0,977	0,988	0,988

Die Schätzung mit den Originaldaten ergibt sehr hohe t-Werte und ein für Querschnittsdaten unüblich hohes Bestimmtheitsmaß von 0,988. Dies ist möglicherweise auf Endogenitätsprobleme zurückzuführen. Dennoch eröffnet das Beispiel Erkenntnisse zur Wirkungsweise von datenverändernden Anonymisierungsverfahren in linearen Modellen, insbesondere auch deshalb, weil durch die nichtlineare Transformation der Logarithmierung ein zusätzliches Problem bei der Anonymisierung sichtbar wird.

21.2 Binäres Probit-Modell zur Erklärung der Tarifbindung mit den Daten des IAB-Betriebspanels

In Deutschland können Lohnvereinbarungen, die Arbeitszeit und auch die Arbeitsbedingungen sowohl auf der individuellen als auch auf der kollektiven Ebene getroffen werden. Vereinbarungen auf der individuellen Ebene basieren auf einem Vertrag zwischen Arbeitgeber und Arbeitnehmer. Bezüglich der kollektiven Ebene kann man zwischen Verträgen auf Branche-

Tabelle 21.4: Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für die Originalwerte der KSE 1999 (ohne Wirtschaftszweig Recycling)

Inputfaktoren	OLS-Regression	OLS-Regression mit robusten Standardfehlern	OLS-Regression	OLS-Regression mit robusten Standardfehlern
	(Daten nicht bereinigt)		(Daten bereinigt)	
Materialeinsatz	0,415	0,415	0,435	0,435
(t-Werte)	(191,87)	(78,59)	(229,77)	(149,26)
Personalkosten	0,339	0,339	0,322	0,322
(t-Werte)	(99,72)	(58,61)	(116,16)	(100,1)
Externe Dienstleistungen	0,058	0,058	0,052	0,052
(t-Werte)	(39,94)	(29,02)	(44,21)	(39,26)
Sonstige Kosten	0,113	0,113	0,105	0,105
(t-Werte)	(50,34)	(34,9)	(56,84)	(46,83)
Kapitalkosten	0,055	0,055	0,071	0,071
(t-Werte)	(22,17)	(16,01)	(33,82)	(32,4)
Konstante	1,804	1,804	1,718	1,718
(t-Werte)	(78,1)	(66,19)	(99,19)	(97,7)
Anzahl Beobachtungen	16.251	16.251	14.934	14.934
R^2	0,977	0,977	0,988	0,988

nebene zwischen einer Gewerkschaft und einem Arbeitgeberverband (Branchentarifvertrag, Flächentarifvertrag) und Verträgen zwischen einer Gewerkschaft und dem einzelnen Betrieb (Haustarifvertrag) unterscheiden.

In Deutschland existiert bereits in umfangreichem Maße Literatur über die möglichen Determinanten der Tarifbindung. Datengrundlagen der empirischen Studien sind das IAB-Betriebspanel oder das Hannoveraner Firmenpanel (siehe z.B. Bellmann et al. (1999), Kohaut und Schnabel (2003) und Lehmann (2002) oder für Baden-Württemberg, Strotmann (2002)).

Die Entscheidung eines Betriebes, sich der Tarifbindung zu unterziehen, wird in diesen Studien unter Verwendung von einfachen Maximum-Likelihood-Probit-Schätzungen untersucht. Es wird angenommen, dass der latente Nutzen des Betriebes, einen Tarifvertrag zu akzeptieren (Y_i^*) durch die lineare Kombination der beobachtbaren Determinanten und eine i.i.d verteilte Störgröße U_i (für die unbeobachtete Heterogenität) beschrieben werden kann. Da Y_i^* unbeobachtbar ist, können wir nur beobachten ob sich ein Betrieb letztlich für die Tarifbindung entscheidet oder nicht. Es wird angenommen, dass der Betrieb sich für die Tarifbindung entscheidet, d.h. die beobachtete abhängige Variable Y_i den Wert Eins annimmt, wenn die latente Variable einen bestimmten Grenzwert überschreitet, der für alle Betriebe identisch ist und ohne Verlust der Allgemeingültigkeit gleich Null gesetzt wird.

$$Y_i = \begin{cases} 0 & \text{wenn } Y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + U_i \leq 0 \\ 1 & \text{wenn } Y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + U_i > 0 \end{cases} \qquad U_i \sim N(0,1) \tag{21.3}$$

Die Wahrscheinlichkeit $P(Y_i = 1)$ kann dann als Wert der Verteilungsfunktion F_{u_i} ausgedrückt werden. Unter Verwendung der Verteilungsfunktion der Standardnormalverteilung für F_{u_i} erhält man ein Probit-Modell, das mit Hilfe der Maximum-Likelihood-Methode geschätzt werden kann.

Die wichtigsten Determinanten der Tarifbindung sind die Größe des Betriebes und die Branchenzugehörigkeit. Weitere Determinanten sind z.B. Betriebsalter, das Vorhandensein eines Betriebsrats oder der Anteil von qualifizierten Angestellten.

Für die Replikationsstudie wird ein „Basismodell" geschätzt. Dabei dienen die Betriebsgröße (gemessen durch den natürlichen Logarithmus der Beschäftigung) und ein Satz von Dummy-Variablen für verschiedene Branchen als erklärende Variablen. Die abhängige Variable nimmt den Wert Eins an, falls der Betrieb tarifgebunden ist und den Wert Null, falls der Betrieb nicht tarifgebunden ist.

Für die Schätzungen und die späteren Simulationen werden vor allem die Daten des IAB-Betriebspanels für Deutschland verwendet, das für die hier betrachteten Branchen insgesamt 14.757 Betriebe umfasst. Tabelle 21.5 zeigt exemplarisch die deskriptiven Statistiken für die bei der Probit-Schätzung für den deutschen Gesamtdatensatz berücksichtigten Variablen unter Verwendung unmaskierter Daten.

Tabelle 21.5: Deskriptive Statistik, ungewichtete Daten

Variable	Mittelwert	Standard-abweichung	Minimum	Maximum
Tarifvertrag (1 = ja, 0 = nein)	0,567	0,495	0	1
Log. Beschäftigtenzahl	3,470	1,850	0	10,830
Branchen				
Verarbeitendes Gewerbe	0,274	0,446	0	1
Baugewerbe	0,094	0,292	0	1
Handel und Reperatur	0,133	0,340	0	1
Dienstleistungen	0,390	0,488	0	1
Öffentlicher Sektor	0,109	0,312	0	1

In der Tabelle 21.6 sind in der ersten Spalte die Ergebnisse der Maximum-Likelihood-Schätzung des Probit-Modells mit den Originaldaten des IAB-Betriebspanels dargestellt. Die Unternehmensgröße hat, wie erwartet, einen hochsignifikant positiven Einfluss auf die Tarifbindung. Die Dummy-Variablen für die Branchen sind auf dem 1%-Signifikanzniveau

gemeinsam signifikant. Im Vergleich zum Verarbeitenden Gewerbe sind Betriebe in der Öffentlichen Verwaltung, dem Baugewerbe und dem Handel häufiger tarifgebunden, während zwischen dem Dienstleistungssektor und dem Verarbeitenden Gewerbe keine Unterschiede zu verzeichnen sind.

Tabelle 21.6: ML-Probitschätzung zur Erklärung der Tarifbindung mit Orignaldaten

	Abhängige Variable: Tarifbindung (1 = ja, 0 = nein)			
	Deutschland		Baden-Württemberg	
	vollständig	50%-Stichprobe	vollständig	50%-Stichprobe
	14.757 Betriebe	7.386 Betriebe	1.201 Betriebe	601 Betriebe
Konstante	-1,126	-1,143	-0,988	-1,044
(t-Werte)	(-33,76)	(-24,28)	(7,46)	(5,63)
Log. Beschäftigung	0,318	0,319	0,301	0,301
(t-Werte)	(46,16)	(32,80)	(12,24)	(8,94)
Branchen				
Bau	0,668	0,609	0,748	0,683
(t-Werte)	(15,86)	(10,25)	(4,46)	(10,25)
Handel	0,306	0,311	0,379	0,535
(t-Werte)	(8,29)	(6,00)	(2,88)	(2,93)
Dienstleistungen	0,081	0,054	0,040	0,102
(t-Werte)	(2,92)	(1,38)	(0,40)	(0,71)
Verwaltung	0,966	1,007	0,796	0,748
(t-Werte)	(20,49)	(15,06)	(4,60)	(3,24)
Beobachtungen	14.757	7.386	1.201	601
Pseudo R^2	0,161	0,164	0,140	0,140
LR-test	3242,89	1658,89	222,78	111,45
Log likelihood	-8479,31	-4243,18	-686,46	-343,68

Unter anderem um einen möglichen Einfluss der Größe des Datensatzes auf die inhaltlichen Schlussfolgerungen herauszuarbeiten, wird alternativ auch mit der Baden-Württemberg-Stichprobe des IAB-Betriebspanels gearbeitet, die 1.201 Betriebe enthält. Sowohl für Deutschland als auch für Baden-Württemberg werden ergänzend auch jeweils 50%-Zufallsstichproben gezogen und die entsprechenden Analysen für diese durchgeführt. Wie man Tabelle 21.6 entnehmen kann, resultieren hieraus einige Unterschiede in den Koeffizientenschätzungen, insbesondere bei den Dummy-Variablen. Man erkennt, dass Stichprobenziehungen einen relevanten Einfluss auf die Schätzergebnisse (auch auf die Koeffizientenschätzer) haben können (vgl. auch Kapitel 7). Dennoch bleiben die Schlussfolgerungen hinsichtlich der Einflüsse der einzelnen Regressoren hier unverändert. Die geringere Stichprobengröße schlägt sich insbesondere darin nieder, dass die geschätzten Koeffizienten größere Standardfehler aufweisen und die t-Werte daher geringer ausfallen.

Kapitel 22

Stochastische Überlagerungen in linearen und nichtlinearen ökonometrischen Modellen

22.1 Theoretische Eigenschaften

22.1.1 Stochastische Überlagerungen in linearen Modellen

Das lineare Regressionsmodell ist durch

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{22.1}$$

gegeben.

Dabei sind $\mathbf{y}$ und $\mathbf{u}$ n-dimensionale Zufallsvektoren, $\boldsymbol{\beta}$ ist ein K-dimensionaler Parametervektor und $\mathbf{X}$ ist eine nichtstochastische $(n \times K)$-Matrix, die die Werte der K Einflussvariablen enthält.

Im klassischen Modell gelten für den Störvektor $\mathbf{u}$ die folgenden Annahmen:

$$E[\mathbf{u}] = \mathbf{0}, \tag{22.2}$$

$$cov[\mathbf{u}] \equiv \boldsymbol{\Sigma}_{uu} = \sigma^2\mathbf{I}. \tag{22.3}$$

Dabei ist $\mathbf{I}$ die Einheitsmatrix.

Der Kleinste-Quadrate-Schätzer ergibt sich als

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{22.4}$$

Der Schätzer ist konsistent, das heißt es gilt:

$$plim\left(\hat{\boldsymbol{\beta}}\right) = \boldsymbol{\beta}. \tag{22.5}$$

Dies kann man zeigen, indem der KQ-Schätzer (Gleichung (22.4)) wie folgt umgeschrieben wird:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u} \tag{22.6}$$

oder

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\left(\frac{\mathbf{X}'\mathbf{u}}{n}\right). \tag{22.7}$$

Der Schätzer $\hat{\boldsymbol{\beta}}$ ist somit dann konsistent, wenn die rechte Seite der Gleichung (22.7) gegen Null konvergiert.

Für den ersten Term auf der rechten Seite der Gleichung (22.7) kann man folgendes asymptotische Verhalten annehmen:

$$\lim_{n\to\infty}\frac{\mathbf{X}'\mathbf{X}}{n} = \mathbf{Q}. \tag{22.8}$$

Dabei ist $\mathbf{Q}$ eine feste (nicht stochastische) positiv definite Matrix vom Rang K.

Damit lässt sich zeigen, dass der Schätzer dann konsistent ist, wenn für den Wahrscheinlichkeitsgrenzwert des zweiten Terms gilt:

$$plim\left(\frac{\mathbf{X}'\mathbf{u}}{n}\right) = \mathbf{0}. \tag{22.9}$$

Es lässt sich zeigen, dass dies genau dann gilt, wenn die Regressormatrix $\mathbf{X}$ und der Störterm $\mathbf{u}$ asymptotisch unkorreliert sind. Das heißt, es muss gelten:

$$\lim_{n\to\infty} cov\left(\mathbf{X}'\mathbf{u}\right) = \mathbf{0}, \tag{22.10}$$

während für die Unverzerrtheit des Schätzers im Kleinstichprobenfall die strengere Anforderung

$$cov\,(\mathbf{X}'\mathbf{u}) = \mathbf{0} \tag{22.11}$$

gelten muss. Beide Anforderungen sind im klassischen Regressionsmodell erfüllt.

Die unbekannte Restvarianz σ^2 wird erwartungstreu durch

$$\hat{\sigma}^2 = \frac{1}{n-K}\hat{\mathbf{u}}'\hat{\mathbf{u}} \tag{22.12}$$

geschätzt.

Dabei ist $\hat{\mathbf{u}}$ der Vektor der KQ-Residuen.

Demgegenüber gilt im verallgemeinerten Modell für die Varianz-Kovarianzmatrix statt Gleichung (22.3):

$$cov\,[\mathbf{u}] \equiv \mathbf{\Sigma}_{uu} = \sigma^2\mathbf{\Omega}. \tag{22.13}$$

Damit ist der KQ-Schätzer aus (22.4) nach wie vor erwartungstreu und konsistent, jedoch nicht mehr effizient. Diese Eigenschaft besitzt nun der GLS-Schätzer

$$\hat{\boldsymbol{\beta}}^{GLS} = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}. \tag{22.14}$$

Für die Schätzung der unbekannten Restvarianz σ^2 werden dann die Residuen aus der GLS-Schätzung verwendet. Wird die Matrix Ω geschätzt, so wird der Schätzer als FGLS-Schätzer (feasible GLS-Schätzer) bezeichnet.

Die Effekte der stochastischen Überlagerungen entsprechen nun den Auswirkungen von zufälligen Messfehlern in den Variablen. Dabei wird im Folgenden wie bereits bei der Vorstellung der Verfahren in Unterabschnitt 6.2.3 zwischen additiven und multiplikativen stochastischen Überlagerungen unterschieden.

a) Additive Stochastische Überlagerungen in linearen Modellen

Für einen additiv stochastisch überlagerten Vektor aus Variablenwerten gilt:

$$\mathbf{x}^a = \mathbf{x} + \mathbf{w}. \tag{22.15}$$

Wird der gesamte Datensatz oder ein Teildatensatz mit mehreren Variablen stochastisch überlagert, so gilt (vgl. Gleichung (6.13)):

$$\mathbf{X}^a = \mathbf{X} + \mathbf{W}. \tag{22.16}$$

Dabei werden für die zur Anonymisierung verwendeten Zufallsfehler die Annahmen aus Unterabschnitt 6.2.3 zugrundegelegt. Insbesondere wird ein Erwartungswert von Null angenommen. Die Überlagerungen sind mit den Regressoren sowie mit dem Störterm des Modells unkorreliert. Ferner wird zunächst unterstellt, dass die Störterme, mit denen die einzelnen Variablen überlagert werden, stochastisch unabhängig sind. Das Absolutglied wird nicht überlagert. Damit weisen die Komponenten des ersten Spaltenvektors der Überlagerungsmatrix $\mathbf{W}$ eine Varianz von Null auf.

Werden die additiv überlagerten Werte für die Schätzung verwendet, so schätzt man statt des „wahren“ Modells das folgende Modell:

$$\mathbf{y}^a = \mathbf{X}^a\boldsymbol{\beta}^a + \mathbf{u}^a. \tag{22.17}$$

Dabei bezeichnet $\mathbf{y}^a(\mathbf{y}+\mathbf{v})$ den Vektor der anonymisierten abhängigen Variablen, $\mathbf{X}^a(\mathbf{X}+\mathbf{W})$ die Matrix der anonymisierten Regressoren und $\mathbf{u}^a$ den Vektor der Störterme bei Verwendung der anonymisierten Werte.

Setzt man dieses Ergebnis in das wahre Modell (Gleichung (22.1)) ein, so erhält man

$$\mathbf{y}^a - \mathbf{v} = (\mathbf{X}^a - \mathbf{W})\boldsymbol{\beta} + \mathbf{u} \tag{22.18}$$

oder auch

$$\mathbf{y}^a = \mathbf{X}^a\boldsymbol{\beta} + \mathbf{u} - \mathbf{v} - \mathbf{W}\boldsymbol{\beta}. \tag{22.19}$$

Vergleicht man dieses Ergebnis mit der Gleichung (22.17), so kann man erkennen, dass nach wie vor der „richtige“ Parametervektor modelliert wird, allerdings der Störterm nun mit

$$\mathbf{u}^a = (\mathbf{u} - \mathbf{v} - \mathbf{W})\boldsymbol{\beta} \tag{22.20}$$

nicht mehr unabhängig vom Parametervektor $\boldsymbol{\beta}$ und der Regressormatrix $\mathbf{X}$ ist.

Additive Überlagerung der abhängigen Variablen

Wird lediglich die abhängige Variable Y anonymisiert, so schätzt man das Modell

$$\mathbf{y}^a = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}^a \tag{22.21}$$

mit

$$\mathbf{u}^a = \mathbf{u} - \mathbf{v}. \tag{22.22}$$

Dabei ist der Störterm in der Gleichung (22.22) wiederum unabhängig von der Regressormatrix. Und es gilt:

$$cov\,(\mathbf{X}'\mathbf{u}^a) = \mathbf{0} \tag{22.23}$$

und somit auch

$$\lim_{n \to \infty} cov\,(\mathbf{X}'\mathbf{u}^a) = \mathbf{0}. \tag{22.24}$$

Damit ist der KQ-Schätzer in diesem Fall sowohl konsistent als auch im Kleinstichprobenfall erwartungstreu.

Additive Überlagerung der Regressoren

Werden hingegen die Einflussgrößen in Form der Regressormatrix $\mathbf{X}$ anonymisiert, so ergibt sich das zu schätzende Modell mit

$$\mathbf{y} = \mathbf{X}^a\boldsymbol{\beta}^a + \mathbf{u}^a. \tag{22.25}$$

Der KQ-Schätzer für dieses Modell lautet somit:

$$\hat{\boldsymbol{\beta}}^a = \left(\mathbf{X}^{a'}\mathbf{X}^a\right)^{-1}\mathbf{X}^{a'}\mathbf{y}. \tag{22.26}$$

oder

$$\hat{\boldsymbol{\beta}}^a = ((\mathbf{X}+\mathbf{W})'(\mathbf{X}+\mathbf{W}))^{-1}(\mathbf{X}+\mathbf{W})'\mathbf{y}. \tag{22.27}$$

Um die asymptotischen Eigenschaften des KQ-Schätzers zu untersuchen, kann man Gleichung (22.27) auch wie folgt umschreiben:

$$\hat{\boldsymbol{\beta}}^a = \left(\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{X}+\mathbf{W})}{n}\right)^{-1}\frac{(\mathbf{X}+\mathbf{W})'\mathbf{y}}{n}. \tag{22.28}$$

Für den ersten Term auf der rechten Seite gilt:

$$\left(\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{X}+\mathbf{W})}{n}\right) = \frac{1}{n}\left[\mathbf{X}'\mathbf{X}+\mathbf{X}'\mathbf{W}+\mathbf{W}'\mathbf{X}+\mathbf{W}'\mathbf{W}\right]. \tag{22.29}$$

Für den Wahrscheinlichkeitsgrenzwert dieses Terms gilt somit:

$$\begin{aligned} plim\left(\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{X}+\mathbf{W})}{n}\right) &= plim\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)+plim\left(\frac{\mathbf{X}'\mathbf{W}}{n}\right)+ \\ &\quad +plim\left(\frac{\mathbf{W}'\mathbf{X}}{n}\right)+plim\left(\frac{\mathbf{W}'\mathbf{W}}{n}\right). \end{aligned} \tag{22.30}$$

Da die zur Anonymisierung verwendeten additiven Zufallsfehler Erwatungswert Null haben und mit den Regressoren unkorreliert sind, gilt:

$$\lim_{n\to\infty} E(\mathbf{X}'\mathbf{W}) = \lim_{n\to\infty} E(\mathbf{W}'\mathbf{X}) = \mathbf{0} \tag{22.31}$$

und

$$\lim_{n\to\infty} var(\mathbf{X}'\mathbf{W}) = \lim_{n\to\infty} var(\mathbf{W}'\mathbf{X}) = \mathbf{0} \tag{22.32}$$

Damit konvergieren die Ausdrücke $\mathbf{X}'\mathbf{W}$ und $\mathbf{W}'\mathbf{X}$ im quadratischen Mittel gegen Null und es gilt:

$$plim\left(\frac{\mathbf{X}'\mathbf{W}}{n}\right) = plim\left(\frac{\mathbf{W}'\mathbf{X}}{n}\right) = \mathbf{0} \tag{22.33}$$

Für den Wahrscheinlichkeitsgrenzwert von $\frac{\mathrm{W}'\mathrm{W}}{n}$ gilt:

$$plim\left(\frac{\mathbf{W}'\mathbf{W}}{n}\right) = \mathbf{\Sigma}_{ww}. \tag{22.34}$$

Und somit ergibt sich für den Wahrscheinlichkeitsgrenzwert des ersten Terms auf der rechten Seite der Gleichung (22.28):

$$plim\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right) + plim\left(\frac{\mathbf{X}'\mathbf{W}}{n}\right) + plim\left(\frac{\mathbf{W}'\mathbf{X}}{n}\right) + plim\left(\frac{\mathbf{W}'\mathbf{W}}{n}\right) = \mathbf{Q} + \mathbf{\Sigma}_{ww}. \tag{22.35}$$

Für den zweiten Term auf der rechten Seite der Gleichung (22.28) gilt:

$$\begin{aligned} \frac{1}{n}(\mathbf{X}+\mathbf{W})'\mathbf{y} &= \frac{1}{n}\left[(\mathbf{X}+\mathbf{W})'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})\right] \\ &= \frac{1}{n}\left[\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + \mathbf{W}'\mathbf{X}\boldsymbol{\beta} + \mathbf{X}'\mathbf{u} + \mathbf{W}'\mathbf{u}\right] \end{aligned} \tag{22.36}$$

Es wurde bereits gezeigt, dass

$$plim\left(\frac{\mathbf{W}'\mathbf{X}}{n}\right) = \mathbf{0} \tag{22.37}$$

und damit auch

$$plim\left(\frac{\mathbf{W}'\mathbf{X}\boldsymbol{\beta}}{n}\right) = \mathbf{0} \tag{22.38}$$

gilt.

Ebenso wurde gezeigt, dass

$$plim\left(\frac{\mathbf{X}'\mathbf{u}}{n}\right) = \mathbf{0} \tag{22.39}$$

gilt.

Da die zur Anonymisierung verwendete stochastische Fehlermatrix $\mathbf{W}$ unabhängig vom Fehlervektor $\mathbf{u}$ ist und alle Fehlerkomponenten einen Erwartungswert Null haben, gilt auch hier,

dass der Ausdruck $\mathbf{W}'\mathbf{u}$ im quadratischen Mittel gegen Null konvergiert. Daraus wiederum folgt:

$$plim\,(\mathbf{W}'\mathbf{u}) = \mathbf{0}. \tag{22.40}$$

Somit folgt für den Wahrscheinlichkeitsgrenzwert des zweiten Terms auf der rechten Seite von Gleichung (22.28):

$$plim\left(\frac{(\mathbf{X}+\mathbf{W})'\,\mathbf{y}}{n}\right) = \mathbf{Q}\boldsymbol{\beta} \tag{22.41}$$

Und damit gilt für den Wahrscheinlichkeitsgrenzwert des Schätzers $\hat{\boldsymbol{\beta}}^a$

$$plim\,\hat{\boldsymbol{\beta}}^a = (\mathbf{Q}+\boldsymbol{\Sigma}_{ww})^{-1}\,\mathbf{Q}\boldsymbol{\beta} \tag{22.42}$$

oder

$$plim\,\hat{\boldsymbol{\beta}}^a = \boldsymbol{\beta} - (\mathbf{Q}+\boldsymbol{\Sigma}_{ww})^{-1}\,\boldsymbol{\Sigma}_{ww}\boldsymbol{\beta}. \tag{22.43}$$

Somit ist der KQ-Schätzer inkonsistent, wenn die Regressoren mit einem additiven stochastischen Fehler überlagert werden. Dabei ist es für die Zerstörung der Inkonsistenz des KQ-Schätzers ausreichend, wenn ein Regressor mit einem additiven Fehler überlagert wird.

Brand (2000) zeigt, dass die asymptotische Verzerrung der Koeffizientenschätzung mit der Stärke der Überlagerung zunimmt. Hierzu schreibt sie Gleichung (22.42) wie folgt um:

$$plim\,\hat{\boldsymbol{\beta}}^a = \left(\mathbf{Q}+\alpha\;diag(0,\sigma_2^2,\ldots,\sigma_K^2)\right)^{-1}\mathbf{Q}\boldsymbol{\beta} \tag{22.44}$$

Dabei wird ausgenutzt, dass die Störterme für die einzelnen Variablen voneinander unabhängig sind.

Somit gilt auch

$$plim\,\hat{\boldsymbol{\beta}}^a = \left(\mathbf{I}_K+\alpha\mathbf{Q}^{-1}diag(0,\sigma_2^2,\ldots,\sigma_K^2)\right)^{-1}\boldsymbol{\beta} \tag{22.45}$$

Dabei ist $\mathbf{I}_K$ die $(K\times K)$-Einheitsmatrix.

Brand (2000) nutzt nun aus, dass die Schätzung des $((K-1)\times K)$-Vektors der Steigungsparameter $\boldsymbol{\beta}$ in inhomogenen Regressionsmodellen identisch zur Schätzung der Parameter

im zentrierten Modell ist. Für den Wahrscheinlichkeitsgrenzwert des Koeffizientenschätzers im zentrierten Modell erhält man aufgrund der Identität der Momentenmatrix der zentrierten Werte zur Kovarianzmatrix der nicht zentrierten Werte:

$$plim\hat{\boldsymbol{\beta}}_1^a = \left(\mathbf{I}_{K-1} + \alpha\boldsymbol{\Sigma}_{x_r x_r}^{-1} diag(0, \sigma_2^2, \ldots, \sigma_K^2)\right)^{-1}\boldsymbol{\beta}_1 \tag{22.46}$$

mit $\boldsymbol{\Sigma}_{x_r x_r} = \frac{1}{n}\left(\mathbf{X}_r - \overline{\mathbf{X}}_r\right)'\left(\mathbf{X}_r - \overline{\mathbf{X}}_r\right)$.

Für den Fall unkorrelierter Regressoren ergibt sich somit (Brand 2000):

$$plim\hat{\boldsymbol{\beta}}_1^a = \frac{1}{1+\alpha}\boldsymbol{\beta}_1 \tag{22.47}$$

Damit wird die absolute Höhe der Koeffizienten tendenziell unterschätzt. Das bedeutet, dass positive Koeffizienten tendenziell unterschätzt, negative Koeffizienten hingegen tendenziell überschätzt werden.

Additive Überlagerung der Regressoren und der abhängigen Variablen

Werden sowohl die abhängige Variable als auch die Regressoren additiv überlagert, so ergibt sich für den Schätzer:

$$\hat{\boldsymbol{\beta}}^{a2} = \left(\mathbf{X}^{a\prime}\mathbf{X}^a\right)^{-1}\mathbf{X}^{a\prime}\mathbf{y}^a. \tag{22.48}$$

oder

$$\hat{\boldsymbol{\beta}}^{a2} = \left(\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{X}+\mathbf{W})}{n}\right)^{-1}\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{y}+\mathbf{v})}{n}. \tag{22.49}$$

Daraus ergibt sich für den Wahrscheinlichkeitsgrenzwert des Schätzers $\hat{\boldsymbol{\beta}}^{a2}$:

$$plim\hat{\boldsymbol{\beta}}^{a2} = plim\left(\left(\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{X}+\mathbf{W})}{n}\right)^{-1}\right)\ plim\left(\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{y}+\mathbf{v})}{n}\right). \tag{22.50}$$

Für den ersten Term auf der rechten Seite der Gleichung gilt wiederum:

$$plim\left(\left(\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{X}+\mathbf{W})}{n}\right)^{-1}\right) = \left(\mathbf{Q}+\boldsymbol{\Sigma}_{xx}\right)^{-1}. \tag{22.51}$$

Für den zweiten Term gilt nun:

$$\begin{aligned} plim\left(\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{y}+\mathbf{v})}{n}\right) &= plim\left(\frac{\mathbf{X}'\mathbf{y}}{n}+\frac{\mathbf{X}'\mathbf{v}}{n}+\frac{\mathbf{W}'\mathbf{y}}{n}+\frac{\mathbf{W}'\mathbf{v}}{n}\right) \\ &= plim\left(\frac{\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})}{n}+\frac{\mathbf{X}'\mathbf{v}}{n}+\frac{\mathbf{W}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u})}{n}+\frac{\mathbf{W}'\mathbf{v}}{n}\right) \\ &= plim\left(\frac{\mathbf{X}'\mathbf{X}\boldsymbol{\beta}}{n}+\frac{\mathbf{X}'\mathbf{u}}{n}+\frac{\mathbf{X}'\mathbf{v}}{n}+\frac{\mathbf{W}'\mathbf{X}\boldsymbol{\beta}}{n}+\frac{\mathbf{W}'\mathbf{u}}{n}+\frac{\mathbf{W}'\mathbf{v}}{n}\right) \\ &= \mathbf{Q}\boldsymbol{\beta}+plim\left(\frac{\mathbf{W}'\mathbf{v}}{n}\right). \end{aligned} \tag{22.52}$$

Geht man wieder davon aus, dass die additiven Störterme unkorreliert sind, so erhält man das Ergebnis aus Gleichung (22.41) und damit das gleiche Ergebnis wie im Fall der ausschließlichen Überlagerung der Regressoren. Auch in diesem Fall ist der KQ-Schätzer nicht konsistent.

Sind die Störterme hingegen korreliert, so ergibt sich:

$$plim\left(\frac{(\mathbf{X}+\mathbf{W})'(\mathbf{y}+\mathbf{v})}{n}\right)=\mathbf{Q}\boldsymbol{\beta}+\boldsymbol{\Sigma}_{wv}. \tag{22.53}$$

Damit ergibt sich in diesem Fall für den Wahrscheinlichkeitsgrenzwert des Schätzers $\hat{\boldsymbol{\beta}}^{a2}$:

$$plim\,\hat{\boldsymbol{\beta}}^{a2}=\left(\mathbf{Q}+\boldsymbol{\Sigma}_{ww}\right)^{-1}\left(\mathbf{Q}\boldsymbol{\beta}+\boldsymbol{\Sigma}_{wv}\right). \tag{22.54}$$

In Unterabschnitt 6.2.3 wurde bereits darauf hingewiesen, dass eine Anpassung der Varianz-Kovarianzmatrix der Überlagerungen an die Varianz-Kovarianzmatrix der zu anonymisierenden Variablen dazu führt, dass der KQ-Schätzer im linearen Modell konsistent ist, sofern alle Variablen additiv überlagert werden.

Es gilt dann für die Varianz-Kovarianzmatrix der Störterme gemäß Gleichung (6.15):

$$\boldsymbol{\Sigma}_{ww}=d\,\boldsymbol{\Sigma}_{xx} \tag{22.55}$$

und damit für die Varianz-Kovarianzmatrix der überlagerten Variablen gemäß Gleichung (6.16):

$$\boldsymbol{\Sigma}_{x^a x^a}=(1+d)\,\boldsymbol{\Sigma}_{xx}. \tag{22.56}$$

Setzt man dies in Gleichung (22.54) ein, so folgt daraus für den Wahrscheinlichkeitsgrenzwert des KQ-Schätzers:

$$plim\hat{\boldsymbol{\beta}}^{a2} = (\mathbf{Q} + d\boldsymbol{\Sigma}_{xx})^{-1}(\mathbf{Q}\boldsymbol{\beta} + d\boldsymbol{\Sigma}_{xy}). \tag{22.57}$$

Außerdem gilt:

$$\mathbf{Q}\boldsymbol{\beta} = \mathbf{M}_{xy}. \tag{22.58}$$

Dabei ist $\mathbf{M}_{xy}$ der Momentenvektor $plim\left(\frac{\mathbf{X}'\mathbf{y}}{n}\right)$ (Brand 2000).

Somit gilt:

$$plim\hat{\boldsymbol{\beta}}^{a2} = (\mathbf{Q} + d\boldsymbol{\Sigma}_{xx})^{-1}(\mathbf{M}_{xy} + d\boldsymbol{\Sigma}_{xy}). \tag{22.59}$$

Brand (2000) greift nun wiederum auf die Identität der Kovarianzmatrix der echten Regressoren des Ausgangsmodells mit der Momentenmatrix der Regressoren des zentrierten Modells zurück und erhält:

$$plim\hat{\boldsymbol{\beta}}_1^{a2} = (\boldsymbol{\Sigma}_{x_r x_r} + d\boldsymbol{\Sigma}_{x_r x_r})^{-1}(\boldsymbol{\Sigma}_{x_r y} + d\boldsymbol{\Sigma}_{x_r y}) \tag{22.60}$$

und somit

$$plim\hat{\boldsymbol{\beta}}_1^{a2} = ((1+d)\boldsymbol{\Sigma}_{x_r x_r})^{-1}((1+d)\boldsymbol{\Sigma}_{x_r y}) = \boldsymbol{\beta}_1. \tag{22.61}$$

Dabei bezeichnen der Index r die zentrierten Regressorvariablen und $\boldsymbol{\beta}_1$ den Koeffizientenvektor ohne das Absolutglied.

Folglich ist der KQ-Schätzer im Fall additiver stochastischer Überlagerungen konsistent, wenn alle Variablen überlagert werden, und zwar mit einer Matrix von Störgrößen, deren Varianz-Kovarianzmatrix proportional zur Varianz-Kovarianzmatrix der Ausgangsvariablen ist.[23]

23) Würde man die Momentenmatrix und damit die Varianz-Kovarianzmatrix der Überlagerungen proportional zur Momentenmatrix der Ausgangsvariablen wählen, würde also gelten $\Sigma_{ww} = d\mathrm{Q}$ und $\Sigma_{wv} = d\mathrm{M}_{xy}$, so könnte man die Konsistenz des Schätzers auch direkt aus Gleichung (22.54) ablesen. Für diese würde dann nämlich gelten: $plim\hat{\boldsymbol{\beta}}^{a2} = (\mathrm{Q} + d\mathrm{Q})^{-1}(\mathrm{M}_{xy} + d\mathrm{M}_{xy})$ und damit $plim\hat{\boldsymbol{\beta}}^{a2} = \mathrm{Q}^{-1}\mathrm{M}_{xy}$.

Für die Beurteilung der Wirkung eines Anonymisierungsverfahrens in linearen Modellen ist jedoch nicht nur die Erwartungstreue beziehungsweise Konsistenz des Schätzers von Interesse, sondern auch das Verhalten der Teststatistiken. In die Teststatistiken geht jedoch neben den geschätzten Koeffizientenwerten die geschätzte Varianz der Störgrößen ein.

Üblicherweise wird die Varianz der Störgrößen erwartungstreu und konsistent geschätzt durch (vgl. Gleichung (22.12))

$$\hat{\sigma}_u^2 = \frac{1}{n-K}\hat{\mathbf{u}}'\hat{\mathbf{u}}. \tag{22.62}$$

Brand (2000) verwendet nun, dass die Residuen $\hat{\mathbf{u}}$ identisch sind mit den Residuen des zentrierten Modells:

$$\hat{\mathbf{u}} = \hat{\mathbf{u}}_r = (\mathbf{y} - \overline{\mathbf{y}}) - \left(\mathbf{X}_r - \overline{\mathbf{X}}_r\right)'\hat{\boldsymbol{\beta}}_1. \tag{22.63}$$

Damit gilt für die Schätzgleichung:

$$\hat{\sigma}_u^2 = \frac{1}{n-K}\hat{\mathbf{u}}_r'\hat{\mathbf{u}}_r. \tag{22.64}$$

Brand (2000) schreibt im Folgenden den Schätzer weiter um zu

$$\hat{\sigma}_u^2 = \left(\hat{\sigma}_y^2 - \hat{\boldsymbol{\Sigma}}'_{x_r y}\left(\hat{\boldsymbol{\Sigma}}_{x_r x_r}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{x_r y}\right) \tag{22.65}$$

mit $\hat{\sigma}_y^2 = \frac{1}{n-K}(\mathbf{y} - \overline{\mathbf{y}})'(\mathbf{y} - \overline{\mathbf{y}})$.

Der Wahrscheinlichkeitsgrenzwert ergibt sich damit als

$$plim\,\hat{\sigma}_u^2 = \left(\sigma_y^2 - \boldsymbol{\Sigma}'_{x_r y}\left(\boldsymbol{\Sigma}_{x_r x_r}\right)^{-1}\boldsymbol{\Sigma}_{x_r y}\right) = \sigma_u^2. \tag{22.66}$$

Im Fall der additiven Überlagerung aller Variablen wird jedoch die Varianz der Störgrößen aus dem fehlerbehafteten Modell geschätzt. Es ergibt sich:

$$\hat{\sigma}_u^{a2^2} = \frac{1}{n-K}\hat{\mathbf{u}}^{a2'}\hat{\mathbf{u}}^{a2} \tag{22.67}$$

mit $\hat{\mathbf{u}}^{a2} = \mathbf{y}^a - \mathbf{X}^{a'}\hat{\boldsymbol{\beta}}$.

Analog zum Originalmodell schreibt Brand (2000) den Schätzer um zu

$$\left(\widehat{\sigma_u^{a2}}\right)^2 = \left(\hat{\sigma}_{y^a}^2 - \hat{\mathbf{\Sigma}}'_{x_r^a y^a} \left(\hat{\mathbf{\Sigma}}_{x_r^a x_r^a}\right)^{-1} \hat{\mathbf{\Sigma}}_{x_r^a y^a}\right). \tag{22.68}$$

Für den Wahrscheinlichkeitsgrenzwert ergibt sich somit:

$$plim\left(\widehat{\sigma_u^{a2}}\right)^2 = \left(\sigma_{y^a}^2 - \mathbf{\Sigma}'_{x_r^a y^a} \left(\mathbf{\Sigma}_{x_r^a x_r^a}\right)^{-1} \mathbf{\Sigma}_{x_r^a y^a}\right). \tag{22.69}$$

Berücksichtigt man, dass die Varianz-Kovarianzmatrix der Überlagerungen proportional zur Varianz-Kovarianzmatrix der Originalvariablen gewählt wurde, so erhält man schließlich

$$plim\left(\widehat{\sigma_u^{a2}}\right)^2 = (1+d)\left(\sigma_y^2 - \mathbf{\Sigma}'_{x_r y} \left(\mathbf{\Sigma}_{x_r x_r}\right)^{-1} \mathbf{\Sigma}_{x_r y}\right) = (1+d)\sigma_u^2. \tag{22.70}$$

Damit wird die Varianz der Störgrößen umso stärker überschätzt, je höher die relative Stärke der Überlagerung ist (Brand 2000). Allerdings lässt sich – zumindest für den Fall großer Stichproben und damit auch für die Projektstatistiken – eine Korrektur durchführen, wenn dem Nutzer die Stärke der Überlagerung d bekannt ist. Wird – wie von Kim (1986) vorgeschlagen – die Varianz-Kovarianzmatrix der additiv überlagerten Werte durch Transformation so verändert, dass sie der Varianz-Kovarianzmatrix der Originalvariablen entspricht, so kann im linearen Modell die Varianz der Störgrößen asymptotisch erwartungstreu geschätzt werden. Damit bleiben die Teststatistiken für diesen Fall erhalten.

Korrekturverfahren bei additiver Überlagerung

Wird die Varianz-Kovarianzmatrix der Überlagerungen nicht proportional zur Varianz-Kovarianzmatrix der Originalvariablen gewählt oder werden nicht alle Variablen additiv überlagert, so existieren unterschiedliche Möglichkeiten, um den verzerrten Schätzer zu korrigieren.

Aus der Gleichung (22.19) ergibt sich für den Fall, dass ausschließlich die Regressoren additiv überlagert werden:

$$\mathbf{y} = \mathbf{X}^a\boldsymbol{\beta} + \mathbf{u} - \mathbf{W}\boldsymbol{\beta}. \tag{22.71}$$

Eine Möglichkeit besteht nun in der Verwendung des so genannten Instrumentvariablen-Schätzers (IV-Schätzer). Dabei werden alle Terme mit den in einer $(n \times K)$-Matrix $\mathbf{Z}$ zusammengefassten Instrumenten prämultipliziert. Dabei müssen die Instrumentvariablen

asymptotisch unkorreliert mit den Messfehlern u_i und den Überlagerungen w_{ij} und hoch korreliert mit den Regressoren sein.

Es muss also gelten:

$$plim\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right) = \mathbf{Q}_{zx} \neq \mathbf{0}, \tag{22.72}$$

$$plim\left(\frac{\mathbf{Z}'\mathbf{u}}{n}\right) = \mathbf{0} \tag{22.73}$$

und

$$plim\left(\frac{\mathbf{Z}'\mathbf{W}}{n}\right) = \mathbf{0}. \tag{22.74}$$

Für die Prämultiplikation der Gleichung (22.71) gilt somit:

$$\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}^a\boldsymbol{\beta} + \mathbf{Z}'\mathbf{u} - \mathbf{Z}'\mathbf{W}\boldsymbol{\beta}. \tag{22.75}$$

Geht man zu den Wahrscheinlichkeitsgrenzwerten über, so erhält man:

$$plim\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = plim\left(\frac{\mathbf{Z}'\mathbf{X}\boldsymbol{\beta}}{n}\right) + plim\left(\frac{\mathbf{Z}'\mathbf{W}\boldsymbol{\beta}}{n}\right) + plim\left(\frac{\mathbf{Z}'\mathbf{u}}{n}\right) - plim\left(\frac{\mathbf{Z}'\mathbf{W}\boldsymbol{\beta}}{n}\right). \tag{22.76}$$

Unter den obigen Annahmen ergibt sich:

$$plim\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = plim\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\boldsymbol{\beta}. \tag{22.77}$$

Der IV-Schätzer ergibt sich somit als

$$\hat{\boldsymbol{\beta}}^{IV} \quad = \quad (\mathbf{Z}'\,\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{22.78}$$

Bei der Anonymisierung wirtschaftsstatistischer Einzeldaten kann den Nutzern ein Datensatz mit den Instrumentvariablen ganz einfach dadurch zur Verfügung gestellt werden, dass der Originaldatenbestand zweimal mit unkorrelierten Störtermen überlagert wird und die Nutzer somit zwei anonymisierte Datensätze erhalten. Allerdings kann dieses Vorgehen zu einer Erhöhung des Reidentifikationsrisikos führen.

Alternativ kann eine Korrektur des Fehler-in-den-Variablen-Schätzers $\hat{\boldsymbol{\beta}}^{a}$ vorgenommen werden (Lechner und Pohlmeier 2003).

$$\hat{\boldsymbol{\beta}}^{a^{Corr}} = \left(\mathbf{I}_K - (\mathbf{M}_{xx} + \boldsymbol{\Sigma}_{ww})^{-1}\boldsymbol{\Sigma}_{ww}\right)^{-1} \hat{\boldsymbol{\beta}}^{a} \tag{22.79}$$

mit $\mathbf{M}_{xx} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ und $\mathbf{I}_K$ der K-dimensionalen Einheitsmatrix.

Dabei könnte ausgenutzt werden, dass

$$\frac{1}{n}\mathbf{X}^{a'}\,\mathbf{X}^{a} = \hat{\mathbf{M}}_{xx} + \hat{\boldsymbol{\Sigma}}_{ww} \tag{22.80}$$

gilt.

Also müsste zusätzlich nur eine Information beziehungsweise Schätzung für $\boldsymbol{\Sigma}_{ww}$ zur Verfügung gestellt werden. Wenn alle Störvariablen unkorreliert sind, würde dies bedeuten, dass nur Varianzen aus $\boldsymbol{\Sigma}_{ww}$, bekannt sein müssten.

Eine operationale Form dieses Schätzers lautet demnach

$$\hat{\boldsymbol{\beta}}^{a^{CorrOp}} = \left(\mathbf{I}_K - \left(\frac{1}{n}\,\mathbf{X}^{a'}\,\mathbf{X}^{a}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{ww}\right)^{-1} \hat{\boldsymbol{\beta}}^{a}. \tag{22.81}$$

Für den Fall, dass sowohl die abhängige Variable als auch die Regressormatrix additiv stochastisch überlagert wird, der Schätzer dennoch (asymptotisch) verzerrt ist, kann analog zu Gleichung (22.78) der Instrumentvariablen-Schätzer verwendet werden.

Alternativ kann auch für diesen Fall analog zu Gleichung (22.79) eine Korrektur des verzerrten Fehler-in-den-Variablen-Schätzers wie folgt vorgenommen werden:

$$\hat{\boldsymbol{\beta}}^{a2^{Corr}} = \left(\mathbf{I}_K - (\mathbf{M}_{xx} + \boldsymbol{\Sigma}_{ww})^{-1}\boldsymbol{\Sigma}_{ww}\right)^{-1} \hat{\boldsymbol{\beta}}^{a2} - (\mathbf{M}_{xx})^{-1}\boldsymbol{\Sigma}_{wv} \tag{22.82}$$

und somit in operationalisierbarer Form:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{a2^{CorrOp}} = {} & \left(\mathbf{I}_K - \left(\frac{1}{n}\mathbf{X}^{a'}\mathbf{X}^{a} + \hat{\boldsymbol{\Sigma}}_{ww}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{ww}\right)^{-1}\hat{\boldsymbol{\beta}}^{a2} \\ & - \left(\frac{1}{n}\mathbf{X}^{a'}\mathbf{X}^{a} - \hat{\boldsymbol{\Sigma}}_{ww}\right)^{-1}\hat{\boldsymbol{\Sigma}}_{wv}.\end{aligned} \tag{22.83}$$

Fuller (1987) schlägt für diesen Fall folgenden Schätzer vor:

$$\hat{\boldsymbol{\beta}}^{Fu} = \left(\mathbf{M}_{x^a x^a} - \boldsymbol{\Sigma}_{ww}\right)^{-1} \left(\mathbf{M}_{x^a y^a} - \boldsymbol{\Sigma}_{wv}\right). \tag{22.84}$$

Eine operationalisierbare Form des Schätzers lautet somit:

$$\hat{\boldsymbol{\beta}}^{Fu^{Op}} = \left(\frac{1}{n}\mathbf{X}^{a\prime}\mathbf{X}^a - \hat{\boldsymbol{\Sigma}}_{ww}\right)^{-1} \left(\frac{1}{n}\mathbf{X}^{a\prime}\mathbf{y}^a - \hat{\boldsymbol{\Sigma}}_{wv}\right). \tag{22.85}$$

Da $\boldsymbol{\Sigma}_{ww}$ nicht vollen Rang hätte, wenn die Konstante des linearen Modells in Gleichung (22.82) enthalten wäre, wird der Schätzwert für β_0 bestimmt durch (Gottschalk 2004):

$$\hat{\beta}_0^{Fu} = \overline{y}^a - \overline{\mathbf{X}}^{a\prime}\hat{\boldsymbol{\beta}}^{Fu}. \tag{22.86}$$

Dabei ist $\overline{\mathbf{X}}^{a\prime}$ der $(1 \times K)$-Zeilenvektor der arithmetischen Mittel der überlagerten Regressoren.

Ein konsistenter Schätzer für die Varianz von $\hat{\boldsymbol{\beta}}^{Fu}$ lautet (Fuller 1987):

$$\begin{aligned}\widehat{var}\left(\hat{\boldsymbol{\beta}}^{Fu}\right) \;=\; & \frac{1}{n}\Big[\left(\mathbf{M}_{x^a x^a} - \boldsymbol{\Sigma}_{ww}\right)^{-1}\sigma_a^2 + \left(\mathbf{M}_{x^a x^a} - \boldsymbol{\Sigma}_{ww}\right)^{-1} \\ & \left(\boldsymbol{\Sigma}_{ww}\sigma_a^2 + \left(\boldsymbol{\Sigma}_{wv} - \boldsymbol{\Sigma}_{ww}\hat{\boldsymbol{\beta}}^{Fu}\right)\left(\boldsymbol{\Sigma}_{wv} - \boldsymbol{\Sigma}_{ww}\hat{\boldsymbol{\beta}}^{Fu}\right)'\right) \\ & \left(\mathbf{M}_{x^a x^a} - \boldsymbol{\Sigma}_{ww}\right)^{-1}\Big]\end{aligned} \tag{22.87}$$

mit

$$\sigma_a^2 = (n-K)^{-1}\sum_{i=1}^{n}(y_i^a - \mathbf{X}^{a_i\prime}\hat{\boldsymbol{\beta}}^{Fu})^2. \tag{22.88}$$

Dabei ist $\mathbf{X}_i^{a\prime}$ der i-te $(1 \times K)$-Zeilenvektor der additiv überlagerten Regressormatrix.

b) Multiplikative Stochastische Überlagerungen in linearen Modellen

Im Falle einer multiplikativen stochastischen Überlagerung gilt für eine anonymisierte Datenmatrix (vgl. Gleichung (6.56)):

$$\mathbf{X}^a = \mathbf{X} \odot \mathbf{W}. \tag{22.89}$$

Dabei ist $\mathbf{W}$ eine Matrix aus identisch verteilten Störgrößen. Für die nachfolgenden Ergebnisse ist wesentlich, dass diese Matrix zwar spaltenweise korreliert sein kann, zeilenweise jedoch unkorreliert sein muss.

$\mathbf{X} \odot \mathbf{W}$ ist das so genannte Hadamard-Produkt der Matrizen $\mathbf{X}$ und $\mathbf{W}$ und stellt die elementweise Multiplikation beider Matrizen dar (vgl. z.B. Lütkepohl (1997)).

Multiplikative Überlagerung der abhängigen Variablen

Wird lediglich die abhängige Variable Y multiplikativ überlagert, so ergibt sich für den anonymisierten Merkmalsvektor:

$$\mathbf{y}^a = \mathbf{y} \odot \mathbf{v}. \tag{22.90}$$

Für das geschätzte Modell ergibt sich:

$$\mathbf{y}^a = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}^a \tag{22.91}$$

oder

$$\mathbf{y} \odot \mathbf{v} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}^a. \tag{22.92}$$

Damit ergibt sich für den OLS-Schätzer:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}^a \tag{22.93}$$

und mit Gleichung (22.90)

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{y} \odot \mathbf{v}). \tag{22.94}$$

Für den Wahrscheinlichkeitsgrenzwert des Schätzers ergibt sich somit:

$$plim\,\hat{\boldsymbol{\beta}} = plim\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} plim\left(\frac{\mathbf{X}'(\mathbf{y} \odot \mathbf{v})}{n}\right). \tag{22.95}$$

Dies kann weiter umgeformt werden zu

$$plim\,\hat{\boldsymbol{\beta}} = plim\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\; plim\left(\frac{\mathbf{X}'\left((\mathbf{X}\boldsymbol{\beta}+\mathbf{u})\odot\mathbf{v}\right)}{n}\right) \tag{22.96}$$

und damit zu

$$plim\,\hat{\boldsymbol{\beta}} = \mathbf{Q}^{-1}\;\left[plim\left(\frac{\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}\odot\mathbf{v})}{n}\right) + plim\left(\frac{\mathbf{X}'(\mathbf{u}\odot\mathbf{v})}{n}\right)\right]. \tag{22.97}$$

In Rosemann (2005) wird in Analogie zu Lin (1989) gezeigt, dass $plim\left(\frac{\mathrm{X}'(\mathrm{X}\boldsymbol{\beta}\odot\mathrm{v})}{n}\right) = \mathbf{Q}\boldsymbol{\beta}$ und $plim\left(\frac{\mathrm{X}'(\mathrm{u}\odot\mathrm{v})}{n}\right) = \mathbf{0}$ gilt.

Daraus folgt für den Wahrscheinlichkeitsgrenzwert des Schätzers:

$$plim\,\hat{\boldsymbol{\beta}} = \mathbf{Q}^{-1}\,\mathbf{Q}\boldsymbol{\beta}\; = \boldsymbol{\beta}. \tag{22.98}$$

Damit ist der Schätzer auch für den Fall einer multiplikativen Überlagerung der abhängigen Variablen konsistent.

Multiplikative Überlagerung der Regressoren

Anders verhält sich dies für den Fall, dass nur die Regressoren multiplikativ überlagert werden. In diesem Fall gilt für das zu schätzende Modell:

$$\mathbf{y} = \mathbf{X}^a\boldsymbol{\beta}^a + \mathbf{u}^a. \tag{22.99}$$

Daraus folgt für den Schätzer:

$$\hat{\boldsymbol{\beta}}^a = \left(\mathbf{X}^{a'}\mathbf{X}^a\right)^{-1}\mathbf{X}^{a'}\mathbf{y}. \tag{22.100}$$

Für den Wahrscheinlichkeitsgrenzwert des Schätzers ergibt sich daraus:

$$plim\,\hat{\boldsymbol{\beta}}^a = plim\left(\frac{\mathbf{X}^{a'}\mathbf{X}^a}{n}\right)^{-1}\; plim\left(\frac{\mathbf{X}^{a'}\mathbf{y}}{n}\right). \tag{22.101}$$

Für den zweiten Term auf der rechten Seite der Gleichung (22.101) ergibt sich:

$$plim\left(\frac{\mathbf{X}^{a'}\mathbf{y}}{n}\right) = plim\left(\frac{\mathbf{X}^{a'}\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right)}{n}\right) \tag{22.102}$$

und somit

$$plim\left(\frac{\mathbf{X}^{a'}\mathbf{y}}{n}\right) = plim\left(\frac{\mathbf{X}^{a'}\mathbf{X}\boldsymbol{\beta}}{n}\right) + plim\left(\frac{\mathbf{X}^{a'}\mathbf{u}}{n}\right) \tag{22.103}$$

oder

$$plim\left(\frac{\mathbf{X}^{a'}\mathbf{y}}{n}\right) = plim\left(\frac{(\mathbf{X}\odot\mathbf{W})'\mathbf{X}\boldsymbol{\beta}}{n}\right) + plim\left(\frac{(\mathbf{X}\odot\mathbf{W})'\mathbf{u}}{n}\right). \tag{22.104}$$

Für Gleichung (22.104) gilt aber (Hwang 1986):

$$plim\left(\frac{\mathbf{X}^{a'}\mathbf{y}}{n}\right) = \mathbf{Q}\boldsymbol{\beta}. \tag{22.105}$$

(Dies wird in Rosemann (2005) analog zu Lin (1989) hergeleitet.)

Für den ersten Term auf der rechten Seite der Gleichung (22.101) ergibt sich hingegen:

$$plim\left(\frac{\mathbf{X}^{a'}\mathbf{X}^{a}}{n}\right) = plim\left(\frac{(\mathbf{X}\odot\mathbf{W})'(\mathbf{X}\odot\mathbf{W})}{n}\right) = \mathbf{M}\odot\mathbf{Q}. \tag{22.106}$$

Dabei ist $\mathbf{M} = E\left(\mathbf{w}_i\mathbf{w}_i'\right)$ mit $\mathbf{w}_i'$ der i-ten Zeile der Matrix $\mathbf{W}$ (Hwang 1986; Lin 1989).

Damit gilt für den Wahrscheinlichkeitsgrenzwert des Schätzers:

$$plim\,\hat{\boldsymbol{\beta}}^{a} = (\mathbf{M}\odot\mathbf{Q})^{-1}\mathbf{Q}\boldsymbol{\beta}. \tag{22.107}$$

Der KQ-Schätzer ist somit bei multiplikativer Überlagerung der Regressoren nicht konsistent. Hwang (1986) zeigt, dass gilt:

$$\left|(\mathbf{M}\odot\mathbf{Q})^{-1}\mathbf{Q}\boldsymbol{\beta}\right| \leq |\boldsymbol{\beta}|. \tag{22.108}$$

Damit gilt auch im Fall der multiplikativen Überlagerung, dass die absolute Höhe der Koeffizienten unterschätzt wird.

Hwang (1986) und Lin (1989) gehen davon aus, dass die multiplikativen Fehler unkorreliert sind. Ihre Ergebnisse treffen jedoch auch zu, wenn die Überlagerungsfaktoren für unterschiedliche Variablen korreliert sind, allerdings müssen die Faktoren für unterschiedliche Unternehmen unkorreliert sein. Dies bedeutet, dass die Matrix **W** spaltenweise korreliert sein kann, jedoch zeilenweise unkorreliert ist.

Multiplikative Überlagerung der Regressoren und der abhängigen Variablen

Für den Fall, dass sowohl die Regressoren als auch die abhängige Variable multiplikativ überlagert werden, ergibt sich für das zu schätzende Modell:

$$\mathbf{y}^a = \mathbf{X}^a \boldsymbol{\beta}^{a2} + \mathbf{u}^a. \tag{22.109}$$

Damit gilt für den Schätzer:

$$\hat{\boldsymbol{\beta}}^{a2} = \left(\mathbf{X}^{a'}\mathbf{X}^a\right)^{-1}\mathbf{X}^{a'}\mathbf{y}^a \tag{22.110}$$

oder

$$\hat{\boldsymbol{\beta}}^{a2} = \left((\mathbf{X} \odot \mathbf{W})'(\mathbf{X} \odot \mathbf{W})\right)^{-1}(\mathbf{X} \odot \mathbf{W})'(\mathbf{y} \odot \mathbf{v}) \tag{22.111}$$

und damit

$$\hat{\boldsymbol{\beta}}^{a2} = \left(\frac{(\mathbf{X} \odot \mathbf{W})'(\mathbf{X} \odot \mathbf{W})}{n}\right)^{-1}\frac{(\mathbf{X} \odot \mathbf{W})'(\mathbf{y} \odot \mathbf{v})}{n}. \tag{22.112}$$

Geht man wiederum zum Wahrscheinlichkeitsgrenzwert über, so erhält man:

$$plim\,\hat{\boldsymbol{\beta}}^{a2} = plim\left(\frac{(\mathbf{X} \odot \mathbf{W})'(\mathbf{X} \odot \mathbf{W})}{n}\right)^{-1} plim\left(\frac{(\mathbf{X} \odot \mathbf{W})'(\mathbf{y} \odot \mathbf{v})}{n}\right). \tag{22.113}$$

Für den ersten Term auf der rechten Seite dieser Gleichung gilt wiederum:

$$plim\left(\frac{(\mathbf{X} \odot \mathbf{W})'(\mathbf{X} \odot \mathbf{W})}{n}\right)^{-1} = (\mathbf{M} \odot \mathbf{Q})^{-1}. \tag{22.114}$$

Das Ergebnis für den zweiten Term auf der rechten Seite der Gleichung hängt nun davon ab, ob die Überlagerungen der Regressoren und der abhängigen Variablen korreliert oder unkorreliert sind.

Sind die Überlagerungen der Regressoren und der abhängigen Variablen unkorreliert, so ergibt sich für den zweiten Term auf der rechten Seiten der Gleichung (22.113) (Lin 1989):

$$plim\left(\frac{(\mathbf{X}\odot\mathbf{W})'(\mathbf{y}\odot\mathbf{v})}{n}\right)=\frac{\mathbf{X}'\mathbf{X}}{n}\boldsymbol{\beta}=\mathbf{Q}\boldsymbol{\beta}. \tag{22.115}$$

Damit folgt für den Wahrscheinlichkeitsgrenzwert des Schätzers in diesem Fall:

$$plim\hat{\boldsymbol{\beta}}^{a2}=(\mathbf{M}\odot\mathbf{Q})^{-1}\mathbf{Q}\boldsymbol{\beta}. \tag{22.116}$$

Sind die Überlagerungen der Regressoren und der abhängigen Variablen hingegen korreliert, so ergibt sich für den zweiten Term auf der rechten Seiten der Gleichung (22.113):

$$plim\left(\frac{(\mathbf{X}\odot\mathbf{W})'(\mathbf{y}\odot\mathbf{v})}{n}\right)=\mathbf{Q}\boldsymbol{\beta}\odot\mathbf{K} \tag{22.117}$$

mit $\mathbf{K}=E(\mathbf{w}_i v_i)$.

Damit gilt für den Wahrscheinlichkeitsgrenzwert des Schätzers:

$$plim\hat{\boldsymbol{\beta}}^{a2}=(\mathbf{M}\odot\mathbf{Q})^{-1}(\mathbf{Q}\boldsymbol{\beta}\odot\mathbf{K}). \tag{22.118}$$

Wesentlich für das Ergebnis ist die Annahme, dass die Überlagerungen der abhängigen Variablen und der Regressoren zwar korreliert sein können, die Korrelationen zwischen den Überlagerungen unterschiedlicher Unternehmen jedoch Null sind (zeilenweise unkorreliert). (Zu den Herleitungen vgl. Rosemann (2005).)

Somit entspricht das Ergebnis für den Fall, dass die zur Überlagerung der Regressoren und der abhängigen Variablen verwendeteten Überlagerungen unkorreliert sind ($\mathbf{K}=\mathbf{1}_K$), dem Ergebnis für die ausschließliche Überlagerung der Regressoren aus Gleichung (22.107) (vgl. hierzu auch Lin (1989)). Der Schätzer ist somit auch bei der multiplikativen Überlagerung von Regressoren und abhängiger Variable inkonsistent.

Allerdings ist er für den Fall, dass alle Merkmale einer Einheit mit einem konstanten Faktor überlagert werden, konsistent und auch im Kleinstichprobenfall erwartungstreu, falls eine der beiden folgenden Bedingungen zutrifft:

- Es wird ein Modell ohne Absolutglied geschätzt (beziehungsweise das Absolutglied ist Null).

- Es gilt für alle Regressoren: $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} w_i x_i = 0$.

Im ersten Fall (Fall ohne Absolutglied) gilt für das zu schätzende Modell:

$$\mathbf{y} \odot \mathbf{w} = (\mathbf{X} \odot \mathbf{W})\boldsymbol{\beta} + \mathbf{u} \odot \mathbf{w} \tag{22.119}$$

und somit

$$\mathbf{y} \odot \mathbf{w} = (\mathbf{X} \odot (\mathbf{w} \otimes \boldsymbol{\iota}'))\boldsymbol{\beta} + \mathbf{u} \odot \mathbf{w} \tag{22.120}$$

oder

$$\begin{pmatrix} y_1 w_1 \\ \vdots \\ y_n w_n \end{pmatrix} = \begin{pmatrix} x_{11} w_1 & \dots & x_{1K} w_1 \\ \vdots & \dots & \vdots \\ x_{n1} w_n & \dots & x_{nK} w_n \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_K \end{pmatrix} + \begin{pmatrix} u_1 w_1 \\ \vdots \\ u_K w_K \end{pmatrix}. \tag{22.121}$$

Bei einem konstanten Überlagerungsfaktor für alle Merkmale kann das Verhalten der Schätzer allgemeingültig anhand des einfachen linearen Regressionsmodells näher betrachtet werden. Für die Schätzer gilt in diesem Fall:

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}\frac{1}{n}\sum_{i=1}^{n} x_i \tag{22.122}$$

und

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - \hat{\alpha}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i^2}. \tag{22.123}$$

oder

$$\hat{\beta} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}. \tag{22.124}$$

Werden nun y_i und x_i jeweils mit dem gleichen Faktor w_i überlagert, so ergibt sich für den Schätzer des Absolutglieds α:

$$\hat{\alpha}^a = \frac{1}{n}\sum_{i=1}^{n} w_i y_i - \hat{\beta}\frac{1}{n}\sum_{i=1}^{n} w_i x_i. \tag{22.125}$$

Da $E(w_i) = 1$ gilt, ist der Schätzer für α immer dann erwartungstreu beziehungsweise konsistent, wenn auch $\hat{\beta}$ ein erwartungstreuer beziehungsweise konsistenter Schätzer für β ist.

Für $\hat{\beta}$ gilt aber im Fall einer konstanten Überlagerung beider Variablen:

$$\begin{aligned}
\hat{\beta}^a &= \frac{n\sum_{i=1}^{n} w_i^2 x_i y_i - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i y_i}{n\sum_{i=1}^{n} w_i^2 x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2} \\
&= \frac{n\sum_{i=1}^{n} w_i^2 x_i(x_i\beta + \alpha + u_i) - \sum_{i=1}^{n} w_i x_i \sum_{i=1}^{n} w_i(x_i\beta + \alpha + u_i)}{n\sum_{i=1}^{n} w_i^2 x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2} \\
&= \frac{n\sum_{i=1}^{n} w_i^2 x_i^2\beta + n\alpha\sum_{i=1}^{n} w_i^2 x_i + n\sum_{i=1}^{n} w_i^2 x_i u_i - (\sum_{i=1}^{n} w_i x_i)^2\beta - \alpha\sum_{i=1}^{n} w_i x_i - \sum_{i=1}^{n} w_i x_i u_i}{n\sum_{i=1}^{n} w_i^2 x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2} \\
&= \beta + \alpha\frac{n\sum_{i=1}^{n} w_i^2 x_i - \sum_{i=1}^{n} w_i x_i}{n\sum_{i=1}^{n} w_i^2 x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2} + \frac{n\sum_{i=1}^{n} w_i^2 x_i u_i - \sum_{i=1}^{n} w_i x_i u_i}{n\sum_{i=1}^{n} w_i^2 x_i^2 - (\sum_{i=1}^{n} w_i x_i)^2}.
\end{aligned} \tag{22.126}$$

Für den Wahrscheinlichkeitsgrenzwert von $\hat{\beta}^a$ gilt somit:

$$
\begin{aligned}
plim(\hat{\beta}^a) &= \beta + \alpha \frac{plim\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2 x_i - \frac{1}{n^2}\sum_{i=1}^{n} w_i x_i\right)}{plim\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2 x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} w_i x_i)^2\right)} + \\
&+ \frac{plim\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2 x_i u_i\right)}{plim\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2 x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} w_i x_i)^2\right)} - \\
&- \frac{plim\left(\frac{1}{n^2}\sum_{i=1}^{n} w_i x_i u_i\right)}{plim\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2 x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} w_i x_i)^2\right)}
\end{aligned}
\tag{22.127}
$$

und somit

$$
plim(\hat{\beta}^a) = \beta + \alpha \frac{\left(\frac{1}{n}E(w_i^2) - \frac{1}{n^2}E(w_i)\right)\sum_{i=1}^{n} x_i}{plim\left(\frac{1}{n}\sum_{i=1}^{n} w_i^2 x_i^2 - (\frac{1}{n}\sum_{i=1}^{n} w_i x_i)^2\right)}. \tag{22.128}
$$

Somit ist $\hat{\beta}$ in diesem Fall nur dann konsistent, wenn entweder $\alpha = 0$ oder $\sum_{i=1}^{n} x_i = 0$ gilt. Diese Bedingungen sind relativ restriktiv.

Korrekturverfahren bei multiplikativer Überlagerung

Im Falle eines inkonsistenten Schätzers kann eine Korrektur analog zum Fall der additiven stochastischen Überlagerung mit Hilfe des IV-Schätzers erfolgen. Dabei werden die Instrumente wiederum so gewählt, dass sie mit den Störgrößen und den Überlagerungen asymptotisch unkorreliert und mit den Regressoren hoch korreliert sind. Für die Prämultiplikation der Gleichung (22.99) gilt somit:

$$
\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X}^a\boldsymbol{\beta} + \mathbf{Z}'\mathbf{u}^a \tag{22.129}
$$

oder

$$
\mathbf{Z}'\mathbf{y} = \mathbf{Z}'\mathbf{X} \odot \mathbf{W}\boldsymbol{\beta} + \mathbf{Z}'\mathbf{u}^a. \tag{22.130}
$$

Geht man zu den Wahrscheinlichkeitsgrenzwerten über, so erhält man:

$$plim\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = plim\left(\frac{\mathbf{Z}'\mathbf{X}\odot\mathbf{W}}{n}\right)\boldsymbol{\beta} + plim\left(\frac{\mathbf{Z}'\mathbf{u}^a}{n}\right). \tag{22.131}$$

Unter den obigen Annahmen ergibt sich:

$$plim\left(\frac{\mathbf{Z}'\mathbf{y}}{n}\right) = plim\left(\frac{\mathbf{Z}'\mathbf{X}}{n}\right)\boldsymbol{\beta}. \tag{22.132}$$

Der IV-Schätzer ergibt sich somit wiederum als

$$\hat{\boldsymbol{\beta}}^{IV} \quad = \quad (\mathbf{Z}'\,\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}. \tag{22.133}$$

Hwang (1986) schlägt alternativ für den Fall, dass ausschließlich die Regressoren multiplikativ überlagert werden, einen ebenfalls konsistenten Korrekturschätzer vor:

$$\hat{\boldsymbol{\beta}}^{Hwa} \quad = \quad \left[\left(\mathbf{X}^{a'}\mathbf{X}^{a}\right) \div \mathbf{M}\right]^{-1}\mathbf{X}^{a'}\mathbf{y}, \tag{22.134}$$

dessen operationalisierbare Form sich wie folgt darstellen lässt:

$$\hat{\boldsymbol{\beta}}^{Hwa^{Op}} \quad = \quad \left[\left(\mathbf{X}^{a'}\mathbf{X}^{a}\right) \div \hat{\mathbf{M}}\right]^{-1}\mathbf{X}^{a'}\mathbf{y} \tag{22.135}$$

mit $\hat{\mathbf{M}} = \frac{\mathrm{W}'\mathrm{W}}{n}$.

Dabei bezeichnet $\div$ die Hadamard-Division, also die elementweise Division.

Notwendige Bedingung für die Verwendung dieses Schätzers ist jedoch, dass die Überlagerungen für unterschiedliche Unternehmen, also zeilenweise, unkorreliert sind. Der Schätzer kann auch verwendet werden, wenn auch die abhängige Variable multiplikativ überlagert wird, allerdings nur, sofern die Überlagerungsfaktoren der abhängigen Variablen mit den Überlagerungsfaktoren der Regressoren unkorreliert sind (Lin 1989).

Wird lediglich die Regressormatrix $\mathbf{X}$ multiplikativ überlagert, so lautet ein konsistenter Schätzer für die Varianz-Kovarianzmatrix des Schätzers (Hwang 1986):

$$\begin{aligned}\widehat{var}\left(\hat{\boldsymbol{\beta}}^{Hwa}\right) &= \frac{1}{n}\left(\frac{\mathbf{X}^{a\prime}\mathbf{X}^{a}}{n} \div \mathbf{M}\right)^{-1} \\ &\times \left\{\sum_{i=1}^{n}\left[\mathbf{x}_i^a y_i - \left(\mathbf{x}_i^a \mathbf{x}_i^{a\prime} \odot \mathbf{M}\right)\hat{\boldsymbol{\beta}}^{Hwa}\right]\left[\mathbf{x}_i^a y_i - \left(\mathbf{x}_i^a \mathbf{x}_i^{a\prime} \odot \mathbf{M}\right)\hat{\boldsymbol{\beta}}^{Hwa}\right]\right\} \\ &\times \left(\frac{\mathbf{X}^{a\prime}\mathbf{X}^{a}}{n} \div \mathbf{M}\right)^{-1}. \end{aligned} \tag{22.136}$$

Werden hingegen sowohl die Regressoren als auch die abhängige Variable multiplikativ überlagert, so ergibt sich ein konsistenter Schätzer für die Varianz-Kovarianzmatrix des Schätzers (Lin 1989):[24]

$$\begin{aligned}\widehat{var}\left(\hat{\boldsymbol{\beta}}^{Hwa}\right) &= \left(\mathbf{X}^{a\prime}\mathbf{X}^{a} \div \mathbf{M}\right)^{-1}\left[\hat{\sigma}_u^2\left(1+\sigma_v^2\right)\left(\mathbf{X}^{a\prime}\mathbf{X}^{a}\right) + \left(\mathbf{I}_K \otimes \hat{\boldsymbol{\beta}}^{Hwa\prime}\right)\right. \\ &\times \left[E\left[vec\left(\mathbf{w}_i v_i \mathbf{1}_K' - \mathbf{w}_i \mathbf{w}_i' \odot \mathbf{M}\right)\right]\left[vec\left(\mathbf{w}_i v_i \mathbf{1}_K' - \mathbf{w}_i \mathbf{w}_i' \odot \mathbf{M}\right)\right]'\right. \\ &\odot \sum\sum_{i=1}^{n}\left[vec\left(\mathbf{x}_i^a \mathbf{x}_i^{a\prime}\right)\right]\left[vec\left(\mathbf{x}_i^a \mathbf{x}_i^{a\prime}\right)\right]' \odot E\left[vec\left(\mathbf{w}_i \mathbf{w}_i'\right)\right]\left[vec\left(\mathbf{w}_i \mathbf{w}_i'\right)\right]'] \\ &\left.\times \left(\mathbf{I}_K \otimes \hat{\boldsymbol{\beta}}^{Hwa}\right)\right]\left(\mathbf{X}^{a\prime}\mathbf{X}^{a} \div \mathbf{M}\right)^{-1} \end{aligned} \tag{22.137}$$

mit

$$\hat{\sigma}_u^2 = \left[\mathbf{y}^{a\prime}\mathbf{y}^{a} \div \left(\sigma_v^2 + 1\right) - \hat{\boldsymbol{\beta}}^{Hwa}\mathbf{X}^{a\prime}\mathbf{y}^{a}\right] / (n-K). \tag{22.138}$$

Werden abhängige Variable und Regressoren überlagert und sind die Überlagerungsfaktoren nicht spaltenweise (wohl aber zeilenweise) unkorreliert, so ergibt sich ein Korrekturschätzer durch:

$$\hat{\boldsymbol{\beta}}^{a2Korr} = \left(\mathbf{X}^{a\prime}\mathbf{X}^{a} \div \mathbf{M}\right)\left(\mathbf{X}^{a\prime}\mathbf{y}^{a} \div \mathbf{K}\right). \tag{22.139}$$

Eine operationalisierbare Form des Schätzers lautet:

$$\hat{\boldsymbol{\beta}}^{a2Korr^{Op}} = \left(\mathbf{X}^{a\prime}\mathbf{X}^{a} \div \hat{\mathbf{M}}\right)\left(\mathbf{X}^{a\prime}\mathbf{y}^{a} \div \hat{\mathbf{K}}\right) \tag{22.140}$$

24) $vec(A)$ ordnet die s Spalten der Matrix A untereinander an, so dass A mit der Dimension $t \times s$ in einen $(ts \times 1)$-Vektor transformiert wird, vgl. zum Beispiel Gottschalk (2005).

mit $\hat{\mathbf{K}} = \frac{\mathbf{W}'\mathbf{v}}{n}$.

Eine Alternative zu den dargestellten Korrekturverfahren kann in einigen Fällen darin bestehen, durch Logarithmierung multiplikative in additive Fehler umzuwandeln und damit Korrekturverfahren für additive Fehler verwenden zu können. Allerdings ist dieses Vorgehen nur möglich, wenn für den Erwartungswert des logarithmierten Überlagerungsfaktors $E(log(W)) = 0$ gilt (Gottschalk 2005).

Kennt der Nutzer die Verteilung des Überlagerungsfaktors, so kann er den Erwartungswert des Fehlers bestimmen und diesen von der logarithmierten überlagerten Variablen subtrahieren. Der Erwartungswert des so entstehenden additiven Fehlers ist definitionsgemäß Null. Somit kann beispielsweise der Korrekturschätzer von Fuller (1987) verwendet werden. Allerdings setzt dieses Vorgehen voraus, dass die logarithmierte Form der Variablen modellkonform ist (Gottschalk 2005). Bei der von Kim und Winkler (2001) vorgeschlagenen Form der multiplikativen stochastischen Überlagerung, bei der die Originalvariablen zunächst logarithmiert und dann multiplikativ überlagert werden (vgl. Unterabschnitt 6.2.3), führt eine Logarithmierung der entsprechend behandelten Variablen in jedem Fall zu einem additiven Fehlerterm mit Erwartungswert Null.

22.1.2 Stochastische Überlagerungen in nichtlinearen Modellen

Im vorangegangenen Unterabschnitt 22.1.1 wurde ausführlich hergeleitet, dass stochastische Überlagerungen der Regressoren in der Regel zu inkonsistenten Schätzern in linearen Modellen führen. Dies gilt ebenso für die Schätzung von nichtlinearen Modellen. Einen Überblick über die Messfehlerproblematik in nichtlinearen Modellen geben Carroll et al. (1995).

Allerdings ist die Beschaffenheit der Verzerrung und damit auch deren Korrektur in nichtlinearen Modellen weitaus komplexer als in linearen Modellen. Dies liegt inbesondere daran, dass sich unter dem Oberbegriff der nichtlinearen Modelle viele unterschiedliche Zusammenhänge subsumieren lassen. Hierzu gehören beispielsweise:

- Generalisierte lineare Modelle,
- Loglineare Modelle,
- Quadratische Modelle,
- Logit-, Probit- und Tobitmodelle,
- Verweildaueranalysen.

Während die Korrektur von Messfehlern in linearen Modellen eine lange Tradition hat, sind die nichtlinearen Fehler-in-den-Variablen-Modelle eher neueren Datums (Carroll et al.

1996). Aufgrund der Verschiedenheit der nichtlinearen Modelle wurden auch spezielle Korrekturmodelle entwickelt, beispielsweise von Stefanski und Carroll (1985) für die logistische Regression.

Daneben existieren jedoch auch generelle Fehler-Korrektur-Schätzer für nichtlineare Modelle (Carroll et al. 1995), von denen insbesondere der Kalibrationsschätzer und der SIMEX-Schätzer (vgl. hierzu auch Lechner und Pohlmeier (2004)) zu nennen sind, die auch im Programmpaket STATA implementiert sind. SIMEX-Schätzer und Kalibration sind für alle Schätzmethoden anwendbar (Carroll et al. 1995). Somit ist – ebenso wie für stochastische Überlagerungen in linearen Modellen – auch in nichtlinearen Modellen eine Korrektur möglich. Im Folgenden wird die SIMEX-Methode ausführlicher vorgestellt, da dieser ein sehr anschauliches Vorgehen zugrundeliegt.

Die SIMEX-Methode wurde von Cook und Stefanski (1994) zur Korrektur von Messfehlern in nichtlinearen Modellen vorgeschlagen. SIMEX steht für „Simulation Extrapolation". Die Methode kann angewendet werden, sofern die Varianz der Überlagerungen bekannt ist oder diese gut geschätzt werden kann, bei anonymisierten Daten beispielsweise, indem ein zweiter in gleicher Weise (gleiche Verteilungsfamilie, gleicher Mittelwert, gleiche Varianz der Überlagerungen) stochastisch überlagerter Datensatz zur Verfügung gestellt wird (Cook und Stefanski 1994).[25]

Die Grundidee des SIMEX-Schätzers wurde für den Fall eines additiven Messfehlers beziehungsweise für die additive stochastische Überlagerung entwickelt. Dabei wird zunächst davon ausgegangen, dass die Variable X additiv mit einem normalverteilten Fehler V überlagert wird. Dabei ist V unabhängig von X. Es gilt somit für die anonymisierte Variable:

$$X^a = X + V \tag{22.141}$$

mit $V \sim N(0, \sigma_v^2)$.

Die Idee des SIMEX-Schätzers lässt sich zunächst am besten anhand des einfachen linearen Regressionsmodells $Y = \beta_1 + \beta_x X + U$ beschreiben. Der Wahrscheinlichkeitsgrenzwert des OLS-Schätzers ohne Korrektur der Überlagerung lautet in diesem Fall (Carroll et al. 1995):

$$plim\,\hat{\beta}_x = \frac{\beta_x \sigma_x^2}{\sigma_x^2 + \sigma_v^2}. \tag{22.142}$$

25) Für den Schätzer der Varianz-Kovarianzmatrix der Überlagerungen gilt bei k_i Replikationen:

$$\hat{\mathbf{\Sigma}}_{vv} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{k_i}\left(\mathrm{x}_{ij}^a - \overline{\mathrm{x}}_i^a\right)\left(\mathrm{x}_{ij}^a - \overline{\mathrm{x}}_i^a\right)'}{\sum_{i=1}^{n}(k_i - 1)}$$

(Carroll et al. 1995).

Dem SIMEX-Schätzer liegt nun die Idee zugrunde, die Effekte der stochastischen Überlagerung experimentell mit Hilfe von Simulationen zu bestimmen. Ist die Varianz der additiven Überlagerung bekannt, so kann die anonymisierte Variable mit einem zusätzlichen Fehler überlagert werden, dessen Varianz das λ-fache der ursprünglichen Fehlervarianz beträgt:

$$X^a_{b,i}(\lambda) = X^a_i + \lambda^{1/2} V_{b,i} = X_i + V_i + \lambda^{1/2} V_{b,i} \tag{22.143}$$

mit $\lambda \geq 0$.

Damit gilt für die Varianz der zweifach überlagerten Variablen:

$$var(X^a_{b,i}(\lambda)) = \sigma^2_x + \sigma^2_v + \lambda\sigma^2_v = \sigma^2_x + (1+\lambda)\sigma^2_v \tag{22.144}$$

und für den Wahrscheinlichkeitsgrenzwert des Schätzers unter Verwendung dieser Variablen:

$$plim\,\hat{\beta}_x(\lambda) = \frac{\beta_x \sigma^2_x}{\sigma^2_x + (1+\lambda)\sigma^2_v}. \tag{22.145}$$

Variiert man nun den Parameter λ zwischen 0 und λ_M (in der Regel wird $\lambda_M = 2$ gewählt (Cook und Stefanski 1994)), so erhält man einen funktionalen Zusammenhang zwischen dem Grad der Überlagerung λ und dem sich ergebenden „naiven" Schätzer. Für das einfache lineare Regressionsmodell stellt Gleichung (22.145) diesen funktionalen Zusammenhang dar.

Der wahre Wert für den Schätzer würde sich nun ergeben, wenn λ den Wert -1 annehmen würde. Deshalb kann der wahre Schätzwert durch eine Extrapolation des funktionalen Zusammenhangs zwischen λ und $\hat{\beta}$ auf $\lambda = -1$ ermittelt werden. Diese Überlegung wird bei der Berechnung des SIMEX-Schätzers auf nichtlineare Zusammenhänge übertragen.

Das Vorgehen bei der Berechnung des SIMEX-Schätzers besteht aus zwei Schritten, dem Simulationsschritt und dem Extrapolationsschritt (Carroll et al. 1995). Zunächst wird für $\lambda \geq 0$ eine zusätzliche Überlagerung der anonymisierten Variablen vorgenommen durch:

$$x^a_{b,i}(\lambda) = x^a_i + \lambda^{1/2} v_{b,i} \tag{22.146}$$

mit $i = 1, \ldots, n$ der Anzahl der Beobachtungen, $b = 1, \ldots, B$ der Anzahl der Replikationen im Simulationsschritt und $v_{b,i}$ den Realisationen unabhängig und identisch normalverteilter Zufallsvariablen mit $E(V_{b,i}) = 0$ und $var(V_{b,i}) = \sigma^2_v$.[26]

26) Cook und Stefanski (1994) weisen darauf hin, dass obwohl die IID-Pseudofehler $V_{b,i}$ unter diesen

Anschließend wird für jedes λ der „naive“ Schätzer $\hat{\theta}_b(\lambda)$ bestimmt. Auf jeder Stufe für λ werden somit b „naive“ Schätzer ermittelt, um möglichst gute Simulationsergebnisse zu erhalten. Das arithmetische Mittel hieraus ergibt den Schätzer in Abhängigkeit von λ:

$$\hat{\theta}(\lambda) = \frac{1}{B} \sum_{b=1}^{B} \hat{\theta}_b(\lambda). \tag{22.147}$$

Der zweite und schwierigere Schritt besteht nun in der Extrapolation. Als besondere Herausforderung erweist sich dabei die Wahl einer geeigneten Extrapolationsfunktion. Die Extrapolationsfunktion muss so gewählt werden, dass sie die im ersten Schritt berechneten Schätzer $\hat{\theta}(\lambda)$ optimal verbindet und den wahren Schätzer an der Stelle $\lambda = -1$ möglichst genau trifft.

Cook und Stefanski (1994) halten drei Formen der Extrapolationsfunktion für denkbar:

- Eine lineare Extrapolationsfunktion

$$G_L(\lambda) = \gamma_1 + \gamma_2 \lambda,$$

- eine quadratische Extrapolationsfunktion

$$G_Q(\lambda) = \gamma_1 + \gamma_2 \lambda + \gamma_3 \lambda^2,$$

- eine nichtlineare Funktion, die als Quotient aus zwei linearen Funktionen gebildet wird und deshalb „rational lineare“ Extrapolationsfunktion genannt wird

$$G_{RL}(\lambda) = \gamma_1 + \frac{\gamma_2}{\gamma_3 + \lambda}.$$

Sofern die Überlagerung normalverteilt ist, ist jede dieser drei Extrapolationsfunktionen exakt für verschiedene Schätzer (Cook und Stefanski 1994). Die lineare Extrapolationsfunktion ist exakt für Schätzer, die eine lineare Funktion der ersten und zweiten Momente der Daten darstellen. Es kann gezeigt werden, dass die quadratische Extrapolationsfunktion exakt ist für Schätzer, die lineare Funktionen der ersten vier Momente sind. Die nichtlineare Extrapolationsfunktion ist exakt für verschiedene lineare und nichtlineare Regressionsschätzer, beispielsweise die einfache und die multiple lineare Regression sowie die multiple lineare

Annahmen erzeugt werden, die simulierten Realisationen dieser Pseudofehler trotzdem Korrelationen mit den beobachteten Daten, keinen exakten Mittelwert von 0 und keine exakte Varianz von 1 aufweisen. Sie schlagen deshalb vor, die Pseudofehler so zu erzeugen, dass sie mit den beobachteten Daten exakt unkorreliert sind sowie einen Stichprobenmittelwert von 0 und eine Stichprobenvarianz von 1 aufweisen. Man nennt diese erzeugten Fehler dann „NON-IID“-Pseudofehler.

Regression mit Interaktionsterm (Cook und Stefanski 1994). Für die einfache lineare Regression erkennt man dies in Gleichung (22.145). Die rational lineare Extrapolationsfunktion ist außerdem asymptotisch exakt für den Fall der generalisierten KQ-Schätzung eines Exponentialmodells (Cook und Stefanski 1994).

Die Auswahl der geeigneten Extrapolationsfunktion ist wohl das größte Problem bei der Anwendung des SIMEX-Schätzers. Sie ist deshalb sorgfältig zu treffen. Zu beachten ist, dass bei der rational linearen Funktion numerische Instabilitäten auftreten können. Auf der anderen Seite sind sowohl die lineare als auch die quadratische Extrapolationsfunktion hinsichtlich der Korrektur der Schätzung konservativer (Carroll et al. 1995). Carroll et al. (1995) empfehlen, den Extrapolationsschritt wie jedes andere Modellierungsproblem zu behandeln und die Extrapolationsfunktion auf Basis theoretischer Überlegungen zu wählen.

Für den einfachsten Fall des einfachen linearen Regressionsmodells (mit $\beta_1 = 1$, $\beta_x = 0,5$, einem standardnormalverteilten Störterm U, einer ebenfalls standardnormalverteilten additiven Überlagerung V und einem Stichprobenumfang von 1.000) ist das Vorgehen für alle drei möglichen Extrapolationsfunktionen in den Abbildungen 22.1, 22.2 und 22.3 grafisch veranschaulicht.

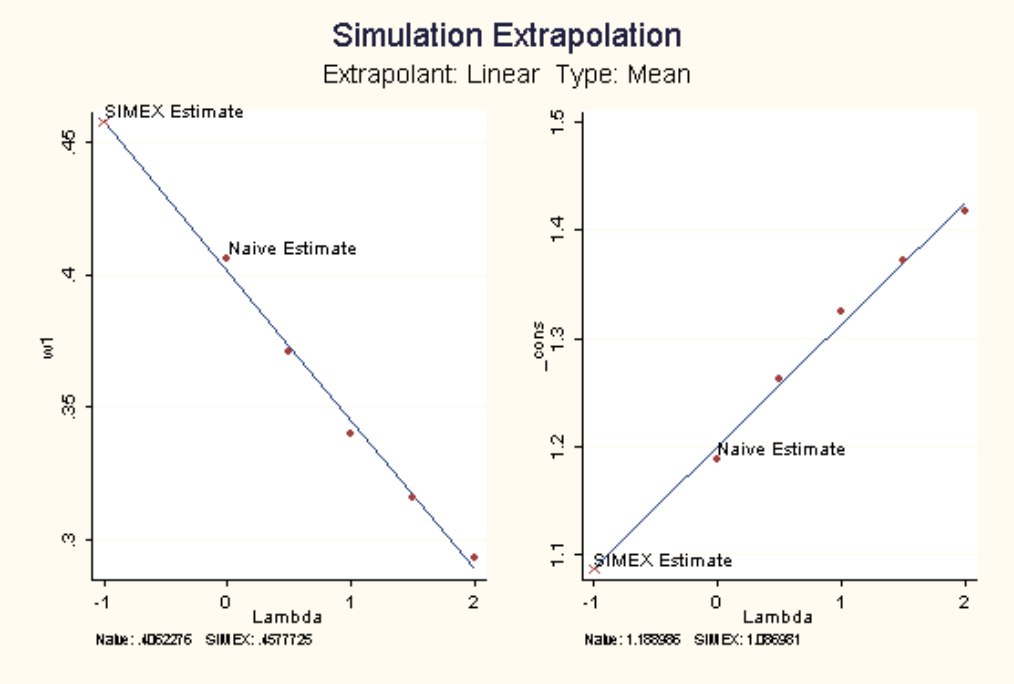

Abbildung 22.1: SIMEX-Schätzer im einfachen linearen Modell – Lineare Extrapolationsfunktion

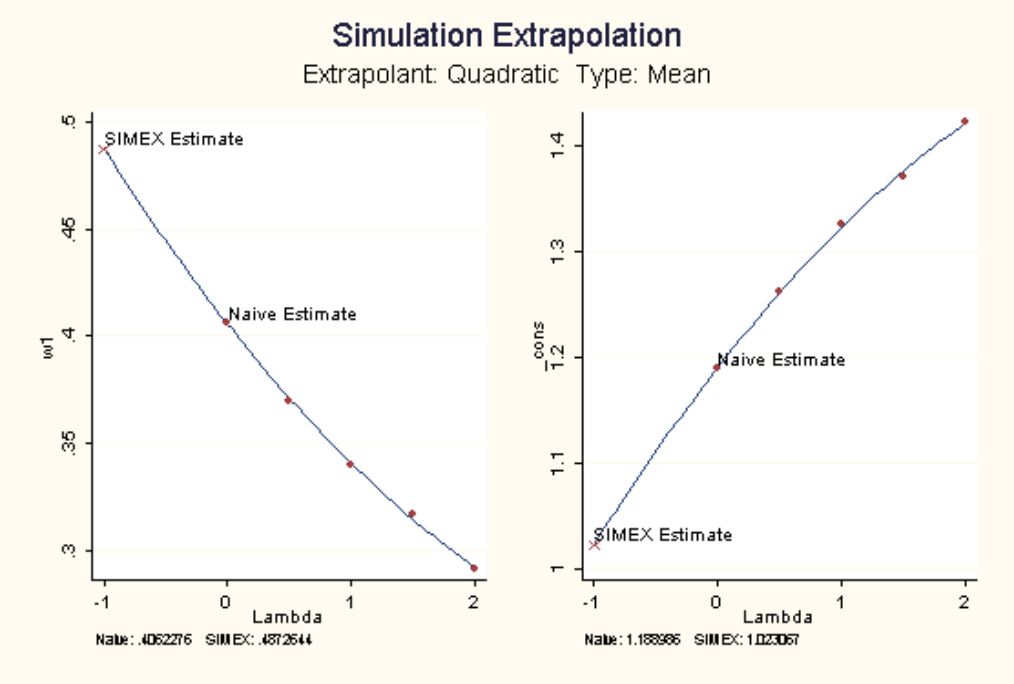

Abbildung 22.2: SIMEX-Schätzer im einfachen linearen Modell – Quadratische Extrapolationsfunktion

Man kann den SIMEX-Schätzer jedoch auch formal darstellen (Cook und Stefanski 1994). Hierzu unterstellt man eine Schätzprozedur, die den Datensatz in den Parameterraum $\boldsymbol{\Theta}$ überführt. Es gilt für den Parametervektor $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Außerdem definiert man $\mathbf{T}$ als diejenige Funktion, durch die die Schätzprozedur abgebildet wird. Dann kann man die folgenden Schätzer definieren:

$$\hat{\boldsymbol{\theta}}_{wahr} = \mathbf{T}(\{y_i, x_i\}_1^n) \tag{22.148}$$

und

$$\hat{\boldsymbol{\theta}}_{naiv} = \mathbf{T}(\{y_i, x_i^a\}_1^n). \tag{22.149}$$

Außerdem gilt:

$$\hat{\boldsymbol{\theta}}_b(\lambda) = \mathbf{T}(\{y_i, x_{b,i}^a(\lambda)\}_1^n) \tag{22.150}$$

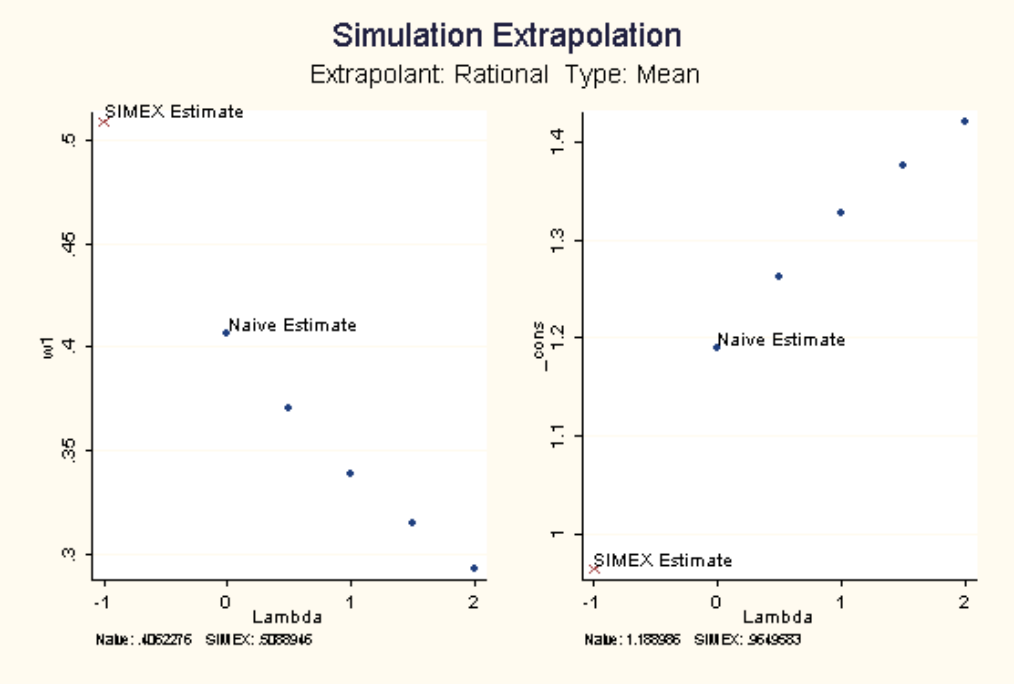

Abbildung 22.3: SIMEX-Schätzer im einfachen linearen Modell – Rational lineare (nichtlineare) Extrapolationsfunktion (Kurve wird von STATA nicht ausgewiesen)

und

$$\hat{\boldsymbol{\theta}}(\lambda) = E\left(\hat{\boldsymbol{\theta}}_b(\lambda)|\,\{y_i, x_i^a\}_1^n\right). \tag{22.151}$$

Nun wird angenommen, dass die Schätzer $\hat{\boldsymbol{\theta}}_{wahr}$, $\hat{\boldsymbol{\theta}}_{naiv}$, $\hat{\boldsymbol{\theta}}_b(\lambda)$ und $\hat{\boldsymbol{\theta}}(\lambda)$ endliche Erwartungswerte haben und dass die Schätzer aufgrund des schwachen Gesetzes der großen Zahlen in Wahrscheinlichkeit gegen die jeweiligen Erwartungswerte konvergieren (Cook und Stefanski 1994).

Aus Gleichung (22.151) folgt, dass gilt $E\left(\hat{\boldsymbol{\theta}}(\lambda)\right) = E\left(\hat{\boldsymbol{\theta}}_b(\lambda)\right)$. Annahmegemäß konvergiert $\hat{\boldsymbol{\theta}}_b(\lambda)$ in Wahrscheinlichkeit gegen seinen Erwartungswert, den man als Funktion des wahren Parametervektors $\boldsymbol{\theta}_0$ und der Varianz des gesamten Fehlers (der gesamten Überlagerung) $(\sigma_v^2(1+\lambda))$ betrachten kann. Man kann diese Funktion somit als $\mathbf{T}(\boldsymbol{\theta}_0, \sigma_v^2(1+\lambda))$ schreiben. Somit konvergieren annahmengemäß sowohl $\hat{\boldsymbol{\theta}}_b(\lambda)$ als auch $\hat{\boldsymbol{\theta}}(\lambda)$ in Wahrscheinlichkeit gegen $\mathbf{T}(\boldsymbol{\theta}_0, \sigma_v^2(1+\lambda))$. Falls nun $\sigma_v^2 = 0$ gilt, also keine Überlagerung vorliegt, so

gilt $\hat{\boldsymbol{\theta}}_{wahr} = \hat{\boldsymbol{\theta}}_{naiv} = \hat{\boldsymbol{\theta}}_b(\lambda) = \hat{\boldsymbol{\theta}}(\lambda)$. Somit folgt, dass, falls $\hat{\boldsymbol{\theta}}_{wahr}$ ein konsistenter Schätzer für $\boldsymbol{\theta}_0$ ist, die Schlussfolgerung gezogen werden kann, dass $\boldsymbol{\theta}_0 = \mathbf{T}(\boldsymbol{\theta}_0, 0)$ gilt.

Nimmt man weiter an, dass $\mathbf{T}(.\ ,\ .)$ eine stetige Funktion ist, so ergibt sich:

$$\lim_{\lambda \searrow -1} E\left(\hat{\boldsymbol{\theta}}(\lambda)\right) = \lim_{\lambda \searrow -1} \mathbf{T}(\boldsymbol{\theta}_0, \sigma_v^2(1+\lambda)) = \mathbf{T}(\boldsymbol{\theta}_0, 0) = \boldsymbol{\theta}_0. \tag{22.152}$$

Die Extrapolationsfunktion approximiert somit $E\left(\hat{\boldsymbol{\theta}}(\lambda)\right)$. Ihre Extrapolation an der Stelle $\lambda = -1$ approximiert $\boldsymbol{\theta}_0$. Insofern stellt der SIMEX-Schätzer eine Approximation von $\boldsymbol{\theta}_0$ dar. Er ist generell lediglich approximativ konsistent, weil es sich bei der geschätzten Extrapolationsfunktion ebenfalls um eine Approximation handelt. Lediglich in den Spezialfällen, in denen die Extrapolationsfunktion exakt ist, ist auch der SIMEX-Schätzer exakt konsistent (Cook und Stefanski 1994).

Bisher wurde von normalverteilten Überlagerungen ausgegangen. Cook und Stefanski (1994) finden jedoch, dass der SIMEX-Schätzer gegenüber der Verletzung dieser Annahme robust ist. Außerdem wurde bisher lediglich der Fall einer additiven stochastischen Überlagerung betrachtet. Von Interesse ist jedoch auch in nichtlinearen Modellen die multiplikative stochastische Überlagerung, insbesondere wegen ihrer größeren Schutzwirkung bei Großunternehmen.

Generell gilt, dass der SIMEX-Schätzer nicht auf den Fall der additiven Überlagerung beschränkt ist (Carroll et al. 1995). Die überlagerte Variable kann durch eine Transformation H in ein Modell mit additiver Überlagerung überführt werden, so dass $H(X^a) = H(X) + V$ gilt. Falls zu H eine inverse Funktion F existiert, gilt für den Simulationsschritt:

$$X_{b,i}^a(\lambda) = F\left(H(X_i^a) + \lambda^{1/2} V_{b,i}\right). \tag{22.153}$$

Im Fall einer multiplikativen stochastischen Überlagerung gilt $H = log()$ und $F = exp()$ (Carroll et al. 1995) und somit

$$X_{b,i}^a(\lambda) = exp\left(log(X_i^a) + \lambda^{1/2} V_{b,i}\right). \tag{22.154}$$

Allerdings kann die multiplikative Überlagerung mit einem Überlagerungsfaktor W auch behandelt werden, als ob es sich um eine additive Überlagerung mit V handeln würde. Dies kann wie folgt veranschaulicht werden: Im Fall der additiven Überlagerung (mit einem Erwartungswert von Null) gilt für den Erwartungswert der anonymisierten Variablen

$$E(X^{a(add)}) = \mu_x. \tag{22.155}$$

Für die Varianz der anonymisierten Variablen ergibt sich (vgl. Abschnitt 19.1, Gleichung (19.10)):

$$var(X^{a(add)}) = \sigma_x^2 + \sigma_v^2. \tag{22.156}$$

Für den Erwartungswert einer multiplikativ überlagerten Variablen (mit Erwartungswert des Überlagerungsfaktors von Eins) ergibt sich ebenfalls Erwartungstreue (vgl. Abschnitt 19.1):

$$E(X^{a(mult)}) = \mu_x. \tag{22.157}$$

Für die Varianz hingegen gilt im Fall der multiplikativen Überlagerung (vgl. Abschnitt 19.1, Gleichung (19.11)) :

$$var(X^{a(mult)}) = \sigma_x^2 + \left(\sigma_x^2 + \mu_x^2\right)\sigma_w^2. \tag{22.158}$$

Setzt man die beiden Varianzformeln (Gleichungen (22.156) und (22.158)) gleich und löst nach σ_v^2 auf, so ergibt sich:

$$\sigma_v^2 = \left(\sigma_x^2 + \mu_x^2\right)\sigma_w^2. \tag{22.159}$$

Das gleiche Ergebnis erhält man, wenn man die multiplikative Überlagerung direkt als additiven Fehler betrachtet. Es ergibt sich dann der folgende Zusammenhang:

$$X + V = XW \tag{22.160}$$

und daraus

$$V = XW - X = X(W - 1). \tag{22.161}$$

Für den Erwartungswert von V ergibt sich somit bei Unabhängigkeit von X und W:

$$E(V) = E(X(W - 1)) = E(X)E(W - 1) = 0 \tag{22.162}$$

und für die Varianz von V gilt:

$$
\begin{aligned}
var(V) = var(X(W-1)) &= E\left[(X(W-1)-\mu_x)(X(W-1)-\mu_x)\right] \\
&= E\left[(X^2(W-1)^2 - 2X(W-1)\mu_x + \mu_x^2\right] \\
&= E(X^2)E\left[(W-1)^2\right] - 2\mu_x E(X)E(W-1) + \mu_x^2 \\
&= E(X^2)E\left[W^2 - 2W + 1\right] - 2\mu_x^2 \cdot 0 + \mu_x^2 \\
&= E(X^2)\left(E(W^2) - 2E(W) + 1\right) + \mu_x^2 \\
&= E(X^2)\left(E(W^2) - 1\right) + \mu_x^2 \\
&= E(X^2)E(W^2) - E(X^2) + \mu_x^2 \\
&= (\sigma_x^2 + \mu_x^2)(\sigma_w^2 + \mu_w^2) - E(X^2) + \mu_x^2 \\
&= (\sigma_x^2 + \mu_x^2)(\sigma_w^2 + 1) - \sigma_x^2 \\
&= \sigma_x^2\sigma_w^2 + \sigma_x^2 + \mu_x^2\sigma_w^2 - \sigma_x^2 \\
&= (\sigma_x^2 + \mu_x^2)\sigma_w^2. \qquad (22.163)
\end{aligned}
$$

Somit kann der multiplikative Fehler W wie ein additiver Fehler V mit der Fehlervarianz $var(V) = (\sigma_x^2 + \mu_x^2)\sigma_w^2$ betrachtet werden. Allerdings ist der Fehler nun nicht mehr von der Ausgangsvariable X unabhängig. Entsprechend ist die Fehlervarianz auch durch Erwartungswert und Varianz von X determiniert.

Im Rahmen des Simulationsschritts zur Berechnung des SIMEX-Schätzers werden nun normalverteilte Pseudo-Zufallsvariablen mit Erwartungswert Null und Varianz $(\sigma_x^2 + \mu_x^2)\sigma_w^2$ erzeugt. Es ergibt sich somit für die Varianz von $X_{b,i}^a(\lambda)$:

$$
var(X_{b,i}^a(\lambda)) = \sigma_x^2 + (1+\lambda)(\sigma_x^2 + \mu_x^2)\sigma_w^2. \qquad (22.164)
$$

Folglich kann auch im Fall einer multiplikativen stochastischen Überlagerung mit einem SIMEX-Schätzer, der für additive Überlagerungen programmiert wurde, eine Korrektur durchgeführt werden.

Lechner und Pohlmeier (2005) zeigen, dass der SIMEX-Schätzer auch bei nichtparametrischen Regressionsschätzern zu akzeptablen Ergebnissen führt, so dass die Benutzung des SIMEX-Schätzers nicht davon abhängt, ob der Forscher die Form des wahren Regressionszusammenhangs kennt.

Die Ausführungen zeigen, dass der SIMEX-Schätzer sehr einfach und anschaulich ist. Allerdings sind die Eigenschaften des Schätzers sehr komplex. Deshalb ist auch die Berechnung der Standardfehler und Teststatistiken schwierig (Carroll et al. 1995). Carroll et al. (1996) untersuchen die asymptotische Verteilung des SIMEX-Schätzers für parametrische Modelle. Unter der Annahme, dass die Variablen identisch und unabhängig verteilt sind, zeigen sie, dass der SIMEX-Schätzer asymptotisch normalverteilt ist und leiten einen Schätzer für dessen asymptotische Varianz-Kovarianzmatrix her.

Stefanski und Cook (1996) leiten eine Methode der Varianzschätzung her, die angewendet werden kann, wenn σ_v^2 bekannt ist oder so gut geschätzt werden kann, dass man diese Annahme treffen kann. Diese Methode wird in Carroll et al. (1995) beschrieben und ist angelehnt an Tukeys Jackknife-Varianz-Schätzer.

Ausgangspunkt ist die Funktion $\mathbf{T}$, die bereits in den Gleichungen (22.148) bis (22.151) eingeführt wurde und die eine funktionale Beschreibung der Schätzprozedur darstellt. Man definiert zusätzlich die Funktion $\mathbf{T}_{var}$ als diejenige Funktion, welche allgemein die Berechnung des Varianzschätzers darstellt. Beispielsweise kann $\mathbf{T}_{var}$ die Inverse der Informationsmatrix darstellen, sofern $\hat{\boldsymbol{\theta}}$ ein Maximum-Likelihood-Schätzer ist. Häufig handelt es sich auch um einen Sandwich-Schätzer. Bei der praktischen Umsetzung kann es sich auch um einen Bootstrap-Schätzer handeln. Es gilt somit:

$$\mathbf{T}_{var}\left(\{Y_i, X_i\}_1^n\right) = \widehat{var}\left(\hat{\boldsymbol{\theta}}_{wahr}\right) = \widehat{var}\left[\mathbf{T}\left(\{Y_i, X_i\}_1^n\right)\right]. \tag{22.165}$$

Nun werden folgende Definitionen vorgenommen:

$$\tau^2 = var\left(\hat{\boldsymbol{\theta}}_{wahr}\right), \tag{22.166}$$

$$\hat{\tau}^2_{wahr} = \mathbf{T}_{var}\left(\{Y_i, X_i\}_1^n\right) \tag{22.167}$$

und

$$\hat{\tau}^2_{naiv} = \mathbf{T}_{var}\left(\{Y_i, X_i^a\}_1^n\right). \tag{22.168}$$

Für den Erwartungswert des SIMEX-Schätzers gilt, dass er approximativ dem wahren Schätzer entspricht:

$$E\left(\hat{\boldsymbol{\theta}}_{SIMEX} | \{Y_i, X_i\}_1^n\right) \approx \hat{\boldsymbol{\theta}}_{wahr}. \tag{22.169}$$

Hieraus folgt für Varianz des SIMEX-Schätzers:

$$var\left(\hat{\boldsymbol{\theta}}_{SIMEX}\right) \approx var\left(\hat{\boldsymbol{\theta}}_{wahr}\right) + var\left(\hat{\boldsymbol{\theta}}_{SIMEX} - \hat{\boldsymbol{\theta}}_{wahr}\right). \tag{22.170}$$

Die Varianz des SIMEX-Schätzers kann somit in eine Stichprobenvariabilität und eine Messfehlervariabilität zerlegt werden.

Die Stichprobenvariabilität $var\left(\hat{\boldsymbol{\theta}}_{wahr}\right) = \tau^2$ kann im Rahmen des Simulationsschritts zur Bestimmung des SIMEX-Schätzers wie folgt approximiert werden. Für die Varianz der einzelnen simulierten Schätzer in Abhängigkeit von λ gilt:

$$\hat{\tau}_b^2(\lambda) = \mathbf{T}_{var}\left[\left\{Y_i, X_{b,i}^a(\lambda)\right\}_1^n\right] \tag{22.171}$$

und damit für die mittlere Varianz des Schätzers in Abhängigkeit von λ

$$\hat{\tau}^2(\lambda) = \frac{1}{B}\sum_{b=1}^{B}\hat{\tau}_b^2(\lambda). \tag{22.172}$$

Somit stellt wiederum die Extrapolation an der Stelle $\lambda = -1$ eine Approximation von $\tau^2 = var\left(\hat{\boldsymbol{\theta}}_{wahr}\right)$ dar.

Auch die Messfehlervariabilität kann mit Hilfe der im Rahmen des Simulationsschritts erzeugten Schätzers approximiert werden. Hierzu wird zunächst folgende Differenz definiert:

$$\boldsymbol{\Delta}_b(\lambda) = \hat{\boldsymbol{\theta}}_b(\lambda) - \hat{\boldsymbol{\theta}}(\lambda) \qquad b = 1, \ldots, B. \tag{22.173}$$

Für die Varianz dieser Differenz kann man schreiben:

$$s_\Delta^2(\lambda) = (B-1)^{-1}\sum_{b=1}^{B}\left(\hat{\boldsymbol{\theta}}_b(\lambda) - \hat{\boldsymbol{\theta}}(\lambda)\right)\left(\hat{\boldsymbol{\theta}}_b(\lambda) - \hat{\boldsymbol{\theta}}(\lambda)\right)'. \tag{22.174}$$

und damit gilt für den Erwartungswert dieser Varianz:

$$E\left(s_\Delta^2(\lambda)\right) = var\left(\hat{\boldsymbol{\theta}}_b(\lambda) - \hat{\boldsymbol{\theta}}(\lambda)\right). \tag{22.175}$$

Stefanski und Cook (1996) zeigen, dass gilt:

$$\begin{aligned} var\left(\hat{\boldsymbol{\theta}}_{SIMEX} - \hat{\boldsymbol{\theta}}_{wahr}\right) &= -\lim_{\lambda \searrow -1} var\left(\hat{\boldsymbol{\theta}}_b(\lambda) - \hat{\boldsymbol{\theta}}(\lambda)\right) \\ &= -\lim_{\lambda \searrow -1} E\left(s_\Delta^2(\lambda)\right). \end{aligned} \tag{22.176}$$

Somit kann die Varianz des SIMEX-Schätzers in folgender Weise approximiert werden:

$$var\left(\hat{\boldsymbol{\theta}}_{SIMEX}\right) \approx \lim_{\lambda \searrow -1} \hat{\tau}^2(\lambda) - \lim_{\lambda \searrow -1} s^2_{\Delta}(\lambda). \tag{22.177}$$

Bei der Implementation des SIMEX-Schätzers im Programmpaket STATA kann die Varianz auf jeder Stufe durch einen Bootstrap-Schätzer bestimmt werden. Die Varianz der Überlagerungen kann insbesondere durch den Nutzer angegeben werden, was voraussetzt, dass die datenproduzierende Institution diese den Nutzern preisgibt, oder geschätzt werden, sofern dem Nutzer mindestens zwei überlagerte Datenfiles zur Verfügung stehen. Dabei muss die Varianz der Überlagerungen in beiden Fällen gleich gewählt werden.

22.2 Monte-Carlo-Simulationen

22.2.1 Stochastische Überlagerungen in linearen Modellen

a) Stochastische Überlagerungen in einem linearen Modell mit nicht transformierten Variablen

Grundlage der folgenden Simulationsexperimente zur Wirkung von stochastischen Überlagerungen ist zunächst das folgende lineare Regressionsmodell:

$$Y = 1 + X_1 - X_2 + 0,5X_3 + U. \tag{22.178}$$

Dabei sind X_1, X_2 und X_3 normalverteilt mit Erwartungswert Null und Varianz-Kovarianzmatrix

$$\mathbf{C} = \begin{pmatrix} 1 & 0,2 & 0,4 \\ 0,2 & 1 & 0,1 \\ 0,4 & 0,1 & 1 \end{pmatrix}.$$

Der Störterm U ist standardnormalverteilt. Es werden jeweils 1.000 Replikationen von Monte-Carlo-Simulationen auf der Grundlage von jeweils 1.000 Beobachtungen durchgeführt.[27]

Mit Hilfe der Simulationsexperimente werden insbesondere die in den vorangegangenen Abschnitten theoretisch hergeleiteten Ergebnisse zur Wirkung von stochastischen Über-

27) Die gleichen Simulationsergebnisse wurden auch mit 1.000 Beobachtungen und 100 Replikationen sowie 10.000 Beobachtungen und 100 Replikationen durchgeführt. Da die Ergebnisse jedoch grundsätzlich gleich sind, wird auf eine Darstellung dieser Simulationsvarianten verzichtet.

lagerungen in linearen Modellen überprüft. Daneben bieten die Monte-Carlo-Simulationen auch zusätzliche Erkenntnisse.[28]

Die Tabellen 22.1 bis 22.13 enthalten die mittleren Schätzergebnisse aus den 1.000 Replikationen für unterschiedliche Verfahren der stochastischen Überlagerung im Vergleich zu den Schätzungen mit den Originaldaten. Mit Hilfe von t-Tests zum Signifikanzniveau von 5% wird jeweils untersucht, ob die Parameterschätzer bei Anonymisierung signifikant von den Schätzwerten mit Originaldaten sowie von den Parametern des theoretischen Modells abweichen.

Tabelle 22.1 zeigt zunächst, dass die Parameterschätzer verzerrt sind, wenn alle Variablen mit einem additiven Fehler überlagert werden. In einem Fall werden die Variablen Y, X_1, X_2 und X_3 mit einem einheitlichen Fehler überlagert, im anderen Fall mit unterschiedlichen Fehlern. Der Erwartungswert des Fehlers beträgt jeweils Null, die Varianz 1/3.

Man erkennt, dass die Abweichungen der Parameterschätzer größer sind, wenn die Variablen mit unterschiedlichen Fehlern überlagert werden. Lediglich der Koeffizientenschätzer für die Konstante weicht nicht signifikant vom Originalwert beziehungsweise dem theoretischen Wert ab. In der rechten Spalte der Tabelle wird jeweils eine Korrektur mit Hilfe des Instrumentvariablen-Schätzers vorgenommen. Die Instrumentvariablen wurden ebenfalls durch eine additive Überlagerung der Originalvariablen durch Fehler mit gleicher Varianz erzeugt. Durch die Korrektur mit Hilfe des IV-Schätzers sind alle Parameterschätzer nicht mehr signifikant von den Originalschätzern und den theoretischen Parametern verschieden. Allerdings ergeben sich in jedem Fall verzerrte t-Werte, so dass die Qualität der Schätzung durch die Anonymisierung Schaden nimmt. Die Verzerrungen der Teststatistiken fallen bei der einfachen additiven Überlagerung mit unterschiedlichem Fehler höher aus als bei der einfachen additiven Überlagerung mit gleichem Fehler.

Tabelle 22.2 zeigt die Ergebnisse für eine additive Überlagerung aller Variablen, bei der die Varianz-Kovarianzmatrix der Überlagerungen proportional zur Varianz-Kovarianzmatrix der Originalvariablen gewählt wird. Der Proportionalitätsfaktor beträgt $d = 0,1$. Daneben sind die Ergebnisse des Verfahrens von Kim dargestellt, bei dem die derart überlagerten Werte zusätzlich einer Transformation unterzogen werden, so dass die ersten und zweiten Momente der Originaldaten erhalten bleiben. Beide Verfahren erhalten die Parameterschätzer gegenüber den Originaldaten und gegenüber dem theoretischen Modell. Das Kim-Verfahren erhält zusätzlich die Teststatistiken, wobei in diesem Fall auch die Überlagerung mit einer proportionalen Varianz-Kovarianzmatrix zu einem annäherenden Erhalt der Teststatistiken führt.

Für den Fall einer multiplikativen Überlagerung aller Variablen mit einer Gleichverteilung im Intervall (0,5;1,5) finden sich die Ergebnisse der Monte-Carlo-Simulationen in den Tabellen

28) Es muss darauf hingewiesen werden, dass für jede durchgeführte Schätzung die Daten wieder neu simuliert wurden. Deshalb stimmen die Parameterschätzer auch in den Fällen, in denen durch die Anonymisierung keine Verzerrung hervorgerufen wird, nicht exakt mit den Originalschätzern überein.

22.3 und 22.4. In Tabelle 22.3 findet man zunächst das Ergebnis aus Unterabschnitt 22.1.1 bestätigt, dass die multiplikative Überlagerung aller Variablen mit einem einheitlichen Faktor zu unverzerrten Schätzer und unverzerrten Teststatistiken führt, weil für alle Regressoren ein einheitlicher Mittelwert von Null unterstellt wurde ($\sum_{i=1}^{n} x_i = 0$). Dagegen führt die multiplikative Überlagerung mit unterschiedlichen Faktoren auch in diesem Fall zu verzerrten Schätzern und verzerrten Teststatistiken. Die verzerrten Parameterschätzer können, wie ebenfalls in Tabelle 22.3 dargestellt ist, beispielsweise mit Hilfe des IV-Schätzers korrigiert werden.

Während die Korrektur der ersten und zweiten Momente bei der additiven Überlagerung zu unverzerrten Schätzern, gar zu unverzerrten Teststatistiken, führt, ist dies bei der multiplikativen Überlagerung nicht der Fall. Werden vorher konstante Faktoren entsprechend transformiert, so wird dadurch eine Verzerrung der Parameterschätzer hervorgerufen, die t-Werte bleiben hingegen offenbar erhalten.

Tabelle 22.5 zeigt, dass die multiplikative Überlagerung mit der „Mischungsverteilung“ nach dem Verfahren von Höhne wegen der geringen Varianz der Überlagerung und der annährend gleichen Überlagerungsfaktoren bei Überlagerung aller Variablen nur zu kaum merklichen Verzerrungen führt, die ebenfalls durch eine IV-Schätzung korrigiert werden können. Bei diesem Verfahren handelt es sich um keine echte Mischungsverteilung. Vielmehr wird zunächst für jede Einheit mit einer Wahrscheinlichkeit von 50 Prozent entschieden, ob eine Vergrößerung oder Verkleinerung der Merkmalswerte vorgenommen wird. Hierzu werden Grundüberlagerungsfaktoren in der Höhe von $1 + / - f$ festgelegt. Anschließend werden diese mit einem additiven Fehler mit Standardabweichung s überlagert. In diesem Fall werden $f = 0,11$ und $s = 0,03$ gewählt.

Tabelle 22.6 zeigt, dass keine Verzerrung der Parameterschätzer zu beobachten ist, wenn lediglich die abhängige Variable additiv oder multiplikativ überlagert wird. Durch den zusätzlichen Fehler steigt jedoch die Stichprobenvarianz und somit auch die Varianz der Schätzer, was eine Verzerrung (Reduzierung) der Teststatistiken nach sich zieht.

In den Tabellen 22.7 und 22.8 wird gezeigt, dass die additive stochastische Überlagerung mit einer proportionalen Varianz-Kovarianzmatrix und die multiplikative Überlagerung mit einem konstanten Faktor, die bei der Anonymisierung aller Variablen noch zu unverzerrten Parameterschätzungen geführt haben, ebenfalls verzerrte Parameterschätzer produzieren, sofern nur die Regressoren (Tabelle 22.7) beziehungsweise die abhängigen Variablen und ein Teil der Regressoren (Tabelle 22.8) anonymisiert werden. Das gleiche Phänomen beobachtet man, wenn die überlagerten Variablen zusätzlich der Kim-Transformation unterzogen werden. Allerdings sind die Abweichungen der Parameterschätzer dann geringer (Tabellen 22.9 und 22.10). In diesen Fällen sind dann wieder Korrekturverfahren anwendbar, wie am Beispiel der IV-Schätzung bei der multiplikativen Überlagerung gezeigt wird. Dennoch ist die Anonymisierung dann in jedem Fall durch die Verzerrung der Teststatistiken mit einem Qualitätsverlust verbunden.

Überlagert man beim Verfahren von Höhne ebenfalls nur einen Teil der Variablen (Tabellen 22.10 bis 22.13), so treten stärkere Verzerrungen auf als bei der Überlagerung aller Variablen. Dies liegt daran, dass nun bei konstanten Faktoren keine Unverzerrtheit mehr vorliegt. Die auftretende Verzerrung kann auch durch den Einsatz des IV-Schätzers nicht exakt korrigiert werden.

Tabelle 22.1: MC-Simulationen – Lineares Modell, einfache additive Überlagerung, alle Variablen anonymisiert, 1.000 Replikationen

	Original	Einfache additive Überlagerung mit gleichem Fehler					
		nicht korrigiert			korrigiert (IV)		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	1,054			1,000	*	*
(t-Werte)	(28,57)	(30,10)			(26,31)		
X_2	-0,999	-0,918			-1,000	*	*
(t-Werte)	(-30,91)	(-29,73)			(-26,90)		
X_3	0,499	0,569			0,500	*	*
(t-Werte)	(14,45)	(16,78)			(13,05)		
Konst.	1,001	0,999	*	*	1,001	*	*
(t-Werte)	(31,66)	(30,87)			(30,37)		

	Original	Einfache additive Überlagerung mit unterschiedlichem Fehler					
		nicht korrigiert			korrigiert (IV)		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	0,725			0,998	*	*
(t-Werte)	(28,57)	(18,22)			(16,16)		
X_2	-0,999	-0,705			-1,000	*	*
(t-Werte)	(-30,91)	(-18,51)			(-18,33)		
X_3	0,499	0,437			0,503	*	*
(t-Werte)	(14,45)	(11,09)			(8,31)		
Konst.	1,001	0,999	*	*	1,001	*	*
(t-Werte)	(31,66)	(22,99)			(21,86)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.2: MC-Simulationen – Lineares Modell, additive Überlagerung mit proportionaler Varianz-Kovarianzmatrix und Kim-Verfahren, alle Variablen anonymisiert, 1.000 Replikationen

	Original	Additive Überlagerung: VCV-Matrix der Fehler proportional zur VCV-Matrix der Originalvariablen			Kim-Verfahren		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	1,000	*	*	1,001	*	*
(t-Werte)	(28,57)	(28,51)			(28,56)		
X_2	-0,999	-1,000	*	*	-0,999	*	*
(t-Werte)	(-30,91)	(-30,93)			(-30,95)		
X_3	0,499	0,501	*	*	0,498	*	*
(t-Werte)	(14,45)	(14,48)			(14,43)		
Konst.	1,001	0,999	*	*	0,999	*	*
(t-Werte)	(31,66)	(30,08)			(31,62)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.3: MC-Simulationen – Lineares Modell, multiplikative Überlagerung (Gleichverteilung (0,5;1,5)), alle Variablen anonymisiert, 1.000 Replikationen

	Original	Multiplikative Überlagerung mit unterschiedlichen Faktoren				Multiplikative Überlagerung mit konstantem Faktor			
		nicht korrigiert		korrigiert (IV)					
			Übereinstimmung zum Signifikanzniveau von 5% mit dem		Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original-wert	theor. Wert	Original-wert	theor. Wert		Original-wert	theor. Wert
X_1	1,001	0,912			1,000	*	0,999	*	*
(t-Werte)	(28,57)	(22,34)			(21,74)		(27,39)		
X_2	-0,999	-0,906			-1,000	*	-1,001	*	*
(t-Werte)	(-30,91)	(-23,75)			(-23,92)		(-29,84)		
X_3	0,499	0,486			0,500	*	0,500	*	*
(t-Werte)	(14,45)	(12,05)			(11,06)		(13,94)		
Konst.	1,001	1,000	*	*	0,998	*	1,000	*	*
(t-Werte)	(31,66)	(25,67)			(25,46)		(29,24)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.4: MC-Simulationen – Lineares Modell, multiplikative Überlagerung (Gleichverteilung (0,5;1,5)) mit Transformation zum Erhalt der ersten und zweiten Momente, alle Variablen anonymisiert, 1.000 Replikationen

	Original	Multiplikative Überlagerung, konstante Faktoren, mit Transformation			Multiplikative Überlagerung, unterschiedliche Faktoren, mit Transformation		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	0,987			0,898		
(t-Werte)	(28,57)	(27,48)			(22,25)		
X_2	-0,999	-0,990			-0,892		
(t-Werte)	(-30,91)	(-29,90)			(-23,70)		
X_3	0,499	0,495			0,480		
(t-Werte)	(14,45)	(14,01)			(12,06)		
Konst.	1,001	0,999	*	*	1,000	*	*
(t-Werte)	(31,66)	(30,84)			(27,02)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.5: MC-Simulationen – Lineares Modell, multiplikative Überlagerung nach dem Verfahren von Höhne, alle Variablen anonymisiert, 1.000 Replikationen

	Original	Multiplikative Überlagerung, Mischungsverteilung nach dem Verfahren von Höhne					
		nicht korrigiert			korrigiert (IV)		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor.Wert		Originalwert	theor. Wert
X_1	1,001	0,998			1,001	*	*
(t-Werte)	(28,57)	(28,21)			(27,85)		
X_2	-0,999	-0,997	*		-1,001	*	*
(t-Werte)	(-30,91)	(-30,62)			(-30,32)		
X_3	0,499	0,501	*	*	0,499	*	*
(t-Werte)	(14,45)	(14,40)			(14,14)		
Konst.	1,001	1,000	*	*	1,000	*	*
(t-Werte)	(31,66)	(31,16)			(31,13)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.6: MC-Simulationen – Lineares Modell, additive (NV) und multiplikative (GV) Überlagerung, nur abhängige Variable anonymisiert, 1.000 Replikationen

	Original	Einfache additive Überlagerung			Einfache multiplikative Überlagerung		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	1,001	*	*	0,999	*	*
(t-Werte)	(28,57)	(24,64)			(24,53)		
X_2	-0,999	-0,999	*	*	-0,999	*	*
(t-Werte)	(-30,91)	(-26,87)			(-26,68)		
X_3	0,499	0,499	*	*	0,502	*	*
(t-Werte)	(14,45)	(12,53)			(12,54)		
Konst.	1,001	1,001	*	*	0,999	*	*
(t-Werte)	(31,66)	(27,39)			(27,24)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.7: MC-Simulationen – Lineares Modell, additive Überlagerung (NV) mit proportionaler Varianz-Kovarianzmatrix und multiplikative Überlagerung (GV) mit konstantem Faktor, nur Regressoren anonymisiert, 1.000 Replikationen

	Original	Additive Überlagerung: VCV-Matrix der Fehler proportional zur VCV-Matrix der Originalvariablen			Multiplikative Überlagerung mit konstantem Faktor					
					nicht korrigiert			korrigiert (IV)		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original-wert	theor. Wert		Original-wert	theor. Wert		Original-wert	theor. Wert
X_1	1,001	0,910			0,924			1,001	*	*
(t-Werte)	(28,57)	(24,83)			(25,38)			(25,24)		
X_2	-0,999	-0,909			-0,924			-1,001	*	*
(t-Werte)	(-30,91)	(-26,94)			(-27,53)			(-27,38)		
X_3	0,499	0,453			0,461			0,500	*	*
(t-Werte)	(14,45)	(12,56)			(12,86)			(12,79)		
Konst.	1,001	1,001	*	*	1,000	*	*	1,000	*	*
(t-Werte)	(31,66)	(28,88)			(29,27)			(29,07)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.8: MC-Simulationen – Lineares Modell, additive Überlagerung (NV) mit proportionaler Varianz-Kovarianzmatrix und multiplikative Überlagerung (GV) mit konstantem Faktor, abhängige und zwei erklärende Variablen anonymisiert, 1.000 Replikationen

	Original	Additive Überlagerung: VCV-Matrix der Fehler proportional zur VCV-Matrix der Originalvariablen			Multiplikative Überlagerung mit konstantem Faktor					
					nicht korrigiert			korrigiert (IV)		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original-wert	theor. Wert		Original-wert	theor. Wert		Original-wert	theor. Wert
X_1	1,001	0,980			0,982			0,999	*	*
(t-Werte)	(28,57)	(25,77)			(25,29)			(25,26)		
X_2	-0,999	-0,981			-0,982			-1,001	*	*
(t-Werte)	(-30,91)	(-29,13)			(-28,35)			(-26,60)		
X_3	0,499	0,541			0,536			0,500	*	*
(t-Werte)	(14,45)	(15,12)			(14,57)			(12,37)		
Konst.	1,001	1,000	*	*	0,999	*	*	1,001	*	*
(t-Werte)	(31,66)	(28,85)			(28,27)			(28,33)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.9: MC-Simulationen – Lineares Modell, Kim-Verfahren und multiplikative Überlagerung (GV), konstante Faktoren mit Kim-Transformation, nur Regressoren anonymisiert, 1.000 Replikationen

	Original	Kim-Verfahren			Multiplikative Überlagerung, konstante Faktoren, mit Transformation		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	0,954			0,960		
(t-Werte)	(28,57)	(24,87)			(25,32)		
X_2	-0,999	-0,953			-0,960		
(t-Werte)	(-30,91)	(-26,94)			(-27,53)		
X_3	0,499	0,476			0,480		
(t-Werte)	(14,45)	(12,58)			(12,84)		
Konst.	1,001	1,000	*	*	1,000	*	*
(t-Werte)	(31,66)	(28,87)			(29,25)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.10: MC-Simulationen – Lineares Modell, Kim-Verfahren und multiplikative Überlagerung (GV), konstante Faktoren mit Kim-Transformation, abhängige und zwei erklärende Variablen anonymisiert, 1.000 Replikationen

	Original	Kim-Verfahren			Multiplikative Überlagerung, konstante Faktoren, mit Transformation		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	0,934			0,933		
(t-Werte)	(28,57)	(25,77)			(25,34)		
X_2	-0,999	-0,890			-0,974		
(t-Werte)	(-30,91)	(-26,43)			(-28,55)		
X_3	0,499	0,495			0,531		
(t-Werte)	(14,45)	(13,84)			(14,61)		
Konst.	1,001	1,001	*	*	0,997		
(t-Werte)	(31,66)	(30,26)			(29,76)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.11: MC-Simulationen – Lineares Modell, multiplikative Überlagerung nach dem Verfahren von Höhne, nur Regressoren anonymisiert, 1.000 Replikationen

	Original	Multiplikative Überlagerung, Mischungsverteilung nach dem Verfahren von Höhne					
		nicht korrigiert			korrigiert (IV)		
			Übereinstimmung zum Signifikanz Niveau von 5% mit dem			Übereinstimmung zum Signifikanz Niveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	0,989			0,987		
(t-Werte)	(28,57)	(27,96)			(27,87)		
X_2	-0,999	-0,986			-0,988		
(t-Werte)	(-30,91)	(-30,23)			(-30,32)		
X_3	0,499	0,493			0,495		
(t-Werte)	(14,45)	(14,16)			(14,19)		
Konst.	1,001	0,999	*	*	1,001	*	*
(t-Werte)	(31,66)	(31,11)			(31,17)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.12: MC-Simulationen – Lineares Modell, multiplikative Überlagerung nach dem Verfahren von Höhne, abhängige und zwei erklärende Variablen anonymisiert, 1.000 Replikationen

	Original	Multiplikative Überlagerung, Mischungsverteilung nach dem Verfahren von Höhne					
		nicht korrigiert			korrigiert (IV)		
			Übereinstimmung zum Signifikanz Niveau von 5% mit dem			Übereinstimmung zum Signifikanz Niveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	0,997			0,996		
(t-Werte)	(28,57)	(27,89)			(27,81)		
X_2	-0,999	-0,997	*		-0,997	*	
(t-Werte)	(-30,91)	(-30,43)			(-30,41)		
X_3	0,499	0,505			0,506		
(t-Werte)	(14,45)	(14,42)			(14,42)		
Konst.	1,001	1,000	*	*	1,000	*	*
(t-Werte)	(31,66)	(30,95)			(30,94)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.13: MC-Simulationen – Lineares Modell, multiplikative Überlagerung nach dem Verfahren von Höhne, zwei Regressoren anonymisiert, 1.000 Replikationen

	Original	Multiplikative Überlagerung, Mischungsverteilung nach dem Verfahren von Höhne					
		nicht korrigiert			korrigiert (IV)		
			Übereinstimmung zum Signifikanz Niveau von 5% mit dem			Übereinstimmung zum Signifikanz Niveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor.Wert
X_1	1,001	1,000	*	*	1,001	*	*
(t-Werte)	(28,57)	(28,26)			(28,35)		
X_2	-0,999	-0,987			-0,989		
(t-Werte)	(-30,91)	(-30,45)			(-30,60)		
X_3	0,499	0,493			0,493		
(t-Werte)	(14,45)	(14,24)			(14,25)		
Konst.	1,001	0,999	*	*	0,999	*	*
(t-Werte)	(31,66)	(31,28)			(31,32)		

* Übereinstimmung zum Signifikanzniveau von 5%

b) Stochastische Überlagerungen in einem linearen Modell mit nichtlinear transformierten Variablen

Nun soll untersucht werden, wie sich die Ergebnisse der Monte-Carlo-Simulationen verändern, wenn die stochastisch überlagerten Variablen, bevor sie in das lineare Schätzmodell eingehen, logarithmiert, also einer nichtlinearen Transformation unterzogen werden. Hierzu wird das folgende linearisierte Modell betrachtet:

$$log(Y) = \alpha + \beta_1 log(X_1) + \beta_2 log(X_2) + \beta_3 log(X_3) + U \tag{22.179}$$

X_1, X_2 und X_3 sind lognormalverteilt. $log(X_1)$, $log(X_2)$ und $log(X_3)$ sind normalverteilt mit Erwartungswerten von Null und Varianz-Kovarianzmatrix

$$\mathbf{C} = \begin{pmatrix} 1 & 0,2 & 0,4 \\ 0,2 & 1 & 0,1 \\ 0,4 & 0,1 & 1 \end{pmatrix}.$$

Wesentlich ist, dass die Anonymisierung vor der Transformation durchgeführt wird, die transformierten Variablen also nicht bei der Anonymisierung berücksichtigt werden. Dies dürfte im Falle der Erstellung eines Scientific-Use-Files auch in der Regel der Fall sein. Würden die logarithmierten Variablen anonymisiert, so würden sich im Prinzip die gleichen Ergebnisse ergeben, wie beim oben betrachteten einfachen linearen Modell.

Im Fall nichtlinear transformierter überlagerter Variablen im linearen Modell können die Fehler-Korrektur-Verfahren für lineare Modelle nur unter bestimmten Umständen eingesetzt werden. Beispielsweise sind Fehler-Korrektur-Modelle für additive Fehler in linearen Modellen verwendbar, wenn ein multiplikativer Fehler vorliegt und sich dieser durch die Logarithmierung in einen additiven Fehler mit Erwartungswert Null umformen lässt. Allerdings kann auch im Fall von nichtlinear transformierten überlagerten Variablen auf Korrekturverfahren für nichtlineare Modelle zurückgegriffen werden. Hier wird der in Unterabschnitt 22.1.2 vorgestellte SIMEX-Schätzer verwendet. Dabei wird eine quadratische Extrapolationsfunktion gewählt[29], die Anzahl der Wiederholungen auf jeder Stufe λ beträgt $b = 50$. Die Varianz des Schätzers wird mit Hilfe eines Bootstrap-Schätzers geschätzt. In der praktischen Umsetzung kann die Varianz der Überlagerungen entweder angegeben werden oder, wie in Unterabschnitt 22.1.2 beschrieben, geschätzt werden.

Bei den Monte-Carlo-Simulationen wird wiederum mit 1.000 Beobachtungen und 1.000 Replikationen gearbeitet. Die Ergebnisse sind in den Tabellen 22.14 bis 22.18 dargestellt.

29) Eine linear rationale Extrapolationsfunktion erwies sich als ungeeignet. Zum einen traten häufiger Polstellen auf, die zu völlig falschen Ergebnissen führten, zum anderen war die Schätzung der Varianz deutlich schlechter.

Diejenigen Verfahren – wie das Kim-Verfahren –, die noch im einfachen linearen Modell zu unverzerrten Schätzern geführt haben, ziehen beim Vorliegen nichtlinear transformierter Variablen eine Verzerrung nach sich. Man erkennt, dass durch die additive Überlagerung, unabhängig davon, ob eine einfache Überlagerung aller Variablen einer Einheit mit dem gleichen Fehler, unterschiedlichen Fehlern (jeweils mit Varianz 1/3) oder mit dem Verfahren von Kim ($d = 0,1$) erfolgt, eine Verzerrung entsteht, die sich weder durch eine Instrumentvariablen-Schätzung noch den SIMEX-Schätzer ausreichend korrigieren lässt. Der Grund hierfür besteht darin, dass durch eine additive Überlagerung vorher positive Werte negativ werden können. Damit ist der Logarithmus nicht mehr definiert und die Beobachtung steht für die Schätzung nicht mehr zur Verfügung. Dies führt zu einer zusätzlichen Verzerrung der Schätzergebnisse.

Dieses Problem tritt bei multiplikativen Überlagerungen nicht auf, da sie die Nichtnegativität der Werte grundsätzlich erhalten. Falls durch die Logarithmierung ein additiver Fehler mit Erwartungswert Null auftritt, können sowohl Fehler-Korrektur-Schätzer für lineare Modelle als auch Fehler-Korrektur-Verfahren für nichtlineare Modelle angewendet werden. Während bei einer multiplikativen Überlagerung mit einer Gleichverteilung im Intervall (0,5;1,5) kein Erwartungswert von Null auftritt, ist dies bei einer Überlagerung mit einer „Mischungsverteilung“ nach dem Verfahren von Höhne annähernd der Fall.

Beim Einsatz des SIMEX-Schätzers ist zu bedenken, dass die Varianz-Kovarianzmatrix der Überlagerungen durch die Transformation verzerrt wird. Somit ist durch die Angabe der Varianz-Kovarianzmatrix der ursprünglichen Überlagerungen keine adäquate Korrektur zu erreichen. Deshalb bietet sich lediglich der Weg an, die Varianz der Überlagerungen der transformierten Variablen durch die Angabe von Instrumenten zu schätzen. Eine solche Schätzung ist jedoch für den Fall konstanter Faktoren nicht möglich, wenn die überlagerten Werte logarithmiert werden und somit konstante additive Fehler entstehen, weil die additiven Fehler der Instrumente für jeweils einen Merkmalsträger konstante Vielfache der additiven Fehler der anonymisierten Merkmale sind.

Entsprechend diesen Überlegungen kann für die multiplikative stochastische Überlagerung mit einem konstanten Faktor keine SIMEX-Korrektur vorgenommen werden. Obwohl der nach der Logarithmierung entstehende additive Fehler keinen Erwartungswert von Null aufweist, wird eine IV-Korrektur vorgenommen, die auch für die meisten Koeffizienten (Ausnahme: Konstante) zu verbesserten Schätzergebnissen führt (Tabelle 22.17).

Erfolgt die multiplikative Überlagerung mit unterschiedlichen Faktoren aus einer Gleichverteilung im Intervall (0,5;1,5) so ist eine SIMEX-Korrektur möglich, die auch im Vergleich zur IV-Schätzung zu den besseren Ergebnissen führt (Tabelle 22.17).

Multiplikative Überlagerungen nach dem Verfahren von Höhne mit den Parametern ($f = 0,11$ und $s = 0,03$) verursachen im Simulationsbeispiel nur eine geringfügige Verzerrung. In diesem Fall führt eine Korrektur mittels IV-Schätzung im Vergleich zum SIMEX-Schätzer zu den besseren Ergebnissen. Dies ist möglicherweise auf zwei Gründe zurückzuführen: Zum einen ist nun der Mittelwert der durch Logarithmierung entstandenen additiven Fehler un-

gefähr Null, was den Einsatz des IV-Schätzers rechtfertigt. Zum anderen stimmt aufgrund der besonderen Konstruktion des Verfahrens die Varianz-Kovarianzmatrix der Überlagerungen der Instrumente beim Höhne-Verfahren nur annähernd mit der Varianz-Kovarianzmatrix der eigentlichen Überlagerungen überein.

Tabelle 22.14: MC-Simulationen – Lineares Modell mit logarithmierten Variablen, additive Überlagerung (NV) mit gleichen Zufallszahlen, 1.000 Replikationen

	Original	Einfache additive Überlagerung mit gleichen Zufallszahlen								
		nicht korrigiert			korrigiert (IV)			korrigiert (SIMEX, Varianz geschätzt)		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert		Originalwert	theor. Wert
log X_1	1,000	0,865			0,954			0,998	*	*
(t-Werte)	(28,53)	(-20,35)			(15,37)			(16,75)		
log X_2	-0,998	-0,794			-0,933			-0,888		
(t-Werte)	(-30,96)	(-20,69)			(-15,01)			(-15,73)		
log X_3	0,499	0,459			0,489			0,534		
(t-Werte)	(14,46)	(11,20)			(7,79)			(9,27)		
a (Konstante)	1,002	1,070			1,060			1,030		
(t-Werte)	(31,69)	(27,12)			(20,24)			(21,78)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.15: MC-Simulationen – Lineares Modell mit logarithmierten Variablen, additive Überlagerung (NV) mit unterschiedlichen Zufallszahlen, 1.000 Replikationen

	Original	Einfache additive Überlagerung mit unterschiedlichen Zufallszahlen								
		nicht korrigiert			korrigiert (IV)			korrigiert (SIMEX, Varianz geschätzt)		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original-wert	theor. Wert		Original-wert	theor. Wert		Original-wert	theor. Wert
log X_1	1,000	0,569			0,943			0,889	*	*
(t-Werte)	(28,53)	(11,69)			(9,20)			(11,53)		
log X_2	-0,998	-0,513			-0,955			-0,886		
(t-Werte)	(-30,96)	(-10,82)			(-9,67)			(-11,63)		
log X_3	0,499	0,359			0,477			0,481		
(t-Werte)	(14,46)	(7,45)			(4,75)			(6,51)		
Konst.	1,002	1,175			1,115			1,117		
(t-Werte)	(31,69)	(22,18)			(15,21)			(16,99)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.16: MC-Simulationen – Lineares Modell mit logarithmierten Variablen, additive Überlagerung (NV) nach dem Kim-Verfahren, 1.000 Replikationen

	Original	Kim-Verfahren									
		nicht korrigiert			korrigiert (IV)			korrigiert (SIMEX, Varianz geschätzt)			
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original-wert	theor. Wert		Original-wert	theor. Wert		Original-wert	theor. Wert
log X_1	1,000	0,320			0,428			0,439		
(t-Werte)	(28,53)	(6,19)			(4,21)			(5,25)		
log X_2	-0,998	-0,227			-0,386			-0,370		
(t-Werte)	(-30,96)	(-4,88)			(-3,87)			(-4,62)		
log X_3	0,499	0,179			0,208			0,226		
(t-Werte)	(14,46)	(3,55)			(2,09)			(2,77)		
a (Konstante)	1,002	2,405			2,346			2,358		
(t-Werte)	(31,69)	(44,06)			(30,03)			(34,05)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.17: MC-Simulationen – Lineares Modell mit logarithmierten Variablen, multiplikative Überlagerung (GV) mit konstanten Faktoren, 1.000 Replikationen

	Original	Multiplikative Überlagerung mit konstanten Faktoren					
		nicht korrigiert			korrigiert (IV)		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
log X_1	1,000	1,021			1,000	*	*
(t-Werte)	(28,53)	(29,10)			(27,81)		
log X_2	-0,998	-0,968			-1,001		*
(t-Werte)	(-30,96)	(-30,45)			(-29,66)		
log X_3	0,499	0,528			0,500	*	*
(t-Werte)	(14,46)	(15,36)			(14,01)		
Konst.	1,002	0,982			0,977		
(t-Werte)	(31,69)	(30,65)			(30,39)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.18: MC-Simulationen – Lineares Modell mit logarithmierten Variablen, multiplikative Überlagerung (GV) mit unterschiedlichen Faktoren, 1.000 Replikationen

	Original	Multiplikative Überlagerung mit unterschiedlichen Faktoren									
		nicht korrigiert				korrigiert (IV)				korrigiert (SIMEX, Varianz geschätzt)			
			Übereinstimmung zum Signifikanzniveau von 5% mit dem				Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert			Originalwert	theor. Wert		Originalwert	theor. Wert
log X_1	1,000	0,899			0,998		*	*	1,002	*	*
(t-Werte)	(28,53)	(24,03)			(23,33)				(25,43)		
log X_2	-0,998	-0,894			-0,999		*	*	-0,999	*	*
(t-Werte)	(-30,96)	(-25,59)			(-25,74)				(-27,72)		
log X_3	0,499	0,484			0,502			*	0,500	*	*
(t-Werte)	(14,46)	(13,11)			(11,93)				(12,95)		
Konst.	1,002	0,977			0,978				0,978		
(t-Werte)	(31,69)	(27,14)			(26,94)				(28,59)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.19: MC-Simulationen – Lineares Modell mit logarithmierten Variablen, multiplikative Überlagerung nach dem Varfahren von Höhne, 1.000 Replikationen

	Original	Multiplikative Überlagerung nach dem Verfahren von Höhne							
		nicht korrigiert			korrigiert (IV)			korrigiert (SIMEX, Varianz geschätzt)	
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem
			Original-wert	theor. Wert		Original-wert	theor. Wert		Original-wert	theor. Wert
log X_1	1,000	1,002	*		1,000	*	*	1,003		
(t-Werte)	(28,53)	(28,54)			(28,35)			(29,35)		
log X_2	-0,998	-0,994			-1,000	*	*	-0,994		
(t-Werte)	(-30,96)	(-30,81)			(-30,70)			(-31,43)		
log X_3	0,499	0,504			0,501	*	*	0,504		
(t-Werte)	(14,46)	(14,59)			(14,38)			(14,96)		
Konst.	1,002	0,996			0,996			0,997		
(t-Werte)	(31,69)	(31,37)			(31,33)			(32,21)		

* Übereinstimmung zum Signifikanzniveau von 5%

22.2.2 Stochastische Überlagerungen in nichtlinearen Modellen

Als Beispiel für ein nichtlineares Modell wird ein Probit-Modell mit einer binären abhängigen Variablen geschätzt. Die drei Regressoren sind normalverteilt mit Erwartungswert Null und Varianz-Kovarianzmatrix

$$\mathbf{C} = \begin{pmatrix} 1 & 0,2 & 0,4 \\ 0,2 & 1 & 0,1 \\ 0,4 & 0,1 & 1 \end{pmatrix}.$$

Für die latente Variable Y^* gelte

$$Y^* = 1 + X_1 - X_2 + 0,5X_3 + U \tag{22.180}$$

Beobachtbar ist jedoch nur die binäre Variable Y, die den Wert 1 annimmt, falls Y^* größer Null ist und sonst den Wert Null.

Das Modell wird auf der Basis von Monte-Carlo-Simulationen mit 1.000 Beobachtungen und 1.000 Replikationen geschätzt. Sowohl für die Originalschätzung als auch für die einzelnen Anonymisierungsverfahren werden getrennte Simulationsexperimente durchgeführt, so dass auch die Originaldaten für jedes Anonymisierungsverfahren immer neu erzeugt werden.

Getestet werden sowohl additive als auch multiplikative stochastische Überlagerungen. In einem Fall werden alle drei Einflussfvariablen überlagert, im anderen lediglich die Regressoren X_2 und X_3. Die Ergebnisse für den ersten Fall sind in den Tabellen 22.20 bis 22.22 dargestellt, diejenigen für den zweiten Fall in den Tabellen 22.23 bis 22.25. Allerdings gibt es keine wesentlichen Unterschiede zwischen diesen beiden Fällen.

Tabelle 22.20: MC-Simulationen – Probit-Modell, additive stochastische Überlagerung (Normalverteilung), alle Regressoren überlagert, 1.000 Replikationen

	Original	Einfache additive Überlagerung mit gleicher Zufallszahl, Normalverteilung (Varianz 1/3), unkorrigiert			Einfache additive Überlagerung mit unterschiedlichen Zufallszahlen, Normalverteilung, (Varianz 1/3), unkorrigiert			Einfache additive Überlagerung mit unterschiedlichen Zufallszahlen, Normalverteilung, (Varianz 1/3), SIMEX-Korrektur, Angabe der Varianz der Überlagerung,			Einfache additive Überlagerung mit unterschiedlichen Zufallszahlen, Normalverteilung, (Varianz 1/3), SIMEX-Korrektur, Varianz der Überlagerung geschätzt		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original	theor. Wert		Original	theor. Wert		Original	theor. Wert		Original	theor. Wert
X_1	1,009	0,933			0,587			0,822			0,937		
(t-Werte)	(12,88)	(12,59)			(11,26)			(10,13)			(10,76)		
X_2	-1,008	-1,066			-0,569			-0,815			-0,932		
(t-Werte)	(-13,51)	(-14,57)			(-11,37)			(-10,58)			(-11,28)		
X_3	0,504	0,426			0,349			0,459			0,485		
(t-Werte)	(7,77)	(7,04)			(7,35)			(6,56)			(6,73)		
Konst.	1,005	0,985			0,807			0,915			0,970		
(t-Werte)	(14,91)	(14,93)			(14,73)			(13,30)			(13,58)		

Tabelle 22.21: MC-Simulationen – Probit-Modell, multiplikative stochastische Überlagerung (Gleichverteilung), alle Regressoren überlagert, 1.000 Replikationen

	Original	Multiplikative Überlagerung mit konstantem Faktor, Gleichverteilung (0,5;1,5), unkorrigiert			Multiplikative Überlagerung mit konstantem Faktor, Gleichverteilung (0,5;1,5), SIMEX-Korrektur, Varianz der Überlagerung geschätzt			Multiplikative Überlagerung mit unterschiedlichen Faktoren, Gleichverteilung (0,5;1,5), unkorrigiert			Multiplikative Überlagerung mit unterschiedlichen Faktoren, Gleichverteilung (0,5;1,5), SIMEX-Korrektur, Varianz der Überlagerung geschätzt,		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original	theor. Wert		Original	theor. Wert		Original	theor. Wert		Original	theor. Wert
X_1	1,009	0,941			1,067			0,880			1,033		
(t-Werte)	(12,88)	(12,59)			(11,65)			(12,40)			(11,63)		
X_2	-1,008	-0,947			-1,064			-0,868			-1,030		
(t-Werte)	(-13,51)	(-13,26)			(-12,16)			(-12,93)			(-12,29)		
X_3	0,504	0,470			0,531			0,462			0,511		
(t-Werte)	(7,77)	(7,53)			(7,31)			(7,70)			(7,32)		
Konst.	1,005	0,954			1,027			0,938			1,014		
(t-Werte)	(14,91)	(15,00)			(14,65)			(14,96)			(14,45)		

Tabelle 22.22: MC-Simulationen – Probit-Modell, multiplikative stochastische Überlagerung (Mischungsverteilung nach dem Verfahren von Höhne), alle Regressoren überlagert, 1.000 Replikationen

	Original	Multiplikative Überlagerung nach dem Verfahren von Höhne unkorrigiert			Multiplikative Überlagerung nach dem Verfahren von Höhne, SIMEX-Korrektur, Varianz geschätzt		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original	theor. Wert		Original	theor. Wert
X_1	1,009	0,996		*	0,999		*
(t-Werte)	(12,88)	(12,82)			(12,94)		
X_2	-1,008	-0,995			-0,997		*
(t-Werte)	(-13,51)	(-13,45)			(-13,61)		
X_3	0,504	0,497		*	0,498		*
(t-Werte)	(7,77)	(7,72)			(7,82)		
Konst.	1,005	0,999		*	1,000	*	*
(t-Werte)	(14,91)	(14,94)			(15,40)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.23: MC-Simulationen – Probit-Modell, additive stochastische Überlagerung (NV), Teil der Regressoren überlagert, 1.000 Replikationen

	Original	Einfache additive Überlagerung mit gleicher Zufallszahl, Normalverteilung (Varianz 1/3), unkorrigiert			Einfache additive Überlagerung mit unterschiedlichen Zufallszahlen, Normalverteilung (Varianz 1/3), unkorrigiert			Einfache additive Überlagerung mit unterschiedlichen Zufallszahlen, Normalverteilung (Varianz 1/3), SIMEX-Korrektur, Angabe der Varianz der Überlagerung			Einfache additive Überlagerung mit unterschiedlichen Zufallszahlen, Normalverteilung (Varianz 1/3), SIMEX-Korrektur, Varianz der Überlagerung geschätzt		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem		
			Original	theor. Wert		Original	theor. Wert		Original	theor. Wert		Original	theor. Wert	
X_1	1,009	0,924			0,891			0,962			0,992			
(t-Werte)	(12,88)	(12,61)			(12,89)			(11,82)			(12,11)			
X_2	-1,008	-0,888			-0,652			-0,886			-0,968			
(t-Werte)	(-13,51)	(-13,49)			(-12,00)			(-10,95)			(-11,60)			
X_3	0,504	0,604			0,312			0,433			0,482			
(t-Werte)	(7,77)	(10,06)			(6,26)			(6,07)			(6,66)			
Konst.	1,005	0,984			0,881			0,956			0,988			
(t-Werte)	(14,91)	(14,92)			(14,86)			(13,73)			(14,05)			

Tabelle 22.24: MC-Simulationen – Probit-Modell, multiplikative stochastische Überlagerung (Gleichverteilung), Teil der Regressoren überlagert, 1.000 Replikationen

	Original	Multiplikative Überlagerung mit konstantem Faktor, Gleichverteilung (0,5;1,5) unkorrigiert			Multiplikative Überlagerung mit konstantem Faktor, Gleichverteilung (0,5;1,5), SIMEX-Korrektur, Varianz der Überlagerung geschätzt			Multiplikative Überlagerung mit unterschiedlichen Faktoren, Gleichverteilung (0,5;1,5) unkorrigiert			Multiplikative Überlagerung mit unterschiedlichen Faktoren, Gleichverteilung (0,5;1,5) SIMEX-Korrektur, Varianz der Überlagerung geschätzt		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original	theor. Wert		Original	theor. Wert		Original	theor. Wert		Original	theor. Wert
X_1	1,009	0,976			1,014	*		0,973			1,011	*	
(t-Werte)	(12,88)	(12,87)			(12,75)			(12,89)			(12,72)		
X_2	-1,008	-0,929			-1,031			-0,908			-1,024		
(t-Werte)	(-13,51)	(-13,17)			(-12,56)			(-13,08)			(-12,61)		
X_3	0,504	0,472			0,518			0,444			0,508	*	
(t-Werte)	(7,77)	(7,59)			(7,42)			(7,29)			(7,38)		
Konst.	1,005	0,973			1,015			0,969			1,009	*	
(t-Werte)	(14,91)	(14,93)			(14,93)			(14,95)			(14,83)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 22.25: MC-Simulationen – Probit-Modell, multiplikative stochastische Überlagerung (Mischungsverteilung nach dem Verfahren von Höhne), zwei der Regressoren überlagert, 1.000 Replikationen

	Original	Multiplikative Überlagerung nach dem Verfahren von Höhne, unkorrigiert			Multiplikative Überlagerung nach dem Verfahren von Höhne, SIMEX-Korrektur, Varianz geschätzt		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original	theor. Wert		Original	theor. Wert
X_1	1,009	0,999		*	1,005	*	
(t-Werte)	(12,88)	(12,84)			(13,16)		
X_2	-1,008	-0,994			-0,999		*
(t-Werte)	(-13,51)	(-13,43)			(-13,67)		
X_3	0,504	0,499	*	*	0,499	*	*
(t-Werte)	(7,77)	(7,76)			(7,89)		
Konst.	1,005	0,998		*	1,003	*	*
(t-Werte)	(14,91)	(14,92)			(15,33)		

* Übereinstimmung zum Signifikanzniveau von 5%

Als Korrekturverfahren wird der SIMEX-Schätzer eingesetzt. Dabei wird eine quadratische Extrapolationsfunktion verwendet. Die Anzahl der Schätzungen auf einer Stufe beträgt $b = 50$. Die Varianzschätzung erfolgt mittels eines Bootstrap-Schätzers.

Bei der SIMEX-Korrektur kommt auch für multiplikative Überlagerungen der eigentlich für additive Fehler programmierte SIMEX-Schätzer zum Einsatz. Dabei wird auf die Überlegungen in Unterabschnitt 22.1.2 zurückgegriffen.

In den Tabelle 22.20 und 22.23 ist zunächst zu erkennen, dass die additive Überlagerung mit unterschiedlichen Zufallszahlen bei gleicher Varianz der Überlagerungen (hier: 1/3) zu stärkeren Verzerrungen führt als die Überlagerung aller Merkmalswerte einer Einheit mit der gleichen Zufallszahl. Allerdings ist die SIMEX-Korrektur bei einer einheitlichen Zufallszahl nicht möglich, so dass korrigierte Ergebnisse nur für den Fall unterschiedlicher Zufallszahlen zur Verfügung stehen. Dabei kann alternativ die Varianz der Überlagerungen vorgegeben oder geschätzt werden. Die bessere Korrektur ergibt sich im zweiten Fall, wobei auch hier eine recht starke Abweichung von den Orginalschätzern verbleibt.

Auch bei der multiplikativen Überlagerung mit Fehlern aus einer Gleichverteilung (Tabellen 22.21 und 22.24) ergeben sich bei konstanter Varianz stärkere Abweichungen bei einer Überlagerung mit unterschiedlichen Faktoren. Hier wird ausschließlich eine SIMEX-Korrektur eingesetzt, bei der die Varianz der Fehler durch die Zugabe von Instrumenten geschätzt wird. Dabei gelingt eine bessere Anpassung an die Originalschätzer als bei der additiven Überlagerung. Allerdings weichen auch hier die korrigierten Schätzer nur in Einzelfällen nicht signifikant von den Originalschätzern ab.

Nur sehr geringe Abweichungen der Schätzer gegenüber dem Original werden durch die multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne (Tabellen 22.22 und 22.25) hervorgerufen, bei dem die Überlagerungen allerdings im Vergleich zur verwendeten Gleichverteilung eine weitaus geringere Varianz aufweisen. Hier weist bereits ein Teil der Schätzer ohne Korrektur keine signaifikante Abweichung zum Originalschätzer auf. Diese Ergebnisse werden durch die SIMEX-Korrektur noch leicht verbessert.

22.3 Praxisbeispiele

22.3.1 Stochastische Überlagerungen in linearen Modellen

a) Das geschätzte Modell

Zur Untersuchung der Auswirkungen von stochastischen Überlagerungen in linearen Modellen wird als Praxisbeispiel die in Abschnitt 21.1 beschriebene linearisierte Cobb-Douglas-Produktionsfunktion analog zu Fritsch und Stephan (2003) geschätzt. Für die Untersu-

chung der stochastischen Überlagerung wird vor der Analyse der Wirtschaftszweig 37 (Recycling) entfernt; zudem wird zunächst auf die Ausreißerbereinigung verzichtet.

Im Folgenden wird zunächst der Fall betrachtet, dass die in das Modell eingehenden transformierten Variablen direkt mittels stochastischer Überlagerungen anonymisiert werden. Anschließend wird die Schätzung durchgeführt, nachdem die Inputfaktoren sowie der Output überlagert werden. Zuletzt wird der in der Realität bei Scientific-Use-Files realistischere Fall untersucht, bei dem die Ausgangsvariablen anonymisiert und anschließend für die Modellschätzung transformiert werden.

b) Die Schätzung einer linearisierten Cobb-Douglas-Produktionsfunktion mit stochastisch überlagerten logarithmierten Variablen

Im ersten Schritt werden die transformierten Inputfaktoren und der transformierte Output, wie sie in die Regressionsgleichung eingehen, direkt stochastisch überlagert. Damit erhält man anhand der Projektdaten Erkenntnisse darüber, wie die unterschiedlichen Varianten der stochastischen Überlagerung in der Praxis in linearen Modellen wirken.

Dabei werden folgende Varianten der additiven Überlagerung getestet:

- Additive Überlagerung der sechs Variablen mit voneinander unabhängigen standardnormalverteilten Zufallsfehlern;
- Additive Überlagerung der sechs Variablen mit Zufallsfehlern, deren Varianz-Kovarianzmatrix proportional (Faktor $d = 0,1$) zur Varianz-Kovarianzmatrix der Originalvariablen ist;
- Additive Überlagerung nach dem Verfahren von Kim (dabei werden die überlagerten Werte einer korrigierenden Transformation unterzogen, die zum Erhalt der ersten beiden Momente führt).

Im Bereich der multiplikativen stochastischen Überlagerung werden folgende Varianten in die Untersuchung einbezogen:

- Multiplikative Überlagerung der sechs Variablen mit einem für jedes Unternehmen konstanten Faktor, dabei entstammen die Faktoren einer Gleichverteilung mit dem Intervall (0,5;1,5);
- Multiplikative Überlagerung der sechs Variablen mit unterschiedlichen Faktoren, die ebenfalls einer Gleichverteilung mit dem Intervall (0,5;1,5) entstammen;

- Multiplikative Überlagerung der sechs Variablen mit einem für jedes Unternehmen konstanten Faktor, dabei entstammen die Faktoren einer Gleichverteilung mit dem Intervall (0,8;1,2);
- Multiplikative Überlagerung der sechs Variablen mit unterschiedlichen Faktoren, die ebenfalls einer Gleichverteilung mit dem Intervall (0,8;1,2) entstammen;
- Multiplikative Überlagerung der sechs Variablen mit Faktoren aus einer zweigipfligen Mischungsverteilung nach dem Verfahren von Höhne. Dabei wird zunächst mit einer Wahrscheinlichkeit von 0,5 festgelegt, ob die Werte für ein Unternehmen vergrößert oder verkleinert werden. Als „Grundfaktoren" werden $0,89$ und $1,11$ ($f = 0,11$) gewählt. Diese werden zusätzlich jeweils additiv mit einem Zufallsfehler aus einer Normalverteilung mit Mittelwert Null und Standardabweichung $s = 0,03$ überlagert. Anschließend werden die einzelnen Merkmalswerte mit den so entstandenen Faktoren multiplikativ überlagert.

Die Untersuchung erfolgt im Rahmen von Monte-Carlo-Simulationen mit 1.000 Replikationen, d.h. die stochastische Überlagerung und die anschließende Modellschätzung wird für jede Anonymisierungsvariante 1.000 mal wiederholt. Die Ergebnisse sind in den Tabellen 22.26 bis 22.29 dargestellt.

Für die additiven Überlagerungen bestätigen sich die Ergebnisse der Simulationsexperimente aus Unterabschnitt 22.2.1. Die additive Überlagerung mit standardnormalverteilten Zufallsfehlern führt zu verzerrten Schätzern, die sich beispielsweise durch eine Instrumentvariablen-Schätzung wieder korrigieren lassen, allerdings sind die Teststatistiken verzerrt. Die additive Überlagerung mit Zufallsfehlern, deren Varianz-Kovarianzmatrix proportional zur Varianz-Kovarianzmatrix der Originalvariablen gewählt wird und das Kim-Verfahren erhalten sowohl die Koeffizientenschätzer als auch die Teststatistiken (vgl. Tabelle 22.26).

Bei der multiplikativen Überlagerung bestätigt sich zunächst die in Unterabschnitt 22.1.1 getätigte Herleitung, dass bei konstanten Überlagerungsfaktoren für alle Merkmale eines Merkmalsträgers die Schätzer im linearen Modell nur dann konsistent sind, wenn kein Absolutglied vorhanden ist oder alle Regressoren einen Mittelwert von Null aufweisen. Während das Zutreffen der zweiten Bedingung bei den Simulationsexperimenten in 22.2.1 zum Erhalt der Originalschätzer geführt hat, trifft im vorliegenden Praxisbeispiel keine der beiden Bedingungen zu. Somit ist es folgerichtig, dass bei der Schätzung der linearisierten Cobb-Douglas-Produktionsfunktion auch die multiplikative Überlagerung aller Merkmale eines Unternehmens mit einem konstanten Faktor zu verzerrten Schätzern führt. Dabei ist es unerheblich, ob die Überlagerungsfaktoren einer Gleichverteilung im Intervall (0,5;1,5) entstammen (vgl. Tabelle 22.27) oder einer Gleichverteilung im Intervall (0,8;1,2) (vgl. Tabelle 22.28). Ohnehin in beiden Fällen verzerrt sind die Schätzer bei einer Überlagerung mit unterschiedlichen Faktoren.

Will man die entstandenen Verzerrungen mittels eines Instrumentvariablen-Schätzers korrigieren, so treten ebenfalls Probleme auf:

- Die durch multiplikative Überlagerung mit konstantem gleichverteilten Fehler aus dem Intervall (0,5;1,5) verursachte Verzerrung lässt sich zwar im Mittel der 1.000 Replikationen korrigieren, allerdings treten in Einzelfällen recht hohe Abweichungen vom Originalschätzer auf. Dieses Problem reduziert sich deutlich, wenn die Varianz der Überlagerungen dadurch vermindert wird, dass die Überlagerungsfaktoren aus dem Intervall (0,8;1,2) gezogen werden.
- Sind die Faktoren nicht für alle Merkmalswerte einer Einheit konstant, so lässt sich bei Fehlern aus dem Intervall (0,5;1,5) eine funktionierende Korrektur mittels IV-Schätzung gar nicht gewährleisten. Bei einer Beschränkung des Intervalls der Überlagerungsfaktoren auf (0,8;1,2) gelingt zwar im Durchschnitt über alle 1.000 Replikationen die Fehlerkorrektur ganz gut, nicht jedoch für viele der Simulationsläufe. Vielmehr weisen auch hier die Schätzer häufig eine beträchtliche Abweichung vom Originalwert und über die 1.000 Replikationen eine hohe Varianz auf.

Damit erweist sich eine multiplikative Überlagerung durch Fehler aus einer Gleichverteilung mit den hier gewählten Intervallen für eine Anonymisierung eher als unbrauchbar. Zu besseren Ergebnissen führt hingegen die multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne (vgl. Tabelle 22.29, wobei die Überlagerungsfaktoren dabei eine weitaus geringere Varianz aufweisen).

Insgesamt sind die Auswirkungen von multiplikativen Überlagerungen auf die Schätzergebnisse stark vom Startwert des Zufallsgenerators bei der Erzeugung der Zufallsfehler abhängig. Dies gilt auch für das Verfahren von Höhne. Deshalb sollte auch bei Anwendung dieses Verfahrens durch Beispielrechnungen sichergestellt werden, dass eine aus Analysesicht günstige Konstellation gewählt wurde.

Tabelle 22.26: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für additiv überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz bereinigt, 1.000 Replikationen, Logarithmen überlagert, robuste Standardfehler

	Additive Überlagerung (standardnormalverteilt)		Additive Überlagerung (standardnormalverteilt), Korrektur IV		Additve Überlagerung VCV-Matrix der Überlagerungen proportional zur VCV-Matrix der Originalvariablen		Kim-Verfahren	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Materialeinsatz	0,268	(41,50)	0,415	(20,77)	0,415	(84,97)	0,415	(84,96)
Personalkosten	0,177	(23,78)	0,339	(8,27)	0,339	(62,24)	0,339	(62,25)
Externe Dienstleistungen	0,116	(21,88)	0,058	(5,36)	0,058	(30,30)	0,058	(30,29)
Sonstige Kosten	0,178	(27,71)	0,113	(5,34)	0,113	(36,58)	0,113	(36,57)
Kapitalkosten	0,136	(20,05)	0,055	(2,25)	0,055	(16,72)	0,055	(16,72)
Konst.	3,887	(43,10)	1,803	(9,12)	1,804	(67,82)	1,804	(67,86)
R^2	0,607		0,574		0,977		0,977	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	35,35	47,19	0,03	73,57	0,09	8,12	0,10	8,11
Personalkosten	47,83	59,43	0,13	85,89	0,10	6,19	0,11	6,21
Externe Dienstleistungen	99,35	24,60	0,01	81,53	0,00	4,41	0,01	4,38
Sonstige Kosten	57,25	20,60	0,33	84,70	0,34	4,81	0,31	4,79
Kapitalkosten	146,63	25,23	0,71	85,95	0,06	4,43	0,07	4,43
Konst.	115,45	34,88	0,04	86,22	0,01	2,46	0,01	2,52
Durchschn.	83,64	35,32	0,21	82,98	0,10	5,07	0,10	5,07

Tabelle 22.27: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für multiplikativ (Gleichvert.(0,5;1,5)) überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz bereinigt, 1.000 Replikationen, Logarithmen überlagert, robuste Standardfehler

	Multiplikative Überlagerung (Konstanter Faktor, Gleichverteilung (0,5;1,5))		Multiplikative Überlagerung (Konstanter Faktor, Gleichverteilung (0,5;1,5)), Korrektur IV		Multiplikative Überlagerung (Unterschiedliche Faktoren, Gleichverteilung (0,5;1,5))		Multiplikative Überlagerung (Unterschiedliche Faktoren, Gleichverteilung (0,5;1,5)), Korrektur IV	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Materialeinsatz	0,428	(66,57)	0,414	(48,40)	0,070	(8,51)	0,676	(0,62)
Personalkosten	0,502	(82,57)	0,335	(14,18)	0,054	(6,30)	-0,756	(0,24)
Externe Dienstleistungen	0,029	(12,30)	0,059	(10,16)	0,089	(9,49)	0,165	(0,28)
Sonstige Kosten	0,074	(19,41)	0,115	(13,28)	0,081	(8,91)	0,339	(0,28)
Kapitalkosten	0,043	(10,19)	0,056	(7,40)	0,072	(7,62)	0,332	(0,16)
Konst.	0,129	(19,00)	1,844	(7,86)	11,635	(47,37)	6,275	(0,50)
R^2	0,997		0,987		0,031			
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	3,24	15,29	0,19	38,41	83,04	89,17	62,99	99,21
Personalkosten	48,02	40,88	1,20	75,81	84,02	89,25	322,94	99,59
Externe Dienstleistungen	49,39	57,62	1,28	64,99	54,05	67,30	183,97	99,04
Sonstige Kosten	34,23	44,38	1,43	61,95	28,22	74,47	200,40	99,20
Kapitalkosten	20,93	36,35	1,08	53,78	30,82	52,40	503,65	99,00
Konst.	92,86	71,29	2,24	88,13	544,94	28,43	247,86	99,24
Durchschn.	41,44	44,30	1,24	63,84	137,51	66,84	253,63	99,21

Tabelle 22.28: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für multiplikativ (Gleichvert.(0,8;1,2)) überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz bereinigt, 1.000 Replikationen, Logarithmen überlagert, robuste Standardfehler

	Multiplikative Überlagerung (Konstanter Faktor, Gleichverteilung (0,8;1,2))		Multiplikative Überlagerung (Konstanter Faktor, Gleichverteilung (0,8;1,2)), Korrektur IV		Multiplikative Überlagerung (Unterschiedliche Faktoren, Gleichverteilung (0,8;1,2))		Multiplikative Überlagerung (Unterschiedliche Faktoren, Gleichverteilung (0,8;1,2)), Korrektur IV	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Materialeinsatz	0,425	(73,92)	0,415	(71,04)	0,171	(22,43)	0,415	(6,71)
Personalkosten	0,460	(81,19)	0,339	(41,44)	0,121	(14,47)	0,340	(2,50)
Externe Dienstleistungen	0,037	(17,16)	0,058	(22,40)	0,146	(18,93)	0,057	(1,99)
Sonstige Kosten	0,084	(24,48)	0,114	(27,59)	0,168	(20,41)	0,115	(1,83)
Kapitalkosten	0,046	(12,20)	0,055	(13,01)	0,142	(16,21)	0,054	(0,77)
Konst.	0,559	(37,52)	1,807	(30,80)	5,964	(39,32)	1,793	(2,98)
R^2	0,989		0,984		0,274		0,167	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	2,43	5,94	0,10	9,61	58,73	71,46	0,04	91,46
Personalkosten	35,67	38,53	0,03	29,30	64,16	75,31	0,20	95,73
Externe Dienstleistungen	36,63	40,87	0,18	22,81	151,67	34,77	1,77	93,14
Sonstige Kosten	25,39	29,86	0,49	20,95	48,86	41,52	1,80	94,76
Kapitalkosten	15,66	23,80	0,19	18,74	157,56	1,25	0,93	95,19
Konst.	68,99	43,31	0,15	53,47	230,58	40,60	0,63	95,50
Durchschn.	30,79	30,38	0,19	25,81	118,59	44,15	0,89	94,30

Tabelle 22.29: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für multiplikativ (Verfahren von Höhne) überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz bereinigt, 1.000 Replikationen, Logarithmen überlagert, robuste Standardfehler

	Multiplikative Überlagerung (Mischungsverteilung nach dem Verfahren von Höhne)		Multiplikative Überlagerung (Mischungsverteilung nach dem Verfahren von Höhne), Korrektur IV	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Materialeinsatz	0,370	(61,31)	0,415	(45,62)
Personalkosten	0,347	(51,48)	0,338	(19,51)
Externe Dienstleistungen	0,062	(15,94)	0,058	(11,52)
Sonstige Kosten	0,144	(26,13)	0,114	(13,26)
Kapitalkosten	0,120	(19,77)	0,055	(6,02)
Konst.	1,001	(27,88)	1,809	(14,33)
R^2	0,933		0,929	
	Relative Abweichungen von den Originalwerten in %			
Materialeinsatz	10,92	21,99	0,10	41,95
Personalkosten	2,27	12,17	0,19	66,71
Externe Dienstleistungen	6,80	45,07	0,31	60,30
Sonstige Kosten	27,37	25,13	0,63	62,01
Kapitalkosten	118,63	23,49	0,47	62,40
Konst.	44,49	57,88	0,29	78,35
Durchschn.	35,08	30,95	0,33	61,95

c) Die Schätzung einer linearisierten Cobb-Douglas-Produktionsfunktion mit stochastisch überlagerten Outputwerten und Inputfaktoren

Nun wird untersucht, wie sich die nichtlineare Transformation des Logarithmus einer überlagerten Variablen auf die Schätzung der linearisierten Cobb-Douglas-Produktionsfunktion auswirkt. Hierzu werden anstatt der logarithmierten Inputfaktoren und des logarithmierten Outputs die Inputfaktoren und der Output vor der Logarithmierung stochastisch überlagert. Allerdings werden vorher diejenigen Unternehmen ausgeschlossen, bei denen ein Inputfaktor oder der Output den Wert Null aufweist, damit die Verzerrungen in den Schätzergebnissen ausschließlich auf die stochastische Überlagerung und die nichtlineare Transformation zurückzuführen sind. Dabei werden dieselben Anonymisierungsverfahren untersucht, die im vorangegangenen Schritt auf die logarithmierten Inputfaktoren beziehungsweise den logarithmierten Output angewendet wurden. Die Ergebnisse sind in den Tabellen 22.30 bis 22.34 dargestellt.

Wesentlich zur Erklärung der Ergebnisse ist, dass im hier betrachteten Fall der linearisierten Cobb-Douglas-Produktionsfunktion mit ungefähr konstanten Skalenerträgen die multiplikative stochastische Überlagerung mit einem (für das Unternehmen) konstanten Faktor zu unveränderten Schätzern führt. Dies kann wie folgt hergeleitet werden. Für die Cobb-Douglas-Produktionsfunktion gilt allgemein:

$$Y = A \prod_{k=1}^{K} X_k^{\beta_k} \tag{22.181}$$

und im anonymisierten Fall

$$Y^a = A \prod_{k=1}^{K} X_k^{a^{\beta_k}} \tag{22.182}$$

und im Fall der multiplikativen Überlagerung mit einem konstanten Faktor

$$WY = A \prod_{k=1}^{K} W X_k^{\beta_k}. \tag{22.183}$$

Wird nun die Funktion durch Logarithmierung auf beiden Seiten linearisiert, so ergibt sich:

$$log(Y) + log(W) = C + \sum_{k=1}^{K} (log(X_k \beta_k) + (log(W \beta_k)). \tag{22.184}$$

Für den Fall konstanter Skalenerträge gilt aber $\sum_{k=1}^{K}(log(W\beta_k)) = log(W)$, somit kürzt sich $log(W)$ auf beiden Seiten der Gleichung (22.184) weg und es ergibt sich die Gleichung, die auch für Originalwerte Gültigkeit besitzt.

Wird eine multiplikative stochastische Überlagerung mit unterschiedlichen Faktoren gewählt, so werden durch die Logarithmierung auf beiden Seiten der Regressionsgleichung die multiplikativen Fehler mit Erwartungswert Eins in additive Fehler umgewandelt:

$$log(Y) + log(W_y) = C + \sum_{k=1}^{K}\beta_k(log(X_k) + log(W_k)). \tag{22.185}$$

Somit lassen sich die klassischen Korrekturschätzer für additive Fehler einsetzen, sofern der Erwartungswert des additiven Fehlers Null ist. Diese Bedingung ist für die Gleichverteilung nicht erfüllt, für die Mischungsverteilung nach dem Verfahren von Höhne gilt dies aber ungefähr. Insofern kann bei diesem Verfahren auch mit Instrumentvariablen-Schätzungen korrigiert werden.

Bei additiven Überlagerungen besteht, wie bereits bei den Simulationsexperimenten gezeigt wurde, im Gegensatz zur multiplikativen Überlagerung das Problem, dass es durch die Überlagerung zu Vorzeichenwechseln kommen kann. Da die überlagerten Inputfaktoren sowie der Ouptut anschließend logarithmiert werden, können im anonymisierten Datensatz zusätzliche „Missing Values“ auftreten. Hierdurch hervorgerufene Verzerrungen der Koeffizientenschätzer können nicht korrigiert werden, was in den Tabellen 22.30 und 22.31 illustriert ist. Der SIMEX-Schätzer führt in diesem Fall sogar zu einer Erhöhung der Verzerrung, da durch die Veränderung der Zusammensetzung des Datensatzes die Richtung der eigentlichen Verzerrung durch die Überlagerung nicht mehr erkennbar ist.

Wie bereits erwähnt, führt die multiplikative Überlagerung aller Merkmalswerte einer Einheit mit einem konstanten Faktor in diesem Fall zu einem nahezu unverzerrten Schätzer (jeweils erste Spalte der Tabellen 22.32 und 22.33). Verwendet man jedoch unterschiedliche Faktoren, so ergibt sich eine Verzerrung, die mit steigender Varianz der Überlagerungen zunimmt.

Die Überlagerung mit einer Mischungverteilung nach dem Höhne-Verfahren (Tabelle 22.34) verursacht nicht nur wegen der relativ geringen Varianz eine fast vernachlässigbare Verzerrung, sondern insbesondere deshalb, weil aufgrund der besonderen Vorgehensweise alle Merkmalswerte eines Unternehmens mit einem annähernd gleichen Faktor überlagert werden.

Alternativ werden für die multiplikativen Überlagerungen eine Korrektur mit Instrumentvariablen und SIMEX-Schätzer getestet, wobei für die SIMEX-Korrektur die Varianz der Überlagerungen ebenfalls mit Hilfe von Instrumenten geschätzt wird. Dabei kann man feststellen, dass der IV-Schätzer die besseren Ergebnisse liefert. Dies entspricht auch den Er-

wartungen, da es sich beim IV-Schätzer um eine exakte Korrektur, beim SIMEX-Schätzer hingegen nur um ein approximatives Verfahren handelt.

Tabelle 22.30: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für additiv (NV) überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz unbereinigt, 1.000 Replikationen, Inputfaktoren und Output überlagert, robuste Standardfehler

	Additve Überlagerung VCV-Matrix der Überlagerungen proportional zur VCV-Matrix der Originalvariablen		Additve Überlagerung VCV-Matrix der Überlagerungen proportional zur VCV-Matrix der Originalvariablen, Korrektur IV		Additve Überlagerung VCV-Matrix der Überlagerungen proportional zur VCV-Matrix der Originalvariablen, SIMEX-Korrektur, Varianz geschätzt	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Materialeinsatz	0,569	(31,23)	0,561	(11,77)	0,951	(19,41)
Personalkosten	0,150	(11,14)	0,172	(3,24)	0,208	(4,78)
Externe Dienstleistungen	0,041	(8,71)	0,089	(2,47)	0,023	(1,19)
Sonstige Kosten	0,137	(15,86)	0,148	(3,35)	0,169	(5,71)
Kapitalkosten	0,038	(3,96)	0,060	(1,08)	0,052	(1,49)
Konst.	2,448	(13,56)	0,859	(2,39)	-6,526	(-11,25)
R^2	0,880		0,896			
	Relative Abweichungen von den Originalwerten in %					
Materialeinsatz	37,21	60,26	35,10	85,02	129,23	75,30
Personalkosten	55,87	80,99	49,30	94,47	38,60	91,84
Externe Dienstleistungen	29,02	69,99	53,06	91,49	60,26	95,90
Sonstige Kosten	20,89	54,56	30,78	90,40	49,56	83,64
Kapitalkosten	31,49	75,27	8,34	93,25	6,06	90,69
Konst.	35,68	79,51	52,40	96,39	461,75	117,00
Durchschn.	35,03	70,10	38,17	91,84	124,24	92,40

Tabelle 22.31: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für mit dem Kim-Verfahren additiv (NV) überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz unbereinigt, 1.000 Replikationen, Inputfaktoren und Output überlagert, robuste Standardfehler

	Kim-Verfahren		Kim-Verfahren, Korrektur IV		Kim-Verfahren, SIMEX-Korrektur, Varianz geschätzt	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Materialeinsatz	0,569	(31,60)	0,558	(11,80)	0,952	(19,44)
Personalkosten	0,149	(11,15)	0,174	(3,27)	0,208	(4,80)
Externe Dienstleistungen	0,041	(8,77)	0,089	(2,48)	0,023	(1,17)
Sonstige Kosten	0,137	(15,92)	0,146	(3,43)	0,169	(5,69)
Kapitalkosten	0,038	(4,00)	0,061	(1,09)	0,053	(1,54)
Konst.	2,463	(13,89)	0,869	(2,42)	-6,535	(-11,31)
R^2	0,880		0,895			
	Relative Abweichungen von den Originalwerten in %					
Materialeinsatz	37,05	59,79	34,54	84,99	129,38	75,26
Personalkosten	56,02	80,98	48,66	94,42	38,55	91,81
Externe Dienstleistungen	28,93	69,78	53,95	91,45	60,45	95,97
Sonstige Kosten	21,06	54,38	29,22	90,17	49,37	83,70
Kapitalkosten	31,52	75,02	10,38	93,19	4,13	90,38
Konst.	36,55	79,01	51,84	96,34	462,27	117,09
Durchschn.	35,19	69,83	38,10	91,76	124,02	92,37

Tabelle 22.32: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für multiplikativ (Gleichvert. (0,5;1,5)) überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz unbereinigt, 1.000 Replikationen, Inputfaktoren und Output überlagert, robuste Standardfehler

	Multiplikative Überlagerung konstante Faktoren, Gleichverteilung (0,5;1,5),		Multiplikative Überlagerung unterschiedliche Faktoren, Gleichverteilung (0,5;1,5),		Multiplikative Überlagerung unterschiedliche Faktoren, Gleichverteilung (0,5;1,5), Korrektur IV		Multiplikative Überlagerung unterschiedliche Faktoren, Gleichverteilung (0,5;1,5), SIMEX-Korrektur, Varianz geschätzt	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Materialeinsatz	0,415	(78,56)	0,386	(72,46)	0,415	(63,87)	0,414	(67,46)
Personalkosten	0,341	(59,14)	0,283	(45,67)	0,339	(38,41)	0,335	(42,73)
Externe Dienstleistungen	0,058	(28,95)	0,070	(24,88)	0,058	(18,51)	0,058	(19,78)
Sonstige Kosten	0,113	(34,86)	0,139	(33,09)	0,114	(22,01)	0,115	(24,15)
Kapitalkosten	0,055	(16,05)	0,085	(18,68)	0,055	(9,68)	0,057	(10,95)
Konst.	1,786	(67,48)	2,213	(53,76)	1,803	(37,37)	1,821	(41,29)
R^2	0,978		0,925		0,924			
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	0,12	0,04	7,07	7,80	0,10	18,73	0,17	14,16
Personalkosten	0,52	0,90	16,60	22,08	0,07	34,47	1,09	27,09
Externe Dienstleistungen	0,33	0,24	20,82	14,27	0,03	36,22	0,59	31,84
Sonstige Kosten	0,12	0,11	22,82	5,19	0,48	36,93	1,64	30,80
Kapitalkosten	0,37	0,25	54,30	16,68	0,10	39,54	3,32	31,61
Konst.	1,01	1,95	22,70	18,78	0,03	43,54	0,93	37,62
Durchschn.	0,41	0,58	24,05	14,13	0,13	34,90	1,29	28,85

Tabelle 22.33: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für multiplikativ (Gleichvert. (0,8;1,2)) überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz unbereinigt, 1.000 Replikationen, Inputfaktoren und Output überlagert, robuste Standardfehler

	Multiplikative Überlagerung, konstante Faktoren, Gleichverteilung (0,8;1,2)		Multiplikative Überlagerung, unterschiedliche Faktoren, Gleichverteilung (0,8;1,2)		Multiplikative Überlagerung, unterschiedliche Faktoren, Gleichverteilung (0,8;1,2), Korrektur IV		Multiplikative Überlagerung, unterschiedliche Faktoren, Gleichverteilung (0,8;1,2), SIMEX-Korrektur, Varianz geschätzt	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Materialeinsatz	0,415	(78,58)	0,410	(77,77)	0,415	(78,59)	0,415	(78,69)
Personalkosten	0,340	(58,69)	0,329	(56,12)	0,339	(58,58)	0,339	(56,85)
Externe Dienstleistungen	0,058	(29,01)	0,060	(27,98)	0,058	(29,02)	0,058	(27,44)
Sonstige Kosten	0,113	(34,89)	0,118	(34,50)	0,113	(34,89)	0,113	(33,15)
Kapitalkosten	0,055	(16,01)	0,061	(16,63)	0,055	(16,01)	0,055	(15,13)
Konst.	1,801	(66,37)	1,875	(62,88)	1,804	(66,02)	1,804	(60,94)
R^2	0,977		0,969		0,977			
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	0,10	0,01	1,21	1,04	0,10	0,00	0,11	0,13
Personalkosten	0,18	0,14	3,09	4,25	0,11	0,05	0,09	3,00
Externe Dienstleistungen	0,06	0,03	3,50	3,58	0,00	0,00	0,05	5,44
Sonstige Kosten	0,30	0,03	4,46	1,15	0,33	0,03	0,37	5,01
Kapitalkosten	0,12	0,00	10,34	3,87	0,07	0,00	0,16	5,50
Konst.	0,17	0,27	3,95	5,00	0,02	0,26	0,01	7,93
Durchschn.	0,16	0,08	4,43	3,15	0,11	0,06	0,13	4,50

Tabelle 22.34: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für multiplikativ (Verfahren von Höhne) überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz unbereinigt, 1.000 Replikationen, Inputfaktoren und Output überlagert, robuste Standardfehler

	Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne (f=0,11; s=0,03)		Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne (f=0,11; s=0,03), Korrektur IV		Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne (f=0,11; s=0,03), SIMEX-Korrektur, Varianz geschätzt	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Materialeinsatz	0,414	(78,54)	0,415	(78,54)	0,414	(79,31)
Personalkosten	0,339	(58,50)	0,339	(58,50)	0,344	(59,19)
Externe Dienstleistungen	0,058	(28,93)	0,058	(28,93)	0,057	(27,99)
Sonstige Kosten	0,114	(34,87)	0,113	(34,87)	0,112	(34,13)
Kapitalkosten	0,055	(16,06)	0,055	(16,06)	0,056	(15,80)
Konst.	1,806	(66,10)	1,803	(66,10)	1,743	(61,60)
R^2	0,976		0,976			
	Relative Abweichungen von den Originalwerten in %					
Materialeinsatz	0,18	0,06	0,10	0,22	0,24	0,92
Personalkosten	0,06	0,19	0,12	0,51	1,47	0,99
Externe Dienstleistungen	0,19	0,31	0,02	0,65	1,72	3,55
Sonstige Kosten	0,61	0,09	0,32	0,63	0,88	2,21
Kapitalkosten	0,84	0,31	0,11	0,62	1,82	1,31
Konst.	0,12	0,14	0,04	1,13	3,38	6,93
Durchschn.	0,33	0,18	0,12	0,63	1,59	2,65

d) Die Schätzung einer linearisierten Cobb-Douglas-Produktionsfunktion mit stochastisch überlagerten Ausgangsvariablen

Nun werden die 30 Ausgangsvariablen der KSE stochastisch überlagert und anschließend transformiert. Dabei werden auch solche Unternehmen anonymisiert, für die durch die Transformation bei den Originalwerten „Missing Values" entstehen. Durch die Überlagerung können die Werte jedoch so verändert werden, dass die Transformation durchführbar ist und keine „Missing Values" mehr vorhanden sind.

Bei einer additiven Überlagerung kann bei allen Variablen das Problem auftreten, dass sich die Werte vom positiven in den negativen Bereich verschieben. Bei einer multiplikativen Überlagerung kann dies vorkommen, weil sowohl der Output als auch die Kapitalkosten als Differenzen aus mehreren Ausgangsvariablen berechnet werden. Konkret gilt für den Output:

Y = Bruttoproduktionswert - (Gesamtumsatz - Umsatz aus eigenen Erzeugnissen - Umsatz aus Handelsware)

und für die Kapitalkosten:

Kapitalkosten = Abschreibungen + Mieten und Pachten
= Bruttowertschöpfung - Nettowertschöpfung zu Faktorkosten + Mieten und Pachten,

da die Abschreibungen nicht direkt in den Ausgangsdaten vorhanden sind.

Gilt nun für eine transformierte Variable allgemein $x^{trans} = x_1 - x_2$ und ist bei Verwendung der Originaldaten $x_i^{trans} = x_{1i} - x_{2i} > 0$, so ist nicht zwingend, dass auch $x_i^{trans^a} = x_{1i}^a - x_{2i}^a > 0$ gilt. Vielmehr hängt dies entscheidend von der Art der Anonymisierung ab. Bei der multiplikativen Überlagerung gilt, dass diese Bedingung immer dann automatisch erfüllt ist, wenn alle Merkmalswerte einer Einheit mit demselben Faktor überlagert werden. Somit werden auch in diesem Fall bei der Überlagerung mit einem konstanten Faktor die Originalergebnisse erhalten, weil keine Beobachtung verloren geht und sich – wie oben gezeigt wurde – im Falle konstanter Skalenerträge die Überlagerungen auf beiden Seiten der Gleichung aufheben.

Folgende Verfahren der additiven und multiplikativen stochastischen Überlagerung werden untersucht:

- Additive Überlagerung nach dem Kim-Verfahren: Zunächst wird eine additive Überlagerung gewählt, bei der die Varianz-Kovarianzmatrix der Überlagerungen proportional zur Varianz-Kovarianzmatrix der Originalvariablen ist ($d = 0,1$). Anschließend wird die von Kim (1986) vorgeschlagene Transformation zum Erhalt der ersten und zweiten Momente vorgenommen.
- Multiplikative Überlagerung mit einem (für das Unternehmen) konstanten Überlage-

rungsfaktor, der aus einer Mischungsverteilung aus zwei Normalverteilungen stammt. Die Gipfel der Mischungsverteilung liegen bei 0,92 und 1,08 ($f = 0,08$). Die Standardabweichung beträgt jeweils $s = 0,03$.

- Multiplikative Überlagerung mit Transformation der überlagerten Merkmalswerte, so dass die ersten und zweiten Momente wieder hergestellt werden. Dabei wird zunächst wieder ein konstanter Überlagerungsfaktor gewählt, der aus einer Mischungsverteilung aus zwei Normalverteilungen stammt. Die Gipfel der Mischungsverteilung liegen bei 0,92 und 1,08 ($f = 0,08$). Die Standardabweichung beträgt jeweils $s = 0,03$. Anschließend werden die Merkmalswerte so transformiert, dass die ersten beiden Momente wieder hergestellt werden. Dadurch liegt kein konstanter Überlagerungsfaktor mehr vor.
- Multiplikative Überlagerung mit Faktoren aus einer zweigipfligen Mischungsverteilung nach dem Verfahren von Höhne. Dabei wird zunächst mit einer Wahrscheinlichkeit von 0,5 festgelegt, ob die Werte für ein Unternehmen vergrößert oder verkleinert werden. Als „Grundfaktoren" werden $0,89$ und $1,11$ ($f = 0,11$) gewählt. Diese werden zusätzlich jeweils additiv mit einem Zufallsfehler aus einer Normalverteilung mit Mittelwert Null und Standardabweichung $s = 0,03$ überlagert. Anschließend werden die einzelnen Merkmalswerte mit den so entstandenen Faktoren multiplikativ überlagert.
- Multiplikative Überlagerung nach dem Verfahren von Höhne mit anschließender Transformation der Merkmalswerte, so dass die ersten und zweiten Momente wieder hergestellt werden.

In diesem Fall werden keine Monte-Carlo-Simulationen durchgeführt. Vielmehr erfolgt lediglich eine einmalige Anonymisierung der KSE-Daten. In den Tabellen 22.35 für unbereinigte Daten und 22.36 für um Ausreißer bereinigte Daten erkennt man zunächst, dass bei Anwendung des Verfahrens von Kim oder einer multiplikativen Überlagerung mit anschließender Transformation der Variablen zum Erhalt der ersten und zweiten Momente deutliche Verzerrungen der Koeffizientenschätzer auftreten. Diese sind – wie bereits erwähnt – nur teilweise direkt auf die Überlagerungen zurückzuführen. Ein Teil der Verzerrung erklärt sich vielmehr durch die Veränderung der Zusammensetzung der in die Schätzung einbezogenen Unternehmen. Eine Korrektur der Verzerrung durch den SIMEX-Schätzer wäre auch nur dann möglich, wenn die gleiche Gesamtheit zur Schätzung herangezogen würde.

Erwartungsgemäß ergibt sich bei der multiplikativen Überlagerung mit einem pro Unternehmen konstanten Faktor (für alle Merkmale) in diesem Fallbeispiel eine unverzerrte Schätzung. Von besonderem Interesse ist, welche Auswirkungen sich durch die multiplikative Überlagerung nach dem Höhne-Verfahren ergeben. Hier erfolgt die Überlagerung zwar nicht mit einem konstanten Faktor, wohl aber mit einem ähnlichen Faktor, weil jeder Merkmalswert eines Unternehmens den gleichen „Grundüberlagerungsfaktor" (0,89 oder 1,11) erhält und somit in jedem Fall in die gleiche Richtung verzerrt wird. Dieses Vorgehen reduziert die Anzahl der Fälle, bei denen der Output oder die Kapitalkosten negativ werden und somit die Logarithmen nicht mehr definiert sind.

In den Tabellen 22.37 und 22.38 sind die Ergebnisse für multiplikativ stochastisch überlagerte KSE-Daten mit Überlagerungen aus einer Mischungsverteilung nach dem Verfahren von Höhne dargestellt. Man erkennt, dass trotz zweifacher Verzerrung durch die Überlagerung – unmittelbar und mittelbar, indem sich die Zusammensetzung der einbezogenenen Unternehmen verändert – nur geringfügige Verzerrungen auftreten. Im Durchschnitt ergibt sich durch die Instrumentvariablen-Schätzung beziehungsweise die SIMEX-Korrektur, bei der die Varianz der Überlagerungen durch die Zugabe von Instrumenten geschätzt wird, im Fall der nicht bereinigten Daten eine Verschlechterung, im Fall der bereinigten Daten hingegen eine Verbesserung der Ergebnisse. Allerdings ist die Veränderung der Verzerrung bezogen auf die einzelnen Koeffizienten jeweils unterschiedlich. Somit wird anschaulich, dass die Korrekturverfahren natürlich nur die unmittelbar durch die Überlagerung hervorgerufenen Verzerrungen korrigieren können.

Um den indirekten Effekt der Überlagerung zu eliminieren wird ein Fall konstruiert, in dem vor der Schätzung alle Unternehmen entfernt werden, die bei Originaldaten, bei den anonymisierten Daten oder bei den Instrumenten für eine der im Modell berücksichtigten Variablen einen „Missing Value“ aufweisen. Anschließend werden Schätzungen mit den Originaldaten, mit den entsprechend dem Verfahren von Höhne multiplikativ stochastisch überlagerten Daten sowie Korrekturschätzungen mit Instrumentvariablen und dem SIMEX-Schätzer durchgeführt. Die Ergebnisse finden sich in Tabelle 22.39.

Nun ist eindeutig, dass die Korrekturverfahren zu einer Verbesserung der Schätzergebnisse führen. Die Instrumentvariablen-Schätzung führt zu einer besseren Korrektur, was daran liegt, dass es sich hierbei um ein exaktes Korrekturverfahren handelt (allerdings ist der Erwartungswert der additiven Überlagerungen nicht exakt Null), während die SIMEX-Korrektur lediglich approximativ exakt ist.

Tabelle 22.35: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für stochastisch überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz unbereinigt, Ausgangsvariablen überlagert, robuste Standardfehler

	Additive Überlagerung nach dem Verfahren von Kim (d=0,1)		Multiplikative Überlagerung mit konstanten Faktoren (Mischungsverteilung mit f=0,08 und s=0,018)		Multiplikative Überlagerung mit konst. Faktoren (MV mit f=0,08 und s=0,018) und Kim-Transformation		Multiplikative Überlagerung (MV nach dem Verfahren von Höhne mit f=0,11 und s=0,03) und Kim-Transformation	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,489	(27,53)	0,415	(78,57)	0,511	(116,44)	0,405	(52,05)
Personalkosten	0,184	(13,59)	0,340	(58,72)	0,258	(55,94)	0,356	(43,81)
Externe Dienstleistungen	0,059	(13,11)	0,058	(29,01)	0,055	(32,33)	0,062	(19,59)
Sonstige Kosten	0,151	(16,19)	0,113	(34,90)	0,128	(32,34)	0,119	(25,92)
Kapitalkosten	0,062	(5,89)	0,055	(16,01)	0,052	(14,80)	0,039	(12,11)
Konst.	2,385	(18,33)	1,800	(66,57)	1,334	(60,53)	1,769	(45,06)
R^2	0,903		0,977		0,985		0,955	
Anzahl Beob.	6.620		16.251		16.277		13.240	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	17,83	64,97	0,00	0,03	23,13	48,16	2,41	33,77
Personalkosten	45,72	76,81	0,29	0,19	23,89	4,56	5,01	25,25
Externe Dienstleistungen	1,72	54,82	0,00	0,03	5,17	11,41	6,9	32,49
Sonstige Kosten	33,63	53,61	0,00	0,00	13,27	7,34	5,31	25,73
Kapitalkosten	12,73	63,21	0,00	0,00	5,45	7,56	29,09	24,36
Konst.	32,21	72,31	0,22	0,57	26,05	8,55	1,94	31,92
Durchschnitt	23,97	64,29	0,09	0,14	16,16	14,59	8,44	28,92

Tabelle 22.36: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für stochastisch überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling), Datensatz bereinigt, Ausgangsvariablen überlagert, robuste Standardfehler

	Additive Überlagerung nach dem Verfahren von Kim (d=0,1)		Multiplikative Überlagerung mit konstanten Faktoren (Mischungsverteilung mit f=0,08 und s=0,018)		Multiplikative Überlagerung mit konst. Faktoren (MV mit f=0,08 und s=0,018) und Transformation zum Erhalt der ersten beiden Momente		Multiplikative Überlagerung MV nach dem Verfahren von Höhne mit f=0,11 und s=0,03) und Transformation zum Erhalt der ersten beiden Momente	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,478	(27,75)	0,435	(149,24)	0,507	(206,11)	0,426	(104,07)
Personalkosten	0,185	(13,92)	0,322	(100,32)	0,275	(101,04)	0,339	(66,88)
Externe Dienstleistungen	0,062	(14,08)	0,052	(39,23)	0,041	(39,97)	0,064	(27,89)
Sonstige Kosten	0,144	(17,18)	0,105	(46,83)	0,113	(49,66)	0,119	(33,71)
Kapitalkosten	0,071	(6,96)	0,071	(32,44)	0,060	(30,40)	0,031	(12,00)
Konst.	2,533	(20,45)	1,715	(97,85)	1,411	(107,07)	1,757	(59,31)
R^2	0,911		0,988		0,992		0,968	
Anzahl Beob.	6.605		14.934		15.069		12.189	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	9,89	81,41	0,00	0,01	16,55	38,09	2,07	30,28
Personalkosten	42,55	86,09	0,00	0,22	14,60	0,94	5,28	33,19
Externe Dienstleistungen	19,23	64,14	0,00	0,08	21,15	1,81	23,08	28,96
Sonstige Kosten	37,14	63,31	0,00	0,00	7,62	6,04	13,33	28,02
Kapitalkosten	0,00	78,54	0,00	0,03	15,49	6,26	56,34	63,00
Konst.	47,44	79,07	0,17	0,15	17,87	9,59	2,27	39,29
Durchschnitt	26,04	75,43	0,03	0,08	15,55	10,45	17,06	37,12

Tabelle 22.37: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für multiplikativ überlagerte KSE-Daten, Mischungsverteilung nach dem Verfahren von Höhne (ohne Wirtschaftszweig Recycling), Datensatz unbereinigt, Ausgangsvariablen überlagert, robuste Standardfehler

	Multiplikative Überlagerung nach dem Verfahren von Höhne (f=0,11, s=0,03)		Multiplikative Überlagerung nach dem Verfahren von Höhne (f=0,11, s=0,03), SIMEX-Korrektur, Varianz geschätzt		Multiplikative Überlagerung nach dem Verfahren von Höhne (f=0,11, s=0,03), Korrektur IV	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,416	(80,41)	0,419	(75,62)	0,417	(81,47)
Personalkosten	0,341	(60,57)	0,335	(60,96)	0,329	(56,90)
Externe Dienstleistungen	0,059	(29,23)	0,056	(28,86)	0,058	(28,11)
Sonstige Kosten	0,115	(35,22)	0,11	(29,83)	0,112	(34,33)
Kapitalkosten	0,049	(16,22)	0,066	(16,04)	0,065	(16,09)
Konst.	1,793	(64,93)	1,724	(57,49)	1,807	(64,66)
R^2	0,976				0,976	
Anzahl Beob.	16.024		15.839		15.839	
	Relative Abweichungen von den Originalwerten in %					
Materialeinsatz	0,24	2,32	0,96	3,78	0,54	3,66
Personalkosten	0,59	3,34	1,18	4,01	2,92	2,92
Externe Dienstleistungen	1,72	0,72	3,45	0,55	0,51	3,14
Sonstige Kosten	1,77	0,92	2,65	14,53	1,15	1,63
Kapitalkosten	10,91	1,31	20,00	0,19	18,31	0,50
Konst.	0,61	1,90	4,43	13,14	0,14	2,31
Durchschnitt	2,64	1,75	5,45	6,03	3,93	2,36

Tabelle 22.38: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für multiplikativ überlagerte KSE-Daten, Mischungsverteilung nach dem Verfahren von Höhne (ohne Wirtschaftszweig Recycling), Datensatz bereinigt, Ausgangsvariablen überlagert, robuste Standardfehler

	Multiplikative Überlagerung nach dem Verfahren von Höhne (f=0,11, s=0,03)		Multiplikative Überlagerung nach dem Verfahren von Höhne (f=0,11, s=0,03), SIMEX-Korrektur, Varianz geschätzt		Multiplikative Überlagerung nach dem Verfahren von Höhne (f=0,11, s=0,03), Korrektur IV	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,436	(143,22)	0,436	(137,51)	0,436	(140,34)
Personalkosten	0,335	(99,65)	0,332	(77,35)	0,327	(91,44)
Externe Dienstleistungen	0,054	(39,70)	0,046	(27,24)	0,047	(32,49)
Sonstige Kosten	0,108	(46,40)	0,106	(43,69)	0,107	(45,12)
Kapitalkosten	0,050	(23,98)	0,068	(23,80)	0,066	(24,17)
Konst.	1,702	(91,86)	1,643	(75,44)	1,716	(90,13)
R^2	0,987				0,986	
Anzahl Beob.	14.845		14.824		14.824	
	Relative Abweichungen von den Originalwerten in %					
Materialeinsatz	0,23	4,05	0,23	7,87	0,18	5,98
Personalkosten	4,04	0,45	3,11	22,73	1,58	8,65
Externe Dienstleistungen	3,85	1,12	11,54	30,62	10,20	17,24
Sonstige Kosten	2,86	0,92	0,95	6,71	2,03	3,65
Kapitalkosten	29,58	26,06	4,23	26,61	6,71	25,47
Konst.	0,93	5,98	4,37	22,78	0,12	7,75
Durchschnitt	6,91	6,43	4,07	19,55	3,47	11,46

Tabelle 22.39: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für multiplikativ überlagerte KSE-Daten (ohne Wirtschaftszweig Recycling) Mischungsverteilung nach dem Verfahren von Höhne, Beschränkung auf für Originaldaten, anonymisierte Daten und Instrumente definierte Logarithmen, Ausgangsvariablen überlagert, robuste Standardfehler

	Original		Multiplikative Überlagerung (Mischungsverteilung nach dem Verfahren von Höhne)		Multiplikative Überlagerung (Mischungsverteilung nach dem Verfahren von Höhne), SIMEX-Korrektur, Varianz geschätzt		Multiplikative Überlagerung (Mischungsverteilung nach dem Verfahren von Höhne), Korrektur IV	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,418	(82,16)	0,420	(82,89)	0,419	(79,73)	0,418	(82,14)
Personalkosten	0,328	(58,41)	0,336	(60,80)	0,334	(65,55)	0,328	(57,08)
Externe Dienstleistungen	0,057	(28,22)	0,059	(28,75)	0,056	(29,65)	0,058	(28,01)
Sonstige Kosten	0,111	(34,52)	0,114	(35,01)	0,111	(33,97)	0,112	(34,34)
Kapitalkosten	0,067	(18,27)	0,052	(16,95)	0,066	(17,02)	0,066	(16,18)
Konst.	1,801	(65,70)	1,787	(64,40)	1,725	(60,68)	1,805	(64,65)
R^2	0,978		0,977				0,977	
Anzahl Beob.	15.833		15.833		15.833		15.833	
			Relative Abweichungen von den Originalwerten in %					
Materialeinsatz			0,41	0,89	0,16	2,96	0,04	0,02
Personalkosten			2,64	4,09	1,86	12,22	0,02	2,28
Externe Dienstleistungen			2,56	1,88	2,39	5,07	0,35	0,74
Sonstige Kosten			2,18	1,42	0,56	1,59	0,60	0,52
Kapitalkosten			21,80	7,22	0,36	6,84	1,63	11,44
Konst.			0,82	1,98	4,23	7,64	0,20	1,60
Durchschnitt			5,07	2,91	1,59	6,05	0,47	2,77

22.3.2 Stochastische Überlagerungen in nichtlinearen Modellen

Zur Veranschaulichung der Wirkung stochastischer Überlagerungen in nichtlinearen Modellen wird auf das in Abschnitt 21.2 eingeführte binäre Probit-Modell zur Erklärung der Tarifbindung von Betrieben mit den Daten der Welle 2002 des IAB-Betriebspanels für Baden-Württemberg zurückgegriffen. Dabei wird in einem Fall die einzige in das Modell eingehende metrische Variable (die logarithmierte Beschäftigung) direkt stochastisch überlagert, im anderen Fall wird die Beschäftigung stochastisch überlagert und anschließend logarithmiert.

Folgende Varianten der additiven und multiplikativen stochastischen Überlagerung kommen zum Einsatz:

- additive stochastische Überlagerung durch eine Normalverteilung, wobei die Varianz das 0, 1-fache der Varianz der jeweiligen Originalvariablen beträgt,
- multiplikative Überlagerung durch eine Gleichverteilung im Intervall (0,5;1,5),
- multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne (Abweichung der „Gipfel“ der Mittelwerte der beiden Normalverteilungen: $f = 0,11$; Standardabweichung der Normalverteilungen: $s = 0,03$).

Die Ergebnisse der Schätzungen sind in den Tabellen 22.40 bis 22.42 dargestellt.

Tabelle 22.40: Probit-Schätzung zur Erklärung der Tarifbindung – Schätzergebnisse für additiv stochastisch überlagerte Daten des IAB-Betriebspanels 2002 für Baden-Württemberg, 500 Replikationen

	Additive Überlagerung der logarithmierten Beschäftigung (Normalverteilung, Überlagerungsfaktor 0,1)		Additive Überlagerung der logarithmierten Beschäftigung (Normalverteilung, Überlagerungsfaktor 0,1) SIMEX-Korrektur, Varianz geschätzt		Additive Überlagerung der Beschäftigung (Normalverteilung, Überlagerungsfaktor 0,1)		Additive Überlagerung der Beschäftigung (Normalverteilung, Überlagerungsfaktor 0,1) SIMEX-Korrektur, Varianz geschätzt	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Log. Beschäftigung	0,195	(10,10)	0,291	(10,69)	0,274	(7,82)	0,368	(7,84)
Baugewerbe	0,540	(3,31)	0,729	(4,32)	0,471	(2,33)	0,552	(3,02)
Handel	0,184	(1,46)	0,362	(2,76)	0,140	(0,93)	0,211	(1,54)
Dienstleistungssektor	-0,108	(-1,14)	0,027	(0,26)	-0,151	(-1,42)	-0,113	(-1,14)
Öffentliche Verwaltung	0,754	(4,44)	0,792	(4,38)	0,736	(3,88)	0,785	(4,19)
Konst.	-0,496	(-4,42)	-0,945	(-6,57)	-0,923	(-4,72)	-1,425	(-5,73)
	Relative Abweichungen von den Originalwerten in %							
Log. Beschäftigung	35,13	17,48	3,13	12,68	8,92	36,12	22,31	35,92
Baugewerbe	27,84	25,85	2,44	3,16	37,05	47,66	26,15	32,37
Handel	51,28	49,44	4,47	4,14	62,95	67,56	44,21	46,41
Dienstleistungssektor	371,51	384,16	32,21	33,85	479,30	455,22	383,40	383,89
Öffentliche Verwaltung	5,21	3,56	0,51	4,70	7,57	15,62	1,35	8,82
Konst.	49,85	40,70	4,39	11,93	6,55	36,71	44,18	23,22
Durchschn.	90,14	86,87	7,86	11,74	100,39	109,81	86,93	88,44

Tabelle 22.41: Probit-Schätzung zur Erklärung der Tarifbindung – Schätzergebnisse für multiplikativ stochastisch überlagerte Daten des IAB-Betriebspanels 2002 für Baden-Württemberg, logarithmierte Beschäftigung überlagert, 500 Replikationen

	Multiplikative Überlagerung der logarithmierten Beschäftigung, Gleichverteilung im Intervall (0,5;1,5)		Multiplikative Überlagerung der logarithmierten Beschäftigung, Gleichverteilung im Intervall (0,5;1,5), SIMEX-Korrektur, Varianz geschätzt		Multiplikative Überlagerung der logarithmierten Beschäftigung, nach dem Verfahren von Höhne, (f=0,11, s=0,03)		Multiplikative Überlagerung der logarithmierten Beschäftigung, nach dem Verfahren von Höhne, (f=0,11, s=0,03) SIMEX-Korrektur, Varianz geschätzt	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Log. Beschäftigung	0,201	(10,04)	0,303	(10,27)	0,281	(11,81)	0,303	(12,12)
Baugewerbe	0,541	(3,32)	0,744	(4,44)	0,704	(4,22)	0,750	(4,55)
Handel	0,19	(1,50)	0,376	(2,88)	0,339	(2,60)	0,382	(3,00)
Dienstleistungssektor	-0,102	(-1,08)	0,040	(0,39)	0,01	(0,10)	0,042	(0,43)
Öffentliche Verwaltung	0,754	(4,43)	0,800	(4,39)	0,788	(4,57)	0,797	(4,44)
Konst.	-0,515	(-4,55)	-0,988	(-6,64)	-0,891	(-6,92)	-0,996	(-7,43)
	Relative Abweichungen von den Originalwerten in %							
Log. Beschäftigung	33,14	17,97	0,58	16,09	6,76	3,51	0,66	0,98
Baugewerbe	27,59	25,56	0,49	0,45	5,79	5,38	0,32	2,02
Handel	49,93	47,92	0,61	0,00	10,45	9,72	0,81	4,17
Dienstleistungssektor	358,02	370,00	0,07	2,50	74,35	75,00	6,68	7,50
Öffentliche Verwaltung	5,21	3,70	0,46	4,57	1,03	0,65	0,14	3,48
Konst.	47,83	39,01	0,03	10,99	9,86	7,24	0,85	0,40
Durchschn.	86,96	84,03	0,37	5,77	18,04	16,92	1,58	3,09

Tabelle 22.42: Probit-Schätzung zur Erklärung der Tarifbindung – Schätzergebnisse für multiplikativ stochastisch überlagerte Daten des IAB-Betriebspanels 2002 für Baden-Württemberg, Beschäftigung überlagert, 500 Replikationen

	Multiplikative Überlagerung der Beschäftigung, Gleichverteilung im Intervall (0,5;1,5)		Multiplikative Überlagerung der Beschäftigung, Gleichverteilung im Intervall (0,5;1,5) SIMEX-Korrektur, Varianz geschätzt		Multiplikative Überlagerung der Beschäftigung, nach dem Verfahren von Höhne (f=0,11, s=0,03)		Multiplikative Überlagerung der Beschäftigung, nach dem Verfahren von Höhne (f=0,11, s=0,03) SIMEX-Korrektur, Varianz geschätzt	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Log. Beschäftigung	0,291	(12,06)	0,301	(12,43)	0,3	(12,22)	0,301	(12,58)
Baugewerbe	0,727	(4,35)	0,748	(4,55)	0,745	(4,44)	0,748	(4,57)
Handel	0,36	(2,75)	0,379	(2,99)	0,376	(2,86)	0,379	(2,98)
Dienstleistungssektor	0,025	(0,25)	0,04	(0,41)	0,038	(0,38)	0,04	(0,41)
Öffentliche Verwaltung	0,791	(4,58)	0,797	(4,45)	0,795	(4,60)	0,79	(4,46)
Konst.	-0,927	(-7,15)	-0,974	(-7,44)	-0,98	(-7,42)	-0,986	(-7,55)
	Relative Abweichungen von den Originalwerten in %							
Log. Beschäftigung	3,44	1,47	0,07	1,55	0,48	0,16	0,01	2,78
Baugewerbe	2,72	2,47	0,02	2,02	0,40	0,45	0,02	2,47
Handel	5,06	4,51	0,03	3,82	0,74	0,69	0,04	3,47
Dienstleistungssektor	37,87	37,50	0,03	2,50	5,01	5,00	0,18	2,50
Öffentliche Verwaltung	0,63	0,43	0,08	3,26	0,09	0,00	0,02	3,04
Konst.	6,21	4,16	1,40	0,27	0,85	0,54	0,17	1,21
Durchschn.	9,32	8,42	0,27	2,24	1,26	1,14	0,07	2,58

Bei der additiven Überlagerung kann die Verzerrung mit Hilfe des SIMEX-Schätzers nur korrigiert werden, wenn die logarithmierte Beschäftigung überlagert wird. Allerdings bleibt auch in diesem Fall eine sichtbare Abweichung von den Originalwerten bestehen. Die höchste Abweichung tritt allerdings bei der ohnehin nicht signifikanten Dummy für den Dienstleistungssektor auf.

Bei der multiplikativen Überlagerung fällt die Verzerrung deutlich geringer aus, wenn direkt die Beschäftigung überlagert wird, was daran liegt, dass durch die Logarithmierung dann ein (additiver) Fehler mit kleinerer Varianz entsteht. Zudem ergibt sich eine deutlich geringere Verzerrung bei Anwendung des Verfahrens von Höhne (Mischungsverteilung aus zwei Normalverteilungen) im Vergleich zur Überlagerung mit gleichverteilten Faktoren im Intervall (0,5;1,5). Dies ist allerdings leicht einsichtig, weil bei letzterer die Varianz der Überlagerungen auch deutlich größer ist. Die SIMEX-Korrektur führt in allen vier Fällen dazu, dass die geschätzten Parameter im Durchschnitt nur noch geringfügig von den originalen Parametern abweichen, allerdings weisen die korrigierten Schätzer bei der Gleichverteilung über die 1.000 Replikationen eine höhere Varianz auf als beim Verfahren von Höhne.

Hinsichtlich der statistischen Signifikanz der Einflüsse führen diejenigen Korrekturschätzer, die auch im Durchschnitt der 500 Replikationen zu guten Ergebnissen führen, in jeder der 500 Wiederholungen zu eindeutigen und mit dem Original übereinstimmenden Ergebnissen.

Kapitel 23

Mikroaggregationsverfahren in linearen und nichtlinearen Modellen

23.1 Theoretische Eigenschaften

23.1.1 Mikroaggregationsverfahren in linearen Modellen

Für das lineare Modell gelte wiederum:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}. \tag{23.1}$$

a) Zufällige Mikroaggregation oder abstandsorientierte Mikroaggregation ohne Einbeziehung der abhängigen Variablen

Zunächst wird der Fall betrachtet, dass eine zufällige Mikroaggregation vorgenommen wird, bei der die Aggregationsmatrix von den zu anonymisierenden Variablen unabhängig ist, oder eine abstandsorientierte Mikroaggregation, bei der die Aggregationsmatrix ausschließlich von den Regressoren abhängt und nicht von der abhängigen Variablen ($\mathbf{D} = \mathbf{D}(\mathbf{X})$).[30]

Nur die abhängige Variable wird mikroaggregiert

Wird nur die abhängige Variable mikroaggregiert, so lässt sich das zu schätzende Modell

30) Wie in Unterabschnitt 6.2.4 gezeigt wurde, ist die Aggregationsmatrix $\mathbf{D}$ außer bei der Bootstrap-Mikroaggregation stets symmetrisch idempotent.

schreiben als

$$\mathbf{Dy} = \mathbf{X}\boldsymbol{\beta}^a + \mathbf{u}^a. \tag{23.2}$$

Für den OLS-Schätzer ergibt sich:

$$\hat{\boldsymbol{\beta}}^a = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Dy}. \tag{23.3}$$

Damit gilt für den Erwartungswert des Schätzers:

$$E\left(\hat{\boldsymbol{\beta}}^a\right) = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Dy}\right] \tag{23.4}$$

oder

$$E\left(\hat{\boldsymbol{\beta}}^a\right) = E\left[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\left(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}\right)\right], \tag{23.5}$$

$$E\left(\hat{\boldsymbol{\beta}}^a\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}E\left(\mathbf{u}\right) \tag{23.6}$$

und mit $E(\mathbf{u}) = \mathbf{0}$

$$E\left(\hat{\boldsymbol{\beta}}^a\right) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}\boldsymbol{\beta}. \tag{23.7}$$

Das bedeutet, dass der Schätzer in diesem Fall stets verzerrt ist. Die Höhe der Verzerrung ergibt sich dabei durch

$$\begin{aligned} E\left(\mathbf{Bias}^a\right) &= E\left(\hat{\boldsymbol{\beta}}^a\right) - E\left(\hat{\boldsymbol{\beta}}\right) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta} \\ &= \left((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}\mathbf{X} - \mathbf{I}_K\right)\boldsymbol{\beta}. \end{aligned} \tag{23.8}$$

Für den Wahrscheinlichkeitsgrenzwert des Schätzers ergibt sich:

$$\begin{aligned} plim\left(\hat{\boldsymbol{\beta}}^a\right) &= plim\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1} plim\left(\frac{\mathbf{X}'\mathbf{D}\mathbf{X}}{n}\right)\boldsymbol{\beta} \\ &= \mathbf{Q}\, plim\left(\frac{\mathbf{X}'\mathbf{D}\mathbf{X}}{n}\right)\boldsymbol{\beta}. \end{aligned} \tag{23.9}$$

Damit ist der Schätzer auch nicht konsistent.

Nur die Regressoren werden gemeinsam mikroaggregiert

Wird nur die Regressormatrix durch gemeinsame zufällige Mikroaggregation oder gemeinsame abstandsorientierte Mikroaggregation (bei der die abhängige Variable unberücksichtigt bleibt) anonymisiert, so ergibt sich für das zu schätzende Modell:

$$\mathbf{y} = \mathbf{DX}\boldsymbol{\beta}^{a2} + \mathbf{u}^{a2}. \tag{23.10}$$

Für den Kleinste-Quadrate-Schätzer ergibt sich:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{a2} &= ((\mathbf{DX})'\mathbf{DX})^{-1}(\mathbf{DX})'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{D}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{D}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{Dy}. \end{aligned} \tag{23.11}$$

Für den Erwartungswert dieses Schätzers ergibt sich:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}^{a2}\right) &= E\left[(\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{Dy}\right] \\ &= E\left[(\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{D}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})\right] \\ &= E\left[(\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{DX}\boldsymbol{\beta} + (\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{Du}\right] \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{D}E(\mathbf{u}) = \boldsymbol{\beta}. \end{aligned} \tag{23.12}$$

Damit ist der Schätzer in diesem Fall erwartungstreu. Allerdings ist er nicht effizient, denn es gilt für die Varianz-Kovarianzmatrix des Schätzers:

$$\begin{aligned} var\left(\hat{\boldsymbol{\beta}}^{a2}\right) &= E\left[\left(\hat{\boldsymbol{\beta}}^{a2} - E\left(\hat{\boldsymbol{\beta}}^{a2}\right)\right)\left(\hat{\boldsymbol{\beta}}^{a2} - E\left(\hat{\boldsymbol{\beta}}^{a2}\right)\right)'\right] \\ &= E\left[\left((\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{Du}\right)\left((\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{Du}\right)'\right] \\ &= E\left[(\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{Duu}'\mathbf{D}'\mathbf{X}(\mathbf{X}'\mathbf{DX})^{-1}\right] \\ &= \sigma_u^2\left[(\mathbf{X}'\mathbf{DX})^{-1}\mathbf{X}'\mathbf{DX}(\mathbf{X}'\mathbf{DX})^{-1}\right] \\ &= \sigma_u^2(\mathbf{X}'\mathbf{DX})^{-1}. \end{aligned} \tag{23.13}$$

Für die Verzerrung der Varianz-Kovarianzmatrix des Schätzers ergibt sich somit:

$$\begin{aligned} var\left(\hat{\boldsymbol{\beta}}^{a2}\right) - var\left(\hat{\boldsymbol{\beta}}\right) &= \sigma_u^2 \left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1} - \sigma_u^2 \left(\mathbf{X}'\mathbf{X}\right)^{-1} \\ &= \sigma_u^2 \left(\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1} - \left(\mathbf{X}'\mathbf{X}\right)^{-1}\right). \end{aligned} \tag{23.14}$$

Diese Differenz ist stets nichtnegativ definit (Lechner und Pohlmeier 2003).

Lechner und Pohlmeier (2003) betrachten diese Differenz genauer für den Fall ohne Absolutglied. Sie zeigen, dass in diesem Fall

$$\left(var\left(\hat{\boldsymbol{\beta}}\right)\right)^{-1} - \left(var\left(\hat{\boldsymbol{\beta}}^{a2}\right)\right)^{-1} = \frac{1}{\sigma_u^2} \sum_{j=1}^{M} \sum_{i=1}^{A} \left(x_{ij} - \bar{x}_j\right)^2 \tag{23.15}$$

gilt.

Man erkennt, dass die Verzerrung geringer ist, wenn die einzelnen Werte näher an dem jeweiligen Mittelwert liegen (Lechner und Pohlmeier 2003). Dies bedeutet, dass der Effizienzverlust mit sinkender Gruppengröße abnimmt und außerdem bei geordneten Werten (abstandsorientierte Mikroaggregation) geringer ist.

Schmid et al. (2005) leiten her, dass der Effizienzverlust asymptotisch gegen Null strebt. Hierfür kann man die Gleichung (23.14) wie folgt umschreiben:

$$var\left(\hat{\boldsymbol{\beta}}^{a2}\right) - var\left(\hat{\boldsymbol{\beta}}\right) = \frac{\sigma_u^2}{n} \left(\left(\frac{\mathbf{X}'\mathbf{D}\mathbf{X}}{n}\right)^{-1} - \left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}\right). \tag{23.16}$$

Da die beiden Ausdrücke $\left(\frac{\mathbf{X}'\mathbf{D}\mathbf{X}}{n}\right)^{-1}$ und $\left(\frac{\mathbf{X}'\mathbf{X}}{n}\right)^{-1}$ für $n \rightharpoonup \infty$ jeweils gegen feste Matrizen konvergieren, konvergiert der Ausdruck auf der rechten Seite der Gleichung (23.16) insgesamt gegen Null.

Die Erwartungstreue des KQ-Schätzers gilt in diesem Fall jedoch nicht für das verallgemeinerte lineare Modell. Hier gilt für die Varianz-Kovarianzmatrix der Residuen $var(\mathbf{u}) = E\left(\mathbf{u}\mathbf{u}'\right) = \sigma_u^2\boldsymbol{\Omega}$ und für den GLS-Schätzer ergibt sich:

$$\hat{\boldsymbol{\beta}}^{GLS} = \left(\mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\boldsymbol{\Omega}^{-1}\mathbf{y}. \tag{23.17}$$

Für den Fall der gemeinsamen Mikroaggregation der Regressoren gilt für den GLS-Schätzer:

$$\hat{\boldsymbol{\beta}}^{GLS,a2} = \left(\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{y}. \tag{23.18}$$

Damit ergibt sich für den Erwartungswert des GLS-Schätzers bei gemeinsamer Mikroaggregation der Regressoren:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}^{GLS,a2}\right) &= E\left[\left(\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right)\right] \\ &= \left(\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{23.19}$$

Abhängige Variable und Regressoren werden gemeinsam mikroaggregiert

Werden die abhängige Variable und die Regressoren gemeinsam mikroaggregiert, so gilt für das zu schätzende Modell:

$$\mathbf{D}\mathbf{y} = \mathbf{D}\mathbf{X}\boldsymbol{\beta}^{a3} + \mathbf{u}^{a3}. \tag{23.20}$$

Für den KQ-Schätzer gilt:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^{a3} &= \left((\mathbf{D}\mathbf{X})'\mathbf{D}\mathbf{X}\right)^{-1}(\mathbf{D}\mathbf{X})'\mathbf{D}\mathbf{y} \\ &= \left(\mathbf{X}'\mathbf{D}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}'\mathbf{D}\mathbf{y} \\ &= \left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{y}. \end{aligned} \tag{23.21}$$

Für den Erwartungswert des Schätzers ergibt sich in diesem Fall:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}^{a3}\right) &= E\left[\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{y}\right] \\ &= E\left[\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right)\right] \\ &= E\left[\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{u}\right] \\ &= \boldsymbol{\beta} + \left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}E(\mathbf{u}) = \boldsymbol{\beta}. \end{aligned} \tag{23.22}$$

Es zeigt sich also, dass der Schätzer auch für diesen Fall erwartungstreu ist.

Die Varianz-Kovarianzmatrix des Schätzers ergibt sich wiederum als:

$$
\begin{aligned}
var\left(\hat{\boldsymbol{\beta}}^{a3}\right) &= E\left[\left(\hat{\boldsymbol{\beta}}^{a3} - E\left(\hat{\boldsymbol{\beta}}^{a3}\right)\right)\left(\hat{\boldsymbol{\beta}}^{a3} - E\left(\hat{\boldsymbol{\beta}}^{a3}\right)\right)'\right] \\
&= E\left[\left(\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{u}\right)\left(\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\mathbf{u}\right)'\right] \\
&= \sigma_u^2\left(\mathbf{X}'\mathbf{D}\mathbf{X}\right)^{-1}. \qquad (23.23)
\end{aligned}
$$

Somit entspricht der Effizienzverlust exakt dem Fall, bei dem lediglich die Regressoren gemeinsam mikroaggregiert werden.

Durch die Mikroaggregation wird auch der herkömmliche Schätzer für die Varianz des Fehlerterms und somit der Standardfehler des KQ-Schätzers von $\boldsymbol{\beta}$ verzerrt (Lechner und Pohlmeier 2003).

Lechner und Pohlmeier (2003) leiten einen unverzerrten Schätzer für σ_u^2 für den Fall der gemeinsamen Mikroaggregation aller Variablen wie folgt her. Das zu schätzende Modell aus der Gleichung (23.20) kann auch geschrieben werden als:

$$\mathbf{D}\mathbf{y} = \mathbf{D}\mathbf{X}\boldsymbol{\beta}^{a3} + \mathbf{D}\mathbf{u}. \qquad (23.24)$$

Im klassischen linearen Regressionsmodell gilt für den Vektor der geschätzten Residuen $\hat{\mathbf{u}} = \mathbf{M}\mathbf{u}$ mit $\mathbf{M} = \mathbf{I}_n - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$. Somit gilt im Fall der gemeinsamen Mikroaggregation aller Variablen:

$$\hat{\mathbf{u}}^{a3} = \mathbf{M}^a\mathbf{u}^{a3} = \mathbf{M}^a\mathbf{D}\mathbf{u} \qquad (23.25)$$

mit $\mathbf{M}^a = \mathbf{I}_n - \mathbf{X}^a\left(\mathbf{X}^{a'}\mathbf{X}^a\right)^{-1}\mathbf{X}^{a'}$.

Somit gilt für den Erwartungswert von $\hat{\mathbf{u}}^{a3'}\hat{\mathbf{u}}^{a3}$:

$$\begin{aligned}
E\left(\hat{\mathbf{u}}^{a3\prime}\hat{\mathbf{u}}^{a3}\right) &= E\left(\mathbf{u}^{a3\prime}\mathbf{M}^{a}\mathbf{u}^{a3}\right) \\
&= E\left((\mathbf{Du})'\mathbf{M}^{a}\mathbf{Du}\right) \\
&= E\left(\mathbf{u}'\mathbf{D}'\mathbf{M}^{a}\mathbf{Du}\right) \\
&= E\left[tr\left(\mathbf{u}'\mathbf{D}'\mathbf{M}^{a}\mathbf{Du}\right)\right] \\
&= tr\left(E\left(\mathbf{uu}'\right)\mathbf{D}'\mathbf{M}^{a}\mathbf{D}\right) \\
&= \sigma_u^2\, tr\left(\mathbf{D}'\mathbf{M}^{a}\mathbf{D}\right) \\
&= \sigma_u^2\, tr\left(\mathbf{M}^{a}\mathbf{D}\right) \\
&= \sigma_u^2\, tr\left(\mathbf{D}-\mathbf{X}^{a}\left(\mathbf{X}^{a\prime}\mathbf{X}^{a}\right)^{-1}\mathbf{X}^{a\prime}\right) \\
&= \sigma_u^2\,(M-K). \qquad (23.26)
\end{aligned}$$

Damit ist $\frac{\hat{\mathbf{u}}^{a3\prime}\hat{\mathbf{u}}^{a3}}{M-K}$ ein erwartungstreuer Schätzer für die Varianz der Residuen.

Die Erwartungstreue des Schätzers gilt in diesem Fall auch für das verallgemeinerte lineare Modell.

Für den Fall der gemeinsamen Mikroaggregation aller Variablen gilt für den GLS-Schätzer:

$$\hat{\boldsymbol{\beta}}^{GLS,a3} = \left(\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{DX}\right)^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{Dy}. \qquad (23.27)$$

Damit ergibt sich für den Erwartungswert des GLS-Schätzers bei gemeinsamer Mikroaggregation:

$$\begin{aligned}
E\left(\hat{\boldsymbol{\beta}}^{GLS,a3}\right) &= E\left[\left(\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{DX}\right)^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{D}\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right)\right] \\
&= \left(\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{DX}\right)^{-1}\mathbf{X}'\mathbf{D}\boldsymbol{\Omega}^{-1}\mathbf{DX}\boldsymbol{\beta} = \boldsymbol{\beta}. \qquad (23.28)
\end{aligned}$$

Getrennte Mikroaggregation von abhängiger Variablen einerseits und Regressoren andererseits

Werden abhängige Variable und Regressoren getrennt mikroaggregiert (wobei angenommen wird, dass die einzelnen Regressoren gemeinsam mikroaggregiert werden), so ergibt sich das zu schätzende Modell als

$$\mathbf{D}_y\mathbf{y} = \mathbf{D}_x\mathbf{X}\boldsymbol{\beta}^b + \mathbf{u}^b \tag{23.29}$$

mit $\mathbf{D}_x \neq \mathbf{D}_y$.

Für den KQ-Schätzer folgt daraus:

$$\begin{aligned} \hat{\boldsymbol{\beta}}^b &= ((\mathbf{D}_x\mathbf{X})'\mathbf{D}_x\mathbf{X})^{-1}(\mathbf{D}_x\mathbf{X})'\mathbf{D}_y\mathbf{y} \\ &= (\mathbf{X}'\mathbf{D}_x\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_x\mathbf{D}_y\mathbf{y}. \end{aligned} \tag{23.30}$$

Für dessen Erwartungswert gilt:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}^b\right) &= E\left[(\mathbf{X}'\mathbf{D}_x\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_x\mathbf{D}_y\mathbf{y}\right] \\ &= E\left[(\mathbf{X}'\mathbf{D}_x\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_x\mathbf{D}_y(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})\right] \\ &= E(\mathbf{X}'\mathbf{D}_x\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_x\mathbf{D}_y\mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{23.31}$$

Damit ist der KQ-Schätzer in diesem Fall nicht erwartungstreu. Er ist auch nicht konsistent, da die Aggregationsmatrix auch für n gegen unendlich nicht verschwindet.

Abhängige Variable wird getrennt von den teilweise gemeinsam mikroaggreggierten Regressoren mikroaggregiert

Bisher wurde stets eine gemeinsame Mikroaggregation der einzelnen Regressoren unterstellt. Nun soll der Fall betrachtet werden, dass auch die Regressoren teilweise gemeinsam mikroaggregiert werden.

Zu diesem Zweck wird das lineare Regressionsmodell zunächst in der folgenden Form dargestellt:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}. \tag{23.32}$$

Dabei beinhaltet die $(n \times K_1)$-Matrix $\mathbf{X}_1$ die Variablenwerte für K_1 Regressoren, die $(n \times K_2)$-Matrix $\mathbf{X}_2$ die Variablenwerte der anderen K_2 Regressoren. Dabei gilt $K_1 + K_2 = K$.

Wird keine der Variablen mikroaggregiert, so lassen sich die beiden Teil-Schätzer schreiben als:

$$\hat{\boldsymbol{\beta}}_1 = \left(\mathbf{X}_1'\mathbf{M}_2\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y} \tag{23.33}$$

und

$$\hat{\boldsymbol{\beta}}_2 = \left(\mathbf{X}_2'\mathbf{M}_1\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{M}_1\mathbf{y} \tag{23.34}$$

mit

$$\mathbf{M}_1 = \mathbf{I}_n - \mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1' \tag{23.35}$$

und

$$\mathbf{M}_2 = \mathbf{I}_n - \mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'. \tag{23.36}$$

Damit lassen sich die beiden Schätzer weiter umschreiben zu:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= \left(\mathbf{X}_1'\left(\mathbf{I}_n - \mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\right)\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{M}_2\mathbf{y} \\ &= \left[\left(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1' - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\right)\right] \\ &\quad \times \left(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}\right)\end{aligned} \tag{23.37}$$

und damit

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= \left(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1' - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\right)\mathbf{X}_1\boldsymbol{\beta}_1 \\ &+ \left(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1' - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\right)\mathbf{X}_2\boldsymbol{\beta}_2 \\ &+ \left(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{X}_1\right)^{-1}\left(\mathbf{X}_1' - \mathbf{X}_1'\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\right)\mathbf{u}\end{aligned} \tag{23.38}$$

sowie

$$\begin{aligned}\hat{\boldsymbol{\beta}}_2 &= \left(\mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\right)^{-1}\left(\mathbf{X}_2' - \mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right)\mathbf{X}_1\boldsymbol{\beta}_1 \\ &+ \left(\mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\right)^{-1}\left(\mathbf{X}_2' - \mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right)\mathbf{X}_2\boldsymbol{\beta}_2 \\ &+ \left(\mathbf{X}_2'\mathbf{X}_2 - \mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\right)^{-1}\left(\mathbf{X}_2' - \mathbf{X}_2'\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\right)\mathbf{u}. \quad (23.39)\end{aligned}$$

Zunächst wird der allgemeine Fall betrachtet, bei dem die abhängige Variable mit der Aggregationsmatrix $\mathbf{D}_y$, K_1 erklärende Variablen mit der Aggregationsmatrix $\mathbf{D}_1$ und K_2 Regressoren mit der Aggregationsmatrix $\mathbf{D}_2$ mikroaggregiert werden.

Für den ersten Summanden auf der rechten Seite der Gleichung (23.38) gilt dann:

$$\begin{aligned}&\left((\mathbf{D}_1\mathbf{X}_1)'\mathbf{D}_1\mathbf{X}_1 - (\mathbf{D}_1\mathbf{X}_1)'\mathbf{D}_2\mathbf{X}_2\left((\mathbf{D}_2\mathbf{X}_2)'\mathbf{D}_2\mathbf{X}_2\right)^{-1}(\mathbf{D}_2\mathbf{X}_2)'\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\ &\quad\times\left((\mathbf{D}_1\mathbf{X}_1)' - (\mathbf{D}_1\mathbf{X}_1)'\mathbf{D}_2\mathbf{X}_2\left((\mathbf{D}_2\mathbf{X}_2)'\mathbf{D}_2\mathbf{X}_2\right)^{-1}(\mathbf{D}_2\mathbf{X}_2)'\right)\mathbf{D}_y\mathbf{X}_1\boldsymbol{\beta}_1 \\ &\qquad= \left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\ &\qquad\quad\times\left(\mathbf{X}_1'\mathbf{D}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\right)\mathbf{D}_y\mathbf{X}_1\boldsymbol{\beta}_1 \\ &\qquad= \left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\ &\quad\times\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_y\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_1\right)\boldsymbol{\beta}_1. \quad (23.40)\end{aligned}$$

Für den zweiten Summanden auf der rechten Seite der Gleichung (23.38) ergibt sich:

$$\begin{aligned}&\left((\mathbf{D}_1\mathbf{X}_1)'\mathbf{D}_1\mathbf{X}_1 - (\mathbf{D}_1\mathbf{X}_1)'\mathbf{D}_2\mathbf{X}_2\left((\mathbf{D}_2\mathbf{X}_2)'\mathbf{D}_2\mathbf{X}_2\right)^{-1}(\mathbf{D}_2\mathbf{X}_2)'\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\ &\quad\times\left((\mathbf{D}_1\mathbf{X}_1)' - (\mathbf{D}_1\mathbf{X}_1)'\mathbf{D}_2\mathbf{X}_2\left((\mathbf{D}_2\mathbf{X}_2)'\mathbf{D}_2\mathbf{X}_2\right)^{-1}(\mathbf{D}_2\mathbf{X}_2)'\right)\mathbf{D}_y\mathbf{X}_2\boldsymbol{\beta}_2 \\ &\qquad= \left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\ &\qquad\quad\times\left(\mathbf{X}_1'\mathbf{D}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\right)\mathbf{D}_y\mathbf{X}_2\boldsymbol{\beta}_2 \\ &\qquad= \left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\ &\quad\times\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_y\mathbf{X}_2 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_2\right)\boldsymbol{\beta}_2. \quad (23.41)\end{aligned}$$

Und für den dritten Summanden gilt:

$$\begin{aligned}
&\left((\mathbf{D}_1\mathbf{X}_1)'\,\mathbf{D}_1\mathbf{X}_1 - (\mathbf{D}_1\mathbf{X}_1)'\,\mathbf{D}_2\mathbf{X}_2\left((\mathbf{D}_2\mathbf{X}_2)'\,\mathbf{D}_2\mathbf{X}_2\right)^{-1}(\mathbf{D}_2\mathbf{X}_2)'\,\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\
&\qquad\times\left((\mathbf{D}_1\mathbf{X}_1)' - (\mathbf{D}_1\mathbf{X}_1)'\,\mathbf{D}_2\mathbf{X}_2\left((\mathbf{D}_2\mathbf{X}_2)'\,\mathbf{D}_2\mathbf{X}_2\right)^{-1}(\mathbf{D}_2\mathbf{X}_2)'\right)\mathbf{u} \\
&\qquad=\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\
&\qquad\qquad\times\left(\mathbf{X}_1'\mathbf{D}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\right)\mathbf{D}_y\mathbf{u} \\
&\qquad=\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\
&\qquad\qquad\times\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_y - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\right)\mathbf{u}. \qquad (23.42)
\end{aligned}$$

Da der Erwartungswert des dritten Summanden in jedem Fall $\mathbf{0}$ ist, ergibt sich der Erwartungswert des KQ-Schätzers für $\boldsymbol{\beta}_1$ allgemein als

$$\begin{aligned}
E\left(\hat{\boldsymbol{\beta}}_1^c\right) &= \left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\
&\quad\times\Big[\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_y\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_1\right)\boldsymbol{\beta}_1 \\
&\quad+\Big(\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_y\mathbf{X}_2 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1} \\
&\quad\times\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_2\Big)\boldsymbol{\beta}_2\Big] \qquad (23.43)
\end{aligned}$$

und der KQ-Schätzer für $\boldsymbol{\beta}_2$ als

$$\begin{aligned}
E\left(\hat{\boldsymbol{\beta}}_2^c\right) &= \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\right)^{-1} \\
&\quad\times\Big[\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_1 - \mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_y\mathbf{X}_1\right)\boldsymbol{\beta}_1 \\
&\quad+\Big(\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\
&\quad\times\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_y\mathbf{X}_2\Big)\boldsymbol{\beta}_2\Big]. \qquad (23.44)
\end{aligned}$$

Für die Varianzen der beiden Teil-Schätzer gilt:

$$
\begin{aligned}
var\left(\hat{\boldsymbol{\beta}}_1^c\right) &= E\left[\left(\hat{\boldsymbol{\beta}}_1^a - E\left(\boldsymbol{\beta}_1^a\right)\right)\left(\hat{\boldsymbol{\beta}}_1^a - E\left(\boldsymbol{\beta}_1^a\right)\right)'\right] \\
&= E\left[\left(\mathbf{M}_{D_1}\mathbf{u}\right)\left(\mathbf{M}_{D_1}\mathbf{u}\right)'\right] \\
&= E\left[\mathbf{M}_{D_1}\mathbf{u}\mathbf{u}'\mathbf{M}_{D_1}'\right] \\
&= \sigma_u^2\mathbf{M}_{D_1}\mathbf{M}_{D_1}' \qquad (23.45)
\end{aligned}
$$

mit

$$
\begin{aligned}
\mathbf{M}_{D_1} &= \left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)^{-1} \\
&\quad \times \left(\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_y - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\right) \qquad (23.46)
\end{aligned}
$$

und

$$
var\left(\hat{\boldsymbol{\beta}}_2^c\right) = \sigma_u^2\mathbf{M}_{D_2}\mathbf{M}_{D_2}' \qquad (23.47)
$$

mit

$$
\begin{aligned}
\mathbf{M}_{D_2} &= \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\right)^{-1} \\
&\quad \times \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y - \mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_y\right) \qquad (23.48)
\end{aligned}
$$

Werden die abhängige Variable sowie jeweils ein Teil der erklärenden Variablen unabhängig voneinander gemeinsam mikroaggregiert, gilt also $\mathbf{D}_y \neq \mathbf{D}_1 \neq \mathbf{D}_2$, so sind die beiden Teilschätzer nicht erwartungstreu. Wiederum gilt, dass die Aggregationsmatrizen auch für n gegen unendlich nicht verschwinden. Somit sind die Schätzer auch nicht konsistent.

Nur ein Teil der Regressoren wird gemeinsam mikroaggregiert

Wird nur ein Teil der Einflussgrößen gemeinsam mikroaggregiert, so gilt $\mathbf{D}_1 = \mathbf{D}_y = \mathbf{I}_n \neq \mathbf{D}_2$. Daraus ergibt sich für die Erwartungswerte der beiden Teilschätzer:

$$
\begin{aligned}
E\left(\hat{\boldsymbol{\beta}}_1^{c1}\right) &= \boldsymbol{\beta}_1 + \left(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\right)^{-1} \\
&\quad \left(\mathbf{X}_1'\mathbf{X}_2 - \mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\right)\boldsymbol{\beta}_2. \qquad (23.49)
\end{aligned}
$$

und

$$E\left(\hat{\boldsymbol{\beta}}_2^{c1}\right) = \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\right)^{-1} \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\right)\boldsymbol{\beta}_2. \quad (23.50)$$

Auch der Schätzer für $\boldsymbol{\beta}_2$ ist nicht erwartungstreu. Auch hier gilt wiederum, dass die Aggregationsmatrix nicht für n gegen unendlich verschwindet, so dass die beiden Teilschätzer nicht konsistent sind.

Regressoren werden teilweise gemeinsam mikroaggregiert

Nun werden zwar alle Regressoren mikroaggregiert, allerdings nur teilweise gemeinsam. Zur Vereinfachung wird davon ausgegangen, dass die Regressoren in zwei Gruppen jeweils gemeinsam mikroaggregiert werden. Es gilt dann für die Aggregationsmatrizen $\mathbf{D}_1 \neq \mathbf{D}_2$. Die abhängige Variable wird zunächst nicht mikroaggregiert.

Aus den Gleichungen (23.43) und (23.44) folgt somit für die Erwartungswerte der beiden Teilschätzer:

$$E\left(\hat{\boldsymbol{\beta}}_1^{c2}\right) = \left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)^{-1} \times \left[\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\right)\boldsymbol{\beta}_1 + \left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_2 - \mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\right)\boldsymbol{\beta}_2\right] \quad (23.51)$$

und

$$E\left(\hat{\boldsymbol{\beta}}_2^{c2}\right) = \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_1\mathbf{D}_2\mathbf{X}_2\right)^{-1} \times \left[\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1 - \mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\right)\boldsymbol{\beta}_1 + \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{D}_1\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_1\mathbf{X}_2\right)\boldsymbol{\beta}_2\right]. \quad (23.52)$$

Beide Teilschätzer sind somit nicht erwartungstreu und bleiben auch asymptotisch verzerrt.

Abhängige Variable und Teil der Regressoren werden gemeinsam mikroaggregiert

Wird die abhängige Variable und ein Teil der Regressoren gemeinsam mikroaggregiert, gilt also $\mathbf{D}_y = \mathbf{D}_2$, so gilt für die Erwartungswerte der beiden Teilschätzer:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}_1^{c3}\right) &= \left(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\right)^{-1} \\ &\quad \times \left(\mathbf{X}_1'\mathbf{D}_2\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\right)\boldsymbol{\beta}_1 \end{aligned} \tag{23.53}$$

und

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}_2^{c3}\right) &= \boldsymbol{\beta}_2 + \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\right)^{-1} \\ &\quad \times \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1 - \mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_2\mathbf{X}_1\right)\boldsymbol{\beta}_1. \end{aligned} \tag{23.54}$$

Auch in diesem Fall sind also die Teilschätzer verzerrt.

Die abhängige Variable und ein Teil der Regressoren werden unabhängig voneinander mikroaggregiert

Zuletzt wird noch der Fall betrachtet, dass die abhängige Variable getrennt von einem Teil der Regressoren mikroaggregiert wird, wobei diese Regressoren gemeinsam mikroaggregiert werden. Es gilt also $\mathbf{D}_y \neq \mathbf{D}_2$ und damit für die Erwartungswerte der beiden verzerrten Teilschätzer:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}_1^{c4}\right) &= \left(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\right)^{-1} \\ &\quad \times \Big[\left(\mathbf{X}_1'\mathbf{D}_y\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_1\right)\boldsymbol{\beta}_1 \\ &\quad + \left(\mathbf{X}_1'\mathbf{D}_y\mathbf{X}_2 - \mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_2\right)\boldsymbol{\beta}_2\Big] \end{aligned} \tag{23.55}$$

und

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}_2^{c4}\right) &= \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\right)^{-1} \\ &\quad \times \Big[\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_1 - \mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_y\mathbf{X}_1\right)\boldsymbol{\beta}_1 \\ &\quad + \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{D}_y\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_y\mathbf{X}_2\right)\boldsymbol{\beta}_2\Big]. \end{aligned} \tag{23.56}$$

Zusammengefasst ergibt sich damit nur in zwei Fällen ein erwartungstreuer beziehungsweise konsistenter KQ-Schätzer, zum einen wenn ausschließlich die Regressoren (und zwar alle gemeinsam) mikroaggregiert werden, zum anderen wenn abhängige Variable und Regressoren gemeinsam mikroaggregiert werden. Dieses Ergebnis gilt sowohl für die zufällige Mikroaggregation als auch für eine deterministische Mikroaggregation, bei der die abstandsorientierte Gruppenbildung nur von den Regressoren, nicht aber von der abhängigen Variablen, abhängt. Grundsätzlich gilt, dass der KQ-Schätzer im Fall mikroaggregierter Variablen nicht effizent ist. Dabei ist der Effizienzverlust geringer, wenn die Werte näher am jeweiligen Gruppenmittelwert liegen. Dies bedeutet, dass eine abstandsorientierte Mikroaggregation zu einem geringeren Effizienzverlust führt. Aus diesem Grund führt eine getrennte Mikroaggregation zu einem geringeren Effizienzverlust als die gemeinsame Mikroaggregation, allerdings unter Inkaufnahme eines verzerrten Schätzers. Sind die Variablen allerdings sehr stark korreliert, so nähert sich die getrennte Mikroaggregation der gemeinsamen Mikroaggregation an, damit sinkt auch der Bias des Schätzers.

b) Abstandsorientierte Mikroaggregation unter Einbeziehung der abhängigen Variablen

Wie bereits erwähnt, gelten diese Ergebnisse nicht, falls das Gewichtungsschema der Aggregation von der abhängigen Variablen abhängt ($\mathbf{D} = \mathbf{D}(\mathbf{y})$ oder $\mathbf{D} = \mathbf{D}(\mathbf{y}, \mathbf{X})$). Der Hauptunterschied zu der bisher betrachteten Situation besteht darin, dass die Aggregationsmatrix, beziehungsweise das zur Gruppenbildung verwendete Abstandsmaß, nun vom Fehlerterm des Modells abhängt (Schmid et al. 2005). In diesem Fall ist der KQ-Schätzer nichtlinear mit zunächst unbekannten Verteilungseigenschaften (Lechner und Pohlmeier 2003).

Schmid et al. (2005) untersuchen diesen Fall für das einfache lineare Modell mit einem normalverteilten Regressor X mit Erwartungswert μ_x und Varianz σ_x^2 sowie einem standardnormalverteilten Störterm mit Varianz σ_u^2. Hieraus folgt, dass die abhängige Variable Y ebenfalls normalverteilt ist mit Mittelwert $\mu_y = \alpha + \beta\mu_x$ und Varianz $\sigma_y^2 = \beta^2\sigma_x^2 + \sigma_u^2$. Dabei gehen sie von n unabhängigen und identisch verteilten Beobachtungen (x_i, y_i) mit $i = 1, \ldots, n$ aus. Die Gruppenbildung bei der Mikroaggregation erfolgt ausschließlich nach der abhängigen Variablen. Es gilt also $\mathbf{D} = \mathbf{D}(\mathbf{y})$.

Für das zu schätzende Modell ergibt sich in diesem Fall:

$$\mathbf{D}(\mathbf{y})\mathbf{y} = \alpha^d + \mathbf{D}(\mathbf{y})\mathbf{x}\beta^d + \mathbf{u}^d. \tag{23.57}$$

Für den KQ-Schätzer des Parameters β^d gilt im Fall der Einfachregression:

$$\hat{\beta}^d = \frac{s_{x^a y^a}}{s^2_{x^a}}, \tag{23.58}$$

wobei

$$s_{x^a y^a} = \frac{1}{n}\sum_{i=1}^{n}(x_i^a - \bar{x}^a)(y_i^a - \bar{y}^a) \tag{23.59}$$

die empirische Kovarianz zwischen den beiden mikroaggregierten Variablen x^a und y^a sowie

$$s^2_{x^a} = \frac{1}{n}\sum_{i=1}^{n}(x_i^a - \bar{x}^a)^2 \tag{23.60}$$

die empirische Varianz der Variablen x^a darstellt.

Schmid et al. (2005) untersuchen nun das asymptotische Verhalten von $\hat{\beta}^d$, indem sie zunächst das asymptotische Verhalten von $s_{x^a y^a}$ und $s^2_{x^a}$ betrachten. Sie zeigen, dass Folgendes gilt:

- $s^2_{y^a}$ konvergiert in Wahrscheinlichkeit gegen σ^2_y.
- $s^2_{x^a}$ konvergiert in Wahrscheinlichkeit gegen $\sigma^2_{x^a} = \frac{\sigma^2_x}{f(\rho)}$ mit $f(\rho) = \frac{1}{\frac{1}{A}+(1-\frac{1}{A})\rho^2}$.
- $s_{x^a y^a}$ konvergiert in Wahrscheinlichkeit gegen $\sigma_{xy} = \rho\sigma_x\sigma_y$.

Daraus folgt, dass der KQ-Schätzer $\hat{\beta}^d$ in Wahrscheinlichkeit gegen $\beta f(\rho)$ konvergiert.

$$\hat{\beta}^d = \frac{s_{x^a y^a}}{s^2_{x^a}} \rightharpoonup \frac{\sigma_{xy}}{\sigma^2_x / f(\rho)} = \beta f(\rho). \tag{23.61}$$

Hieraus leiten Schmid et al. (2005) die folgenden wesentlichen Ergebnisse her:

- $|\beta|$ wird durch $\hat{\beta}^d$ systematisch überschätzt, auch für großes n. Lediglich für $\beta = 0$ ist $\hat{\beta}^d$ ein konsistenter Schätzer für β.
- Wenn X und Y unkorreliert sind ($\rho = 0$), gilt $f(\rho) = A$. Wenn $|\rho| \rightharpoonup 1$, gilt $f(\rho) \rightharpoonup 1$. Das bedeutet, dass für eine hohe Korrelation zwischen X und Y die Verzerrung verschwindet. Eine hohe Korrelation bedeutet nämlich, dass die Gruppenbildung nach Y zum gleichen Ergebnis führt wie die Gruppenbildung nach X. Eine gemeinsame Mikroaggregation beider Variablen nach der erklärenden Variablen X führt aber zu einem erwartungstreuen beziehungsweise konsistenten Schätzer.

Da für ρ^2 auch gilt:

$$\rho^2 = \frac{\beta^2}{\beta^2 + \sigma_u^2/\sigma_x^2}, \tag{23.62}$$

schreiben Schmid et al. (2005) die asymptotische relative Verzerrung des Schätzers $\hat{\beta}^d$ auch als

$$f(\rho) = \frac{A(\beta^2 + v^2)}{A\beta^2 + v^2} \tag{23.63}$$

mit $v^2 = \sigma_u^2/\sigma_x^2$.

Damit stellen sie die asymptotische Verzerrung des Schätzers dar als

$$\begin{aligned} Bias(\hat{\beta}^d) &= \beta(f(\rho) - 1) \\ &= (A-1)\frac{\beta}{1 + \frac{A}{v^2}\beta^2}. \end{aligned} \tag{23.64}$$

Man kann erkennen, dass die Verzerrung für kleine Werte von β zunächst proportional mit β steigt, für große Werte von β hingegen gegen Null strebt.

Für den „naiven“ Schätzer für das Absolutglied α ergibt sich:

$$\hat{\alpha}^d = \bar{y}^a - \hat{\beta}^d \bar{x}^a. \tag{23.65}$$

Daraus folgt für die asymptotische Verzerrung des Schätzers:

$$\begin{aligned} Bias(\hat{\alpha}^d) &= plim(\bar{y}^a - \hat{\beta}^d \bar{x}^a) - \alpha \\ &= (\mu_y - \beta f(\rho)\mu_x) - (\mu_y - \beta\mu_x) \\ &= -Bias(\hat{\beta}^d)\mu_x. \end{aligned} \tag{23.66}$$

Für den Wahrscheinlichkeitsgrenzwert des „naiven“ Schätzers für die Varianz der Residuen $\sigma_u^2 = s_{y^a}^2 - \hat{\beta}^{d^2} s_{x^a}^2$ ergibt sich (Schmid et al. 2005):

$$\begin{aligned} plim\, \hat{\sigma}_u^{d^2} &= \sigma_y^2 - \beta^2 f(\rho)^2 \frac{\sigma_x^2}{f(\rho)} \\ &= \beta^2\sigma_x^2 + \sigma_u^2 - f(\rho)\beta^2\sigma_x^2 \\ &= (1 - f(\rho))\beta^2\sigma_x^2 + \sigma_u^2 \\ &= \frac{v^2 + \beta^2}{v^2 + A\beta^2}\sigma_u^2 \\ &= \frac{1}{A} f(\rho)\sigma_u^2. \end{aligned} \tag{23.67}$$

Aus diesen Ergebnissen leiten Schmid et al. (2005) korrigierte Schätzer für die Parameter des Modells ab. Dabei ist $r_{x^a y^a}$ der empirische Korrelationskoeffizient zwischen den Variablen X und Y. Und es gilt $plim\, r_{x^a y^a} = \rho^2 f(\rho)$.

Ein konsistenter Schätzer für ρ^2 ergibt sich somit als:

$$\begin{aligned} \hat{\rho}^{d^{Corr^2}} &= \frac{r_{x^a y^a}^2}{f(\hat{\rho}^{d^{Corr}})} \\ &= r_{x^a y^a}^2 \left(\frac{1}{A} + \left(1 - \frac{1}{A}\right) \hat{\rho}^{d^{Corr^2}} \right) \\ &= \frac{r_{x^a y^a}^2}{A} + r_{x^a y^a}^2 \left(1 - \frac{1}{A}\right) \hat{\rho}^{d^{Corr^2}} \\ &= \frac{r_{x^a y^a}^2}{A - r_{x^a y^a}^2 (A-1)}. \end{aligned} \tag{23.68}$$

Ein konsistenter Schätzer für β ergibt sich als:

$$\begin{aligned} \hat{\beta}^{d^{Corr}} &= \frac{\hat{\beta}^d}{f(\hat{\rho}^{d^{Corr}})} \\ &= \frac{\hat{\beta}^d \left(1 + (A-1)\hat{\rho}^{d^{Corr^2}}\right)}{A} \\ &= \frac{\hat{\beta}^d}{A - (A-1) r_{x^a y^a}^2}. \end{aligned} \tag{23.69}$$

Ein konsistenter Schätzer für das Absolutglied α ergibt sich als:

$$\hat{\alpha}^{d^{Corr}} = \hat{\alpha}^d + \left(\hat{\beta}^d - \hat{\beta}^{d^{Corr}}\right) \bar{x}^a. \tag{23.70}$$

Und schließlich kann die Residuenvarianz σ_u^2 konsistent durch

$$\hat{\sigma}_u^{d^{Corr^2}} = \frac{A\left(s_{y^a}^2 - \hat{\beta}^{d^2} s_{x^a}^2\right)}{f(\hat{\rho}^{d^{Corr}})}. \tag{23.71}$$

geschätzt werden.

Wird im Unterschied zu dem von Schmid et al. (2005) betrachteten Fall ausschließlich die erklärende Variable – allerdings abstandsorientiert nach der abhängigen Variablen – mikroaggregiert, so gelten dieselben Ergebnisse wie für den Fall, dass beide Variablen in dieser Weise mikroaggreggiert werden. Der Grund hierfür besteht darin, dass die zunächst von Schmid et al. (2005) durchgeführten Herleitungen über die Wahrscheinlichkeitsgrenzwerte von $s_{y^a}^2$, $s_{x^a}^2$ und $s_{x^a y^a}$ erhalten bleiben. $s_{y^a}^2$ konvergiert nun (da nicht anonymisiert) ohnehin gegen σ_y^2. Das Ergebnis für den Wahrscheinlichkeitsgrenzwert von $s_{x^a}^2$ bleibt unverändert. Somit konvergiert auch $s_{x^a y^a}$ in diesem Fall gegen σ_{xy}.

Schmid et al. (2005) beschränken sich auf den Fall eines Regressors. Weitaus komplizierter wird das Problem, wenn von mehreren Regressoren mit unterschiedlichen Korrelationen zwischen den einzelnen Regressoren und der abhängigen Variablen Y ausgegangen wird. Beschränkt man sich dabei darauf, dass die Gruppenbildung bei der Mikroaggregation ausschließlich nach der abhängigen Variablen vorgenommen wird, so dürften die Ergebnisse in ähnlicher Weise gelten, wie sie für einen Regressor hergeleitet wurden.

Allerdings kann bei der abstandsorientierten Mikroaggregation die Gruppenbildung auch nach allen Variablen gemeinsam erfolgen. Dann dürfte die Stärke der Verzerrung von den Korrelationen zwischen den Variablen sowie von der Beschaffenheit des Parametervektors abhängen. Zudem dürfte der Einfluss der abhängigen Variablen und somit auch die Verzerrung mit der Anzahl der in die Gruppenbildung einbezogenen Variablen zurückgehen.

c) Bootstrap-Mikroaggregation

Wird ausschließlich die abhängige Variable mikroaggregiert, so kann aus den Herleitungen für die gewöhnliche Mikroaggregation auch für die Bootstrap-Mikroaggreagtion eine Verzerrung des Schätzers gefolgert werden.

Werden ausschließlich die Regressoren mikroaggregiert, so gilt für das zu schätzende Modell:

$$\mathbf{y} = \mathbf{X}^{BS}\boldsymbol{\beta}^e + \mathbf{u}^e \tag{23.72}$$

oder

$$\mathbf{y} = \mathbf{D}_{BS}\mathbf{X}\boldsymbol{\beta}^{e} + \mathbf{u}^{e}. \tag{23.73}$$

Da es sich bei der Aggregationsmatrix im Fall der Bootstrap-Aggregation um eine stochastische Matrix handelt, hat man es in diesem Fall mit stochastischen Regressoren zu tun.

Für den Schätzer ergibt sich:

$$\hat{\boldsymbol{\beta}}^{e} = \left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{y}. \tag{23.74}$$

Für seinen Erwartungswert gilt:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}^{e}\right) &= E\left[\left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right)\right] \\ &= \left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{X}\boldsymbol{\beta} \\ &\quad + E\left[\left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{u}\right] \\ &= \left(\mathbf{X}'\mathbf{D}'_{BS}\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}'_{BS}\mathbf{X}\boldsymbol{\beta}. \end{aligned} \tag{23.75}$$

Der Schätzer ist somit in diesem Fall nicht erwartungstreu, weil die Aggregationsmatrix nicht symmetrisch idempotent ist.

Demgegenüber gilt im Falle der Bootstap-Mikroaggreagtion aller Variablen für das einfache lineare Modell:

$$\mathbf{y}^{BS} = \mathbf{X}^{BS}\boldsymbol{\beta} + \mathbf{u}^{BS} \tag{23.76}$$

oder

$$\mathbf{D}_{BS}\mathbf{y} = \mathbf{D}_{BS}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}_{BS}\mathbf{u}. \tag{23.77}$$

Als Schätzer für $\boldsymbol{\beta}$ erhält man:

$$\hat{\boldsymbol{\beta}}^{BS} = \left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{y}. \tag{23.78}$$

Für den Erwartungswert des Schätzers gilt:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}^{BS}\right) &= E\left[\left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\left(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}\right)\right] \\ &= \left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\boldsymbol{\beta} \\ &\quad + E\left[\left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{u}\right] = \boldsymbol{\beta}. \end{aligned} \tag{23.79}$$

Die Erwartungstreue ergibt sich wegen der Unkorreliertheit von $\mathbf{D}_{BS}$ und $\mathbf{u}$.

Für die bedingte Varianz-Kovarianzmatrix des Schätzers gilt somit:

$$\begin{aligned} var_{u|\mathrm{D}_{BS}}\left(\hat{\boldsymbol{\beta}}^{BS}\right) &= E_{u|\mathrm{D}_{BS}}\left[\left(\left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{u}\right)\right. \\ &\quad \left.\times\left(\left(\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{X}\right)^{-1}\left(\mathbf{D}_{BS}\mathbf{X}\right)'\mathbf{D}_{BS}\mathbf{u}\right)'\right]. \end{aligned} \tag{23.80}$$

Eine wesentliche Vereinfachung dieses Ausdrucks würde sich ergeben, wenn $\mathbf{D}_{BS}$ symmetrisch idempotent wäre.

Eine genauere Betrachtung verdient der Störterm $\mathbf{D}_{BS}\mathbf{u}$. Für seinen Erwartungswert gilt:

$$E\left[\mathbf{D}_{BS}\mathbf{u}\right] = E\left[\mathbf{D}_{BS}\right]E\left[\mathbf{u}\right] = \mathbf{0}. \tag{23.81}$$

Für seine Varianz ergibt sich:

$$\begin{aligned} E\left[\left(\mathbf{D}_{BS}\mathbf{u}\right)\left(\mathbf{D}_{BS}\mathbf{u}\right)'\right] &= E_{D_{BS}}\left[E_{\mathrm{u}|D_{BS}}\left[\mathbf{D}_{BS}\mathbf{u}\mathbf{u}'\mathbf{D}_{BS}'|\mathbf{D}_{BS}\right]\right] \\ &= E_{D_{BS}}\left[\mathbf{D}_{BS}\sigma_u^2\mathbf{I}_n\mathbf{D}_{BS}'\right] \\ &= {\sigma_u}^2 E_{D_{BS}}\left[\mathbf{D}_{BS}\mathbf{D}_{BS}'\right]. \end{aligned} \tag{23.82}$$

Wir haben es somit nicht mit einer skalaren Varianz-Kovarianzmatrix, sondern mit einer heteroskedastischen zu tun. Lechner und Pohlmeier (2003) empfehlen bei der Schätzung der Standardfehler die Verwendung des heteroskedastie-robusten Varianz-Kovarianzschätzers, um diese inhärente Heteroskedastie zu berücksichtigen.

23.1.2 Mikroaggregationsverfahren in nichtlinearen Modellen

Bei der Mikroaggregation handelt es sich um eine lineare Transformation g. Einem nichtlinearen Modell liegt ein nichtlinearer Zusammenhang und somit eine nichtlineare Transfor-

mation zugrunde. Lineare und nichtlineare Transformationen sind jedoch nicht vertauschbar.

Gilt für eine lineare Transformation $Y_1 = f_1(X)$, so gilt im Fall der gemeinsamen Mikroaggregation der beiden Variablen X und Y:

$$f_1(X^a) = f_1(g(X)) = g(f_1(X)) = g(Y_1^a). \tag{23.83}$$

Für nichtlineare Zusammenhänge $Y_2 = f_2(X)$ gilt jedoch:

$$f_2(X^a) = f_2(g(X)) \neq g(f_2(X)) = g(Y_2^a). \tag{23.84}$$

Damit werden nichtlineare Zusammenhänge durch die Mikroaggregation zerstört. Beispielsweise gilt $\frac{1}{A}\sum_{i=1}^{A} x_i^2 \neq \left(\frac{1}{A}\sum_{i=1}^{A} x_i\right)^2$ oder $\frac{1}{A}\left(\sum_{i=1}^{A} ln(x_i)\right) \neq ln\left(\frac{1}{A}\left(\sum_{i=1}^{A} x_i\right)\right)$.

Damit werden auch die Schätzer in nichtlinearen Modellen durch die gemeinsame Mikroaggregation grundsätzlich verzerrt. Die getrennte Mikroaggregation von Variablen führt bereits im linearen Modell zu verzerrten Schätzern. Somit werden auch die Schätzer in nichtlinearen Modellen bei einer getrennten Mikroaggregation verzerrt.

Umfang und Art der Verzerrung ist dabei stark von der Art der verwendeten Mikroaggregation sowie der Art des nichtlinearen Zusammenhangs und der konkreten Modellspezifikation abhängig. Die Simulationsexperimente in Unterabschnitt 23.2.2 zeigen unter anderem hierzu anschauliche Beispiele.

23.2 Monte-Carlo-Simulationen

23.2.1 Mikroaggregationsverfahren in linearen Modellen

a) Mikroaggregationsverfahren im linearen Modell mit nicht transformierten Variablen

Den Simulationsexperimenten liegt das gleiche lineare Modell und das gleiche Vorgehen zugrunde, wie es auch zur Untersuchung der Auswirkungen von stochastischen Überlagerungen in linearen Modellen verwendet wurde (vgl. Unterabschnitt 22.2.1).

Analog zum Vorgehen bei den stochastischen Überlagerungen sind in den Tabellen 23.1 bis 23.4 die Ergebnisse der Monte-Carlo-Simulationen für verschiedene Varianten der Mikroaggregation dargestellt. Die Teststatistiken werden unkorrigiert ausgewiesen. Tabelle 23.1

enthält zunächst die durchschnittlichen Schätzergebnisse aus 1.000 Replikationen mit jeweils 1.000 Beobachtungen, bei denen alle Variablen mikroaggregiert werden.[31] Für die Gruppengröße wird $A = 5$ gewählt.

Werden alle Variablen gemeinsam zufällig oder abstandsorientiert nach dem Regressor X_1 mikroaggregiert, so ergeben sich unverzerrte Parameterschätzer. Die t-Werte werden durch die Mikroaggregation überhöht ausgewiesen, weil eine verzerrte Schätzung für die Residuenvarianz zugrundeliegt (vgl. Unterabschnitt 23.1.1). Die korrekten t-Werte lassen sich durch die Multiplikation der hier ausgewiesenen Teststatistiken mit dem Faktor $f = \sqrt{\frac{M-K}{n-K}}$ ermitteln. Dabei ist M die Anzahl der durch die Mikroaggregation gebildeten Gruppen. In diesem Fall ergibt sich mit 1.000 Beobachtungen, $K = 4$ Regressoren (inklusive Konstante) und $M = 200$ Gruppen für den Korrekturfaktor $f \approx 0,44$. Korrigiert man die t-Werte derart, so erhält man für die Konstante einen t-Wert, der dem t-Wert aus der Originalschätzung entspricht. Die anderen t-Werte liegen dann deutlich unterhalb der Originalwerte.

Die getrennte zufällige Mikroaggregation und die abstandsorientierte Mikroaggregation nach der abhängigen Variablen führen erwartungsgemäß zu verzerrten Parameterschätzungen. Etwas überraschend scheint auf den ersten Blick zu sein, dass die getrennte abstandsorientierte Mikroaggregation zu Parameterschätzungen führt, die nicht signifikant von den theoretischen Parametern und den Schätzern mit den Originaldaten abweichen. Dieses Phänomen kann insbesondere auf zwei Ursachen zurückgeführt werden. Zum einen werden bei einer großen Dichte der Datenpunkte die einzelnen Werte durch diese Form der Mikroaggregation nur geringfügig verändert, zum anderen nähert sich bei einer hohen Korrelation zwischen den Variablen die getrennte Mikroaggregation der gemeinsamen Mikroaggregation an. Falls die Korrelation zwischen abhängiger und erklärender Variablen gegen Eins geht, führt außerdem – wie von Schmid et al. (2005) gezeigt wurde – auch die abstandsorientierte Mikroaggregation, bei der die Gruppenbildung nach der abhängigen Variablen erfolgt, zu einem unverzerrten Schätzer.

Die Ergebnisse in Tabelle 23.2 bestätigen, dass die zufällige Mikroaggregation zu verzerrten Parameterschätzern führt, sofern lediglich die abhängige Variable mikroaggregiert wird und dass unverzerrte Parameterschätzer entstehen, wenn die Regressoren gemeinsam und zufällig mikroaggregiert werden (vgl. auch Unterabschnitt 23.1.1).

Werden lediglich die abhängige und zwei der drei erklärenden Variablen mikroaggregiert (Tabelle 23.3), so lassen sich keine unverzerrten Schätzer erzeugen, wenn die Variablen zufällig gemeinsam oder getrennt mikroaggregiert werden. Wohl aber können nahezu unverzerrte Parameterschätzer beobachtet werden, wenn die Variablen getrennt abstandsorientiert mikroaggregiert werden. Auch die t-Werte stimmen fast mit den originalen Teststatistiken überein.

31) Wiederum wurden auch Simulationen mit 1.000 Beobachtungen und 100 Replikationen sowie mit 10.000 Beobachtungen und 100 Replikationen getestet. Auf die Darstellung der Ergebnisse wird verzeichtet, da sie im Wesentlichen mit den hier dargestellten übereinstimmen.

In Tabelle 23.4 sind die Ergebnisse für den Fall dargestellt, dass lediglich zwei der drei Regressoren mikroaggregiert werden, die abhängige Variable verbleibt im Originalzustand. Wiederum lassen sich unverzerrte Schätzer für den Fall der getrennten abstandsorientierten Mikroaggregation und verzerrte Schätzer im Fall der getrennten zufälligen Mikroaggregation beobachten. Im Unterschied zu den in Unterabschnitt 23.1.1 hergeleiteten theoretischen Ergebnissen scheinen die Parameterschätzer für die gemeinsame stochastische Mikroaggregation hier unverzerrt zu sein. Die Erwartungswerte der beiden Teilschätzer wurden in Unterabschnitt 23.1.1 hergeleitet und in den Gleichungen (23.49) und (23.50) dargestellt:

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}_1^{c1}\right) &= \boldsymbol{\beta}_1 + \left(\mathbf{X}_1'\mathbf{X}_1 - \mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2\right)^{-1}\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\right)^{-1} \\ &\quad \times \left(\mathbf{X}_1'\mathbf{X}_2 - \mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\right)\boldsymbol{\beta}_2 \end{aligned} \tag{23.85}$$

und

$$\begin{aligned} E\left(\hat{\boldsymbol{\beta}}_2^{c1}\right) &= \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2\right)^{-1} \\ &\quad \times \left(\mathbf{X}_2'\mathbf{D}_2\mathbf{X}_2 - \mathbf{X}_2'\mathbf{D}_2\mathbf{X}_1\left(\mathbf{X}_1'\mathbf{X}_1\right)^{-1}\mathbf{X}_1'\mathbf{X}_2\right)\boldsymbol{\beta}_2. \end{aligned} \tag{23.86}$$

Offenbar unterscheiden sich $\mathbf{X}_1'\mathbf{X}_2$ und $(\mathbf{X}_1'\mathbf{D}_2\mathbf{X}_2)$ nur geringfügig, so dass die Verzerrung sehr gering ausfällt. Das Phänomen bleibt auch erhalten, wenn die Gruppengröße auf 50 erhöht wird.

Tabelle 23.1: MC-Simulationen – Lineares Modell, unterschiedliche Varianten der Mikroaggregation, alle Variablen anonymisiert, 1.000 Replikationen

	Original	Mikro gemeinsam zufällig			Mikro getrennt zufällig			Mikro getrennt abstandsorientiert		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	1,001	*	*	0,194			0,999	*	*
(t-Werte)	(28,5)	(28,62)			(3,47)			(28,46)		
X_2	-0,999	-0,998	*	*	-0,159			-0,999	*	*
(t-Werte)	(-30,91)	(-30,97)			(-2,86)			(-30,85)		
X_3	0,499	0,496	*	*	0,146			0,500	*	*
(t-Werte)	(14,45)	(14,40)			(2,61)			(14,45)		
Konst.	1,001	1,000	*	*	1,000	*	*	1,000	*	*
(t-Werte)	(31,66)	(70,97)			(40,32)			(31,55)		

	Original	Mikro gemeinsam nach X_1			Mikro gemeinsam nach Y		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	1,000	*	*	1,343		
(t-Werte)	(28,57)	(48,44)			(39,06)		
X_2	-0,999	-1,001	*	*	-1,337		
(t-Werte)	(-30,91)	(-31,09)			(-43,82)		
X_3	0,499	0,501	*	*	0,669		
(t-Werte)	(14,45)	(14,52)			(17,48)		
Konst.	1,001	1,002	*		1,002	*	*
(t-Werte)	(31,66)	(71,14)			(61,63)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 23.2: MC-Simulationen – Lineares Modell, zufällige Mikroaggregation, nur abhängige Variable anonymisiert sowie nur Regressoren anonymisiert, 1.000 Replikationen

	Original	Mikro zufällig, nur abhängige Variable anonymisiert			Mikro gemeinsam zufällig, nur Regressoren anonymisiert		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,001	0,200			1,001	*	*
(t-Werte)	(28,57)	(7,72)			(7,67)		
X_2	-0,999	-0,199			-0,998	*	*
(t-Werte)	(-30,91)	(-8,34)			(-8,29)		
X_3	0,499	0,100			0,496	*	*
(t-Werte)	(14,45)	(3,91)			(3,86)		
Konst.	1,001	1,000	*	*	1,000	*	*
(t-Werte)	(31,66)	(42,98)			(19,01)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 23.3: MC-Simulationen – Lineares Modell, unterschiedliche Varianten der Mikroaggregation, abhängige und zwei erklärende Variablen anonymisiert, 1.000 Replikationen

	Original	Mikro gemeinsam zufällig			Mikro getrennt zufällig			Mikro getrennt abstandsorientiert		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original-wert	theor. Wert		Original-wert	theor. Wert		Original-wert	theor. Wert
X_1	1,001	0,168			0,200			1,000	*	*
(t-Werte)	(28,57)	(9,03)			(8,09)			(28,51)		
X_2	-0,999	-0,872			-0,190			-1,001	*	*
(t-Werte)	(-30,91)	(-21,18)			(-3,48)			(-30,93)		
X_3	0,499	0,822			0,086			0,500	*	*
(t-Werte)	(14,45)	(19,60)			(1,57)			(14,45)		
Konst.	1,001	0,999	*	*	0,999	*	*	1,000	*	*
(t-Werte)	(31,66)	(54,67)			(41,37)			(31,58)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 23.4: MC-Simulationen – Lineares Modell, unterschiedliche Varianten der Mikroaggregation, zwei Regressoren anonymisiert, 1.000 Replikationen

	Original	Mikro gemeinsam zufällig			Mikro getrennt zufällig			Mikro getrennt abstandsorientiert		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Original-wert	theor. Wert		Original-wert	theor. Wert		Original-wert	theor. Wert
X_1	1,001	1,001	*	*	1,007			1,002	*	*
(t-Werte)	(28,57)	(22,38)			(22,39)			(28,55)		
X_2	-0,999	-0,998	*	*	-0,960			-1,000	*	*
(t-Werte)	(-30,91)	(-10,03)			(-9,68)			(-30,92)		
X_3	0,499	0,504	*	*	0,415			0,500	*	*
(t-Werte)	(14,45)	(5,02)			(4,15)			(14,46)		
Konst.	1,001	1,000	*	*	1,000	*	*	1,000	*	*
(t-Werte)	(31,66)	(22,70)			(22,64)			(31,56)		

* Übereinstimmung zum Signifikanzniveau von 5%

b) Mikroaggregationsverfahren im linearen Modell mit nichtlinearen Transformationen

Analog zum Vorgehen bei den Simulationsexperimenten zur Untersuchung der Wirkung von stochastischen Überlagerungen in linearen Modellen wird nun noch der Fall betrachtet, dass die in das Modell eingehenden Variablen logarithmiert werden und somit nach der Anonymisierung einer nichtlinearen Transformation unterzogen werden. Dabei wird auf das gleiche Modell zurückgegriffen wie in Unterabschnitt 22.2.1. Die Anzahl der Replikationen beträgt wiederum 1.000, ebenso die Anzahl der Beobachtungen.

Alle Variablen werden mikroaggregiert. Es werden nur diejenigen Varianten der Mikroaggregation betrachtet, die bei Mikroaggregation aller Variablen im einfachen linearen Modell zu guten Schätzergebnissen geführt haben: die getrennte abstandsorientierte Mikroaggregation, die gemeinsame abstandsorientierte Mikroaggregation nach einer der Regressorvariablen und die gemeinsame stochastische Mikroaggregation. Die Ergebnisse sind in Tabelle 23.5 dargestellt.

Man erkennt, dass die nichtlineare Transformation der Variablen dazu führt, dass die beiden Varianten der gemeinsamen Mikroaggregation nun nicht mehr zu erwartungstreuen Schätzern führen. Damit werden die theoretischen Überlegungen aus Abschnitt 23.1.2 bestätigt. Demgegenüber ist die Verzerrung bei der getrennten abstandsorientierten Mikroaggregation weiterhin sehr gering.

Tabelle 23.5: MC-Simulationen – Lineares Modell mit logarithmierten Variablen, verschiedene Varianten der Mikroaggregation, 1.000 Replikationen

	Original	Mikro getrennt			Mikro gemeinsam nach X_1			Mikro gemeinsam stochastisch		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert		Originalwert	theor. Wert
log X_1	1,000	0,998	*	*	0,955			0,884		
(t-Werte)	(28,53)	(28,37)			(30,54)			(16,77)		
log X_2	-0,998	-0,998	*	*	-0,611			-0,662		
(t-Werte)	(-30,96)	(-30,83)			(-13,42)			(-13,28)		
log X_3	0,499	0,501	*	*	0,426			0,473		
(t-Werte)	(14,46)	(14,45)			(8,61)			(9,04)		
Konst.	1,002	1,002	*	*	1,801			1,734		
(t-Werte)	(31,69)	(31,58)			(57,14)			(47,46)		

* Übereinstimmung zum Signifikanzniveau von 5%

23.2.2 Mikroaggregationsverfahren in nichtlinearen Modellen

Zur Untersuchung der Wirkung von Mikroaggregationsverfahren in nichtlinearen Modellen wird ein binäres Probit-Modell mit drei metrischen Regressoren geschätzt, wie es bereits zur Untersuchung der Wirkung von stochastischen Überlagerungen in Unterabschnitt 22.2.2 verwendet wurde.

Die Mikroaggregationsverfahren werden in einem Fall auf alle Regressoren (Tabelle 23.6), im anderen lediglich auf die beiden Regressoren X_2 und X_3 (Tabelle 23.7) angewendet. Getestet werden jeweils die gemeinsame stochastische, die getrennte stochastische, die gemeinsame abstandsorientierte (nach der Variablen X_1) und die getrennte abstandsorientierte Mikroaggregation.

Tabelle 23.6: MC-Simulationen – Probit-Modell, unterschiedliche Varianten der Mikroaggregation, alle Regressoren mikroaggregiert, 1.000 Replikationen

	Original	Mikro gemeinsam stochastisch			Mikro getrennt stochastisch			Mikro getrennt abstandsorientiert			Mikro gemeinsam nach X_1		
			Übereinstimmung zum Signifkanz-niveau von 5% mit dem			Übereinstimmung zum Signifkanz-niveau von 5% mit dem			Übereinstimmung zum Signifkanz-niveau von 5% mit dem			Übereinstimmung zum Signifkanz-niveau von 5% mit dem	
			Original	theor. Wert		Original	theor. Wert		Original	theor. Wert		Original	theor. Wert
X_1	1,009	0,615			0,589			1,007	*		0,730		
(t-Werte)	(12,88)	(5,53)			(5,86)			(12,87)			(9,67)		
X_2	-1,008	-0,609			-0,489			-1,008	*		-0,727		
(t-Werte)	(-13,51)	(-5,92)			(-4,90)			(-13,49)			(-6,48)		
X_3	0,504	0,296			0,450			0,505	*		0,356		
(t-Werte)	(7,77)	(2,75)			(4,53)			(7,78)			(3,05)		
Konst.	1,005	0,607			0,609			1,009	*		0,7245		
(t-Werte)	(14,91)	(13,68)			(13,72)			(14,94)			(14,41)		

* Übereinstimmung zum Signifikanzniveau von 5%

Tabelle 23.7: MC-Simulationen – Probit-Modell, unterschiedliche Varianten der Mikroaggregation, Teil der Regressoren mikroaggregiert, 1.000 Replikationen

	Original	Mikro gemeinsam stochastisch			Mikro getrennt stochastisch			Mikro getrennt abstandsorientiert			Mikro gemeinsam nach X_1		
			Übereinstimmung zum Signifkanz-niveau von 5% mit dem			Übereinstimmung zum Signifkanz-niveau von 5% mit dem			Übereinstimmung zum Signifkanz-niveau von 5% mit dem			Übereinstimmung zum Signifkanz-niveau von 5% mit dem	
			Original	theor. Wert		Original	theor. Wert		Original	theor. Wert		Original	theor. Wert
X_1	1,009	0,730			0,729			1,008	*		0,724		
(t-Werte)	(12,88)	(12,84)			(12,84)			(12,85)			(9,57)		
X_2	-1,008	-0,727			-0,695			-1,006	*		-0,723		
(t-Werte)	(-13,51)	(-6,56)			(-6,34)			(-13,49)			(-6,44)		
X_3	0,504	0,366			0,302			0,503	*	*	0,367		
(t-Werte)	(7,77)	(3,35)			(2,80)			(7,75)			(3,14)		
Konst.	1,005	0,725			0,725			1,007	*		0,727		
(t-Werte)	(14,91)	(14,40)			(14,43)			(14,94)			(14,44)		

* Übereinstimmung zum Signifikanzniveau von 5%

Man erkennt, dass nun im nichtlinearen Probit-Modell alle Varianten der gemeinsamen Mikroaggregation, ob stochastisch oder abstandsorientiert, zu verzerrten Schätzern führen, wie dies auch im Rahmen der theoretischen Überlegungen in Unterabschnitt 23.1.2 vermutet wurde. Dies gilt im Unterschied zur Schätzung linearer Modelle unabhängig davon, ob alle Regressoren oder nur ein Teil der Regressoren mikroaggregiert wird. Ebenso führen getrennt stochastisch mikroaggregierte Variablen in allen Fällen zu starken Verzerrungen analog zu ihrer Wirkung in linearen Modellen.

Gute Ergebnisse lassen sich wiederum bei Anwendung einer getrennten abstandsorientierten Mikroaggregation erzielen. Zwar sind die Schätzer ebenfalls geringfügig verzerrt, was sich auch daran erkennen lässt, dass die Übereinstimmung mit den theoretischen Parameterwerten bei t-Tests zum Signifikanzniveau von fünf Prozent überwiegend nicht gegeben ist. Allerdings sind die Verzerrungen sehr gering. Die Schätzergebnisse lassen sich fast originalgetreu replizieren. Dies kann auf die nur geringfügige Veränderung der Einzelwerte in Verbindung mit dem annähernden Erhalt der Korrelationsstruktur bei der getrennten abstandsorientierten Mikroaggregation zurückgeführt werden.

23.3 Praxisbeispiele

23.3.1 Mikroaggregationsverfahren in linearen Modellen

a) Das geschätzte Modell: Linearisierte Cobb-Douglas-Produktionsfunktion

Als Praxisbeispiel wird wiederum das in Abschnitt 21.1 eingeführte Modell einer linearisierten Cobb-Douglas-Produktionsfunktion mit den KSE-Daten für 1999 geschätzt. Bei den Untersuchungen zur Wirkung der Mikroaggregationsverfahren wird grundsätzlich der um Ausreißer bereinigte, aber ansonsten vollständige KSE-Datensatz verwendet.

b) Angewendete Varianten der Mikroaggregation

Für die nachfolgenden Untersuchungen der Wirkung von Mikroaggregationsverfahren bei der Schätzung der dargestellten linearisierten Cobb-Douglas-Produktionsfunktion wird die Mikroaggregation auf die Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe in den folgenden Varianten angewendet:

1. **Abstandsorientierte Mikroaggregationsverfahren**
   (a) Getrennte abstandsorientierte Mikroaggregation
      i. Einfache getrennte abstandsorientierte Mikroaggregation mit einer Gruppengröße von drei bis fünf: MA33G.

ii. Getrennte abstandsorientierte Mikroaggregation mit einer vorgegebenen Mindestabweichung zwischen den benachbarten Gruppen von 7 Prozent: MA33G_7p.

iii. Getrennte abstandsorientierte Mikroaggregation mit einer vorgegebenen Mindestabweichung zwischen den benachbarten Gruppen von 15 Prozent: MA33G_15p.

(b) Gemeinsame abstandsorientierte Mikroaggregation (nach allen Merkmalen)

i. Gemeinsam nach allen Merkmalen abstandsorientierte Mikroaggregation mit einer Gruppengröße von drei bis fünf: MA1G.

ii. Gemeinsam nach allen Merkmalen abstandsorientierte Mikroaggregation, dabei werden Gruppen mit jeweils vier Merkmalsträgern gebildet, von denen zwei den um die Standardabweichung innerhalb der Gruppe verminderten Gruppendurchschnitt als neuen Merkmalswert zugewiesen bekommen, die beiden anderen den um die Standardabweichung erhöhten Gruppendurchschnitt: MA1G_var.

(c) Teilweise gemeinsam abstandsorientierte Mikroaggregation

i. Teilweise gemeinsam abstandsorientierte Mikroaggregation, Einteilung der Variablen in elf Gruppen: Zunächst werden die Bravais-Pearson-Korrelationen zwischen den 33 stetigen Variablen berechnet. Anschließend werden die drei Variablen mit den höchsten Korrelationen zusammengefasst, dann die nächsten drei, bis schließlich noch drei stetige Variablen übrig bleiben. Auch diese werden zu einer Gruppe zusammengefasst. Für diese so entstandenen elf Gruppen von Variablen (Tabelle 23.8) werden jeweils die euklidischen Distanzen zwischen den Unternehmen bestimmt. Auf dieser Basis werden für die Gruppen von Variablen getrennt die Unternehmen zu Gruppen mit einer Besetzungsgröße von mindestens drei zusammengefasst und die Einzelwerte der Variablen durch den Durchschnitt der (mindestens) drei Unternehmen ersetzt: MA11G.

ii. Teilweise gemeinsam abstandsorientierte Mikroaggregation, Einteilung der Variablen in acht Gruppen: Wiederum werden die Korrelationskoeffizienten zwischen den Variablen bestimmt. Allerdings werden die Variablen nun so zusammengefasst, dass Gruppen mit unterschiedlicher Größe entstehen, wobei die Variablen in einer Gruppe möglichst stark miteinander korreliert sind. Die so entstandenen Gruppen sind in Tabelle 23.9 dargestellt. Als problematisch erweist sich eine „Restgruppe“ aus den Variablen „Tätige Inhaber“ und „Bestandsveränderungen“, da diese Variablen sowohl miteinander als auch mit den anderen Variablen nur sehr gering korreliert sind: MA8G.

2. **Stochastische Mikroaggregationsverfahren**

(a) Zufällige gemeinsame Mikroaggregation mit einer Gruppengröße von drei bis fünf: MA1G_stoch.

(b) Gemeinsame Bootstrap-Mikroaggregation mit einer Gruppengröße von drei: MA1G_BS.

Tabelle 23.8: Automatische Gruppierung der Merkmale der KSE für gruppierte Mikroaggregation MA11G

Gruppen-nummer	Merkmal 1	Merkmal 2	Merkmal 3
1	Bruttogehalts- und Lohnsumme	Angestellte und Arbeiter	Tätige Personen insgesamt
2	Bruttowertschöpfung zu Faktorkosten	Nettowertschöpfung zu Faktorkosten	Gesetzliche Sozialkosten
3	Gesamtumsatz	Bruttoproduktionswert	Umsatz aus eigenen Erzeugnissen
4	Kosten insgesamt	Sonstige Kosten	Sonstige Sozialkosten
5	Gesamtaufwendung für innerbetriebliche Forschung und Entwicklung	Anzahl der für FuE eingesetzten Lohn- und Gehaltsempfänger	Teilzeitbeschäftigte umgerechnet in Vollzeiteinheiten
6	Verbrauch an Rohstoffen	Mieten und Pachten	Umsatz aus Handelsware
7	Fremdkapitalzinsen	Teilzeitbeschäftigte	Kosten für Reparaturen
8	Endbestand Rohstoffe gemessen am Umsatz aus eigenen Erzeugnissen	Anfangsbestand Rohstoffe gemessen am Umsatz aus eigenen Erzeugnissen	Endbestand Erzeugnisse gemessen am Umsatz aus eigenen Erzeugnissen
9	Einsatz an Handelsware	Kosten für Lohnarbeiten	Energieverbrauch
10	Endbestand an Handelsware gemessen am Umsatz aus Handelswaren	Anfangsbestand an Handelsware gemessen am Umsatz aus Handelswaren	Anfangbestand Erzeugnisse gemessen am Umsatz aus eigenen Erzeugnissen
11	Tätige Inhaber	Bestandsveränderungen	Kosten für Leiharbeitnehmer

Bearbeitet man die Daten der Kostenstrukturerhebung des Jahres 1999 mit diesen Varianten der Mikroaggregation, so können sich im hier zu schätzenden Modell aus den folgenden Gründen verzerrte Schätzer ergeben:[32]

1. Der Schätzer ist durch die Art der Mikroaggregation grundsätzlich verzerrt: Dies gilt für alle hier getesteten Varianten der abstandsorientierten Mikroaggregation. Die Schätzer bei den beiden Varianten der stochastischen Mikroaggregation sind hingegen grundsätzlich unverzerrt.

2. Die Logarithmierung ist eine nichtlineare Transformation. Es gilt daher, dass der Logarithmus einer mikroaggregierten Variable nicht dem mikroaggregierten Logarithmus der Ursprungsvariable entspricht. Müssen also die Logarithmen aus den bereits mikroaggregierten Variablen berechnet werden, so ist der Schätzer in jedem Fall verzerrt.

32) Zusätzlich ist der KQ-Schätzer bei der Mikroaggregation nicht mehr effizient.

3. Durch die Mikroaggregation verändert sich die Zusammensetzung der in die Schätzung eingehenden Unternehmen. Hierfür gibt es mehrere Gründe: Zum einen nehmen weniger Unternehmen bei den Input- und Outputvariablen den Wert Null an. Damit sind auch weniger Logarithmen dieser Variablen nicht definiert. Zudem können im Fall der als Differenzen berechneten Merkmale vorher negative Werte positiv und vorher positive Werte negativ werden. Zuletzt verändert sich nach der Mikroaggregation auch die Ausreißerbereinigung.

Tabelle 23.9: Gruppierung der Merkmale der KSE für gruppierte Mikroaggregation MA8G

Gruppenname	Merkmale	Begründung
Umsatzgruppe	- Angestellte und Arbeiter - Tätige Personen insgesamt - Umsatz aus eigenen Erzeugnissen - Gesamtumsatz - Bruttoproduktionswert - Verbrauch an Rohstoffen - Bruttogehalts- und Lohnsumme - Gesetzliche Sozialkosten - Sonstige Sozialkosten - Kosten insgesamt - Bruttowertschöpfung zu Faktorkosten - Nettowertschöpfung zu Faktorkosten	Alle Merkmale sind mit mindesten 0,9 mit dem Gesamtumsatz korreliert (nach Pearson)
FuE-Gruppe	- Gesamtaufwendung für innerbetriebliche Forschung und Entwicklung - Anzahl der für FuE eingesetzten Lohn- und Gehaltsempfänger	mit 0,95 korreliert und erhält die Struktur FuE (ja/nein) in der Erhebung
Teilzeit	- Teilzeitbeschäftigte - Teilzeitbeschäftigte umgerechnet in Vollzeiteinheiten	mit 0,92 korreliert
Bestände Erzeugnisse	- Anfangbestand / Endbestand gemessen am Umsatz aus eigenen Erzeugnissen	mit 0,79 korreliert
Bestände Rohstoffe	- Anfangbestand / Endbestand gemessen am Umsatz aus eigenen Erzeugnissen	mit 0,92 korreliert
Handel	- Umsatz aus Handelsware - Anfangsbestand / Endbestand an Handelsware gemessen am Umsatz aus Handelsware - Einsatz an Handelsware	Erhält die Struktur Handel (ja/nein) in der Erhebung
Einzelkosten	- Energieverbrauch - Kosten für Leiharbeitnehmer - Kosten für Lohnarbeiten - Kosten für Reparaturen - Mieten und Pachten - Sonstige Kosten - Fremdkapitalzinsen	sind signifikant positiv korreliert
Rest	- Tätige Inhaber - Bestandsveränderungen	sind mit keinen der anderen Merkmale korreliert

Um die unterschiedlichen Ursachen für die Verzerrung der Schätzergebnisse möglichst getrennt voneinander analysieren zu können, werden für die stochastischen Verfahren sowie für einen Teil der abstandsorientierten Varianten jeweils drei Szenarien bei der Anonymisierung durchgeführt:

1. Lediglich die in der KSE von vornherein vorhandenen Variablen stehen in mikroaggregierter Form zur Verfügung. Es erfolgt keine „Bereinigung" des Datensatzes vor der Anonymisierung.
2. Es stehen die ursprünglichen Variablen der KSE in mikroaggregierter Form ergänzt um die mikroaggregierten Logarithmen der Inputfaktoren und des Outputs zur Verfügung. Es erfolgt keine „Bereinigung" des Datensatzes vor der Anonymisierung.
3. Auf Basis der Originaldaten werden die „Ausreißer", wie oben beschrieben, entfernt. Anschließend werden die mikroaggregierten Ausgangsvariablen ergänzt um die mikroaggregierten Logarithmen der Inputfaktoren und des Outputs zur Verfügung gestellt. Dabei werden von vornherein nur diejenigen Merkmalsträger berücksichtigt und somit auch anonymisiert, bei denen im Original alle Logarithmen definiert sind.

Für die gruppierte abstandsorientierte Mikroaggregation, bei der die Variablen teilweise gemeinsam mikroaggregiert werden (MA11G und MA8G), ist dieses Vorgehen nicht möglich, da das Hinzufügen weiterer Variablen (in diesem Fall der Logarithmen der Inputvariablen und des Outputs) eine Neuberechnung der Korrelationen und eine Veränderung der Gruppen nach sich ziehen würde. Auch für die Varianten MA33G_7p, MA33G_15p und MA1G_var wird nur ein Datensatz erstellt, bei dem Ausgangsvariablen für alle Unternehmen mikroaggregiert werden. Zusätzlich erfolgt für einen Teil der Verfahren auch zunächst eine Blockung des Datensatzes nach Wirtschaftszweigen (Zweisteller der WZ93) und Regionaltypen.[33] Anschließend wird die Mikroaggregation zellenweise durchgeführt. In Tabelle 23.10 ist dargestellt, für welche Varianten der Mikroaggregation welche Arten des Vorgehens getestet werden.

c) Auswirkungen unterschiedlicher Varianten der Mikroaggregation bei der Schätzung einer linearisierten Cobb-Douglas-Produktionsfunktion

In Tabelle 23.11 sind zunächst die Ergebnisse für verschiedene abstandsorientierte Mikroaggregationsvarianten dargestellt. Es handelt sich um die getrennte abstandsorientierte Mikroaggregation (MA33G), die beiden Varianten der gruppierten (teilweise gemeinsam abstandsorientierten) Mikroaggregation (MA11G und MA8G) und die gemeinsame abstandsorientierte Mikroaggregation (MA1G). Dabei wurden die Ausgangsvariablen mikroaggregiert und die Daten erst nach der Anonymisierung um Ausreißer bereinigt. Dieses

33) Herangezogen wird in diesem Fall der Regionsgrundtyp des BBR(BBR3).

Tabelle 23.10: Getestete Varianten der Mikroaggregation

Verfahren	Ausgangsvariablen anonymisiert		Ausgangsvariablen + Logarithmen anonymisiert		Ausgangsvariablen + Logarithmen anonymisiert, Datensatz vorher bereinigt	
	Über gesamten Datensatz	Nach Zellen	Über gesamten Datensatz	Nach Zellen	Über gesamten Datensatz	Nach Zellen
Abstandsorientierte Verfahren						
MA33G	×	×	×	×	×	×
MA33G_7p	×	-	-	-	-	-
MA33G_15p	×	-	-	-	-	-
MA11G	×	-	-	-	-	-
MA8G	×	-	-	-	-	-
MA1G	×	×	×	×	×	×
MA1G_var	×	-	-	-	-	-
Stochastische Verfahren						
MA1G_stoch	×	×	×	×	×	×
MA1G_BS	×	×	×	×	×	×

Szenario ist realistisch, wenn man bedenkt, dass spätere spezielle Nutzungen bei der Erstellung eines Scientific-Use-Files nicht oder nur sehr verallgemeinert berücksichtigt werden können.

Tabelle 23.11: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für abstandsorientiert mikroaggregierte KSE-Daten, Datensatz bereinigt, robuste Standardfehler

	MA33G		MA11G		MA8G		MA1G	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,434	(148,75)	0,435	(150,6)	0,474	(200,75)	0,492	(230,19)
Personalkosten	0,322	(100,45)	0,321	(101,3)	0,326	(109,42)	0,272	(112,06)
Externe Dienstleistungen	0,053	(39,57)	0,047	(36,37)	0,030	(24,87)	0,052	(49,25)
Sonstige Kosten	0,105	(46,89)	0,108	(48,52)	0,077	(36,78)	0,111	(57,68)
Kapitalkosten	0,070	(32,4)	0,073	(34,73)	0,083	(36,12)	0,058	(31,85)
Konst.	1,723	(97,16)	1,720	(94,29)	1,522	(94,59)	1,591	(135,54)
Anzahl Beobachtungen	15.049		15.349		15.531		15.616	
R^2	0,988		0,987		0,989		0,995	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	0,23	0,80	0	0,43	8,97	33,88	13,10	53,51
Personalkosten	0	0,24	0,31	0,61	1,24	8,67	15,53	11,29
Externe Dienstleistungen	1,92	0,66	9,62	7,48	42,31	36,73	0	25,29
Sonstige Kosten	0	0,04	2,86	3,43	26,67	21,59	5,71	22,96
Kapitalkosten	0	0,12	4,29	7,06	18,57	11,34	17,14	1,82
Konst.	0,35	0,74	0,17	3,67	11,36	3,36	7,34	38,48
Durchschn.	0,42	0,43	2,87	3,78	18,19	19,26	9,80	25,56

Es ist zu erkennen, dass sich die Ergebnisse in der Tendenz bei keinem der vier Verfahren

verändern. Es bestätigt sich die Erkenntnis aus den Simulationen des vorangegangenen Unterabschnitts, dass die getrennte abstandsorientierte Mikroaggregation nur zu einer sehr geringen Veränderung der Parameterwerte und der Teststatistiken führt und im Vergleich mit den anderen Verfahren am besten abschneidet. Die durchschnittliche Abweichung der Koeffizientenwerte ist mit 0,4% fast zu vernachlässigen. Ähnlich gut sind die Ergebnisse für das Verfahren, bei dem die Variablen in elf Gruppen gemeinsam mikroaggregiert wurden (MA11G). Die durchschnittliche Abweichung der Koeffizientenwerte ist mit rund 3% ebenfalls gering. Kein Parameter weicht um mehr als 10 Prozent vom Originalwert ab. Deutlich schlechter sind die Ergebnisse hingegen für das Verfahren, bei dem die Variablen in acht Gruppen mikroaggregiert wurden (MA8G). Die durchschnittliche Abweichung der Parameterwerte liegt immerhin bei rund 18%. Einzelne Parameterwerte weichen um deutlich über zehn Prozent vom Originalwert ab. Hinsichtlich der Abweichung der Parameterwerte schneidet dieses Verfahren somit sogar noch schlechter ab als die gemeinsame abstandsorientierte Mikroaggregation, bei der ebenfalls Abweichungen einzelner Parameterwerte von über zehn Prozent beobachtet werden. Hinsichtlich der Vergleichbarkeit der t-Werte muss erwähnt werden, dass diese bei der gemeinsamen abstandsorientierten Mikroaggregation eigentlich noch korrigiert werden müssten. Dann ergeben sich für diese Variante der Mikroaggregation deutlich geringere t-Werte.

Ausgehend von diesen Ergebnissen wird im Folgenden zunächst dargestellt, welche Maßnahmen zu einer Verbesserung beziehungsweise Verschlechterung der Resultate für die gemeinsame und die getrennte abstandsorientierte Mikroaggregation führen. Tabelle 23.12 zeigt zunächst für die getrennte abstandsorientierte Mikroaggregation, dass sich die Abweichungen von den Ergebnissen der Originalschätzung reduzieren, wenn statt der Ausgangsvariablen die in der Regressionsschätzung berücksichtigten logarithmierten Produktionsfaktoren sowie der logarithmierte Output mikroaggregiert werden. Gleichzeitig zeigt sich, dass sich die Veränderung der Ergebnisse erhöht, wenn die Mikroaggregation nicht für den Gesamtdatensatz, sondern getrennt für einzelne Zellen vorgenommen wird. Dieses Ergebnis ist auch einsichtig, weil sich bei einem solchen Vorgehen die Abweichungen der Einzelwerte erhöhen.

Für die gemeinsame abstandsorientierte Mikroaggregation ist in Tabelle 23.13 gezeigt, dass sich die Ergebnisse auch bei diesem Verfahren verbessern, sofern statt der Ausgangsvariablen die in das Modell eingehenden transformierten Variablen mikroaggregiert werden. Dies ist auch einsichtig, weil die nichtlineare Transformation einer der Gründe für die Verzerrung des Schätzers bei gemeinsam mikroaggregierten Daten ist. Die Wirkung der zellenweisen Mikroaggregation ist bei der gemeinsamen abstandsorientierten Mikroaggregation hingegen nicht eindeutig. Während sie bei der Mikroaggregation der Ausgangsvariablen zu einer Verbesserung der Ergebnisse führt, ergibt sich bei der Mikroaggregation der transformierten Variablen eine Verschlechterung, falls die Mikroaggregation zellenweise erfolgt. Auch sind die Ergebnisse bei der zellenweisen Mikroaggregation schlechter, sofern die transformierten Variablen mikroaggregiert werden.

Die zufällige gemeinsame Mikroaggregation hat gegenüber der abstandsorientierten ge-

Tabelle 23.12: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für getrennt abstandsorientiert mikroaggregierte KSE-Daten, Datensatz bereinigt, robuste Standardfehler

	MA33G Ausgangsvariablen mikroaggregiert		MA33G nach Zellen Ausgangsvariablen mikroaggregiert		Mikro_getrennt Logarithmen der Inputfaktoren und des Outputs mikroaggregiert		Mikro_getrennt nach Zellen Logarithmen der Inputfaktoren und des Outputs mikroaggregiert	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,434	(148,75)	0,437	(138,98)	0,434	(148,81)	0,432	(131,19)
Personalkosten	0,322	(100,45)	0,328	(90,33)	0,322	(100,53)	0,328	(87,82)
Externe Dienstleistungen	0,053	(39,57)	0,054	(38,28)	0,052	(39,53)	0,052	(37,35)
Sonstige Kosten	0,105	(46,89)	0,108	(45,05)	0,105	(46,89)	0,108	(44,37)
Kapitalkosten	0,070	(32,40)	0,056	(21,16)	0,070	(32,25)	0,06	(24,49)
Konst.	1,723	(97,16)	1,734	(81,85)	1,721	(97,01)	1,76	(81,49)
Anzahl Beobachtungen	15.049		14.937		15.049		15.030	
R^2	0,988		0,985		0,988		0,985	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	0,23	0,80	0,46	7,32	0,23	0,76	0,69	12,51
Personalkosten	0	0,24	1,86	10,29	0	0,16	1,86	12,78
Externe Dienstleistungen	1,92	0,66	3,85	2,62	0	0,56	0	4,99
Sonstige Kosten	0	0,04	2,86	3,97	0	0,04	2,86	5,41
Kapitalkosten	0	0,12	20	34,77	0	0,59	14,29	24,51
Konst.	0,35	0,74	0,99	16,38	0,23	0,89	2,50	16,74
Durchschn.	0,42	0,43	5,00	12,56	0,08	0,50	3,70	12,82

meinsamen Mikroaggregation den Vorteil, dass in linearen Modellen grundsätzlich keine Verzerrung der Schätzer auftritt. Allerdings ergeben sich im hier verwendeten Anwendungsbeispiel auch bei dieser stochastischen Variante Abweichungen von den Parameterwerten der Originalschätzung, die auf die nichtlineare Transformation der mikroaggregierten Merkmale und die Ausreißerbereinigung zurückzuführen sind. In der ersten Spalte der Tabelle 23.14 ist zu erkennen, dass diese Abweichungen über denen liegen, die durch die abstandsorientierte gemeinsame Mikroaggregation hervorgerufen werden.

Tabelle 23.13: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für gemeinsam abstandsorientiert mikroaggregierte KSE-Daten, Datensatz bereinigt, robuste Standardfehler

	Ausgangs-variablen mikro-aggregiert		nach Zellen Ausgangs-variablen mikroaggregiert		Logarithmen der Inputfaktoren und des Outputs mikroaggregiert		nach Zellen Logarithmen der Inputfaktoren und des Outputs mikroaggregiert	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,492	(230,19)	0,463	(190,2)	0,439	(153,83)	0,43	(137,6)
Personalkosten	0,272	(112,06)	0,296	(110,54)	0,325	(108,71)	0,345	(86,12)
Externe Dienstleistungen	0,052	(49,25)	0,053	(46,42)	0,053	(46,02)	0,034	(26,26)
Sonstige Kosten	0,111	(57,68)	0,111	(56,09)	0,104	(51,61)	0,116	(50,92)
Kapitalkosten	0,058	(31,85)	0,066	(35,53)	0,066	(33,44)	0,062	(18,26)
Konst.	1,591	(135,54)	1,585	(125,99)	1,664	(110,02)	1,643	(100,60)
Anzahl Beobachtungen	15.616		15.594		15.037		15.291	
R^2	0,995		0,994		0,992		0,992	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	13,1	53,51	6,44	26,84	0,92	2,59	1,15	8,24
Personalkosten	15,53	11,29	8,07	9,78	0,93	7,97	7,14	14,47
Externe Dienstleistungen	0	25,29	1,92	18,09	1,92	17,07	34,62	33,20
Sonstige Kosten	5,71	22,96	5,71	19,57	0,95	10,02	10,48	8,55
Kapitalkosten	17,14	1,82	5,71	9,53	5,71	3,08	11,43	43,71
Konst.	7,34	38,48	7,69	28,72	3,09	12,40	4,31	2,78
Durchschn.	9,8	25,56	5,93	18,75	2,25	8,85	11,52	18,49

Tabelle 23.14: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für gemeinsam zufällig mikroaggregierte KSE-Daten, Datensatz bereinigt, robuste Standardfehler

	Ausgangs-variablen mikro-aggregiert		Nach Zellen Ausgangs-variablen mikroaggregiert		Logarithmen der Inputfaktoren und des Outputs mikroaggregiert		Nach Zellen Logarithmen der Inputfaktoren und des Outputs mikroaggregiert	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,493	(222,56)	0,491	(209,97)	0,405	(107,65)	0,404	(112,80)
Personalkosten	0,267	(101,83)	0,264	(104,69)	0,356	(70,19)	0,34	(86,42)
Externe Dienstleistungen	0,053	(44,19)	0,054	(47,75)	0,046	(29,50)	0,047	(30,98)
Sonstige Kosten	0,117	(57,40)	0,120	(59,83)	0,114	(45,24)	0,123	(47,87)
Kapitalkosten	0,058	(28,27)	0,058	(29,06)	0,060	(20,78)	0,064	(24,71)
Konst.	1,568	(114,70)	1,589	(112,90)	1,769	(80,51)	1,837	(86,17)
Anzahl Beobachtungen	15.589		15.580		15.589		15.580	
R^2	0,992		0,991		0,979		0,98	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	13,33	48,42	12,87	40,03	6,9	28,21	7,13	24,77
Personalkosten	17,08	1,13	18,01	3,97	10,56	30,29	5,59	14,17
Externe Dienstleistungen	1,92	12,41	3,85	21,47	11,54	24,96	9,62	21,19
Sonstige Kosten	11,43	22,36	14,29	27,54	8,57	3,56	17,14	2,05
Kapitalkosten	17,14	12,85	17,14	10,42	14,29	35,94	8,57	23,83
Konst.	8,68	17,18	7,45	15,35	3,03	17,75	6,99	11,96
Durchschn.	11,6	19,06	12,27	19,80	9,15	23,45	9,17	16,33

Die Ergebnisse verschlechtern sich durch die Durchführung einer zellenweisen zufälligen Mikroaggregation nur geringfügig. Geringere Abweichungen in den Parameterwerten ergeben sich wiederum, wenn statt der Ausgangsvariablen die in das Modell eingehenden transformierten Variablen zufällig mikroaggregiert werden. Die in diesem Fall verbleibende Verzerrung ist darauf zurückzuführen, dass der Datensatz erst nach Durchführung der Mikroaggregation um Ausreißer bereinigt wird.

In der Tendenz gleiche Ergebnisse wie bei der zufälligen Mikroaggregation ergeben sich für die Bootstrap-Mikroaggregation (Tabelle 23.15). Allerdings sind die Ergebnisverbesserungen durch den Übergang von der Anonymisierung der Ausgangsvariablen zur Anonymisierung der im Modell berücksichtigten transformierten Variablen hier eindeutiger.

Tabelle 23.17 zeigt, wie sich die Ergebnisse optimieren lassen, wenn die Ausreißerbereinigung vor der Anonymisierung vorgenommen wird und anschließend die transformierten im Modell berücksichtigten Variablen mikroaggregiert werden, sowie die Unternehmen mit „Missing Values“ bei den originalen Logarithmen nicht berücksichtigt werden. Für diesen Fall sind, wie in den theoretischen Herleitungen in Unterabschnitt 23.1.1 abgeleitet wurde, die Schätzer bei den stochastischen Mikroaggregationsverfahren unverzerrt. Die Schätzer bei Anwendung der getrennten abstandsorientierten und der gemeinsamen abstandsorientierten Mikroaggregation sind hingegen theoretisch verzerrt, letztere aufgrund der Tatsache, dass die abhängige Variable bei der Berechnung des Abstandsmaßes, das zur Gruppenbildung

Tabelle 23.15: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für mit Bootstrap-Mikroaggregation bearbeitete KSE-Daten, Datensatz bereinigt, robuste Standardfehler

	Ausgangs-variablen mikro-aggregiert		nach Zellen Ausgangs-variablen mikroaggregiert		Logarithmen der Inputfaktoren und des Outputs mikroaggregiert		nach Zellen Logarithmen der Inputfaktoren und des Outputs mikroaggregiert	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,497	(235,96)	0,488	(203,59)	0,414	(110,94)	0,413	(117,72)
Personalkosten	0,270	(105,92)	0,261	(101,31)	0,349	(82,12)	0,324	(86,75)
Externe Dienstleistungen	0,053	(47,43)	0,055	(49,50)	0,052	(32,74)	0,053	(33,09)
Sonstige Kosten	0,111	(58,02)	0,120	(59,36)	0,109	(42,26)	0,122	(47,76)
Kapitalkosten	0,059	(28,76)	0,064	(32,19)	0,058	(21,14)	0,066	(25,51)
Konst.	1,543	(110,56)	1,592	(112,29)	1,735	(74,36)	1,853	(83,72)
Anzahl Beobachtungen	15.607		15.606		14.065		14.157	
R^2	0,992		0,991		0,98		0,981	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	14,25	57,36	12,18	35,77	4,83	26,02	5,06	21,49
Personalkosten	16,15	5,19	18,94	0,62	8,39	18,44	0,62	13,84
Externe Dienstleistungen	1,92	20,66	5,77	25,92	0	16,71	1,92	15,82
Sonstige Kosten	5,71	23,68	14,29	26,54	3,81	9,91	16,19	1,81
Kapitalkosten	15,71	11,34	8,57	0,77	17,14	34,83	5,71	21,36
Konst.	10,13	12,95	7,28	14,72	1,05	24,03	7,92	14,47
Durchschn.	10,65	21,87	11,17	17,39	5,87	21,66	6,24	14,80

herangezogen wird, berücksichtigt wird.

Die in Tabelle 23.17 dargestellten Schätzergebnisse zeigen, dass sich die Veränderungsraten bei den stochastischen Mikroaggregationsvarianten tatsächlich deutlich reduzieren. Allerdings sind, obwohl die Schätzer theoretisch verzerrt sind, für die abstandsorientierten Varianten der Mikroaggregation bei der durchgeführten Schätzung noch geringere Abweichungen im Vergleich zur Originalschätzung zu beobachten. Bei der gemeinsamen Mikroaggregation kann dieses Ergebnis darauf zurückgeführt werden, dass der Einfluss der abhängigen Variablen bei der Gruppenbildung gering ist. Vor allem aber geht die Verzerrung der Schätzer gegen Null, falls abhängige und erklärende Variable hoch korreliert sind, was in diesem Fall zutrifft (vgl. hierzu Tabelle 23.16).

Tabelle 23.16: Korrelationskoeffizienten zwischen dem logarithmierten Output und den logarithmierten Inputgrößen

	Logarithmierter Output
Logarithmierter Materialeinsatz	0,9663
Logarithmierte Personalkosten	0,9493
Logarithmierte Externe Dienstleistungen	0,8169
Logarithmierte Sonstige Kosten	0,9212
Logarithmierte Kapitalkosten	0,9030

Tabelle 23.17: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für mit verschiedenen Varianten der Mikroaggregation bearbeitete KSE-Daten, Mikroaggregation nach Bereinigung, Logarithmen des Outputs und der Inputfaktoren mikroaggregiert, robuste Standardfehler

	Getrennte abstands-orientierte Mikro-aggregation		Gemeinsame abstands-orientierte Mikro-aggregation		Gemeinsame zufällige Mikro-aggregation		Gemeinsame Bootstrap-Mikro-aggregation	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,435	(149,88)	0,435	(168,01)	0,432	(190,02)	0,435	(196,97)
Personalkosten	0,322	(100,57)	0,322	(115,71)	0,326	(112,58)	0,320	(113,59)
Externe Dienstleistungen	0,052	(39,31)	0,052	(49,86)	0,051	(41,70)	0,052	(42,43)
Sonstige Kosten	0,105	(46,91)	0,105	(55,21)	0,107	(58,83)	0,106	(55,07)
Kapitalkosten	0,070	(32,41)	0,072	(39,38)	0,067	(31,27)	0,072	(34,88)
Konst.	1,717	(96,83)	1,703	(123,90)	1,726	(100,65)	1,715	(100,33)
Anzahl Beobachtungen	15.017		15.017		15.017		15.017	
R^2	0,988		0,993		0,988		0,988	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	0	0,05	0	12,04	0,69	26,72	0	31,36
Personalkosten	0	0,12	0	14,92	1,24	11,81	0,62	12,81
Externe Dienstleistungen	0	0,00	0	26,84	1,92	6,08	0	7,94
Sonstige Kosten	0	0,00	0	17,69	1,9	25,41	0,95	17,40
Kapitalkosten	0	0,09	2,86	21,39	4,29	3,61	2,86	7,52
Konst.	0	1,07	0,82	26,58	0,52	2,83	0,12	2,50
Durchschn.	0	0,22	0,61	19,91	1,76	12,74	0,76	13,25

Bei der getrennten Mikroaggregation gelten die gleichen Gründe, die bereits bei den Simulationsexperimenten im vorangegangenen Unterabschnitt angeführt wurden, um die nahezu perfekten Ergebnisse zu erklären. Zum einen werden durch diese Form der Mikroaggregation die einzelnen Merkmalswerte nur geringfügig verändert (insbesondere bei den kleineren und mittleren Unternehmen), zum anderen führen hohe Korrelationen zwischen erklärenden Variablen und abhängiger Variable auch bei der getrennten Mikroaggregation dazu, dass die Schätzergebnisse im linearen Modell auch theoretisch nahezu unverzerrt sind. Tabelle 23.18 zeigt, dass sich die Ergebnisse verschlechtern, wenn die Mikroaggregation nach Ausreißerbereinigung und Transformation zellenweise durchgeführt wird.

Nun wird erneut die Ausgangsvariante betrachtet, bei der die Ausgangsvariablen mikroaggregiert werden und die Ausreißerbereinigung erst nach der Anonymisierung vorgenommen wird. Für diesen Fall wird untersucht, inwiefern sich die Ergebnisse bei der getrennten abstandsorientierten Mikroaggregation verändern, wenn ein Mindestabstand von sieben beziehungsweise 15 Prozent zwischen den Gruppenmittelwerten und damit zwischen den mikroaggregierten Werten eingeführt wird (MA33G_7p und MA33G_15p). Die entsprechenden Ergebnisse in Tabelle 23.19 zeigen, dass mit wachsendem Mindestabstand die Abweichungen der Parameterwerte (und auch der t-Werte) erwartungsgemäß steigen. In beiden hier betrachteten Fällen treten bei einzelnen Parametern Abweichungen von über zehn Prozent auf.

Tabelle 23.18: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für mit verschiedenen Varianten der Mikroaggregation bearbeitete KSE-Daten, Mikroaggregation nach Bereinigung, Logarithmen des Outputs und der Inputfaktoren mikroaggregiert, Mikroaggregation nach Zellen, robuste Standardfehler

	Getrennte abstands-orientierte Mikro-aggregation		Gemeinsame abstands-orientierte Mikro-aggregation		Gemeinsame zufällige Mikro-aggregation		Gemeinsame Bootstrap-Mikro-aggregation	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,434	(142,49)	0,427	(163,89)	0,421	(164,74)	0,422	(172,84)
Personalkosten	0,317	(85,76)	0,316	(113,45)	0,322	(124,57)	0,318	(121,67)
Externe Dienstleistungen	0,052	(37,10)	0,055	(48,94)	0,052	(41,13)	0,052	(42,49)
Sonstige Kosten	0,107	(44,43)	0,113	(53,71)	0,115	(54,21)	0,112	(53,95)
Kapitalkosten	0,073	(29,43)	0,073	(37,64)	0,074	(35,66)	0,077	(38,64)
Konst.	1,734	(82,83)	1,742	(129,42)	1,759	(103,93)	1,778	(107,79)
Anzahl Beobachtungen	15.017		15.017		15.017		15.017	
R^2	0,986		0,994		0,988		0,988	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	0,23	4,97	1,84	9,30	3,22	9,86	2,99	15,27
Personalkosten	1,55	14,83	1,86	12,67	0	23,72	1,24	20,84
Externe Dienstleistungen	0	5,62	5,77	24,50	0	4,63	0	8,09
Sonstige Kosten	1,9	5,29	7,62	14,50	9,52	15,56	6,67	15,01
Kapitalkosten	4,29	9,28	4,29	16,03	5,71	9,93	10,00	19,11
Konst.	0,99	15,38	1,46	32,22	2,45	6,18	3,55	10,12
Durchschn.	1,49	9,23	3,81	18,20	3,48	11,65	4,08	14,74

Zuletzt wird in Tabelle *23.20* gezeigt, wie sich die Ergebnisse der gemeinsamen abstandsorientierten Mikroaggregation verändern, wenn statt einer Gruppengröße von drei (bis maximal fünf) Einheiten, bei denen die Einzelwerte komplett durch die Gruppendurchschnitte ersetzt werden, Gruppen mit jeweils vier Merkmalsträgern gebildet werden, von denen zwei den um die Standardabweichung innerhalb der Gruppe verminderten Gruppendurchschnitt als neuen Merkmalswert zugewiesen bekommen, die beiden anderen den um die Standardabweichung erhöhten Gruppendurchschnitt (MA1G_var). Offenbar ergibt sich bei diesem Verfahren eine deutlich stärkere Abweichung der Ergebnisse vom Original als bei der normalen gemeinsamen abstandsorientierten Mikroaggregation.

Tabelle 23.19: Lineariserte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für mit verschiedenen Varianten der getrennten abstandsorientierten Mikroaggregation bearbeitete KSE-Daten, Daten bereinigt, Ausgangsvariablen anonymisiert, robuste Standardfehler

	MA33G		MA33G_7p		MA33G_15p	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,434	(148,75)	0,438	(148,12)	0,439	(141,66)
Personalkosten	0,322	(100,45)	0,328	(101,27)	0,330	(94,15)
Externe Dienstleistungen	0,053	(39,57)	0,053	(39,87)	0,053	(37,61)
Sonstige Kosten	0,105	(46,89)	0,105	(46,37)	0,110	(45,00)
Kapitalkosten	0,070	(32,4)	0,058	(27,95)	0,050	(21,76)
Konst.	1,723	(97,16)	1,715	(94,53)	1,710	(88,06)
Anzahl Beobachtungen	15.049		15.036		14.582	
R^2	0,988		0,987		0,985	
	Relative Abweichungen von den Originalwerten in %					
Materialeinsatz	0,23	0,80	0,69	1,22	0,92	5,53
Personalkosten	0	0,24	1,86	0,58	2,48	6,50
Externe Dienstleistungen	1,92	0,66	1,92	1,42	1,92	4,32
Sonstige Kosten	0	0,04	0	1,15	4,76	4,07
Kapitalkosten	0	0,12	17,14	13,84	28,57	32,92
Konst.	0,35	0,74	0,12	3,42	0,41	10,03
Durchschn.	0,42	0,43	3,62	3,61	6,51	10,56

Tabelle 23.20: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für mit verschiedenen Varianten der gemeinsamen abstandsorientierten Mikroaggregation bearbeitete KSE-Daten, Daten bereinigt, Ausgangsvariablen anonymisiert, robuste Standardfehler

	MA1G		MA1G_var	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,492	(230,19)	0,384	(110,89)
Personalkosten	0,272	(112,06)	0,273	(54,38)
Externe Dienstleistungen	0,052	(49,25)	0,103	(55,44)
Sonstige Kosten	0,111	(57,68)	0,143	(44,33)
Kapitalkosten	0,058	(31,85)	0,056	(18,85)
Konst.	1,591	(135,54)	2,338	(72,13)
Anzahl Beobachtungen	15.616		15.153	
R^2	0,995		0,958	
	Relative Abweichungen von den Originalwerten in %			
Materialeinsatz	13,10	53,51	11,72	26,05
Personalkosten	15,53	11,29	15,22	45,99
Externe Dienstleistungen	0	25,29	98,08	41,03
Sonstige Kosten	5,71	22,96	36,19	5,50
Kapitalkosten	17,14	1,82	20,00	41,89
Konst.	7,34	38,48	36,17	26,31
Durchschn.	9,80	25,56	36,23	31,13

23.3.2 Mikroaggregationsverfahren in nichtlinearen Modellen

Zur Untersuchung der Wirkung von Mikroaggregationsverfahren in nichtlinearen Modellen wird das Probit-Modell zur Erklärung der Tarifbindung mit den Daten des IAB-Betriebspanels für Baden-Württemberg (2002) geschätzt, wie es in Abschnitt 21.2 eingeführt wurde. Dabei ist die logarithmierte Beschäftigung die einzige erklärende metrische Variable.

In einem Fall wird direkt die logarithmierte Beschäftigung mikroaggregiert, im anderen Fall die Beschäftigung. Zur Anwendung kommt ebenso die abstandsorientierte wie die zufällige Mikroaggregation. Die Ergebnisse sind in Tabelle 23.21 dargestellt.

Tabelle 23.21: Probit-Schätzung zur Erklärung der Tarifbindung – Schätzergebnisse für mikroaggregierte Daten des IAB-Betriebspanels 2002 für Baden-Württemberg, bei stochastischen Verfahren 1.000 Replikationen

	Abstandsorientierte Mikroaggregation der log. Beschäftigung		Zufällige Mikroaggregation der log. Beschäftigung		Abstandsorientierte Mikroaggregation der Beschäftigung		Zufällige Mikroaggregation der Beschäftigung	
Variablen	Koeff.	(t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)	Koeff.	(t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Log. Beschäftigung	0,301	(12,24)	0,240	(5,19)	0,30	(12,24)	0,152	(4,06)
Baugewerbe	0,747	(4,45)	0,244	(1,56)	0,75	(4,46)	0,205	(1,32)
Handel	0,377	(2,87)	-0,086	(-0,72)	0,38	(2,88)	-0,123	(-1,03)
Dienstleistungs-sektor	0,039	(0,40)	-0,309	(-3,43)	0,04	(0,40)	-0,339	(-3,79)
Öffentliche Verwaltung	0,795	(4,60)	0,686	(4,13)	0,79	(4,60)	0,682	(4,12)
Konst.	-0,987	(-7,46)	0,546	(5,19)	-0,99	(-7,46)	-0,358	(-1,78)
	Relative Abweichungen von den Originalwerten in %							
Log. Beschäftigung	0,09	0	20,12	57,6	0,10	0	49,54	66,83
Baugewerbe	0,11	0,22	67,41	65,02	0,21	0	72,53	70,4
Handel	0,34	0,35	122,8	125,00	0,17	0	132,55	135,76
Dienstleistungs-sektor	0,85	0	878,48	957,50	1,53	0	954,72	1047,5
Öffentliche Verwaltung	0,14	0	13,79	10,22	0,17	0	14,35	10,43
Konst.	0,11	0	44,79	62,73	0,09	0	63,75	76,14
Durchschn.	0,27	0,10	191,23	213,01	0,38	0	214,57	234,51

Man erkennt, dass sich bei der abstandsorientierten Mikroaggregation annähernd die gleichen Ergebnisse erzielen lassen wie mit den Originaldaten, unabhängig davon, ob die Beschäftigung zuerst logarithmiert oder erst mikroaggregiert wird. Mit der abstandsorientierten getrennten Mikroaggregation lassen sich somit offenbar auch in nichtlinearen Modellen und bei nichtlinearen Transformationen recht gute Ergebnisse erzielen.

Demgegenüber ergeben sich durch die zufällige Mikroaggregation bei der Probit-Schätzung starke Abweichungen der geschätzten Koeffizienten sowie der Teststatistiken. Es kommt sogar zu Vorzeichenwechseln. Zufällige Mikroaggregationsverfahren erweisen sich somit in nichtlinearen Modellen als für die Anonymisierung nicht geeignet.

Die Unterscheidung zwischen getrennter und gemeinsamer abstandsorientierter Mikroaggregation lässt sich bei einer metrischen Regressorvariablen natürlich nicht treffen. Allerdings haben bereits die Simulationsergebnisse in Unterabschnitt 23.2.2 gezeigt, dass die gemeinsame abstandsorientierte Mikroaggregation zu sehr starken Verzerrungen bei der Schätzung nichtlinearer Modelle führt.

Kapitel 24

Resampling in linearen Modellen

24.1 Ergebnisse von Monte-Carlo-Simulationen

Gottschalk (2005) testet mit Hilfe von Monte-Carlo-Simulationen die Wirkungsweise der verschiedenen Resamplingvarianten in linearen Regressionsmodellen. Dazu werden 2.500 synthetische Beobachtungen erzeugt, indem Pseudo-Zufallszahlen aus theoretischen Verteilungen gezogen werden: Die Variable Y ist eine Linearkombination von zwei normalverteilten Merkmalen $X_1 \sim N(0,1)$ und $X_2 \sim N(0,1)$ ($\rho(X_1, X_2) = 0,5$) und einer normalverteilten Zufallszahl:

$$Y = 0,3 + X_1 + 0,5X_2 + U, \qquad U \sim N(0,1). \tag{24.1}$$

Wie bereits in Unterabschnitt 6.2.5 erläutert, bestimmt die Wahl der Bandbreiten das Ausmaß der Anonymisierung der Daten, da sie im Wesentlichen die Stärke der Glättung der geschätzten Kerndichte der Variablen festlegt. Je größer die Bandbreite ist, desto stärker wird die Ursprungsverteilung geglättet, bei zu starker Glättung auch verzerrt. Verteilungsparameter können dann nicht mehr erwartungstreu geschätzt werden. Auf der anderen Seite steigt durch eine stärkere Glättung aber der Vertrauensschutz, weil die einzelnen Punkte der inversen Verteilungsfunktion nur noch näherungsweise den Ursprungswerten entsprechen. Eine Reidentifizierung originaler Merkmalsträger wird unwahrscheinlicher. Eine optimale Bandbreite berücksichtigt beide Punkte: die Bewahrung der Analysefähigkeit und die Gewährleistung des Vertrauensschutzes.

Ausgangspunkt dieser Untersuchung sind die quasi-optimalen Bandbreiten für univariate Verteilungen, die Silverman (1986) vorgeschlagen hat:

$$
\begin{aligned}
h &= \frac{0,9m}{n^{1/5}} \qquad (24.2)\\
m &= \min\left[\sqrt{\sigma_x^2}, \frac{quant(p75)_x - quant(p25)_x}{1,349}\right].
\end{aligned}
$$

Für jede der drei Variablen wird eine eigene Bandbreite geschätzt. Sie können als eine Approximation der optimalen Bandbreiten hinsichtlich der Güte der Schätzung univariater Verteilungen betrachtet werden. Kann für das jeweilige Resample mit diesen Bandbreiten faktische Anonymität erreicht werden, sind quasi-optimale Bandbreitenwerte hinsichtlich Schutzwirkung *und* Analysefähigkeit des Resamples – bezüglich der Abbildung der univariaten Verteilungen[34] – gefunden. Dabei wird folgendes Optimierungsproblem zur Bestimmung der Bandbreiten herangezogen: Maximiere die Bandbreiten und erhalte dabei die Analysefähigkeit der Daten. An diesem kritischen Punkt führen kleine Erweiterungen der Bandbreiten zu einer deutlichen Verschlechterung der Datenqualität. Das bedeutet, dass die Verzerrung der Ursprungsverteilungen bis zu dem kritischen Punkt zwar iterativ erhöht wird, aber dennoch sichergestellt ist, dass valide wissenschaftliche Auswertungen mit den anonymisierten Daten möglich sind. Eine weitere Anoymisierung bzw. Glättung der Daten würde die Datensätze unbrauchbar werden lassen.

Die Restriktion dieses Optimierungsproblems – die Erhaltung der Analysefähigkeit – kann ebenfalls nicht eindeutig definiert werden. Erste Experimente haben gezeigt, dass eindimensionales Resampling univariate Verteilungsparameter gut reproduziert, aber mehrdimensionale Strukturen zum Teil verzerrt. Multivariates Resampling, bei dem die Korrelationsstruktur in den Algorithmus einfließt, bildet dagegen mehrdimensionale Beziehungen besser ab als univariate Parameter (Gottschalk 2004).

Aus diesem Grund formuliert Gottschalk (2005) folgendes Maximierungsproblem: Die Koeffizienten des linearen Regressionsmodells in Gleichung (24.1) sollen mit denjenigen Resamples, die durch Einbeziehung der Korrelationsstruktur der Daten entstehen, nicht signifikant von den originalen Koeffizienten abweichen:

$$
\begin{aligned}
&\text{Max} \quad \rightharpoonup \quad h\\
&\text{u.d.N.} \quad \beta_{X^a} = \beta_X\\
&\text{mit} \qquad Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + U, \quad U \sim N(0,1)\\
&\text{und} \qquad Y^a = \beta_0^a + \beta_1^a X_1^a + \beta_2^a X_2^a + U, \quad U \sim N(0,1).
\end{aligned}
$$

Der Maximierungsprozess wird iterativ durchlaufen.[35] Die Startwerte für die Bandbreiten für

34) Das heißt, nicht unbedingt optimal zur Abbildung der multivariaten Verteilung.

jede Variable sind die optimalen Bandbreiten laut Gleichung (24.2). In der ersten Iteration werden alle Bandbreiten mit einen Faktor $f = 1,5$ multipliziert und das Resample berechnet (100 Wiederholungen). Nach Schätzung der Modellparameter wird mit Hilfe von t-Tests überprüft, ob die Koeffizienten signifikant von ihren theoretischen Werten abweichen. Da dies nicht der Fall ist, wird der Algorithmus wiederholt, wobei die Bandbreitenfaktoren um 0,2 auf $f = 1,7$ erhöht werden.

Mehrere Durchläufe des Algorithmus führen im Durchschnitt zu einem maximalen Bandbreitenfaktor von $f = 2,1$. Eine Verdoppelung des optimalen Bandbreitenwertes führt also im Resample weiterhin zu sinnvollen Ergebnissen einer linearen Regressionsanalyse. Um die Auswirkung einer Bandbreitenerhöhung auf die Analysefähigkeit der Resamples zu untersuchen, verwendet Gottschalk (2005) in der Monte-Carlo-Studie jeweils fünf verschiedene Bandbreiten, die sich hinsichtlich ihrer Faktoren f unterscheiden: f nimmt den Wert 1, 1,5, 1,7, 1,9 oder 2,1 an.

Von den drei Variablen werden in jedem Durchlauf der Simulation (insgesamt 100) aufgrund der verschiedenen Werte für f je fünf univariate Resamples und $5 * 3$ multivariate Resamples gezogen. Die drei multivariaten Resamplingversionen unterscheiden sich hinsichtlich ihrer Kernmatrix, wie in Unterabschnitt 6.2.5 dargestellt. Eine OLS-Schätzung der Parameter des linearen Modells in Gleichung (24.1) wird mit jedem der vier Resamples durchgeführt. Nach 100 Wiederholungen werden die Mittelwerte der Koeffizienten, die Monte-Carlo-Schätzer, ermittelt. Anhand von t-Tests wird geklärt, ob die Monte-Carlo-Schätzer signifikant von den theoretischen Parameterwerten des linearen Modells abweichen.

Der Resamplingalgorithmus besteht im Wesentlichen aus zwei Komponenten: der Bootstrap-Stichprobenziehung und der stochastischen Fehlerüberlagerung durch die Kerne. Um zu erkennen, wie stark diese beiden Komponenten jeweils den Informationsgehalt der Daten einschränken, werden im ersten Schritt Schätzungen mit dem Bootstrapsample, d.h. mit den unveränderten Beobachtungen, durchgeführt. In Tabelle 24.1 sind die Ergebnisse der Monte-Carlo-Analyse dokumentiert. In der ersten Spalte sind die Schätzergebnisse mit dem Bootstrapsample eingetragen. Die unterschiedlichen multivariaten Resamplingvarianten sind in der Tabelle folgendermaßen bezeichnet:

- Multivariat-A: Die Kerne werden mit der Varianz-Kovarianzmatrix der Originalvariablen gewichtet.
- Multivariat-V: Die mehrdimensionale Kernmatrix wird so umskaliert, dass sie dieselbe Kovarianzstruktur aufweist wie die originale Datenmatrix.
- Multivariat-C: Die mehrdimensionale Kernmatrix wird so umskaliert, dass sie dieselbe Korrelationsmatrix aufweist wie die originale Datenmatrix.

35) Auf eine ausführliche Darstellung des Algorithmus wird an dieser Stelle verzichtet. In Gottschalk (2005) wird er aber detailliert beschrieben.

Die Koeffizienten sowohl der Bootstrapsamples als auch der Resamples stimmen größtenteils signifikant mit den Originalkoeffizienten überein. Die Schätzer mit den univariat und durch die multivariate Version C generierten Resamples weichen nur wenig von den Bootstrap-Werten ab. Für diese Resampling-Varianten wird demnach durch die Fehlerüberlagerung keine deutliche Veränderung der Ergebnisse induziert. Das Bestimmtheitsmaß R^2 der Schätzmodelle liegt bei allen Version-C-Resamples höher als bei den anderen. Die t-Werte der Koeffizienten bei den Versionen C weichen als einzige nur wenig von den Werten, die mit dem Bootstrapsample berechnet wurden, ab. Das deutet darauf hin, dass die Skalierung der Kernmatrix mit der Korrelationsmatrix das lineare Modell am wenigsten von allen Varianten verfälscht. Bei Version A werden teilweise ungenaue Parameterschätzungen beobachtet, z.B. der Koeffizient von der Variablen X_1 bei erhöhtem Bandbreitenfaktor (1,5; 1,7; 1,9; 2,1). Die Bestimmtheitsmaße sind deutlich niedriger als bei allen anderen Schätzmodellen. Version-V-Resamples produzieren weniger deutliche Abweichungen von den Originalwerten, können aber nicht die Genauigkeit der Resampling-Varianten C und der univariaten Version erreichen.

Tabelle 24.1: MC-Simulationen, Resampling – OLS-Schätzergebnisse im Vergleich, Resampling-Verfahren mit verschiedenen Bandbreitenfaktoren (BBF) (100 Wiederholungen)

Variable	Bootstrap-sample	Univariat BBF=1	Multivariat-A BBF=1	Multivariat-V BBF=1	Multivariat-C BBF=1
X_1	0,996*	0,989	0,965	0,988	0,996*
(t-Werte)	(43,09)	(42,47)	(40,42)	(42,27)	(43,08)
X_2	0,498*	0,498*	0,502*	0,496*	0,499*
(t-Werte)	(21,56)	(21,39)	(21,01)	(21,25)	(21,60)
Konst.	0,302*	0,301*	0,301*	0,302*	0,302*
(t-Werte)	(15,10)	(14,91)	(14,53)	(14,91)	(15,09)
R^2	0,635	0,628	0,608	0,627	0,635

Variable	Bootstrap-sample	Univariat		Multivariat-A		Multivariat-V		Multivariat-C	
		BBF=1,5	BBF=1,7	BBF=1,5	BBF=1,7	BBF=1,5	BBF=1,7	BBF=1,5	BBF=1,7
X_1		0,981	0,977	0,929	0,911	0,982	0,979	0,996*	0,996*
(t-Werte)		(41,72)	(41,36)	(37,56)	(36,25)	(41,54)	(41,25)	(43,08)	(43,08)
X_2		0,498*	0,498*	0,504*	0,506*	0,494*	0,493	0,499*	0,499*
(t-Werte)		(21,18)	(21,08)	(20,40)	(20,11)	(20,93)	(20,80)	(21,61)	(21,62)
Konst.		0,301*	0,301*	0,301*	0,301*	0,301*	0,301*	0,302*	0,302*
(t-Werte)		(14,72)	(14,62)	(13,96)	(13,71)	(14,75)	(14,68)	(15,09)	(15,09)
R^2		0,619	0,614	0,577	0,561	0,619	0,616	0,635	0,635

Variable	Bootstrap-sample	Univariat		Multivariat-A		Multivariat-V		Multivariat-C	
		BBF=1,9	BBF=2,1	BBF=1,9	BBF=2,1	BBF=1,9	BBF=2,1	BBF=1,9	BBF=2,1
X_1		0,972	0,967	0,892	0,871	0,976	0,974	0,996*	0,996*
(t-Werte)		(40,96)	(40,52)	(34,89)	(33,50)	(40,97)	(40,71)	(43,08)	(43,08)
X_2		0,498*	0,497*	0,506*	0,507	0,492	0,491	0,499*	0,499*
(t-Werte)		(20,97)	(20,84)	(19,81)	(19,49)	(20,68)	(20,56)	(21,63)	(21,63)
Konst.		0,301*	0,301*	0,301*	0,301*	0,301*	0,301*	0,302*	0,302*
(t-Werte)		(14,52)	(14,41)	(13,45)	(13,19)	(14,61)	(14,55)	(15,09)	(15,08)
R^2		0,609	0,603	0,545	0,527	0,612	0,609	0,635	0,635

*Koeffizienten weichen nicht-signifikant von dem Originalwert ab (Signifikanzniveau mind. 95%, Quelle: Gottschalk (2005, S. 152-153))

24.2 Anwendung mit Daten der Kostenstrukturerhebung

In diesem Abschnitt wird die Wirkungsweise des Resamplings in linearen Modellen einem praktischen Test mit den Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe (KSE) unterzogen. Dabei wird neben dem univariaten Resampling diejenige Variante getestet, bei der eine Anpassung des Kerns an die Korrelationsmatrix der Originalvariablen – Variante C – erfolgt. Diese multivariate Resamplingversion wird hier gewählt, weil sie sich aufgrund der Monte-Carlo-Studie[36] als die am besten geeignete mehrdimensionale Resampling-Variante zur Abbildung multivariater Strukturen herausgestellt hat.

Für die Anwendung mit der Kostenstrukturerhebung wird das Ergebnis der Bandbreitenoptimierung der Monte-Carlo-Simulation im vorangegangenen Abschnitt übernommen und die Bandbreite nur bis zu diesem kritischen Wert erhöht. Dadurch wird sichergestellt, dass das Analysepotenzial der Resamples nicht unverhältnismäßig stark eingeschränkt wird. Der Bandbreitenfaktor f variiert hier ebenfalls zwischen den Werten 1, 1,5, 1,7, 1,9 und 2,1. Pro Resampling-Version (uni- oder multivariat) werden insgesamt fünf verschiedene Resamples (jeweils ein Resample pro Bandbreitenfaktor f) berechnet, also insgesamt resultieren 10 unterschiedliche Resamples pro Variable. Somit können sowohl die Auswirkungen der Wahl unterschiedlicher Bandbreiten als auch die der verschiedenen Resampling-Versionen untersucht werden.

24.2.1 Notwendige Anpassung der Bandbreiten

Werden die verschiedenen Resampling-Varianten, wie beschrieben, berechnet, gibt es eine positive Wahrscheinlichkeit, dass negative Werte generiert werden, auch wenn die Ursprungsvariable nur im positiven Wertebereich definiert ist. Die Kernfunktionen sind symmetrisch mit einem Erwartungswert von Null. Für die empirische Anwendung werden die Bereiche mit hohen Dichtewerten der schief verteilten Merkmale, also sehr kleine Werte, betrachtet. Alle Variablen, deren Verteilungen nur im nichtnegativen Wertebereich definiert sind, was für den Großteil der Merkmale der Kostenstrukturerhebung der Fall ist, durchlaufen einen Anpassungsalgorithmus. Für alle Variablen des Resamples, die per Definition keine negativen Werte annehmen dürfen, wird eine adaptive Bandbreite analog zum Vorgehen bei Gottschalk (2005) berechnet: Die Bandbreiten werden für kleine Werte der Originalangaben solange iterativ verringert, bis keine negativen Merkmalsausprägungen im Resample verbleiben.

1. Suche den kleinsten negativen Wert X^a_{min} der Variablen X^a im Resample.
2. Ersetze den Wert der Variablen X^a an der Stelle X^a_{min} mit einem kleinen positiven

36) Eine Anwendung mit dem Mannheimer Innovationspanel kommt zu derselben Schlussfolgerung (Gottschalk 2005).

Wert, der nur wenig größer als 0 ist.

3. Berechne an der Stelle X^a_{min} eine neue Bandbreite h_{min} durch Lösen der Gleichung (6.81)[37] nach h. Da die Gleichung bezüglich h quadratisch ist, wird jeweils die kleinste positive Lösung h_{min} übernommen.

4. Generiere einen Vektor h_{X^a}, der den ursprünglichen Wert der Bandbreite h annimmt, wenn $X^a_i \geq 0$ und wenn $X^a_i < 0$, dann setze $h_{X^a_i} = h_{min}$, $\forall\ i = 1, 2, ..., n$ Beobachtungen.

5. Ersetze X^a durch Gleichung (6.81) mit $h = h_{X^a}$.

6. Sind weiterhin Werte von X^a kleiner als Null, wiederhole den Algorithmus.

Durch die Verringerung der Bandbreiten an diesen Stellen der Datensätze werden die entsprechenden Werte gegenüber dem Original weniger stark verändert. Da die Bandbreiten für kleine Werte angepasst werden, ist der Grad an Glättung bzw. Anonymisierung in diesem Bereich also geringer. Für die bei der KSE relevanten Variablen liegt dieser Bereich in einem Intervall mit hohen Dichtewerten. Merkmalsträger, deren Angaben einen hohen Dichtewert haben, sind aber einer geringeren Reidentifikationsgefahr ausgesetzt (dies ist relevant bei kleinen Unternehmen, deren monetäre Merkmale klein sind). Demzufolge ist eine schwächere Anonymisierung an diesen Stellen im Hinblick auf den Vertrauensschutz als nicht kritisch zu bewerten.

Die Anpassung der Bandbreiten führt zu einer geringeren Glättung der geschätzten Verteilungsfunktion im Bereich hoher Dichtewerte, also für die Angaben der kleineren Unternehmen, wohingegen Werte von mittleren und großen Unternehmen nur mit einer sehr geringen Wahrscheinlichkeit negativ werden und angepasst werden müssen und demnach stärker verzerrt sind als die der kleineren. Das Analysepotenzial wird durch die Bandbreitenanpassung verbessert, da die Informationen von kleineren Firmen weniger stark verfälscht werden. Verzerrungen der Verteilungen resultieren demnach hauptsächlich aus Veränderungen der Angaben mittlerer und großer Unternehmen.

24.2.2 Technische Umsetzung

Die Implementierung der beschriebenen Resamplingmethode mit gängiger Software ist zunächst unproblematisch. Die Berechnungen der diversen univariaten und multivariaten Resamples der KSE des Jahres 1999 wurden mit dem Statistik-Analyseprogramm STATA durchgeführt. Die praktische Umsetzung der Methode mit den Daten der KSE wirft jedoch aufgrund ihres Umfangs einige Probleme auf. Die rechentechnische Umsetzung mit Querschnittsdaten des Mannheimer Innovationspanels im Verarbeitenden Gewerbe und Bergbau

37) Jeweils angepasst für die univariate bzw. die multivariate Variante.

(MIP) erwies sich dagegen als wenig problematisch (Gottschalk 2004, 2005). Dies liegt zum einen an dem mit im Durchschnitt etwa 2.200 Beobachtungen sehr viel geringeren Umfang eines MIP-Querschnitts. Zum anderen besteht die KSE hauptsächlich aus stetigen Merkmalen. Dies ist bei der ZEW-Innovationserhebung nicht der Fall. Das bedeutet, dass die Berechnung eines Resamples der KSE deutlich länger dauert.

Eine erfolgreiche Umsetzung des multivariaten Verfahrens ist ferner von der Korrelationsstruktur der Daten abhängig. Um Korrelationen zwischen den Variablen im Resample zu erhalten, müssen die Werte im Resample, wie oben beschrieben, transformiert werden. Dies geschieht mit Hilfe einer Cholesky-Transformation der Korrelationsmatrix, die notwendigerweise voraussetzt, dass die Korrelationsmatrix nicht singulär ist. Das heißt, bei zu großen Abhängigkeiten zwischen den betroffenen Variablen, kann die Matrix nicht invertiert werden und eine Transformation ist nicht durchführbar. Dieses Problem tritt insbesondere dann auf, wenn für Analysezwecke notwendige Linearkombinationen oder Transformationen von Variablen der KSE gebildet werden. Es kann gelöst werden, indem die Variablen in zwei oder mehrere Gruppen unterteilt werden, für die getrennt voneinander die Cholesky-Transformation durchgeführt wird. Auf diese Weise können Resamples von Linearkombinationen oder Transformationen von den ursprünglich im Datenbestand vorhandenen Variablen in einem getrennten Rechenvorgang berechnet werden. Bei geschickter Gruppeneinteilung sind nur geringe Auswirkungen auf das Analysepotenzial der multivariaten Resamples zu erwarten.

Die Anpassung der Bandbreiten, die durch die Generierung negativer Merkmalsausprägungen notwendig wird, verursacht mit zunehmender Größe der Datenbasis einen Anstieg der Rechenzeit. Bis der Anpassungsalgorithmus ausschließlich positive Werte hervorbringt, endet die Berechnung der Resamples der KSE erst nach mehreren Stunden. Um die Rechenzeit zu verkürzen bzw. „Endlosschleifen“ zu durchbrechen, wurde das Abbruchkriterium abgeschwächt: Wenn nur noch wenige Beobachtungen pro Variable negativ sind (i.d.R. sind das ein bis drei) wird der Anpassungsalgorithmus für die Bandbreiten gestoppt. Die verbleibenden negativen Werte im Resample dürfen dann allerdings für Analysen nicht verwendet werden.

24.3 Praxisbeispiele

Die Auswirkung der verschiedenen Varianten des Resampling auf das Analysepotenzial der Mikrodaten der Kostenstrukturerhebung soll im Folgenden untersucht werden. Der Einfluss der Bandbreitenwahl auf die Analysefähigkeit der Resamples wird dargestellt. Allerdings wird hier nicht gemessen, inwieweit die Bandbreitenanpassung bei negativen Merkmalsausprägungen im Resample die Analysefähigkeit der Daten verbessert.

Anhand einer linearisierten Cobb-Douglas-Produktionsfunktion werden die Effekte auf die Ergebnisse von linearen Regressionen beispielhaft verdeutlicht. Die Spezifikation der

Schätzgleichung wurde in Abschnitt 21.1 beschrieben. Die Parameter der Cobb-Douglas-Produktionsfunktion in logarithmierter Form werden im Folgenden sowohl mit den Originaldaten der Kostenstrukturerhebung im Verarbeitenden Gewerbe des Jahres 1999 als auch mit deren verschiedenen Resample-Varianten geschätzt und verglichen. Die Determinanten des Produktionsoutputs bestehen demnach aus den fünf Inputfaktoren Materialeinsatz, Personalkosten, Externe Dienstleistungen, Sonstige Kosten und Kapitalkosten.

Die Ergebnisse von vier OLS-Schätzungen mit den Originaldaten ist in Tabelle 24.2 einzusehen (vgl. Abschnitt 21.1). Bei den bereinigten Modell-Varianten werden extreme Merkmalsausprägungen aus der Berechnung ausgeschlossen. Das bedeutet, dass Unternehmen, bei denen die Produktionsanteile eines Inputfaktors weniger als das 1%- oder mehr als das 99%-Quantil der Verteilung der Produktionsanteile über alle Unternehmen betragen, aus der Untersuchung gestrichen werden. Gleiches gilt für Beobachtungen, bei denen ein Inputfaktor den Wert Null annimmt. Durch die Extremwertbereinigung soll ein repräsentatives Schätzergebnis gewährleistet werden. In einer weiteren Modifikation werden jeweils robuste Standardfehler geschätzt (Huber-White-Sandwich-Schätzer der Varianz).

Tabelle 24.2: Cobb-Douglas-Produktionsfunktion – OLS-Schätzergebnisse mit den Originaldaten der KSE 1999

Schätzverfahren	OLS		OLS mit robusten Standardfehlern		OLS (um Extremwerte bereinigte Daten)		OLS mit robusten Standardfehlern (um Extremwerte bereinigte Daten)	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,414	(192,24)	0,414	(79,02)	0,435	(230,41)	0,435	(149,95)
Personalkosten	0,339	(100,11)	0,339	(58,87)	0,322	(116,99)	0,322	(100,69)
Externe DL	0,058	(40,13)	0,058	(29,27)	0,052	(44,39)	0,052	(39,31)
Sonstige Kosten	0,114	(50,52)	0,114	(35,01)	0,105	(56,95)	0,105	(46,91)
Kapitalkosten	0,055	(22,29)	0,055	(16,14)	0,070	(33,85)	0,070	(32,44)
Konst.	1,805	(78,37)	1,805	(66,42)	1,717	(99,49)	1,717	(97,88)

Die verschiedenen Resamplingversionen werden, wie oben beschrieben, generiert. Da die endogene und die exogenen Variablen des Modells in den Originaldaten nicht in logarithmierter Form vorliegen, gibt es zwei Möglichkeiten, Resamples dieser Merkmale zu berechnen:

1. Es werden die jeweiligen Resamples des originalen Datenbestandes ohne zusätzliche transformierte Variablen gezogen. Die Logarithmen der für die Schätzungen benötigten Größen werden *nach* dem Resampling berechnet. Da die Verteilungen von Funktionen von Variablen aber nicht erwartungstreu mit den Resamples geschätzt werden können, sind auch die Regressionsergebnisse verzerrt.

2. Transformationen von Variablen fließen zusätzlich zu den Basisdaten in den Resampling-Prozess ein. Die geglätteten Verteilungen der logarithmierten Modellgrößen sind geeignete Schätzer für die empirischen Verteilungsfunktionen der Ursprungsdaten. Auch Regressionsergebnisse werden demnach weniger stark verändert als im ersten Fall.

Tabelle 24.3 zeigt die Ergebnisse der OLS-Schätzungen der ersten Variante, bei der die Extremwerte nicht ausgeschlossen werden, für die univariaten Resamples mit verschiedenen Bandbreiten. Die relativen Abweichungen von den geschätzten Parameterwerten der Originalschätzung (Erste Spalte in Tabelle 24.2) sind im unteren Abschnitt der Tabelle dargestellt. Der durchschnittliche Fehler bei der Parameterschätzung beträgt etwa 16% (bei einfachem) bis ca. 19% (bei 2,1-fachem Bandbreitenfaktor). Die Verzerrung der Koeffizienten nimmt erwartungsgemäß im Durchschnitt mit steigendem Bandbreitenfaktor (fast) stetig zu. Allerdings beträgt dieser Anstieg insgesamt nur knapp 3 Prozentpunkte. Die Erhöhung der Bandbreiten führt demnach nicht zu einer erheblichen Verschlechterung der Schätzresultate gegenüber den optimalen Bandbreiten (BBF= 1).

Betrachtet man die Schätzwerte im Einzelnen, fällt auf, dass erhebliche Unterschiede bezüglich der Verzerrung bestehen. Die geschätzten Parameter der „externen Dienstleistungen“ und der „Kapitalkosten“ werden stärker verzerrt als die übrigen Koeffizienten. Die Relation zwischen den Abweichungsmaßen variiert darüber hinaus mit dem Bandbreitenfaktor. Während beispielsweise der Parameter der „Personalkosten“ bei einem Faktor von 1 um etwa 6% vom Original abweicht, sind es nur 0,5% im Fall der höchsten Bandbreite. Bei den „sonstigen Kosten“ ist eine umgekehrte Entwicklung zu beobachten. Eine eindeutige positive Beziehung zwischen Schätzfehler und Bandbreitenerhöhung ist folglich nicht zu konstatieren. Die Standardabweichungen bzw. t-Werte der Koeffizienten weichen deutlich stärker von ihren Originalwerten ab als die Parameterschätzer. Es werden durchgängig Fehlerquoten erzielt, die für fast alle Koeffizienten mehr als 50% betragen. Die Standardabweichungen der Schätzparameter werden deutlich unterschätzt. Das deutet darauf hin, dass im Allgemeinen durch Regressionen mit univariaten Resamples die im Originaldatenmaterial signifikanten Zusammenhänge nicht mehr erkannt werden würden. Offensichtlich können Varianzen von Modellschätzern nicht durch univariates Resampling abgebildet werden.

Tabelle 24.4 und Tabelle 24.5 zeigen die Ergebnisse der OLS-Schätzung für multivariates Resampling ohne und mit Clusterbildung bei der Berechnung der Kernmatrizen. Auch hier werden zunächst die transformierten Modellvariablen erst nach dem Resampling berechnet und die Daten nicht um Extremwerte bereinigt. Gegenüber den eindimensional generierten Resamples werden bei allen mehrdimensionalen Varianten deutlich geringere Fehler sowohl bei den Parameterschätzern als auch deren Standardabweichungen gemessen. Die Version ohne Clusterbildung erzeugt die dem Original ähnlichsten Regressionsresultate. Etwa 6% durchschnittliche Abweichung werden bei der Verwendung der jeweils kleinsten Bandbreiten

erzielt. Durch die Bandbreitenerhöhung steigert sich der Fehler insgesamt auf etwa 10%. Durch die Clusterbildung entstehen leicht stärkere Verzerrungen. Die Schätzung der Parametervarianzen kann durch multivariates Resampling gegenüber der univariaten Variante merklich verbessert werden. Der Einsatz multivariater Kerne, die die Korrelationsstruktur der Originaldaten übernehmen, führt zu einer wesentlich genaueren Reproduktion mehrdimensionaler Verteilungen. Und je mehr Informationen über lineare Zusammenhänge zwischen den Variablen übernommen werden, d.h. keine Cluster gebildet werden, desto präziser werden Originalstrukturen abgebildet.

Werden robuste Standardfehler der Modellparameter geschätzt, verringern sich die Fehler, die durch univariates Resampling entstehen, im Mittel um etwa 11 Prozentpunkte (vgl. Tabelle 24.6).[38] Mit 39% bei kleinster und 44% bei größter Bandbreitenwahl sind sie dennoch deutlich größer als beim multivariaten Resampling.

Durch eine Extremwertbereinigung des Beobachtungsraums liegen die Abweichungen zwischen Original- und Resampleschätzern in den meisten Fällen um einige Prozentpunkte höher (vgl. Tabelle 24.6 mit Tabelle 24.7). Eine Verbesserung der Analyseergebnisse, d.h. eine Annäherung an das Originalergebnis, kann also durch diese Maßnahme nicht erzielt werden.

Die Abbildung der originalen Schätzergebnisse durch Resampling lässt sich verbessern, wenn alle für die Analyse benötigten Transformationen schon vor dem Resampling berechnet werden. Univariate Verteilungen können bei moderater Bandbreitenwahl erwartungstreu abgebildet werden. Ein positiver Effekt auf die Bewahrung mehrdimensionaler Strukturen ist zu erwarten. Eine Zusammenstellung der Ergebnisse wird in Tabelle 24.8 dargestellt. Die durchschnittlichen Fehler verringern sich gegenüber den vorherigen Analysen bei allen Varianten. Die Qualität der multivariaten Resamples wird dabei stärker verbessert. Wenn einfache Bandbreitenfaktoren verwendet werden, weichen die Koeffizientenschätzer im Mittel nur um etwa 3% von ihren Originalwerten ab.[39] Um die Eignung der Resamples für ökonometrische Analysen zu optimieren, ist es demnach sinnvoll, notwendige Variablentransformationen schon mit den Originaldaten durchzuführen.

38) Um Übersichtlichkeit zu bewahren, werden im Weiteren nur noch die Resultate für die jeweils kleinsten und größten Bandbreitenfaktoren aufgeführt, denn an der Dynamik der Abweichungen in Abhängigkeit von der Bandbreitenerhöhung ändert sich durch die im Folgenden betrachteten Modifikationen nichts.

39) Mit Extremwertbereinigung werden etwas höhere Fehlerquoten generiert, analog zu den Ergebnissen in Tabelle 24.7.

Tabelle 24.3: Cobb-Douglas-Produktionsfunktion – OLS-Schätzergebnisse mit den univariaten Resamples (ohne Extremwertbereinigung) der KSE 1999

Band-breiten-faktor	BBF= 1		BBF= 1,5		BBF= 1,7		BBF= 1,9		BBF= 2,1	
Variablen	Koeff.	(t-Wert)	Koeff.	(t-Wert)	Koeff.	(t-Wert)	Koeff.	(t-Wert)	Koeff.	(t-Wert)
Material-einsatz	0,425	(98,80)	0,416	(88,76)	0,422	(87,43)	0,422	(84,38)	0,420	(84,18)
Personal-kosten	0,361	(56,61)	0,358	(51,93)	0,355	(50,42)	0,346	(47,73)	0,341	(47,12)
Externe DL	0,070	(23,73)	0,071	(22,23)	0,071	(21,48)	0,074	(21,57)	0,075	(21,81)
Sonstige Kosten	0,113	(25,75)	0,125	(26,21)	0,120	(24,23)	0,126	(24,46)	0,130	(25,37)
Kapital-kosten	0,026	(6,76)	0,023	(5,70)	0,028	(6,78)	0,025	(5,98)	0,023	(5,61)
Konst.	1,524	(32,04)	1,551	(29,31)	1,507	(27,57)	1,574	(27,80)	1,640	(28,86)
	Relative Abweichungen von den Originaldaten in %									
Material-einsatz	2,48	48,61	0,38	53,83	1,78	54,52	1,87	56,11	1,26	56,21
Personal-kosten	6,49	43,45	5,51	48,13	4,78	49,64	1,94	52,32	0,50	52,93
Externe DL	19,45	40,87	22,72	44,61	22,39	46,47	27,22	46,25	28,45	45,65
Sonstige Kosten	0,16	49,03	10,33	48,12	5,32	52,04	10,53	51,58	14,42	49,78
Kapital-kosten	52,89	69,67	58,00	74,43	49,35	69,58	55,06	73,17	57,49	74,83
Konst.	15,55	59,12	14,06	62,60	16,50	64,82	12,77	64,53	9,12	63,17
Durchsch.	16,17	51,79	18,50	55,28	16,69	56,18	18,23	57,33	18,54	57,10

Tabelle 24.4: Cobb-Douglas-Produktionsfunktion – OLS-Schätzergebnisse mit den multivariaten Resamples (ohne Extremwertbereinigung) der KSE 1999

Band-breiten-faktor	BBF= 1		BBF= 1,5		BBF= 1,7		BBF= 1,9		BBF= 2,1	
Variablen	Koeff.	(t-Wert)	Koeff.	(t-Wert)	Koeff.	(t-Wert)	Koeff.	(t-Wert)	Koeff.	(t-Wert)
Material-einsatz	0,422	(180,10)	0,423	(173,34)	0,421	(176,64)	0,422	(174,83)	0,423	(171,99)
Personal-kosten	0,336	(92,75)	0,338	(89,79)	0,336	(91,50)	0,334	(90,04)	0,339	(90,29)
Externe DL	0,064	(40,86)	0,063	(38,43)	0,067	(41,56)	0,067	(41,04)	0,067	(40,39)
Sonstige Kosten	0,111	(45,42)	0,110	(43,36)	0,117	(46,76)	0,116	(45,82)	0,116	(45,00)
Kapital-kosten	0,044	(17,98)	0,046	(18,99)	0,037	(15,64)	0,038	(15,70)	0,033	(13,80)
Konst.	1,840	(72,44)	1,785	(67,29)	1,851	(71,29)	1,850	(70,28)	1,835	(68,30)
	Relative Abweichungen von den Originaldaten in %									
Material-einsatz	1,87	6,32	2,02	9,83	1,51	8,11	1,94	9,06	2,02	10,53
Personal-kosten	0,80	7,35	0,17	10,31	0,97	8,60	1,46	10,06	0,02	9,81
Externe DL	9,55	1,82	8,18	4,24	14,65	3,56	14,88	2,27	15,27	0,65
Sonstige Kosten	2,15	10,10	2,90	14,17	2,60	7,44	2,04	9,30	1,95	10,93
Kapital-kosten	19,95	19,34	16,21	14,80	32,54	29,83	31,89	29,56	41,10	38,09
Konst.	1,92	7,57	1,10	14,14	2,55	9,03	2,51	10,32	1,66	12,85
Durchschn.	6,04	8,75	5,10	11,25	9,14	11,10	9,12	11,76	10,34	13,81

Tabelle 24.5: Cobb-Douglas-Produktionsfunktion – OLS-Schätzergebnisse mit den multivariaten Resamples (mit Clusterbildung und ohne Extremwertbereinigung) der KSE 1999

Band-breiten-faktor	BBF= 1		BBF= 1,5		BBF= 1,7		BBF= 1,9		BBF= 2,1	
Variablen	Koeff.	t-Wert	Koeff.	(t-Wert)	Koeff.	(t-Wert)	Koeff.	(t-Wert)	Koeff.	(t-Wert)
Material-einsatz	0,424	(174,46)	0,429	(173,01)	0,428	(175,23)	0,428	(173,28)	0,429	(170,01)
Personal-kosten	0,329	(87,33)	0,327	(85,76)	0,328	(86,88)	0,326	(85,85)	0,331	(85,25)
Externe DL	0,064	(39,34)	0,062	(37,30)	0,063	(38,17)	0,064	(38,48)	0,063	(36,87)
Sonstige Kosten	0,122	(48,18)	0,122	(47,28)	0,126	(49,44)	0,127	(49,14)	0,126	(47,89)
Kapital-kosten	0,037	(14,52)	0,034	(13,97)	0,032	(13,10)	0,029	(11,97)	0,028	(11,69)
Konst.	1,867	(70,77)	1,883	(69,59)	1,848	(68,94)	1,869	(68,93)	1,834	(66,03)
	Relative Abweichungen von den Originaldaten in %									
Material-einsatz	2,34	9,25	3,51	10,00	3,36	8,85	3,41	9,86	3,47	11,56
Personal-kosten	2,90	12,77	3,66	14,33	3,30	13,22	3,71	14,24	2,35	14,84
Externe DL	9,42	1,97	6,64	7,05	7,57	4,88	9,85	4,11	7,53	8,12
Sonstige Kosten	7,29	4,63	7,47	6,41	10,79	2,14	11,81	2,73	11,25	5,21
Kapital-kosten	33,36	34,86	37,60	37,33	42,38	41,23	47,10	46,30	49,23	47,55
Konst.	3,44	9,70	4,31	11,20	2,38	12,03	3,55	12,05	1,63	15,75
Durchschn.	9,79	12,20	10,53	14,39	11,63	13,72	13,24	14,88	12,58	17,17

Tabelle 24.6: Cobb-Douglas-Produktionsfunktion – OLS-Schätzergebnisse mit robusten Standardfehlern (ohne Extremwertbereinigung) der KSE 1999

Verfahren	Univariat				Multivariat				Multivariate Cluster			
Band-breiten-faktor	BBF= 1		BBF= 2, 1		BBF= 1		BBF= 2, 1		BBF= 1		BBF= 2, 1	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Material-einsatz	0,425	(59,81)	0,420	(53,67)	0,422	(82,87)	0,423	(80,81)	0,424	(81,69)	0,429	(80,23)
Personal-kosten	0,361	(45,80)	0,341	(38,38)	0,336	(56,57)	0,339	(57,07)	0,329	(55,14)	0,331	(54,31)
Externe DL	0,070	(20,21)	0,075	(18,61)	0,064	(24,82)	0,067	(29,27)	0,064	(21,84)	0,063	(20,48)
Sonstige Kosten	0,113	(21,34)	0,130	(21,27)	0,111	(30,90)	0,116	(30,97)	0,122	(33,51)	0,126	(33,59)
Kapital-kosten	0,026	(6,03)	0,023	(5,27)	0,044	(13,26)	0,033	(11,80)	0,037	(11,48)	0,028	(9,86)
Konst.	1,524	(29,59)	1,640	(29,82)	1,840	(60,22)	1,835	(58,19)	1,867	(61,73)	1,834	(57,00)
	Relative Abweichungen von den Originaldaten in %											
Material-einsatz	2,48	24,31	1,26	32,08	1,87	4,87	2,02	2,27	2,34	3,38	3,47	1,53
Personal-kosten	6,49	22,20	0,50	34,81	0,80	3,91	0,02	3,06	2,90	6,34	2,35	7,75
Externe DL	19,45	30,95	28,45	36,42	9,55	15,20	15,27	0,00	9,42	25,38	7,53	30,03
Sonstige Kosten	0,16	39,05	14,42	39,25	2,15	11,74	1,95	11,54	7,29	4,28	11,25	4,06
Kapital-kosten	52,89	62,64	57,49	67,35	19,95	17,84	41,10	26,89	33,36	28,87	49,23	38,91
Konst.	15,55	55,45	9,12	55,10	1,92	9,33	1,66	12,39	3,44	7,06	1,63	14,18
Durchschn.	16,17	39,10	18,54	44,17	6,04	10,48	10,34	9,36	9,79	12,55	12,58	16,08

Tabelle 24.7: Cobb-Douglas-Produktionsfunktion – OLS-Schätzergebnisse mit robusten Standardfehlern (mit Extremwertbereinigung) der KSE 1999

Verfahren	Univariat				Multivariat				Multivariate Cluster			
Band-breiten-faktor	BBF= 1		BBF= 2,1		BBF= 1		BBF= 2,1		BBF= 1		BBF= 2,1	
Variablen	Koeff.	t-Werte	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Material-einsatz	0,422	(89,60)	0,414	(57,04)	0,441	(143,29)	0,437	(132,08)	0,442	(135,77)	0,446	(132,72)
Personal-kosten	0,334	(57,86)	0,324	(41,26)	0,327	(91,52)	0,335	(89,97)	0,318	(86,47)	0,322	(86,95)
Externe DL	0,066	(26,63)	0,073	(21,42)	0,059	(41,21)	0,063	(41,38)	0,062	(40,22)	0,061	(38,84)
Sonstige Kosten	0,120	(30,52)	0,130	(24,93)	0,106	(42,59)	0,114	(44,41)	0,117	(45,21)	0,122	(46,59)
Kapital-kosten	0,029	(8,96)	0,032	(8,35)	0,045	(16,87)	0,029	(13,78)	0,038	(16,68)	0,026	(12,78)
Konst.	1,934	(53,05)	1,943	(42,18)	1,779	(90,91)	1,767	(85,19)	1,823	(87,63)	1,798	(84,81)
	Relative Abweichungen von den Originaldaten in %											
Material-einsatz	3,06	40,25	4,92	61,96	1,47	4,44	0,57	11,92	1,55	9,46	2,43	11,49
Personal-kosten	3,71	42,54	0,66	59,02	1,55	9,11	4,13	10,65	1,15	14,12	0,15	13,65
Externe DL	27,65	32,26	41,15	45,51	14,13	4,83	21,97	5,27	18,26	2,31	17,67	1,20
Sonstige Kosten	14,49	34,94	23,95	46,86	1,69	9,21	8,74	5,33	11,69	3,62	16,65	0,68
Kapital-kosten	58,74	72,38	54,90	74,26	36,59	48,00	58,65	57,52	45,95	48,58	63,65	60,60
Konst.	12,64	45,80	13,16	56,91	3,63	7,12	2,94	12,96	6,19	10,47	4,76	13,35
Durchschn.	20,05	44,69	23,12	57,42	9,84	13,78	16,17	17,27	14,13	14,76	17,55	16,83

Tabelle 24.8: Cobb-Douglas-Produktionsfunktion – OLS-Schätzergebnisse mit vor dem Resampling logarithmierten Variablen (ohne Extremwertbereinigung) der KSE 1999

Verfahren	Univariat				Multivariat				Multivariate Cluster			
Band-breiten-faktor	BBF= 1		BBF= 2,1		BBF= 1		BBF= 2,1		BBF= 1		BBF= 2,1	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Material-einsatz	0,430	(101,95)	0,422	(88,62)	0,412	(176,93)	0,416	(174,94)	0,423	(181,90)	0,431	(166,91)
Personal-kosten	0,339	(54,09)	0,333	(48,51)	0,329	(90,06)	0,331	(90,30)	0,338	(91,93)	0,340	(84,69)
Externe DL	0,066	(23,26)	0,071	(21,66)	0,062	(38,67)	0,063	(38,74)	0,058	(37,22)	0,064	(37,13)
Sonstige Kosten	0,127	(29,16)	0,133	(27,21)	0,121	(49,95)	0,122	(49,37)	0,112	(45,92)	0,122	(44,85)
Kapital-kosten	0,025	(6,57)	0,025	(6,16)	0,056	(22,69)	0,045	(19,47)	0,048	(19,43)	0,020	(7,95)
Konst.	1,667	(36,02)	1,726	(31,96)	1,840	(73,04)	1,867	(72,31)	1,812	(70,97)	1,824	(64,28)
	Relative Abweichungen von den Originaldaten in %											
Material-einsatz	3,67	46,97	1,79	53,90	0,49	7,96	0,35	9,00	2,17	5,38	3,91	13,18
Personal-kosten	0,13	45,97	1,68	51,54	3,07	10,04	2,45	9,80	0,19	8,17	0,35	15,40
Externe DL	14,15	42,04	21,23	46,03	6,03	3,64	8,53	3,46	0,91	7,25	9,87	7,48
Sonstige Kosten	12,15	42,28	17,26	46,14	6,50	1,13	7,65	2,28	1,54	9,11	7,23	11,22
Kapital-kosten	55,52	70,52	55,00	72,36	1,04	1,79	18,31	12,65	13,92	12,83	64,29	64,33
Konst.	7,66	54,04	4,38	59,22	1,97	6,80	3,45	7,73	0,41	9,44	1,08	17,98
Durchschn.	15,55	50,30	16,89	54,87	3,18	5,23	6,79	7,49	3,19	8,70	14,46	21,60

24.4 Zusammenfassung

Nicht-parametrisches Resampling wurde in unterschiedlichen Varianten als Anonymisierungsmaßnahme für Mikrodaten vorgestellt und im Hinblick auf die Bewahrung des Analysepotenzials getestet. Es wurde zwischen univariatem und multivariatem Resampling differenziert und die Effekte steigender Bandbreiten, d.h. stärkerer Glättung, untersucht. Der multivariate Resampling-Algorithmus berücksichtigt mehrdimensionale Strukturen der Originaldaten, indem die Kernmatrix umskaliert wird, so dass sie der Korrelationsstruktur der Originaldaten folgt. Zwei alternative Methoden werden angewendet: Zum einen wird die gesamte Korrelationsmatrix als Skalierungsinstrument eingesetzt, zum anderen werden die Datenmerkmale in Cluster, d.h. in homogene Gruppen, unterteilt. Innerhalb dieser Cluster werden getrennt voneinander, also unabhängig, mehrdimensionale Resamples gezogen.

Multivariate Verteilungsparameter werden genauer reproduziert, wenn die Resamples mit dem mehrdimensionalen Algorithmus gezogen werden. Die Qualität von ökonometrischen Schätzungen wird dadurch verbessert. Dabei ist zu konstatieren, dass bei Einbeziehung möglichst vieler Informationen eine Verbesserung interdependenter Strukturen erreicht werden kann. In diesem Fall zeigt sich, dass eine einzige Korrelationsmatrix als Skalierungsinstrument vorzuziehen ist gegenüber einer Berechnung mehrerer Matrizen für die einzelnen Cluster. Ferner wird deutlich, dass bei vorheriger Berechnung von für die Schätzgleichung notwendigen Variablentransformationen, die Ergebnisse verbessert werden können. Allerdings ist die (allgemeine) Verfügbarkeit von Transformationen bei der praktischen Arbeit mit anonymisierten Daten keine realistische Annahme.

Kapitel 25

Latin Hypercube Sampling in linearen Modellen

25.1 Theoretische Eigenschaften

Latin Hypercube Sampling (LHS) ist ein Simulationsverfahren, bei dem die originalen Merkmalswerte so vertauscht werden, dass die Rangkorrelationsmatrix möglichst gut erhalten bleibt. Somit werden die univariaten Verteilungsmaße – wie Mittelwerte und Varianzen – erhalten. Allerdings wird die Korrelation zwischen den Variablen beeinträchtigt.

Explizite formale Ableitungen über das Verhalten der Schätzer in linearen Modellen – wie bei stochastischen Überlagerungen ober Mikroaggregationsverfahren – sind für das LHS-Verfahren nicht möglich. Es kann allerdings festgehalten werden, dass die Verzerrung von Schätzergebnissen in linearen – und vermutlich auch in nichtlinearen – Modellen entscheidend davon abhängt, wie stark die Korrelationsstruktur beziehungsweise die Varianz-Kovarianzmatrix durch die Anonymisierung mit LHS verändert wird.

25.2 Monte-Carlo-Simulationen

Den Simulationsexperimenten[40] liegt das folgende linearisierte Modell zugrunde:

$$log(Y) = \alpha + \beta_1 log(X_1) + \beta_2 log(X_2) + \beta_3 log(X_3) + U \qquad (25.1)$$

X_1, X_2 und X_3 sind lognormalverteilt. $log(X_1)$, $log(X_2)$ und $log(X_3)$ sind normalverteilt mit Erwartungswerten von Null und Varianz-Kovarianzmatrix

40) Die Ergebnisse zum Latin Hypercube Sampling verdankt das Projektteam der Diplomarbeit von Holger Herrmann. Das Programm zur Anwendung von LHS wurde von Kersten Magg, Universität Tübingen, bereitgestellt.

$$\mathbf{C} = \begin{pmatrix} 1 & 0,2 & 0,4 \\ 0,2 & 1 & 0,1 \\ 0,4 & 0,1 & 1 \end{pmatrix}.$$

Es wird nun verglichen, welche Wirkung die Anonymisierung mit Latin Hypercube Sampling auf die Schätzung dieses linearisierten Modells hat, wenn in einem Fall die Ausgangsvariablen (X_1, X_2 und X_3) und im anderen Fall die logarithmierten Variablen ($log(X_1)$, $log(X_2)$ und $log(X_3)$) anonymisiert werden. Die Zahl der Beobachtungen beträgt wiederum 1.000. Die Anzahl der Replikationen wird in diesem Fall auf 100 festgesetzt, um die Rechenzeit zu reduzieren. Analog zum Vorgehen bei den Simulationsexperimenten zur Untersuchung der Wirkung von stochastischen Überlagerungen und Mikroaggregationsverfahren wird mit Hilfe von t-Tests getestet, ob die mittels der mit LHS anonymisierten Daten geschätzten Koeffizienten denjenigen bei der Verwendung der Originaldaten geschätzten Koeffizienten entsprechen. Hierfür wird ein Signifikanzniveau von 5 Prozent zugrundegelgt. Die Ergebnisse sind in Tabelle 25.1 dargestellt.

Tabelle 25.1: MC-Simulationen – Lineares Modell mit logarithmierten Variablen, LHS, alle Variablen anonymisiert, 100 Replikationen

	Original	Ausgangsvariablen mit LHS anonymisiert			Logarithmen mit LHS anonymisiert		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert
X_1	1,0001	0,950			0,949		
X_2	-1,0014	-0,955			-0,954		
X_3	0,5001	0,544			0,546		
Konstante	0,9999	1,000	*	*	1,000	*	*

* Übereinstimmung zum Signifikanzniveau von 5%

Zunächst ist zu erkennen, dass die Ergebnisse in beiden Varianten ähnlich ausfallen. Unabhängig davon, ob die Ausgangsvariablen oder die Logarithmen anonymisiert werden, ergibt sich nur für die Konstante eine signfikante Übereinstimmung mit dem theoretischen Wert beziehungsweise mit dem Originalschätzer. Die Abweichungen der Koeffizientenschätzer bewegen sich für beide Varianten in einem ähnlichen Rahmen. Die Abweichungen der Koeffizientenschätzer sind zumindest im Durchschnitt – obwohl signifikant – nicht besonders hoch.

Etwas überraschend ist zunächst, dass in beiden Fällen ähnliche Ergebnisse erzeugt werden. Schließlich wäre eher zu vermuten gewesen, dass die nichtlineare Transformation des Logarithmus nach der Anonymisierung zu einer zusätzlichen Verzerrung der Schätzergebnisse führt. Hier ist allerdings die Funktionsweise des LHS zu bedenken. Da beim LHS für jedes

Merkmal die originalen Merkmalswerte repliziert werden, werden auch die Originalwerte der nichtlinearen Transformationen erhalten. Entscheidend ist nun, ob die Korrelationsstruktur der transformierten Variablen der Korrelationsstruktur der Ausgangsvariablen entspricht. Ist das der Fall, so wird die Korrelationsstruktur der transformierten Variablen in gleicher Weise verändert wie die Korrelationsstruktur der Ausgangsvariablen. Damit sind die Ergebnisveränderungen in beiden Fällen ähnlich.

Beim Logarithmus handelt es sich um eine streng monotone Transformation, somit entspricht die Rangkorrelation der Logarithmen exakt der Rangkorrelation der Ausgangsvariablen. Damit entspricht auch die Rangkorrelationsstruktur der anonymisierten Logarithmen der Rangkorrelationsstruktur der anonymisierten Ausgangsvariablen. Aus diesem Grund sind die Ergebnisse der Simulationsexperimente in beiden Varianten fast identisch.

25.3 Praxisbeispiele

Zur Untersuchung der Wirkung des Latin Hypercube Samplings in linearen Modellen wird beispielhaft wiederum die linearisierte Cobb-Douglas-Produktionsfunktion geschätzt, die bereits zur Untersuchung von stochastischen Überlagerungen, Mikroaggregationsverfahren und des Resamplings verwendet wurde. Die Datengrundlage ist die Kostenstrukturerhebung des Jahres 1999 ohne Wirtschaftszweig 37 (Recycling). Auf die Ausreißerbereinigung wird verzichtet.

Tabelle 25.2: Linearisierte Cobb-Douglas-Produktionsfunktion – Schätzergebnisse für mit LHS bearbeitete KSE-Daten, Daten nicht bereinigt, robuste Standardfehler

	Ausgangsvariablen mit LHS anonymisiert		Inputfaktoren und Output mit LHS anonymisiert		Inputfaktoren und Output mit LHS anonymisiert (Unternehmen vorher entfernt, für die bei einem Faktor der Logarithmus nicht definiert ist)		Logarithmen der Inputfaktoren und des Outputs mit LHS anonymisiert	
Variablen	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)	Koeff.	(t-Werte)
Materialeinsatz	0,478	(121,50)	0,348	(75,67)	0,354	(78,49)	0,425	(100,14)
Personalkosten	0,319	(50,02)	0,396	(63,76)	0,376	(60,24)	0,322	(50,24)
Externe Dienstleistungen	0,028	(15,05)	0,054	(25,94)	0,067	(29,43)	0,062	(26,44)
Sonstige Kosten	0,097	(30,11)	0,136	(32,71)	0,131	(33,87)	0,121	(27,60)
Kapitalkosten	0,06	(21,34)	0,066	(15,05)	0,069	(16,40)	0,055	(11,39)
Konstante	1,66	(32,38)	1,567	(28,48)	1,636	(30,30)	1,741	(32,48)
Anzahl Beobachtungen	13.700		16.272		16.251		16.251	
R^2	0,963		0,950		0,953		0,941	
	Relative Abweichungen von den Originalwerten in %							
Materialeinsatz	15,18	54,60	16,14	3,72	14,70	0,13	2,41	27,42
Personalkosten	5,90	14,66	16,81	8,79	10,91	2,78	5,01	14,28
Externe Dienstleistungen	51,72	48,14	6,90	10,61	15,52	1,41	6,90	8,89
Sonstige Kosten	14,16	13,72	20,35	6,28	15,93	2,95	7,08	20,92
Kapitalkosten	9,09	33,29	20,00	6,00	25,45	2,44	0,00	28,86
Konstante	17,34	51,08	13,14	56,97	9,31	54,22	4,15	50,93
Durchschn.	15,18	35,92	15,56	15,39	15,30	10,66	2,41	25,22

Verglichen werden die Schätzergebnisse für vier unterschiedliche Vorgehensweisen bei der Anonymisierung:

- 30 Ausgangsvariablen der KSE werden mit LHS anonymisiert.[41]
- Die fünf Inputfaktoren und der Output werden mit LHS anonymisiert.
- Die fünf Inputfaktoren und der Output werden mit LHS anonymisiert, nachdem alle Unternehmen entfernt wurden, für die bei einem der Inputfaktoren oder beim Output der Logarithmus nicht definiert ist.
- Die definierten Logarithmen der Inputfaktoren und des Outputs werden mit LHS anonymisiert.

Hintergrund dieses Vorgehens ist die Tatsache, dass sich nun die Ergebnisse bei der Anonymisierung der Ausgangsvariablen (erste Spalte) und der Anonymisierung der in die Schätzung eingehenden transformierten Variablen (letzte Spalte) deutlich unterscheiden. Der Grund hierfür besteht darin, dass durch die Transformation der bereits mit LHS bearbeiteten Daten zusätzliche „Missing Values“ in beträchtlichem Umfang entstehen, was man in Tabelle 25.2 daran erkennen kann, dass die Anzahl der Beobachtungen gegenüber der Originalschätzung von 16.251 auf 13.700 abnimmt. Ursächlich hierfür ist, dass sich der Output und die Kapitalkosten als Differenz aus mehreren Ausgangsvariablen berechnen. Durch das LHS werden zwar die einzelnen Werte der Ausgangsvariablen erhalten, die Zuordnung jedoch zerstört. Somit ändern sich auch die Differenzen. Bei einer größeren Anzahl von Unternehmen weisen Kapitalkosten und Ouptut nichtpositive Werte auf. Da der Logarithmus dann nicht definiert ist, können diese Unternehmen bei der Schätzung nicht berücksichtigt werden. Dies führt zu einer zusätzlichen Verzerrung der Schätzergebnisse. Entsprechend sind die Abweichungen der Koeffizientenschätzer von den Originalwerten bei der Anonymisierung der Ausgangsvariablen deutlich höher als bei der Anonymisierung der in die Regression eingehenden transformierten Variablen.

Als Zwischenstufe wird deshalb der Fall betrachtet, dass weder die Ausgangsvariablen noch die logarithmierten Merkmale, sondern die Inputfaktoren und der Ouptut direkt mit LHS anonymisiert werden. Hierbei ergeben sich jedoch ähnlich schlechte Ergebnisse wie bei der Anonymisierung der Ausgangsvariablen. Durch die Anwendung des LHS auf Inputfaktoren und Output entstehen Merkmalsträger mit einer neuen Zusammensetzung an Merkmalswerten. Die Zuordnung der einzelnen Merkmalswerte auf die Merkmalsträger wird gestört. Da einzelne Merkmalswerte für die Inputvariablen oder den Output kleiner gleich Null sind und somit die Logarithmen nicht definiert sind, verändert sich nicht nur die Korrelationsstruktur, sondern auch die Zusammensetzung der in die Schätzung einbezogenen Werte.

41) Gegenüber dem Projektbeginn wurden die Variablen *Tätige Inhaber*, *Angestellte und Arbeiter* sowie *Bestandsveränderungen an unfertiger und fertiger Ware aus eigener Produktion* entfernt.

Insgesamt werden nun weniger Merkmalsträger entfernt als bei der Berechnung der Logarithmen mit den Originaldaten.

Anonymisiert man zwar die Inputfaktoren und den Output vor der Logarithmierung, entfernt aber bereits vor der Anonymisierung diejenigen Unternehmen, für die durch die Logarithmierung „Missing Values“ entstehen, so sollte man wiederum ähnliche Ergebnisse erhalten wie bei der Anonymisierung der Logarithmen, weil es sich beim Logarithmus um eine streng monotone Tranformation handelt. Dies ist aber offenbar hier nicht der Fall (dritte Spalte der Tabelle 25.2). Möglicherweise hängt die Güte der Ergebnisse bei mit LHS anonymisierten Daten in Regressionsmodellen auch vom Zufall ab. Hier wären ergänzende Monte-Carlo-Simulationsexperimente mit Echtdaten hilfreich.

Zusammengefasst lässt sich somit festhalten, dass das Latin Hypercube Sampling für das lineare Modell dann zu einigermaßen vertretbaren Abweichungen der Koeffizientenschätzer gegenüber den Originalergebnissen führen kann, wenn die anonymisierten Merkmale auch direkt im zu schätzenden Modell berücksichtigt werden oder es sich bei den Modellvariablen um monotone Transformationen der Ausgangsvariablen handelt. Ein Modell mit transformierten Variablen, die einer nicht monotonen Transformation entstammen, kann bei einem mit LHS anonymisierten Datensatz nur geschätzt werden, wenn die transformierten Merkmale bereits bei der Anonymisierung berücksichtigt wurden. Da keine Korrekturverfahren vorhanden sind, stellt dies ein wesentliches Hindernis für die Verwendung von LHS zur Erstellung eines Scientific-Use-Files dar, der möglichst viele Nutzungsmöglichkeiten eröffnen soll.

Ein weiteres Problem besteht darin, dass Analysen für abgrenzbare Teilgesamtheiten nur möglich sind, wenn diese bei der Anwendung des LHS berücksichtigt wurden, die Anonymisierung also für die einzelnen Teilgesamtheiten getrennt erfolgt oder die zur Abgrenzung verwendeten kategorialen Variablen ebenfalls mit LHS behandelt wurden. Es ist leicht einsichtig, dass bei der Anwendung von LHS auf alle Unternehmen zwar die univariaten Verteilungsmaße und annähernd auch die Rangkorrelationen im Gesamtdatensatz erhalten bleiben, diese Verteilungsmaße in den einzelnen Teilgesamtheiten jedoch zerstört werden. Somit scheidet LHS als Verfahren zur Erzeugung von Scientific-Use-Files aus, wohl aber ist es ein Anonymisierungsverfahren, das ohne größeren Aufwand für die Anonymisierung von speziell auf bestimmte Analysen zugeschnittene Daten verwendet werden kann.

Kapitel 26

Ein Fazit für den Einsatz von datenverändernden Verfahren auf metrische Variablen in linearen und nichtlinearen Modellen

26.1 Bewertung einzelner datenverändernder Anonymisierungsverfahren

Im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ wurden für metrische Variablen die folgenden datenverändernden Verfahrensgruppen untersucht:

- Stochastische Überlagerungen
- Mikroaggregationsverfahren
- Simulationsverfahren (Resampling und Latin-Hypercube Sampling)

Aus ökonometrischer Sicht ergibt sich aufgrund der angestellten theoretischen Überlegungen, der durchgeführten Monte-Carlo-Simulationsexperimente und der empirischen Beispielrechnungen mit den Projektdaten folgende Bewertung der Verfahren (für lineare Modelle vgl. auch die Tabellen 26.1 für additive stochastische Überlagerungen, 26.2 für multiplikative stochastische Überlagerungen und 26.3 für Mikroaggregationsverfahren):

Tabelle 26.1: Additive stochastische Überlagerung und lineare Modelle

	Nur abhängige Variable stochastisch überlagert	Alle Variablen stochastisch überlagert	Nur alle Regressoren oder Teil der Regressoren oder Teil der Regressoren und abhängige Variable stochastisch überlagert
Einfache additive Überlagerung	Erwartungstreu	Nicht (asymptotisch) erwartungstreu (korrigierbar: z.B. Fuller, IV)	Nicht (asymptotisch) erwartungstreu (korrigierbar: z.B. Fuller, IV)
Additive Überlagerung (VCV-Matrix der Überlagerungen proportional zur VCV-Matrix der Originaldaten)	Erwartungstreu	Asymptotisch erwartungstreu	Nicht (asymptotisch) erwartungstreu (korrigierbar: z.B. Fuller, IV)
Additive Überlagerung und Transformationen zum Erhalt der ersten und zweiten Momente (Verfahren von Kim)	Erwartungstreu	Asymptotisch erwartungstreu (t-Werte erhalten)	Nicht (asymptotisch) erwartungstreu (korrigierbar) z.B. Fuller, IV)

Tabelle 26.2: Multiplikative stochastische Überlagerung und lineare Modelle

	Nur abhängige Variable stochastisch überlagert	Alle Variablen stochastisch überlagert	Nur alle Regressoren oder Teil der Regressoren oder Teil der Regressoren und abhängige Variable stochastisch überlagert
Multiplikative Überlagerung (mit Erwartungswert Eins), Kein konstanter Faktor	Erwartungstreu	Nicht (asymptotisch) erwartungstreu (korrigierbar: z.B. Hwang, IV)	Nicht (asymptotisch) erwartungstreu (korrigierbar: z.B. Hwang, IV)
Multiplikative Überlagerung (mit Erwartungswert Eins), Konstanter Faktor für jedes Merkmal eines Merkmalsträgers	Erwartungstreu	Unter bestimmten Bedingungen konsistent (kein Absolutglied oder Mittelwerte aller Regressoren gleich Null)	Nicht (asymptotisch) erwartungstreu (korrigierbar: z.B. Hwang, IV)

Tabelle 26.3: Mikroaggregation und lineare Modelle

	alle Variablen mikroaggregiert	Nur abhängige Variable mikroaggregiert	Nur alle unabhängigen Variablen mikroaggregiert	Nur Teil der unabh. oder Teil der unabh. und abh. mikroaggregiert
Getrennte Mikroaggregation (det. oder stoch.)	Nicht (asymptotisch) erwartungstreu	Nicht (asymptotisch) erwartungstreu	Nicht (asymptotisch) erwartungstreu. Ausnahme: nur eine unabh. Variable	Nicht (asymptotisch) erwartungstreu
Gemeinsame Mikroaggregation (det. von X oder stoch.)	Erwartungstreu	Nicht (asymptotisch) erwartungstreu	Erwartungstreu (Ausnahme: Bootstrap-Mikroaggregation)	Nicht (asymptotisch) erwartungstreu
Gemeinsame Mikroaggregation nach abhängiger Variablen oder nach Regressoren und abhängiger Variablen	Nicht (asymptotisch) erwartungstreu Korrigierbar, falls nur nach abhängiger Variablen	Nicht (asymptotisch) erwartungstreu	Nicht (asymptotisch) erwartungstreu Korrigierbar, falls nur nach abhängiger Variablen	Nicht (asymptotisch) erwartungstreu

1. Stochastische Überlagerungen

   (a) Additive stochastische Überlagerungen

      i. Keine Transformation zum Erhalt der ersten und zweiten Momente (nach Kim)

         - Wird lediglich die abhängige Variable mit einem Fehler überlagert, so sind die Schätzer erwartungstreu beziehungsweise konsistent.
         - Wird die Kovarianzmatrix der Überlagerungen nicht proportional zur Kovarianzmatrix der Originalwerte gewählt, so sind die Schätzer in linearen Regressionsmodellen verzerrt. Hierfür existieren jedoch Korrekturmöglichkeiten, wie Instrumentenvariablen-Schätzer und Fehler-in-den-Variablen-Modelle.
         - Wird die Kovarianzmatrix der Überlagerungen hingegen proportional zur Kovarianzmatrix der Originalwerte gewählt, so sind die Schätzer in linearen Regressionsmodellen asymptotisch erwartungstreu. Die Varianz der Störgrößen wird jedoch überschätzt.
         - Die Schätzer in nichtlinearen Modellen oder bei nichtlinearen Transformationen in linearen Modellen sind grundsätzlich verzerrt. Jedoch existieren auch hier Korrekturmöglichkeiten wie z.B. der SIMEX-Schätzer. Als problematisch kann sich dabei erweisen, dass durch die additive stochastische Überlagerung Vorzeichenwechsel auftreten können und nichtlineare Transformationen nicht mehr definiert sind. Hierdurch auftretende Verzerrungen können nicht korrigiert werden.
         - Bei einer additiven Überlagerung mit der gleichen Zufallszahl über alle Merkmalswerte einer Einheit ist der SIMEX-Schätzer nicht anwendbar.

      ii. Transformation zum Erhalt der ersten und zweiten Momente (Verfahren von Kim)

         - Wird lediglich die abhängige Variable mit einem Fehler überlagert, so sind die Schätzer erwartungstreu beziehungsweise konsistent.
         - Werden alle Variablen anonymisiert, so sind die Schätzer in linearen Regressionsmodellen asymptotisch erwartungstreu. Die Varianzen der Störgrößen bleiben ebenfalls erhalten.
         - Die Schätzer in nichtlinearen Modellen oder bei nichtlinearen Transformationen in linearen Modellen sind grundsätzlich verzerrt. Die Verzerrungen können jedoch durch Korrekturverfahren für nichtlineare Modelle korrigiert werden, allerdings verschärfen sich die ohne die Kim-Transformation auftretenden Probleme, weil die Häufigkeit von Vorzeichenwechseln ansteigt.

   (b) Multiplikative stochastische Überlagerungen

      - Für multiplikative stochastische Überlagerungen gelten grundsätzlich die gleichen Ergebnisse wie für additive stochastische Überlagerungen: Die

Schätzer sind erwartungstreu bzw. konsistent, sofern ausschließlich die abhängige Variable überlagert wird. Ansonsten sind sie verzerrt.

- Erfolgt die Überlagerung mit einem pro Merkmalsträger über alle Merkmale konstanten Faktor, so sind die Schätzer im linearen Regressionsmodell konsistent, wenn ein Modell ohne Absolutglied geschätzt wird oder die Mittelwerte aller Regressoren Null sind.
- Wird eine durch Logarithmierung linearisierte Funktion geschätzt und wird der multiplikative Fehler mit Erwartungswert Eins zum additiven Fehler mit Erwartungswert Null, so sind Korrekturverfahren für additive Fehler anwendbar.
- Multiplikative stochastische Überlagerungen haben gegenüber additiven stochastischen Überlagerungen den Vorteil, dass bei einer hier einfach möglichen Beschränkung des Überlagerungsfaktors auf positive Werte keine Vorzeichenwechsel möglich sind. Außerdem bleiben (strukturelle) Nullen grundsätzlich erhalten. Beides führt dazu, dass die Definitionsbereiche für nichtlineare Transformationen erhalten bleiben. Somit können die Fehler in linearen und nichtlinearen Regressionsmodellen in der Regel auch dann korrigiert werden, wenn die Variablen vorher transformiert wurden.
- Bei der Anwendung von Korrekturverfahren bei multiplikativen Fehlern kann es in der Praxis jedoch zu Problemen kommen, wenn die Varianz der Überlagerungen zu hoch gewählt ist. Daneben zeigen Simulationsexperimente, dass die Wirkung der multiplikativen stochastischen Überlagerung auch vom Startwert bei der Erzeugung der Zufallsfehler abhängt.
- Bei einem konstanten Überlagerungsfaktor und Linearisierung des Modells durch Logarithmierung entstehen gleiche additive Fehler für alle Merkmalswerte einer Einheit. Dann ist der SIMEX-Schätzer nicht anwendbar.

2. Mikroaggregationsverfahren

   (a) Abstandsorientierte Mikroaggregationsverfahren

      i. Gemeinsame Mikroaggregation

         - Die Schätzer in linearen ökonometrischen Modellen sind erwartungstreu, sofern nicht die abhängige Variable bei der Berechnung des Abstandsmaßes zur Gruppenbildung für die Mikroaggregation berücksichtigt wird. Werden alle Variablen berücksichtigt, so ist die Verzerrung aufgrund des geringeren Einflusses der abhängigen Variablen gering, allerdings ist das Problem bei jeder Modellspezifikation vorhanden. Eine Alternative bestünde darin, den Abstand nur nach einer dominierenden Variable (zum Beispiel „Umsatz“ oder „Beschäftigung“) zu berechnen. Dann wäre der Schätzer nur verzerrt, wenn die abhängige Variable „Umsatz“ oder „Beschäftigung“ wäre.
         - Werden nur erklärende Variablen (unabhängig von der abhängigen Variablen) mikroaggregiert, so sind die Schätzer linearer Modelle in jedem

Fall erwartungstreu. Es ist jedoch bei der Erstellung eines Scientific-Use-Files nicht bekannt, welche Variablen ausschließlich als erklärende eingesetzt werden.

- Die Teststatistiken sind verzerrt, können aber korrigiert werden, sofern erwartungstreue Schätzer vorliegen und das Aggregationsniveau bekannt ist. Die Verzerrung ist umso geringer, je ähnlicher sich die Werte innerhalb einer Gruppe sind.
- Die Schätzer in nichtlinearen Modellen oder bei nichtlinearen Transformationen in linearen Modellen sind grundsätzlich verzerrt.

ii. Teilweise gemeinsame (gruppierte) Mikroaggregation

- Die Schätzer in linearen ökonometrischen Modellen sind nicht erwartungstreu beziehungsweise konsistent. In Modellrechnungen ergeben sich weitaus stärkere Abweichungen als bei der getrennten Mikroaggregation.
- Die Schätzer in nichtlinearen ökonometrischen Modellen oder bei nichtlinearen Transformationen in linearen Modellen sind nicht konsistent beziehungsweise erwartungstreu.
- Die Teststatistiken sind verzerrt, können auch nicht korrigiert werden.

iii. Getrennte Mikroaggregation

- Die Schätzer in linearen ökonometrischen Modellen sind nicht erwartungstreu, allerdings ergeben sich aufgrund der geringeren Veränderung der Einzelwerte insbesondere bei guter Modellspezifikation und einer hohen Korrelation zwischen abhängiger und erklärenden Variablen nur geringfügige Veränderungen der geschätzten Koeffizienten.
- Die Schätzer in nichtlinearen ökonometrischen Modellen oder bei nichtlinearen Transformationen in linearen Modellen sind nicht konsistent beziehungsweise erwartungstreu, allerdings zeigen Simulationsstudien und empirische Untersuchungen, dass sich aufgrund der geringeren Veränderung der Einzelwerte insbesondere bei guter Modellspezifikation und einer hohen Korrelation zwischen abhängiger und erklärenden Variablen nur geringfügige Veränderungen der geschätzten Koeffizienten ergeben.
- Die Teststatistiken sind bei den durchgeführten Modellrechnungen nur geringfügig verzerrt.

(b) Stochastische Mikroaggregationsverfahren

i. Zufällige Mikroaggregation (gemeinsame Mikroaggregation,[42] Gruppenbildung zufällig)

- Die Schätzer in linearen ökonometrischen Modellen sind grundsätzlich erwartungstreu, sofern abhängige und erklärende Variablen gemeinsam

42) Die getrennte zufällige Mikroaggregation führt eindeutig zu nicht akzeptablen Verzerrungen und wird daher vernachlässigt.

mikroaggregiert werden oder lediglich erklärende Variablen gemeinsam mikroaggregiert werden.

- Die Teststatistiken sind verzerrt, können aber korrigiert werden, sofern das Aggregationsniveau bekannt ist. Die Verzerrung ist umso geringer, je ähnlicher sich die Werte innerhalb einer Gruppe sind.
- Die Schätzer in nichtlinearen Modellen oder bei nichtlinearen Transformationen in linearen Modellen sind grundsätzlich verzerrt.

ii. Bootstrap-Mikroaggregation

- Die Schätzer in linearen ökonometrischen Modellen sind grundsätzlich erwartungstreu, sofern abhängige und erklärende Variablen gemeinsam mikroaggregiert werden. Für die Schätzung der Standardfehler sollte der heteroskedastie-robuste Varianz-Kovarianzschätzer verwendet werden. Demgegenüber sind die Schätzer nicht erwartungstreu, sofern lediglich die erklärenden Variablen mikroaggregiert werden.
- Die Schätzer in nichtlinearen Modellen oder bei nichtlinearen Transformationen in linearen Modellen sind grundsätzlich verzerrt.

3. Simulationsverfahren

(a) Resampling

- Für mit Resampling bearbeitete Daten lassen sich theoretisch recht gute Ergebnisse bei linearen und nichtlinearen Schätzungen erzielen, sofern die im Modell berücksichtigten Variablen auch direkt anonymisiert wurden. Werden hingegen nichtlineare Transformationen verwendet, so sind die Schätzergebnisse in jedem Fall verzerrt.
- Bei der multivariaten Variante des Resamplings schneidet dasjenige Vorgehen am besten ab, bei dem die Korrelationsmatrix angepasst wird.
- Allerdings zeigen Modellrechnungen mit der KSE, dass auch bei Berücksichtigung der im Modell verwendeten nichtlinear transformierten Variablen Abweichungen von den Originalergebnissen beobachtet werden können. Diese waren in den durchgeführten Berechnungen durchweg größer als beispielsweise bei der getrennten Mikroaggregation.
- Auch qualitative Variablen, die als Dummy-Variablen in eine Schätzung eingehen können, müssen bei der Durchführung des Resamplings berücksichtigt werden oder der Datensatz muss vor Durchführung des Verfahrens nach diesen qualitativen Variablen aufgeteilt werden, damit sich erwartungstreue beziehungsweise konsistente Schätzergebnisse ermitteln lassen. Hier tritt wiederum das Problem auf, dass zum Zeitpunkt der Anonymisierung nicht bekannt ist, welche qualitativen Variablen hierbei berücksichtigt werden müssen.

(b) Latin Hypercube Sampling

- Für mit LHS bearbeitete Daten lassen sich recht gute Ergebnisse bei linearen und nichtlinearen Schätzungen erzielen, sofern die im Modell berücksichtigten Variablen auch direkt anonymisiert werden. Werden hingegen nicht monotone Transformationen verwendet, so sind die Schätzergebnisse in jedem Fall verzerrt.
- Das Gleiche gilt für qualitative Variablen, die als Dummy-Variablen in eine Schätzung eingehen können. Nur sofern diese qualitativen Variablen bei der Durchführung von LHS berücksichtigt wurden beziehungsweise der Datensatz vor Durchführung des Verfahrens nach diesen qualitativen Variablen aufgeteilt wurde, lassen sich auch erwartungstreue beziehungsweise konsistente Schätzergebnisse ermitteln. Hier ist wiederum zum Zeitpunkt der Anonymisierung nicht bekannt, welche qualitativen Variablen unbedingt berücksichtigt werden müssen.

26.2 Schlussfolgerungen für den Einsatz der Verfahren zur Erstellung von Scientific-Use-Files

Aus der im vorangegangenen Abschnitt vorgenommenen Bewertung der wesentlichen datenverändernden Verfahren und Verfahrensgruppen für metrische Variablen aus Sicht ökonometrischer Auswertungen lassen sich die folgenden Schlussfolgerungen für den Einsatz dieser Verfahren zur faktischen Anonymisierung wirtschaftsstatistischer Einzeldaten und damit zur Erstellung von Scientific-Use-Files ableiten:

- Die Varianten der gemeinsamen Mikroaggregation (abstandsorientierte gemeinsame Mikroaggregation, zufällige Mikroaggregation, Bootstrap-Mikroaggregation) sind ausschließlich für lineare Regressionsmodelle geeignet. Die Konsistenz der Schätzer geht in nichtlinearen Modellen oder bei nichtlinearen Transformationen verloren. Die teilweise gemeinsame (gruppierte) Mikroaggregation ist weder für lineare noch für nichtlineare Modelle geeignet. Die abstandsorientierte getrennte Mikroaggregation führt zwar nicht zu erwartungstreuen bzw. konsistenten Schätzern, allerdings aufgrund der gut erhaltenen Korrelationsstruktur und der geringen Abweichung der Einzelwerte sowohl bei linearen als auch bei nichtlinearen Regressionsmodellen nur zu geringen Abweichungen der Koeffizientenwerte, das Gleiche gilt für die Teststatistiken. Damit können sowohl lineare als auch nichtlineare Modelle in der Regel mit derart anonymisierten Daten geschätzt werden.
- Aufgrund der existierenden Korrekturverfahren in linearen und nichtlinearen Modellen sind sowohl additive als auch multiplikative stochastische Überlagerungen grundsätzlich aus der Sicht der ökonometrischen Nutzung für die Anonymisierung geeignet. Werden ausschließlich lineare Modelle mit nicht oder nur linear transformierten Variablen geschätzt, so ist das Verfahren von Kim zu präferieren, weil sowohl die Parameterschätzer als auch die Teststatistiken erhalten werden. Werden allerdings auch

nichtlineare Transformationen verwendet und nichtlineare Modelle geschätzt, so ist eine multiplikative stochastische Überlagerung vorzuziehn. Um in jedem Fall die Anwendbarkeit der SIMEX-Korrektur sicherzustellen und trotzdem mit annähernd gleichen Faktoren überlagern zu können, ist eine zweigipflige Mischungsverteilung nach dem Verfahren von Höhne zu empfehlen.

- Simulationsverfahren sind aus Sicht ökonometrischer Auswertungen zur Anonymisierung geeignet, sofern die in eine Untersuchung eingehenden transformierten Variablen und die bei der Untersuchung zu berücksichtigenden Teilgesamtheiten bekannt sind und bei der Anonymisierung berücksichtigt werden können. Dies ist jedoch bei einem für eine Vielzahl von Nutzungen gedachten Scientific-Use-File nicht möglich. Simulationsverfahren sind daher eher geeignet, wenn ein Nutzer einen bestimmten Datensatz für eine bestimmte vorab festgelegte Auswertung bestellt. Da das Resampling-Verfahren jedoch als eine Art der additiven stochastischen Überlagerung intepretiert werden kann, ist zu untersuchen, ob hierfür entwickelte Fehler-Korrektur-Modelle auch beim Resampling angewendet werden können. Dann wäre das Verfahren möglicherweise auch für die Erzeugung von Scientific-Use-Files einsetzbar.

Teil IX

Wirkung der Post-Randomisierung in ausgewählten Modellen

In diesem Teil wird die Wirkung der Post-Randomisierung in ausgewählten Modellen untersucht. Kapitel 27 behandelt die Post-Randomisierung der abhängigen Variablen in einem binären Probit-Modell. Kapitel 28 beschäftigt sich mit der Post-Randomisierung von erklärenden diskreten Variablen im Modell der Varianzanalyse. Zuletzt werden in Kapitel 29 die Auswirkungen der Post-Randomisierung von erklärenden Dummy-Variablen im Probit-Modell untersucht. Kapitel 30 zieht ein Fazit für den Einsatz der Post-Randomisierung zur Anonymisierung unter Beachtung der untersuchten Modelle.

Kapitel 27

Post-Randomisierung der abhängigen Variablen im Probit-Modell

Im Rahmen dieses Kapitels werden die Auswirkungen der Post-Randomisierung im Probit-Modell untersucht. Betrachtet wird zunächst in Abschnitt 27.1 die ausschließliche Anonymisierung der binären abhängigen Variablen durch die Post-Randomisierung. Hierzu werden in Unterabschnitt 27.1.1 einige theoretische Eigenschaften abgeleitet, die anschließend mittels Monte-Carlo-Simulationsexperimenten in Unterabschnitt 27.1.2 und mit Daten des IAB-Betriebspanels in Unterabschnitt 27.1.3 überprüft werden.

In Abschnitt 27.2 wird die Post-Randomisierung der abhängigen Variablen mit einer Anonymisierung der erklärenden Variablen mittels stochastischer Überlagerung oder Mikroaggregation verbunden. Hierzu werden ebenfalls Monte-Carlo-Simulationen und Untersuchungen mit dem IAB-Betriebspanel durchgeführt.

27.1 Ausschließliche Post-Randomisierung der abhängigen Variablen im Probit-Modell

27.1.1 Theoretische Eigenschaften

Das binäre Probit-Modell (Ronning 1991; Greene 2003) beschreibt den Effekt einer oder mehrerer erklärender Variablen X_i auf eine latente metrische Variable Y^*. Für den Fall einer Einflussgröße X handelt es sich um ein latentes lineares Modell der Form

$$Y^* \quad = \quad \alpha + \beta X + U \tag{27.1}$$

Dabei ist U wiederum ein standardnormalverteilter Störterm.

Allerdings ist Y^* lediglich als dichotome Variable Y beobachtbar. Für Y gilt:

$$Y \quad = \quad \begin{cases} 1 & \text{if} \quad Y^* > 0 \\ 0 & \text{if} \quad Y^* \leq 0 \end{cases} \quad . \tag{27.2}$$

Dabei nimmt Y mit der Wahrscheinlichkeit ϕ den Wert 1 an, mit der Wahrscheinlichkeit $1-\phi$ den Wert Null. Die für die Schätzung notwendigen Informationen sind durch n Paare $(x_i\,,\,y_i)$ gegeben, wobei $y_i \in \{0,1\}$.

Die Schätzung der beiden unbekannten Parameter α und β erfolgt durch Maximierung der Likelihoodfunktion

$$\mathcal{L}(\alpha,\beta|(y_i,x_i), i=1,\ldots,n) \;=\; \prod_{i=1}^{n} \Phi_i^{y_i}\,(1-\Phi_i)^{(1-y_i)}\,. \tag{27.3}$$

Φ_i bezeichnet die bedingte Wahrscheinlichkeit unter der Normalverteilungsannahme, dass die Variable Y_i für gegebenes X_i den Wert 1 annimmt, $\Phi_i \;\equiv\; \Phi(\alpha + \beta X_i) \;=\; P(Y_i^* > 0 \mid X_i)$.

Wird nun die dichotome abhängige Variable Y durch eine Post-Randomisierung mit symmetrischer Übergangsmatrix, wie in Unterabschnitt 6.1.2 dargestellt, anonymisiert, so gilt für die beobachtbare anonymisierte Variable Y_i^a:

$$Y_i^a \quad = \begin{cases} 1 & \text{mit Wahrscheinlichkeit} \quad \Phi_i\,\pi \;+\; (1-\Phi_i)\,(1-\pi) \\ & \\ 0 & \text{mit Wahrscheinlichkeit} \quad \Phi_i\,(1-\pi) \;+\; (1-\Phi_i)\,\pi \end{cases} \tag{27.4}$$

Verwendet man trotz der Post-Randomisierung der abhängigen Variablen den „naiven" Probit-Schätzer, berücksichtigt also die Übergangsmatrix nicht in der Likelihoodfunktion, so erhält man einen verzerrten Schätzer. Die Verzerrung kann für die marginalen Effekte wie folgt gezeigt werden (Ronning und Rosemann 2004).

Der wahre marginale Effekt lautet (Greene 2003):

$$\frac{\partial\, P(Y \,=\, 1\,|\,x)}{\partial\, x} \quad = \quad \phi(\alpha + \beta x)\,\beta \tag{27.5}$$

Aus Gleichung (27.4) wissen wir, dass im Fall der mit Post-Randomisierung anonymisierten Daten für die bedingte Wahrscheinlichkeit, dass Y^a für gegebenes X den Wert Eins annimmt, gilt:

$$P(Y^a = 1 \mid X) \quad = \quad \Phi_i \pi \; + \; (1 - \Phi_i)(1 - \pi) \; = \; (2\pi - 1)\Phi_i \; + \; (1 - \pi) \qquad (27.6)$$

Daraus folgt für den marginalen Effekt des „naiven" Schätzers:

$$\frac{\partial \, P(Y^a = 1 \mid X)}{\partial \, X} \quad = \quad (2\,\pi \; - \; 1)\; \phi(\alpha + \beta X)\,\beta \qquad (27.7)$$

oder

$$\frac{\partial \, P(Y^a = 1 \mid X)}{\partial \, X} \quad = \quad (2\,\pi \; - \; 1)\; \frac{\partial \, P(Y = 1 \mid X)}{\partial \, X} \qquad (27.8)$$

Daraus folgt, dass der „naive" Schätzer den marginalen Effekt unterschätzt, sofern für die Bleibewahrscheinlichkeit π gilt:

$$\frac{1}{2} \quad < \quad \pi \quad < \quad 1. \qquad (27.9)$$

Man erhält sogar das verkehrte Vorzeichen des Effekts, wenn die Bleibewahrscheinlickeit geringer als 50 Prozent ist, also für π gilt:

$$0 \quad < \quad \pi \quad < \quad \frac{1}{2} \quad . \qquad (27.10)$$

Allerdings kann man die Randomisierung in der Schätzung des Probit-Modells berücksichtigen. Dann gilt für die zu maximierende Likelihoodfunktion (Ronning 2005; Ronning und Rosemann 2004; Ronning et al. 2005):

$$\begin{aligned} &\mathcal{L}(\alpha, \beta | (y_i^a, x_i), i = 1, \ldots, n) \\ &= \prod_{i=1}^{n} \left(\Phi_i \pi + (1 - \Phi_i)(1 - \pi)\right)^{y_i^a} \left(\Phi_i (1 - \pi) + (1 - \Phi_i)\pi\right)^{(1 - y_i^a)} . \end{aligned} \qquad (27.11)$$

und für die entsprechende Log-Likelihoodfunktion:

$$\begin{aligned} &L \equiv log(\mathcal{L}) \\ &= \sum_{i=1}^{n} y_i^a \, log\left[\Phi_i \pi + (1 - \Phi_i)\,(1 - \pi)\right] + (1 - y_i^a)\, log\left([\Phi_i (1 - \pi) + \right. \\ &\left. (1 - \Phi_i)\,\pi)\right] \end{aligned} \qquad (27.12)$$

Ronning (2005) ermittelt die folgende Hesse-Matrix aus den partiellen Ableitungen nach den Parametern α und β:

$$H^a = -(2\pi - 1) \sum_i \frac{g_i \phi_i}{(W_i(1 - W_i))^2} \mathbf{u}_i \mathbf{u}_i' \tag{27.13}$$

mit

$$W_i \equiv \pi\Phi_i + (1 - \pi)(1 - \Phi_i) = (2\pi - 1)\Phi_i + (1 - \pi), \tag{27.14}$$

$$g_i = (2\pi - 1)\left(y_i^a - 2y_i^a W_i + W_i^2\right)\phi_i + (y_i^a - W_i)\, W_i\, (1 - W_i) \times (\alpha + \beta x_i)\,, \tag{27.15}$$

$$\mathbf{u}_i = (\mathbf{1} \quad \mathbf{x}_i) \tag{27.16}$$

und ϕ_i der zu Φ_i korrespondierenden Dichte.

Man erhält damit durch die Berücksichtigung der Randomisierung in der Likelihoodfunktion konsistente Schätzer für die Parameter α und β. Allerdings ist die Hesse-Matrix für eine Bleibewahrscheinlichkeit von $\pi = 0,5$ nicht definiert. Für diesen Fall existiert folglich keine Lösung für das Maximierungsproblem.

Für die Informationsmatrix gilt im Fall der Post-Randomisierung:

$$\mathcal{I}^a = (2\pi - 1)^2 \sum_i \frac{\phi_i^2}{W_i(1 - W_i)} \mathbf{u}_i \mathbf{u}_i', \tag{27.17}$$

während im Fall unmaskierter Daten ($\pi = 1$) für die Informationsmatrix gilt:

$$\mathcal{I} = (2\pi - 1)^2 \sum_i \frac{\phi_i^2}{\Phi_i(1 - \Phi_i)} \mathbf{u}_i \mathbf{u}_i'. \tag{27.18}$$

Ronning (2005) zeigt, dass die Differenz $\mathcal{I} - \mathcal{I}^a$ nichtnegativ definit ist, folglich mit der Post-Randomisierung ein Effizienzverlust einhergeht. Dieser ist am größten, wenn π Werte nahe bei 1/2 annimmt.

Die Berücksichtigung der Post-Randomisierung in der Likelihoodfunktion ist auch möglich, wenn die Übergangsmatrix nicht symmetrisch gewählt wird. Es gilt dann für zwei ungleiche Bleibewahrscheinlichkeiten π_1 und π_0:

$$\begin{aligned}&\mathcal{L}(\alpha, \beta | (y_i^a, x_i), i = 1, \ldots, n) \\ &= \prod_{i=1}^{n} (\Phi_i \, \pi_1 + (1 - \Phi_i)(1 - \pi_0))^{y_i^a} \, (\Phi_i \, (1 - \pi_1) + (1 - \Phi_i)\pi_0)^{(1 - y_i^a)} \, . \quad (27.19)\end{aligned}$$

Eine besondere Variante der nicht symmetrischen Post-Randomisierung ist das in Unterabschnitt 6.1.2 ebenfalls dargestellte invariante PRAM. Hier gilt für die Bleibewahrscheinlichkeiten (Ronning und Rosemann 2004):

$$\pi_1 \quad = \quad P(Y^* = 1 | Y = 1) \quad = \quad \lambda \, + \, (1 \, - \, \lambda) \, \theta \qquad (27.20)$$

und

$$\pi_0 \quad = \quad P(Y^* = 0 | Y = 0) \quad = \quad 1 \, - \, (1 \, - \, \lambda) \, \theta. \qquad (27.21)$$

θ ist dabei der Anteil der Beobachtungen, die den Wert 1 annehmen. Dieser Anteil hängt im Probit-Modell von den erklärenden Variablen ab. In der Praxis muss θ_j für die einzelnen Kategorien durch den jeweiligen Anteil in der Stichprobe geschätzt werden:

$$\hat{\theta}_j \quad = \quad \frac{m_j}{n} \qquad (27.22)$$

mit m_j als dem Anteil der Merkmalsträger in der Stichprobe in Kategorie j.

Für den hier dargestellten Fall mit zwei Kategorien gilt für die Schätzung des konstanten Parameters θ:

$$\hat{\theta} \quad = \quad \frac{m}{n}. \qquad (27.23)$$

Dabei ist m der Anteil der Merkmalsträger mit dem Wert 1 in der Stichprobe mit Umfang n.

Eine entsprechende konsistente Schätzung der unbekannten Parameter durch die Berücksichtigung der Post-Randomisierung in der zu maximierenden Likelihoodfunktion ist analog auch für eine höhere Anzahl an Kategorien ($r > 2$) möglich. Die gleichen Ergebnisse gelten auch für mehrere Einflussgrößen $X_1, ..., X_n$.

Im Folgenden werden die theoretischen Ergebnisse mit Hilfe von Monte-Carlo-Simulationen und Praxisbeispielen anhand des IAB-Betriebspanels überprüft.

27.1.2 Monte-Carlo-Simulationen

In diesem Unterabschnitt wird eine kleine Simulationsstudie zur Wirkung des in Unterabschnitt 27.1.1 vorgestellten Korrekturschätzers bei Post-Randomisierung der abhängigen Variablen im Probit-Modell dargestellt.

Zu schätzen ist das „latente" Modell

$$Y^* = \alpha + \beta X + U$$

mit $E(U) = 0, var(U) = \sigma_u^2 = 1$. Die beobachtete abhängige Variable Y ist durch

$$Y \quad = \quad \begin{cases} 1 & \text{falls } Y^* > 0 \\ 0 & \text{sonst} \end{cases}$$

gegeben.

In den Simulationsstudien wurden folgende Parameter festgelegt:

X	$\sim N(4,35\,,\ 1,75^2)$	
U	$\sim N(0,1)$	
α	unbek. Parameter	$\alpha = 2,500$
β	unbek. Parameter	$\beta = 0,600$

Die abhängige binäre Variable wird mit einfachem sowie mit invariantem PRAM anonymisiert. Dabei sollen die Parameter π beziehungsweise λ als bekannt vorausgesetzt werden.

Alternativ werden zwei Stichprobenumfänge $n = 3.000$ and $n = 10.000$ verwendet. Die Ergebnisse basieren auf Monte-Carlo-Simulationen mit 1.000 Replikationen. Für jede der beiden Stichprobengrößen wird ein Vektor $\mathbf{x} = (x_1\,,\ x_2,\ \ldots\ ,x_n)$ erzeugt, der für alle Simulationsläufe konstant gehalten wird.

Für einfaches PRAM werden die Bleibewahrscheinlichkeiten $\pi = 0,80$ und $\pi = 0,60$ verwendet.

Für invariantes PRAM werden alternativ $\lambda = 0,80$ und $\lambda = 0,60$ getestet.

Aus

$$E[Y^*|X = x] \ = \ -2,5 \ + \ 0,6\,x$$

erhält man

$$E[Y^*] \ = \ -2,5 \ + \ 0,6 \cdot 4,35 \ = \ 0,11.$$

Weiterhin gilt

$$var[Y^*] \;=\; Var_x[E[Y^*|X=x]] + E_x[var[Y^*|X=x]] = 0,6^2 \cdot 1,75^2 \;+\; 1,0 \;=\; 2,1025\;.$$

Daraus folgt, dass die Verteilung der Zufallsvariable Y^* gegeben ist durch

$$Y^* \quad \sim \quad N(0,11\;,\;2,1025)$$

und es gilt

$$P(Y^* \;\leq\; 0) \quad = \quad \Phi\left(\frac{0,11}{\sqrt{2,1025}}\right) \quad = \quad \Phi(0,077) \;=\; 0,53067.$$

Damit ergibt sich die Wahrscheinlichkeit eine 1 zu beobachten als $\theta \;=\; 0,4693$.

Mit $\lambda \;=\; 0,8$ und $\lambda \;=\; 0,6$ erhält man für die Bleibewahrscheinlichkeiten im Fall des invarianten PRAM

$$\pi_1 \quad = \quad \begin{cases} 0,8939, & \text{falls} \quad \lambda = 0,8 \\ 0,7877, & \text{falls} \quad \lambda = 0,6 \end{cases} \qquad , \qquad \pi_0 \quad = \quad \begin{cases} 0,9061, & \text{falls} \quad \lambda = 0,8 \\ 0,8123, & \text{falls} \quad \lambda = 0,6 \end{cases}$$

Es ist zu beachten, dass zur Bestimmung der Bleibewahrscheinlichkeiten beim invarianten PRAM der geschätzte Anteil an Einheiten mit $Y = 1$ ($\hat{\theta}$, geschätzt durch die Anteile in der Stichprobe) verwendet wird und nicht der wahre Anteil aus der Grundgesamtheit (θ).

Um den Effekt der Post-Randomisierung auf die Ergebnisse der Parameterschätzung zu isolieren, werden zwei Simulationsdesigns getestet, die als „Y fix“ und „Y stochastisch“ bezeichnet werden: Im ersten Fall werden auch die Ausprägungen der Variablen Y lediglich einmal simuliert und eine Stichprobe mit festen Werten $(y_i\;,\;x_i)\;,\ldots\;,i=1,\ldots,\;n$ erzeugt. Anschließend werden in jedem Simulationslauf beide Varianten der Post-Randomisierung neu auf die feste Datenmatrix angewendet. In der zweiten Variante werden hingegen die Realisationen von Y bei festem Vektor $\mathbf{x}$ mit jedem Simulationslauf neu erzeugt.

Tabelle 27.1: Ergebnisse der Monte-Carlo-Simulationen für einfaches PRAM im Probit-Modell

Einfaches PRAM

a) Y stochastisch

n	π	$\bar{\hat{\alpha}}$	$s_{\hat{\alpha}}$	$\bar{\hat{\beta}}$	$s_{\hat{\beta}}$
3.000	0,8	-2,500	0,216	0,600	0,049
3.000	0,6	-2,640	0,845	0,639	0,198
10.000	0,8	-2,506	0,116	0,601	0,027
10.000	0,6	-2,533	0,607	0,607	0,089

b) Y fix

3.000	0,8	-2,501	0,187	0,602	0,044
3.000	0,6	-2,635	0,738	0,636	0,172
10.000	0,8	-2,526	0,103	0,603	0,024
10.000	0,6	-2,559	0,366	0,612	0,084

Simulationsergebnisse basieren auf 1.000 Replikationen.
Wahre Parameterwerte: $\alpha = -2,5$ und $\beta = 0,6$.

Die Simulationsergebnisse sind in den Tabellen 27.1 und 27.2 dargestellt. $\bar{\hat{\alpha}}$ bezeichnet den Schätzer für α und $s_{\hat{\alpha}}$ ist die Standardabweichung dieses Schätzers bei 1.000 Simulationsläufen. Die analoge Notation gilt für β.

Das erste in Tabelle 27.1 zu erkennende Ergebnis ist, dass die Ergebnisse für einen größeren Stichprobenumfang und eine höhere Bleibewahrscheinlichkeit eindeutig besser werden. Beinahe keine Verzerrung der Schätzergebnisse ist zu erkennen, wenn $\pi = 0,80$ und $n = 10.000$ gewählt wird. Die Standardabweichung steigt dramatisch, wenn der Stichprobenumfang auf $n = 3.000$ reduziert wird. Wenn zusätzlich die Bleibewahrscheinlichkeit π auf 0,60 reduziert wird, scheinen beide Schätzer (für α und β) „weg von Null“ verzerrt zu sein.

Die Ergebnisse für invariantes PRAM sind in Tabelle 27.2 dargestellt. Es ist zu beachten,

Tabelle 27.2: Ergebnisse der Monte-Carlo-Simulationen für invariantes PRAM im Probit-Modell

Invariantes PRAM

a) Y stochastisch

n	λ	$\bar{\hat{\alpha}}$	$s_{\hat{\alpha}}$	$\bar{\hat{\beta}}$	$s_{\hat{\beta}}$
3.000	0,8	-2,502	0,146	0,601	0,033
3.000	0,6	-2,517	0,215	0,604	0,049
10.000	0,8	-2,499	0,079	0,600	0,018
10.000	0,6	-2,511	0,109	0,603	0,029

b) Y fix

3.000	0,8	-2,505	0,109	0,603	0,025
3.000	0,6	-2,505	0,186	0,603	0,043
10.000	0,8	-2,533	0,060	0,605	0,014
10.000	0,6	-2,526	0,099	0,603	0,023

Simulationsergebnisse basieren auf 1.000 Replikationen.
Wahre Parameterwerte: $\alpha = -2,5$ und $\beta = 0,6$.

dass ein direkter Vergleich zwischen den Ergebnissen für invariantes und einfaches PRAM schwierig ist, weil die Bleibewahrscheinlichkeiten nicht direkt vergleichbar sind und in diesem Fall für das invariante PRAM generell geringer gewählt wurden. Allerdings sind sie asymmetrisch. Beide Parameterschätzer und die dazugehörigen Standardabweichungen deuten darauf hin, dass die Schätzer in diesem Fall unverzerrt sind.

27.1.3 Praxisbeispiele

Strotmann (2004) ergänzt die theoretischen Untersuchungen von Ronning (2005) und Ronning und Rosemann (2004), indem Post-Randomisierungsmethoden auf Betriebsdaten des IAB-Betriebspanels angewendet werden (vgl. auch Ronning et al. (2005)). Die im

Folgenden dargestellten Ergebnisse basieren auf diesen Untersuchungen.

Als Praxisbeispiel zur Untersuchung der Post-Randomisierung im binären Probit-Modell wird das in Abschnitt 21.2 beschriebene Modell zur Erklärung der Tarifbindung herangezogen.

Für den Fall einer dichotomen Variable kann die Methode der Post-Randomisierung mit Hilfe der (2 x 2)-Übergangsmatrix P beschrieben werden, wobei diese die Wechselwahrscheinlichkeiten zwischen den beiden Ausprägungen „0“ und „1“ sowie die entsprechenden Bleibewahrscheinlichkeiten enthält (vgl. hierzu Unterabschnitt 6.1.2). Y^a repräsentiert die anonymisierte Variable.

$$P = \begin{pmatrix} P(Y^a = 0 \,|Y = 0) & P(Y^a = 0 \,|Y = 1) \\ P(Y^a = 1 \,|Y = 0) & P(Y^a = 1 \,|Y = 1) \end{pmatrix} \tag{27.24}$$

Sind die Bleibewahrscheinlichkeiten der Ausprägungen für Y $p_{00} = P(Y^a = 0 \,|Y = 0)$ und $p_{11} = P(Y^a = 1 \,|Y = 1)$ symmetrisch, z.B. $p_{00} = p_{11} = \pi$, so wird die Methode „symmetrisches PRAM“ oder auch „einfaches PRAM“ genannt; sind die Bleibewahrscheinlichkeiten hingegen asymmetrisch, so wird sie „asymmetrisches PRAM“ genannt. Ein spezieller Fall des asymmetrischen PRAM ist das „invariante PRAM“, bei dem die Randverteilung der binären Variablen Y unverändert bleibt (vgl. hierzu Unterabschnitt 6.1.2).

Um die Auswirkungen des einfachen PRAM zu untersuchen, wurden Monte-Carlo-Simulationen mit 500 Replikationen durchgeführt. Dabei wurden die symmetrischen Tauschwahrscheinlichkeiten jeweils in Schritten von zwei Prozentpunkten systematisch von 2% auf 30% erhöht, um zu untersuchen, wie sich eine Erhöhung der Wechselwahrscheinlichkeit $(1-\pi)$ auf die Qualität der Schätzungen auswirkt. Die möglichen Auswirkungen der Größe der Stichprobe auf die inhaltlichen Ergebnisse werden dadurch berücksichtigt, dass die Schätzungen einerseits für Deutschland insgesamt (knapp 15.000 Beobachtungen), andererseits für Baden-Württemberg (rund 1.200 Beobachtungen) durchgeführt werden.

In Tabelle 27.3 sind die Ergebnisse für Deutschland dargestellt, Tabelle 27.4 gibt die entsprechenden Ergebnisse für Baden-Württemberg wieder. Neben den Originalschätzungen, die jeweils auf die unveränderten Daten angewendet werden, sind in beiden Tabellen die Ergebnisse naiver Probitschätzungen für anonymisierte Daten sowie die Resultate PRAM-korrigierter Schätzungen wiedergegeben (zur PRAM-Korrektur vgl. Unterabschnitt 27.1.1).

Tabelle 27.3: Vergleich von Originalschätzung, naiver Probitschätzung bei anonymisierten Daten und PRAM-korrigierter ML-Probitschätzung für variierende Tauschwahrscheinlichkeiten, MC-Simulationen mit 500 Replikationen, Deutschland

		Abhängige Variable: Tarifvertrag (1 = ja, 0 = nein)						
		Original-	Naive Schätzungen			PRAM korrigiert		
		daten	2%	14%	30%	2%	14%	30%
Konstante		-1,126	-1,061	-0,735	-0,387	-1,129	-1,137	-1,145
		(-33,76)	(-32,18)	(-23,40)	(-12,71)	(-31,82)	(-22,23)	(-11,76)
Log. Beschäftigtenzahl		0,318	0,299	0,207	0,109	0,318	0,318	0,318
		(46,16)	(44,29)	(33,14)	(18,32)	(43,08)	(29,35)	(15,31)
	Bau-	0,668	0,633	0,448	0,240	0,672	0,686	0,670
	gewerbe	(15,86)	(15,13)	(11,00)	(6,00)	(15,15)	(11,68)	(6,05)
	Handel und	0,306	0,288	0,200	0,105	0,308	0,318	0,326
	Reparatur	(8,29)	(7,85)	(5,58)	(2,98)	(7,97)	(5,98)	(3,31)
	Dienst-	0,081	0,075	0,049	0,026	0,083	0,091	0,098
	leistungen	(2,92)	(2,73)	(1,85)	(0,94)	(2,85)	(2,25)	(1,31)
	Öffentlicher	0,966	0,890	0,565	0,279	0,982	1,033	1,049
	Sektor	(20,49)	(19,49)	(13,91)	(7,35)	(18,71)	(12,39)	(6,37)
Beobachtungen		14.757	14.757	14.757	14.757	14.757	14.757	14.757

Bemerkung: t-Werte in Klammern; t-Werte sind definiert als Verhältnis des Mittelwerts der geschätzten Koeffizienten und des Mittelwerts der geschätzten Standardfehler. Das Verarbeitende Gewerbe ist Referenzkategorie. Für detaillierte Ergebnisse zur Originalschätzung vgl. Tabelle 21.6.

Tabelle 27.4: Vergleich von Originalschätzung, naiver Probitschätzung bei anonymisierten Daten und PRAM-korrigierter ML-Probitschätzung für variierende Tauschwahrscheinlichkeiten, MC-Simulationen mit 500 Replikationen, Baden-Württemberg

	Abhängige Variable: Tarifvertrag (1 = ja, 0 = nein)						
	Original-	Naive Schätzungen			PRAM korrigiert		
	daten	2%	14%	30%	2%	14%	30%
Konstante	-0,988	-0,928	-0,644	-0,338	-0,992	-0,978	-1,044
	(-7,46)	(-7,08)	(-5,12)	(-2,76)	(-7,05)	(-4,87)	(-2,62)
Log. Beschäftigtenzahl	0,301	0,283	0,197	0,104	0,301	0,297	0,310
	(12,24)	(11,72)	(8,74)	(4,84)	(11,48)	(7,77)	(4,03)
Bau-	0,748	0,708	0,492	0,248	0,748	0,749	0,799
gewerbe	(4,46)	(4,26)	(3,08)	(1,60)	(4,22)	(3,00)	(0,66)
Handel und	0,379	0,356	0,252	0,127	0,378	0,380	0,416
Reparatur	(2.88)	(2.73)	(1.98)	(1,02)	(2,75)	(1,99)	(1,14)
Dienst-	0,040	0,034	0,016	0,001	0,041	0,045	0,066
leistungen	(0,40)	(0,34)	(0,16)	(0,01)	(0,39)	(0,31)	(0,24)
Öffentlicher	0,796	0,736	0,478	0,751	0,823	0,875	1,226
Sektor	(4,60)	(4,37)	(3,15)	(1,77)	(4,22)	(2,75)	(0,14)
Beobachtungen	1.201	1.201	1.201	1.201	1.201	1.201	1.201

Bemerkung: t-Werte in Klammern; t-Werte sind definiert als Verhältnis des Mittelwerts der geschätzten Koeffizienten und des Mittelwerts der geschätzten Standardfehler. Das Verarbeitende Gewerbe ist Referenzkategorie. Für detaillierte Ergebnisse zur Originalschätzung vgl. Tabelle 21.6.

Abbildung 27.1 stellt zusätzlich die Ergebnisse für die PRAM-korrigierten Schätzer graphisch dar, wobei jeweils für variierende Weschelwahrscheinlichkeiten der relative Fehler ("Bias") der Koeffizientenschätzungen abgetragen ist.

Zusammenfassend lassen sich folgende Ergebnisse festhalten:

- Sowohl für die Deutschland- als auch für die Baden-Württemberg-Stichprobe des IAB-Betriebspanels zeigt sich, dass der Bias der PRAM-korrigierten Schätzung in der Tendenz mit zunehmender Tauschwahrscheinlichkeit steigt.

- Die Koeffizienten der Konstanten, der logarithmierten Beschäftigung und der Dummy-Variablen für das Baugewerbe und den Handel werden sowohl für Deutschland als auch für Baden-Württemberg selbst bei größeren Tauschwahrscheinlichkeiten recht gut geschätzt. So liegt die Verzerrung bei dem Koeffizienten der Betriebsgröße für Deutschland für sämtliche Tauschwahrscheinlichkeiten bis 30% deutlich unter 0,5%. Auch für die 1.201 Beobachtungen der Baden-Württemberg Stichprobe übersteigt der relative Fehler erst für sehr große Werte von $(1-\pi)$ den 1%-Fehlerbereich.

- Die Verzerrungen der Koeffizienten bei den Dummy-Variablen für den Dienstleistungssektor und den öffentlichen Sektor sind dagegen bereits für kleine Werte der Tauschwahrscheinlichkeit recht erheblich. So beträgt für den Deutschland-Datensatz der durchschnittliche relative Fehler für die Dienstleistungsdummy und eine Tauschwahrscheinlichkeit von 2% bereits 2,7%, für eine Tauschwahrscheinlichkeit von 8% bereits knapp 9%.

- Hinsichtlich der relativen Fehler für Deutschland und Baden-Württemberg im Vergleich lässt sich keine eindeutige Aussage ableiten. So ist die relative Verzerrung für die Koeffizienten der logarithmierten Beschäftigung, der Dienstleistungs-Dummy und der Dummy für den öffentlichen Sektor für den kleineren Beobachtungsumfang jeweils tendenziell größer, während für den Bereich Handel die Verzerrung für die Deutschland-Stichprobe größer ist.

- Abbildung 27.1 deutet an, dass die Verzerrung bei PRAM-korrigierter Schätzung systematischer Natur sein könnte. Während die Konstante unterschätzt wird, gibt es eine positive Verzerrung bei den Koeffizienten sämtlicher erklärender Variablen. Offenbar wird also der absolute Koeffizientenwert im Korrekturmodell bei hohen Wechselwahrscheinlichkeiten überschätzt. Dies spiegelt sich auch darin wieder, dass die Verteilung der in den 500 Replikationen geschätzten Koeffizienten mit wachsender Tauschwahrscheinlichkeit schief wird. Die Ergebnisse der Simulationen stimmen insoweit mit den Ergebnissen der theoretischen Herleitungen in Unterabschnitt 27.1.2 überein.

- Für Werte der Tauschwahrscheinlichkeit, die größer sind als 30%, konnte für eine erhebliche Anzahl der PRAM-korrigierten ML-Schätzungen keine Konvergenz erreicht werden. Für die Baden-Württemberg-Stichprobe konnten z.B. für $(1-\pi) = 36\%$

in 500 Replikationen nur 175 Schätzungen, für $(1 - \pi) = 38\%$ nur 90 Schätzungen erfolgreich realisiert werden. Die Schätzungen werden dabei regelmäßig extrem verzerrt. Dies erklärt auch, warum der Bias für Tauschwahrscheinlichkeiten ab 30% teilweise sogar explodieren kann. Für Baden-Württemberg und die Dummy für den öffentlichen Sektor beträgt der relative Fehler bei einer Tauschwahrscheinlichkeit von 34% zum Beispiel fast 72%.

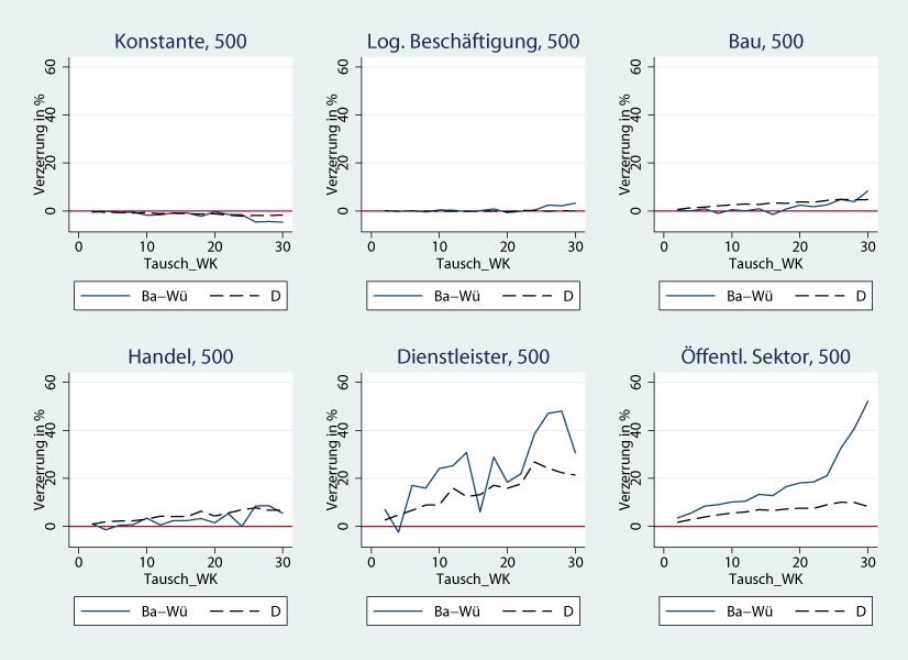

Abbildung 27.1: Relative Fehler von PRAM-korrigierten ML-Schätzungen in Abhängigkeit von variierenden Wechselwahrscheinlichkeiten. Schätzungen für Deutschland und Baden-Württemberg im Vergleich

Die theoretische Analyse der Eigenschaften des PRAM-Schätzers (siehe Ronning (2005) und Ronning und Rosemann (2004) sowie Unterabschnitt 27.1.1) zeigt, dass ein Effizienzverlust bei der Verwendung von PRAM besteht, der – für eine Wechselwahrscheinlichkeit von weniger als 0,5 – positiv von der Wechselwahrscheinlichkeit abhängt. In empirischen Analysen, bei denen anonymisierte Daten verwendet werden, kann dieser Effizienzverlust zum Ziehen falscher Rückschlüsse hinsichtlich der statistischen Signifikanz der Einflüsse führen. Abbildung 27.2 zeigt die Boxplots für die geschätzten Koeffizienten der logarithmierten Beschäftigung für Wechselwahrscheinlichkeiten von 2%, 14% und 30%, wenn 500 Replikationen simuliert werden. Wie man leicht sehen kann, erhöht sich die Varianz des einfachen PRAM-Schätzers mit steigenden Wechselwahrscheinlichkeiten beträchtlich. Sowohl

das Niveau der geschätzten Standardfehler als auch die Streuung des Schätzers steigen mit zunehmenden Wechselwahrscheinlichkeiten stark an. Dies deutet auf die ernste Gefahr fehlerhafter Schlussfolgerungen in Bezug auf die statistische Signifikanz in empirischen Analysen mit anonymisierten Daten hin. Es ist allerdings bemerkenswert, dass die Verzerrung für hohe Wechselwahrscheinlichkeiten geringer auszufallen scheint als in der Simulation aus Ronning und Rosemann (2004) (vgl. auch Unterschnitt 27.1.2). Dies könnte daran liegen, dass das empirische Modell zur Erklärung der Tarifbindung mit mehr als einer erklärenden Variable geschätzt wurde.

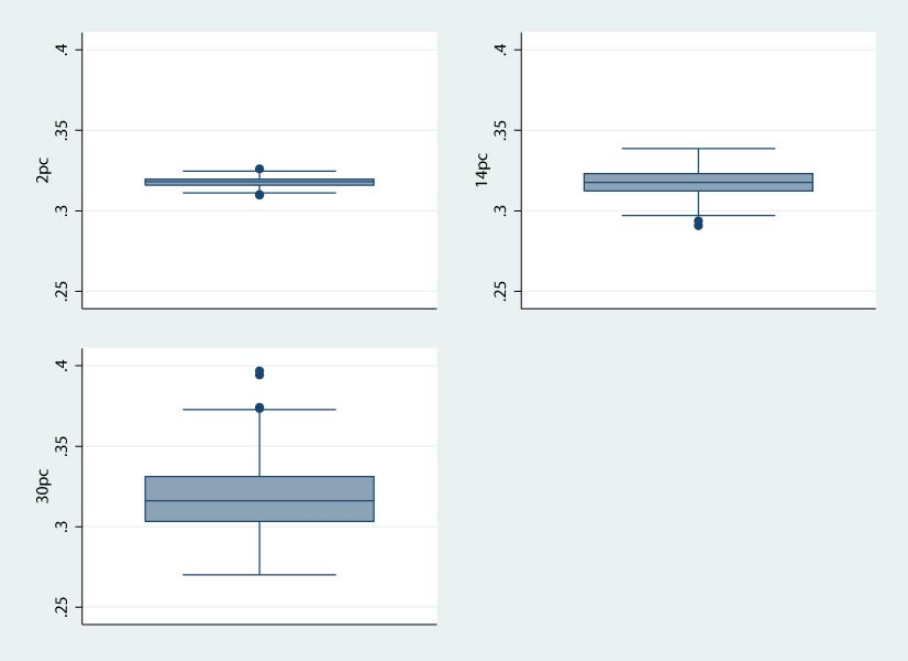

Abbildung 27.2: Boxplots der Verteilungen der geschätzten Koeffizienten für die logarithmierte Beschäftigung – einfaches PRAM mit Wechselswahrscheinlichkeiten von 2%, 14% und 30%, Deutschland, 500 Replikationen

Bislang wurden die Mittelwerte der geschätzten Koeffizienten und deren Standardfehler für die Berechnung der t-Statistiken verwendet. Eine durchschnittliche Signifikanz in 500 Replikationen muss jedoch keineswegs bedeuten, dass man sich in jedem konkreten Anwendungsfall auf die Schlussfolgerungen verlassen kann. Um einen besseren Eindruck über das Risiko der irrtümlichen Annahme einer falschen Nullhypothese (H_0: Koeffizient ist Null) zu bekommen, wird im Folgenden dieser „Fehler zweiter Art" für alle im Originalmodell signifikanten Variablen berechnet. Abbildung 27.3 zeigt folglich den Anteil der „richti-

gen“ Entscheidungen in Bezug auf die Signifikanz der Koeffizienten (basierend auf dem 5%-Signifikanzniveau) für variierende Werte der Wechselwahrscheinlichkeiten. Dabei sind gleichzeitig die Ergebnisse für Baden-Württemberg und Deutschland insgesamt sowie jeweils die beiden 50%-Stichproben dargestellt, um die Auswirkungen der Fallzahl auf die Ergebnisse zu illustrieren.

Auffällig ist, dass im vorliegenden Fall für Deutschland und den vollständigen Datensatz bei einem Signifikanzniveau von 5% selbst bei Tauschwahrscheinlichkeiten bis immerhin fast 30% in keiner Replikation eine fehlerhafte Schlussfolgerung hinsichtlich der Signifikanz der Variablen resultierte. Eine Ausnahme stellt die Dummyvariable für den Dienstleistungssektor dar, deren Koeffizient für eine Tauschwahrscheinlichkeit von 10% nur noch in rund 80% der Fälle richtigerweise als signifikant von Null verschieden erkannt wird.

Für den kleineren Baden-Württemberg-Datensatz fallen die Ergebnisse aufgrund der bereits in den Originalmodellen geringeren t-Werte weniger günstig aus. Für Tauschwahrscheinlichkeiten von bis zu 20% werden die Signifikanzen der Koeffizienten der logarithmierten Beschäftigung sowie der Dummy-Variablen für das Baugewerbe und den Handel zwar praktisch ausnahmslos erkannt. Für die Dummy-Variable Handel und für eine Wechselwahrscheinlichkeit von 10% besteht bereits ein Risiko von etwa 35%, dass die Nullhypothese fälschlicherweise angenommen wird. Für eine Wechselwahrscheinlichkeit von 20% erhöht sich das Risiko auf 60% und für eine Wechselwahrscheinlichkeit von knapp 40% tritt der „Fehler zweiter Art“ mit knapp 100% fast sicher ein.

Die Simulationsergebnisse zeigen, dass das Ziehen von Rückschlüssen hinsichtlich der statistischen Signifikanz der Koeffizienten riskant ist, wenn – wie in der Realität der Fall – nur ein Datensatz vorhanden ist. Dies gilt insbesondere, sofern die Wechselwahrscheinlichkeiten groß und die Anzahl der Beobachtungen eher gering sind. Fasst man die Ergebnisse zusammen, so sollte für diese Auswertung mit dem IAB-Betriebspanel für Baden-Württemberg maximal eine Wechselwahrscheinlichkeit von 5% gewählt werden, für Deutschland könnte die Wechselwahrscheinlichkeit auch etwas höher angesetzt werden.

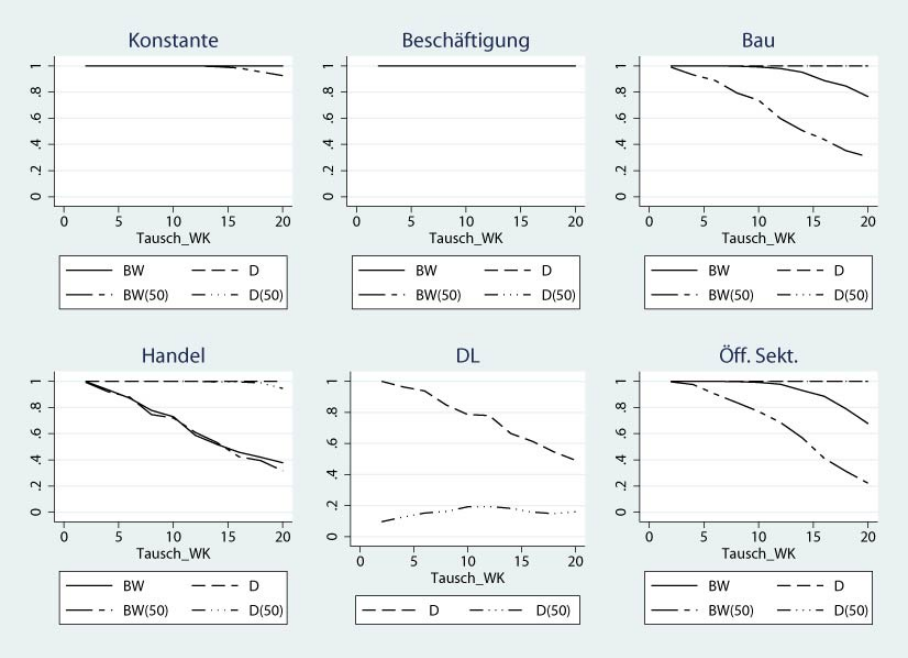

Abbildung 27.3: Anteile der „richtigen Entscheidungen“ hinsichtlich der statistischen Signifikanz der Koeffizienten (5%-Niveau) für variierende Werte der Wechselwahrscheinlichkeiten, einfaches PRAM, Deutschland und Baden-Württemberg im Vergleich, jeweils vollständige Stichprobe und 50%-Sample, 500 Replikationen

27.2 Post-Randomisierung der abhängigen diskreten Variablen und Bearbeitung der erklärenden metrischen mit datenverändernden Verfahren im Probit-Modell

Nachdem im vorangegangenen Abschnitt 27.1 ausschließlich die abhängige Variable im binären Probit-Modell mit Post-Randomisierung anonymisiert wurde, soll nun in einem zweiten Schritt untersucht werden, wie die Korrektur der Post-Randomisierung funktioniert, wenn gleichzeitig die erklärenden metrischen Variablen mit stochastischen Überlagerungen oder Mikroaggregation behandelt werden.

In Unterabschnitt 27.2.1 werden hierzu zunächst ein paar theoretische Überlegungen angestellt. Unterabschnitt 27.2.2 beinhaltet die Ergebnisse von Monte-Carlo-Simulationsexperimenten mit simulierten Daten. Die Ergebnisse praktischer Beispielrechnungen werden in Unterabschnitt 27.2.3 dargestellt.

27.2.1 Theoretische Eigenschaften

Grundlage der folgenden Überlegungen sind die in den vorangegangenen Teilen und Kapiteln abgeleiteten Ergebnisse über die isolierte Wirkung der einzelnen Verfahren im Probit-Modell. Dabei wurde hergeleitet, dass sich die Verzerrung des Schätzers bei einer Post-Randomisierung der abhängigen Variablen im binären Probit-Modell durch die Berücksichtigung der Post-Randomisierung bei der Maximum-Likelihood-Schätzung korrigieren lässt, sofern die Tauschwahrscheinlichkeit nicht zu nahe an 50 Prozent und die Anzahl der Beobachtungen groß genug ist (so genannte PRAM-Korrektur). Für stochastische Überlagerungen wurde abgeleitet, dass in nichtlinearen Modellen in jedem Fall eine Verzerrung auftritt, die sich jedoch durch Korrekturschätzer für nichtlineare Modelle, wie zum Beispiel den SIMEX-Schätzer, in der Regel approximativ korrigieren lässt. Daraus kann gefolgert werden, dass auch eine Verzerrung, die durch die Verbindung von Post-Randomisierung der abhängigen binären Variablen einerseits und stochastischer Überlagerung der erklärenden metrischen Variablen andererseits hervorgerufen, korrigiert werden kann, indem die SIMEX-Korrektur anstatt auf die „naive“ Probit-Schätzung auf die PRAM-korrigierte Probit-Schätzung angewendet wird.

Mikroaggregationsverfahren verursachen hingegen in nichtlinearen Modellen eine Verzerrung der Schätzer, die nicht korrigiert werden kann. Allerdings hat die abstandsorientierte getrennte Mikroaggregation von erklärenden Variablen auch bei Probit-Schätzungen nur zu geringen Verzerrungen der Schätzer und der Teststatistiken geführt. Da unabhängig von der Mikroaggregation eine PRAM-korrigierte Probit-Schätzung möglich ist, kann gefolgert werden, dass dann gute Ergebnisse erzielbar sind, wenn sowohl die PRAM-Korrektur erfolgreich ist als auch die Mikroaggregation nur zu geringen Verzerrungen führt. Dies bedeutet, dass bei einer abstandsorientierte getrennte Mikroaggregation der erklärenden metrischen Variablen in Verbindung mit Post-Randomisierung der abhängigen binären Variablen mit einer

nicht zu hoch gewählten Wechselwahrscheinlichkeit bei ausreichendem Beobachtungsumfang die Verwendung des PRAM-korrigierenden Probit-Schätzers zu akzeptablen Schätzergebnissen führen dürfte. Wird demgegenüber eine zu hohe Wechselwahrscheinlichkeit gewählt, die zu einer Verzerrung des PRAM-korrigierten Schätzers führen kann, oder wird eine Variante der stochastischen Mikroaggregation beziehungsweise die abstandsorientierte gemeinsame oder die abstandsorientierte gruppierte Mikroaggregation gewählt, so dürften die Schätzergebnisse in der Regel unzufriedenstellend hohe Verzerrungen aufweisen.

27.2.2 Monte-Carlo-Simulationen

a) Post-Randomisierung der abhängigen Variablen und stochastische Überlagerung der Regressoren

Im vorliegenden Fall werden wir den „korrigierten" Maximum-Likelihood-Schätzer für das Probitmodell, in dem die abhängige Variable randomisiert ist, mit dem SIMEX-Schätzer zur Berücksichtigung der stochastischen Überlagerung der rechtsstehenden Variablen kombinieren. Wir beschränken uns dabei auf den Fall einer additiven stochastischen Überlagerung.[43]

Betrachtet wird das Simulations-Design, das auch bereits in Unterabschnitt 27.1.2 betrachtet wurde, in dem lediglich die abhängige Variable mit Post-Randomisierung anonymisiert wurde. Die Tabellen 27.5 und 27.6 geben Simulationsergebnisse für den Fall an, dass das Modell 500 mal simuliert und geschätzt wird. Man beachte, dass hier das Design „Y stochastisch" verwendet wird (siehe dazu Unterabschnitt 27.1.2).

Die abhängige beobachtete Variable wurde mittels PRAM-Parameter π anonymisiert, d.h. es wird Y^a statt Y beobachtet, und die unabhängige Variable wurde additiv mit einem Fehler V überlagert.

Im Fall der additiven stochastischen Überlagerung wird X^a statt X beobachtet, wobei

$$X^a = X + V \text{ mit } E(V) = 0, var(V) = \sigma_v^2$$

gilt.

Für die im Simulationsexperiment getesteten Anonymisierungsparameter gilt:

$$\begin{array}{lll} V & \sim N(0, \sigma_v^2) & \sigma_v^2 = 0,01\ 0,25 \\ \pi & \text{PRAM-Parameter} & \pi \in [0,80\,;\ 1,000] \end{array}$$

43) Wir danken Sandra Lechner, Universität Konstanz, für die Überlassung eigener GAUSS-Programme für den SIMEX-Schätzer, der eine **quadratische** Extrapolationsfunktion verwendet.

Tabelle 27.5: MC-Simulationen – Post-Randomisierung der abhängigen Variablen und additive stochastische Überlagerung im Probit-Modell ($\sigma_v^2 = 0,01, n = 500$, 500 Replikationen), SIMEX-Schätzer und PRAM-Korrektur

π	Schätzwert	Stand.Abw.	Varianz	Minimum	Median	Maximum
1,000	-2,522	0,242	0,059	-3,245	-2,509	-1,934
	0,604	0,054	0,003	0,468	0,600	0,777
0,950	-2,548	0,257	0,066	-3,355	-2,522	-1,862
	0,610	0,059	0,003	0,465	0,610	0,781
0,900	-2,542	0,256	0,066	-3,468	-2,527	-1,757
	0,610	0,058	0,003	0,445	0,606	0,799
0,850	-2,557	0,273	0,075	-3,710	-2,531	-1,984
	0,612	0,062	0,004	0,488	0,606	0,863
0,800	-2,611	0,301	0,091	-3,646	-2,582	-1,947
	0,625	0,068	0,005	0,453	0,619	0,848

Hinweise
Die erste Zeile gibt die Resultate bezüglich α und die zweite bezüglich β an.

Tabelle 27.6: MC-Simulationen – Post-Randomisierung der abhängigen Variablen und additive stochastische Überlagerung im Probit-Modell ($\sigma_v^2 = 0,25, n = 500$, 500 Replikationen), SIMEX-Schätzer und PRAM-Korrektur

π	Schätzwert	Stand.Abw.	Varianz	Minimum	Median	Maximum
1,000	-2,513	0,266	0,071	-3,385	-2,493	-1,723
	0,603	0,060	0,004	0,441	0,600	0,827
0,950	-2,525	0,282	0,080	-3,881	-2,494	-1,842
	0,605	0,063	0,004	0,469	0,600	0,922
0,900	-2,498	0,278	0,077	-3,492	-2,467	-1,800
	0,600	0,064	0,004	0,453	0,595	0,863
0,850	-2,537	0,285	0,081	-3,537	-2,510	-1,702
	0,608	0,066	0,004	0,419	0,603	0,867
0,800	-2,584	0,316	0,100	-3,805	-2,558	-1,891
	0,622	0,073	0,005	0,464	0,616	0,923

Hinweise
Die erste Zeile gibt die Resultate bezüglich α und die zweite bezüglich β an.

Ein Vergleich mit den Ergebnissen aus Unterabschnitt 27.1.2 ist nur bedingt möglich, weil dort die Stichprobenumfänge größer gewählt wurden. Auch lassen sich nur die Ergebnisse für $\pi = 0,80$ vergleichen.

Betrachtet man zunächst den Fall, bei dem ausschließlich stochastisch überlagert wurde ($\pi = 1$), so ergibt sich ein Bias, der allerdings nicht mit der Varianz der Überlagerung steigt. Angesichts der ausgewiesenen Standardabweichungen der Schätzer ist ohnehin diese Abweichung als recht gering anzusehen. Für $\pi = 0,80$ ergibt sich in diesem Fall (mit $n = 500$ deutlich geringerer Beobachtungsumfang als in Unterabschnitt 27.1.2) ein deutlich stärkerer Bias, der allerdings wieder weitgehend unbeeinflusst von der Höhe der Varianz der stochastischen Überlagerung ist. Insgesamt deuten die Simulationsergebnisse darauf hin, dass eine Korrektur der durch simultane Anwendung von Post-Randomisierung und additiver stochastischer Überlagerung hervorgerufenen Verzerrung möglich ist, indem der SIMEX-Schätzer auf die PRAM-korrigierte Probit-Schätzung angewendet wird.

b) Post-Randomisierung der abhängigen Variablen und Mikroaggregation der Regressoren

Zur Überprüfung der in Unterabschnitt 27.2.1 aufgestellten Hypothesen zum Zusammenspiel von Post-Randomisierung der abhängigen Variablen und Mikroaggregation der Regressvariablen wird dasselbe latente Modell herangezogen wie bereits bei der Untersuchung der Verbindung von Post-Randomisierung und stochastischer Überlagerung.

Die abhängige Variable wird mit einfachem PRAM mit den Bleibewahrscheinlichkeiten $\pi = 0,80$ und $\pi = 0,60$ anonymisiert. Die Regressorvariable wird alternativ abstandsorientiert und zufällig mikroaggregiert. Die Anzahl der Beobachtungen beträgt 3.000. Die Gruppengröße bei der Mikroaggregation wird auf drei gesetzt. Die Ergebnisse basieren auf Monte-Carlo-Simulationen mit 1.000 Replikationen. Dabei werden die Beobachtungen in jedem Simulationslauf und für jede Art der Anonymisierung neu erzeugt. Die Durchschnittsergebnisse der Monte-Carlo-Simulationen sind in Tabelle 27.7 dargestellt.

Man erkennt, dass sich die Hypothesen aus Unterabschnitt 27.2.1 bestätigen. Gute Ergebnisse lassen sich erzielen, sofern die abstandsorientierte (getrennte) Mikroaggregation angewendet wird und die Tauschwahrscheinlichkeit bei der Post-Randomisierung mit 0,2 so gering gewählt wird, dass die PRAM-Korrektur gelingt.

Unabhängig vom Gelingen der PRAM-Korrektur ergeben sich hingegen bei der zufälligen Gruppenbildung starke Abweichungen der Schätzergebnisse vom Original. Ebenso treten auch bei der abstandsorientierten Mikroaggregation Verzerrungen auf, wenn gleichzeitig die Tauschwahrscheinlichkeit bei der Post-Randomisierung mit 0,4 für den gewählten Beobachtungsumfang von 3.000 Einheiten zu hoch gewählt wird.

Tabelle 27.7: MC-Simulationen – Post-Randomisierung der abhängigen Variablen und Mikroaggregation der Regressorvariablen im Probit-Modell, 1.000 Replikationen

	Original	PRAM 0,20 / Mikro			PRAM 0,20 / Mikro_stoch			PRAM 0,40 / Mikro			PRAM 0,40 / Mikro_stoch		
			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem			Übereinstimmung zum Signifikanzniveau von 5% mit dem	
			Originalwert	theor. Wert		Originalwert	theor. Wert		Originalwert	theor. Wert		Originalwert	theor. Wert
X	0,601	0,603	*	*	0,456			0,657			0,471		
Konst.	-2,507	-2,51	*	*	-1,900			-2,74			-1,96		

* Übereinstimmung zum Signifikanzniveau von 5%

27.2.3 Praxisbeispiele

Als Praxisbeispiel wird wiederum das Probit-Modell zur Erklärung der Tarifbindung herangezogen. Dabei liegen bisher ausschließlich Ergebnisse für eine Verbindung der Post-Randomisierung der erklärenden abhängigen Variablen mit der Mikroaggregation der Regressoren vor. Dargestellt werden Schätzergebnisse mit dem IAB-Betriebspanel 2002 für Baden-Württemberg (Strotmann 2004; Ronning et al. 2005).

Im Weiteren wird für das IAB-Betriebspanel Baden-Württemberg mit 1.201 Beobachtungen zunächst ausschließlich die Variable *Logarithmierte Beschäftigung* mikroaggregiert. Es muss daher nicht zwischen gemeinsamer und getrennter Mikroaggregation unterschieden werden. Es wird allerdings zwischen der abstandsorientierten und der zufälligen Mikroaggregation differenziert. Im Einzelnen werden die folgenden Varianten der Mikroaggregation auf die logarithmierte Beschäftigung angewendet:

- abstandsorientierte Mikroaggregation mit einer Gruppengröße von drei
- abstandsorientierte Mikroaggregation mit einer Gruppengröße von fünf
- abstandsorientierte Mikroaggregation mit einer Gruppengröße von 15
- zufällige Mikroaggregation mit einer Gruppengröße von fünf

Die Ergebnisse einer Kombination der Mikroaggregationsvarianten mit der Post-Randomisierung der abhängigen Variablen (Tarifbindung) werden wiederum mit Hilfe von 500 Replikationen mit den Ergebnissen der ausschließlichen Post-Randomisierung der abhängigen Variablen verglichen (Tabelle 27.8).

Die wesentlichen Ergebnisse dieses Vergleichs, die auch in den Abbildungen 27.4 und 27.5 dargestellt werden, sind:

- Die zusätzliche *abstandsorientierte* Mikroaggregation der erklärenden Variablen *L*ogarithmierte Beschäftigung führt unabhängig von der gewählten Gruppengröße im Vergleich zur ausschließlichen Bearbeitung der abhängigen Variablen mit der Post-Randomisierung zu keiner nennenswerten Verzerrung der Koeffizientenwerte.
- Das gleiche Ergebnis erhält man auch für die t-Werte (Teststatistiken) mit Ausnahme des t-Wertes für die Dummy-Variable Öffentliche Verwaltung.
- Demgegenüber führt die *zufällige* Mikroaggregation der Variablen *L*ogarithmierte Beschäftigung zu erheblichen systematischen Verzerrungen der Koeffizientenwerte und der Teststatistiken (vgl. Tabelle 27.8).

Die Verzerrungen der Schätzergebnisse durch Mikroaggregationsverfahren in nichtlinearen Modellen – also auch im Probit-Modell – sind nicht korrigierbar. Insofern erweist sich eine

Tabelle 27.8: Vergleich des einfachen PRAM und des einfachen PRAM mit zusätzlicher Mikroaggregation der logarithmierten Beschäftigung, PRAM-korrigierte ML-Probit Schätzung für eine Wechselwahrscheinlichkeit von 20% - Ergebnisse von MC-Simulationen mit 500 Replikationen

	Einfaches PRAM der abh. Variable Wechselwahrscheinlichkeit 20%					
	Abhängige Variable: Tarifvertrag (1 = ja, 0 = nein)					
	Original-daten	Naive Schätzungen	Abstandsorientierte Mikroaggregation			Zufällige Mikroaggregation
			Gruppen-größe 3	Gruppen-größe 5	Gruppen-größe 15	Gruppen-größe 5
Konstante	-0,988	-1,009	-1,009	-1,023	-1,008	-0,556
	(-7,46)	(-4,04)	(-4,05)	(-4,08)	(-4,04)	(-1,62)
Log. Beschäftigtenzahl	0,301	0,303	0,302	0,305	0,303	0,244
	(12,24)	(6,32)	(6,34)	(6,36)	(6,35)	(2,96)
Baugewerbe	0,748	0,770	0,774	0,769	0,765	0,282
	(4,46)	(2,49)	(2,49)	(2,49)	(2,47)	(0,97)
Handel und Reparatur	0,379	0,395	0,390	0,407	0,392	-0,079
	(2.88)	(1,69)	(1,69)	(1,74)	(1,68)	(-0,38)
Dienstleistungen	0,040	0,051	0,052	0,051	0,050	-0,311
	(0,40)	(0,28)	(0,29)	(0,29)	(0,28)	(-2,01)
Öffentlicher Sektor	0,796	0,924	0,965	0,954	0,947	0,809
	(4,60)	(1,48)	(0,64)	(2,10)	(0,34)	(0,29)
Beobachtungen	1.201	1.201	1.201	1.201	1.201	1.201

Bemerkung: t-Werte in Klammern; t-Werte sind definiert als Verhältnis des Mittelwerts der geschätzten Koeffizienten und des Mittelwerts der geschätzten Standardfehler. Das Verarbeitende Gewerbe ist Referenzkategorie. Für detaillierte Ergebnisse zur Originalschätzung vgl. Tabelle 21.6.

zufällige Mikroaggregation hier als ungeeignet, während die abstandsorientierte Mikroaggregation bei der Entwicklung einer Anonymisierungsstrategie weiterverfolgt werden kann.

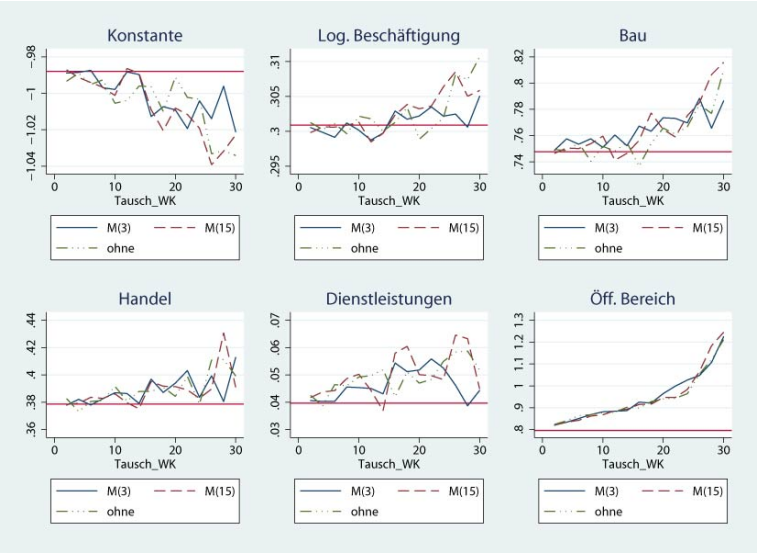

Abbildung 27.4: Vergleich von einfachem PRAM und einfachem PRAM mit Mikroaggregation, 500 Replikationen

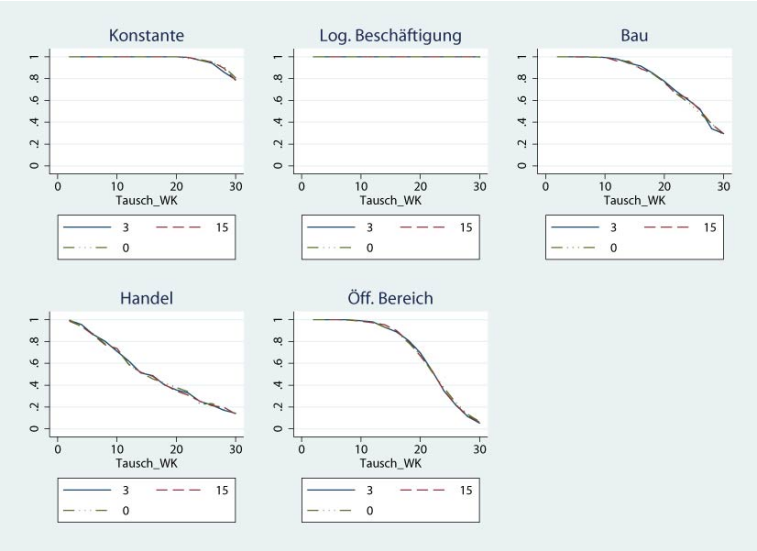

Abbildung 27.5: Vergleich der Anteile der „richtigen Entscheidungen“ hinsichtlich der statistischen Signifikanz der Koeffizienten auf dem 5%-Niveau, einfache PRAM ohne und mit Mikroaggregation (von 3 oder 15 Beobachtungen)

Kapitel 28

Post-Randomisierung der erklärenden diskreten Variablen in der Varianzanalyse

28.1 Einleitung

In Abschnitt 27.1 werden die formalen Eigenschaften einer mittels PRAM „randomisierten“ bzw. anonymisierten Zufallsvariablen im Einzelnen dargestellt. Wir verwenden dort die Bezeichnung Y^a, weil diese Variable dort als abhängige Variable im Probitmodell Verwendung findet. Im Folgenden verwenden wir stattdessen die Bezeichnung X^a für die randomisierte Regressorvariable, für die insbesondere

$$E[X_i^a] \quad = \quad (2\,\pi \,-\, 1)\,\theta \,+\, (1 \,-\, \pi) \tag{28.1}$$

sowie

$$var[X_i^a] \quad = \quad \pi\,(1-\pi) \,+\, (2\pi \,-\, 1)^2\,\theta\,(1 \,-\, \theta) \tag{28.2}$$

gilt. Dabei gibt θ die 'Erfolgswahrscheinlichkeit' für die binäre Variable X an und π ist die Wahrscheinlichkeit, den Originalwert zu erhalten, d.h.

$$\pi \;=\; P(X^a = 1|X = 1) \;=\; P(X^a = 0|X = 0)$$

Siehe Abbildung 28.1. Wir nennen dies die „symmetrische“ Randomisierung und werden uns in diesem Abschnitt auch auf diese Variante beschränken.[44]

---

44) Alternativ kann man „invariantes“ PRAM anwenden. Siehe dazu Unterabschnitt 6.1.2.

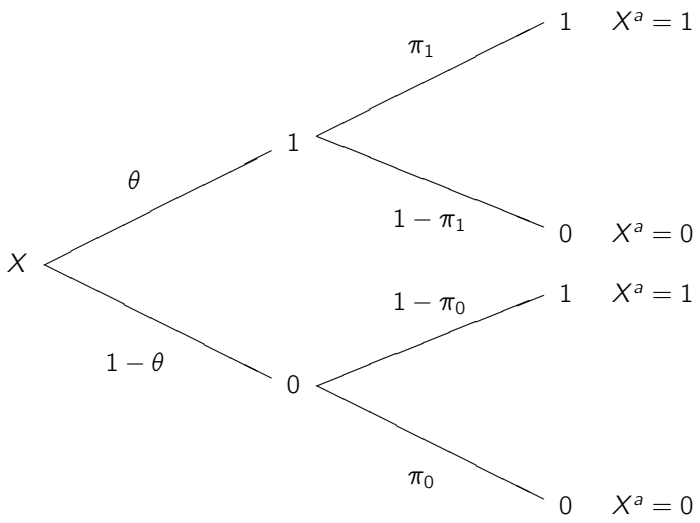

Abbildung 28.1: Randomisierte binäre Regressorvariable X^a

28.2 Einfache Varianzanalyse mit festen und stochastischen diskreten Effekten

28.2.1 Feste Effekte

Das Modell schreiben wir wie folgt:

$$Y_i \quad = \quad \mu \; + \; \beta_1 \, x_{i1} \; + \; \beta_2 \, x_{i2} \; + \; U_i \quad , \quad i = 1, 2, \ldots, n \, . \tag{28.3}$$

Dabei sind x_{i1} und x_{i2} zwei binäre Einflussvariablen mit

$$x_{ij} \quad = \begin{cases} 1 & \text{falls Beobachtung } i \text{ aus Kategorie } j \text{ bzw. } i \, \epsilon \, G_j \\ 0 & \text{sonst} \end{cases}$$

Bekanntlich lassen sich nur zwei der drei Parameter μ, β_1 und β_2 schätzen, wobei den drei Parametern eine beliebige lineare Restriktion auferlegt wird. Üblich ist

$$\beta_2 \quad = \quad 0 \text{ (Restkategorienormierung).}$$

Im vorliegenden Fall ist es einfacher, die Normierung

$$\mu = 0$$

zu verwenden.

Im Folgenden sollen n_1 der n Beobachtungen aus der Kategorie 1 stammen und die restlichen $n_2 = n - n_1$ Beobachtungen aus der Kategorie 2. Wir erhalten dann in Matrizenschreibweise

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \tag{28.4}$$

mit[45]

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix}$$

sowie

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} n_1 & 0 \\ 0 & n_2 \end{pmatrix} \quad , \quad \mathbf{X}'\mathbf{y} = \begin{pmatrix} n_1 \bar{y}_1 \\ n_2 \bar{y}_2 \end{pmatrix}$$

und damit für die Kleinstquadrateschätzung

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X}')^{-1}\mathbf{X}'\mathbf{y} = \begin{pmatrix} \bar{y}_1 \\ \bar{y}_2 \end{pmatrix} \tag{28.5}$$

Demnach werden die Koeffizienten durch die jeweiligen Gruppenmittel

$$\bar{y}_j = \frac{1}{n_j} \sum_{i \epsilon G_j} y_i$$

geschätzt. Die Schätzung ist erwartungstreu und konsistent.

45) Die Matrix X enthält in der ersten Spalte n_1 Einsen.

28.2.2 Stochastische (diskrete) Effekte

Für stochastische (**diskrete**) Effekte[46] erhalten wir folgendes Modell der einfachen Varianzanalyse:

$$Y_i \quad = \quad \mu \, + \, \beta_1 \, X_i \, + \, \beta_2 \, (1 - X_i) \, + \, U_i \quad , \quad i = 1, 2, \ldots, n \, . \tag{28.6}$$

Dabei ist X_i eine dichotome Zufallsvariable mit $P(X_i = 1) \, = \, \theta$, die unabhängig von der Störvariablen U_i verteilt ist. Wieder haben wir die Normierung

$$\mu \quad = \quad 0$$

gewählt. Für den bedingten Erwartungswert von Y_i gegeben X_i ergibt sich dann

$$E[Y_i | X_i = 1] \, = \, \beta_1 \quad \text{und} \quad E[Y_i | X_i = 0] \, = \, \beta_2 \tag{28.7}$$

und damit (nach dem Satz des iterierten Erwartungswertes)

$$E[Y_i] \quad = \quad \theta \, \beta_1 \, + \, (1 - \theta) \, \beta_2 \quad . \tag{28.8}$$

Die (stochastische) Regressormatrix lautet nun wie folgt:

$$\begin{pmatrix} X_1 & 1 - X_1 \\ X_2 & 1 - X_2 \\ \vdots & \vdots \\ X_n & 1 - X_n \end{pmatrix}$$

Damit erhalten wir

$$\mathbf{X}' \mathbf{X} \quad = \quad \begin{pmatrix} \sum\limits_i X_i^2 & \sum\limits_i X_i \, (1 - X_i) \\ \sum\limits_i X_i \, (1 - X_i) & \sum\limits_i (1 \, - \, X_i)^2 \end{pmatrix}$$

Da für die dichotomen Zufallsvariablen

$$E[X_i] \, = \, E[X_i^2] \, = \, \theta$$

gilt, erhalten wir

$$E \left[\frac{1}{n} \mathbf{X}' \mathbf{X} \right] \quad = \quad \begin{pmatrix} \theta & 0 \\ 0 & 1 - \theta \end{pmatrix}$$

46) Dies ist vom Fall stochastischer Effekte im Sinne der Varianzkomponenten-Modelle zu unterscheiden, bei denen die stochastischen Effekte als **stetige** Variable modelliert werden. Dieses Modell hat Relevanz für die Evaluationsproblematik: Falls die „Treatments“ zufällig zugeordnet werden, sprich Arbeitslose zufällig für eine Weiterbildungsmassnahme ausgewählt werden („random assignment“), haben wir es mit dem hier betrachteten Modell zu tun.

Ferner erhalten wir

$$\mathbf{X}'\mathbf{Y} \quad = \quad \begin{pmatrix} \sum_i X_i Y_i \\ \sum_i (1 - X_i) Y_i \end{pmatrix}$$

wobei $\mathbf{Y}$ den n-dimensionalen **Vektor** der abhängigen Variablen bezeichnet.

Unter Verwendung der obigen Ergebnisse für den bedingten und unbedingten Erwartungswert von Y_i erhalten wir

$$E\left[\frac{1}{n}\mathbf{X}'\mathbf{Y}\right] \quad = \quad \begin{pmatrix} \theta\beta_1 \\ (1-\theta)\beta_2 \end{pmatrix}$$

Daraus erhalten wir für den Wahrscheinlichkeitsgrenzwert des Kleinstquadrate Schätzers

$$\text{plim}\hat{\boldsymbol{\beta}} \quad = \quad \left(\text{plim } \frac{1}{n}\mathbf{X}'\mathbf{X}\right)^{-1} \text{plim } \frac{1}{n}\mathbf{X}'\mathbf{Y} \quad = \quad \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \quad , \qquad (28.9)$$

d.h. der Schätzer ist konsistent.

28.2.3 Formulierung mittels $\tau = \beta_1 - \beta_2$

Die im vorigen Abschnitt gewählte „symmetrische“ Formulierung in Gleichung (28.6) führt zwar zu einem leicht interpretierbaren Ergebnis (bezüglich des Bias bzw. der Inkonsistenz), ist aber nicht die übliche. Deshalb wird jetzt die Analyse auch für den Fall dargestellt, dass statt der genannten Gleichung geschrieben wird:

$$\begin{aligned} Y_i &= \mu + \beta_1 X_i + \beta_2 (1 - X_i) + U_i \quad , \quad i = 1, 2, \ldots, n \\ &= (\mu + \beta_2) + (\beta_1 - \beta_2) X_i + U_i \\ &= \gamma + \tau X_i + U_i \end{aligned} \qquad (28.10)$$

mit

$$\gamma = \mu + \beta_2 \quad \text{und} \quad \tau = \beta_1 - \beta_2 \quad .$$

Daraus ergibt sich als bedingter Erwartungswert von Y

$$E(Y|X) \quad = \quad \begin{cases} \gamma & \text{wenn} \quad X = 0 \\ \gamma + \tau & \text{wenn} \quad X = 1 \end{cases} \qquad (28.11)$$

und für den unbedingten Erwartungswert erhalten wir (nach den Regeln des iterierten Erwartungswertes)

$$E(Y) \quad = \quad \theta(\gamma + \tau) + (1 - \theta)\tau \quad = \quad \gamma + \theta\tau \quad . \qquad (28.12)$$

Weiter erhalten wir wegen

$$E(XY|X) \quad = \quad \gamma X + \tau X^2$$

$$E(XY) \quad = \quad \theta\,(\gamma + \tau) \tag{28.13}$$

wobei $E(X) = E(X^2) = \theta$ verwendet wurde.

In diesem Fall lautet der OLS-Schätzer für den Parameter τ:

$$\hat{\tau} \quad = \quad \frac{s_{xy}}{s_{xx}} \tag{28.14}$$

Wegen

$$\text{plim}\,(s_{xx}) = var(X) = \theta(1-\theta)$$

und

$$\text{plim}\,(s_{xy}) = cov(XY) = \theta\,(\gamma + \tau) \; - \; \theta[\gamma + \theta\,\tau] = \theta\,(1 \; - \; \theta)\,\tau$$

ergibt sich unmittelbar

$$\text{plim}\,(\hat{\tau}) = \frac{\theta\,(1 \; - \; \theta)\,\tau}{\theta(1-\theta)} = \tau \tag{28.15}$$

d.h. der OLS-Schätzer ist konsistent.

28.3 Randomisierte Dummy-Variable (Einfache Varianzanalyse mit randomisierten Effekten)

28.3.1 Einige nützliche Resultate

Für die folgenden Ausführungen ist es nützlich, die Methode der Post-Randomisierung (PRAM) als Fehlermodell zu interpretieren. Dies soll hier etwas ausführlicher dargestellt werden, weil diese Ergebnisse nicht ganz üblich sind.

Alternativ zur üblichen PRAM-Analyse unterstellen wir im Folgenden die Beziehung

$$X^a \quad = \quad X \quad + \quad V \tag{28.16}$$

wobei X die binäre Regressorvariable und V ebenfalls eine diskrete Zufallsvariable ist, die stochastisch unabhängig von U ist.[47] Die Fehlervariable V nimmt Werte aus der Menge $\{-1\,,\,0\,,\,+1\}$ an und es gilt

$$\begin{array}{lcl}
P(X^a = 1|X = 0) = \text{P(V = 1 | X = 0)} & = & 1 \;\text{-}\; \pi_1 \\
P(X^a = 1|X = 1) = \text{P(V = 0 | X = 1)} & = & \pi_1 \\
P(X^a = 0|X = 0) = \text{P(V = 0 | X = 0)} & = & \pi_0 \\
P(X^a = 0|X = 1) = \text{P(V = -1 | X = 1)} & = & 1 \;\text{-}\; \pi_0
\end{array}$$

47) Frazis und Loewenstein (2003, S.153) weisen darauf hin, dass dies eine starke Annahme sein kann. Sie betrachten allerdings nicht die Anonymisierung, sondern die falsche Beantwortung von Fragen (mit zwei Antwortkategorien).

Dabei wird im Fall der Anonymisierung im Allgemeinen (siehe oben) die „symmetrische“ Formulierung verwendet, d.h. es gilt

$$\pi_1 \quad = \quad \pi_0 \quad = \quad \pi \quad ,$$

was auch für das Folgende angenommen wird.

Es sollte beachtet werden, dass die Realisationen von X und V nicht beliebig kombinierbar sind. Beispielsweise würde sich für $X = 1$, $V = 1$ der Wert 2 für die anonymisierte Zufallsvariable ergeben. In der folgenden Tabelle 28.1 ist die gemeinsame Verteilung von X und V dargestellt.[48]

Tabelle 28.1: Gemeinsame Verteilung von X und Y

	X = 1	X = 0	
V = -1	$(1 - \pi)\ \theta$	**0**	$(1 - \pi)\ \theta$
V = 0	$\pi\ \theta$	$\pi\ (1 - \theta)$	π
V = +1	**0**	$(1 - \pi)\ (1 - \theta)$	$(1 - \pi)\ (1 - \theta)$
	θ	$1 - \theta$	1

Daraus ergibt sich die Kovarianz zwischen X und V wie folgt: Zunächst erhalten wir aus der Randverteilung für V (letzte Spalte in Tabelle 28.1)

$$E(V) \quad = \quad \theta\,(1-\pi)\,(1-2\theta)$$

sowie aus der gemeinsamen Verteilung von X und V

$$E[X\,V] \;=\; -\,(1-\pi)\,\theta \quad .$$

Da $E(X) = \theta$ gilt, erhalten wir somit für die Kovarianz:

$$\begin{aligned} cov[X,V] \quad &= \quad E[XV] \;-\; E[X]\,E[V] \\ &= \quad -(1-\pi)\,\theta \;-\; \theta\,(1-\pi)\,(1-2\,\theta)) \\ &= \quad -\,2\,(1-\pi)\,\theta\,(1-\theta) \\ &< \quad 0\,. \end{aligned} \tag{28.17}$$

Man beachte, dass im Gegensatz zum „klassischen Fall“ hier die Kovarianz zwischen Fehler und Regressor *nie* gleich Null und stets negativ sein wird. Dies wird für die Instrumentvariablen-Schätzung relevant werden. Außerdem hat die Fehlervariable V nur im speziellen Fall $\theta = 1/2$ eine symmetrische Verteilung und damit Erwartungswert Null.

Entsprechend lässt sich die gemeinsame Verteilung von X^a und V ableiten. Die einzelnen Resultate dazu finden sich in Tabelle 28.2.[49]

48) Für eine Ableitung dieser Ergebnisse siehe Ronning (2004a).

49) Für eine Ableitung dieser Ergebnisse siehe wieder Ronning (2004a).

Tabelle 28.2: Gemeinsame Verteilung von X und V

	$X^a = 1$	$X^a = 0$	
V = -1	**0**	(1 - π) θ	(1 - π) θ
V = 0	π θ	π (1 - θ)	π
V = +1	(1 - π) (1 - θ)	**0**	(1 - π) (1 - θ)
	(2 π - 1) θ + (1 - π)	π - (2 π - 1) θ	1

Daraus bestimmen wir die Kovarianz zwischen X^a und V. Aus der Tabelle ist unmittelbar

$$E(X^a) \quad = \quad (2\pi - 1)\,\theta \;+\; (1 - \pi)$$

zu entnehmen. Außerdem wurde $E(V)$ bereits weiter oben bestimmt. Ferner folgt aus (28.2):

$$E[X^a V] \quad = \quad (1 - \pi)\,(1 - \theta)$$

Unter Verwendung dieser Ergebnisse erhalten wir dann als Kovarianz von X^a und V:

$$\begin{aligned} cov(X^a, V) &= E[X^a V] \;-\; E[X^a]\,E[V] \\ &= (1 - \pi)\,(1 - \theta) \quad - \quad [(2\pi - 1)\,\theta \;+\; (1 - \pi)]\;[(1 - \pi)\,(1 - 2\,\theta)] \\ &= (1 - \pi)[2\theta(1 - \theta) + (2\theta - 1)^2\pi] \\ &> 0 \quad . \end{aligned} \tag{28.18}$$

Man beachte, dass diese Kovarianz stets nichtnegativ ist, was direkt aus der Nichtnegativität der Summanden in der eckigen Klammer folgt.[50]

50) Das obige Ergebnis in der zweiten Zeile von (28.18) wurde wie folgt gewonnen:

$$\begin{aligned} (1 - \pi)(1 - \theta) \;-\; & [(2\pi - 1)\,\theta + (1 - \pi)][(1 - \pi)(1 - 2\theta)] \\ & = (1 - \pi)((1 - \theta) - (2\pi - 1)\theta(1 - \pi)(1 - 2\theta) - (1 - \pi)^2(1 - 2\theta) \\ & = (1 - \pi)[1 - \theta - (2\pi - 1)\theta(1 - 2\theta) - (1 - \pi)(1 - 2\theta)] \\ & = (1 - \pi)[1 - \theta - (1 - 2\theta)\{(2\pi - 1)\theta + 1 - \pi\}] \\ & = (1 - \pi)[1 - \theta - (1 - 2\theta)\{\pi(2\theta - 1) + 1 - \theta\}] \\ & = (1 - \pi)[(1 - \theta) + (2\theta - 1)^2\pi + (2\theta - 1)(1 - \theta)] \\ & = (1 - \pi)[2\theta(1 - \theta) + (2\theta - 1)^2\pi] \end{aligned}$$

Da beide Summanden in der eckigen Klammer positiv sind (sofern $0 < \pi < 1$ und $0 < \theta < 1$ gelten), ist der Ausdruck insgesamt positiv.

28.3.2 Inkonsistenz des OLS-Schätzers bei Randomisierung

Wir betrachten wieder das Modell (28.10), d.h.

$$Y_i \quad = \quad \gamma \quad + \quad \tau\, X_i \quad + \quad U_i \quad .$$

Allerdings ist der Regressor nicht beobachtbar. Beobachtet wird stattdessen die randomisierte binäre Zufallsvariable X^a, für die (28.16) gilt, d.h.

$$X^a \quad = \quad X \quad + \quad V \quad .$$

Demnach lautet das zu schätzende Modell

$$Y_i \quad = \quad \gamma \quad + \quad \tau\,(X_i^a - V_i) \quad + \quad U_i \tag{28.19}$$

oder

$$\begin{aligned} Y_i \quad &= \quad \gamma \quad + \quad \tau\, X_i^a \quad + \quad (U_i \,-\, \tau\, V_i) \\ &= \quad \gamma \quad + \quad \tau\, X_i^a \quad + \quad \varepsilon_i \end{aligned} \tag{28.20}$$

mit

$$\varepsilon_i \quad = \quad U_i \,-\, \tau\, V_i \quad .$$

Dabei ist – wie üblich – der Fehlerterm ε mit der fehlerbehafteten Regressorvariablen X^a korreliert. Wir wollen dies für den hier betrachteten Fall zeigen.

Wir beachten zunächst, dass

$$cov(X^a, \varepsilon) \;=\; cov(X^a, U) \,-\, \tau\, cov(X^a, V) \;=\; -\,\tau\, cov(X^a, V),$$

wobei die Annahme[51] ausgenutzt wird, dass sowohl X als auch V und somit auch X^a von U unabhängig sind. Dann folgt aus Gleichung (28.18), dass (gegeben $\tau > 0$) die Kovarianz zwischen X^a und ε stets negativ ist:

$$cov(X^a, \varepsilon) \quad = \quad -\,\tau\; cov(X^a, V) \quad < \quad 0 \quad \text{wenn } \tau \;>\; 0 \quad . \tag{28.21}$$

Als Wahrscheinlichkeitsgrenzwert von $\hat{\tau}$ ergibt sich

$$\text{plim}\,(\hat{\tau}) \quad = \quad \frac{cov(Y, X^a)}{var(X^a)} \quad .$$

Dabei gilt

$$cov(Y, X^a) \quad = \quad \tau\, var(X^a) \quad + \quad cov(X^a, \varepsilon)$$

und somit unter Verwendung der Gleichungen (28.2), (28.18) und (28.21)

$$cov(Y, X^a) \;=\; \tau\,[(2\,\pi \,-\, 1)\,\theta \;+\; (1 \,-\, \pi)] \;-\; \tau\,(1 \,-\, \pi)\,\left[2\,\theta\,(1 \,-\, \theta) \;+\; (2\,\theta \,-\, 1)^2\,\pi\,\right]$$

51) Siehe Unterabschnitt 28.2.2.

Allerdings ist diese Formel sehr unübersichtlich. Wesentlich ist nur, dass sich wegen der negativen Kovarianz zwischen X^a und ε (bzw. wegen der positiven Kovarianz zwischen X^a und V) für den Wahrscheinlichkeitsgrenzwert des OLS-Schätzers

$$\text{plim}\,(\hat{\tau}) \quad = \quad \tau \quad + \quad \frac{cov(X^a, \varepsilon)}{var(X^a)} \quad = \quad \tau \left(1 - \frac{cov(X^a, V)}{var(X^a)}\right) \quad < \quad \tau \quad (28.22)$$

ergibt, d.h. bei Randomisierung oder Fehlklassifikation der binären Einflussvariablen wird der „Treatment-Effekt“ *unter*schätzt.

Wir wollen noch untersuchen, wie der Bias-Term

$$\frac{cov(X^a, V)}{var(X^a)} \quad = \quad \frac{(1 - \pi)\,\left[2\,\theta\,(1 - \theta) \,+\, (2\,\theta - 1)^2\,\pi\,\right]}{\theta\,(1 - \theta)}$$

von den Parametern π und θ abhängt. Dies ergibt sich aus Tabelle 28.3. Man erkennt deutlich, dass der Bias zunimmt, wenn sich die Verbleibewahrscheinlichkeit π von Eins entfernt. Außerdem ist die Verzerrung größer, wenn der Anteil θ der „Behandelten“ („treated“) von 0,5 abweicht.

Tabelle 28.3: Bias-Faktor $cov(X^a, V) \,/\, var(X^a)$

	π									
θ	0,99	0,95	0,90	0,85	0,80	0,75	0,70	0,65	0,60	0,55
0,980	0,331	0,722	0,847	0,900	0,930	0,949	0,963	0,974	0,984	0,992
0,900	0,084	0,327	0,512	0,633	0,719	0,786	0,840	0,885	0,926	0,964
0,820	0,046	0,205	0,360	0,483	0,584	0,671	0,747	0,816	0,880	0,941
0,740	0,031	0,148	0,278	0,393	0,496	0,592	0,680	0,764	0,845	0,923
0,660	0,024	0,119	0,232	0,338	0,441	0,539	0,635	0,728	0,820	0,910
0,580	0,021	0,104	0,207	0,309	0,410	0,510	0,609	0,707	0,805	0,903
0,500	0,020	0,100	0,200	0,300	0,400	0,500	0,600	0,700	0,800	0,900
0,420	0,021	0,104	0,207	0,309	0,410	0,510	0,609	0,707	0,805	0,903
0,340	0,024	0,119	0,232	0,338	0,441	0,539	0,635	0,728	0,820	0,910
0,260	0,031	0,148	0,278	0,393	0,496	0,592	0,680	0,764	0,845	0,923
0,180	0,046	0,205	0,360	0,483	0,584	0,671	0,747	0,816	0,880	0,941
0,100	0,084	0,327	0,512	0,633	0,719	0,786	0,840	0,885	0,926	0,964
0,020	0,331	0,722	0,847	0,900	0,930	0,949	0,963	0,974	0,984	0,992

28.4 Instrumentvariablen-Schätzung

28.4.1 Binäre Instrumentvariable

Wir betrachten für die beiden binären Zufallsvariablen X und Z die folgende gemeinsame Verteilung:

Tabelle 28.4: Gemeinsame Verteilung von X und Z

		Z	
		1	0
X	1	α_{11}	α_{10}
	0	α_{01}	α_{00}

Dabei gilt

$$\alpha_{00} \;=\; 1 \,-\, \alpha_{11} \,-\, \alpha_{10} \,-\, \alpha_{01}$$

sowie

$$E[X] \;=\; \alpha_{11} \,+\, \alpha_{10} \text{ und } E[Z] \;=\; \alpha_{11} \,+\, \alpha_{01}$$

und

$$E[X\,Z] \;=\; \alpha_{11}$$

Demnach gilt für die Kovarianz

$$cov[X\,,\,Z] \quad = \quad \alpha_{11} \,-\, (\alpha_{11} \,+\, \alpha_{10})\,(\alpha_{11} \,+\, \alpha_{01}) \;=\; \alpha_{11}\,\alpha_{00} \,-\, \alpha_{01}\,\alpha_{10}$$

Diese Kovarianz (und damit die Korrelation) ist maximal, wenn

$$\alpha_{10} \;=\; \alpha_{01} \;=\; 0 \tag{28.23}$$

gilt, d.h. wenn die Tabelle Diagonalgestalt aufweist. Man beachte: Dies ist - empirisch argumentiert - äquivalent mit der Forderung, dass Z immer dann eine „1“ anzeigt, wenn auch X dies tut. Theoretisch gesehen bedeutet dies, dass die Wahrscheinlichkeit dafür, dass Z eine „0“ anzeigt, wenn X eine „1“ anzeigt, gleich Null ist. Aus der Diagonalform folgt, dass die beiden Zufallsvariablen einen identischen Erwartungswert besitzen, d.h. es muss $E[X] = E[Z] \equiv \theta$ gelten.[52]

52) Dagegen ist die Korrelation Null, wenn

$$\alpha_{11}\,\alpha_{22} \;=\; \alpha_{12}\,\alpha_{21}$$

gilt. In diesem Fall sind die beiden Zufallsvariablen auch stochastisch unabhängig

Im Folgenden[53] soll Z eine „optimale“ Instrumentvariable sein, die sehr hoch mit der Regressorvariablen X korreliert ist und gleichzeitig möglichst geringe Korrelation mit dem Fehler V aufweisen soll. Falls wir den obigen „Idealfall“ aus Gleichung (28.23) verlangen, d.h. eine maximale Korrelation zwischen X und Z voraussetzen, muss (siehe oben)

$$E[X] \quad = \quad E[Z] \quad \equiv \quad \theta \tag{28.24}$$

gelten. Andererseits sollte nach Möglichkeit die Kovarianz zwischen Z und V gleich Null sein. Im Folgenden wird demonstriert, dass dies nicht der Fall ist, vielmehr diese Kovarianz stets negativ ist.

Um die allgemeine Formel für die Kovarianz zwischen V und Z abzuleiten, benötigen wir die gemeinsame Verteilung zwischen diesen beiden Zufallsvariablen. Dazu verwenden wir die Tabelle 28.5. Dort sind die **Randverteilungen** für beide Zufallsvariablen, die ja bekannt sind, eingetragen. Für V siehe dazu Unterabschnitt 28.3.1 und für Z die oben abgeleitete Tatsache, dass die binäre Variable Z denselben Erwartungswert wie X aufweist (und damit auch dieselbe Verteilung). Ferner sind – wie bei der gemeinsamen Verteilung von X und V – die beiden „strukturellen Nullen“ in der gemeinsamen Verteilung zu berücksichtigen. Siehe dazu die gemeinsame Verteilung von X und V in Tabelle 28.2.

Tabelle 28.5: Gemeinsame Verteilung von V und Z

	Z = 1	Z = 0	
V = -1	?	**0**	$(1 - \pi)\ \theta$
V = 0	?	?	π
V = +1	**0**	?	$(1 - \pi)\ \ (1 - \theta)$
	θ	$1 \ - \ \theta$	1

Damit sind aber auch alle anderen Zellen, die durch ein „?“ gekennzeichnet sind, bereits fixiert: In der ersten und dritten Zeile ergeben sich die Wahrscheinlichkeiten $P(V = -1\,,\, X = 0)$ und $P(V = +1\,,\, X = 1)$ durch die vorgegebenen Randwahrscheinlichkeiten für V. Damit sind aber dann auch die gemeinsamen Wahrscheinlichkeiten $P(V = 0\,,\, X = 0)$ und $P(V = 0\,,\, X = 1)$ bestimmt. Wir gelangen also zu dem Ergebnis, dass in diesem Idealfall die Verteilung von Z und V der Verteilung von X und V entspricht, d.h. diese Verteilung ist durch Tabelle 28.2 gegeben. Dies bedeutet insbesondere, dass für die Kovarianz von Z und V gilt:

53) Siehe zum Folgenden die Diskussion bei (Kane et al. 1999, section II)

$$\begin{aligned} cov[Z,V] &\equiv cov[X,V] \\ &= E[XV] - E[X]\,E[V] \\ &= -(1-\pi)\,\theta - \theta\,(1-\pi)\,(1-2\,\theta)) \\ &= -2\,(1-\pi)\,\theta\,(1-\theta) \\ &< 0\,. \end{aligned} \tag{28.25}$$

Kane et al. (1999, S.7) bemerken dazu: „The problem is that any variable which is correlated with a categorical indicator of 'true' schooling will generally also be correlated with the measurement error, since the measurement itself is related to schooling."

28.4.2 Der IV-Schätzer überschätzt den Treatment-Effekt

Wenn man den Parameter τ in dem Modell (28.10) mit dem üblichen IV-Schätzer schätzt, so läuft das auf einen Schätzer[54] hinaus, der das Verhältnis der empirischen Kovarianz zwischen Z und Y in Beziehung setzt zur empirischen Kovarianz zwischen Z und X^a, d.h.

$$\hat{\tau}^{IV} = \frac{s_{yz}}{s_{X^a z}} \tag{28.26}$$

Demnach erhalten wir für den Wahrscheinlichkeitsgrenzwert von $\hat{\tau}^{IV}$ unter Beachtung der Gleichungen (28.10) und (28.16) :

$$\text{plim}\,(\hat{\tau}^{IV}) = \frac{cov[Z,Y]}{cov[Z,X^a]} = \tau\,\frac{cov[Z,X]}{cov[Z,X] + cov[Z,V]} > \tau \quad . \tag{28.27}$$

Da einerseits die Kovarianz zwischen dem Regressor und der Instrumentvariablen positiv gemäß Konstruktion bzw. Voraussetzung ist und andererseits, wie oben gezeigt, die Kovarianz zwischen Z und V *negativ* ist, ergibt sich eine *Über*schätzung des Treatment-Effektes durch die IV-Schätzung.

---

54) In Matrixschreibweise lautet der Schätzvektor für den Parametervektor

$$\boldsymbol{\beta} = \begin{pmatrix} \gamma \\ \tau \end{pmatrix}$$

wie folgt:

$$\hat{\boldsymbol{\beta}}^{IV} = (\mathrm{Z}'\mathrm{X})^{-1}\,\mathrm{Z}'\mathrm{Y}$$

Dabei ist X die übliche $(n \times 2)-$Regressormatrix im Modell der Einfachregression (mit Einsen in der ersten Spalte) und Z die entsprechende Matrix der Instrumentvariablen, ferner Y der $(n \times 1)-$ **Vektor** der abhängigen Variablen.

28.4.3 Ein explizites Resultat

In Anlehnung an die Arbeit von Kane et al. (1999), allerdings mit der hier verwendeten Symbolik, soll nun ein explizites Resultat abgeleitet werden. Es sollen zwei fehlerbehaftete binäre Zufallsvariablen verfügbar sein, die beide stochastisch von der binären Einflussvariablen X abhängen und die hier mit X^a und Z bezeichnet werden.[55]. Für die Beziehung zwischen den drei erwähnten Variablen soll gelten:

$$X^a \quad = \quad \lambda_{10} \, + \, \lambda_{11} \, X \, + \, W_1 \tag{28.28}$$

und

$$Z \quad = \quad \lambda_{20} \, + \, \lambda_{21} \, X \, + \, W_2 \tag{28.29}$$

wobei W_1 und W_2 zwei Störvariablen mit den üblichen Eigenschaften sind. Kane et al. (1999) bemerken, dass im klassischen Fall des „Fehler in den Variablen-Modells" die beiden additiven Konstanten Null und die beiden multiplikativen Konstanten gleich 1 sind. Dagegen sind im hier betrachteten Fall die additiven Konstanten ungleich Null und die multiplikativen Konstanten sind *kleiner als Eins*. Wir wählen speziell

$$\lambda_{10} \, = \, 1 \, - \, \pi_0 \qquad \text{und} \qquad \lambda_{11} \, = \, \pi_1 \, + \, \pi_0 \, - \, 1$$

und unterstellen dabei, dass $\lambda_{11} > 0$ gilt, d.h. dass

$$\pi_1 \, + \, \pi_0 \quad > \quad 1 \tag{28.30}$$

gilt. Die Gleichung (28.28) lautet demnach nun wie folgt:

$$X^a \quad = \quad (1 \, - \, \pi_0) \, + \, (\pi_1 \, + \, \pi_0 \, - \, 1) \, X \, + \, W_1 \tag{28.31}$$

Man kann leicht nachprüfen, dass wir damit wieder bei dem Modell der Randomisierung sind, allerdings mit „asymmetrischen" Übergangswahrscheinlichkeiten.[56]

Insbesondere erhalten wir als Erwartungswert von X^a den Ausdruck

$$E(X^a) \quad = \quad (\pi_1 + \pi_0 \, - \, 1) \, \theta \, + \, (1 \, - \, \pi_0)$$

was dem vorher verwendeten Ausdruck $E(X^a) = (2\pi - 1)\theta + (1 - \pi)$ für den zuvor behandelten Fall $\pi_1 = \pi_0 = \pi$ entspricht.

Die Parameter der Gleichung (28.29) für Z bleiben unspezifiziert, sollen jedoch ebenfalls

$$\lambda_{20} \, > \, 0 \qquad \text{und} \qquad \lambda_{21} \, < \, 1$$

55) Bei Kane et al. (1999) ist S^* der wahre Regressor und S_1 und S_2 sind die beiden fehlerbehafteten Variablen.

56) Diese allgemeinere Formulierung wurde auch bereits von Hausman et al. (1998) verwendet. Allerdings betrachten sie eine randomisierte *abhängige* Variable in einem Probitmodell. Siehe auch Ronning (2005) und Ronning et al. (2005)

erfüllen.

In diesem Modell lässt sich nun die Kovarianz zwischen X^a und Z explizit angeben. Es ergibt sich

$$cov[X^a, Z] = cov[\{(1-\pi_0) + (\pi_1 + \pi_0 - 1)\, X + W_1\}, Z] = (\pi_1 + \pi_0 - 1)\, cov[X, Z]$$

Wenn wir dieses Ergebnis in die obige Gleichung (28.27) einsetzen, ergibt sich[57]

$$\begin{aligned} \text{plim}\,(\hat{\tau}^{IV}) &= \frac{cov[Z,Y]}{cov[Z,X^a]} = \tau\, \frac{cov[Z,X]}{(\pi_1 + \pi_0 - 1)\, cov[Z,X]} \\ &= \tau\, \frac{1}{\pi_1 + \pi_0 - 1} > \tau\,. \end{aligned} \tag{28.32}$$

Man beachte, dass der geschätzte Wert von τ beliebig groß werden kann, wenn die Summe der beiden Wahrscheinlichkeiten gegen 1/2 tendiert. Und es kann sich sogar ein Vorzeichenwechsel einstellen, wenn $\pi_1 + \pi_0 < 1$ gilt, was dem Fall $\pi < 1/2$ im „symmetrischen" Fall entspricht.

Frazis und Loewenstein (2003, Proposition 3) erhalten dasselbe Resultat, allerdings auf anderem Wege. Dabei betrachten sie das allgemeinere Modell, in dem neben dem Treatment-Effekt auch beliebige andere Regressoren zugelassen werden. Sie zeigen, dass nur der Treatment-Effekt unterschätzt wird, während alle anderen Regressionskoeffizienten konsistent geschätzt werden. Sie konstruieren auf der Basis dieses Resultates einen Momentenschätzer, für den Konsistenz gegeben ist. Außerdem betrachten sie den Fall, dass die binäre Regressorvariable endogen ist. „When the mismeasured variable is endogenous, the IV estimate and the measurement errors can be used to bound its coefficient."(Frazis und Loewenstein 2003, S. 152)

57) Siehe Kane et al. (1999), Formel (5).

Kapitel 29

Post-Randomisierung der erklärenden diskreten Variablen im Probit-Modell

In Kapitel 27 wurden die Auswirkungen der Post-Randomisierung der abhängigen diskreten Variablen in einem binären Probit-Modell anhand theoretischer Überlegungen sowie von Simulationsstudien mit simulierten Daten und Querschnittsdaten des IAB-Betriebspanels untersucht. Anschließend wurden zusätzlich die metrische Einflussgröße mit Mikroaggregationsverfahren und stochastischen Überlagerungen anonymisiert. Die diskreten Einflussgrößen blieben jedoch unmaskiert.

Nun wird in diesem Abschnitt der Fall betrachtet, dass im Probit-Modell mit metrischen und diskreten Einflussgrößen nicht die abhängige diskrete Variable, sondern die erklärenden diskreten Variablen mit Post-Randomisierung anonymisiert werden.

29.1 Theoretische Überlegungen

Aufgrund der im vorangegangenen Abschnitt durchgeführten Herleitung, dass die Post-Randomisierung einer erklärenden diskreten beziehungsweise kategorialen Variablen in linearen Modellen zu verzerrten Schätzungen führt und diese auch nicht durch Instrumentenvariablenschätzer korrigiert werden kann, ist klar, dass auch die Post-Randomisierung von erklärenden diskreten Variablen im nichtlinearen Probit-Modell zu verzerrten Schätzern führt. Dies soll im Folgenden am Beispiel der Probit-Schätzung zur Erklärung der Tarifbindung mit den Daten des IAB-Betriebspanels für Baden-Württemberg illustriert werden.

29.2 Praxisbeispiel

Es wird das Probit-Modell zur Erklärung der Tarifbindung mit den Daten des IAB-Betriebspanels 2002 für Baden-Württemberg geschätzt, wie es in Abschnitt 21.2 beschrieben wurde. Dabei werden zwei der Dummy-Variablen für einzelne Wirtschaftszweige – diejenige für das Verarbeitende Gewerbe und diejenige für die Baubranche – mit Post-Randomisierung so bearbeitet, dass die Ausprägungen mit Wechselwahrscheinlichkeiten von fünf beziehungsweise zehn Prozent zwischen diesen beiden Ausprägungen getauscht werden. Die Ergebnisse der Probit-Schätzung sind in Tabelle 29.1 dargestellt. In beiden Fällen ergibt sich eine Verzerrung der geschätzten Koeffizienten und der Teststatistiken, wobei die Verzerrung mit zunehmender Wechselwahrscheinlichkeit ansteigt.

Tabelle 29.1: Probit-Schätzung zur Erklärung der Tarifbindung – Dummy-Variablen für Verarbeitendes Gewerbe und Baugewerbe mit PRAM anonymisiert, IAB-Betriebspanel 2002 für Baden-Württemberg, 500 Replikationen

	PRAM mit Wechselwahrscheinlichkeit 5%		PRAM mit Wechselwahrscheinlichkeit 10%	
Variablen	Durchschn. Koeff.	(Durchschn. t-Werte)	Durchschn. Koeff.	(Durchschn. t-Werte)
Log. Beschäftigung	-0,941	(-7,16)	-0,901	(-6,90)
Baugewerbe	0,292	(12,06)	0,286	(11,93)
Handel	0,598	(3,81)	0,483	(3,24)
Dienstleistungssektor	0,355	(2,70)	0,333	(2,52)
Öffentliche Verwaltung	0,019	(0,20)	0,001	(0,01)
Konst.	0,783	(4,52)	0,769	(4,43)
	Relative Abweichungen von den Originalwerten in %			
Log. Beschäftigung	4,8	4,02	8,78	7,51
Baugewerbe	2,83	1,47	4,93	2,53
Handel	20,03	14,57	35,47	27,35
Dienstleistungssektor	6,38	6,25	12,13	12,5
Öffentliche Verwaltung	51,05	50,00	98,56	97,5
Konst.	1,68	1,74	3,39	3,70
Durchschn.	14,46	13,01	27,21	25,18

Es besteht Forschungsbedarf dahingehend, inwiefern sich die Idee der SIMEX-Korrektur auf den Fall einer randomisierten Dummy-Variablen übertragen lässt.

Kapitel 30

Ein Fazit für den Einsatz der Post-Randomisierung in ökonometrischen Modellen

Wird eine diskrete abhängige Variable in einem ökonometrischen Modell mit Post-Randomisierung anonymisiert, wie die abhängige Variable in einem binären Probit-Modell, so lassen sich, wie in Unterabschnitt 27.1.1 gezeigt wurde, Korrekturverfahren konstruieren, indem die Randomisierung in der zu maximierenden Likelihoodfunktion berücksichtigt wird.

Diese Korrektur lässt sich unabhängig davon realisieren, ob auch die metrischen Regressoren mit datenverändernden Anonymisierungsverfahren bearbeitet werden. Werden die Regressoren stochastisch überlagert, so muss ergänzend zur PRAM-Korrektur eine Korrektur der Überlagerungen durchgeführt werden, beispielsweise, indem die SIMEX-Korrektur nicht auf die naive Probit-Schätzung, sondern auf die PRAM-korrigierende Probit-Schätzung angewendet wird. Für die additive stochastische Überlagerung liegen hierzu erste Ergebnisse vor.

Dabei ergeben sich die gleichen Einschränkungen, die auch bei der isolierten Anwendung der beiden Anonymisierungsverfahren und der beiden Korrekturschätzer gelten. Für die Post-Randomisierung gilt insbesondere, dass die Korrekturschätzung nur zu guten Ergebnissen führt, wenn die Wechselwahrscheinlichkeit nicht zu hoch gewählt wird. Dabei hängt die maximal korrigierbare Wechselwahrscheinlichkeit allerdings sowohl vom Beobachtumgsumfang als auch von der Beschaffenheit des zu schätzenden Modells ab. Für das gewählte Fallbeispiel des Probit-Modells zur Erklärung der Tarifbindung ergibt sich bei einem kleinen Beobachtungsumfang (Baden-Württemberg-Daten) von rund 1.200 eine maximale Wechselwahrscheinlichkeit von fünf Prozent. Auch durch die zusätzliche Mikroaggregation der Regressoren ergeben sich keine Einschränkungen für die Korrigierbarkeit der durch die Post-Randomisierung hervorgerufenen Verzerrung. Für multiplikative stochastische Überlagerungen besteht noch Forschungsbedarf.

Als problematischer erweist sich die Anonymisierung von diskreten Merkmalen durch Post-

Randomisierung, wenn diese als erklärende Variable in einem linearen oder nichtlinearen Modell eingesetzt werden. Im linearen Modell ergibt sich eine Verzerrung, die auch nicht durch den Einsatz des Instrumentenvariablen-Schätzers korrigiert werden kann. In nichtlinearen Modellen, wie dem Probit-Modell, tritt ebenfalls eine Verzerrung auf. Es ist noch zu untersuchen, ob solche durch Post-Randomisierung von Dummy-Variablen hervorgerufene Verzerrungen in linearen und nichtlinearen Modellen ebenfalls durch ein Vorgehen in Analogie zum SIMEX-Schätzer korrigierbar sind.

Da in der Regel nicht klar ist, ob eine mit Post-Randomisierung anonymisierte Variable als abhängige oder erklärende Variable in einem Regressionsmodell verwendet wird, ist das Verfahren der Post-Randomisierung nach jetzigem Wissensstand eher nicht geeignet, um zur Erstellung von Scientific-Use-Files eingesetzt zu werden. Allerdings besteht hier, wie bereits angemerkt wurde, weiterer Forschungsbedarf.

Teil X

Möglichkeiten und Auswirkungen einer Beschränkung der Anonymisierung auf die Überschneidungsmerkmale

Im Folgenden befassen wir uns mit den Auswirkungen von Anonymisierungsstrategien, die zum Ziel haben, die Anonymisierungsmaßnahmen auf die Überschneidungsmerkmale zu beschränken. Das Motiv für ein solches Vorgehen ist, die restlichen im Datensatz vorhandenen Merkmale unberührt zu lassen und damit für die wissenschaftlichen Analysen uneingeschränkt zu erhalten.

In Kapitel 31 wird zunächst eine Abgrenzung der Überschneidungsmerkmale versucht. Außerdem werden mögliche Anonymisierungsstrategien vorgestellt, die sich ausschließlich auf Überschneidungsmerkmale richten. Anschließend wird in Kapitel 32 zu den Implikationen dieser Strategien auf die Datensicherheit Stellung genommen. Ausführungen zu ihren Auswirkungen auf das Analysepotenzial folgen in Kapitel 33. Kapitel 34 zieht abschließend aus Sicht beider Aspekte ein Fazit zur Beurteilung der Strategie, die Anonymisierung auf die Überschneidungsmerkmale zu beschränken.

Kapitel 31

Abgrenzung der Überschneidungsmerkmale und mögliche Anonymisierungsstrategien

Eine Anonymisierungsstrategie, die die Anonymisierungsmaßnahmen auf die Überschneidungsmerkmale beschränkt, bedeutet aus Sicht der Datensicherheit, dass im Falle einer Reidentifikation eines Merkmalsträgers durch einen potenziellen Datenangreifer sämtliche Informationen – möglicherweise abgesehen von den Überschneidungsmerkmalen selbst – für den Datenangreifer im Original vorliegen und sich somit zur Erreichung der faktischen Anonymität das in Kapitel 12 eingeführte Enthüllungsrisiko auf den Anteil der reidentifizierten Merkmalsträger reduziert. Dies wiederum macht, verglichen mit den auf sämtliche Merkmale angewendeten Verfahren, eine weit stärkere Veränderung der Überschneidungsmerkmale nötig. Bei der Anonymisierung steht man daher vor der Frage, inwieweit es vorteilhaft ist, auf Kosten des Informationsgehalts der Überschneidungsmerkmale, den Informationsgehalt der Nicht-Überschneidungsmerkmale voll zu erhalten, oder ob besser der Informationsgehalt der Nicht-Überschneidungsmerkmale zu Gunsten der Überschneidungsmerkmale eingeschränkt werden soll. Wie diese Frage zu entscheiden ist, hängt unter anderem davon ab, auf welchem Weg das Analysepotenzial in der Summe am besten erhalten bleibt. Hierbei ist zu beachten, dass die Überschneidungsmerkmale oftmals für wissenschaftliche Analysen wertvoll sind (z.B. Umsatz oder Beschäftigte). Eine stärkere Veränderung dieser Merkmale führt daher zu einer besonders starken Reduzierung des Analysepotenzials. Die Auswirkungen einer alleinigen Behandlung solcher Merkmale auf das Analysepotenzial werden in Kapitel 33 verdeutlicht.

Darüber hinaus haben die Ergebnisse des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ für die Projektstatistiken gezeigt, dass es Merkmalsträger gibt, die auch nach der Anwendung starker Anonymisierungsmaßnahmen leicht zu reidentifizieren sind (z.B. Unternehmen mit mehr als 1.000 Beschäftigten in einer dünn besetzten Branche). Besonders bei diesen Merkmalsträgern ist davon auszugehen, dass sich ein Datenangreifer der Korrektheit seiner Zuordnung nahezu sicher sein und er bei einer alleinigen Behandlung der Überschneidungsmerkmale für sämtliche verbleibende Zielmerkmale die Originalwerte enthüllen würde. In diesem Fall besteht aus Vertraulichkeitsgesichtspunkten

die Notwendigkeit, sämtliche Merkmale mit Anonymisierungsverfahren zu bearbeiten. Bei den Merkmalsträgern, die auf Basis der Korrektheit der Zuordnung zwischen Zusatzwissen und Zieldaten als ungeschützt einzustufen wären, kann der insgesamt nötige Schutz erst durch eine Verfremdung der Einzelwerte erreicht werden, indem die gefundenen Informationen für einen potenziellen Datenangreifer mehrheitlich uninteressant werden.

31.1 Zur Abgrenzung der Überschneidungsmerkmale

Bei einer Beschränkung der Anonymisierung auf die Überschneidungsmerkmale ist zunächst zu beachten, dass es schwierig ist, die Teilmenge der Überschneidungsmerkmale gegenüber den verbleibenden Merkmalen abzugrenzen. Aus diesem Grund wird in den nachfolgenden Abschnitten zwischen einem *engen Feld* und einem *weiten Feld* von Überschneidungsmerkmalen unterschieden. Das enge Feld beinhaltet solche Merkmale, die sowohl in den Zieldaten als auch in handelsüblichen Unternehmensdatenbanken enthalten sind und somit zu einem Massenfischzug herangezogen werden können. Das weite Feld beinhaltet zusätzlich Merkmale, welche dem Datenangreifer für einzelne Unternehmen, etwa als Ergebnis von Internetrecherchen, vorliegen können.

Während das enge Feld durch gewissenhafte Untersuchung verfügbarer Unternehmensdatenbanken abgegrenzt werden kann (siehe Unterabschnitt 11.3.1), können einem Datenangreifer für einen Einzelangriff Informationen über Merkmale vorliegen, welche weder anhand der Unternehmensdatenbanken noch durch Internetrecherche abgrenzbar sind (siehe die Ausführungen zum Expertenwissen in Unterabschnitt 11.3.3). Aus diesem Grund ist eine Abgrenzung des weiten Feldes von Überschneidungsmerkmalen schwer möglich und basiert letztlich auf einem Werturteil. Damit ist aber auch der Ansatz einer alleinigen Anonymisierung mittels der Überschneidungsmerkmale werturteilsbehaftet.

Ist ein Überschneidungsmerkmal hoch mit anderen im Datensatz vorhandenen Merkmalen korreliert, so kann eine separate Behandlung dieses Merkmals zu einer inakzeptablen Verringerung des Analysepotenzials bis hin zu Inplausibilitäten führen (z.B. können die eigentlich bestehenden Summenbeziehungen zwischen Merkmalen aufgelöst werden). Ein Spezialfall ergibt sich, wenn eine Erhebung aus einer Vielzahl von Untermerkmalen eines Merkmals besteht und dieses „Hauptmerkmal“ gleichzeitig ein Überschneidungsmerkmal ist. Dann sind aus Sicherheitsaspekten die Untermerkmale genauso als Überschneidungsmerkmale zu behandeln wie das Hauptmerkmal. Dies trifft bei den untersuchten Erhebungen besonders bei der Umsatzsteuerstatistik zu. Diese weist als Hauptmerkmal den Umsatz der Unternehmen auf und darüber verschiedene Unterumsatzgruppen (z.B. Umsatz zu 16 Prozent Umsatzsteuer). Eine Beschränkung der Anonymisierung auf Überschneidungsmerkmale ergibt daher bei der Umsatzsteuerstatistik wenig Sinn, da aufgrund der teilweise sehr hohen Korrelationen bis hin zu identischen Werten zweier Merkmale (z.B. bei Unternehmen, die lediglich Umsatz zu 16 Prozent Umsatzsteuer erwirtschaften) von vornherein bereits zahlreiche Merkmale als (zumindest indirekte) Überschneidungsmerkmale mit in die Betrach-

tung einbezogen werden müssten. Auch diese Gesichtspunkte zeigen, dass die Abgrenzung zwischen Überschneidungsmerkmalen und restlichen Merkmalen sehr schwer sein kann.

31.2 Mögliche Anonymisierungsstrategien

Will man die Anonymisierungsmaßnahmen auf die Überschneidungsmerkmale beschränken, so steigt c.p. das Enthüllungsrisiko an, weil ein höherer Anteil an gefundenen Werten für den Datenangreifer brauchbar wird (höherer Nutzen der Reidentifikation). Wie bereits ausgeführt wurde, kann dieser Effekt nur ausgeglichen werden, wenn auf der anderen Seite der Anteil der Zuordnungen reduziert wird. Dies setzt eine „stärkere" Anonymisierung der Überschneidungsmerkmale voraus. Bei additiver stochastischer Überlagerung ist eine alleinige Behandlung der Überschneidungsmerkmale wenig aussichtsreich, da dieses Verfahren gerade in den oben angesprochenen kritischen Datenbereichen weder eine korrekte Zuordnung verhindern noch die Brauchbarkeit von Einzelwerten für den Datenangreifer nennenswert reduziert. Auch bei der getrennten Mikroaggregation mit einer Gruppengröße von drei bis fünf wird das Reidentifikationsrisiko kaum verringert. Deshalb könnten sich folgende Strategien anbieten, um bei einer Beschränkung der Anonymisierungsverfahren auf die Überschneidungsmerkmale die faktische Anonymität sicherzustellen:

1. Vergrößerung der Gruppengröße bei der getrennten/eindimensionalen Mikroaggregation
2. Anwendung einer Form der Mikroaggregation auf die Überschneidungsmerkmale, die zu einer stärkeren Veränderung der Einzelwerte führt (z.B. gruppierte Mikroaggregation, gemeinsame/mehrdimensionale Mikroaggregation)
3. Anwendung von informationsreduzierenden Maßnahmen (wie z.B Umwandlung der metrischen Überschneidungsmerkmale in kategoriale Variable oder eine weitere Vergröberung der kategorialen Merkmale)
4. Anwendung einer speziellen Anonymisierungsmaßnahme ausschließlich auf die Überschneidungsmerkmale der besonders gefährdeten Unternehmen: z.B. Replacement oder Censoring für besonders gefährdete Unternehmen
5. Multiplikative stochastische Überlagerung der Überschneidungsmerkmale

Die möglichen Anonymisierungsstrategien bei Beschränkung der Anonymisierung auf die Überschneidungsmerkmale werden im Folgenden sowohl hinsichtlich ihrer Implikationen auf die Datensicherheit als auch auf das Analysepotenzial beurteilt.

Kapitel 32

Implikationen für die Datensicherheit

32.1 Allgemeine Beurteilung verschiedener Anonymisierungsstrategien

Generell ist eine Beschränkung der Anonymisierung auf die Überschneidungsmerkmale aus Sicht der Datensicherheit aus mehreren Gründen kritisch zu bewerten:

- Die Abgrenzung der Überschneidungsmerkmale ist bei Unternehmensdaten aufgrund der Vielzahl an Quellen des Zusatzwissens nicht eindeutig.
- Besonders gefährdete Teilgesamtheiten sind bei Unternehmensdaten nur schwer vor einer Reidentifikation zu schützen. Hier bietet die Reduzierung des Nutzens der gefundenen Werte für einen Datenangreifer eine wesentliche Möglichkeit zur Sicherstellung der faktischen Anonymität.

Unabhängig von diesen generellen Überlegungen kann für die im vorangegangenen Kapitel vorgestellten möglichen Anonymisierungsstrategien bei der Beschränkung der Anonymisierung auf die Überschneidungsmerkmale folgende Bewertung hinsichtlich ihrer Implikationen auf die Datensicherheit vorgenommen werden:

- Mit einer Umwandlung der metrischen Überschneidungsmerkmale in ausreichend vergröberte kategoriale Merkmale (Strategie 3) sollte in jedem Fall die faktische Anonymität erreicht werden. Allerdings stellt eine solche Maßnahme einen recht starken Eingriff in die Analysemöglichkeiten dar.
- Ähnliches gilt für Censoring oder Replacement besonders gefährdeter Unternehmen (Strategie 4).

- Die Beschränkung einer multiplikativen stochastischen Überlagerung auf die Überschneidungsmerkmale (Strategie 5) ist aus Sicht der Datensicherheit kritisch zu bewerten, weil sich bei der Simulation von Massenfischzügen mit den Projektdaten gezeigt hat, dass das Verfahren vor allem zu einer verlässlichen Verringerung des Nutzens gefundener Werte und weniger zu einer Verringerung des Reidentifikationsrisikos führt. Das Verfahren weist dabei den besonderen Vorteil auf, dass sich durch die Wahl der Parameter die vorgegebenen Abweichungsschwellen ziemlich exakt realisieren lassen. Sicherlich ließe sich durch eine Erhöhung der Varianz der Überlagerungen das Reidentifikationsrisiko reduzieren, allerdings sind einer solchen Erhöhung Grenzen gesetzt, weil die Überlagerungsfaktoren in jedem Fall positiv sein müssen, um Vorzeichenwechsel zu vermeiden.

Im folgenden Abschnitt wird am Beispiel der Kostenstrukturerhebung im Verarbeitenden Gewerbe die Anwendbarkeit unterschiedlicher Mikroaggregationsverfahren für eine alleinige Anonymisierung der Überschneidungsmerkmale (Strategien 1 und 2) näher beleuchtet.

32.2 Beurteilung von auf die Überschneidungsmerkmale beschränkten Mikroaggregationsverfahren

Auf Basis von Recherchen zum Zusatzwissen sind die vier Merkmale Regionalkennung, Wirtschaftszweigklassifikation, Gesamtumsatz und Anzahl der Beschäftigten für alle Unternehmen als Überschneidungsmerkmale anzusehen. Über Internetrecherchen konnte herausgefunden werden, dass auch die Merkmale Aufwand in F&E, Einsatz an Handelsware und Abschreibungen für einige Unternehmen bezüglich der Reidentifikationsgefahr eine Rolle spielen. Aus diesem Grunde wird ein „enges“[58] und ein „weites Feld“[59] an Überschneidungsmerkmalen betrachtet.

Es werden folgende Varianten der eindimensionalen und mehrdimensionalen Mikroaggregation auf die Daten der Kostenstrukturerhebung angewendet:

1.) Mikroaggregation für jedes Überschneidungsmerkmal einzeln.

   a) Enges Feld an Überschneidungsmerkmalen (MA30G_eng)
   b) Weites Feld an Überschneidungsmerkmalen (MA30G_weit)

   Über die Erhöhung der Anzahl k der jeweils gemittelten Merkmalsausprägungen kann der Anonymisierungsgrad entsprechend erhöht werden.[60]

58) Regionalkennung (BBR3), Wirtschaftszweigklassifikation (Zweistellerebene), Gesamtumsatz und Anzahl der Beschäftigten

59) Regionalkennung (BBR3), Wirtschaftszweigklassifikation (Zweistellerebene), Gesamtumsatz, Anzahl der Beschäftigten, F&E, Einsatz an Handelsware und Abschreibungen

2.) Mehrdimensionale Mikroaggregation über zehn Gruppen mit je drei metrischen Merkmalen.

   a) Enges Feld an Überschneidungsmerkmalen (MA10G_eng)

   b) Weites Feld an Überschneidungsmerkmalen (MA10G_weit)

   Im Vergleich zum ersten Verfahren werden die Überschneidungsmerkmale wesentlich stärker verändert. Die Gruppierung der Merkmale erfolgte durch eine Clusteranalyse. Dadurch werden die Überschneidungsmerkmale auf mehrere Gruppen verteilt. Nach der Mikroaggregation werden die veränderten Werte für die jeweiligen Überschneidungsmerkmale beibehalten, während für die restlichen Merkmale ihre ursprünglichen Werte wieder zugespielt werden.

Nachfolgend werden die Ergebnisse von Massenfischzügen zwischen der MARKUS-Datenbank und den Zieldaten einerseits (siehe Tabelle 32.1) und zwischen den Originaldaten und den Zieldaten (worst-case Szenario) andererseits (siehe Tabelle 32.2) dargestellt. Da in der MARKUS-Datenbank nur das enge Feld von Überschneidungsmerkmalen zur Verfügung stand, wurde zur Schätzung des Risikozuwachses bei Erhöhung der Anzahl von Überschneidungsmerkmalen das worst-case Szenario herangezogen.

Tabelle 32.1 dient dazu, die Veränderung des Schutzes in den Zieldaten beim Übergang von den klassischen Mikroaggregationsvarianten, in welchen alle Merkmale (d.h. sowohl Überschneidungs- als auch Nicht-Überschneidungsmerkmale) anonymisiert werden, zu den entsprechenden Varianten, in welchen allein die Überschneidungsmerkmale anonymisiert werden, zu verdeutlichen. Bei der alleinigen Behandlung der Überschneidungsmerkmale ist bei einer Reidentifikation eines Merkmalsträgers jede Ausprägung eines Nicht-Überschneidungsmerkmals brauchbar, während bei den klassischen Varianten eine Einzelinformation als unbrauchbar angesehen wird, wenn der anonymisierte Wert um mehr als γ von seinem Originalwert relativ abweicht. In den untenstehenden Berechnungen wurde beispielhaft $\gamma = 0,05$ (zur Definition vgl. Abschnitt 12.2) angenommen. Tabelle 32.2 soll die Veränderung des Schutzes beim Übergang von dem engen Feld zu dem weiten Feld an Überschneidungsmerkmalen beschreiben. Obwohl hier das worst-case Szenario, welches nicht als realistisch einzustufen ist, durchgeführt werden musste, wird die Veränderungsrate des Risikos mit wachsender Anzahl an Überschneidungsmerkmalen sichtbar.

Tabelle 32.1: Enthüllungsrisiken (MARKUS-Datenbank)

	MA10G	MA10G_eng	MA30G	MA30G_eng
Enthüllungsrisiko	0,16	0,21	0,32	0,32

60) D.h., dass jeweils k Merkmalsträger denselben Wert nach erfolgter Anonymisierung aufweisen.

Bei der mehrdimensionalen Mikroaggregationsvariante ist der Verlust an Schutz in den Daten beim Übergang von MA10G zu MA10G_eng deutlich ausgefallen. Weit weniger deutlich ist dieser Schutzverlust bei der eindimensionalen Variante, was an den sehr geringen Abweichungen der anonymisierten von den originalen Einzelwerten und dem damit verbundenen sehr hohen Anteil an für den Datenangreifer brauchbaren Informationen liegt. Wie obige Tabelle zeigt, ist der Schutzverlust für die gesamte Datei bei der Rundung auf zwei Nachkommastellen nicht erkennbar. Die Veränderung des Schutzes wird hier erst sichtbar, wenn man das Enthüllungsrisiko auf Teilmassen (wie z.B. Unternehmen einer dünn besetzten Branche) herunterbricht.

Tabelle 32.2: Enthüllungsrisiken (Worst-Case)

	MA10G	MA10G_eng	MA10G_weit
Enthüllungsrisiko	0,43	0,54	0,56

Erwartungsgemäß steigen die Trefferquoten mit der Anzahl der Überschneidungsmerkmale.[61] Allerdings ist die Wachstumsrate (und der damit einhergehende Schutzgewinn) erstaunlich gering. Bei den eindimensionalen Mikroaggregationsvarianten ist das Risiko bereits bei einem engen Feld an Überschneidungsmerkmalen so nah bei Eins, dass hier der Übergang kaum spürbar ist.

Während die Varianten der eindimensionalen Mikroaggregation mangels ausreichender Schutzwirkung nicht als faktisch anonym eingestuft werden konnten, wurde festgestellt, dass die Varianten der mehrdimensionalen Mikroaggregation zwar allen durchgeführten Datenangriffssimulationen standhielten, aber ein nicht ausreichendes Potenzial für wissenschaftliche Analysen bereithielten (siehe nachfolgender Abschnitt). Zusätzlich wurde untersucht, wie sich eine Erhöhung der Gruppengröße k bei der Variante Mikro30G auf den Schutz auswirkt. Hier hat sich gezeigt, dass selbst bei einer aus Nutzersicht inakzeptablen Gruppengröße von $k = 200$ das Risiko nur geringfügig sank und weit oberhalb der mit den mehrdimensionalen Varianten verbundenen Risiken lag. Hier sei nochmals darauf hingewiesen, dass in diesem Falle blockweise 200 Unternehmen dieselbe Ausprägung bezüglich eines Überschneidungsmerkmals hätten und damit der erforderliche Grad an Anonymität auf Kosten wissenschaftlicher Analysen, welche auf dieses Merkmal stützten, erreicht würde.

Im Projekt wurde schon frühzeitig die Möglichkeit eines Datenangreifers getestet, große Unternehmen mittels Einzelangriffen zu reidentifizieren. So wurden bei der KSE 19 Unternehmen in den geschützten Daten gesucht, die jeweils mindestens 1.000 Beschäftigte aufweisen, vier von ihnen hatten mehr als 5.000 Beschäftigte.[62]

Von den 15 zu reidentifizierenden Unternehmen der KSE mit 1.000 bis 4.999 Beschäftigten

61) Vergleich der Varianten (a) gegenüber (b).

62) Bei der Umsatzsteuerstatistik wurde in einer weiteren Simulation Einzelangriffe speziell auf 6 so genannte Marktführer durchgeführt.

konnten 9 einem Datensatz in der KSE richtig zugeordnet werden. 3 Unternehmen wurden nicht und 3 wurden falsch zugeordnet (vgl. Tabelle 32.3). Das größte nicht reidentifizierte Unternehmen hatte gut 2.300 Beschäftigte. Die drei Unternehmen dieser Größenklasse mit mehr als 2.300 Beschäftigten wurden alle reidentifiziert. Die höhere Trefferquote war begleitet von einem leichteren Zugang bei gleichzeitig geringerem Bedarf an Zusatzwissen.

Zu allen vier Unternehmen mit mehr als 5.000 Beschäftigten konnten eindeutige und richtige Zuordnungen zu einem Datensatz der KSE hergestellt werden (vgl. Tabelle 32.3). Nötig war zur Reidentifikation dieser Unternehmen lediglich die vierstellige Wirtschaftszweigklassifikation und ungefähre Angaben über Umsatz und Beschäftigung. Auch bei vergröberter Wirtschaftszweigklassifikation sind diese Unternehmen für einen Datenangreifer relativ leicht zu reidentifizieren. Hierbei ist das Risiko für einen Datenangreifer, eine falsche Zuordnung zu erzielen, gering (siehe Tabelle 32.4). Mit anderen Worten, er kann sehr gut abschätzen, inwieweit seine gemachte Zuordnung korrekt bzw. falsch ist.

Tabelle 32.3: Gesamtergebnis der Reidentifikationsversuche

	alle Unternehmen		große Unternehmen[1]		sehr große Unternehmen[2]	
	Anzahl	Anteil (%)	Anzahl	Anteil (%)	Anzahl	Anteil (%)
Gesuchte Unternehmen	19,0	100,0	15,0	100,0	4,0	100,0
Eindeutige Zuordnungen	16,0	84,2	12,0	80,0	4,0	100,0
Richtige Zuordnungen	13,0	68,4	9,0	60,0	4,0	100,0
Falsche Zuordnungen	3,0	15,7	3,0	20,0	0,0	0,0
keine Zuordnungen	3,0	15,7	3,0	20,0	0,0	0,0
nicht reidentifiziert[3]	6,0	31,5	6,0	40,0	0,0	0,0

1) 1.000-4.999 Beschäftigte 2) ab 5.000 Beschäftigte

3) Als nicht reidentifiziert gelten alle Unternehmen, die falsch zugeordnet wurden oder bei denen keine eindeutige Zuordnung gelang.

Wie schwierig es ist, die Zuordnungsquoten bei den großen Unternehmen so zu senken, dass eine faktische Anonymisierung als gelungen anzusehen ist, zeigt ein Blick auf Tabelle 32.5. In dieser sind die Trefferquoten im Rahmen des erwähnten Szenarios für unterschiedliche Anonymisierungsverfahren enthalten. Die Zuordnungsquoten konnten nur mit solchen Anonymisierungsmaßnahmen reduziert werden, die aufgrund ihrer sehr starken Veränderung und der damit verbundenen Reduzierung des Analysepotenzials im Projektverlauf abgelehnt werden mussten.

Tabelle 32.4: Falschzuordnungsquoten nach Unternehmensgröße

	Falschzuordnungsquoten[1] in %
Alle Unternehmen	18,8
Große Unternehmen	25,0
Sehr große Unternehmen	0,0

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

Die von der Konzeption her mit den zuvor betrachteten Varianten MA10G und MA30G vergleichbaren Varianten MA11G und MA33G sind dagegen nicht in der Lage, die Zuordnungen im ausreichenden Maße zu reduzieren. Zur Erreichung faktischer Anonymität unter Verwendung dieser Anonymisierungsverfahren ist es daher notwendig, den Nutzen der durch die Zuordnung enthüllten Informationen bzw. Merkmale soweit zu reduzieren, dass über diesen Weg die faktische Anonymität gewährleistet ist. Dies kann nur unter Einbeziehung sämtlicher Merkmale in die Anonymisierung geschehen.

Noch eindeutiger ist das Ergebnis beim Angriff auf die Marktführer bei der Umsatzsteuerstatistik ausgefallen. Hier konnten alle sechs gesuchten Unternehmen aus fünf unterschiedlichen Branchen reidentifiziert werden (aus einer Branche wurden die zwei größten Unternehmen gesucht). Hierzu waren lediglich die Kenntnis der Branche und das Wissen um die Marktführerschaft zur Reidentifikation notwendig. Dies kann nur durch sehr weitreichende Anonymisierungsmaßnahmen, wie Entfernung der Wirtschaftszweigklassifikation, verhindert werden. Da dies aus Analysesicht nicht gewollt sein kann, muss auch in diesem Fall der Nutzen der enthüllten Informationen durch andere Maßnahmen so weit gesenkt werden, dass die faktische Anonymität gewährleistet wird.

Die beiden Beispiele machen deutlich, dass eine Reduzierung der Anonymisierung auf die Überschneidungsmerkmale bei großen Unternehmen nur unter erheblicher Einschränkung des Analysepotenzials möglich erscheint. Wird der Nutzen für den Datenangreifer durch eine Reidentifikation mit in die Betrachtung gezogen, können die Überschneidungsmerkmale deutlich schwächer anonymisiert werden; dies allerdings nur unter der Bedingung, dass alle Zielmerkmale ebenfalls mit anonymisiert werden.

Tabelle 32.5: Vergleich der Anzahl an Reidentifikationen

Daten	Reidentifikationen		
	insgesamt	groß[1]	sehr groß[2]
Formale Anonymisierung	19	9	4
MA33G	19	9	4
MA33G V1*	14	6	4
MA33G V2*	12	5	3
MA33G PRAM20*	13	6	3
MA11G	19	9	4
MA11G V1*	13	5	4
MA11G V2*	12	5	2
MA8G*	16	8	2
MA8G V1*	10	5	2
MA8G V2*	9	4	1
MA1G*	10	8	1
MA1G V1*	5	4	1
MA1G V2*	4	3	0
SAFE2A*	15	7	4
RSWP 1%*	7	2	1
LHS1*	6	5	1

1) 1.000-4.999 Beschäftigte

2) wenigstens 5.000 Beschäftigte

* Diese Verfahren wurden mangels Analysepotenzial bereits während der Projektlaufzeit verworfen und sollen daher an dieser Stelle nicht näher spezifiziert werden. Eine detaillierte Beschreibung findet sich in Statistische Ämter des Bundes und der Länder und IAW (2003).

Kapitel 33

Implikationen einer Beschränkung der Anonymisierung auf die Überschneidungsmerkmale für das Analysepotenzial

33.1 Allgemeine Vorbemerkungen

Aus der Sicht des Analysepotenzials bietet der Vorschlag, nur die Überschneidungsmerkmale mit Anonymisierungsverfahren zu behandeln, natürlich den Vorteil, dass Analysen, in die nur nicht behandelte Merkmale einbezogen werden, durch die Anonymisierung in keiner Weise tangiert sind. Hier ist also in jedem Fall eine Verbesserung aus Sicht des Analysepotenzials zu konstatieren, wenn nur die Überschneidungsmerkmale anonymisiert werden. Allerdings handelt es sich bei den metrischen Überschneidungsmerkmalen im engeren Sinne („Beschäftigte", „Umsatz") ebenso um für Analysen wesentliche Merkmale, wie auch bei den Überschneidungsmerkmalen im weiteren Sinne (bei der KSE: „FuE-Aufwendungen" und „Abschreibungen", weniger „Einsatz an Handelsware"). Somit muss die Einbeziehung gerade dieser Überschneidungsmerkmale in Analysen beachtet werden. Deshalb wäre es – vor allem vor dem Hintergrund, dass die Beschränkung der Anonymisierung auf die Überschneidungsmerkmale „stärkere" Anonymisierungsmaßnahmen am einzelnen Merkmal erfordert, – zu einfach, diese Beschränkung aus Sicht des Analysepotenzials generell für besser zu halten.

Weitere Überschneidungsmerkmale sind diskrete beziehungsweise kategoriale Merkmale, wie Rechtsform, Wirtschaftszweig oder Regionalangabe. Die Bewertung einer Anonymisierung dieser Merkmale durch Informationseinschränkungen hängt von den Nutzerinteressen ab. In der Regel wird man um eine Informationseinschränkung bei diesen Variablen nicht herumkommen. Das konkrete Vorgehen ist dabei, wie bereits in den Teilen IV und V beschrieben, in einem diskursiven Prozess zwischen Datenanbietern und Datennutzern zu klären. Die folgenden Abschnitte konzentrieren sich deshalb auf die Überlegungen, die Anonymisierungsmaßnahmen bei den metrischen Variablen auf die Überschneidungsmerkmale zu beschränken.

33.2 Beurteilungsschema

Das in Abbildung 33.1 dargestellte Schema soll bei der Beurteilung des Vorschlags einer Beschränkung der Anonymisierungsmaßnahmen auf die Überschneidungsmerkmale aus der Sicht des Analysepotenzials helfen. Dabei soll grundsätzlich zwischen univariaten und multivariaten Auswertungen unterschieden werden.

Wird eine univariate Auswertung (inklusive Tabellenauswertungen nach kategorialen Merkmalen) vorgenommen, so ist die betreffende Variable entweder von der Anonymisierungsmaßnahme betroffen, weil sie zu den Überschneidungsmerkmalen gehört, oder nicht. Ist sie von der Anonymisierungsmaßnahme nicht betroffen, so ergibt sich durch die Anonymisierung keine Beeinträchtigung und somit eine Verbesserung des Analysepotenzials gegenüber dem Fall, dass alle Variablen anonymisiert werden.

Handelt es sich bei dem auszuwertenden Merkmal hingegen um ein Überschneidungsmerkmal, und ist es somit von der Anonymisierung tangiert, so hängt die Bewertung davon ab, ob bei einer Beschränkung der Anonymisierung auf die Überschneidungsmerkmale „stärkere" Maßnahmen ergriffen werden müssen, die auch zu stärkeren Beeinträchtigungen des Analysepotenzials führen. Daneben ist für die Bewertung jedoch auch noch relevant, ob durch die Beschränkung auf die Überschneidungsmerkmale möglicherweise solche Merkmale von der Anonymisierung ausgenommen werden können, die für Analysen wichtiger sind oder bei denen die Anonymisierungsmaßnahme stärkere Beeinträchtigungen von Analyseergebnissen hervorruft.

Damit stellen sich für univariate Auswertungen drei Fragen:

1. Ist die durch die Beschränkung auf die Überschneidungsmerkmale erforderliche stärkere Anonymisierung bei univariaten Auswertungen mit nicht hinnehmbaren Beeinträchtigungen bei Auswertungen mit den von der Anonymisierung betroffenen Merkmalen verbunden?
2. Können durch die Beschränkung der Anonymisierung auf die Überschneidungsmerkmale für die Analyse besonders wichtige Merkmale von der Anonymisierung ausgenommen werden?
3. Können durch die Beschränkung auf die Überschneidungsmerkmale solche Merkmale von der Anonymisierung ausgenommen werden, bei denen univariate Auswertungen durch die Anonymisierungsmaßnahme besonders beeinträchtigt werden?

Werden bei multivariaten Auswertungen alle für die Analyse benötigten Merkmale nicht anonymisiert, so ergibt sich in jedem Fall eine Verbesserung des Analysepotenzials durch die Beschränkung der Anonymisierung auf die Überschneidungsmerkmale. Handelt es sich bei allen in die Analyse einbezogenen Merkmalen um Überschneidungsmerkmale, die von der Anonymisierung direkt tangiert sind, so ist wiederum zu unterscheiden, ob die gleiche

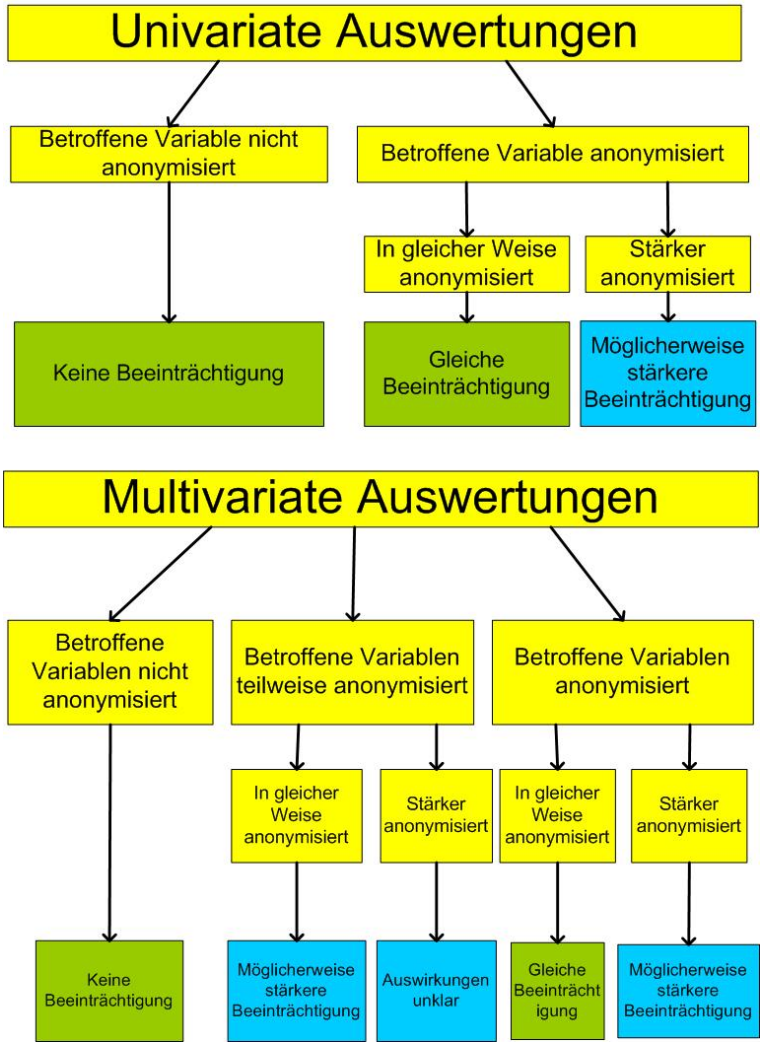

Abbildung 33.1: Schema uni- und multivariate Auswertungen

Maßnahme angewendet wird wie bei der Anonymisierung aller Merkmale oder ob bei der Beschränkung auf die Überschneidungsmerkmale „stärkere“ Anonymisierungsmaßnahmen erforderlich sind. Bei Anwendung der gleichen Anonymisierungsmaßnahme ergibt sich auch die gleiche Beeinträchtigung der Analyseergebnisse wie bei der Anonymisierung aller Merkmale. Sind hingegen „stärkere“ Maßnahmen erforderlich, so kann es auch zu einer stärkeren Beeinträchtigung der Analyseergebnisse kommen. Dies ist zu untersuchen.

Bei multivariaten Auswertungen existiert jedoch noch ein dritter Fall, bei dem die in die Analyse eingehenden Merkmale teilweise in die Anonymisierung einbezogen werden, weil es sich nur bei einem Teil dieser Merkmale um Schlüsselmerkmale handelt. In diesem Fall sind die Auswirkungen auf das Analysepotenzial auf den ersten Blick völlig unklar. Möglicherweise führt die Behandlung mit den gleichen Maßnahmen wie im Fall der Anonymisierung aller Merkmale sogar zu einer schwächeren Beeinträchtigung des Analysepotenzials. Aber auch die Wirkung „stärkerer“ Maßnahmen kann in diesem Fall nicht eindeutig beurteilt werden. Beide Fragen sind daher zu untersuchen.

Es bleiben somit für multivariate Auswertungen folgende drei Fragen:

1. Führen „stärkere“ Anonymisierungsmaßnahmen für den Fall, dass alle in die Analyse einbezogenen Merkmale anonymisiert werden, zu einer stärkeren Beeinträchtigung des Analysepotenzials?
2. Führt die gleiche Anonymisierungsmaßnahme (wie bei der Anonymisierung aller Merkmale) für den Fall, dass ein Teil der Merkmale in die Anonymisierung einbezogen wird, zu einer Verbesserung der Analyseergebnisse?
3. Führen „stärkere“ Anonymisierungsmaßnahmen für den Fall, dass ein Teil der in die Analyse einbezogenen Merkmale anonymisiert wird, zu einer stärkeren Beeinträchtigung des Analysepotenzials?

Eine Beschränkung der Anonymisierung auf metrische Überschneidungsmerkmale bei Simulationsverfahren ist nicht sinnvoll, da die Idee von Simulationsverfahren gerade darin besteht, komplett neue Merkmalsträger zu generieren. Die Beschränkung könnte jedoch insbesondere bei stochastischen Überlagerungen und Mikroaggregationsverfahren sinnvoll sein. Es wird daher versucht, die offenen Fragen für diese beiden Verfahrensgruppen zu beantworten und damit zu einer Beurteilung zu gelangen.

Die Verstärkung einer Anonymisierungsmaßnahme aus Sicht der Datensicherheit besteht bei den Mikroaggregationsverfahren zum einen im Übergang von der getrennten Mikroaggregation zur gruppierten beziehungsweise zur gemeinsamen Mikroaggregation, zum anderen in einer Erhöhung der Gruppengröße. Bei stochastischen Überlagerungen wird eine Verstärkung durch eine Erhöhung der Varianz erreicht.

33.3 Wirkung einer „stärkeren“ Anonymisierung bei univariaten Auswertungen

33.3.1 Wirkung einer „stärkeren“ Mikroaggregation

Wesentliche univariate Auswertungen sind die Berechnung von Mittelwerten, Streuungsmaßen und Konzentrationsmaßen. Die Mittelwerte im Gesamtdatensatz werden bei der

Mikroaggregation grundsätzlich erhalten, Streuungsmaße ebenso wie Konzentrationsmaße werden reduziert. Dabei ist die Verzerrung umso größer, je größer die Gruppengröße und je unkorrelierter gemeinsam mikroaggregierte Variablen sind. Somit bedeutet eine Erhöhung der Gruppengröße oder der Wechsel von der getrennten zur gemeinsamen oder gruppierten Mikroaggregation diesbezüglich eine Verschlechterung des Analysepotenzials. Zusätzlich werden auch die Mittelwerte von Teilgesamtheiten verzerrt, wenn die Mikroaggregation nicht für die einzelnen Teilgesamtheiten getrennt angewendet wird.

Betrachtet man die Auswirkungen unterschiedlicher Varianten der Mikroaggregation auf die Mittelwerte und Standardabweichungen der Überschneidungsmerkmale im weiteren Sinne für die KSE nach Wirtschaftszweigen (Tabelle 33.1), so erkennt man, dass die Abweichung der Mittelwerte bei den in Kapitel 18 festgelegten Abweichungsschwellen von 10 Prozent bei der getrennten Mikroaggregation in dem vorgegebenen Rahmen liegt,[63] während sie für die gruppierte sowie für die gemeinsame Mikroaggregation teilweise sehr deutlich überschritten wird. Wird bei der getrennten Mikroaggregation die Gruppengröße von drei bis fünf auf generell fünf beziehungsweise sieben erhöht, so werden ebenfalls die Abweichungsschwellen häufiger überschritten. Ähnliches gilt für die Standardabweichungen.

Tabelle 33.1: Anteil der Überschreitungen der Abweichungsschwellen bei den Überschneidungsmerkmalen der KSE im weiteren Sinne bei Mikroaggregationsverfahren

Anteil der Zellen mit einer Überschreitung der Abweichungsschwelle von 10%		
	bei den arithmetischen Mitteln	bei den Standardabweichungen
Getrennte abstandsorientierte Mikroaggregation (Gruppengröße 3 bis 5)	2,75%	10,99%
Getrennte abstandsorientierte Mikroaggregation (Gruppengröße 5)	10,99%	19,23%
Getrennte abstandsorientierte Mikroaggregation (Gruppengröße 7)	10,44%	17,58%
Gruppierte abstandsorientierte Mikroaggregation (11 Gruppen)	17,03%	26,37%
Gruppierte abstandsorientierte Mikroaggregation (8 Gruppen)	12,64%	27,47%
Gemeinsame abstandsorientierte Mikroaggregation	37,36%	69,23%

63) Höchstens 10 Prozent der anonymisierten Werte weichen um mehr als 10 Prozent relativ von ihrem Original ab

33.3.2 Wirkung einer „stärkeren“ stochastischen Überlagerungen

Auch bei stochastischen Überlagerungen gilt, dass die arithmetischen Mittel im Gesamtdatensatz unabhängig von der Varianz der Überlagerung erhalten bleiben. Allerdings werden die Streuungsmaße entsprechend stärker verzerrt. Mit Hilfe einer im Anschluss an die eigentliche Überlagerung durchführbaren Transformation kann jedoch auch die Varianz im Gesamtdatensatz erhalten bleiben, jedoch wirkt sich dies negativ auf den Erhalt der Verteilungsmaße in Teilgesamtheiten aus.

Da sich aus der Sicht der Datensicherheit vor allem multiplikative stochastische Überlagerungen zur Erstellung eines Scientific-Use-Files anbieten und diese auch aus Analysesicht ein paar wesentliche Vorteile gegenüber der additiven Überlagerung aufweisen, wird hier für die Daten der KSE gezeigt, wie sich die Erhöhung der Varianz einer multiplikativen Überlagerung aus einer zweigipfligen Mischungsverteilung auf die Überschreitungshäufigkeiten der Abweichungsschwellen bei arithmetischen Mitteln und Standardabweichungen nach Wirtschaftszweigen (Zweisteller) auswirkt.

Hierzu werden zwei verschiedene Mischungsverteilungen verglichen. Bei der ersten beträgt die Verschiebung der beiden Mittelwerte 8 Prozent und die Standardabweichung der einzelnen Komponenten jeweils 1,8 Prozent. Bei der zweiten liegt die Verschiebung bei 11 Prozent und die Standardabweichung bei 3 Prozent. Die Ergebnisse sind in Tabelle 33.2 dargestellt.

Tabelle 33.2: Anteil der Überschreitungen der Abweichungsschwellen bei den Überschneidungsmerkmalen der KSE im weiteren Sinne bei multiplikativen stochastischen Überlagerungen

Anteil der Zellen mit einer Überschreitung der Abweichungsschwelle von 10%		
	bei den arithmetischen Mitteln	bei den Standardabweichungen
Multiplikative stochastische Überlagerung mit einer Mischungsverteilung (f=8%, s=1,8%)	0%	0%
Multiplikative stochastische Überlagerung mit einer Mischungsverteilung (f=11%, s=3%)	4,57%	19,43%

Man erkennt, dass die Erhöhung der Varianz zu einem deutlichen Anstieg der Überschreitungsanteile für die Abweichungsschwellen führt. Der Anstieg fällt bei den Standardabweichungen deutlich höher aus als bei den arithmetischen Mitteln. Nur bei den Standardabweichungen wird die Toleranzgrenze von 10 Prozent Überschreitungen nicht eingehalten. Somit wird klar, dass auch bei den stochastischen Überlagerungen mit zunehmender Varianz sehr schnell Probleme hinsichtlich der Einhaltung der definierten Abweichungsschwellen

auftreten können. Dabei stehen gerade bei der Mischungsverteilung mit der Festlegung der Mittelwerte und der Standardabweichungen der Komponenten zwei Instrumente zur Feinsteuerung zur Verfügung.

33.4 Bedeutung von Überschneidungsmerkmalen und anderen Merkmalen für die Analyse

Letztlich ist damit die Frage verbunden, ob die Überschneidungsmerkmale oder die restlichen Merkmale für die Analysen wichtiger sind. Diese Frage ist jedoch objektiv nicht beantwortbar. Es ist aber nochmals darauf hinzuweisen, dass die als Überschneidungsmerkmale ausgemachten Variablen, wie *Umsatz*, *Beschäftigte*, *FuE-Aufwand* oder *Abschreibungen* auch für Analysen wichtige Merkmale sind. Deshalb sollte eine abschließende Bewertung dieser Frage von der jeweiligen Konstellation abhängen: Ist eine Anonymisierung bei Bewahrung eines ausreichenden Analysepotenzials für alle Variablen durch die Anwendung von datenverändernden Anonymisierungsmaßnahmen auf alle metrischen Variablen möglich und besteht die Alternative darin, dass bei der Beschränkung der Anonymisierungsmaßnahmen auf die Überschneidungsmerkmale, diese so anonymisiert werden müssten, dass sie für Analysen unbrauchbar sind, so sollte die Anonymisierung auf alle Merkmale angewendet werden. Würden jedoch die zur Sicherstellung der faktischen Anonymität erforderlichen Anonymisierungsmaßnahmen bei allen Merkmalen zu einer zu starken Einschränkung des Analysepotenzials führen, so sollten gegebenfalls stärkere Anonymisierungsmaßnahmen angewendet werden, die sich ausschließlich auf die Überschneidungsmerkmale erstrecken und trotzdem die faktische Anonymität sicherstellen. Dann wären die Überschneidungsmerkmale zwar noch weniger brauchbar, die anderen Merkmale jedoch unbeeinträchtigt.

33.5 Unterschiedliche Wirkung von Anonymisierungsmaßnahmen auf Überschneidungsmerkmale und andere Merkmale

Möglicherweise wirken sich Anonymisierungsverfahren bei bestimmten Merkmalen stärker negativ auf Analyseergebnisse aus als bei anderen. Die Frage wurde am Beispiel der Kostenstrukturerhebung für kleinere und mittlere Unternehmen untersucht. Beispielhaft ist in Tabelle 33.3 dargestellt, welche maximalen relativen Abweichungen der arithmetischen Mittel von 30 Merkmalen der KSE in einzelnen Zellen auftreten können, wenn die Merkmale mit getrennter Mikroaggregation anonymisiert werden. Die Auswertungen erfolgte nach Wirtschaftszweigen, Beschäftigtengrößenklassen sowie nach Ost/West. Tendenziell weisen solche Merkmale stärkere Abweichungen auf, die viele (strukturelle) Nullen aufweisen. Die Merkmale, bei denen eine Überschreitung der Abweichungsschwelle festgestellt wurde, sind kursiv, die Überschneidungsmerkmale im weiteren Sinne fett gedruckt.

Man erkennt, dass in der Tat durch eine Beschränkung der Anonymisierung auf die Über-

Tabelle 33.3: Maximale relative Abweichung der arithmetischen Mittel für die einzelnen Variablen in den kleinstmöglichen Zellen (nach Wirtschaftszweigen, Beschäftigtengrößenklassen, Ost/West)

Variable	Größte relative Abweichung in %	Zweitgrößte relative Abweichung in %
Teilzeitbeschäftigte	0,5	0,5
Teilzeitbeschäftigte in Vollzeiteinheiten	*21,5*	*18,5*
Tätige Personen	**0,1**	**0,1**
Umsatz aus eigenen Erzeugnissen	12,4	1,3
Umsatz aus Handelsware	*28,3*	*6,0*
Gesamtumsatz	**12,8**	**1,3**
Bruttoproduktionswert	13,3	1,3
AB Fertige und unfertige Erzeugnisse	2,2	1,1
EB Fertige und unfertige Erzeugnisse	3,7	3,5
AB Roh-, Hilfs- und Betriebsstoffe	*23,8*	*2,7*
EB Roh-, Hilfs- und Betriebsstoffe	11,9	3,0
Verbrauch Rohstoffe	1,3	0,6
Energieverbrauch	*45,2*	*12,5*
AB Handelsware	*40,1*	*22,5*
EB Handelsware	*33,8*	*32,2*
Einsatz Handelsware	**36,8**	**7,2**
Bruttolohn- und Gehaltssumme	0,2	0,2
Gesetzliche Sozialkosten	0,1	0,1
Sonstige Sozialkosten	6,9	4,7
Kosten für Leiharbeit	3,6	2,4
Kosten für Lohnarbeit	*29,0*	*5,0*
Kosten für Reparaturen	*90,1*	*73,9*
Mieten und Pachten	12,6	7,2
Sonstige Kosten	3,9	3,7
Fremdkapitalzinsen	2,1	1,8
Kosten insgesamt	2,1	0,9
Bruttowertschöpfung	**1,8**	**1,2**
Nettowertschöpfung zu Faktorkosten	**4,2**	**0,8**
FuE-Aufwendungen	**3,1**	**2,7**
FuE-Beschäftigte	**10,8**	**1,9**

schneidungsmerkmale größere Abweichungen vermieden werden können, weil von starken Überschreitungen der Abweichungsschwelle von 10 Prozent von den Überschneidungsmerkmalen im weiteren Sinne nur der Einsatz an Handelsware betroffen ist. Legt man lediglich die Überschneidungsmerkmale im engeren Sinne zugrunde, so kann in diesem Fall bei kei-

nem Überschneidungsmerkmal eine Überschreitung der Abweichungsschwelle festgestellt werden.

Wichtig ist darauf hinzuweisen, dass bei dieser Form der Anonymisierung der Anteil an Überschreitung der Abweichungsschwellen innerhalb der vorgegebenen Toleranzgrenze liegt.

Allerdings sind aus diesen Ergebnissen keine generellen Schlussfolgerungen möglich. Vielmehr muss die Frage, ob das Analysepotenzial bestimmter Merkmale von einer Anonymisierungsmaßnahme stärker betroffen ist als das Analysepotenzial anderer Merkmale und zu welcher Gruppe gegebenenfalls die Überschneidungsmerkmale gehören, bei jedem Anonymisierungsverfahren und jedem Datenbestand neu geprüft werden.

33.6 Wirkung einer „stärkeren" Anonymisierung aller in multivariate Analysen einbezogenen Merkmale

33.6.1 Mikroaggregationsverfahren

In Abschnitt 19.2 wurde bereits gezeigt, dass die gruppierte Mikroaggregation und insbesondere die gemeinsame abstandsorientierte Mikroaggregation zu einer stärkeren Verzerrung der Korrelationskoeffizienten führt als die getrennte abstandsorientierte Mikroaggregation.

In Kapitel 23 wurden die Auswirkungen von Mikroaggregationsverfahren auf die Schätzung linearer und nichtlinearer Modelle ausführlich hergeleitet. Sofern die Verstärkung der Anonymisierung einen Übergang von der getrennten abstandsorientierten Mikroaggregation zur teilweise gemeinsamen beziehungsweise gruppierten Mikroaggregation (ebenfalls abstandsorientiert) vorsieht, ist mit einer Verschlechterung der Schätzergebnisse zu rechnen. Erfolgt der Übergang zur gemeinsamen abstandsorientierten Mikroaggregation, so ist im linearen Modell ebenfalls nur mit einer geringen Verzerrung zu rechnen, die ausschließlich deshalb zustande kommt, weil die Gruppenbildung bei der Mikroaggregation auch von der abhängigen Variablen mitbestimmt wird. Allerdings ergeben sich beim Vorliegen nichtlinearer Transformationen und in nichtlinearen Modellen deutlich stärkere Verzerrungen. Dies waren auch die Gründe, warum sowohl gruppierte als auch gemeinsame Mikroaggregation als Anonymisierungsverfahren für Scientific-Use-Files ausgeschlossen wurden.

Eine Erhöhung der Gruppengröße führt in jedem Fall zu einem höheren Effizienzverlust. Die bei der getrennten abstandsorientierten Mikroaggregation kaum sichtbare Verzerrung nimmt mit zunehmender Gruppengröße zu. Welche Gruppengröße gerade noch akzeptabel ist, hängt auch vom untersuchten Zusammenhang ab.

33.6.2 Stochastische Überlagerungen

In Abschnitt 19.2 wurde bereits gezeigt, dass die Verzerrung der Korrelationskoeffizienten umso höher ist, je größer die Varianz der Überlagerungen gewählt wird.

In Kapitel 22 wurden die Auswirkungen von stochastischen Überlagerungen auf die Schätzung linearer und nichtlinearer Modelle ausführlich dargestellt. Grundsätzlich gelten die Herleitungen zur asymptotischen Verzerrung der Schätzer und ihrer Korrigierbarkeit unabhängig von der Varianz der Überlagerungen. Es hat sich jedoch in Simulationsexperimenten gezeigt, dass in der Praxis bei multiplikativen Überlagerungen die Funktionsfähigkeit von Korrekturverfahren mit höherer Varianz nachlassen kann. Aus der Sicht multivariater Auswertungen muss dieses Problem vermieden werden. Zudem ergibt sich mit wachsender Varianz der Überlagerungen ein höherer Effizienzverlust.

33.7 Wirkung der Anonymisierung eines Teils der in multivariate Analysen einbezogenen Merkmale

33.7.1 Mikroaggregationsverfahren

Zur Untersuchung der Wirkung einer Beschränkung der Mikroaggregation auf die Überschneidungsmerkmale in multivariaten Analysen wird beispielhaft mit den Daten der Kostenstrukturerhebung für kleinere und mittlere Unternehmen die bereits in vorangegangenen Kapiteln verwendete linearisierte Cobb-Douglas-Produktionsfunktion geschätzt. Dabei werden die Daten der kleineren und mittleren Unternehmen der KSE sowohl mit der getrennten abstandsorientierten Mikroaggregation als auch mit der im vorangegangenen Kapitel vorgestellten Variante der gruppierten Mikroaggregation mit zehn Gruppen anonymisiert.

Es werden drei Fälle unterschieden: Im ersten Fall werden alle Variablen anonymisiert, im zweiten Fall nur die Überschneidungsmerkmale der weiteren Definition und im dritten Fall sogar nur die Überschneidungsmerkmale der engeren Definition. Tabelle 33.4 kann entnommen werden, dass die enge Definition der Überschneidungsvariablen nur den Output betrifft, während bei der weiten Definition der Überschneidungsvariablen zusätzlich die Inputvariablen Materialeinsatz und Kapitalkosten betroffen sind.

Tabelle 33.5 zeigt die Höhe der Abweichung der Koeffizientenschätzer für die um Ausreißer bereinigten sowie die unbereinigten Daten und die drei Varianten der beiden Mikroaggregationsverfahren. Dabei ist zu beachten, dass bei dieser Auswertung zusätzlich eine 80-Prozent-Stichprobe gezogen wurde, die zu zusätzlichen Verzerrungen führt. Allerdings lassen sich trotzdem Tendenzaussagen zur Wirkung der Anzahl der einbezogenen Variablen machen. Die durchschnittliche Höhe der Verzerrung der Schätzer nimmt mit der Anzahl der von der Anonymisierung betroffenen Variablen ab. Dies gilt nicht für die getrennte abstandsorientierte Mikroaggregation bei Verwendung der bereinigten Daten. Hier bleibt die

durchschnittliche Abweichung, die ohnehin sehr gering ist, etwa konstant.

Wird nur ein Teil der Variablen mikroaggregiert, so können sich jedoch unter Umständen auch schlechtere Ergebnisse bei der Schätzung linearer multivariater Modelle ergeben. Dies ist dann der Fall, wenn entweder eine gemeinsame stochastische Mikroaggregation oder eine abstandsorientierte Mikroaggregation durchgeführt wird, bei der die Gruppenbildung nicht von der abhängigen Variablen beeinflusst ist. In diesen beiden Fällen ergeben sich, wie in Kapitel 23 gezeigt wurde, erwartungstreue Schätzer in linearen Modellen, sofern entweder alle Variablen oder zumindest alle erklärenden Variablen (gemeinsam) mikroaggreggiert werden. Ist dies nicht der Fall, so sind die Schätzer verzerrt.

33.7.2 Stochastische Überlagerungen

Wird ausschließlich die abhängige Variable überlagert, so tritt keine Verzerrung auf. Es tritt allerdings eine Verzerrung auf, wenn mindestens ein Regressor überlagert ist. Die Korrektur der Verzerrung ist von der Anzahl der überlagerten Variablen unabhängig. Da man bei der

Tabelle 33.4: Modellvariablen und ihre Behandlung durch die Mikroaggregation

Defintion der Überschneidungsmerkmale	Eng	Weit
Output		
Bruttoproduktionswert	Nein	Nein
Gesamtumsatz	*Ja*	*Ja*
Umsatz aus eigenen Erzeugnissen	Nein	Nein
Umsatz aus Handelsware	Nein	Nein
Materialeinsatz		
Verbrauch an Rohstoffen	Nein	Nein
Einsatz an Handelsware	*Nein*	*Ja*
Energieverbrauch	Nein	Nein
Personalkosten		
Bruttolohn- und Gehaltssumme	Nein	Nein
Gesetzliche Sozialkosten	Nein	Nein
Kosten für Leiharbeit	Nein	Nein
Externe Dienstleistungen		
Kosten für Reparaturen	Nein	Nein
Kosten für Lohnarbeiten	Nein	Nein
Sonstige Kosten	Nein	Nein
Kapitalkosten		
Abschreibungen	*Nein*	*Ja*
Mieten und Pachten	Nein	Nein

Tabelle 33.5: Abweichung der Parameterschätzer, linearisierte Cobb-Douglas-Produktionsfunktion, KSE-Daten für KMU, unterschiedlich viele mikroaggregierte Merkmale

Durchschnittliche relative Abweichung der Parameterschätzer in %		
	Daten unbereinigt	Daten bereingt
Getrennte abstandsorientierte Mikroaggregation, alle Variablen anonymisiert	2,2	1
Getrennte abstandsorientierte Mikroaggregation, nur Überschneidungsvariablen anonymisiert (weit)	1,7	1
Getrennte abstandsorientierte Mikroaggregation, nur Überschneidungsvariablen anonymisiert (eng)	1,1	0,9
Gruppierte abstandsorientierte Mikroaggregation, alle Variablen anonymisiert	15,3	7,1
Gruppierte abstandsorientierte Mikroaggregation, nur Überschneidungsvariablen anonymisiert (weit)	10,6	5,3
Gruppierte abstandsorientierte Mikroaggregation, nur Überschneidungsvariablen anonymisiert (eng)	0,7	0,8

Erstellung eines Scientific-Use-Files in der Regel nicht weiß, ob eine Variable später als abhängige oder als Regressorvariable eingesetzt wird, ist bei multivariaten Auswertungen eine geringere Anzahl von überlagerten Variablen nicht unbedingt von Vorteil.

Ein Nachteil durch die Beschränkung auf die Überschneidungsmerkmale kann bei einer Kette mehrerer hintereinander erfolgender Transformationen entstehen, wie sie im Anwendungsbeispiel der Cobb-Douglas-Produktionsfunktion auftritt. Werden zunächst zwei Merkmale subtrahiert und anschließend logarithmiert, so hängt die Frage, ob der Logarithmus definiert ist, davon ab, ob sich bei der Differenzenbildung ein positiver oder ein negativer Wert ergibt. Es wurde im Zusammenhang mit diesem Beispiel argumentiert, dass multiplikative Überlagerungen mit (annähernd) gleichen Überlagerungsfaktoren sinnvoll sind, um Vorzeichenwechsel bei Differenzenbildungen zu vermeiden. Wird nun aber auf die Überlagerung einer der beiden an der Differenzenbildung beteiligten Merkmale verzichtet, so tritt das Problem wieder auf.

33.8 Wirkung einer „stärkeren“ Anonymisierung eines Teils der in multivariate Analysen einbezogenen Merkmale

33.8.1 Mikroaggregationsverfahren

Für die Mikroaggregationsverfahren kann eine „stärkere“ Anonymisierung sowohl die Erhöhung der Gruppengröße als auch den Übergang zu einer Variante bedeuten, die zu einer stärkeren Veränderung der Einzelwerte führt. Hier wird nun alternativ die Vergrößerung der Gruppengröße bei einer getrennten abstandsorientierten Mikroaggregation und der Übergang von der getrennten zur teilweise gemeinsamen abstandsorientierten Mikroaggregtion auf der Basis von zehn Variablengruppen untersucht. Dabei werden diese Verfahren auf die Überschneidungsmerkmale der KSE für kleinere und mittlere Unternehmen angewendet. Anschließend werden die Ergebnisse der Schätzung der linearisierten Cobb-Douglas-Produktionsfunktion mit den Originalschätzern verglichen. Dabei ist zu beachten, dass durch eine Stichprobenziehung eine zusätzliche Verzerrung entsteht.

Die Ergebnisse sind in Tabelle 33.6 dargestellt. Eine geringfügig erhöhte Gruppengröße (auf 5 beziehungsweise 7) führt bei der engen Definition der Überschneidungsmerkmale zu geringeren (nicht bereinigte Daten) beziehungsweise zu in etwa gleichen (bereinigte Daten) Abweichungen der Koeffizientenschätzer im Vergleich zur getrennten Mikroaggregation aller Variablen mit Gruppengröße 3. Dahingegen ergeben sich bei der weiten Definition der Überschneidungsmerkmale größere (nicht bereinigte Daten) beziehungsweise in etwa gleiche (bereinigte Daten) Abweichungen durch die Beschränkung auf die Überschneidungsmerkmale bei gleichzeitiger Erhöhung der Gruppengröße.

Der Übergang zur gruppierten Mikroaggregation führt bei Anwendung auf den weiten Kreis der Überschneidungsmerkmale im Anwendungsbeispiel zu einer Verschlechterung der Ergebnisse. Demgegenüber sind bei der Anwendung der gruppierten Mikroaggregation auf den engen Kreis der Überschneidungsmerkmale im Vergleich zur Anwendung der getrennten Mikroaggregation auf alle Merkmale geringere durchschnittliche Abweichungen zu erkennen.

Somit kann die Frage, ob die Anwendung von „stärkeren“ Mikroaggregationsverfahren bei gleichzeitiger Beschränkung auf die Überschneidungsmerkmale in multivariaten Auswertungen, in denen zum Teil auch Überschneidungsmerkmale einbezogen werden, zu größeren oder kleineren Abweichungen der Analyseergebnisse führt, nicht eindeutig beantwortet werden, vielmehr hängt dies von der konkreten Auswertung, dem jeweiligen Anonymisierungsverfahren und der Anzahl der von der Anonymisierung betroffenen Merkmale ab. Diese Frage wird sich somit bei keinem Anonymisierungsvorhaben im vorhinein beantworten lassen, weil nicht alle möglichen multivariaten Auswertungen bei der Anonymisierung beachtet werden können.

Tabelle 33.6: Abweichung der Parameterschätzer, linearisierte Cobb-Douglas-Produktionsfunktion, KSE-Daten für KMU, für unterschiedliche Vorgehensweisen bei der Mikroaggregation

Durchschnittliche relative Abweichung der Parameterschätzer in %		
	Daten unbereinigt	Daten bereingt
Getrennte abstandsorientierte Mikroaggregation, alle Variablen anonymisiert, Gruppengröße 3	2,2	1
Getrennte abstandsorientierte Mikroaggregation, nur Überschneidungsmerkmale anonymisiert (eng), Gruppengröße 5	1,8	1,1
Getrennte abstandsorientierte Mikroaggregation, nur Überschneidungsmerkmale anonymisiert (eng), Gruppengröße 7	1,8	1
Getrennte abstandsorientierte Mikroaggregation, nur Überschneidungsmerkmale anonymisiert (weit), Gruppengröße 5	2,4	1,1
Getrennte abstandsorientierte Mikroaggregation, nur Überschneidungsmerkmale anonymisiert (weit), Gruppengröße 7	3,4	1
Gruppierte abstandsorientierte Mikroaggregation, nur Überschneidungsvariablen anonymisiert (weit)	10,6	5,3
Gruppierte abstandsorientierte Mikroaggregation, nur Überschneidungsvariablen anonymisiert (eng)	0,7	0,8

33.8.2 Stochastische Überlagerungen

Wird ausschließlich die abhängige Variable überlagert, so tritt keine Verzerrung auf, unabhängig davon, wie stark die Überlagerung ist. Da aber nicht davon auszugehen ist, dass die Überschneidungsmerkmale ausschließlich als abhängige Variablen verwendet werden, ist in jedem Fall mit einer Verzerrung zu rechnen. Dabei gilt wiederum, dass grundsätzlich die Korrektur von verzerrten Schätzern unabhängig von der Varianz der Überlagerungen möglich ist, allerdings im Fall multiplikativer Überlagerungen in der Praxis Probleme bei zu hohen Varianzen auftreten können. Außerdem ist eine „stärkere“ Überlagerung mit einem höheren Effizienzverlust verbunden.

Kapitel 34

Zusammenfassung zur Beschränkung der Anonymisierung auf die Überschneidungsmerkmale

34.1 Bewertung einzelner Anonymisierungsstrategien beschränkt auf die Überschneidungsmerkmale

Die in Abschnitt 31.2 aufgelisteten möglichen Strategien bei der Beschränkung der Anonymisierung auf die Überschneidungsmerkmale werden auf Basis der Ergebnisse bei der Verfahrensbewertung und der in den in den vorangegangenen Abschnitten präsentierten Überlegungen und Ergebnissen zur Beschränkung der Anonymisierung auf die Überschneidungsmerkmale wie folgt bewertet:

Zu Strategie 1: Eine moderate Erhöhung der Gruppengröße bei der getrennten Mikroaggregation führt zu bei den vorgenommenen univariaten Auswertungen zu geringfügig schlechteren Analyseergebnissen, bei den multivariaten Auswertungen war dies nicht zu erkennen. Allerdings ist mit der Maßnahme nicht die erforderliche höhere Schutzwirkung verbunden.

Zu Strategie 2: Die Verwendung der gruppierten Mikroaggregation, bei der immer ein Teil der (Überschneidungs-)Merkmale gemeinsam mikroaggregiert wird, führt bei der KSE sowohl zu vergleichsweise starken Veränderungen univariater deskriptiver Auswertungen (Dabei werden Abweichungsschwellen überschritten) als auch zu vergleichsweise starken Abweichungen von Analyseergebnissen (insbesondere beim getesteten Beispiel der Schätzung einer Cobb-Douglas-Produktionsfunktion), zumindest wenn eine weitere Definition der Überschneidungsmerkmale gewählt wird.

Die gemeinsame Mikroaggregation, bei der alle mikroaggregierten Merkmale (hier: Überschneidungsmerkmale) gemeinsam mikroaggregiert werden, führt unter der Voraussetzung, dass die Gruppenbildung zufällig erfolgt oder nur von den erklärenden Variablen abhängt, zwar in linearen Modellen zu erwartungstreuen Schätzern, allerdings nur dann, wenn entweder alle Variablen gemeinsam mikroaggregiert werden oder

die unabhängigen Merkmale gemeinsam mikroaggregiert wurden und die abhängige nicht.

Erfolgt die gemeinsame Mikroaggregation hingegen abstandsorientiert unter Einbeziehung der abhängigen Variablen bei der Gruppenbildung, so sind die Schätzer grundsätzlich verzerrt, allerdings haben Schmid et al. (2005) für den Fall, dass die Gruppenbildung ausschließlich nach der abhängigen erfolgt, hergeleitet, dass eine Korrektur der Schätzer möglich ist.

Da bei der gemeinsamen Mikroaggregation folglich am besten alle Variablen oder zumindest die erklärenden Variablen gemeinsam mikroaggregiert werden sollten, man aber bei einem Scientific-Use-File nicht von vornherein weiß, welche Variablen als abhängige und welche als unabhängige verwendet werden sollen, sollte Vorgehensweise 2. aus Sicht des Analysepotenzials ausgeschlossen werden.

Es kommt hinzu, dass die Erwartungstreue bzw. Konsistenz des Schätzers für die gemeinsame Mikroaggregation bei nichtlinearen Transformationen der Variablen in linearen Modellen bzw. in nichtlinearen Modellen nicht gegeben ist, die Überschneidungsmerkmale Umsatz und Beschäftigte aber häufiger zur Quotientenbildung und damit für nichtlineare Transformationen eingesetzt werden.

Tabelle 34.1: Gemeinsame Mikroaggregation und Erwartungstreue im linearen Modell

	Konstellation bei gemeinsamer Mikroaggregation beschränkt auf die Überschneidungsmerkmale (Gruppenbildung stochastisch oder abhängig von den Regressoren)	
K1	Alle Variablen gemeinsam mikroaggregiert	Erwartungstreu
K2	Abhängige Variable mikroaggregiert, Einflussgrößen nicht mikroaggregiert	Nicht erwartungstreu
K3	Nur erklärende Variablen mikroaggregiert	Erwartungstreu
K4	Nur Teil der unabhängigen Variablen oder Teil der unabhängigen Variablen und abhängige Variable gemeinsam mikroaggregiert	Nicht erwartungstreu

Zu Strategie 3: Informationsreduzierende Maßnahmen haben gegenüber datenverändernden den Vorteil, dass die Auswirkungen auf das Analysepotenzial besser abschätzbar sind. Wird etwa das kategoriale Merkmal *Wirtschaftszweigklassifikation* vergröbert, so kann es passieren, dass eine den Wissenschaftler interessierende Branche auf der nächst höheren Gliederungsebene mit einer anderen Branche zusammengefasst wird und die angestrebten branchenspezifische Analysen nun nicht mehr möglich sind. Daher sollte eine Diskussion mit der Wissenschaft der Entscheidung über informationsreduzierende Anonymisierungsmaßnahmen vorausgehen. Dies gilt insbesondere dann, wenn verschiedene kategoriale Überschneidungsmerkmale vorhanden sind. Steht man z.B. vor der Entscheidung, die Merkmale *Wirtschaftszweigklassifikation* und *Regionalkennung* zu vergröbern, so muss nach Kriterien des Erhaltes wissenschaftlicher Analysemöglichkeiten entschieden werden, ob der erforderliche Schutz mehr zu Lasten des einen oder des anderen Merkmals erreicht werden soll. Im allgemeinen Falle werden solche Maßnahmen zur Erreichung der faktischen Anonymität

nicht ausreichen und die Anwendung zusätzlicher datenverändernder Verfahren auf die metrischen (Überschneidungs)merkmale erforderlich sein. Ein Gegenbeispiel liefert allerdings die in Abschnitt 37.6 vorgestellte faktische Anonymisierung der kleinen und mittleren Unternehmen der Einzelhandelsstatistik des Jahres 1999.

Zu Strategie 4: Die Anwendung von Replacement oder Censoring auf die Schlüsselmerkmale besonders gefährdeter (großer) Unternehmen hat grundsätzlich den Vorteil, dass die Werte nur bei wenigen Unternehmen und Merkmalen überhaupt verändert werden. (Eine Ausnahme besteht unter Umständen, wenn andere Merkmale mit Überschneidungsmerkmalen stark korreliert sind, wie bei der Umsatzsteuerstatistik.)

Replacement hat den Vorteil eines Erhalts der Mittelwerte, Censoring erlaubt hingegen die Verwendung von gestutzten Tobit-Modellen. Nachteile bestehen bei beiden Verfahren dann, wenn die betroffenen Merkmale als erklärende Variablen in Regressionsmodellen verwendet werden. Hier wäre im Zweifel das Entfernen der betroffenen Unternehmen erforderlich, was aber Einfluss auf die Schätzergebnisse hätte. Ist davon auszugehen, dass die Nutzer ohnehin eine Ausreißerbereinigung durchführen, kann eine solche Maßnahme vertreten werden, sofern die betroffenen Unternehmen gesondert gekennzeichnet werden. Sollen jedoch alle Unternehmen bei der Analyse berücksichtigt werden, so hängt die Stärke der negativen Beeinflussung der Analyseergebnisse durch diese Maßnahme davon ab, wie viele Unternehmen in das Censoring oder Replacement einbezogen werden müssen und wie stark die Einzelwerte von den entsprechenden Originalwerten abweichen. Da hier davon ausgegangen werden muss, dass die Anzahl der einbezogenen Einheiten die Gruppengröße bei der getrennten Mikroaggregation deutlich übersteigt, ist diese Variante grundsätzlich gegenüber einer getrennten Mikroaggregation auf alle metrischen Merkmale oder dem Einsatz von stochastischen Überlagerungen auf alle Merkmale abzulehnen, sofern man nicht auf die betroffenen Unternehmen verzichten will.

Zu Strategie 5: Die multiplikative stochastische Überlagerung führt zu einer höheren Schutzwirkung als eine getrennte abstandsorientierte Mikroaggregation. Zudem weist sie aus Sicht des Analysepotenzials weniger Probleme auf als die Mikroaggregation einzelner Variablen mit gruppierten oder gemeinsamer Mikroaggregation, sofern diese Verfahren nur auf die Überschneidungsmerkmale angewendet werden. Wird nur die abhängige Variable in einem linearen Regressionsmodell überlagert, so ist der Schätzer von vornherein erwartungstreu. Werden einzelne oder alle unabhängigen Variablen stochastisch überlagert, so kann der verzerrte Schätzer beispielsweise durch Instrumentenvariablen-Schätzungen korrigiert werden. Werden nur einzelne Einflussgrößen anonymisiert, so sind lediglich für diese Instrumente erforderlich. Das gleiche gilt grundsätzlich auch für nichtlineare Modelle (Korrektur zum Beispiel durch den SIMEX-Schätzer). Allerdings ist zwingend, dass eine Beschränkung multiplikativer stochastischer Überlagerungen auf die Überschneidungsmerkmale eine höhere Varianz der Überlagerungsfaktoren erforderlich macht, wenn faktische Anonymität sichergestellt werden soll. Einer Erhöhung der Varianz ist jedoch bereits von vornherein dadurch eine Grenze gesetzt, dass die Überlagerungsfaktoren ausschließlich

positiv sein dürfen, um Vorzeichenwechsel zu vermeiden. Zudem hat sich bei den durchgeführten Untersuchungen gezeigt, dass eine zu hohe Varianz der Überlagerungen zu schlechteren Ergebnissen bei der Korrektur von Schätzfehlern führt, bis dahin, dass die Korrektur gar nicht gelingt. Zuletzt ist eine höhere Fehlervarianz stets mit einem größeren Effizienzverlust der Schätzer verbunden. Auch bei deskriptiven Auswertungen, insbesondere nach Teilgesamtheiten, erweist sich eine höhere Varianz der Überlagerungen als problematisch.

34.2 Fazit zur Beschränkung der Anonymisierung auf die Überschneidungsmerkmale

Ungeachtet dessen, dass es Spezialfälle gibt, in denen eine alleinige Behandlung der Überschneidungsmerkmale a priori auszuschließen ist – beispielsweise ist bei der im Projekt untersuchten Umsatzsteuerstatistik ein solches Vorgehen aufgrund der Abhängigkeiten nahezu aller metrischen Überschneidungsmerkmale von dem Merkmal *Gesamtumsatz* nicht möglich –, kann eine Beschränkung der Anonymisierung auf die Überschneidungsmerkmale sinnvoll sein, wenn entweder aufgrund eines hohen natürlichen Schutzes in Daten nur eine sehr schwache Anonymisierung erforderlich ist oder die faktische Anonymität nur mit sehr einschneidenden Maßnahmen sichergestellt werden kann. Folgende Fälle sind denkbar:

1. Sehr schwaches Anonymisierungserfordernis:
   In diesem Fall ist es möglich, bei Beschränkung der Anonymisierung auf die Überschneidungsmerkmale einen Scientific-Use-File zu erstellen, d.h. faktisch anonyme Daten, die über alle Merkmale ein ausreichendes Analysepotenzial aufweisen. Der Fall tritt ein, wenn bereits durch eine formale Anonymisierung (d.h. es werden direkte Identifikatoren wie Name und Adresse aus der Datei entfernt) ein hoher Schutz entsteht und somit eine schwache Veränderung der Überschneidungsmerkmale zur Unterschreitung der vorab definierten oberen Risikoschwelle ausreicht. Eine Behandlung der Nicht-Überschneidungsmerkmale ist dann nicht mehr nötig, da der erforderliche Schutz bereits durch die Verhinderung einer Reidentifikation (ungeachtet der Brauchbarkeit gewonnener Einzelinformationen) erreicht wird.

2. Sehr starkes Anonymisierungserfordernis:
   In diesem Fall ist es auch dann nicht möglich, einen Scientific-Use-File zu erstellen, wenn alle Merkmale anonymisiert werden. Hier ist in jedem Fall mit einer starken Beeinträchtigung des Analysepotenzials zumindest für einen Teil der Variablen zu rechnen. Daher ist es vorzuziehen, den Verlust an Analysepotenzial auf die Überschneidungsmerkmale zu beschränken. Allerdings muss dabei auch beachtet werden, dass wesentliche lineare und nichtlineare Zusammenhänge hierbei nicht zerstört werden dürfen. Insbesondere auswertungsrelevante Quotienten sollten annähernd erhalten bleiben.

Es ist davon auszugehen, dass beide oben genannten Fälle eher theoretischer Natur und in der Praxis sehr unwahrscheinlich sind. Bei realen Daten ist es meist so, dass bei einer sehr schwachen Anonymisierung mit besonders gefährdeten und unzureichend geschützten Bereichen in den Daten zu rechnen ist. Können Daten mit schwachen Maßnahmen ausreichend geschützt werden, so ist dies ein Glücksfall für die anonymisierende Institution wie auch die Datennutzer; die Frage der Beschränkung auf die Überschneidungsmerkmale wird dann kaum zu einem Dissens führen. Wenn hingegen eine sehr starke Anonymisierung einen nicht tolerierbaren Verlust an wissenschaftlicher Analysefähigkeit mit sich bringt, dann haben weder die anonymisierende Institution noch die Datennutzer ein Interesse, solche Daten zu erstellen bzw. zu analysieren. Beide Fälle beschreiben daher nur die theoretischen Extrema, nicht aber die üblicherweise vorkommenden Fälle der Erstellung von Scientific-Use-Files. Dies hat sich im regelmäßigen Diskurs zwischen Datenhaltern und Wissenschaftlern im Zuge der Projektarbeiten gezeigt. In der Praxis gilt es, mit der zu entwickelnden Anonymisierungsmaßnahme gleichzeitig für den Schutzaspekt und das Analysepotenzial die vorab definierten Toleranzen einzuhalten. Versteht man die beiden obigen Fälle als entgegengesetzte Pole, so besteht die anspruchsvolle Aufgabe darin, einen schmalen Grad zwischen diesen Polen zu finden, auf welchem keine der beiden Toleranzgrenzen überschritten wird.

Teil XI

Die Anonymisierung der Projektstatistiken

In diesem Teil wird anhand der Projektstatistiken vorgeführt, wie das in Teil VI theoretisch beschriebene Vorgehen bei der Anonymisierung in die Praxis umgesetzt werden kann. Es wird verdeutlicht, wie der diskursive Prozess bei der Generierung eines Scientific-Use-Files bei realen Daten unterschiedlicher Struktur aussehen kann.

Die mit den Projektstatistiken durchgeführten Untersuchungen sind als Fallbeispiele und Handlungsvorschläge für andere Datenhalter zu verstehen. Insbesondere im Falle der drei Erhebungen der statistischen Ämter, der Kostenstrukturerhebung im Verarbeitenden Gewerbe und Bergbau, Umsatzsteuerstatistik und Einzelhandelsstatistik, werden zudem Dateien und deren dazugehörige Anonymisierungsmaßnahmen vorgestellt, die sich bereits während der Projektlaufzeit als Scientific-Use-Files durchgesetzt haben.

Kapitel 35

Anonymisierung der Kostenstrukturerhebung im Verarbeitenden Gewerbe

35.1 Besonderheiten bei der Anonymisierung der Kostenstrukturerhebung

Bei der Kostenstrukturerhebung im Verarbeitenden Gewerbe des Jahres 1999 handelt es sich um eine Stichprobe im Umfang von ungefähr 43%. Aufgrund der vergleichsweise geringen Anzahl von etwa 17.000 Einheiten – zum Beispiel besitzt die ebenfalls im Projekt untersuchte Umsatzsteuerstatistik des Jahres 2000 etwa 2,9 Millionen Einheiten – und der Tatsache, dass nur Unternehmen mit wenigstens 20 Beschäftigten berücksichtigt werden, stellt die Anonymisierung der Kostenstrukturerhebung eine besondere Herausforderung dar. Während z.B. bei der Einzelhandelsstatistik sämtliche Einheiten gemäß der Klassifikation WZ93 einer einzigen Wirtschaftsabteilung zuzuordnen sind, verteilen sich die Einheiten der Kostenstrukturerhebung auf 28 Wirtschaftsabteilungen. Aus diesem Grunde treten mitunter sehr kleine Fallzahlen bei der Tabellierung nach den im Datensatz enthaltenen kategorialen Merkmalen auf. Diese Problematik wird anhand untenstehendem Ausschnitt der Kostenstrukturerhebung in Tabelle 35.1 deutlich. Hier werden die WZ93-Klassifikation und der zu Projektbeginn im Merkmalskanon vorgesehene siedlungsstrukturelle Kreistyp BBR9[64] verwendet.

Aus dem obigen Ausschnitt wird bereits deutlich, dass sehr dünn besetzte Wirtschaftsabteilungen durch die Koppelung mit der Regionalinformation einer besonderen Berücksichtigung bei der Geheimhaltung bedürfen. Ein weiterer Zuwachs des Reidentifikationsrisikos ist mit wachsender Beschäftigtenanzahl eines Unternehmens zu erwarten, was durch Tabelle 35.2 angedeutet wird, welche die Verteilung der Daten auf Beschäftigtengrößenklassen enthält.

64) Dieser Schlüssel dient dem intraregionalen Vergleich. Die Typisierung der Kreise und Kreisregionen erfolgt außerhalb der Kernstädte nach der Bevölkerungsdichte. Insgesamt ergeben sich neun Kreistypen.

Tabelle 35.1: Ausschnitt der Kostenstrukturerhebung

WZ93\BBR9	1	2	3	4	5	6	7	8	9	Summe
10	5	5	2	4	0	2	14	7	0	39
14	7	19	15	4	2	46	32	24	8	157
⋮									⋮	
20	38	54	50	15	8	129	111	57	42	504
22	356	154	57	23	91	147	50	54	18	950
24	267	174	82	32	37	166	63	66	14	901
25	97	187	90	25	16	212	114	85	41	867
26	116	108	73	49	35	228	184	100	72	965
27	120	152	44	21	18	132	61	29	16	593
30	33	28	11	2	12	43	11	13	0	153
⋮									⋮	
37	13	15	6	9	9	22	7	11	2	94
Summe	2.920	2.994	1.379	486	788	4.199	1.987	1.488	677	16.918

Tabelle 35.2: Verteilung der Unternehmen auf Beschäftigtengrößenklassen

Größenklasse	Abs. Häufigkeit	Relative Häufigkeit	Kum. Häufigkeit
20-49	5.294	31,29	31,29
50-99	4.119	24,35	55,64
100-249	3.906	23,09	78,73
250-499	1.758	10,39	89,12
500-999	1.085	6,41	95,53
1.000 und mehr	756	4,47	100,00

Aus den genannten Gründen wurde der Kostenstrukturerhebung in den Projektarbeiten besonders große Aufmerksamkeit geschenkt. Die Erfahrungen konnten später bei den anderen Projektstatistiken eingebracht werden. In diesem Kapitel wird die Vorgehensweise im Projekt zur Generierung einer faktisch anonymen Datei mit bestmöglichem Erhalt an Potenzial für wissenschaftliche Analysen am Beispiel der Verfahrensgruppe Mikroaggregation (zur Beschreibung siehe Unterabschnitt 6.2.4) illustriert.

35.2 Verfügbares Zusatzwissen

Wie in Abschnitt 11.3 beschrieben wurde, kann das Zusatzwissen eines potenziellen Datenangreifers aus verschiedenen Quellen stammen. Während sich für einen Massenfischzug kommerzielle Datenbanken besonders eignen, kommen bei einem Einzelangriff individuelle Kenntnisse über ein gesuchtes Unternehmen, welche etwa über Internetrecherchen ermittelbar sind, hinzu.

Als Überschneidungsmerkmale zwischen kommerziellen Datenbanken und der Kostenstrukturerhebung im Verarbeitenden Gewerbe wurden folgende Merkmale beobachtet:

- Gesamtumsatz
- Anzahl der Beschäftigten
- Regionalkennung
- Wirtschaftszweigklassifikation

Mittels persönlicher Informationsquellen kann ein Datenangreifer vereinzelt Informationen über den Aufwand an Forschung und Entwicklung eines Unternehmens oder Handelsaktivitäten einbringen.

Die im Projekt für Massenfischzugsimulationen verwendeten Quellen werden im Folgenden näher beschrieben. Diese sind die Umsatzsteuerstatistik (unter Verwendung des Gesamtumsatzes des Jahres 1999), die MARKUS-Datenbank (mit kumulierten Jahresangaben für 1999) und die Originaldaten der Kostenstrukturerhebung, letztere zur Abschätzung einer Obergrenze für das mit den anonymisierten Daten der Kostenstrukturerhebung verbundene Reidentifikationsrisiko.

35.2.1 Umsatzsteuerstatistik als Zusatzwissen

Zunächst wurde die Umsatzsteuerstatistik verwendet, da diese von Beginn an innerhalb der Projektdaten verfügbar war. Die Ergebnisse von Angriffsszenarien mithilfe der Um-

satzsteuerstatistik dienten einer ersten Einschätzung des mit den anonymisierten Daten verbundenen Reidentifikationsrisikos. Als Überschneidungsmerkmale standen hier zur Verfügung:

- Gesamtumsatz
- Regionalkennung
- Wirtschaftszweigklassifikation

Nachfolgende Tabelle 35.3 enthält die Verteilung der 9.283 überprüfbaren Unternehmen (d.h. solche Einheiten, für welche nach einem Reidentifikationsversuch die Richtigkeit der Zuordnung mit einer Einheit der Kostenstrukturerhebung überprüft werden konnte) auf Beschäftigtengrößenklassen:[65]

Tabelle 35.3: Verteilung der überprüfbaren Unternehmen auf Beschäftigtengrößenklassen

Größenklasse	abs. Häufigkeit	rel. Häufigkeit	kum. rel. Häufigkeit
20-49	3.120	0,34	0,34
50-99	2.351	0,25	0,59
100-249	2.107	0,23	0,82
250-499	848	0,09	0,91
500-999	513	0,06	0,96
1.000 und mehr	344	0,04	100,00

35.2.2 MARKUS-Datenbank als Zusatzwissen

Für ein realistisches Szenario wurde die so genannte MARKUS-Datenbank verwendet. Sie besteht aus ausgewählten Unternehmen der Creditreform. Sie ist im Handel frei erhältlich als CD-ROM und wird vierteljährlich herausgegeben, wobei nur jeweils ca. 4% aller Unternehmen von einer Ausgabe zur nächsten ausgetauscht werden.
Mit dem Merkmal *Anzahl der Beschäftigten* stand hier gegenüber der Umsatzsteuerstatistik ein weiteres Überschneidungsmerkmal zur Verfügung:

- Anzahl der Beschäftigten
- Gesamtumsatz
- Regionalkennung

65) Es sei angemerkt, dass obige Struktur nicht allein aus den Daten der Umsatzsteuerstatistik abgelesen werden kann, da das Merkmal Beschäftigte nicht im Datensatz enthalten ist.

- Wirtschaftszweigklassifikation

Darüber hinaus enthält die MARKUS-Datenbank folgende Informationen:

- Firmenname und Adresse
- Bilanzangaben
- Stammkapital
- Tätigkeitsbeschreibung
- Beteiligungsstruktur
- Angaben zur Geschäftsführung

Die folgende Tabelle 35.4 zeigt die Verteilung der 9.349 überprüfbaren Einheiten aus der MARKUS-Datenbank auf Beschäftigtengrößenklassen:

Tabelle 35.4: Verteilung der MARKUS-Unternehmen auf Beschäftigtengrößenklassen

Größenklasse	Abs. Häufigkeit	Relative Häufigkeit	Kum. Häufigkeit
20-49	2.692	0,29	0,29
50-99	2.329	0,25	0,54
100-249	2.300	0,25	0,78
250-499	1.032	0,11	0,89
500-999	591	0,06	0,96
1.000 und mehr	450	0,05	100,00

35.3 Anonymisierungsmaßnahmen

Bei den Daten der Kostenstrukturerhebung wurden zahlreiche Anonymisierungsvarianten getestet. Hervorzuheben sind die Varianten der Mikroaggregation, der Zufallsüberlagerung sowie deren Verknüpfungen mit traditionellen, auf die kategorialen Merkmale angewendeten Methoden.

Zum Verständnis der Vorgehensweise im Projekt und der Anwendung des in Kapitel 12 beschriebenen Schutzwirkungskonzeptes wird das Beispiel der Mikroaggregation (zur Beschreibung siehe Unterabschnitt 6.2.4) detailliert beschrieben. Es werden fünf Anonymisierungsvarianten betrachtet:

FORMAL: Diese Variante, die so genannte formale Anonymisierung, entsteht allein aus der Herausnahme direkter Identifikatoren wie Name und Adresse.

MA33G: Dies ist die schwächste Form der Mikroaggregation, bei der jedes metrische Merkmal seine eigene Gruppe definiert und somit separat mikroaggregiert wird.

MA11G: Hier werden 11 Gruppen mit jeweils drei Merkmalen gebildet, wobei bei der Einteilung hoch korrelierte Merkmale zusammen gruppiert wurden.

MA8G: Hier werden acht Gruppen von einer Größe zwischen 2 und 12 Merkmalen gebildet, wobei die Merkmale aus inhaltlichen Gesichtspunkten zusammen gruppiert wurden.

MA1G: Dies ist die stärkste Form der Mikroaggregation, wo sämtliche metrische Merkmale zusammen gruppiert werden und somit Tripel von Einheiten entstehen, welche sich höchstens durch die Ausprägungskombinationen in ihren katgorialen Merkmalen unterscheiden können.

35.4 Überprüfung der Schutzwirkung

Im Folgenden werden die Ergebnisse von Massenfischzügen unter Verwendung der in Abschnitt 11.3 vorgestellten Quellen des Zusatzwissens dargestellt. In Unterabschnitt 35.4.1 werden Szenarien mittels der Umsatzsteuerstatistik und der kommerziell verfügbaren MARKUS-Datenbank (ein realistisches Szenario) simuliert. In Unterabschnitt 35.4.2 werden die Ergebnisse dieser Szenarien mit dem so genannten Worst-Case Szenario, bei welchem die Originaldaten als bestmögliches Zusatzwissen eines Datenangreifers angenommen werden, verglichen. Zwar ist dieses Szenario sehr realitätsfern, jedoch liefert es eine Obergrenze (nicht die kleinste obere Grenze) für das mit den anonymisierten Daten verbundene Reidentifikationsrisiko. In der Praxis sollte dieses Szenario daher nicht oder nur in geringem Maße in die Bewertung der Vertraulichkeit einer Datei einfließen. Die Wahl der Überschneidungsmerkmale für das Worst-Case Szenario – hier könnte man naturgemäß alle Merkmale in den Originaldaten festlegen – fällt auf die „üblichen Verdächtigen“, die in realistischen Szenarien zur Verfügung stehen. Dies ermöglicht eine sinnvollere Gegenüberstellung der verschiedenen Simulationen.[66]

35.4.1 Realistische Massenfischzugszenarien

In diesem Abschnitt werden Szenarien mit der Umsatzsteuerstatistik (ca. 9.400 überprüfbare Einheiten) und der MARKUS-Datenbank (ca. 9.300 überprüfbare Einheiten)

66) Einem Datenangreifer ist eine Abschätzung des Risikos kaum möglich. Zwar könnte er Simulationen mit verschiedenen externen Quellen laufen lassen, ein Vergleich verschiedener Simulationen ist jedoch bestenfalls auf Basis der berechneten, wenig zuverlässigen Gesamtdistanzen für alle Zuordnungen denkbar.

als Zusatzwissen durchgeführt. Obwohl die Umsatzsteuerstatistik einem potenziellen Datenangreifer nicht als Zusatzwissen zur Verfügung stehen wird, werden die entsprechenden Simulationen als (pseudo)realistisch eingestuft, da dieses Zusatzwissen eine mit kommerziellen Datenbanken vergleichbare Qualität hat.

Umsatzsteuerstatistik versus anonymisierte Kostenstrukturerhebung

Als Überschneidungsmerkmale zwischen Umsatzsteuerstatistik und Kostenstrukturerhebung werden im Folgenden beispielhaft[67] verwendet:

- Gesamtumsatz,
- Siedlungsstruktureller Kreistyp (BBR9),
- Wirtschaftszweigklassifikation (WZ93), Zweistellerebene.

Die nachfolgende Tabelle 35.5 enthält die Verteilung der Reidentifikation(srisik)en auf Beschäftigtengrößenklassen.

Tabelle 35.5: Reidentifikationen (Umsatzsteuerstatistik) nach Beschäftigtengrößenklassen

Varianten	Total	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000
MA1G	404	103	61	55	64	47	74
	0,0435	0,0330	0,0259	0,0261	0,0755	0,0916	0,2151
MA8G	1.177	366	223	246	137	96	109
	0,1270	0,1173	0,0949	0,1168	0,1616	0,1871	0,3169
MA11G	2.551	824	602	570	238	180	137
	0,2748	0,2641	0,2561	0,2705	0,2807	0,3509	0,3983
MA33G	2.695	894	639	580	246	189	147
	0,2903	0,2865	0,2718	0,2753	0,2901	0,3684	0,4273
FORMAL	2.677	890	635	574	247	189	142
	0,2884	0,2853	0,2701	0,2724	0,2913	0,3684	0,4128

Erste Zeile: Absolute Häufigkeit; zweite Zeile: Relative Häufigkeit

Die Tabelle 35.5 enthält je Zelle die absolute (erste Zeile) und relative (zweite Zeile) Häufigkeit erfolgreicher Reidentifikationsversuche. Die relative Häufigkeit bezieht sich hier auf die Besetzungszahl der jeweiligen Merkmalskombination im Zusatzwissen. Es wird erwartungsgemäß beobachtet, dass sich die relative Häufigkeitsverteilung der Reidentifikationen mit sinkendem Anonymisierungsgrad der Verteilung bei formal anonymisierten Daten (letzte Zeile in Tabelle 35.5) annähert. Die Schutzwirkung formaler Anonymisierung, die allein

67) Hier sind auch andere Gliederungstiefen bei den beiden kategorialen Merkmalen, vgl. hierzu Abschnitt 35.6, denkbar.

aus der Herausnahme direkter Identifikatoren wie Name und Adresse besteht, kann als natürlicher Schutz in den Daten interpretiert werden. Dieser Schutz ist bereits durch die Dateninkompatibilitäten wie z.B. Abweichungen im Merkmal *Gesamtumsatz* oder verschiedene Branchenzuordnung eines Unternehmens in den beiden Datenquellen gegeben. Die geringsten Risiken treten bei Unternehmen der Größenklasse 50 - 249 Beschäftigte auf. An dieser Stelle muss darauf hingewiesen werden, dass die Einteilung in Größenklassen mit Bedacht zu wählen ist, da sich die Gestalt der Risikoverteilung grundlegend bei einer anderen Einteilung ändern kann.

Obwohl die Mikroaggregationsverfahren naturgemäß stärker in den weniger dicht besetzten Bereichen der Merkmale wirken, zeigt die letzte Spalte in Tabelle 35.5, dass besonders die Großunternehmen in der Klasse mit wenigstens 1.000 Beschäftigten auch nach der Anonymisierung stärker als Unternehmen der restlichen Größenklassen gefährdet sind. Sogar im Falle der für Analysezwecke nur bedingt tauglichen Variante MA1G konnten circa 21 % der Großunternehmen reidentifiziert werden.

Wie erwartet stieg der Anteil der Reidentifikationen beim Übergang von Variante MA8G zu MA11G sehr deutlich. Dies ist darauf zurück zu führen, dass in der Variante MA8G das metrische Überschneidungsmerkmal *Gesamtumsatz* in einer 12-elementigen Gruppe (u.a. gemeinsam mit den Merkmalen *Anzahl der Beschäftigten*, *Umsatz aus eigenen Erzeugnissen* und *Kosten insgesamt*) mikroaggregiert und damit stark verändert wurde. In der Variante MA11G fand sich das Merkmal *Gesamtumsatz* in einer dreielementigen Gruppe mit den Merkmalen *Umsatz aus eigenen Erzeugnissen* und *Gesamtleistung* wieder.

Als Hauptursache für Fehlzuordnungen können Dateninkompatibilitäten, die bereits vor der Anonymisierung bestanden, zwischen den beiden Erhebungen angesehen werden. Während nur etwa 1 % der Unternehmen bzgl. des siedlungsstrukturellen Kreistyps verschieden klassifiziert wurden, fanden sich nahezu 25 % der Unternehmen der Kostenstrukturerhebung in der Umsatzsteuerstatistik in einer anderen Wirtschaftsabteilung wieder. Das Merkmal *Gesamtumsatz* hingegen wies nur geringe Unterschiede in den beiden Erhebungen auf. Etwa 18,8 % der Unternehmen zeigten hier Abweichungen von mehr als 10 % in den beiden Erhebungen.

MARKUS-Datenbank versus anonymisierte Kostenstrukturerhebung

Unter den Überschneidungsmerkmalen zwischen MARKUS-Datenbank und Kostenstrukturerhebung findet sich im Folgenden gegenüber dem vorherigen Abschnitt ein weiteres Merkmal:

- Anzahl der Beschäftigten,
- Gesamtumsatz,
- Siedlungsstruktureller Kreistyp (BBR9),
- Wirtschaftszweigklassifikation (WZ93), Zweistellerebene.

In Variante MA8G wurden, wie im vorherigen Abschnitt bereits erwähnt, die beiden metrischen Überschneidungsmerkmale *Gesamtumsatz* und *Anzahl der Beschäftigten* in einer gemeinsamen Gruppe mikroaggregiert. Das bedeutet, dass feinere Unterschiede zwischen diesen Merkmalen im Zuge der Anonymisierung verloren gegangen sind. Anders verhält es sich mit der Variante MA11G, wo die Merkmale *Anzahl der Beschäftigten* und *Gesamtumsatz* in verschiedenen Gruppen mikroaggregiert wurden.

Analog zu Tabelle 35.5 erhalten wir

Tabelle 35.6: Reidentifikationen (MARKUS) nach Beschäftigtengrößenklassen

Varianten	total	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000
MA1G	353	59	35	71	60	53	75
	0,0376	0,0219	0,0150	0,0309	0,0581	0,0897	0,1667
MA8G	1.845	343	347	503	279	210	163
	0,1964	0,1274	0,1490	0,2187	0,2703	0,3553	0,3622
MA11G	2.273	419	448	609	355	244	198
	0,2420	0,1556	0,1924	0,2648	0,3440	0,4129	0,4400
MA33G	2.289	420	443	609	370	246	201
	0,2437	0,1560	0,1902	0,2648	0,3585	0,4162	0,4467
FORMAL	2.294	420	442	610	373	247	202
	0,2442	0,1560	0,1898	0,2652	0,3614	0,4179	0,4489

Erste Zeile: Absolute Häufigkeit; zweite Zeile: Relative Häufigkeit

Auffallend ist, dass der Verlust an Schutzwirkung beim Übergang von Variante MA8G zu MA11G nicht so groß ist wie im vorherigen Experiment mit der Umsatzsteuerstatistik. Dasselbe gilt für den Übergang von Unternehmen mit 50 – 999 Beschäftigten zu solchen mit wenigstens 1000 Beschäftigten. Bei den schwächeren Mikroaggregationsvarianten MA11G, MA33G und FORMAL fällt der Anteil an Reidentifikationen für kleinere und mittlere Unternehmen (20 – 249 Beschäftigte) geringer aus als im vorherigen Experiment. Es ist etwas überraschend, dass allein bei der Variante MA8G der Anteil an Reidentifikationen gegenüber dem vorherigen Experiment zugenommen, bei allen anderen betrachteten Varianten jedoch abgenommen hat, wo doch mit der MARKUS-Datenbank ein zusätzliches Überschneidungsmerkmal zur Verfügung stand. Als mögliche Begründung hierfür könnten die deutlichen Differenzen in den beiden Erhebungen im Merkmal *Gesamtumsatz* dienen. Etwa 50% aller Unternehmen der MARKUS-Datenbank weichen im Merkmal *Gesamtumsatz* um mehr als 10 % von den jeweiligen Einträgen in der Kostenstrukturerhebung ab. Die Abweichungen in den Merkmalen *Wirtschaftszweigklassifikation* (hier wurden etwa 24% aller Unternehmen auf Abteilungsebene in den beiden Erhebungen verschieden klassifiziert) und *Siedlungsstruktureller Kreistyp* (hier wurden weniger als 2% aller Unternehmen verschieden klassifiziert) sind mit denen des vorherigen Experimentes vergleichbar. Allerdings hat sich in zusätzlich durchgeführten Simulationen gezeigt, dass eine Herausnahme des Merkmals *Gesamtumsatz* zu geringeren Quoten richtiger Zuordnungen führte und damit dieses Merkmal in jedem Falle reidentifizierende Wirkung hat.

35.4.2 Worst-Case Szenario

In den folgenden Simulationen werden verschiedene Teilmengen der metrischen Merkmale als Überschneidungsmerkmale verwendet. Zum einen werden alle 33 metrischen Merkmale als Überschneidungsmerkmale angenommen. Zum anderen werden Simulationen mit einem Überschneidungsmerkmal, dem *Gesamtumsatz* (um einen Vergleich mit dem Ergebnis der Simulation unter Verwendung der Umsatzsteuerstatistik in Unterabschnitt 35.4.1 zu ermöglichen), mit zwei Überschneidungsmerkmalen, dem *Gesamtumsatz* und der *Anzahl der Beschäftigten* (um einen Vergleich mit dem Ergebnis der Simulation unter Verwendung der MARKUS-Datenbank in Unterabschnitt 35.4.1 zu ermöglichen), und mit drei Überschneidungsmerkmalen, nämlich *Gesamtumsatz, Anzahl der Beschäftigten* und *Gesamtaufwendungen für innerbetriebliche Forschung und Entwicklung*. Allerdings kann in der Praxis letzteres Merkmal nur in Einzelfällen (etwa eher via Internetrecherchen) gewonnen werden und stand in den Simulationen in Unterabschnitt 35.4.1 daher nicht zur Verfügung. Durchgängig wurden die kategorialen Merkmale *Siedlungsstruktureller Kreistyp (BBR9)* und *Wirtschaftszweigklassifikation (WZ93) auf Zweistellerebene* zur Blockung der Daten verwendet.

Untenstehende Tabelle 35.7 zeigt die absoluten und relativen Häufigkeiten der Reidentifikationen unter Verwendung von 1, 2, 3 und 33 metrischen Überschneidungsmerkmalen.

Tabelle 35.7: Reidentifikationen (Worst-Case) nach der Anzahl der Überschneidungsmerkmale

Variante	33 Merkmale	3 Merkmale	2 Merkmale	1 Merkmal
MA1G	8.941	2.156	2.076	1.096
	0,5285	0,1274	0,1227	0,0648
MA8G	16.792	12.820	11.127	3.621
	0,9926	0,7578	0,6577	0,2140
MA11G	16.853	16.732	16.765	12.066
	0,9962	0,9890	0,9910	0,7132
MA33G	16.918	16.918	16.912	16.757
	1,0000	1,0000	0,9996	0,9905
FORMAL	16.918	16.918	16.918	16.918
	1,0000	1,0000	1,0000	1,0000

Erste Zeile: Absolute Häufigkeit; zweite Zeile: Relative Häufigkeit

Der Schutz nimmt beim Übergang von einem (*Gesamtumsatz*) zu zwei metrischen Überschneidungsmerkmalen (*Gesamtumsatz* und *Anzahl der Beschäftigten*) beachtlich ab, wohingegen der Anstieg des Reidentifikationsrisikos beim Übergang von zwei zu drei metrischen Überschneidungsmerkmalen sehr verhalten ausfällt und dieser im Falle der Variante MA11G sogar leicht sinkt. Bestätigt wird die verhältnismäßig schwache Schutzwirkung der Variante MA11G, die bereits im vorhergehenden Unterabschnitt 35.4.1 beobachtet wurde.

Analog zu den Tabellen 35.5 und 35.6 wird in Tabelle 35.8 die Verteilung der Reidentifikationen auf Beschäftigtengrößenklassen betrachtet, beginnend mit der Simulation unter Verwendung eines Überschneidungsmerkmals:

Tabelle 35.8: Reidentifikationen (Worst-Case) mit einem Überschneidungsmerkmal nach Beschäftigtengrößenklassen

Variante	total	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000
MA1G	1.096	243	161	164	151	145	232
	0.0648	0.0459	0.0391	0.0420	0.0859	0.1336	0.3069
MA8G	3.621	1.043	681	765	417	354	361
	0,2140	0,1970	0,1653	0,1959	0,2372	0,3263	0,4775
MA11G	12.066	3.841	2.852	2.706	1.252	800	615
	0,7132	0,7255	0,6924	0,6928	0,7122	0,7373	0,8135
MA33G	16.757	5.236	4.084	3.873	1.741	1.078	745
	0,9905	0,9890	0,9915	0,9916	0,9903	0,9935	0,9854
FORMAL	16.918	5.294	4.119	3.906	1.758	1.085	756
	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Erste Zeile: Absolute Häufigkeit; zweite Zeile: Relative Häufigkeit

Tabelle 35.8 kann der Tabelle 35.5 (Simulation mit der Umsatzsteuerstatistik) gegenüber gestellt werden, da dort dieselben Überschneidungsmerkmale im Zusatzwissen vorhanden waren. Zunächst fällt wieder der deutliche Anstieg des Reidentifikationsrisikos beim Übergang von Variante MA8G zu MA11G auf. Darüber hinaus kann beobachtet werden, dass der Anteil an Reidentifikationen in der obersten Beschäftigtengrößenklasse überraschenderweise rückläufig ist, was aber bei den ohnehin sehr hohen Trefferquoten im Worst-Case Szenario nicht von besonderer Bedeutung sein muss. Vergleichbare Ergebnisse werden bei der Simulation mit zwei metrischen Überschneidungsmerkmalen beobachtet (Tabelle 35.9).

Nicht verwunderlich ist der Anstieg an Reidentifikationen gegenüber der vorherigen Simulation, da hier ein zusätzliches Überschneidungsmerkmal von bester Qualität für dem Datenangreifer zur Verfügung stand.

Ebenso kann Tabelle 35.9 der Tabelle 35.6 (Simulation mit der MARKUS-Datenbank) gegenüber gestellt werden. Auch hier ist ein deutlicher Anstieg an Reidentifikationen beim Übergang von MA1G zu MA8G zu verzeichnen, während der Unterschied zwischen MA11G und MA33G in dieser Simulation nahezu vernachlässigbar erscheint.

35.4.3 Zusammenführung zu einem Gesamtrisikomaß

Sogar einer erfolgreiche Zuordnung eines Merkmalsträgers zu den Zieldaten kann vergeblich sein, wenn der den Datenangreifer interessierende Einzelwert (oder die interessierende Einzelinformation) von dem zugehörigen Originalwert relativ um mehr als eine vorgegebene Nutzenschwelle γ abweicht. Das auf diese Weise reduzierte Risiko wurde in Abschnitt 12.3

Tabelle 35.9: Reidentifikationen (Worst-Case) mit zwei Überschneidungsmerkmalen nach Beschäftigtengrößenklassen

Variante	total	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000
MA1G	2.076	394	344	420	311	275	332
	0,1227	0,0744	0,0835	0,1020	0,0796	0,1564	0,3060
MA8G	11.127	3.344	2.610	2.578	1.206	769	620
	0,6577	0,6317	0,6336	0,6600	0,6860	0,7088	0,8201
MA11G	16.765	5.237	4.076	3.879	1.746	1.079	748
	0,9910	0,9892	0,9896	0,9931	0,9932	0,9945	0,9894
MA33G	16.912	5.294	4.117	3.906	1.756	1.085	754
	0,9996	1,0000	0,9995	1,0000	0,9989	1,0000	0,9974
FORMAL	16.918	5.294	4.119	3.906	1.758	1.085	756
	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000	1,0000

Erste Zeile: Absolute Häufigkeit; zweite Zeile: Relative Häufigkeit

Enthüllungsrisiko genannt. Setzt man die Nutzenschwelle beispielhaft mit $\gamma = 0,05$ an, so erhält man zunächst die globalen, mit den drei in Unterabschnitten 35.4.1 und 35.4.2 durchgeführten Experimenten verbundenen Enthüllungsrisiken. In nachfolgender Tabelle 35.10 und Abbildung 35.1 werden die Simulationen mit den Originaldaten der Kostenstrukturerhebung (ORIGINAL, Szenario A) den Simulationen mit der Umsatzsteuerstatistik (Szenario B, TTS) und der MARKUS-Datenbank (Szenario B, MARKUS) als Zusatzwissen gegenüber gestellt.

Tabelle 35.10: Enthüllungsrisiken auf dem Niveau $\gamma = 0,05$

	Zusatzwissen		
Zieldaten	ORIGINAL	TTS	MARKUS
MA1G	0,035	0,017	0,013
MA8G	0,379	0,072	0,108
MA11G	0,809	0,225	0,194
MA33G	0,997	0,290	0,243
FORMAL	1,000	0,288	0,244

Zur Zusammenführung der in Abbildung 35.1 enthaltenen Risiken zu einem Gesamtrisikomaß kann z.B. eine Konvexkombination dienen. Hierzu bezeichne $\hat{P}_{A/\gamma}(w \text{ enthüllt})$ das mit dem Worst-Case Szenario geschätzte Risiko der Enthüllung des Einzelwertes w. Mit $\hat{P}_{B/\gamma}(w \text{ enthüllt})$ wird das mit den realistischen Szenarien geschätzte Enthüllungsrisiko (möglicherweise ebenfalls durch ein gewichtetes Mittel, gebildet aus den verschiedenen Ergebnissen der realistischen Szenarien, berechnet) bezeichnet. Als Gesamtrisikomaß erhalten wir dann

$$\hat{P}_{\gamma}(w \text{ enthüllt}) := \lambda \cdot \hat{P}_{A/\gamma}(w \text{ enthüllt}) + (1 - \lambda) \cdot \hat{P}_{B/\gamma}(w \text{ enthüllt}) \qquad (35.1)$$

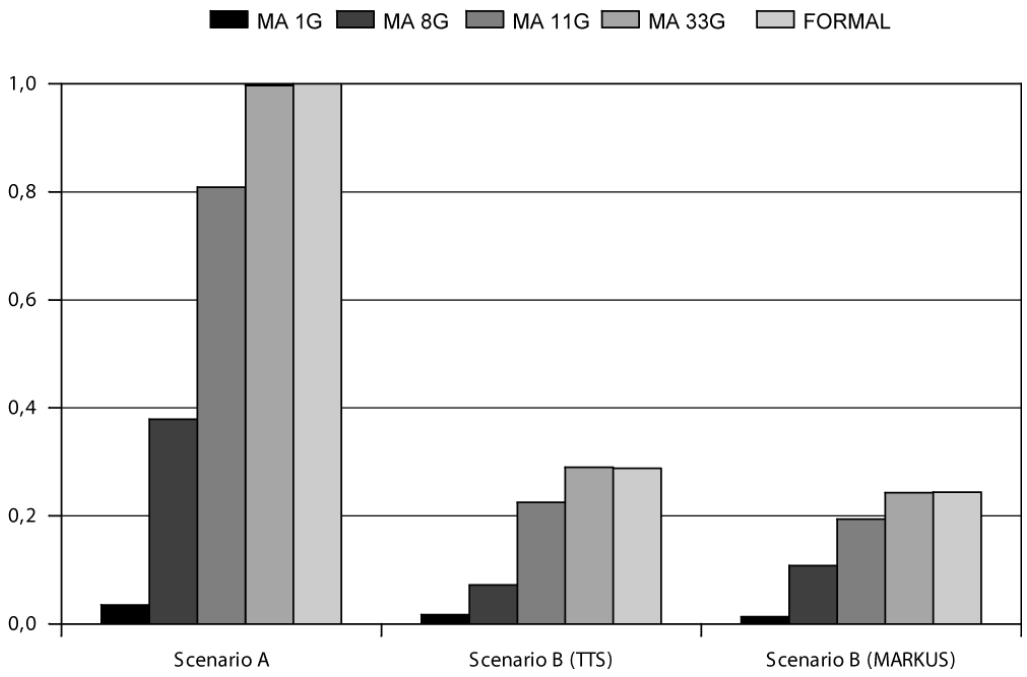

Abbildung 35.1: Enthüllungsrisiken verschiedener Szenarien

wobei der Stellparameter $\lambda \in [0, 1]$ individuell, abhängig von der Qualität bzw. Zuverlässigkeit des vorhandenen Zusatzwissens, gesetzt werden kann. Würde $\lambda = 1$ gewählt, so könnte allein absolut anonymisiertes Datenmaterial die Schutzwirkungsprüfungen passieren, während ein λ nahe bei 0 nahezu uneingeschränktes Vertrauen des Datenanbieters in seine realistischen Angriffssimulationen bedeutete. Der Datenanbieter wäre im letzt genannten Falle sicher, dass einem potenziellen Datenangreifer kein besseres als das in den Simulationen verwendete Zusatzwissen zur Verfügung stehen könnte. Um einen vernünftigen Schätzer für $\hat{P}_{B/\gamma}(w \text{ enthüllt})$ zu erhalten, sollte das realistische Szenario möglichst oft (unter Verwendung verschiedener Quellen möglichen Zusatzwissens) wiederholt werden.

Insgesamt wird dem Datenanbieter empfohlen, das Ergebnis des Worst-Case Szenarions nur geringfügig in das Gesamtrisikomaß einfließen zu lassen, da hier in der Regel sogar bei starker Anonymisierung sehr hohe Risiken zu erwarten sind, welche die jeweiligen Risiken bei den realistischen Szenarien um ein Vielfaches übersteigen. Allerdings kann das Worst-Case Szenario behilflich sein, a priori die Risikoverteilung auf die Daten zu erahnen. Besonders schutzbedürftige Bereiche ragen auch bei diesem Szenario hervor und können so im Vorfeld bei der Entwicklung eines ersten Anonymisierungskonzeptes etwas kritischer behandelt werden.

Wie in Kapitel 17 dargelegt wurde, reicht es im Allgemeinen nicht aus, ein globales Enthüllungsrisiko für die gesamte Zieldatei zu schätzen. Vielmehr sollte das Enthüllungsrisiko für eine gewissenhaft vorgenommene Zerlegung des Datenbestandes mit Bedacht geschätzt werden. Untenstehende Tabelle 35.11 und Abbildung 35.2 enthalten die Verteilung des Enthüllungsrisikos auf Beschäftigtengrößenklassen. Als Stellparameter wurde $\lambda = 0,2$ gewählt, wobei als Schätzer $\hat{P}_{B/\gamma}(w \text{ enthüllt})$ für die betrachteten Probeanonymisierungen das arithmetische Mittel der jeweiligen in 35.4.1 berechneten Enthüllungsrisiken (mit den MARKUS-Daten und der Umsatzsteuerstatistik als Zusatzwissen) bestimmt wurde.

Tabelle 35.11: Enthüllungsrisiken auf dem Niveau $\gamma = 0,05$ nach Beschäftigtengrößenklassen

Zieldaten	total	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000
MA1G	0,0191	0,0094	0,0064	0,0080	0,0174	0,0214	0,0366
MA8G	0,1278	0,1156	0,1112	0,1282	0,1426	0,1350	0,1464
MA11G	0,3474	0,3178	0,3194	0,3364	0,3426	0,3704	0,3528
MA33G	0,4130	0,3756	0,3840	0,4150	0,4611	0,5146	0,5447
FORMAL	0,4127	0,3744	0,3840	0,4136	0,4592	0,5104	0,5464

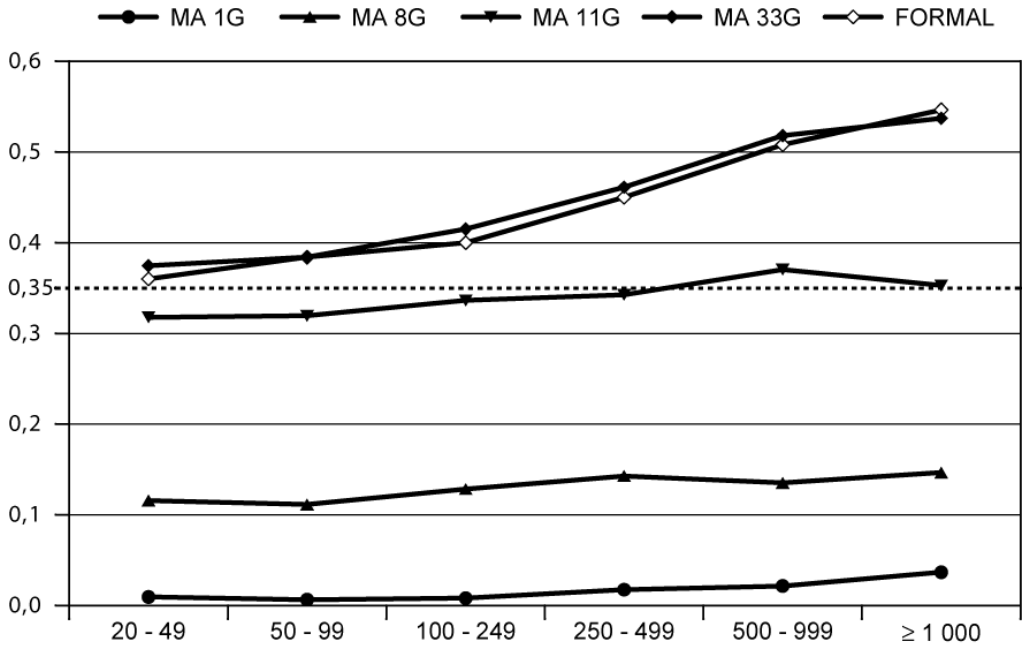

Abbildung 35.2: Vergleich der Schutzwirkung verschiedener Mikroaggregationsverfahren

Während auf der einen Seite mit zunehmender Beschäftigtenanzahl der Anteil korrekt zu-

geordneter Unternehmen steigt, sinkt auf der anderen Seite der Anteil an für den Datenangreifer brauchbaren Informationen, so dass der Zuwachs des Enthüllungsrisikos für größere Unternehmen leicht abgebremst wird. Diese Beobachtung wird besonders am Beispiel der Anonymisierungsvariante MA11G deutlich. Hier bleibt das Enthüllungsrisiko nahezu unverändert über aller Beschäftigtengrößenklassen hinweg und nimmt sogar in der obersten Beschäftigtengrößenklasse gegenüber der nächst kleineren Größenklasse etwas ab.

Liegt das Enthüllungsrisiko für alle Beschäftigtengrößenklassen unterhalb einer vorgegebenen Risikoschwelle τ, d.h.,

$$\hat{P}_\gamma(w \text{ enthüllt}) < \tau,$$

so kann die untersuchte Datei als faktisch anonym eingestuft werden. An dieser Stelle muss der Datenanbieter unter den Dateien, die dieser Bedingung genügen, diejenige mit dem besten Erhalt des Analysepotenzials auswählen. Zunächst aber müssen geeignete Nutzen- und Risikoschwellen γ_i und τ festgelegt werden. Bei einer Nutzenschwelle von $\gamma = 0.05$ für alle metrischen Merkmale und einer oberen Risikoschwelle von 35% (i.e. $\tau = 0.35$), könnten in unserem Beispiel die Anonymisierungsvarianten MA1G und MA8G als faktisch anonym eingestuft werden, während die Einzelwerte von Unternehmen der obersten Beschäftigtengrößenklasse bei der Variante MA11G noch leicht modifiziert werden müssten. Die Varianten MA33G und FORMAL würden bei dieser Schwellenwahl die Geheimhaltungsanforderungen nicht erfüllen. Bei der Erstellung eines Scientific-Use-Files haben sich die Fachleute später für eine globale Nutzenschwelle von $\gamma = 0.1$ und eine obere Risikoschwelle von $\tau = 0.5$ entschieden, wobei eine feinere Gliederung der Daten in Risikobereiche erfolgte (siehe Abschnitt 35.6) und auf das Worst-Case Szenario verzichtet wurde.

35.4.4 Vergleich mit der Verfahrensgruppe multiplikative stochastische Überlagerung

In diesem Abschnitt stellen wir den oben betrachteten Mikroaggregationsvarianten fünf Varianten der multiplikativen Zufallsüberlagerung (zur Beschreibung siehe 6.2.3) gegenüber.

Bei den hier betrachteten Varianten (vgl. Tabelle 35.12) wird mit f der Verschiebungsfaktor der beiden Mischungsverteilungen gegenüber Eins bezeichnet, der Parameter s bezeichnet die Standardabweichung der Überlagerungen. In einigen Varianten wurde nach der Überlagerung eine Varianzkorrektur durchgeführt. Diese Varianzkorrektur verbessert zwar die ersten beiden Momente der Merkmale, bewirkt aber einen größeren Abstand der Einzelwerte von ihren zugehörigen Originalwerten.

Nachfolgende Tabellen enthalten die Reidentifikations- (Tabelle 35.13) und Enthüllungsrisiken (Tabelle 35.14) der oben beschriebenen Varianten der Zufallsüberlagerung unter Verwendung der MARKUS-Datenbank als Zusatzwissen. Zum Vergleich sind die in den vorherigen Abschnitten untersuchten Varianten der Mikroaggregation ebenfalls aufgeführt. Die Tabellen enthalten je Anonymisierungsvariante die jeweilige Risikoverteilung auf Be-

Tabelle 35.12: Varianten multiplikativer stochastischer Überlagerung

Variante	s	f	Korrektur
Mult_f04_s02	0,020	0,04	nein
Mult_f08_s018	0,018	0,08	nein
Mult_f08_s018_trans	0,018	0,08	ja
Mult_f11_s03	0,030	0,11	nein
Mult_f11_s03_trans	0,030	0,11	ja

schäftigtengrößenklassen sowie das mit den Dateien verbundene Gesamtrisiko. Um eine lineare Ordnung zu erhalten, wurden die 10 Anonymisierungsvarianten nach dem Gesamtrisiko aufsteigend sortiert. Nach Einbeziehung der Brauchbarkeit der durch den Datenangreifer gefundenen Werte ergibt sich eine deutliche Reduktion der Risiken sowie eine leichte Verschiebung der vorherigen Sortierung (vgl. Tabelle 35.14).

Tabelle 35.13: Reidentifikationsrisiken aller Anonymisierungsvarianten.

Variante	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	gesamt
MA1G	0,022	0,015	0,031	0,058	0,090	0,167	0,038
Mult_f11_s03_trans	0,025	0,030	0,105	0,198	0,278	0,322	0,094
Mult_f08_s018_trans	0,029	0,044	0,133	0,235	0,308	0,353	0,114
Mult_f11_s03	0,087	0,093	0,130	0,187	0,240	0,336	0,132
Mult_f08_s018	0,108	0,124	0,153	0,229	0,315	0,360	0,162
Mult_f04_s02	0,122	0,156	0,208	0,285	0,354	0,411	0,198
MA8G	0,127	0,149	0,219	0,270	0,355	0,362	0,199
MA11G	0,156	0,192	0,265	0,344	0,413	0,440	0,243
MA33G	0,156	0,190	0,265	0,359	0,416	0,447	0,244
FORMAL	0,156	0,190	0,265	0,361	0,418	0,449	0,245

35.5 Überprüfung des Analysepotenzials

35.5.1 Vorgehensweise

Im Folgenden wird für unterschiedliche Varianten der stochastischen Überlagerung sowie für die getrennte abstandsorientierte Mikroaggregation (MA30G) untersucht, inwiefern die in Kapitel 18 festgelegten Abweichungsschwellen für deskriptive Maße eingehalten werden. Als Grundlage dient die in Abschnitt 35.3 festgelegte Vergröberung der kategorialen Merkmale. Zudem wird der Datensatz der KSE für 1999 ohne den Wirtschaftszweig 37 (Recycling) und ohne die Merkmale *Tätige Inhaber*, *Angestellte und Arbeiter* sowie *Bestandsveränderungen*

Tabelle 35.14: Enthüllungsrisiken aller Anonymisierungsvarianten

Variante	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	gesamt
MA1G	0,006	0,005	0,009	0,015	0,020	0,025	0,011
Mult_f11_s03_trans	0,007	0,008	0,031	0,050	0,078	0,107	0,026
Mult_f11_s03	0,023	0,022	0,025	0,031	0,034	0,038	0,029
Mult_f08_s018_trans	0,008	0,012	0,044	0,073	0,106	0,140	0,036
Mult_f08_s018	0,031	0,032	0,034	0,041	0,050	0,046	0,038
MA8G	0,070	0,089	0,131	0,154	0,192	0,163	0,109
Mult_f04_s02	0,094	0,119	0,154	0,210	0,254	0,291	0,148
MA11G	0,128	0,171	0,222	0,279	0,314	0,290	0,199
MA33G	0,154	0,188	0,265	0,262	0,412	0,438	0,243
FORMAL	0,156	0,190	0,265	0,361	0,418	0,449	0,245

an fertigen und unfertigen Erzeugnissen aus eigener Produktion zugrunde gelegt, so dass der Datensatz über 30 metrische Merkmale verfügt.

Auf die Darstellung von Ergebnissen für andere Varianten der Mikroaggregation wird verzichtet, da sich hier aus den bisherigen Ergebnissen schließen lässt, dass aus Sicht des Analysepotenzials lediglich die getrennte abstandsorientierte Mikroaggregation ein ausreichendes Analysepotenzial sicherstellen kann. Auf die Wiedergabe von Ergebnissen für additive stochastische Überlagerungen (nach dem Verfahren von Kim) wird verzichtet, da sich hier zum einen kein ausreichender Schutz von Großunternehmen sicherstellen lässt, zum anderen das Problem von Vorzeichenwechseln auftritt. Deshalb werden nur Varianten der multiplikativen Überlagerung dargestellt.

Zusätzlich zu den in Tabelle 35.12 aufgeführten Varianten der stochastischen Überlagerung wird die multiplikative stochstische Überlagerung nach dem Verfahren von Höhne getestet:

- Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne mit $f = 0,11$ und $s = 0,03$: Hoe_f11_s03
- Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne mit $f = 0,11$ und $s = 0,03$ und anschließender Transformation zum Erhalt der ersten beiden Momente: Hoe_f11_s03_trans
- Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne mit $f = 0,11$ und $s = 0,03$ und anschließende gruppenweise Kim-Korrektur. Dabei wurde der Datenbestand für jede Variable absteigend sortiert und die Korrektur in Blöcken von 100 Sätzen vorgenommen: Hoe_f11_s03_trans_grupp
- Multiplikative Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne mit $f = 0,11$ und $s = 0,03$ und anschließende gruppenweise Kim-Korrektur.

Dabei wurde der Datenbestand für jede Variable absteigend sortiert und die Korrektur in Blöcken von 100 Sätzen vorgenommen: Hoe_f11_s02_trans_grupp

Dabei werden die in Kapitel 18 definierten Kriterien überprüft:

- Abweichung der arithmetischen Mittel und Mediane im Gesamtdatensatz
- Abweichung der Standardabweichungen im Gesamtdatensatz
- Abweichung der Korrelationen und Rangkorrelationen (sowie Vorzeichenwechsel der Korrelationskoeffizienten) im Gesamtdatensatz
- Abweichung der arithmetischen Mittel, Mediane und Standardabweichungen nach Teilgesamtheiten
- t-Tests auf Mittelwertgleichheit nach Teilgesamtheiten

35.5.2 Überprüfung der einzelnen Abweichungsmaße

In den Tabellen 35.15 bis 35.22 sind die Ergebnisse für die in Kapitel 18 definierten Abweichungsmaße und die genannten Anonymisierungsverfahren dargestellt. Überschreitungen der in Kapitel 18 festgelegten Toleranzgrenze von 10% sind kursiv gekennzeichnet.

Tabelle 35.15: Durchschnittliche Veränderung der Verteilungsmaße

	Durchschnittliche relative Abweichung der arithmetischen Mittel (in %)		Standardabweichungen (in %)	Durchschnittliche absolute Abweichung der Korrelationskoeffizienten (mal 100)	Rangkorrelationen (mal 100)
	Ungewichtet	Gewichtet			
Mult_f04_s02	0,7	0,1	3,2	0,2	0
Mult_f08_s018	1,2	1,1	5,9	0,3	0,1
Mult_f08_s018_trans	0	3	0	0,3	0,1
Mult_f11_s03	1,7	1,5	8,3	0,5	0,2
Mult_f11_s03_trans	0	4,2	0	0,5	0,2
Hoe_f11_s03_trans	0,1	0,2	0,1	0,8	0,3
Hoe_f11_s03	0	5,2	2,6	0,8	0,2
Hoe_f11_s03_trans_grupp	0,1	0,4	0,2	0,6	0,3
Hoe_f11_s02_trans_grupp	0,1	0,5	0,2	0,6	0,3
MA30G	0	0	3,2	2,4	0

Tabelle 35.16: Veränderung der univariaten Verteilungsmaße

	Anteil der arithmetischen Mittel, die um mehr als 10% abweichen (in %)	Anteil der Mediane, die um mehr als 10% abweichen (in %)	Anteil der Standardabweichungen, die um mehr als 10% abweichen (in %)
Mult_f04_s02	0	0	0
Mult_f08_s018	0	0	0
Mult_f08_s018_trans	0	*63,3*	0
Mult_f11_s03	0	3,3	*40,0*
Mult_f11_s03_trans	0	*63,3*	0
Hoe_f11_s03_trans	0	3,3	0
Hoe_f11_s03	0	3,3	3,3
Hoe_f11_s03_trans_grupp	0	3,3	0
Hoe_f11_s02_trans_grupp	0	3,3	0
MA30G	0	0	*13,3*

Anteile bezogen auf die Anzahl der Merkmale

Tabelle 35.17: Veränderung der Korrelationen

	Anteil der Korrelationskoeffizienten, die um mehr als 0,1 abweichen	Anteil der Vorzeichenwechsel bei den Korrelationskoeffizienten	Anteil der Rangkorrelationen, die um mehr als 0,05 abweichen	Anteil der Vorzeichenwechsel bei den Rangkorrelationen
Mult_f04_s02	0	0,5	0	0
Mult_f08_s018	0	0,9	0	0
Mult_f08_s018_trans	0	0,9	0	0
Mult_f11_s03	0	1,2	0	0
Mult_f11_s03_trans	0	1,6	0	0
Hoe_f11_s03_trans	0	0,9	0	0
Hoe_f11_s03	0	0,7	0	0
Hoe_f11_s03_trans_grupp	0	1,2	0	0
Hoe_f11_s02_trans_grupp	0	1,4	0	0
MA30G	2,8	0,7	0	0

Anteile bezogen auf die insgesamt berechneten Korrelationskoeffizienten

Tabelle 35.18: Veränderung der Verteilungsmaße nach Wirtschaftszweigen

	Arithmetische Mittel, ungewichtet (Anteil der Fälle mit einer Abweichung von über 10%)	Arithmetische Mittel, gewichtet (Anteil der Fälle mit einer Abweichung von über 10%)	Mediane (Anteil der Fälle mit einer Abweichung von über 10%)	Standardabweichungen (Anteil der Fälle mit einer Abweichung von über 10%)
Mult_f04_s02	0	0	0	0
Mult_f08_s018	0	0	0,4	0
Mult_f08_s018_trans	7,8	*26,1*	*59,2*	*10,4*
Mult_f11_s03	7,8	0,6	*59,2*	*14,5*
Mult_f11_s03_trans	*16,1*	*32,9*	*63,5*	*25,3*
Hoe_f11_s03_trans	0,6	0,6	1,6	*14,3*
Hoe_f11_s03	0,2	0,2	3,3	*13,5*
Hoe_f11_s03_trans_grupp	0,8	*2,7*	1,4	9,4
Hoe_f11_s02_trans_grupp	0,4	0,8	1,4	8,4
MA30G	3,9	0,4	0	9,6

Anteile bezogen auf die Anzahl Merkmale multipliziert mit den betrachteten Teilgesamtheiten

Tabelle 35.19: Veränderung der Verteilungsmaße nach Ost/West

	Arithmetische Mittel, ungewichtet (Anteil der Fälle mit einer Abweichung von über 10%)	Arithmetische Mittel, gewichtet (Anteil der Fälle mit einer Abweichung von über 10%)	Mediane (Anteil der Fälle mit einer Abweichung von über 10%)	Standardabweichungen (Anteil der Fälle mit einer Abweichung von über 10%)
Mult_f04_s02	0	0	1,7	0
Mult_f08_s018	0	0	1,7	0
Mult_f08_s018_trans	5,0	*28,3*	*61,7*	1,7
Mult_f11_s03	5,0	0	*61,7*	*23,3*
Mult_f11_s03_trans	8,3	*31,7*	*61,7*	3,3
Hoe_f11_s03_trans	0	0	1,7	1,7
Hoe_f11_s03	0	0	1,7	5,0
Hoe_f11_s03_trans_grupp	0	0	1,7	1,7
Hoe_f11_s02_trans_grupp	0	0	5,0	1,7
MA30G	0	0	0	8,3

Anteile bezogen auf die Anzahl Merkmale multipliziert mit den betrachteten Teilgesamtheiten

Tabelle 35.20: Veränderung der Verteilungsmaße nach Ost/West und Wirtschaftszweigen

	Arithmetische Mittel, ungewichtet (Anteil der Fälle mit einer Abweichung von über 10%)	Arithmetische Mittel, gewichtet (Anteil der Fälle mit einer Abweichung von über 10%)	Mediane (Anteil der Fälle mit einer Abweichung von über 10%)	Standardabweichungen (Anteil der Fälle mit einer Abweichung von über 10%)
Mult_f04_s02	0	0	0	0
Mult_f08_s018	0,1	0,1	0,6	1,9
Mult_f08_s018_trans	2,2	*36,4*	*60,9*	*16,2*
Mult_f11_s03	*21,4*	1,9	*60,9*	*19,9*
Mult_f11_s03_trans	*28,5*	*44,6*	*63,1*	*29,2*
Hoe_f11_s03_trans	2,5	2,2	4,7	*20,2*
Hoe_f11_s03	1,8	1,3	6,8	*17,4*
Hoe_f11_s03_trans_grupp	2,3	2,2	4,4	*15,5*
Hoe_f11_s02_trans_grupp	1,7	1,5	4,0	*13,8*
MA30G	2,3	2,1	0	5,2

Anteile bezogen auf die Anzahl Merkmale multipliziert mit den betrachteten Teilgesamtheiten

Tabelle 35.21: t-Tests auf Mittelwertgleichheit (ungewichtet) für Teilgesamtheiten

	Anteil der Fälle (in %) mit einer signifikanten Abweichung der arithmetischen Mittel zum Signifikanzniveau von 10%		
	Nach Wirtschaftszweigen	Nach Ost/West	Nach Wirtschaftszweigen und Ost/West
Mult_f04_s02	4,7	*20,0*	4,5
Mult_f08_s018	4,9	*25,0*	4,4
Mult_f08_s018_trans	*47,8*	*31,7*	*50,1*
Mult_f11_s03	4,7	*21,7*	4,3
Mult_f11_s03_trans	*49,2*	*31,7*	*50,3*
Hoe_f11_s03_trans	*12,2*	8,3	9,4
Hoe_f11_s03	*12,7*	1,7	*11,0*
Hoe_f11_s03_trans_grupp	*13,3*	*13,3*	*11,6*
Hoe_f11_s02_trans_grupp	*13,1*	*11,7*	*11,8*
MA30G	5,7	8,3	6,20%

Anteile bezogen auf die Anzahl Merkmale multipliziert mit den betrachteten Teilgesamtheiten

Tabelle 35.22: t-Tests auf Mittelwertgleichheit (gewichtet) für Teilgesamtheiten

	Anteil der Fälle (in %) mit einer signifikanten Abweichung der arithmetischen Mittel zum Signifikanzniveau von 10%		
	Nach Wirtschaftszweigen	Nach Ost/West	Nach Wirtschaftszweigen und Ost/West
Mult_f04_s02	*6,7*	*18,3*	5,5
Mult_f08_s018	6,5	*21,7*	5,5
Mult_f08_s018_trans	*62,5*	*80,0*	*65,8*
Mult_f11_s03	6,3	*18,3*	5,3
Mult_f11_s03_trans	*64,9*	*80,0*	*65,6*
Hoe_f11_s03_trans	*12,7*	*11,7*	*11,1*
Hoe_f11_s03	*11,2*	1,7	*10,5*
Hoe_f11_s03_trans_grupp	*14,1*	*15,0*	*11,3*
Hoe_f11_s02_trans_grupp	*52,0*	*35,0*	*45,8*
MA30G	5,9	8,3	6,8

Anteile bezogen auf die Anzahl Merkmale multipliziert mit den betrachteten Teilgesamtheiten

Die wesentlichen Ergebnisse aus den Tabellen 35.15 bis 35.22 sind die folgenden:

- Überschreitungen der Abweichungsschwellen treten vorrangig bei denjenigen Varianten der multiplikativen stochastischen Überlagerung auf, bei denen die Werte nach der Überlagerung derart transformiert werden, dass die ersten beiden Momente im Gesamtdatensatz erhalten bleiben. Der Grund hierfür ist, dass durch die varianzkorrigierende Transformation zusätzliche Verzerrungen der Einzelwerte bewirkt werden (teilweise werden Werte sogar negativ). Dies wirkt sich negativ auf den Erhalt der Mittelwerte von Teilgesamtheiten aus. Besonders stark werden offenbar kleinere Unternehmen verzerrt. Dies erklärt, warum die Abweichungen bei diesen Verfahren steigen, wenn man die Hochrechnungsfaktoren berücksichtigt. Dieser Effekt ist bei der blockweisen Kim-Korrektur deutlich geringer. Dennoch treten auch hier in nicht zu vernachlässigendem Umfang Vorzeichenwechsel auf.
- Besonders die Mediane sind von der stochastischen Überlagerung tangiert.
- Grundsätzlich ergibt sich bei einer höheren Varianz der multiplikativen Überlagerungen auch ein höheres Risiko eine Abweichungsschwelle zu überschreiten.
- Neben der abstandsorientierten getrennten Mikroaggregation schneiden die Überlagerungen mit konstanten Faktoren und $f = 0,04/s = 0,02$ bzw. $f = 0,08/s = 0,018$ sowie der Ansatz von Höhne jeweils ohne Transformation am besten ab.

Aufgrund dieser Ergebnisse sowie der generellen Bewertung der im Projekt untersuchten Anonymisierungsverfahren kommen zur Anonymisierung der Kostenstrukturerhebung im Verarbeitenden Gewerbe aus Sicht des Analysepotenzials grundsätzlich zwei Arten von datenverändernden Verfahren in Frage:

- Eine getrennte Mikroaggregation mit einer variablen Gruppengröße von 3 bis 5
- Eine multiplikative stochastische Überlagerung mit einem konstanten Faktor aus einer zweigipfligen Mischungsverteilung oder einer Mischungsverteilung nach dem Verfahren von Höhne jeweils ohne Transformation zum Erhalt der ersten und zweiten Momente.

Mit Hilfe dieser Verfahren kann die Erstellung von Scientific-Use-Files in Angriff genommen werden. Besteht ggf. weiterer Anonymisierungsbedarf, so sollte diesem durch eine Vergröberung der kategorialen Variablen entsprochen werden.

Bei den Untersuchungen zur Datensicherheit wurde deutlich, dass eine Anonymisierung der großen Unternehmen in der KSE mittels der getrennten abstandsorientierten Mikroaggregation nicht möglich ist. Deshalb kann ein Scientific-Use-File für die KSE nur auf der Basis einer multiplikativen stochastischen Überlagerung mit einer Mischungsverteilung erstellt werden. Dabei dürfte das Verfahren von Höhne den größeren Schutz erzeugen, weil

hier auch die Quotienten zweier Merkmale nicht exakt erhalten werden. Zudem spricht für das Verfahren von Höhne auch, dass hier in jedem Fall eine SIMEX-Korrektur angewendet werden kann, während dies bei Überlagerung mit einem konstanten Faktor nicht der Fall ist. Bei einer multiplikativen Überlagerung sollten den Nutzern dann auch in der gleichen Weise anonymisierte Instrumente zur Verfügung gestellt werden. Für die kleinen und mittleren Unternehmen der KSE wurde bereits im Projektverlauf ein Scientific-Use-File erstellt, bei dem die getrennte abstandsorientierte Mikroaggregation zum Einsatz kam. Dieser wird im folgenden Abschnitt dargestellt und in Unterabschnitt 35.6.3 hinsichtlich seines Analysepotenzials gesondert untersucht.

35.6 Ein Scientific-Use-File der Kostenstrukturerhebung im Verarbeitenden Gewerbe

In diesem Abschnitt wird eine Anonymisierung der Kostenstrukturerhebung für kleinere und mittlere Unternehmen vorgestellt, die sich als Scientific-Use-File durchgesetzt hat. Um einerseits möglichst auf datenverändernde Verfahren verzichten, aber andererseits ein akzeptables Schutzniveau erreichen zu können, wurden zunächst traditionelle Verfahren auf die Daten angewendet. Diese bestanden im Wesentlichen in der Entfernung von Merkmalen, die für den Wissenschaftler verzichtbar schienen (wie z.B. das charakterisierende Merkmal *Tätige Inhaber*), und der Vergröberung von Merkmalen so weit, dass die für den Wissenschaftler relevanten Teilmassenauswertungen noch möglich bleiben. Aus der Gruppe der datenverändernden Verfahren wurde die schwächste Form der Mikroaggregation, bei der jedes metrische Merkmal separat mikroaggregiert wird, angewendet.

Als kritische Bereiche wurden Kombinationen der drei Merkmale *Ost-West*, *Beschäftigtengrößenklasse* und *Wirtschaftszweigklassifikation* angesehen. Ein durch den Datenangreifer gefundener Einzelwert wurde als unbrauchbar betrachtet, wenn er relativ um wenigstens $\gamma = 0,1$ von dem dazugehörigen Originalwert abweichte. Als obere Risikoschwelle wurde $\tau = 0,5$ gesetzt, was bedeutet, dass für jedes Tripel von Ausprägungen der drei Merkmale, wie z.B. (*Ost-West*=„Ost“; *Beschäftigtengrößenklasse*=„100-249“; *Wirtschaftszweigklassifikation*=„20“), ein Enthüllungsrisiko für Einzelinformationen von unter 50% bestehen muss und damit ein potenzieller Datenangreifer bei demselben Vorgehen mit einer Wahrscheinlichkeit von über 50% einen unbrauchbaren Einzelwert finden würde. Zur Beurteilung der Vertraulichkeit wurden auch hier Szenarien mit der Umsatzsteuerstatistik und der MARKUS-Datenbank durchgeführt. Das Worst-Case Szenario wurde in diesem Falle nicht berücksichtigt, da für die realistischen Szenarien zwei qualitativ hochwertige Datenquellen zur Verfügung standen.

35.6.1 Informationsreduzierende Maßnahmen

In einem ersten Schritt wurde auf das ursprünglich im Merkmalskanon vorhandene Merkmal *Tätige Inhaber* verzichtet, da es sich im Laufe der Projektarbeiten als besonders reidentifikationsgefährdend und für wissenschaftliche Analysen als wenig wertvoll herausgestellt hat. Das Merkmal *Angestellte und Arbeiter* wurde daher als Differenz aus *Tätige Personen insgesamt* und *Tätige Inhaber* ebenfalls aus dem ursprünglichen Merkmalskanon herausgenommen. Das Merkmal *Bestandsveränderungen* wurde entfernt, da es aus inhaltlichen Gründen verzichtbar war.

Besonders geeignet für Reidentifikationen sind regionale Angaben. Der Erhalt solcher Merkmale in einem Scientific-Use-File stellt daher für die Anonymisierung ein schwieriges Unterfangen dar. Bereits zu Beginn des Projekts wurde die Möglichkeit ausgeschlossen, einen Scientific-Use-File zu erstellen, der administrative Gebietsangaben auf der Ebene der Bundesländer oder gar einer tieferen Gliederungsebene enthält. Da aber die Auswertung nach Regionen einen wichtigen Analysebereich darstellt, wurde nach alternativen Möglichkeiten gesucht und dies vor dem Hintergrund, gleichzeitig auf datenverändernde Maßnahmen (insbesondere bei den metrischen Merkmalen) weitgehend verzichten zu wollen. Als erste Möglichkeit wurde der administrative Gebietsschlüssel durch den nichtadministrativen siedlungsstrukturellen Kreistyp BBR9 und den durch weitere Vergröberung erhältlichen siedlungsstrukturellen Regionstyp BBR3 ersetzt. Die sehr deutliche Verbesserung der Schutzwirkung durch diese Vergröberung der Regionalinformation wurde in (Lenz und Vorgrimler 2005) festgestellt. Allerdings sprach sich der Wissenschaftliche Begleitkreis des Projekts „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ dafür aus, anstelle dieser nicht-administrativen Schlüssel die Ost-West-Klassifizierung einzuführen. Diese zweite Möglichkeit wurde schließlich in den Scientific-Use-File aufgenommen.

Die Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe wurden nach der Klassifikation der Wirtschaftszweige (WZ93) auf der Vierstellerebene (Klasse) erhoben und aufbereitet. Diese Klassifikation ist von der europäischen Klassifikation NACE Rev.1 abgeleitet, die aufgrund der NACE-Verordnung des Rates der Europäischen Gemeinschaften seit 1995 in allen Mitgliedstaaten der Europäischen Union sowohl für die Erhebung als auch für die Darstellung der statistischen Daten anzuwenden ist.[68] Das Kodierungssystem der WZ93 unterscheidet zwischen Abschnitten (Buchstaben A-Q), Unterabschnitten (Buchstaben AA-QA), Abteilungen (Zweisteller), Gruppen (Dreisteller), Klassen (Viersteller) und Unterklassen (Fünfsteller). Der Wirtschaftsbereich „Verarbeitendes Gewerbe sowie Bergbau und Gewinnung von Steinen und Erden“ erstreckt sich über die Abschnitte C und D bzw. – in der numerischen Gliederung – über die Abteilungen 10 bis 37. Im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ haben sich Datenschützer und Datennutzer bei dem hierarchischen Merkmal WZ93 auf die Gliederungstiefe 2 (Zweistellerebe-

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

ne, vereinzelt auch Unterabschnitte) verständigt, da hierdurch zum einen eine beachtliche Schutzwirkung und zum anderen nach Einschätzung der beteiligten Wissenschaftler für einen Scientifc-Use-File eine ausreichende Breite an Analysemöglichkeiten erhalten wird.

In den Veröffentlichungen der statistischen Ämter werden aufgrund von Geheimhaltungsaspekten die Ergebnisse einiger Wirtschaftsabteilungen nicht veröffentlicht. Es handelt sich dabei um Unternehmen der Abteilungen 10, 11, 14, 16, 23, 30, 32, 35 und 37 der WZ93. Bei den im Projekt durchgeführten Simulationen hat sich bestätigt, dass diese Abteilungen neben den Abteilungen 15, 17, 18, 19 , 22 und 34 größerer Geheimhaltung bedürfen. Um diese kritischen Abteilungen im Scientific-Use-File belassen und weitgehend auf datenverändernde Verfahren bei den quantitativen Merkmalen verzichten zu können, wurden die Abteilungen 10 (Kohlenbergbau, Torfgewinnung), 11 (Gewinnung von Erdöl und Erdgas, Erbringung damit verbundener Dienstleistungen) und 14 (Gewinnung von Steinen und Erden, sonstiger Bergbau) zum Abschnitt C, die Abteilungen 15 (Ernährungsgewerbe) und 16 (Tabakverarbeitung) zum Unterabschnitt DA, die Abteilungen 17 (Textilgewerbe) und 18 (Bekleidungsgewerbe) zum Unterabschnitt DB, die Abteilungen 21 (Papiergewerbe) und 22 (Verlags- und Druckgewerbe, Vervielfältigung) zum Unterabschnitt DE, die Abteilungen 30 (Herstellung von Büromaschinen, Dv-Geräten und -einrichtungen) und 31 (Herstellung von Geräten der Elektrizitätserzeugung, -verteilung u.ä.) zum Unterabschnitt DL sowie die Abteilungen 34 (Herstellung von Kraftwagen und Kraftwagenteilen) und 35 (Sonstiger Fahrzeugbau) zum Unterabschnitt DM zusammengefasst. Bei den Abteilungen 19 (Ledergewerbe) und 23 (Kokerei, Mineralölverarbeitung, Herstellung von Brutstoffen) wurde das Merkmal WZ93 unterdrückt. Außerdem wurde die Abteilung 37 (Recycling) aus inhaltlichen und aus Geheimhaltungsgründen herausgenommen. Obwohl die Abteilungen 32 (Rundfunk-, Fernseh- u. Nachrichtentechnik) und 33 (Medizin-, Mess-, Steuer- u. Regelungstechnik, Optik) ebenfalls zum Unterabschnitt DL zu zählen sind, werden Sie im Datensatz separat aufgeführt, da hier die Weitergabe der Zweisteller aus Sicht des Datenschutzes unbedenklich ist. Eine zusammenfassende Aufstellung der im Datensatz vorhandenen Ausprägungen des Merkmals WZ93 findet sich in Tabelle 35.23.

35.6.2 Eindimensionale getrennte Mikroaggregation

Die im Datensatz verbliebenen 30 metrischen Merkmale wurden eindimensional für jedes Merkmal separat mikroaggregiert. Bei dieser Variante der Mikroaggregation werden zunächst die Merkmalsausprägungen je Merkmal absteigend sortiert. Dann werden tripelweise (aus den Merkmalsausprägungen dreier benachbarter Merkmalsträger) die Durchschnittswerte ermittelt, die Originalwerte durch diese Durchschnittswerte ersetzt und wieder an die ursprüngliche Position zurücksortiert (eine detailliertere Beschreibung findet sich in Unterabschnitt 6.2.4). Falls die Anzahl der Merkmalsträger nicht durch die Zahl drei teilbar ist, so ist am Ende der absteigend sortierten Liste von Merkmalsausprägungen auch die Bildung einer Gruppe aus vier oder fünf Merkmalsträgern zulässig. Damit ist jede Merkmalsausprägung bei mindestens drei Merkmalsträgern vorhanden. Das hier skizzierte Verfahren

Tabelle 35.23: Verteilung von Gesamtumsatz und Beschäftigten auf Wirtschaftsabteilungen

Wirtschaftsgliederung	WZ 93 -Angabe	Unternehmen	% Beschäftigte	% Gesamtumsatz
Bergbau und Gew. von Steinen u. Erden	C (10, 11 und 14)	182	1,02	1,17
Ernährungsgewerbe u. Tabakverarbeitung	DA (15 und 16)	1.704	12,90	20,70
Textil- u. Bekleidungs-gewerbe	DB (17 und 18)	955	6,91	5,91
Holzgewerbe (ohne H. v. Möbeln)	20	448	2,81	2,61
Papier-, Verlags- u. Druckgewerbe	DE (21 und 22)	1.063	8,49	8,41
Chemische Industrie	DG (bzw. 24)	628	5,06	7,84
H. v. Gummi- u. Kunststoffwaren	DH (bzw. 25)	659	5,12	4,24
Glasgewerbe, Keramik, Ver. v. Steinen u. Erden	DI (bzw. 26)	783	5,47	5,45
Metallerzg. und -bearbeitung	27	423	3,73	4,46
H. v. Metallerzeugnissen	28	1.638	11,94	9,12
Maschinenbau	DK (bzw. 29)	1.837	14,07	11,57
H. v. Büromasch., Dv-Geräten u. einr., Gerät. d. Elektriz.erzg., -verteilung u. ä.	DL (hier 30 und 31)	757	5,88	5,05
Rundfunk-, Fernseh- u. Nachrichtentechnik	32	219	1,69	1,39
Medizin-, Mess, Steuer- und Regelungstechnik, Optik	33	510	3,71	2,76
Fahrzeugbau	DM (34 und 35)	525	4,24	3,85
H.v.Möbeln,Schmuck, Musikinstrumenten, Sportger. usw	36	731	5,69	4,01
Sonstige	19 und 23	164	1,26	1,46

ist für das Ziel einer möglichst vielseitigen Datennnutzung, sowohl für deskriptive als auch für ökonometrische Auswertungen, das schonendste Verfahren innerhalb der Klasse der Mikroaggregationsverfahren.

35.6.3 Analysepotenzial des Scientific-Use-Files

Da sich aus Sicht des Analysepotenzials die getrennte abstandsorientierte Mikroaggregation grundsätzlich zur Erstellung eines Scientific-Use-Files eignet und hiermit auch in Verbindung mit der Vergröberung der Wirtschaftszweige für die kleinen und mittleren Unternehmen der KSE (bis 250 Beschäftigte) die faktische Anonymität sichergestellt wird, wurde mit diesem Verfahren ein Vorschlag für einen Scientific-Use-File entwickelt, dessen Analysepotenzial im Folgenden bewertet wird.

a) Deskriptive Auswertungen

Der Vorschlag eines Scienfic-Use-Files für die kleinen und mittleren Unternehmen der KSE wird im Folgenden anhand der in Kapitel 18 definierten Abweichungsmaße bewertet. Die Ergebnisse finden sich in den Tabellen 35.24 bis 35.28, Tabelle 35.29 enthält ergänzend beispielhafte Auswertungen nach Beschäftigtengrößenklassen.

Tabelle 35.24: Durchschnittliche Veränderung der Verteilungsmaße

Durchschnittliche relative Abweichung der				Durchschnittliche absolute Abweichung der	
arithmetischen Mittel		Varianzen	Standard-	Korrelations-	Rangkorre-
Ungewichtet	Gewichtet		abweichungen	koeffizienten	lationen
(in %)	(in %)	(in %)	(in %)	(mal 100)	(mal 100)
0,1	0,2	4,3	2,3	0, 786	0, 0166

Tabelle 35.25: Veränderung der univariaten Verteilungsmaße

Anteil der arithmetischen Mittel mit einer Abweichung von mehr als 10%		Anteil der Mediane mit einer Abweichung von mehr als 10%	Anteil der Standard-abweichungen mit einer Abweichung von mehr als 10%	Anteil der Varianzen mit einer Abweichung von mehr als 30%
(in %)		(in %)	(in %)	(in %)
ungewichtet	gewichtet			
0	0	3,3	6,6	3,3

Anteile bezogen auf die Anzahl der Merkmale

Tabelle 35.26: Veränderung der Korrelationen

Anteil der Korrelationskoeffizienten (in %) mit		Anteil der Rangkorrelationen (in %) mit	
einer Abweichung von mehr als 0,1	einem Vorzeichenwechsel	einer Abweichung von mehr als 0,05	einem Vorzeichenwechsel
1,1	0,2	0	0

Anteile bezogen auf die Anzahl der jeweils berechneten Korrelationskoeffizienten

Tabelle 35.27: Veränderung der Verteilungsmaße für Teilgesamtheiten

	Anteil der arithmetischen Mittel (ungewichtet)	Anteil der arithmetischen Mittel (gewichtet)	Anteil der Standardabweichungen	Anteil der Mediane
	mit einer Abweichung von mehr als 10%			
	(in %)	(in %)	(in %)	(in %)
Beschäftigtengrößenklassen	0	0	14,4	1,1
Ost/West	0	0	10	1,7
Wirtschaftszweigen	1,8	1,8	6,9	1,4
Beschäftigtengrößenklassen und Ost/West	0,5	0,5	10,5	0,5
Beschäftigtengrößenklassen und Wirtschaftszweigen	1,8	1,6	3,1	1,5
Ost/West und Wirtschaftszweigen	1,5	1,5	3,5	1,7
Beschäftigtengrößenklassen, Ost/West und Wirtschaftszweigen	1,3	1,3	1,8	1,6

Anteile bezogen auf die Anzahl der betrachteten Merkmale multipliziert mit der Anzahl der Teilgesamtheiten

Alle Toleranzgrenzen für die Überschreitung von Abweichungsschwellen werden mit dem Vorschlag für einen Scientific-Use-File eingehalten. Eine Überschreitung der Abweichungsgrenze von 10% bei den arithmetischen Mitteln tritt überproportional bei den folgenden Variablen auf:

- *Teilzeitbeschäftigte in Vollzeiteinheiten*
- *Anfangs- und Endbestand an Handelsware*
- *Energieverbrauch*
- *Kosten für Reparaturen*

Bei den in Tabelle 35.28 dargestellten t-Tests erweist sich überwiegend die Variable *Teilzeitbeschäftigte in Vollzeiteinheiten* als problematisch.

Tabelle 35.28: t-Tests auf Mittelwertgleichheit für Teilgesamtheiten

	Anteil der signifikanten Abweichungen zum Signifikanzniveau von 10% (in %)	
	ungewichtet	gewichtet
Beschäftigten-größenklassen	6,7	6,7
Ost/West	3,3	8,3
Wirtschaftszweige	8,6	8,2
Beschäftigtengrößen-klassen und Ost/West	8,3	8,3
Beschäftigtengrößen-klassen und Wirtschaftszweige	8,2	8,5
Ost/West und Wirtschaftszweige	7,7	7,8
Beschäftigtengrößen-klassen, Ost/West und Wirtschaftszweige	7,6	8,3

Anteile bezogen auf die Anzahl der betrachteten Merkmale multipliziert mit der Anzahl der Teilgesamtheiten

b) Ökonometrische Schätzungen

Beispielhaft wird mit den als Scientific-Use-File vorgesehenen Daten die linearisierte Cobb-Douglas-Produktionsfunktion geschätzt. In Ergänzung zum bisherigen Vorgehen wird diese nun auch für einzelne Wirtschaftsbereiche sowie für den Gesamtdatensatz unter Berücksichtigung von Dummy-Variablen für die einzelnen Wirtschaftszweige ermittelt. Dabei wird jeweils eine Ausreißerbereinigung vorgenommen. Tabelle 35.30 zeigt die Schätzergebnisse für den gesamten Scientific-Use-File, in Tabelle 35.31 sind die Schätzergebnisse für den Fahrzeugbau, in Tabelle 35.32 sind die Ergebnisse für den Maschinenbau dargestellt. Tabelle 35.33 zeigt die Schätzergebnisse wiederum für den gesamten Datensatz, allerdings unter Berücksichtigung des Einflusses der Wirtschaftszweige.

Die Auswertungen zeigen insgesamt, dass ein gutes Analysepotenzial gewährleistet ist.

Tabelle 35.29: Beispielhafte Auswertungen nach Beschäftigtengrößenklassen

Beschäftigte	20-49		50-99		100-249	
	Original	Anonymisiert	Original	Anonymisiert	Original	Anonymisiert
Anzahl der Unternehmen	5.233	5.232	4.100	4.101	3.894	3.894
Durchschnittliche Beschäftigung	33,3	33,3	69,9	69,9	155,0	155,0
Durchschnittliche Beschäftigung (gewichtet)	33,4	33,3	70,0	69,9	154,6	154,6
Durchschnittlicher Gesamtumsatz	8.464.911	8.467.894	19.804.980	19.896.130	48.481.890	48.378.980
Durchschnittlicher Gesamtumsatz (gewichtet)	7.005.913	7.006.741	17.418.050	17.486.810	44.712.610	44.661.460
Durchschnittliche Anzahl der Teilzeitbeschäftigten in Vollzeiteinheiten	1,74	1,86	3,18	3,29	5,68	5,79
Durchschnittliche Anzahl der Teilzeitbeschäftigten in Vollzeiteinheiten (gewichtet)	1,78	1,90	3,17	3,28	5,70	5,80
Durchschnittlicher Anfangsbestand an Handelsware	0,08098	0,08277	0,08994	0,09139	0,10292	0,09900
Durchschnittliche Kosten für Reparaturen	144.179,1	144.164,7	334.584,9	334.893,3	908.904,2	908.606,7
Durchschnittliche Kosten für Reparaturen (gewichtet)	119.881,0	119.883,1	306.340,1	306.375,8	823.407,4	821.030,8

Tabelle 35.30: Linearisierte Cobb-Douglas-Produktionsfunktion – OLS-Regression, robuste Standardfehler (um Ausreißer bereinigt) für die kleineren und mittleren Unternehmen der KSE

	Originalwerte		Anonymisiert	
Inputfaktoren	Koeffizienten	(t-Werte)	Koeffizienten	(t-Werte)
Materialeinsatz	0,430	(128,84)	0,429	(127,10)
Personalkosten	0,331	(82,32)	0,333	(81,71)
Externe Dienstleistungen	0,051	(34,70)	0,051	(34,64)
Sonstige Kosten	0,102	(39,80)	0,103	(39,76)
Kapitalkosten	0,071	(28,95)	0,070	(28,87)
Konstante	1,698	(54,42)	1,698	(53,84)
Anzahl der Beobachtungen	11.657		11.678	
Bestimmtheitsmaß	0,972		0,972	
Durchschnittliche Abweichung der Koeffizientenwerte			0,5%	
Anzahl der Vorzeichenwechsel			0	
Anzahl der Koeffizienten, die im Original statistisch signifikant sind, bei Anonymisierung nicht			0	
Anzahl der Koeffizienten, die bei Anonymisierung statistisch signifikant sind, im Original nicht			0	
Anzahl der Wechsel zwischen unterschiedlichen Signifikanzniveaus			0	
Durchschnittliche Abweichung der t-Werte			0,6%	

Tabelle 35.31: Linearisierte Cobb-Douglas-Produktionsfunktion Fahrzeugbau – OLS-Regression für die kleineren und mittleren Unternehmen der KSE, robuste Standardfehler (um Ausreißer bereinigt)

	Originalwerte		Anonymisiert	
Inputfaktoren	Koeffizienten	(t-Werte)	Koeffizienten	(t-Werte)
Materialeinsatz	0,417	(23,41)	0,416	(23,45)
Personalkosten	0,365	(15,63)	0,366	(15,71)
Externe Dienstleistungen	0,054	(7,14)	0,054	(7,18)
Sonstige Kosten	0,074	(5,34)	0,073	(5,29)
Kapitalkosten	0,067	(5,42)	0,069	(5,67)
Konstante	1,779	(11,00)	1,778	(11,07)
Anzahl der Beobachtungen	460		460	
Bestimmtheitsmaß	0,96		0,96	
Durchschnittliche Abweichung der Koeffizientenwerte			0,8%	
Anzahl der Vorzeichenwechsel			0	
Anzahl der Koeffizienten, die im Original statistisch signifikant sind, bei Anonymisierung nicht			0	
Anzahl der Koeffizienten, die bei Anonymisierung statistisch signifikant sind, im Original nicht			0	
Anzahl der Wechsel zwischen unterschiedlichen Signifikanzniveaus			0	
Durchschnittliche Abweichung der t-Werte			1,2%	

Tabelle 35.32: Linearisierte Cobb-Douglas-Produktionsfunktion Maschinenbau – OLS-Regression für die kleineren und mittleren Unternehmen der KSE, robuste Standardfehler (um Ausreißer bereinigt)

	Originalwerte		Anonymisiert	
Inputfaktoren	Koeffizienten	(t-Werte)	Koeffizienten	(t-Werte)
Materialeinsatz	0,377	(46,06)	0,376	(45,18)
Personalkosten	0,430	(34,69)	0,430	(34,87)
Externe Dienstleistungen	0,039	(13,92)	0,039	(14,00)
Sonstige Kosten	0,065	(11,04)	0,065	(11,11)
Kapitalkosten	0,075	(12,44)	0,075	(12,94)
Konstante	1,568	(18,13)	1,573	(18,32)
Anzahl der Beobachtungen	1.610		1.618	
Bestimmtheitsmaß	0,976		0,976	
Durchschnittliche Abweichung der Koeffizientenwerte			0,1%	
Anzahl der Vorzeichenwechsel			0	
Anzahl der Koeffizienten, die im Original statistisch signifikant sind, bei Anonymisierung nicht			0	
Anzahl der Koeffizienten, die bei Anonymisierung statistisch signifikant sind, im Original nicht			0	
Anzahl der Wechsel zwischen unterschiedlichen Signifikanzniveaus			0	
Durchschnittliche Abweichung der t-Werte			1,5%	

Tabelle 35.33: Linearisierte Cobb-Douglas-Produktionsfunktion mit Dummy-Variablen für einzelne Wirtschaftszweige - OLS-Regression für die kleineren und mittleren Unternehmen der KSE, robuste Standardfehler (um Ausreißer bereinigt).

	Originalwerte		Anonymisiert	
Inputfaktoren	Koeffizienten	(t-Werte)	Koeffizienten	(t-Werte)
Materialeinsatz	0,431	(128,36)	0,430	(125,89)
Personalkosten	0,334	(76,28)	0,337	(75,77)
Externe Dienstleistungen	0,052	(35,02)	0,052	(34,78)
Sonstige Kosten	0,100	(39,42)	0,100	(39,27)
Kapitalkosten	0,067	(25,7)	0,066	(25,65)
24	0,050	(4,79)	0,050	(4,76)
25	-0,003	(-0,30)	-0,003	(-0,31)
26	-0,028	(-3,07)	-0,028	(-3,08)
27	-0,007	(-0,63)	-0,007	(-0,67)
28	0,024	(3,00)	0,023	(2,87)
29	0,013	(1,61)	0,012	(1,49)
32	0,016	(1,16)	0,017	(1,20)
33	0,067	(6,31)	0,064	(6,03)
36	-0,003	(-0,32)	-0,004	(-0,46)
C (10,11,14)	-0,001	(-0,06)	0	(0,02)
DA (15,16)	0,063	(7,52)	0,062	(7,44)
DB (17,18)	0,016	(1,79)	0,018	(2,01)
DE (21,22)	0,074	(7,46)	0,072	(7,33)
DL (30,31)	0,017	(1,88)	0,017	(1,83)
DM (34,35)	0,031	(2,75)	0,032	(2,83)
19,23 (Restkategorie)	0,02	(1,41)	0,02	(1,37)
Konstante	1,665	(51,85)	1,661	(51,24)
Anzahl der Beobachtungen	11.657		11.678	
Bestimmtheitsmaß	0,973		0,972	
Durchschnittliche Abweichung der Koeffizientenwerte			8,1%	
Anzahl der Vorzeichenwechsel			1 (aber Koeffizient insignifikant)	
Anzahl der Koeffizienten, die im Original statistisch signifikant sind, bei Anonymisierung nicht			0	
Anzahl der Koeffizienten, die bei Anonymisierung statistisch signifikant sind, im Original nicht			0	
Anzahl der Wechsel zwischen unterschiedlichen Signifikanzniveaus			1	
Durchschnittliche Abweichung der t-Werte			10,9%	

Referenzgruppe: Wirtschaftszweig 20 (Holzgewerbe)
Zur Wirtschaftszweigsystematik vgl. Tabelle 35.23

Kapitel 36

Anonymisierung der Umsatzsteuerstatistik

36.1 Besonderheiten bei der Anonymisierung der Umsatzsteuerstatistik

Die Umsatzsteuerstatistik ist als einzige der Projekterhebungen eine Vollerhebung mit einer sehr geringen Abschneidegrenze. Darüber hinaus umfasst sie mit wenigen Ausnahmen sämtliche Wirtschaftsbereiche, die entsprechend bei einer Anonymisierung berücksichtigt werden müssen.

Die Unterscheidung zwischen einem Angreifer mit Teilnahmekenntnis und einem Angreifer ohne entfällt bei der Umsatzsteuerstatistik, da diese durch die Vollerhebung als gegeben angesehen werden kann. Damit entfällt einerseits zwar der Schutz, der durch eine Stichprobenziehung gegeben ist, andererseits stellt ein Angreifer, der Teilnahmekenntnis an der Erhebung besitzt, bei der Umsatzsteuerstatistik kein Problem dar, während bei einer Stichprobenerhebung ein solcher Angreifer ein sehr großes Risiko darstellt.[69]

Ein zweites wesentliches Unterscheidungsmerkmal gegenüber den anderen eingesetzten Projektstatistiken ist die Datenstruktur. Wie in früheren Kapiteln bereits erwähnt, besteht die Umsatzsteuerstatistik bei den metrischen Merkmalen im Wesentlichen aus dem Merkmal Umsatz. Die weiteren Merkmale, wie Umsatz zu 16% Umsatzsteuer, sind Untermerkmale. Für einen Datenangreifer bedeutet dies zweierlei: Einerseits sind seine potenziellen Überschneidungsmerkmale bei den metrischen Merkmalen sehr stark eingeschränkt (er hat z.B. keine Angaben über die Beschäftigten). Andererseits ist der Nutzen einer Reidentifikation als nicht sehr hoch einzustufen, besonders dann nicht, wenn der Umsatz als Überschnei-

69) Wenn z.B. bei einer Grundgesamtheit von 100 Unternehmen nur 10 an der Erhebung teilnehmen, dann stellt dies einen Schutz dar, vorausgesetzt, der Datenangreifer kennt die 10 Erhebungsteilnehmer nicht. Kennt er diese aber (d.h. er hat Teilnahmekenntnis) dann ist das Risiko einer Reidentifikation als sehr hoch anzusehen. Eine solche differenzierte Betrachtung kann bei einer Vollerhebung wie der Umsatzsteuerstatistik entfallen.

dungsmerkmal mit eingesetzt werden muss (was i.d.R. der Fall sein wird), da die zusätzlich enthüllten Merkmale nur eingeschränkt zusätzliche Informationen vermitteln. Dies erleichtert die Anonymisierung im Vergleich zu den anderen Projektstatistiken wesentlich, auch wenn in Betracht gezogen wird, dass mit der Rechtsform ein kategoriales Merkmal vorhanden ist, das weder in der Einzelhandelsstatistik noch in der Kostenstrukturerhebung des Verarbeitenden Gewerbes enthalten ist.

Wie im Teil X bereits dargelegt wurde, macht es aufgrund der hohen Abhängigkeit der Merkmale untereinander keinen Sinn, eine Anonymisierung auf die Überschneidungsmerkmale zu beschränken. Zum einen würde dadurch die innere Plausibilität der einzelnen Datensätze negativ beeinflusst und zum anderen können die mit dem Umsatz hoch korrelierten Merkmale als Ersatzüberschneidungsmerkmal von einem Datenangreifer verwendet werden, womit die Anonymisierung umgangen werden würde.

36.2 Verfügbares Zusatzwissen und Überschneidungsmerkmale

Als Überschneidungsmerkmale aus dem Zusatzwissen, das aus kommerziellen Datenbanken generiert werden kann, sind die Rechtsform, der regionale Gebietsschlüssel, die Wirtschaftszweigklassifikation und der Umsatz zu nennen. Die Umsatzsteuerstatistik unterscheidet sich damit von den anderen Projekterhebungen dahingehend, dass mit der Rechtsform ein kategoriales Überschneidungsmerkmal zusätzlich vorhanden ist, mit dem Umsatz allerdings nur ein metrisches Merkmal. Die Zahl der Beschäftigten liegt als Überschneidungsmerkmal nicht vor. Für einen potenziellen Angreifer bietet dies auf der einen Seite den Vorteil, ein relativ einfach zu generierendes und über den Zeitablauf stabiles Merkmal gegenüber einem schwierigeren und zeitlich variablen Merkmal einzutauschen. Auf der anderen Seite hat dies allerdings den Nachteil, dass die Daten aufgrund des Merkmals Rechtsform wesentlich weniger differenziert werden als bei einem metrischen Merkmal, wie der Anzahl der Beschäftigten. Welcher Vorteil stärker wiegt hängt von unterschiedlichen Faktoren, wie dem Zugang und der Qualität zum Zusatzwissen, ab. Bei einer Vergröberung der Rechtsform (siehe nächsten Abschnitt) und bei der Annahme, dass eine kommerzielle Datenbank einen leichten und verlässlichen Zugang zum Merkmal Beschäftigtenanzahl darstellt, scheint der Nachteil der fehlenden metrischen Variablen den Vorteil der zusätzlichen kategorialen zu überwiegen.

Während bei Massenfischzügen ein standardisierter Kanon an Überschneidungsmerkmalen nötig ist, können bei Einzelangriffen von einem Datenangreifer flexible Lösungen gesucht und gefunden werden. Flexibel heißt in diesem Sinne, dass die individuelle Struktur eines Merkmalsträgers bei der Suche nach passenden Überschneidungen im Zusatzwissen berücksichtigt werden kann. Für die Umsatzsteuerstatistik kämen hierfür Aspekte, wie neugegründetes Unternehmen, Auslandsumsatz des Unternehmens oder Umsatzwachstum, in Frage. Allerdings ist zu beachten, dass die Unternehmen hierdurch nur unzureichend differenziert werden können und es sich bei diesen Überschneidungsmerkmalen um abgeleitete und nicht

direkt erhobene Merkmale handelt. Dadurch sind diese entsprechend fehleranfällig und für Reidentifikationsversuche höchstens in Ausnahmefällen geeignet.

Wie bereits mehrfach dargelegt, sind die meisten metrischen Merkmale sehr stark mit dem Umsatz korreliert, wodurch sich die Eignung als zusätzliches Überschneidungsmerkmal in engen Grenzen hält.

Das alles führt zum Fazit, dass ein Einzelangriff in der Regel über die gleichen Überschneidungsmerkmale durchgeführt werden muss wie bei einem Massenfischzug. Dies hat zur Folge, dass die Ergebnisse, die bei Massenfischzugsimulationen erzielt werden, gute Schätzer für das Risiko darstellen, das von Einzelangriffen ausgeht.

36.3 Anonymisierungsmaßnahmen

Im Folgenden werden die Anonymisierungsmaßnahmen beschrieben, die letztlich zum Scientific-Use-File der Umsatzsteuerstatistik 2000 geführt haben (zum Scientific-Use-File vgl. Vorgrimler et al. (2005))[70]. Darüber hinaus wurden im Projektverlauf noch weitere Varianten getestet, die dann aber entweder wegen mangelnder Sicherheit oder aufgrund eines zu großen Verlusts an Analysepotenzial wieder verworfen wurden.

36.3.1 Kategoriale Merkmale

Zu Beginn des Projekts galt die Umsatzsteuerstatistik als diejenige der Projekterhebungen, bei der eine erfolgreiche faktische Anonymisierung unter Verzicht auf datenverändernde Verfahren am ehesten möglich erschien. Entsprechend konzentrierten sich die ersten Anonymisierungsversuche auf den Einsatz traditioneller Methoden, die wiederum in erster Linie bei den kategorialen Merkmalen ansetzten (eine ausführliche Beschreibung dieser traditionellen Anonymisierung findet sich im Anhang zum Zwischenbericht (Vorgrimler 2003a)). Auch wenn diese traditionelle Anonymisierung aufgrund einer sehr starken Einschränkung des Analysepotenzials verworfen wurde, waren die angewandten Methoden bei den einzelnen kategorialen Merkmalen Grundlage für die beim Scientific-Use-File angewandten Verfahren. Im Einzelnen sehen sie wie folgt aus:

- Der amtliche Gemeindeschlüssel wurde auf Ost/West vergröbert, wobei der Osten die neuen Bundesländer inklusive Berlin umfasst. Alternativ war eine regionale Verschlüsselung nach dem nichtadministrativen Gebietsschlüssel des Bundesamtes für Bauwesen und Raumordnung in der Diskussion. Dieser teilt die Regionen nach unterschiedlichen Siedlungsstrukturen auf. Für den Scientific-Use-File stand ein Schlüssel mit den

70) Eine ausführliche Darstellung der hier verwendeten Anonymisierungsmethoden findet sich in Teil II dieses Handbuchs „Anonymisierungsansätze und Anonymisierungsmethoden“.

drei Ausprägungen Agglomerationsraum, verstädterter Raum und ländlicher Raum zur Verfügung. Die im Wissenschaflichen Begleitkreis vertretenen Wissenschaftler sprachen sich mehrheitlich für eine Verschlüsselung nach Ost/West und gegen den nichtadministrativen Schlüssel aus. Da die Sicherheitsprüfung keine unterschiedlichen Ergebnisse brachte, ist die Entscheidung zu Gunsten der administrativen Einteilung gefallen.

- Die Wirtschaftszweigfklassifizierung (WZ 93) wird in unterschiedlicher Tiefengliederung abhängig von den Besetzungszahlen in die Daten aufgenommen. Dabei werden z.T. auch neu zusammengefasste Positionen gebildet. Mit der unterschiedlichen Tiefengliederung wird zweierlei Aspekten Rechnung getragen. Zum einen sind die einzelnen Wirtschaftsabschnitte unterschiedlich besetzt, so dass ein vergleichbares Sicherheitsniveau bei unterschiedlicher Tiefengliederung der Wirtschaftszweigklassifikation erreicht wird. Zum anderen sind die inhaltlichen Aussagen unterschiedlich von der Tiefe abhängig. Sind z.B. im Abschnitt Land- und Forstwirtschaft bereits auf der obersten Stufe inhaltliche Aussagen möglich, so werden im Bereich des Einzelhandels inhaltliche Aussagen erst ab der dritten Stelle der Wirtschaftszweigklassifikation sinnvoll. Die Besetzungszahlen der beiden Wirtschaftsabschnitte sind genau gegenläufig und entsprechend ist das Sicherheitsrisiko unterschiedlich zu bewerten. Daher erscheint es sinnvoll, den Einzelhandel bis zur dritten Stelle auszuweisen, während bei der Land- und Forstwirtschaft eine Unterschreitung der ersten Stelle aufgrund von Sicherheitsbedenken nicht möglich ist. Die Tabellen 36.1 und 36.2 listen die einzelnen ausgewiesenen Wirtschaftsbereiche des Scienitific-Use-Files auf.

- Aus dem Merkmal *Dauer der Steuerpflicht* wurde das Merkmal **Neugründung** mit den Ausprägung 1 = ja und 0 = nein gebildet. Bei Unternehmen mit mehr als 100 Mio. Euro Umsatz wird das Merkmal generell auf 0 gesetzt. Von über 150.000 als Neugründungen gekennzeichneten Unternehmen haben 118 Unternehmen einen Umsatz von über 100 Mio. Euro. Bei diesen wird dieses Merkmal auf Null gesetzt und damit die Information unterdrückt. Aus Plausibilitätsgründen dürfte diese Informationsreduktion nicht besonders relevant sein, da es sich in der Mehrheit der Fälle um keine echten Neugründungen handeln wird. Zu „unechten“ Neugründungen kommt es bspw. bei Rechtsformänderungen oder Sitzverlagerungen. Wenn es sich aber tatsächlich um echte Neugründungen handelt, dann kann die Sicherheit dieser Merkmalsträger nicht gewährleistet werden, so dass die genannte Maßnahme unumgänglich ist.

  Für unternehmensdemographische Analysen wäre eine Information über die Anzahl der gelöschten bzw. aufgelösten Unternehmen wünschenswert. Diese lässt sich aber leider aus den Daten nicht sicher ableiten.

- Das Merkmal *Organschaft ja/nein* wird sowohl aus Sicherheits- als auch aus Plausibilitätsgründen gestrichen.

- Das Merkmal **Rechtsform** wurde zu folgenden Ausprägungen vergröbert:

  – Personengesellschaften,

- Kapitalgesellschaften,
- Erwerbs- und Wirtschaftsgenossenschaften sowie Betriebe gewerblicher Art von Körperschaften des öffentlichen Rechts und
- Sonstige Rechtsformen.

Tabelle 36.1: Ausgewiesene Wirtschaftsbereiche im Scienitfic-Use-File Umsatzsteuerstatistik 2000, Teil I

WZ	Bezeichnung
A,B	Land- u. Forstwirtschaft, Fischerei u. Fischzucht
C	Bergbau u. Gewinnung von Steinen u. Erden
DA	Ernährungsgewerbe u. Tabakverarbeitung
DB	Textil- u. Bekleidungsgewerbe
19	Ledergewerbe
20	Holzgewerbe (ohne Herstellung von Möbeln)
21	Papiergewerbe
22	Verlags-, Druckgewerbe, Vervielfältigung
23	Kokerei, Mineralölverarbeitung, Herstellung u. Verarbeitung v. Spalt u. Brutstoffen
24	Chemische Industrie
25	Herstellung von Gummi- u. Kunststoffen
26	Glasgewerbe, Keramik, Verarbeitung von Steinen u. Erden
27	Metallerzeugung u. Metallbearbeitung
28	Herststellung von Metallerzeugnissen
29	Maschinenbau
30	Herstellung von Büromaschinen, DV-Geräten u. -einrichtungen
31	Herstellung von Geräten der Elektrizitätserzeugung u. Verteilung u.ä.
32	Rundfunk-, Fernseh- u. Nachrichtentechnik
33	Medizin-, Mess-, Steuer- u. Regelungstechnik, Optik
34	Herstellung von Kraftwagen u. Kraftwagenteilen
35	Sonstiger Fahrzeugbau
36	Herstellung von Möbeln, Schmuck, Musikinstr., Sportgeräten usw.
37	Recycling
40	Energieversorgung
41	Wasserversorgung
45.A	Bauhauptgewerbe
45.B	Bauausbaugewerbe
50	Kfz-Handel; Instandhaltung u. Reparaturen von Kfz; Tankstellen
51	Handelsvermittlung u. Großhandel (ohne Kfz)

Tabelle 36.2: Ausgewiesene Wirtschaftsbereiche im Scienitfic-Use-File Umsatzsteuerstatistik 2000, Teil II

WZ	Bezeichnung
52.1	Einzelhandel mit Waren verschiedener Art (in Verkaufsräumen)
52.2	Facheinzelhandel mit Nahrungsmitteln usw (in Verkaufsräumen)
52.3	Apotheken; Facheinzelhandel mit med. Art. usw. (in Verkaufsräumen)
52.4	Sonstiger Facheinzelhandel (in Verkaufsräumen)
52.5	Einzelhandel mit Antiquitäten u. Gebrauchtwaren (in Verkaufsräumen)
52.6	Einzelhandel (nicht in Verkaufsräumen)
52.7	Reparatur von Gebrauchsgütern
55.A	Beherbergungsgewerbe
55.B	Gaststättengewerbe u. Kantinen
60	Landverkehr; Transport in Rohrfernleitungen
61	Schifffahrt
62	Luftfahrt
63.1	Frachtumschlag u. Lagerei
63.2	Sonstige Hilfs- u. Nebentätigkeiten für den Verkehr
63.3	Reisebüros u. Reiseveranstalter
63.4	Spedition, sonstige Verkehrsvermittlung
64	Nachrichtenübermittlung
J	Kredit- u. Versicherungsgewerbe
70	Grundstücks- u. Wohnungswesen
71	Vermietung beweglicher Sachen
72	Datenverarbeitung u. Datenbanken
73	Forschung u. Entwicklung
74.1	Rechts-, Steuer- u. Unternehmensberatung usw.
74.2	Architekten u. Ingenieurbüros
74.3	Technische, physikalische u. chemische Untersuchungen
74.4	Werbung
74.5	Gewerbsmäßige Vermittlung u. Überlassung v. Arbeitskräften
74.6	Detekteien u. Schutzdienste
74.7	Reinigung von Gebäuden, Inventar u. Verkehrsmitteln
74.8	Erbringung von sonst. Dienstleistungen überwiegend für Unternehmen
L,M,N	Öff. Verw., Verteidigung, Sozialversicherung, Erziehung u. Unterricht, Gesundheits-,Veterinär- u. Sozialwesen
90	Abwasser-, Abfallbeseitigung u. sonstige Entsorgung
91	Interessenvertretung, kirchliche u. sonstige religiöse Vereinigungen
92	Kultur, Sport u. Unterhaltung
93.01	Wäscherei u. chemische Reinigung
93.02	Friseurgewerbe u. Kosmetiksalons
93.03	Bestattungswesen
93.04	Bäder, Saunas, Solarien u.a.
93.05	Erbringung von Dienstleistungen andernorts nicht genannt

36.3.2 Metrische Merkmale

Bei der bereits erwähnten rein traditionellen Anonymisierung wurden aus Sicherheitsgründen besonders die Merkmalsausprägungen der großen Unternehmen mit Maßnahmen des Topcoding sehr stark verändert. Dies hatte eine nichtakzeptable Einschränkung des Ana-

lysepotenzials zur Folge. Aus diesem Grunde wurde im nächsten Schritt versucht, ein datenveränderndes Verfahren zu wählen, das die Merkmalsausprägung so verändert, dass die Merkmalsträger einerseits ausreichend geschützt sind und das Analysepotenzial andererseits nicht unnötig eingeschränkt wird. Die Mikroaggregation wirkt besonders bei den Merkmalsträgern schützend, die auch besonders schutzbedürftig sind. Daher eignet sich diese Verfahrensgruppe bei der Umsatzsteuerstatistik in besonderem Maße zur Anonymsierung. Aufgrund des Analysepotenzials wurde die getrennte Mikroaggregation verwendet.

Eine Untersuchung des Reidentifikationsrisikos bei den Unternehmen der Umsatzsteuerstatistik ergab ein erhöhtes Risiko der marktführenden Unternehmen. Im Gegensatz zu den sonstigen großen Unternehmen konnten diese auch nicht durch eine getrennte Mikroaggregation ausreichend geschützt werden, zumindest nicht, ohne das Analysepotenzial übermäßig stark einzuschränken. Neben der getrennten Mikroaggregation wurde daher noch eine punktuelle Mikroaggregation eingeführt, die nur punktuell bei den marktführenden Unternehmen ansetzt. Die Anonymisierung bei den metrischen Merkmalen besteht daher aus folgender zweistufiger Mikroaggregation:

- Auf der ersten Stufe wird eine für jedes Merkmal getrennte und für alle Unternehmen angewandte Mikroaggregation durchgeführt.

- Zum Schutz der regionalen Branchenmarktführer wurde in der zweiten Stufe eine punktuelle Mikroaggregation durchgeführt. Dabei werden nur speziell die jeweiligen drei regionalen Marktführer einer Branche gemeinsam mikroaggregiert, wobei das Merkmal *Lieferungen und Leistungen* (Umsatz) das bestimmende Merkmal ist[71]. Da dieses Verfahren getrennt nach den beiden Regionen angewandt wird, sind insgesamt 408 Merkmalsträger davon betroffen. Diese sind als solche mit einem zusätzlichen Merkmal kenntlich gemacht. In der getrennten Behandlung der beiden Regionen Ost und West, wird dem besonderen Schutzbedürfnis der wenigen großen ostdeutschen Unternehmen Rechnung getragen.

Tabelle 36.3 enthällt die Datensatzbeschreibung der Umsatzsteuerstatistik, wie sie aufgrund der Anonymisierung entstanden ist.

Wie die folgende Schutzwirkungsanalyse zeigen wird, sind diese Maßnahmen ausreichend, um ein Schutzniveau zu generieren, das den Ansprüchen des §16 Abs. 6 BStatG genügt, so dass diese Daten aus Sicht des Datenschutzes als Scientific-Use-File geeignet sind.

71) Als regionaler Branchenmarktführer gelten die drei Unternehmen, die in einer Branche (abgegrenzt nach der im Scientific-Use-File ausgewiesenen Wirtschaftszweigklassifikation) und einer Region (Ost/West) die höchsten Umsätze aufweisen.

Tabelle 36.3: Datensatzbeschreibung des Scientific-Use-Files der Umsatzsteuerstatistik 2000

Nr.	Merkmal	Untermerkmal	Untermerkmal
1	zufällig vergebene Nummer		
2	Region (Ost/West)		
3	Wirtschaftszweig		
4	Neugründung	1 = ja 0 = nein	
5	Rechtsform 1 = Personengesellschaft 2 = Kapitalgesellschaft 3 = Genossenschaften sowie Betriebe gewerblicher Art öff. Rechts 4 = sonstige Rechtsformen		
6	Lieferungen und Leistungen (LuL)		
7	Steuerpflichtige LuL		
8		zu 16%	
9		zu 7%	
10	Steuerfreie LuL		
11		mit Vorsteuerabzug	
12			innergemeinschaftl. LuL
13			weitere steuerfreie LuL
14		ohne Vorsteuerabzug	
15	Umsatzsteuer vor Abzug der Vorsteuer		
16		für LuL	
17		für innergemeinschaftl. Erwerbe	
18	Abziehbare Vorsteuer		
19		für LuL	
20			aus Rechnungen anderer Unternehmen
21			Einfuhrumsatzsteuer
22		für innergemeinschaftl. Erwerbe	
23	Vorauszahlungssoll		
24	Nachricht.: innergemeinschaftl. Erwerbe		
25	LuL 1999		
26	Vorauszahlungssoll 1999		
27	Punktuell mikroaggregiert	1 = ja 0 = nein	

36.4 Überprüfung der Schutzwirkung der Anonymisierung

Während der Projektlaufzeit wurden parallel zu den verschiedenen Probeanonymisierungen Schutzwirkungstests durchgeführt. Beispiele hierfür finden sich in Lenz und Vorgrimler (2004) und Lenz et al. (2004b). Wie schon im vorhergehenden Abschnitt konzentrieren sich die folgenden Ausführungen auf die Schutzwirkung der Anonymisierung, die zur Erstellung des Scientific-Use-Files geführt hat.

Bei einer Schutzwirkungsanalyse kann zwischen einem natürlichen Schutz und einer Schutzwirkung aufgrund der Anonymisierung unterschieden werden. Ein natürlicher Schutz entsteht, wenn die Merkmalsausprägungen des Zusatzwissens von denen der Zieldaten abweichen, d.h. die Erhebungen nicht kompatibel zueinander sind. Die Rolle dieses natürlichen Schutzes bei der Umsatzsteuerstatistik wird im folgenden Unterabschnitt beschrieben.

36.4.1 Der natürliche Schutz

Um einen Eindruck über den natürlichen Schutz zu gewinnen, wurden die Merkmalsausprägungen der Umsatzsteuerstatistik mit potenziellem Zusatzwissen verglichen. Verwendet wurden hierbei die Wirtschaftszweigklassifikation und der Umsatz. Bei den beiden anderen Überschneidungsmerkmalen Rechtsform und Gebietsschlüssel wurde darauf verzichtet, da zum einen die Rechtsform z.T. künstlich hinzugespielt wurde und damit per se korrekt ist und zum anderen der Gebietsschlüssel im Scientific-Use-File nur noch mit zwei Ausprägungen (Ost/West) in den Zieldaten vorhanden ist.[72] Als potenzielles Zusatzwissen standen die Kostenstrukturerhebung des Verarbeitenden Gewerbes (KSE), die kommerziell erhältliche MARKUS-Datenbank (beschränkt auf Unternehmen des Verarbeitenden Gewerbes) und die Einzelhandelsstatistik zur Verfügung.

Tabelle 36.4 zeigt die Abweichungen in der Ausprägung Wirtschaftszweigklassifikation (in unterschiedlicher Tiefengliederung) und den Anteil derjenigen Merkmalsträger, deren Umsätze in den beiden Erhebungen um weniger als x% voneinander abweichen. Untersucht wurden hierbei 9.283 Unternehmen der Umsatzsteuerstatistik bzw. der KSE.

72) Bei früheren Analysen mit tiefergehendem Regionalschlüssel konnten aber durchaus Abweichungen zwischen Zieldaten und Zusatzwissen festgestellt werden, allerdings waren diese mit 2% sehr gering, so dass es gerechtfertigt erscheint, diese im Folgenden zu vernachlässigen.

Tabelle 36.4: Abweichungen in den Merkmalsausprägungen zwischen Zusatzwissen und Zieldaten

Gegenstand der Nachweisung	Unternehmen absolut	Unternehmen relativ
Abweichung des Umsatzes geringer als...		
1%	3.546	38,2
5%	6.541	70,5
10%	7.539	81,2
25%	8.395	90,4
50%	8.706	93,8
insgesamt	9283	100
identischer WZ-93 Klassifizierung auf Ebene		
4-Steller	5.206	56,1
3-Steller	5.917	63,7
2-Steller	7.007	74,5
1-Steller	7.823	84,3
insgesamt	9.283	100

Wie aus der Tabelle 36.4 ersichtlich, hängt das natürliche Schutzniveau einerseits davon ab, wie genau ein Angreifer den Umsatzwert aus dem Zusatzwissen generieren muss, um richtig reidentifizieren zu können und andererseits davon, in welcher Tiefe der Wirtschaftszweigklassifikation er angreift. Braucht ein Datenangreifer z.B. eine Genauigkeit beim Umsatz von mindestens 10% und greift er mit einer zweistelligen Wirtschaftszweigklassifikation an, dann wird er knapp 40% der Unternehmen deswegen nicht reidentifizieren können, weil das Zusatzwissen mit den Zieldaten nicht kompatibel ist.

Verwendet man anstelle der KSE eine kommerzielle Unternehmensdatenbank als Zusatzwissen (in diesem Fall die MARKUS-Datenbank), kommt dem natürlichen Schutz eine noch größere Bedeutung zu. Bei rund 6.000 untersuchten Unternehmen, die in der Umsatzsteuerstatistik und in der MARKUS-Datenbank enthalten waren, war der Anteil an korrekten Wirtschaftsklassifizierungen zwar nur leicht unter dem Ergebnis, das mit der KSE erzielt wurde (auf Zwei-Steller Ebene stimmen 71% der Ausprägungen überein), das Ergebnis beim Umsatz war aber deutlich schlechter. Lediglich rund die Hälfte der untersuchten Unternehmen hatten Abweichungen von weniger als 10% bei diesem Überschneidungsmerkmal. Der natürliche Schutz ist entsprechend größer.

Beim Vergleich der Unternehmen des Einzelhandels innerhalb der Umsatzsteuerstatistik ist der natürliche Schutz geringer. Zum einen spielt die Wirtschaftszweigklassifikation keine große Rolle, da der Einzelhandel erst auf Dreisteller-Ebene sieben unterschiedliche Ausprägungen aufweist, zum anderen ist die Übereinstimmung der Merkmalsausprägungen zwischen der Umsatzsteuerstatistik als Zieldaten und der Einzelhandelsstatistik als Zusatzwissen etwas größer. 73,5% der Merkmalsträger weisen in den beiden Erhebungen

einen Umsatz auf, der weniger als 5% abweicht. Allerdings zeigten sich besonders bei den größeren Unternehmen zum Teil erhebliche Abweichungen, was an den unterschiedlichen Umsatzdefinitionen der beiden Erhebungen liegt.

Die Analyse des natürlichen Schutzes kann sicherlich nicht als abschließend angesehen werden. Trotzdem lässt sich sagen, dass er eine wesentliche Rolle bei der Erreichung der faktischen Anonymität spielt und entsprechend mit in die Betrachtung einbezogen werden muss. Die Analyse der MARKUS-Datenbank zeigt darüber hinaus, dass, wenn amtliche Daten mit amtlichen Daten als Zusatzwissen angegriffen werden, der natürliche Schutz unter- und das Risiko überschätzt wird. Dies muss bei der Bewertung der folgenden Analyse - wenn die amtliche Umsatzsteuerstatistik mit den amtlichen Daten aus der Kostenstrukturerhebung und der Einzelhandelsstatistik angegriffen wird - beachtet werden.

36.4.2 Schutzwirkung der Anonymisierung

Zum Test der Schutzwirkung wurden Massenfischzugs- und Einzelangriffsszenarien durchgeführt. Als Erstes wurde mit Hilfe der KSE des Verarbeitenden Gewerbes[73] versucht, Merkmalsträger dieses Wirtschaftsabschnittes zu reidentifizieren. Im zweiten Szenario wurden 12.500 Unternehmen der Einzelhandelsstatistik als Zusatzwissen[74] verwendet und ebenfalls ein Massenfischzug durchgeführt.

Szenario I: Massenfischzug mit Hilfe der KSE

Bei dem Matchingexperiment mit Hilfe der KSE als Zusatzwissen wurde versucht, knapp 9.300 Unternehmen der externen KSE-Datei (Zusatzwissen) den Merkmalsträgern der Umsatzsteuerstatistik (Zieldaten) zuzuordnen. In der Kostenstrukturerhebung des Verarbeitenden Gewerbes sind – wie bereits ausführlich behandelt – Unternehmen mit mindestens 20 Beschäftigten als Stichprobe enthalten. Von den 2,9 Mio. Unternehmen der Umsatzsteuerstatistik kamen daher ca. 37.000 Unternehmen als Zielunternehmen in Frage. Da es sich bei der Umsatzsteuerstatistik bei Unternehmen mit Lieferungen und Leistungen von über 16.617 Euro prinzipiell um eine Vollerhebung handelt, konnte von der Teilnahmekenntnis des Datenangreifers ausgegangen werden. Tabelle 36.5 zeigt die Zuordnungsquote, den

73) Zu den Merkmalen der KSE wurde den Merkmalsträgern zusätzlich das Merkmal Rechtsform hinzugespielt. Somit hat das Zusatzwissen die typische Struktur einer kommerziellen Unternehmensdatenbank, mit den Überschneidungsmerkmalen Rechtsform, Gebietsschlüssel, Wirtschaftszweigklassifikation und Umsatz. Darüber hinaus hat es den Vorteil, Anonymisierungsmaßnahmen, die bei der Rechtsform ansetzen, in ihrer Schutzwirkung bewerten zu können. Bei den Ergebnissen muss aber beachtet werden, dass das Zusatzwissen zwar realistisch in dem Sinne ist, dass es die gleiche Struktur aufweist wie eine frei verfügbare Unternehmensdatenbank und es sich um eine völlig andere Erhebung handelt, es aber unrealistisch hinsichtlich seiner hohen Datenqualität, gemessen an der Übereinstimmung zwischen Zieldaten und Zusatzwissen, ist.

74) Als zusätzliches Merkmal wurde den Merkmalsträgern wiederum die Rechtsform hinzugespielt.

Anteil an nützlichen Informationen und das sich daraus ergebende Enthüllungsrisiko.[75] Die Tabelle enthält sowohl die Gesamtwerte als auch nach Umsatzgrößenklassen aufgeteilte Werte. Als Nützlichkeitsschwelle wurde 10% Abweichung der enthüllten von den wahren Werten unterstellt.

Tabelle 36.5: Enthüllungsrisiko bei der Umsatzsteuerstatistik im Szenario I

Umsatzklasse (in Euro)	Zuordnungsquote (in %)	Anteil nützlicher Informationen (in %)	Enthüllungsrisiko (in %)
bis 1 Mio.	9,1	100	9,1
1-10 Mio.	14,2	100	14,2
10-100 Mio.	18,7	99,7	18,6
100-1 Mrd.	27,9	92,6	25,9
über 1 Mrd.	43,1	61,8	26,9
insgesamt	16,7	98,6	16,5

Das durchschnittliche Enthüllungsrisiko aller Unternehmen liegt bei 16,5% wobei in keiner Klasse ein Risiko von über 30% erreicht wird. Weitergehende Analysen zeigen die beträchtliche Schutzwirkung der punktuellen Anonymisierung. Das Enthüllungsrisiko bei den größten Unternehmen sank aufgrund dieser Maßnahme um 11 bzw. 12 Prozentpunkte. Dies lag einerseits an einem Rückgang der Zuordnungsquote (von 47,7% auf 43%) und andererseits an einer Reduzierung des Anteils an nützlichen Informationen (von 78 auf 61%). Bei den mittleren Größenklassen ist zwar der Anteil an unbrauchbaren Informationen bei den zugeordneten Unternehmen sehr gering, allerdings gilt dies auch für die Zuordnungsquoten.

Szenario II: Massenfischzug mit Hilfe der MARKUS-Datenbank

Zu einem früheren Zeitpunkt des Projekts wurde ein Massenfischzug mit Hilfe der MARKUS-Datenbank simuliert. Hierbei wurde versucht, 6.300 Unternehmen der MARKUS-Datenbank, die zum Verarbeitenden Gewerbe zu zählen sind, den 37.000 Zielunternehmen des Szenarios I zuzuordnen. Auch wenn der Massenfischzug nicht bei dem eigentlichen Scientific-Use-File durchgeführt wurde, sondern lediglich bei einer schwächer anonymisierten Datei[76], zeigen die Ergebnisse eindrucksvoll die Rolle des oben beschriebenen natürlichen Schutzes, wenn auf kommerziell vorhandene Datenbanken als Zusatzwissen zurückgegriffen wird. Insgesamt konnten lediglich 5% der Unternehmen richtig zugeordnet werden. Nur in der höchsten Umsatzklasse konnte mit 31% ein für einen Angreifer halbwegs zufriedenstellendes Ergebnis erzielt werden. Diese Quote wird aber wie oben gesehen deutlich geringer, wenn zusätzlich die punktuelle Mikroaggregation betrachtet wird. Ohne auf die Nützlichkeit der enthüllten Informationen eingehen zu

75) Zur Definition des Enthüllungsrisikos siehe Kapitel 12 „Das Konzept zur Schutzwirkung".

76) Bei der hier untersuchten Datei fehlt u.a. die punktuelle Mikroaggregation.

müssen, zeigt bereits die Zuordnungsquote, dass der Massenfischzug im Rahmen dieses Szenarios als gescheitert angesehen werden muss. Ein Hauptgrund hierfür ist in der natürlichen Schutzwirkung zu sehen.

Szenario III: Massenfischzug mit Hilfe der Einzelhandelsstatistik

Die Massenfischzüge, bei denen die KSE als Zusatzwissen verwendet wird, beschränken sich auf Unternehmen des Verarbeitenden Gewerbes mit mindestens 20 Beschäftigten. Um diese eingeschränkte Sichtweise um einen weiteren Wirtschaftsbereich und um Kleinstunternehmen zu erweitern, wurde ein Massenfischzug simuliert, bei dem die Daten der Einzelhandelsstatistik als Zusatzwissen verwendet wurden. Strukturell unterscheidet sich dieses dritte Szenario von den ersten beiden in zweierlei Hinsicht:

- in der geringeren Anzahl an unterschiedlichen Kategorien der Wirtschaftszweigklassifikation (7 Klassen im Vergleich zu 22),
- in der fehlenden „Abschneidegrenze“ für kleine Unternehmen.

Beide Punkte sprechen für wesentlich schlechtere Voraussetzungen für diesen Massenfischzug. Sie führen dazu, dass die gesuchten Unternehmen innerhalb einer wesentlich höheren Zahl an potenziellen Unternehmen gesucht werden müssen, die sich auf weniger disjunkte Gruppen verteilen.

Insgesamt wurden 12.500 Unternehmen der Einzelhandelsstatistik innerhalb von über 300.000 Unternehmen aus der Umsatzsteuerstatistik gesucht. Die riesigen Datenmengen waren nur zu verarbeiten, indem die Unternehmen in beiden Dateien in unterschiedliche Umsatzgrößenklassen geblockt wurden. Durch Blockungsfehler können hierdurch bereits richtige Zuordnungen verhindert worden sein. Allerdings hat bereits dieses Vorgehen eine Rechenlaufzeit von 19 Stunden (in CPU-Zeit) benötigt.

Tabelle 36.6: Enthüllungsrisiko bei der Umsatzsteuerstatistik im Szenario III

Umsatzklasse (in Euro)	Zuordnungsquote (in %)	Anteil nützlicher Informationen (in %)	Enthüllungsrisiko (in %)
bis 1 Mio.	7,0	99,9	7,0
1-10 Mio.	9,9	99,8	9,9
10-100 Mio.	22,9	99,9	22,9
100-1 Mrd.	29,2	94,8	27,7
über 1 Mrd.	30,7	55,1	16,9
insgesamt	8,8	99,5	8,7

In Tabelle 36.6 sind wiederum die Zuordnungsquote, der Anteil an nützlichen Informationen sowie das Enthüllungsrisiko innhalb dieses Szenarios abgetragen. Insgesamt ergibt

sich ein deutlich geringeres Enthüllungsrisiko als bei Szenario I. Dies liegt an der deutlich unterschiedlichen Größenstruktur der Unternehmen, die mit der bereits erwähnten fehlenden Abschneidegrenze zusammenhängt. Bei den Einzelhandelsunternehmen dominieren die kleinsten Unternehmen, daher wirken sie auch dominierend auf das Maß für das Enthüllungsrisiko. Da dieses bei den kleinsten Unternehmen sehr gering ist, wird auch das gesamte Enthüllungsrisiko sehr gering sein.

Geht man von einer Gesamtbetrachtung zu einer Größenbetrachtung über, so wird deutlich, dass sich die Ergebnisse der beiden Szenarien nicht wesentlich unterscheiden. Die schlechteren Ausgangsbedingungen für den Datenangreifer bei den Einzelhandelsunternehmen aufgrund der geringeren Anzahl an Kategorien der Wirtschaftszweigklassifikation werden durch geringere Dateninkompatibilitäten bei der Matchingvariable „Umsatz“ wieder ausgeglichen (vgl. Unterabschnitt über den natürlichen Schutz der Daten). Auffallend ist allerdings das deutlich geringere Risiko bei den größten Unternehmen. Eine Erklärung hierfür könnte darin liegen, dass bei den Einzelhandelsunternehmen weniger Unternehmen der größten Umsatzgrößenklassen existieren und diese dadurch relativ häufiger von der zusätzlichen Anonymisierung aufgrund der punktuellen Mikroaggregation betroffen sind als dies bei den Unternehmen der KSE der Fall ist. Dies äußert sich dann sowohl durch eine geringere Trefferquote (23 zu 43%) als auch durch einen geringeren Anteil an nützlichen Informationen (55 zu 61%).

Die Ergebnisse des Szenarios III bestätigen die des ersten Szenarios. Unter der Beachtung des Ergebnisses aus Szenario II kann daher die Annahme der faktischen Anonymität der Merkmalsträger aufgrund der Massenfischzüge nicht widerlegt werden. Darüber hinaus zeigen die Ergebnisse des Szenarios III deutlich, dass ein Maß für das Enthüllungsrisiko, das alle Unternehmen mit einbezieht, wenig hilfreich ist. Für die Umsatzsteuerstatistik muss aufgrund der Vielzahl an Kleinstunternehmen das Risiko immer in Abhängigkeit zur Unternehmensgröße und der Wirtschaftszweigklassifikation betrachtet werden.

Szenario IV: Einzelangriffe

Bereits frühzeitig im Projekt wurden Einzelangriffe auf Merkmalsträger der Umsatzsteuerstatistik durchgeführt. Mangels kompatiblen Zusatzwissens waren diese aber wenig erfolgreich und mussten als gescheitert betrachtet werden. Von den gesuchten Unternehmen konnte keines richtig zugeordnet werden. Die Auswahl der Unternehmen erfolgte allerdings rein zufällig. Dies hatte aufgrund der Struktur der Umsatzsteuerstatistik die Folge, dass vor allem kleinere Unternehmen, die einen hohen natürlichen Schutz aufweisen, gesucht wurden. Ein gezielter Angriff auf besonders gefährdete Unternehmen fand nicht statt. Eine weitere Testreihe von Einzelangriffen bezog sich speziell auf die Marktführer verschiedener Branchen und es wurden damit gezielt besonders gefährdete Unternehmen für die Reidentifikationsversuche ausgewählt. Von fünf untersuchten Unternehmen aus vier verschiedenen Branchen konnten tatsächlich alle richtig zugeordnet werden. Die faktische Anonymität konnte daher für diese Unternehmen nicht angenommen werden.

Aus diesem Grunde wurde die punktuelle Mikroaggregation eingeführt (s.o.). Diese führt dazu, dass sich die drei regionalen Branchenmarktführer mit Ausnahme der Rechtsform nicht mehr unterscheiden und keine Originalwerte, sondern lediglich Durchschnittswerte der drei regional führenden Unternehmen veröffentlicht werden. Eine eindeutige Zuordnung ist nur noch dann möglich, wenn sich einer der drei Merkmalsträger durch die Rechtsform unterscheidet. Bei den 136 regionalen Branchen (jeweils 68 Wirtschaftszweige in Ost und Westdeutschland) unterscheidet sich in 95 mindestens 1 Unternehmen von den anderen und könnte somit zugeordnet werden. Eine solche eindeutige Zuordnung würde aber nicht gegen die faktische Anonymität verstoßen, da ein Datenangreifer keine zusätzlichen Informationen gewinnen kann. So würde er z.B. weiterhin lediglich den durchschnittlichen Umsatz der drei regionalen Branchenmarktführer kennen und nicht den exakten Umsatzwert seines zugeordneten Unternehmens. Diese durchschnittlichen Werte liegen in gut der Hälfte der Fälle um mindestens 50% von den „wahren“ Werten entfernt. Die Gruppe der regionalen Branchenmarktführer kann daher aufgrund der Maßnahme der punktuellen Mikroaggregation als faktisch anonym angesehen werden.

36.5 Überprüfung des Analysepotenzials

Die Vergröberungen der qualitativen Merkmale wurden in enger Abstimmung zwischen dem Statistischen Bundesamt und dem IAW vorgenommen. Dabei wurde insbesondere darauf geachtet, dass die für eine fundierte Analyse notwendige Tiefengliederung der Branchen insbesondere im Dienstleistungsbereich erhalten bleibt.

Für die Umsatzsteuerstatistik wurden lediglich deskriptive Auswertungen vorgenommen. Das Analysepotenzial des anonymisierten Datensatzes wurde insbesondere gemäß den in Kapitel 18 dargestellten Abweichungsmaßen geprüft (Unterabschnitt 36.5.1). Daneben wurden beispielhaft weitere deskriptive Auswertungen vorgenommen (Unterabschnitt 36.5.2).

36.5.1 Überprüfung der Einhaltung der Abweichungsschwellen bei deskriptiven Maßen

In den Tabellen 36.7 bis 36.11 sind die meisten in Kapitel 18 definierten Abweichungsmaße für die faktisch anonymisierte Umsatzsteuerstatistik dargestellt.

Die arithmetischen Mittel werden durch die vorgenommenen Anonymisierungsmaßnahmen systematisch erhalten. Die Aussagekraft der Veränderungsraten der Streuungsmaße und der Korrelationen des Gesamtdatensatzes ist zum einen dadurch eingeschränkt, dass diese durch die punktuelle Mikroaggregation der Branchenführer beeinträchtigt werden, zum

Tabelle 36.7: Durchschnittliche Veränderung der Verteilungsmaße

Durchschnittliche relative Veränderung der		Durchschnittliche absolute Veränderung der	
arithmetischen Mittel (in %)	Standard-abweichungen (in %)	Korrelations-koeffizienten (x100)	Rangkorrelations-koeffizienten (x100)
0	22,2	9,3	0,5

Tabelle 36.8: Veränderung der univariaten Verteilungsmaße

Anteil der arithmetischen Mittel, die um mehr als 10% abweichen (in %)	Anteill der Mediane, die um mehr als 10% abweichen (in %)	Anteil der Standard-abweichungen, die um mehr als 10% abweichen (in %)
0	0	100

Anteile bezogen auf die Anzahl der Merkmale

anderen führt die stärkere Veränderung der Einzelwerte von Großunternehmen zu einer starken Verzerrung der Streuungsmaße. Etwa die Hälfte der Varianzverzerrung ist auf die punktuelle Mikroaggregation zurückzuführen, die andere Hälfte auf die allgemeine getrennte Mikroaggregation. Dabei fällt die Varianzverzerrung erheblich stärker aus, wenn größere Unternehmen in die Analysen einbezogen werden.

Überschreitungen der definierten Toleranzgrenze (Anteil von 10%) ergeben sich in bei den Mittelwerten (arithmetische Mittel und Mediane) von Teilgesamtheiten nur bei Auswertungen nach der Rechtsform. Dies ist darauf zurückzuführen, dass die punktuelle Mikroaggregation der Branchenführer über die Kategorien dieses Merkmals hinweg erfolgte.

Tabelle 36.9: Veränderung der Korrelationen

Anteil der Korrelations-koeffizienten (in %), die		Anteil der Rang-korrelationen (in %), die	
um mehr als 0,1 abweichen	einen Vorzeichen-wechsel aufweisen	um mehr als 0,05 abweichen	einen Vorzeichen-wechsel aufweisen
25,7	1	1	0

Anteile bezogen auf die Anzahl der berechneten Korrelationen

Tabelle 36.10: Veränderung von Streuungsmaßen für unterschiedliche Arten der Anonymisierung sowie eine Teilgesamtheit der Kleinunternehmen im Vergleich zum Gesamtdatensatz

	Ausschließlich getrennte Mikroaggregation	Getrennte Mikroaggregation und punktuelle Mikroaggregation	Getrennte Mikroaggregation und punktuelle Mikroaggregation (nur Unternehmen mit einem Umsatz von weniger als 10 Mio.)
Durchschnittliche relative Veränderung der Standard abweichungen (in %)	10,7	22,2	2,8
Anteil der Standard-abweichungen mit einer Abweichung von mehr als 10 % (in %)	42,9	100	14,3

Anteile bezogen auf die Anzahl der Merkmale

Tabelle 36.11: Veränderung von Verteilungsmaßen für Teilgesamtheiten

Nach	Anteil der arithmetischen Mittel mit Abweichung von mehr als 10% (in %)	Anteil der Mediane mit Abweichung von mehr als 10% (in %)	Anteil der Standardabweichungen mit Abweichung von mehr als 10% (in %)
Wirtschaftszweigen	7,6	0	42,2
Ost/West	0	0	85,7
Rechtsform	17,9	0	71,4
Neugründung (ja oder nein)	0	0	50,0*
Wirtschaftszweigen und Ost/West	5,5	0,2	38,7
Wirtschaftszweigen und Rechtsform	15,9	0,1	29
Wirtschaftszweigen und Neugründung	4	0,1	22,5
Rechtsform und Ost/West	31,5	0	76,2
Rechtsform und Neugründung	9,5	0	36,3
Ost/West und Neugründung	0	0	44
Wirtschaftszweigen, Ost/West und Rechtsform	14,4	0,4	24,9
Wirtschaftszweigen, Ost/West und Neugründung	3,4	0,1	20,2
Nach Wirtschaftszweigen, Rechtsform und Neugründung	8,2	0,1	16,4
Nach Rechtsform, Ost/West und Neugründung	16,4	0	38,7
Nach Wirtschaftszweigen, Rechtsform, Ost/West und Neugründung	7,4	0,2	15

Anteile bezogen auf die Anzahl der Merkmale multipliziert mit der jeweiligen Anzahl der Teilgesamtheiten

* Abweichungen von über 10% treten ausnahmslos für alle Variablen in der Zelle „keine Neugründung“ auf, Abweichungen von unter 10 % ausnahmslos in der Zelle „Neugründung“. Dies zeigt, dass die Streuungsverzerrung sehr stark von der Größe der Unternehmen abhängt. Zudem sind Neugründungen von der punktuellen Mikroaggregation nicht betroffen.

36.5.2 Darstellung weiterer konkreter Auswertungen

a) Umsatzanalysen

Die häufigsten Auswertungen der Umsatzsteuerstatistik betreffen Auszählungen der Steuerpflichtigen und der *Lieferungen und Leistungen* nach Wirtschaftszweigen und Umsatzgrößenklassen. Tabelle 36.12 zeigt eine Gegenüberstellung der Anzahl der Steuerpflichtigen und deren *Lieferungen und Leistungen* nach Wirtschaftszweigen für die Originaldaten und die anonymisierten Daten.

Tabelle 36.12: Steuerpflichtige, Lieferungen und Leistungen nach Wirtschaftszweigen

Pos. Nr. der WZ93	Wirtschaftsgliederung	Steuerpflichtige 1)		Abweichung	Lieferungen und Leistungen 2)		Abweichung
		Original	Anonymisiert		Original	Anonymisiert	
		Anzahl		%	Mill. Euro		%
A-O	Wirtschaftszweige insgesamt	2.909.150	2.909.150	0,0	4.152.927	4.152.927	0,0
A, B	Land- und Forstwirtschaft, Fischerei und Fischzucht	65.764	65.764	0,0	23.316	23.320	0,0
C	Bergbau und Gewinnung von Steinen und Erden	3.067	3.067	0,0	26.601	26.198	-1,5
D	Verarbeitendes Gewerbe	291.885	291.885	0,0	1.514.702	1.481.318	-2,2
E	Energie- und Wasserversorgung	10.035	10.035	0,0	144.216	142.630	-1,1
F	Baugewerbe	323.116	323.116	0,0	218.928	218.962	0,0
G	Handel; Instandhaltung und Reparatur von Kraftfahrzeugen und Gebrauchsgütern	731.491	731.491	0,0	1.328.683	1.361.060	2,4
H	Gastgewerbe	251.865	251.865	0,0	53.288	53.292	0,0
I	Verkehr und Nachrichtenübermittlung	127.391	127.391	0,0	217.334	220.937	1,7
J	Kredit- und Versicherungsgewerbe	16.052	16.052	0,0	30.917	30.958	0,1
K	Grundstücks- und Wohnungswesen, Vermietung beweglicher Sachen, Erbringung von Dienstleistungen überwiegend für Unternehmen	760.671	760.671	0,0	458.162	457.425	-0,2
L,M,N	Öffentliche Verwaltung, Verteidigung, Sozialversicherung, Erziehung und Unterricht, Gesundheits-, Veterinär- und Sozialwesen	62.901	62.901	0,0	43.247	43.346	0,2
O	Erbringung von sonstigen öffentlichen und persönlichen Dienstleistungen	264.912	264.912	0,0	93.532	93.481	-0,1

1) Mit mehr als 16.617 Euro Jahresumsatz. 2) Umsätze der Unternehmen - ohne Umsatzsteuer.

Während durch die Anonymisierung die Wirtschaftszweigzuordnung nicht verändert wurde und damit die Fallzahlen unverändert geblieben sind, ergeben sich für die Umsätze leichte Veränderungen, die jedoch mit max. 2,4% im Wirtschaftszweig G Handel, Instandhaltung und Reparatur v. Kraftfahrzeugen u. Gebrauchsgütern gering sind. Auch die weiteren quantitativen Merkmale des faktisch anonymisierten Datenfiles weichen nur in wenigen Fällen in größerem Maße von den Originalwerten ab.

Bei einer Betrachtung nach Umsatzgrößenklassen wird der Einfluss der Mikroaggregation auf die quantitativen Merkmale der Großunternehmen deutlich, während die Ergebnisse für die Vielzahl der kleinen und mittleren Unternehmen unverändert bleiben. Tabelle 36.13 zeigt die Anzahl der Steuerpflichtigen und deren *Lieferungen und Leistungen* für Originaldaten und anonymisierte Daten nach Umsatzgrößenklassen. Die Ergebnisse für die Unternehmen mit einem Umsatz von unter 100 Millionen Euro weisen keine nennenswerten Unterschiede zwischen der Umsatzsteuerstatistik und dem faktisch anonymisierten Datenfile auf. Lediglich bei den Unternehmen mit einem Umsatz von mehr als 100 Mill. Euro werden durch das Anonymisierungskonzept leicht unterschiedliche Ergebnisse bewirkt.

Tabelle 36.13: Steuerpflichtige, Lieferungen und Leistungen nach Umsatzgrößenklassen

Größenklassen der Lieferungen und Leistungen von ... bis unter ... EUR	Steuerpflichtige 1)			Lieferungen und Leistungen 2)		
	Original	Anonymisiert	Abweichung	Original	Anonymisiert	Abweichung
	Anzahl		%	Mill. EUR		%
16.617 - 50.000	773.820	773.821	0	24.278	24.278	0
50.000 - 100.000	568.170	568.167	0	40.959	40.959	0
100.000 - 250.000	662.982	662.986	0	105.968	105.969	0
250.000 - 500.000	357.106	357.101	0	126.237	126.235	0
500.000 - 1.Mio.	238.229	238.230	0	167.420	167.419	0
1.Mio. - 2.Mio.	143.908	143.910	0	200.770	200.772	0
2.Mio. - 5.Mio.	93.323	93.321	0	286.776	286.773	0
5.Mio. - 10.Mio.	34.524	34.517	0	240.764	240.700	0
10.Mio. - 25.Mio.	21.591	21.598	0	331.512	331.719	-0,1
25.Mio. - 50.Mio.	7.786	7.769	0,2	270.984	270.345	0,2
50.Mio. - 100.Mio.	3.874	3.879	-0,1	268.392	268.565	-0,1
100.Mio. - 250.Mio.	2.352	2.346	0,3	359.880	358.363	0,4
250.Mio. - und.mehr	1.485	1.505	-1,3	1.728.987	1.730.831	-0,1
Zusammen	2.909.150	2.909.150	0	4.152.927	4.152.927	0

1) Mit mehr als 16.617 Euro Jahresumsatz 2) Umsätze der Unternehmen - ohne Umsatzsteuer.

b) Konzentrationsanalysen

Eine nahe liegende Untersuchungsmöglichkeit, welche die Umsatzsteuerstatistik bietet, ist die Analyse der Umsatzverteilung auf die Unternehmen und davon abgeleitet die Untersuchung der Unternehmenskonzentration. Problematisch ist die Auswahl der geeigneten Messmethoden. Grundsätzlich existieren absolute und relative Konzentrationsmaße. Grob formuliert besteht der Unterschied darin, „dass von einer absoluten Konzentration dann gesprochen wird, wenn ein Großteil des gesamten Merkmalsbetrages auf eine kleine Zahl an Merkmalsträgern entfällt, von einer relativen Konzentration, wenn ein Großteil des gesamten Merkmalsbetrages auf einen kleinen Anteil der Merkmalsträger entfällt".

b1) Absolute Konzentration

Am einfachsten lässt sich die absolute Konzentration als Anteil der m größten Unternehmen am gesamten Merkmal berechnen (Konzentrationsrate CRm). Allerdings bleiben dabei sämtliche Informationen außen vor, die über die anderen (kleineren) Unternehmen bekannt sind. Veränderungen in den Marktanteilen dieser Wettbewerber finden keine Berücksichtigung. Der Vorteil des Maßes liegt in der einfachen Berechenbarkeit, weshalb sich dieses Maß weiterhin großer Beliebtheit erfreuen.

Ein Alternativmaß, um die Informationen aller am Markt tätigen Unternehmen zu nutzen, ist der Herfindahlindex. Dieser Index ist gleich der Summe der quadrierten Marktanteile. Bei absoluter Konzentration erreicht der Index den Wert 1, bei absoluter Gleichverteilung den Wert 1/(Anzahl der Unternehmen). Damit kommt zum Ausdruck, dass die Anzahl der am Markt auftretenden Unternehmen nicht unerheblich ist. Je größer die Anzahl der Unternehmen ist, desto geringer ist der Indexwert bei einer Gleichverteilung. Durch die Definition werden Unternehmen mit hohen Marktanteilen stärker gewichtet als solche mit niedrigeren.

Der größte Vorteil der Umsatzsteuerstatistik liegt in ihrer fast vollständigen Erfassung der Unternehmen. Damit ist der gegenüber den einfacheren Konzentrationsraten aussagefähigere Herfindahlindex, problemlos zu berechnen. Schwierigkeiten bereitet allerdings die aufgrund des Anonymisierungskonzeptes durchgeführte punktuelle Mikroaggregation der drei führenden Unternehmen. Während diese bei den Konzentrationsraten keinen Einfluss auf das Ergebnis hat, solange ein m mit mindestens drei gewählt wird, führt sie beim Herfindahlindex zu einer Unterschätzung. Diese ist umso höher, je höher der Marktanteil des führenden Unternehmens im Verhältnis zu den nächst größeren ist. Je größer dieser Abstand, desto stärker wird sein Marktanteil durch die Mikroaggregation reduziert. Diese Reduzierung senkt überproportional das Ergebnis – wegen der überproportionalen Gewichtung der Großunternehmen – und damit den Indexwert.

Tabelle 36.14: Absolute Konzentrationsmaße der Wirtschaftszweige des Abschnitts K „Grundstücks- und Wohnungswesen, Vermietung beweglicher Sachen, Erbringung von wirtschaftlichen Dienstleistungen, anderweitig nicht genannt"

WZ	CR3 anonymisiert	CR3 original	CR3 Differenz	CR5 anonymisiert	CR5 original	CR5 Differenz	Herfindahlindex anonymisiert	Herfindahlindex Original	Herfindahlindex Differenz
70	2,27	2,59	-0,32	3,31	3,47	-0,16	5,70	6,20	0,46
71	19,25	20,46	-1,21	25,49	26,54	-1,05	182,00	213,9	31,82
72	6,25	6,22	0,03	8,89	8,86	0,03	28,80	30,10	1,36
73	38,77	38,73	0,04	42,61	42,98	-0,37	533,20	1104,90	571,75
741	14,1	13,98	0,12	17,44	17,33	0,11	84,50	103,60	19,13
742	1,66	1,66	0,00	2,55	2,55	0,00	3,70	3,80	0,02
743	35,18	35,11	0,07	48,36	48,32	0,04	554,2	556,4	2,20
744	11,53	11,61	-0,08	15,79	15,86	-0,07	70,1	72,8	2,75
745	13,61	13,58	0,03	18,16	18,12	0,04	85,7	98,8	13,15
746	18,48	18,48	0,00	22,46	22,45	0,01	141,3	205,6	64,3
747	2,36	2,52	-0,16	3,36	3,64	-0,28	9,70	10,00	0,27
748	8,23	8,36	-0,13	11,58	11,86	-0,28	46,60	48,80	2,19

Beispielhaft wurde aus der Umsatzsteuerstatistik sowohl aus den Original- als auch aus den anonymisierten Daten die beschriebenen absoluten Konzentrationsmaße für die Wirtschaftszweige des Abschnitts K der WZ 93 berechnet (vgl. Tabelle 36.14). Die Konzentrationsraten wurden hierbei für $m = 3$ und $m = 5$ ermittelt. Während diese nur minimale bis keine Veränderungen aufgrund der Anonymisierung aufweisen, wird der Herfindahlindex wie erwartet bei steigender Konzentration unterschätzt. Die Korrelation zwischen Unterschätzung des Herfindahlindex und Höhe der Konzentrationsraten liegt über alle Wirtschaftszweige berechnet bei 0,75. Werden daher Konzentrationsuntersuchungen mit absoluten Konzentrationsmaßen mit Hilfe der anonymisierten Umsatzsteuerstatistik durchgeführt, so sollte neben dem theoretisch aussagefähigeren Herfindahlindex ebenfalls Konzentrationsraten berechnet werden, um die Unterschätzung des Herfindahlindex bei konzentrierten Branchen einschätzen zu können.

b2) Relative Konzentration

Mit der relativen Konzentration, die im Folgenden mit Hilfe des Gini-Koeffizienten berechnet wird, kann die Verteilung innerhalb einer Gruppe untersucht werden. Aufgrund seiner Definition reagiert der Gini-Koeffizient weit weniger auf die Veränderung einzelner Merkmalswerte als die absoluten Konzentrationsmaße. Diese These wird durch die Beispielrechnungen gestützt (vgl. Tabelle 36.15).

Tabelle 36.15: Gini-Koeffizienten für die Wirtschaftszweige des Abschnitts K „Grundstücks- und Wohnungswesen, Vermietung beweglicher Sachen, Erbringung von wirtschaftlichen Dienstleistungen, anderweitig nicht genannt“

Branche	Gini anonymisiert	Gini Original	Relative Differenz in %
741	0,8665	0,8663	0,02
742	0,7443	0,7444	0,01
743	0,9308	0,9308	0
744	0,8745	0,8747	0,02
745	0,7577	0,7576	0,01
746	0,8661	0,8661	0
747	0,7714	0,7715	0,01
748	0,8852	0,8856	0,05

Die Anonymisierungsmaßnahmen bewirken beim Gini-Koeffizienten nur minimale bzw. keine Veränderungen. Dies gilt nicht nur für die dargestellten, sondern auch für die weiteren im Scientific-Use-File enthaltenen Branchen. Untersuchungen der relativen Konzentration werden demnach durch die Anonymisierung nicht beeinträchtigt und können daher ohne Einschränkungen durchgeführt werden.

c) Exportquoten

Tabelle 36.16 zeigt die aus den *steuerfreien Lieferungen und Leistungen mit Vorsteuerabzug* und den *Lieferungen und Leistungen* berechneten Exportquoten der im anonymisierten Datenfile nachgewiesenen Wirtschaftszweige des Wirtschaftsabschnitts K „Grundstücks- und Wohnungswesen, Vermietung beweglicher Sachen, Erbringung von wirtschaftlichen Dienstleistungen, anderweitig nicht genannt". Die größte Umsatzbedeutung haben die Exporte in den Bereichen 73 „Forschung und Entwicklung" und den Unternehmen der „Rechts-, Steuer- und Unternehmensberatung" (WZ 74.1).

Tabelle 36.16: Exportquoten der Wirtschaftszweige des Abschnitts K „Grundstücks- und Wohnungswesen, Vermietung beweglicher Sachen, Erbringung von wirtschaftlichen Dienstleistungen, anderweitig nicht genannt“

Pos.-Nr der WZ 93	Wirtschaftsgliederung	Exportquote 1) (%)	
		Original	Anonymisiert
70	Grundstücks- und Wohnungswesen	3,2	3,2
71	Vermietung beweglicher Sachen ohne Bedienungspersonal	7,7	7,8
72	Datenverarbeitung und Datenbanken	5,0	5,0
73	Forschung und Entwicklung	41,3	38,6
741	Rechts-, Steuer- u. Unternehmensberatung	13,8	14,0
742	Architektur- und Ingenieurbüros	3,4	3,4
743	Technische, physikalische und chemische Untersuchung	3,7	3,7
744	Werbung	1,9	1,9
745	Gewerbsmäßige Vermittlung und Überlassung von Arbeitskräften	0,4	0,4
746	Detekteien und Schutzdienste	2,4	2,4
747	Reinigung von Gebäuden, Inventar und Verkehrsmitteln	0,9	0,9
748	Erbringung von sonstigen Dienstleistungen überwiegend für Unternehmen	10,0	9,9
K	Zusammen	7,5	7,5

1) Verhältnis des steuerfreien Umsatzes mit Vorsteuerabzug zum Gesamtumsatz in Prozent.

Interessante Analysemöglichkeiten ergeben sich auch durch die Auswertung der Exportquoten nach Umsatzgrößenklassen. Tabelle 36.17 zeigt die Exportquoten nach Umsatzgrößenklassen über alle Wirtschaftszweige. Dabei wird erwartungsgemäß eine deutlich stärkere Exportorientierung der Großunternehmen deutlich. Bei den Unternehmen mit mehr als 250 Mio. Euro Umsatz wird nahezu jeder vierte Euro mit dem Ausland umgesetzt. Aber auch die mittelständischen Unternehmen der Größenklasse 25 bis unter 50 Mio. Euro erwirtschaften 17% ihrer Umsätze mit dem Ausland.

Tabelle 36.17: Exportquoten nach Größenklassen

Größenklasse der Lieferungen und Leistungen von ... bis unter ... EUR	Exportquote 1)	
	Original %	Anonymisiert %
Unter 1 Mio.	1,8	1,8
1 Mio. - 5 Mio.	5,5	5,5
5 Mio. - 25 Mio.	11,3	11,4
25 Mio. - 50 Mio.	16,1	16,1
50 Mio. - 100 Mio.	17,9	17,9
100 Mio. - 250 Mio.	20,1	20,3
250 Mio. und mehr	23,4	23,3
Insgesamt	16,1	16,1

1) Verhältnis des steuerfreien Umsatzes mit Vorsteuerabzug zum Gesamtumsatz in Prozent.

Neben diesen Auswertungen können die Exportquoten auch mit allen anderen qualitativen Merkmalen des Mikrodatenfiles kombiniert werden (z.B. Rechtsform oder Regionalgliederung). Zudem werden die steuerfreien Umsätze mit Vorsteuerabzug für innergemeinschaftliche Lieferungen an Abnehmer mit Umsatzsteuer-Identifikationsnummer (§4 Nr. 1b UStG) gesondert nachgewiesen, so dass sich auch die EU-Exporte gezielt untersuchen lassen.

d) Bedeutung des ermäßigten Mehrwertsteuersatzes

Die Frage, welche Branchen am meisten von einem ermäßigten Umsatzsteuersatz profitieren, ist ebenfalls eine steuerpolitisch interessante Fragestellung, die mit den Daten der Umsatzsteuerstatistik beantwortet werden kann. Tabelle 36.18 zeigt die 10 Branchen, welche die höchsten Anteile ihres Umsatzes mit Produkten erwirtschaften, die dem ermäßigten Steuersatz von 7% unterliegen. Die Anteile wurden wiederum einmal mit den anonymisierten Daten und einmal mit den Originaldaten berechnet. Die Ergebnisse unterscheiden sich dabei nur unerheblich.

Tabelle 36.18: Die zehn Branchen mit dem höchsten Umsatzanteil zu 7% Umsatzsteuer

WZ	Branche	Anteil anonymisiert	Anteil original
41	Wasserversorgung	0,595	0,595
DA	Ernährungsgewerbe und Tabakverarbeitung	0,577	0,575
522	Facheinzelhandel mit Nahrungsmittel usw. (in Verkaufsräumen)	0,552	0,553
A,B	Land- und Forstwirtschaft, Fischerei u. Fischzucht	0,494	0,494
521	Einzelhandel mit Waren verschiedener Art (in Verkaufsräumen)	0,45	0,462
22	Verlags-, Druckgewerbe, Vervielfältigung	0,26	0,263
91	Interessenvertretung, kirchliche und sonstige religiöse Vereinigungen	0,214	0,214
55B	Gaststättengewerbe und Kantinen	0,201	0,201
92	Kultur, Sport und Unterhaltung	0,199	0,196
51	Handelsvermittlung und Großhandel (ohne Kfz)	0,166	0,166

36.6 Fazit für die faktische Anonymisierung der Umsatzsteuerstatistik

Bei der faktischen Anonymsierung der Umsatzsteuerstatistik 2000 ist großer Wert darauf gelegt worden, die Anonymisierung auf das nötige Maß zu beschränken. Die Attraktivität der Daten für die Wissenschaft sollte nicht unnötig eingeschränkt werden. Aus diesem Grund wurde statt einer reinen traditionellen Anonymisierung mit der Mikroaggregation ein Verfahren der so genannten datenverändernden Verfahrensgruppe gewählt. Eine reine traditionelle Anonymisierung wäre nur unter hohem Verlust an Analysepotenzial möglich gewesen. Die Art der Anonymisierung sichert, dass nur diejenigen Merkmalsträger anonymisiert werden, die auch besonders schutzbedürftig sind.

Die Analyse der Schutzwirkung zeigt, dass die faktische Anonymisierung aufgrund der getroffenen Maßnahmen und des natürlichen Schutzes der Daten erreicht wurde. Dies gilt auch für die größten Unternehmen, die aufgrund der getrennten Mikroaggregation mit Ausnahme der Marktführer geschützt werden konnten. Für die Marktführer wiederum ist eine stärkere punktuelle Mikroaggregation als Schutzmaßnahme nötig. Die Vertraulichkeit der Daten wird durch diese Anonymisierung gewährleistet. Daher können auch die größten Un-

ternehmen Bestandteil des veröffentlichten Scientific-Use-Files sein. Dass die Anonymisierung auf das notwendigste beschränkt ist, bedeutet aber auch, dass eine weitere Lockerung der Anonymisierung ohne ein unakzeptables Enthüllungsrisiko nicht möglich erscheint.

Durch das Vorgehen bei der Anonymisierung, und hier insbesondere durch die punktuelle Mikroaggregation der Branchenführer, die aus Sicht des Datenschutzes unbedingt erforderlich ist, treten aus Sicht des Analysepotenzials drei Probleme auf:

- Auswertungen nach der qualitativen Variablen Rechtsform sind stärker verzerrt als andere mögliche Auswertungen von Teilgesamtheiten. Dieses Problem ist ausschließlich auf die punktuelle Mikroaggregation zurückzuführen. *Lösung*: Keine Auswertungen nach der Variablen „Rechtsform" oder Verzicht auf die drei Branchenführer beziehungsweise Einführen einer Abschneidegrenze nach oben!
- Die Varianzen und auch die Korrelationskoeffizienten sind durch die punktuelle Mikroaggregation der Branchenführer und die getrennte Mikroaggregation bei den großen Unternehmen teilweise sehr stark verzerrt. Etwa die Hälfte der Varianzverzerrung ist auf die punktuelle Mikroaggregation zurückzuführen, die andere Hälfte auf die allgemeine getrennte Mikroaggregation. Die Varianzverzerrung fällt erheblich stärker aus, wenn größere Unternehmen in die Analysen einbezogen werden. *Lösung*: Verzicht auf Streuungsvergleiche bei Einbeziehung der Branchenführer bzw. der Großunternehmen, Einführen einer Abschneidegrenze nach oben ggf. auch bei Regressionsschätzungen (Ausreißerproblematik), Verwendung der Rangkorrelationen statt der Pearson-Korrelationen.
- Absolute Konzentrationsmaße, wie der Herfindahl-Index, bei denen die größten drei Unternehmen einer Branche besonderes Gewicht erhalten, sind verzerrt. *Lösung*: Ausweichen auf andere Konzentrationsmaße, die robust gegenüber der punktuellen Mikroaggregation sind, wie Konzentrationsraten und relative Konzentrationsmaße, Schätzung des Herfindahl-Index aus den Konzentrationsraten.

Durch die Kennzeichnung der punktuell mikroaggregierten Unternehmen im Scientific-Use-File ist für die Datennutzer transparent, in welchen Bereichen mit Ergebnisveränderungen beziehungsweise mit Einschränkungen bei den Nutzungsmöglichkeiten zu rechnen ist. Notwendig ist zudem ein Hinweis an die Nutzer, welcher Art die Einschränkungen des Analysepotenzials sind. Werden die Einschränkungen klar kenntlich gemacht, so steht einer Freigabe des Datensatzes aus der Sicht des Analysepotenzials nichts im Wege.

Kapitel 37

Anonymisierung der Einzelhandelsstatistik

37.1 Besonderheiten bei der Anonymisierung der Einzelhandelsstatistik

Wie bereits in Unterabschnitt 9.3.2 erwähnt, umfasst die Stichprobe der Einzelhandelsstatistik (EHS) für das Jahr 1999 rund 23.500 Merkmalsträger. Damit ist diese Erhebung in der Größenordnung der Kostenstrukturerhebung im Verarbeitenden Gewerbe (knapp 17.000 Merkmalsträger) anzusiedeln, während die Umsatzsteuerstatistik (USt) als Vollerhebung über 2,9 Mio. Merkmalsträger enthält. Von den drei Erhebungen war das Gelingen einer faktischen Anonymisierung bei der EHS im Bereich der mittleren und großen Unternehmen am schwierigsten einzuschätzen. Dies liegt insbesondere an der im Vergleich zur KSE schieferen Größenverteilung der Unternehmen. Bereits in der Beschäftigtengrößenklasse zwischen 100 und 249 Mitarbeitern wird die Besetzungszahl der beschäftigungsstärksten Unternehmen der KSE (mindestens 1.000 Beschäftigte) unterschritten. Bemerkenswert ist auch die Tatsache, dass es sich bei knapp 84 Prozent der Unternehmen mit mindestens 50 Beschäftigten schon um eine Vollerhebung handelt. Das bedeutet, dass bei diesen Unternehmen im Allgemeinen von einer Teilnahmekenntnis eines potenziellen Datenangreifers ausgegangen werden muss.

Im Gegensatz zu den anderen beiden betrachteten Erhebungen sind in der Einzelhandelsstatistik insgesamt 63 Merkmale vorhanden, die Auskunft über Umsatzanteile in Prozentangaben nach folgenden Gesichtspunkten geben:

- Einzelhandelsumsatz nach Absatzformen (4 Merkmale)
- Grobe Gliederung des Umsatzes nach Tätigkeitsbereich (4 Merkmale)
- Feine Gliederung des Umsatzes nach Tätigkeiten bzw. Produkten (55 Merkmale)

Dabei stellt der letzte Gesichtspunkt eine enorme Reidentifikationsgefahr dar. Unternehmen

mit mehreren Tätigkeitsfeldern werden bei Kenntnis eines Datenangreifers im Allgemeinen zu einmaligen Fällen. Beispielsweise lässt sich diese Information mit der im nächsten Kapitel vorgestellten MARKUS-Datenbank gewinnen. Daher wurden diese Merkmale aus Datenschutzgründen entfernt.

37.2 Verfügbares Zusatzwissen und Überprüfung der Datensicherheit

In diesem Abschnitt werden die verfügbaren Informationsquellen an Zusatzwissen und potenzielle Überschneidungsmerkmale vorgestellt. Während sich die M+M-Deutsche Handelsdatenbank als wenig nützlich erweist, stellen die kommerziell erwerbliche MARKUS-Datenbank und persönliche Internetrecherchen brauchbares Zusatzwissen dar. Analog zur Vorgehensweise bei der Kostenstrukturerhebung wurde die Umsatzsteuerstatistik zur ersten Einschätzung von Risiken anonymisierter Daten bei Massenfischzugszenarien als Zusatzwissen verwendet.[77] Die Informationsbeschaffung über das Internet stellt für einen gezielten Einzelangriff eine äußerst wichtige Möglichkeit dar. Darauf wurde in Unterabschnitt 37.4.4 genauer eingegangen. Insbesondere können auf diesem Weg Informationen zum Merkmal *Anzahl der Filialen* gewonnen werden.

37.2.1 M+M Deutsche Handelsdatenbank

Die M+M Deutsche Handelsdatenbank enthält rund 300 Firmenporträts von allen wichtigen Firmen und Organisationen des Lebensmittelhandels. Dabei sind folgende Überschneidungsmerkmale vorhanden:

- Umsatz
- Filialen
- Regionalinformation

Daneben gibt es Informationen zu folgenden Sachverhalten:

- Firmenname und Adresse
- Entscheidungsträger
- Aufteilung Food/Nonfood
- Nationale/internationale Kooperationen

77) Siehe Unterabschnitt 37.4.2.

- Verkaufsflächen
- Zusatzinformationen (Einkaufs-, Sortiments- und Preispolitik, Eigenmarken, Unternehmensstruktur usw.)

Während in der Einzelhandelsstatistik rechtlich selbständige Einheiten erfasst werden (Unternehmensprinzip), ist bei der M+M-Datenbank jedoch „nur" die jeweilige Firma vorhanden (Gruppenprinzip). Zum Beispiel könnte eine Firma in der M+M-Datenbank auch als solche geführt wird, während sie in der EHS mit zehn rechtlich selbständigen Einheiten erscheint. Aus diesem Grund wird eine mögliche Reidentifikation mit dem Zusatzwissen der M+M-Datenbank a priori enorm erschwert. Sowohl durchgeführte Massenfischzugs- als auch Einzelangriffsexperimente haben gezeigt, dass eine korrekte Zuordnung nahezu ausgeschlossen werden kann.

37.2.2 MARKUS-Datenbank

Die MARKUS-Datenbank liefert Geschäftsinformationen zu ausgewählten Unternehmen der Creditreform. Folgende Überschneidungsmerkmale sind hier vorhanden:[78]

- Wirtschaftszweigklassifikation
- Regionalinformation
- Umsatz
- Beschäftigte
- Umsatz nach Tätigkeiten bzw. Produkten (in %)

Darüber hinaus lassen sich folgende Informationen gewinnen:

- Firmenname und Adresse
- Geschäftsführer und Vorstände
- Bilanzangaben
- Stammkapital
- Beteiligungsstruktur
- Tätigkeitsbeschreibung

78) In manchen Fällen lässt sich auch die Anzahl der Filialen eines Unternehmens feststellen.

Für das Gelingen einer erfolgreichen Reidentifikation ist die Qualität der Überschneidungsmerkmale maßgeblich. Bei verschiedener Klassifizierung der kategorialen Merkmale von Merkmalsträgern in Zieldaten und Zusatzwissen wird eine korrekte Zuordnung a priori ausgeschlossen, wenn der Datenangreifer entsprechende Blockungen vornimmt. Durch Abweichungen in den metrischen Merkmalen wird eine korrekte Zuordnung zwar nicht ausgeschlossen, aber erschwert.

Tabelle 37.1 zeigt die Anteile gleich und verschieden klassifizierter Merkmalsträger in den kategorialen Überschneidungsmerkmalen *Regionalinformation* (BBR3, BBR9) und *Wirtschaftszweigklassifikation* (Dreisteller- und Vierstellerebene). Die Anteile beziehen sich auf eine Grundgesamtheit von 8.199 Merkmalsträgern, welche sowohl in der Einzelhandelsstatistik als auch in der MARKUS-Datenbank enthalten sind und für die direkte Identifikatoren ausgemacht werden konnten.

Tabelle 37.1: Vergleich kategorialer Überschneidungsmerkmale bei Originaldaten und MARKUS-Datenbank

Merkmal	BBR3	BBR9	WZ-Dreisteller	WZ-Viersteller
Gleich klassifiziert	8.146 (99%)	8.010 (98%)	6.972 (85%)	6.071 (74%)
Verschieden klassifiziert	53 (1%)	189 (2%)	1.227 (15%)	2.128 (26%)

Während bei den Regionalkennungen BBR9 und BBR3 nur relativ geringe Abweichungen beobachtet werden können, lassen sich bei den Wirtschaftszweigklassifikationen nennenswerte Abweichungen feststellen. Allein die unterschiedlichen Angaben bezüglich des *WZ-Vierstellers* führen in diesem Fall dazu, dass für ein Viertel der gesuchten Unternehmen ein natürlicher Schutz vor einem Datenangriff besteht. Ein Datenangreifer wird deshalb diese Unternehmen entweder gar nicht oder nur falsch zuordnen können.

In die Distanzberechnungen innerhalb des im Projekt verwendeten Matchingalgorithmus gehen die metrischen Merkmale *Umsatz* und *Beschäftigte* ein. Tabelle 37.2 zeigt die relativen Abweichungen in diesen Merkmalen zwischen Zusatzwissen (MARKUS) und Zieldaten (EHS). Sie weist beim Merkmal *Umsatz* nur gut ein Fünftel der Unternehmen eine relative Abweichung unter 10% von ihrem entsprechenden Originalwert auf. Im Falle des Merkmals *Beschäftigte* sind dies gut ein Viertel der Unternehmen. Bemerkenswert ist auch die Tatsache, dass über ein Viertel der Umsatzwerte über 50% relativ zum entsprechenden Originalwert abweichen. In etwas abgeschwächter Form lässt sich diese Aussage auch auf das Merkmal *Beschäftigte* übertragen.

Neben den Abweichungen bei dem kategorialen Merkmal *Wirtschaftszweigklassifikation* besteht durch die Abweichungen der wichtigen Überschneidungsmerkmale *Umsatz* und *Beschäftigte* bereits ein nicht zu vernachlässigender natürlicher Schutz der Originaldaten gegenüber diesem Zusatzwissen.

Tabelle 37.2: Vergleich metrischer Überschneidungsmerkmale bei Originaldaten und MARKUS-Datenbank

Merkmal	Umsatz	Anzahl der Beschäftigten
Rel. Abweichung unter 10%	1.605 (20,5%)	2.089 (26%)
Rel. Abweichung von 10% bis 25%	2.096 (26,5%)	1.654 (20%)
Rel. Abweichung über 25% bis 50%	2.231 (27%)	2.648 (32%)
Rel. Abweichung über 50%	2.267 (28%)	1.808 (22%)

37.2.3 Überschneidungsmerkmale bei der Einzelhandelsstatistik

Unter Berücksichtigung der vorangehenden Unterabschnitte des laufenden Abschnitts sind folgende Überschneidungsmerkmale in externen, realistischen Quellen des Zusatzwissens vorhanden:

- Regionalinformation
- Wirtschaftszweigklassifikation
- Umsatz
- Beschäftigte
- Anzahl der Filialen
- Umsatz nach Tätigkeiten bzw. Produkten (in %)

Dabei ist jedoch nicht auszuschließen, dass ein potenzieller Datenangreifer mit der Hilfe von im Projekt unberücksichtigtem Zusatzwissen über ein oder mehrere weitere Überschneidungsmerkmale verfügen kann.

37.3 Anonymisierungsmaßnahmen

Eine geeignete Anonymisierungsstrategie kann aus traditionellen und/oder datenverändernden Verfahren bestehen.

Getestet wurde in den Untersuchungen die Wirkung folgender datenverändernder Anonymisierungsverfahren:[79]

79) Andere datenverändernde Verfahren, die sich bereits in früheren Simulationen bei der USt und KSE als nicht aussichtsreich erwiesen, wurden nicht verfolgt.

- Formale Anonymisierung (FORMAL),
- eindimensionale getrennte Mikroaggregation über 32 Gruppen (MA32G),
- blockweise Anwendung der mehrdimensionalen Mikroaggregation über 9 Gruppen (MA9G),[80]
- mehrdimensionale gemeinsame Mikroaggregation über 1 Gruppe (MA1G),
- Additive Überlagerung der logarithmierten Werte mit mehrdimensionaler Normalverteilung. Die Varianz-Kovarianz-Matrix der Überlagerung entspricht dem 0,0005-fachen der Kovarianz-Matrix der logarithmierten Werte (Mult_Wink),
- Multiplikative Überlagerung mit einer Mischungsverteilung $W \sim N(1 \pm f, s)$. Konstanter Überlagerungsfaktor für die metrischen Merkmale eines Merkmalträgers mit Erwartungswert $1+f$ bzw. $1-f$ (je nach Mischungskomponente) und der Standardabweichung s (innerhalb der Komponente) (Mult_f_s),
- Multiplikative Überlagerung mit einer „Mischungsverteilung" nach dem Verfahren von Höhne mit den Parametern f für die Mittelwerte der beiden Komponenten und s für deren Streuung (Hoe_f_s).

Die Verfahren der Mikroaggregation werden in Unterabschnitt 6.2.4 erläutert. Da die Varianten der Zufallsüberlagerung erst in der Projektendphase entwickelt wurden, sind sie erst in späteren Analysen getestet worden. Eine detailliertere Beschreibung der Varianten aus der Verfahrensgruppe der Zufallsüberlagerungen wird in Unterabschnitt 6.2.3 vorgenommen.

37.4 Überprüfung der Schutzwirkung

In diesem Abschnitt wird die Schutzwirkung bei der Wahl von verschiedenen Mischungen an Anonymisierungsmaßnahmen – angewendet auf die Daten der Einzelhandelsstatistik – untersucht. In Unterabschnitt 37.4.1 wird zunächst der Einfluss des Merkmals *Anzahl der Filialen* auf das Reidentifikationsrisiko mit Hilfe von simulierten Massenfischzugsszenarien untersucht. Unterabschnitt 37.4.2 befasst sich intensiv mit dem Verfahrenskomplex der Mikroaggregation. Dabei werden Matchingexperimente mit verschiedenen Quellen an Zusatzwissen und Variation der verwendeten Überschneidungsmerkmale beschrieben. Ein Verfahrensvergleich zwischen den im vorigen Unterabschnitt untersuchten Mikroaggregationsvarianten und Varianten der Zufallsüberlagerung wird in Unterabschnitt 37.4.3 erläutert. Schließlich wird in Unterabschnitt 37.4.4 auf die nicht zu vernachlässigende Gefahr von Einzelangriffen eingegangen.

80) Die Einteilung der Gruppen wurde aufgrund der Ergebnisse von Clusteranalysen vorgenommen.

37.4.1 Einfluss des Merkmals Anzahl der Filialen auf das Reidentifikationsrisiko

Da das Merkmal *Anzahl der Filialen* gut über Internetrecherchen gewonnen werden kann, sollte es unbedingt als potenzielles Überschneidungsmerkmal angesehen werden.[81] Bei den hier simulierten Datenangriffen via Massenfischzug wurden die Originaldaten als Zusatzwissen verwendet. Damit ist es möglich, den Einfluss des Merkmals *Anzahl der Filialen* zu untersuchen, obwohl dieses Merkmal nicht bzw. nur eingeschränkt[82] in den anderen beiden verwendeten Quellen des Zusatzwissens vorhanden ist. Wie in Abschnitt 11.2 erwähnt wurde, lässt sich mit einem Massenfischzug die globale Wirkung von Anonymisierungsmaßnahmen feststellen.

Im Folgenden wird untersucht, welche Auswirkungen die Hinzunahme des Merkmals *Anzahl der Filialen* auf das Reidentifikationsrisiko hat. Ebenso wurde geprüft, ob mit einer Einteilung dieses Merkmals in Kategorien[83] das Risiko verringert werden kann. Als Regionalinformation wurde der siedlungsstrukturelle Kreistyp BBR9 und als Wirtschaftszweigklassifikation die Ebene des Dreistellers benutzt. Getestet wurden die drei Mikroaggregationsvarianten MA1G, MA9G und MA32G. Tabelle 37.3 enthält die relativen Trefferquoten des Massenfischzugsszenarios nach Beschäftigtengrößenklassen unter Verwendung des folgenden Kanons an Überschneidungsmerkmalen:

Variante 1: WZ-Dreisteller, Regionalbezug BBR9, Umsatz, Beschäftigte
Variante 2: WZ-Dreisteller, Regionalbezug BBR9, Umsatz, Beschäftigte, Filialkategorien
Variante 3: WZ-Dreisteller, Regionalbezug BBR9, Umsatz, Beschäftigte, Anzahl der Filialen

Da mit dem Verfahren der eindimensionalen getrennten Mikroaggregation schon mit vier Überschneidungsmerkmalen relative Trefferquoten um 100 Prozent (1,00) erzielt werden, ist eine Betrachtung der anderen beiden Mikroaggregationsverfahren für diese Untersuchung aufschlussreicher.

Erwartungsgemäß steigt das Reidentifikationsrisiko mit zunehmender Anzahl von Überschneidungsmerkmalen an.[84] Es lässt sich feststellen, dass die Trefferquoten in den einzelnen Größenklassen im Durchschnitt um das Doppelte ansteigen. Eine Ausnahme bildet die unterste Beschäftigtengrößenklasse. Da diese Klasse sehr stark besetzt ist und gut 93 Prozent der dort vorhandenen Unternehmen nur eine Filiale aufweisen, führt das Merkmal für diese Unternehmen kaum zu einer Reidentifikation.

81) Siehe Unterabschnitt 37.4.4.

82) Vgl. Unterabschnitt 37.2.2.

83) Siehe Tabelle 37.4

84) vgl. Variante 2 bzw. 3 gegenüber 1.

Tabelle 37.3: Relative Trefferquoten nach Beschäftigtengrößenklassen mit den Varianten 1 bis 3

Verfahren	Größenklassen (Beschäftigte)							
	1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
MA1G (1)	0,11	0,10	0,13	0,16	0,21	0,30	0,38	0,11
MA1G (2)	0,13	0,23	0,30	0,40	0,55	0,69	0,64	0,16
MA1G (3)	0,13	0,24	0,33	0,43	0,54	0,66	0,76	0,17
MA9G (1)	0,19	0,24	0,30	0,34	0,49	0,47	0,69	0,21
MA9G (2)	0,23	0,42	0,51	0,64	0,84	0,85	0,85	0,28
MA9G (3)	0,23	0,43	0,55	0,66	0,88	0,80	0,95	0,28

Bei Betrachtung der höchsten Beschäftigtengrößenklasse ist auszumachen, dass die Bildung von Filialkategorien bei dem Verfahren MA9G zu knapp 10 Prozentpunkten und mit MA1G zu rund 12,5 Prozentpunkten das Reidentifikationsrisiko reduziert.

Tabelle 37.4: Kategorien des Merkmals *Anzahl der Filialen*

Zahl der Filialen	Kategorie	Häufigkeit
1	1	11.618
2	2	1.095
3	3	455
4	4	260
5	5	155
6-10	6	360
11-50	7	362
51-100	8	98
$>$ 100	9	79

37.4.2 Detaillierte Analyse der Mikroaggregationsvarianten

In drei Matchingexperimenten werden jeweils die formale Anonymisierung und die drei Mikroaggregationsverfahren mit unterschiedlichem Zusatzwissen angegriffen. Dabei werden zunächst die erfolgreichen Zuordnungsversuche eines potenziellen Datenangreifers betrachtet. Im nächsten Schritt wird die Nützlichkeit der gewonnenen Informationen untersucht. Schließlich werden die sich aus den beiden Sachverhalten ergebenden Enthüllungsrisiken verglichen.

a) Trefferquoten unter Verwendung der Originaldaten

Die Annahme, dass ein Datenangreifer über die Originaldaten verfügt, ist sicherlich nicht realistisch. Dennoch lässt sich damit ein oberes Reidentifikationsrisiko bestimmen. Als Überschneidungsmerkmale wurden die Merkmale *WZ-Dreisteller*, *Regionalbezug BBR9*, *Umsatz* und *Beschäftigte* verwendet. Da maximal diese Merkmale in den anderen beiden benutzten Quellen von Zusatzwissen vorkommen, erscheint dieses Vorgehen aus Gründen der Vergleichbarkeit der Ergebnisse sinnvoll. Tabelle 37.5 stellt die entsprechenden Trefferquoten nach Beschäftigtengrößenklassen dar.

Tabelle 37.5: Relative Trefferquoten nach Beschäftigtengrößenklassen mit Zusatzwissen „Originaldaten“

Verfahren	Größenklassen (Beschäftigte)							
	1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
FORMAL	1	1	1	1	1	1	1	1
MA32G	0,99	1	1	1	0,99	0,98	0,99	0,99
MA9G	0,18	0,25	0,30	0,34	0,50	0,46	0,68	0,20
MA1G	0,09	0,10	0,13	0,16	0,21	0,30	0,38	0,10

Erwartungsgemäß nimmt die Schutzwirkung mit zunehmender Unternehmensgröße und schwächerem Anonymisierungsgrad ab. Während die eindimensionale getrennte Mikroaggregation (MA32G) das Reidentifikationsrisiko kaum verringert, zeigt die mehrdimensionale gemeinsame Mikroaggregation (MA1G) bereits in diesem Experiment eine beachtliche Schutzwirkung. Knapp zwei Drittel der Unternehmen mit mindestens 1.000 Beschäftigten werden nicht korrekt zugeordnet.

b) Trefferquoten unter Verwendung der MARKUS-Datenbank

Mit der Annahme, dass der Datenangreifer über die Unternehmensinformationen der kommerziell erwerblichen MARKUS-Datenbank verfügt, wird ein realistisches Szenario abgebildet. Bei diesem Matchingexperiment konnten 8.199 Unternehmen für den Massenfischzug verwendet werden. Dabei entspricht die Verteilung dieser Unternehmen nach Beschäftigtengrößenklassen weitestgehend der Originalverteilung. Daher sind die in Tabelle 37.6 dargestellten Ergebnisse als aussagekräftig einzustufen. Mit der formalen Anonymisierung (FORMAL) lässt sich der natürliche Schutz der Originaldaten feststellen.

Im Vergleich zum vorigen Szenario nimmt die Schutzwirkung enorm zu. Bereits der natürliche Schutz der Einzelhandelsdaten ist bemerkenswert: Nicht mal jedes Dritte Unternehmen aus der höchsten Beschäftigtengrößenklasse wird reidentifiziert.

Tabelle 37.6: Relative Trefferquoten nach Beschäftigtengrößenklassen mit Zusatzwissen „MARKUS-Datenbank“

Verfahren	Größenklassen (Beschäftigte)							
	1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
FORMAL	0,01	0,04	0,07	0,11	0,19	0,24	0,32	0,03
MA32G	0,01	0,04	0,07	0,11	0,18	0,24	0,31	0,03
MA9G	0,01	0,02	0,06	0,08	0,16	0,19	0,22	0,02
MA1G	0,01	0,01	0,03	0,05	0,12	0,13	0,16	0,02

c) Trefferquoten unter Verwendung der Mikrodaten aus der Umsatzsteuerstatistik

Bei diesem Szenario wird unterstellt, dass dem Datenangreifer Zusatzwissen von der Qualität der amtlichen Einzeldaten aus der Umsatzsteuerstatistik zur Verfügung stehen. Das bisher verwendete Überschneidungsmerkmal *Beschäftigte* kann nicht benutzt werden, da es in der Umsatzsteuerstatistik nicht vorhanden ist. Bei diesem Matchingexperiment konnten 12.102 Unternehmen für den Massenfischzug verwendet werden.

Tabelle 37.7: Relative Trefferquoten nach Beschäftigtengrößenklassen mit Zusatzwissen „Umsatzsteuerstatistik“

Verfahren	Größenklassen (Beschäftigte)							
	1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
FORMAL	0,21	0,23	0,29	0,27	0,39	0,25	0,48	0,22
MA32G	0,21	0,23	0,28	0,29	0,38	0,30	0,45	0,22
MA9G	0,03	0,03	0,06	0,10	0,17	0,13	0,30	0,03
MA1G	0,02	0,02	0,03	0,05	0,07	0,09	0,25	0,02

Tabelle 37.7 zeigt, dass die Trefferquoten einerseits gegenüber dem zweiten Experiment höher sind. Andererseits liegen sie noch deutlich unter denen des ersten Experiments. Dabei sind die Trefferquoten bei den kleineren und mittleren Unternehmen nahezu gleich hoch.

d) Nützlichkeit der gewonnenen Informationen

Neben dem Reidentifikationsrisiko spielt die Brauchbarkeit der gewonnenen Informationen eine entscheidende Rolle bei der Beurteilung der Schutzwirkung von Anonymisierungsverfahren. Dabei wird ein gefundener Einzelwert als nützlich definiert, wenn er weniger als 10 Prozent von seinem Originalwert abweicht.

In den Tabellen 37.8 bis 37.10 sind die relativen Häufigkeiten von brauchbaren Informationen der reidentifizierten Unternehmen aus den vorher beschriebenen drei Szenarien dargestellt.

Tabelle 37.8: Nützlichkeit nach Beschäftigtengrößenklassen mit Zusatzwissen „Originaldaten“

Verfahren	Größenklassen (Beschäftigte)							
	1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
FORMAL	1	1	1	1	1	1	1	1
MA32G	1	1	1	1	1	1	0,97	1
MA9G	0,89	0,84	0,82	0,8	0,79	0,78	0,69	0,87
MA1G	0,86	0,79	0,73	0,64	0,59	0,61	0,55	0,82

Tabelle 37.9: Nützlichkeit nach Beschäftigtengrößenklassen mit Zusatzwissen „MARKUS-Datenbank“

Verfahren	Größenklassen (Beschäftigte)							
	1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
FORMAL	1	1	1	1	1	1	1	1
MA32G	1	1	1	1	1	1	0,95	0,99
MA9G	0,87	0,84	0,82	0,79	0,75	0,76	0,66	0,81
MA1G	0,79	0,72	0,66	0,64	0,56	0,57	0,46	0,67

Tabelle 37.10: Nützlichkeit nach Beschäftigtengrößenklassen mit Zusatzwissen „Umsatzsteuerstatistik“

Verfahren	Größenklassen (Beschäftigte)							
	1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
FORMAL	1	1	1	1	1	1	1	1
MA32G	1	1	1	1	1	1	0,97	1
MA9G	0,89	0,85	0,81	0,78	0,78	0,77	0,67	0,85
MA1G	0,82	0,76	0,71	0,62	0,59	0,66	0,54	0,77

Die Verteilung der Anteile an nützlichen Informationen nach Beschäftigtengrößenklassen bleibt bei dem Wechsel des Zusatzwissens weitestgehend unberührt, wobei mit Verwendung der MARKUS-Datenbank bei den Verfahren MA9G und MA1G systematisch leicht niedrigere Anteile an nützlichen Informationen auszumachen sind. Ausnahme bildet die höchste Größenklasse bei MA1G, bei der dieser Anteil um knapp 10 Prozentpunkte niedriger ausfällt. Die Analyse der drei Tabellen zeigt, dass es für einen Datenangreifer möglich ist, den Anteil der nützlichen Informationen zu schätzen, indem er sein Zusatzwissen entsprechend anonymisiert und eine Nützlichkeitsanalyse durchführt. Dabei steigt die Korrektheit der Schätzung proportional zur Anzahl der Überschneidungsmerkmale.

e) Enthüllungsrisiken bei Variation der kategorialen Überschneidungsmerkmale

Das Risiko der Enthüllung brauchbarer Werte wird wie in Abschnitt 12.3 berechnet. Die Enthüllungsrisiken nach Beschäftigtengrößenklassen für die drei vorher beschriebenen Experimente sind in Tabelle 37.11 dargestellt. Das Gesamtrisiko ergibt sich wie in Kapitel 35 als Konvexkombination zwischen dem Enthüllungsrisiko für das „Worst-Case"-Szenario und dem Enthüllungsrisiko für die beiden realistischen Szenarien:[85]

$$\hat{P}_\gamma = \lambda \hat{P}_{\gamma,wc} + (1-\lambda)\hat{P}_{\gamma,real}$$

für ein geeignetes $\lambda \in [0,1]$.Hier wurde für λ der Wert $0,2$ gewählt. Da die Nützlichkeit nahezu unabhängig von der Wahl des Zusatzwissens ist, reduzieren sich die in den Tabellen 37.5 bis 37.7 dargestellten Risiken systematisch für die drei Experimente.

An dieser Stelle wird darauf eingegangen, wie sich verschiedene Vergröberungen bei den beiden kategorialen Überschneidungsmerkmalen *Regionalinformation* und *Wirtschaftszweig* auf das Enthüllungsrisiko auswirken. Die Ergebnisse dieser simulierten Massenfischzugszenarien sind in den Tabellen 37.11 bis 37.14 zu finden.

Hinsichtlich der Wahl eines geeigneten Mix aus traditionellen und datenverändernden Verfahren lassen sich beispielsweise die beiden folgenden Beobachtungen machen. Mit der Merkmalskombination „*WZ-Viersteller / BBR9*" entfaltet MA1G eine ähnliche Schutzwirkung wie MA9G, wenn bei letzterem nur der *WZ-Dreisteller* und *BBR3* gegeben sind.[86] Als Ausgangspunkt für das zweite Beispiel soll das nutzerfreundlichere datenverändernde Verfahren MA32G und die Merkmalskombination „*WZ-Dreisteller / BBR9*" dienen. Dann lässt sich durch die Vergröberung der Regionalinformation auf den BBR3 – abgesehen vom „Worst-Case"-Fall – mit der Herausgabe der formal anonymisierten Daten in allen Zellenwerten eine Verbesserung des Datenschutzes erzielen.[87]

85) Dabei wurde $\hat{P}_{\gamma,real}$ als Mittelwert der beiden Risiken für die realistischen Experimente berechnet.

86) Dazu betrachte man die Zeilen von MA1G bei Tabelle 37.12 und die Zeilen von MA9G bei Tabelle 37.14, sowie insbesondere die Unternehmen mit mindestens 250 Beschäftigten. Außerdem sollte die Spalte „Total" nicht überbewertet werden, da hier größtenteils die überproportional vorhandenen kleinen Unternehmen einfließen.

87) Vgl. die Zeile MA32G bei Tabelle 37.11 und die Zeile FORMAL bei Tabelle 37.14.

Tabelle 37.11: Enthüllungsrisiken mit *WZ-Dreisteller* und *BBR9*

Variante	Zusatzw.	Größenklassen (Beschäftigte)							
		1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
MA1G	WorstCase	0,08	0,08	0,09	0,10	0,12	0,18	0,21	0,08
MA1G	MARKUS	0,01	0,01	0,02	0,03	0,07	0,07	0,07	0,01
MA1G	USt	0,01	0,01	0,02	0,03	0,04	0,06	0,14	0,02
MA1G	Total	0,02	0,02	0,03	0,04	0,06	0,08	0,12	0,02
MA9G	WorstCase	0,16	0,21	0,25	0,27	0,39	0,36	0,47	0,18
MA9G	MARKUS	0,01	0,02	0,05	0,07	0,12	0,14	0,14	0,02
MA9G	USt	0,02	0,03	0,05	0,08	0,13	0,10	0,20	0,03
MA9G	Total	0,04	0,06	0,09	0,11	0,18	0,17	0,23	0,05
MA32G	WorstCase	0,99	1	1	1	0,99	0,98	0,96	0,99
MA32G	MARKUS	0,01	0,04	0,07	0,11	0,18	0,24	0,29	0,03
MA32G	USt	0,21	0,23	0,28	0,29	0,38	0,30	0,44	0,22
MA32G	Total	0,29	0,30	0,34	0,36	0,42	0,41	0,48	0,30
FORMAL	WorstCase	1	1	1	1	1	1	1	1
FORMAL	MARKUS	0,01	0,04	0,07	0,11	0,19	0,24	0,32	0,03
FORMAL	USt	0,21	0,23	0,29	0,27	0,39	0,25	0,48	0,22
FORMAL	Total	0,29	0,30	0,34	0,35	0,43	0,40	0,52	0,30

Tabelle 37.12: Enthüllungsrisiken mit *WZ-Viersteller* und *BBR9*

Variante	Zusatzw.	Größenklassen (Beschäftigte)							
		1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
MA1G	WorstCase	0,17	0,17	0,19	0,21	0,24	0,28	0,31	0,17
MA1G	MARKUS	0,02	0,03	0,06	0,08	0,08	0,13	0,11	0,03
MA1G	USt	0,04	0,05	0,05	0,06	0,10	0,11	0,19	0,04
MA1G	Total	0,06	0,07	0,08	0,10	0,12	0,15	0,18	0,06
MA9G	WorstCase	0,31	0,39	0,41	0,47	0,54	0,49	0,58	0,33
MA9G	MARKUS	0,03	0,05	0,10	0,12	0,20	0,20	0,18	0,04
MA9G	USt	0,06	0,09	0,13	0,18	0,22	0,17	0,28	0,07
MA9G	Total	0,10	0,13	0,17	0,21	0,28	0,24	0,30	0,11
MA32G	WorstCase	1	1	1	1	1	0,99	0,97	1
MA32G	MARKUS	0,03	0,09	0,15	0,16	0,27	0,28	0,31	0,06
MA32G	USt	0,30	0,39	0,49	0,45	0,50	0,47	0,45	0,32
MA32G	Total	0,33	0,39	0,45	0,44	0,51	0,49	0,50	0,35
FORMAL	WorstCase	1	1	1	1	1	1	1	1
FORMAL	MARKUS	0,03	0,09	0,15	0,16	0,26	0,28	0,34	0,06
FORMAL	USt	0,30	0,39	0,48	0,45	0,52	0,45	0,49	0,32
FORMAL	Total	0,33	0,39	0,45	0,44	0,51	0,49	0,53	0,35

Tabelle 37.13: Enthüllungsrisiken mit *WZ-Viersteller* und *BBR3*

Variante	Zusatzw.	Größenklassen (Beschäftigte)							
		1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
MA1G	WorstCase	0,09	0,10	0,10	0,13	0,14	0,17	0,23	0,09
MA1G	MARKUS	0,01	0,01	0,03	0,04	0,06	0,09	0,09	0,01
MA1G	USt	0,01	0,03	0,03	0,04	0,04	0,08	0,17	0,02
MA1G	Total	0,03	0,04	0,04	0,06	0,07	0,10	0,15	0,03
MA9G	WorstCase	0,18	0,23	0,29	0,33	0,44	0,33	0,47	0,20
MA9G	MARKUS	0,01	0,02	0,07	0,06	0,15	0,14	0,12	0,02
MA9G	USt	0,03	0,05	0,07	0,09	0,17	0,13	0,29	0,03
MA9G	Total	0,05	0,07	0,11	0,12	0,22	0,17	0,26	0,06
MA32G	WorstCase	0,99	1	1	1	0,99	0,98	0,96	0,99
MA32G	MARKUS	0,01	0,05	0,09	0,10	0,20	0,20	0,24	0,03
MA32G	USt	0,20	0,30	0,34	0,37	0,48	0,42	0,50	0,22
MA32G	Total	0,28	0,34	0,37	0,39	0,47	0,44	0,49	0,30
FORMAL	WorstCase	1	1	1	1	1	1	1	1
FORMAL	MARKUS	0,01	0,05	0,09	0,10	0,20	0,20	0,26	0,03
FORMAL	USt	0,20	0,30	0,33	0,37	0,48	0,43	0,51	0,23
FORMAL	Total	0,29	0,34	0,37	0,39	0,47	0,45	0,51	0,30

Tabelle 37.14: Enthüllungsrisiken mit *WZ-Dreisteller* und *BBR3*

Variante	Zusatzw.	Größenklassen (Beschäftigte)							
		1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
MA1G	WorstCase	0,04	0,04	0,05	0,06	0,04	0,12	0,13	0,04
MA1G	MARKUS	0	0,01	0,01	0,02	0,01	0,05	0,05	0
MA1G	USt	0,01	0,01	0,01	0,01	0,01	0,05	0,08	0,01
MA1G	Total	0,01	0,01	0,02	0,02	0,02	0,06	0,08	0,01
MA9G	WorstCase	0,08	0,11	0,15	0,16	0,26	0,21	0,37	0,09
MA9G	MARKUS	0	0,01	0,03	0,03	0,05	0,08	0,13	0,01
MA9G	USt	0,01	0,01	0,01	0,03	0,07	0,10	0,19	0,01
MA9G	Total	0,02	0,03	0,05	0,05	0,10	0,11	0,20	0,03
MA32G	WorstCase	0,98	1	1	1	1	0,99	0,97	0,98
MA32G	MARKUS	0	0,02	0,04	0,08	0,07	0,14	0,22	0,01
MA32G	USt	0,13	0,14	0,17	0,15	0,24	0,21	0,43	0,14
MA32G	Total	0,25	0,26	0,28	0,29	0,32	0,34	0,46	0,26
FORMAL	WorstCase	1	1	1	1	1	1	1	1
FORMAL	MARKUS	0	0,02	0,04	0,07	0,07	0,15	0,28	0,02
FORMAL	USt	0,14	0,15	0,15	0,17	0,26	0,23	0,42	0,15
FORMAL	Total	0,26	0,27	0,28	0,30	0,33	0,35	0,48	0,26

37.4.3 Vergleich von Mikroaggregation und stochastischer Überlagerung

Im Folgenden werden Varianten der Zufallsüberlagerung (Mult_f04_s02, Mult_f08_s02, Mult_f11_s03, Mult_f11_s03_trans, Mult_Wink, Hoe_f04_s02, Hoe_f08_s02, Hoe_f11_s02 und Hoe_f11_s02_trans)[88] mit der formalen Anonymisierung und den drei Mikroaggregationsvarianten MA1G, MA9G und MA32G hinsichtlich ihrer Schutzwirkung verglichen.

Dabei wurden folgende Überschneidungsmerkmale bei den Matchingexperimenten verwendet:

- WZ-Viersteller
- Regionalbezug BBR9
- Umsatz
- Beschäftigte

Als Zusatzwissen wurde die kommerziell erhältliche Markus-Datenbank benutzt. Tabelle 37.15 gibt Auskunft über die Enthüllungsrisiken nach Beschäftigtengrößenklassen. Die Varianten werden absteigend hinsichtlich ihres Enthüllungsrisikos in der obersten Beschäftigtengrößenklasse ($\geq$ 1.000) sortiert.

88) „trans“ in der Bezeichnung der Anonymisierungsvariante bedeutet, dass zusätzlich eine Korrektur der ersten und zweiten Momente vorgenommen wurde (Kim-Korrektur).

Tabelle 37.15: Enthüllungsrisiken nach Beschäftigtengrößenklassen für verschiedene Anonymisierungsverfahren

Verfahren	Größenklassen (Beschäftigte)							
	1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
FORMAL	0,03	0,09	0,15	0,16	0,26	0,28	0,34	0,06
MA32G	0,03	0,09	0,15	0,16	0,27	0,28	0,31	0,06
Mult_f04_s02	0,04	0,06	0,12	0,15	0,20	0,21	0,31	0,06
Hoe_f04_s02	0,04	0,07	0,14	0,17	0,23	0,20	0,31	0,07
Mult_Wink	0,04	0,07	0,11	0,13	0,18	0,18	0,28	0,06
Mult_f08_s02	0,04	0,06	0,11	0,16	0,17	0,13	0,28	0,06
Hoe_f08_s02	0,04	0,06	0,12	0,15	0,20	0,16	0,25	0,06
Mult_f11_s03	0,03	0,03	0,09	0,10	0,12	0,07	0,20	0,04
MA9G	0,03	0,05	0,1	0,12	0,20	0,20	0,18	0,04
Hoe_f11_s02_trans	0,02	0,04	0,08	0,09	0,13	0,10	0,17	0,03
Mult_f11_s03_trans	0,01	0,03	0,08	0,10	0,13	0,09	0,17	0,03
Hoe_f11_s03	0,03	0,04	0,08	0,09	0,12	0,09	0,15	0,04
MA1G	0,02	0,03	0,06	0,08	0,08	0,13	0,11	0,03

Erwartungsgemäß ist die Rangfolge innerhalb der Varianten der Zufallsüberlagerung: Mit höherer Abweichungsschranke vergrößert sich die Schutzwirkung. Die Varianten Mult_f04_s02, Mult_f08_s02, Hoe_f04_s02 und Mult_Wink weisen ähnliche Enthüllungsrisiken wie die eindimensionale Mikroaggregation (MA32G) auf. Demnach leisten sie nur eine geringe Schutzwirkung. Denn bei Betrachtung der Enthüllungsrisiken in den Zellen der Merkmalskombination *Wirtschaftszweig / Regionalinformation / Beschäftigtengrößenklasse* steigen obige Risiken deutlich an und liegen häufig über dem in Abschnitt 37.6 beschriebenen Kriterium für faktische Anonymität. Interessant sind die übrigen Varianten der Zufallsüberlagerung, da sie eine vergleichbare bzw. bessere Schutzwirkung als MA9G erzielen. Bei dieser Variante ist bekannt, dass sie eine weitaus höhere Schutzwirkung gegenüber MA32G besitzt. In den weiteren Untersuchungen stellte sich heraus, dass der Schutz bei der Variante Hoe_f08_s02 ausreicht, um faktische Anonymität gewährleisten zu können, so dass ein Scientific-Use-File für alle Unternehmen der Einzelhandelsstatistik 1999 erstellt werden konnte. Dies wird in Unterabschnitt 37.6.2 beschrieben.

37.4.4 Einzelangriffe

Neben dem Massenfischzug komplettiert der Einzelangriff die beiden relevanten Angriffstechniken. Zur Einschätzung der Gefahr von Einzelangriffen wurde aus dem Originaldatensatz der Einzelhandelsstatistik eine zufällige Stichprobe von 20 Unternehmen mit 1.000 und mehr Beschäftigten entnommen, da diese „großen“ Unternehmen als besonders gefährdet einzustufen sind. Damit wurden Simulationen mit verschiedenen Informationsquellen an Zusatzwissen durchgeführt.

a) Einzelangriffe mit Hilfe des aus dem Internet generierten Zusatzwissens

Um einen Einblick in die Möglichkeiten der eigenen Recherche zu erhalten, wurde versucht, das benötigte Zusatzwissen über eigene Internetrecherchen zu generieren.

Als Überschneidungsmerkmale wurden *WZ-Viersteller, Regionalkennung Ost-West*, *Anzahl der Filialen*, *Beschäftigte* und *Umsatz* verwendet.[89] Bei den Einzelangriffen ging man davon aus, dass der Datenangreifer bereits über die Kenntnis der Branchenzugehörigkeit (vierstelliger Wirtschaftszweig) und Regionalzugehörigkeit (Ost- oder Westdeutschland) des gesuchten Unternehmens verfügt. Diese Annahme ist durchaus realistisch, da der Datenangreifer gezielt nach einem bestimmten Unternehmen sucht. Beispielsweise könnte dies als persönliches Wissen vorliegen, wenn dieser im Auftrag eines Konkurrenzunternehmens agiert. Im Verlauf der Internetrecherche stellte sich heraus, dass es möglich war, die Anzahl der Beschäftigten und die Anzahl der Filialen ohne größeren Aufwand zu generieren. Dabei betrug die Dauer der Internetrecherche im Schnitt ca. 30 Minuten pro Unternehmen.

89) Vgl. Abschnitt 37.2.

Tabelle 37.16 zeigt die Ergebnisse der simulierten Einzelangriffe.

Tabelle 37.16: Einzelangriffe mit Zusatzwissen „Internet“

	alle Unternehmen	Unternehmen mit 1.000 - 4.999 Beschäftigten	Unternehmen mit mindestens 5.000 Beschäftigten
gesuchte Unternehmen	20 (100%)	16 (100%)	4 (100%)
eindeutige Zuordnungen	9 (45%)	6 (37,50%)	3 (75%)
richtige Zuordnungen	8 (40%)	6 (37,50%)	2 (50%)
falsche Zuordnungen	1 (5%)	0 (0%)	1 (25%)
keine Zuordnungen	8 (40%)	8 (50%)	0 (0%)
keine Zuordnungen wegen fehlenden Zusatzwissens (undurchsichtige Unternehmensstrukturen)	3 (15%)	2 (12,50%)	1 (25%)
nicht identifiziert (gesamt)	12 (60%)	10 (62,50%)	2 (50%)

Von den insgesamt 20 Unternehmen konnten acht dem Originaldatensatz richtig zugeordnet werden. Ein weiteres wurde falsch zugeordnet und elf Unternehmen konnten nicht zugeordnet werden. Die Wahrscheinlichkeit, dass ein Datenangreifer ein Unternehmen identifizieren kann, liegt somit in dieser Simulation bei 40%. Da der Datenangreifer die Richtigkeit seiner Zuordnung nicht überprüfen kann, müsste er bei diesem Ergebnis mit einer Wahrscheinlichkeit von ca. 11% damit rechnen, dass er fälschlicherweise zugeordnet hat.

Für diese Untersuchung wurden die eben beschriebenen Einzelangriffe mit demselben Zusatzwissen wiederholt. Einzige Änderung ist die Vergröberung des Merkmals *Anzahl der Filialen* in Kategorien.

Durch diese Anonymisierungsmaßnahme halbiert sich die Anzahl korrekt und eindeutig zugeordneter Unternehmen. Drei der vier zuvor noch richtigen und eindeutigen Zuordnungen wurden uneindeutig; außerdem kam eine falsche Zuordnung hinzu. Demnach erweist sich die erwähnte Maßnahme – analog zur Beobachtung beim Massenfischzug – bei den großen Unternehmen als wirkungsvolle Schutzmaßnahme.

b) Einzelangriffe mit Zusatzwissen MARKUS-Datenbank

Bei diesen Reidentifikationsversuchen wurden die Überschneidungsmerkmale *WZ-Viersteller, Regionalkennung Ost-West*, *Beschäftigte* und *Umsatz* verwendet. Da die MARKUS-Datenbank keine Angaben über die Anzahl der Filialen enthält, konnte dieses Merkmal nicht verwendet werden.

Von den im vorigen Abschnitt untersuchten 20 Unternehmen sind 11 davon in der MARKUS-Datenbank enthalten. Die Ergebnisse der Zuordnungsversuche werden in Tabelle 37.17 wiedergegeben.

Tabelle 37.17: Einzelangriffe mit Zusatzwissen „MARKUS-Datenbank“

	alle Unternehmen	Unternehmen mit 1.000 - 4.999 Beschäftigten	Unternehmen mit mindestens 5.000 Beschäftigten
gesuchte Unternehmen	11 (100%)	7 (100%)	4 (100%)
eindeutige Zuordnungen	11 (100%)	7 (100%)	4 (100%)
richtige Zuordnungen	6 (54,50%)	5 (71,40%)	1 (25%)
falsche Zuordnungen	5 (45,50%)	2 (28,60%)	3 (75%)
keine Zuordnungen	0 (0%)	0 (0%)	0 (0%)
nicht identifiziert (gesamt)	5 (45,50%)	2 (28,60%)	3 (75%)

Demnach konnten sechs Unternehmen dem Originaldatensatz richtig zugeordnet werden. Die Wahrscheinlichkeit, dass ein Datenangreifer ein Unternehmen identifizieren kann, liegt somit bei dieser Simulation bei 54,5%. Da der Datenangreifer die Richtigkeit seiner Zuordnung nicht überprüfen kann, müsste er bei diesem Ergebnis mit einer Wahrscheinlichkeit von 45,5% damit rechnen, dass er falsch zugeordnet hat. Der Unsicherheitsfaktor über die Richtigkeit der Reidentifikation liegt damit deutlich über dem aus der Simulation der Einzelangriffe mit Hilfe der Internetrecherche (11%).

c) Einzelangriffe mit kombiniertem Zusatzwissen

Schließlich wurden in einer weiteren Simulation sowohl die Informationen der Internetrecherche als auch der MARKUS-Datenbank benutzt. Dabei wurde für jedes Überschneidungsmerkmal die qualitativ bessere Information benutzt. Mit dieser Vorgehensweise erhält man die maximal mögliche Trefferquote für einen Datenangreifer. Diese Vorgehensweise führte dazu, dass man die Anzahl der reidentifizierten Unternehmen deutlich steigern konnte. Von den elf gesuchten Unternehmen konnte man acht reidentifizieren, wie in Tabelle 37.18 dargestellt.

Tabelle 37.18: Einzelangriffe mit Zusatzwissen „Internet“ und „MARKUS-Datenbank“

	Gesuchte Unternehmen	Richtige Zuordnungen	Treffer-quote
Internetrecherche	11	4	36%
MARKUS-Datenbank	11	6	55%
Internetrecherche & MARKUS-Datenbank	11	8	73%

Dieses idealtypisch kombinierte Zusatzwissen führt demnach zu einer beachtlichen Trefferquote von 73%. Ein Datenangreifer ist demzufolge in der Lage, ca. drei von vier Unter-

nehmen reidentifizieren zu können. Die hohe Trefferquote verdeutlicht, dass Unternehmen mit mindestens 1.000 Beschäftigten besonders schutzbedürftig sind.

37.5 Überprüfung des Analysepotenzials

37.5.1 Vorgehen

Im Vergleich zu den Untersuchungen des Analysepotenzials bei der KSE wird nun die Anzahl der dargestellten Verfahren auch bei der Gruppe der stochastischen Überlagerungen weiter reduziert. Neben der abstandsorientierten getrennten Mikroaggregation mit einer flexiblen Gruppengröße von drei bis fünf (MA32G) werden zwei Varianten der multiplikativen stochastischen Überlagerung mit einer Mischungsverteilung nach dem Verfahren von Höhne getestet:

- Hoe_f11_s02.
- Hoe_f08_s02.

Analog zum Vorgehen bei der KSE (vgl. Kapitel 35) wird im Folgenden dargestellt, inwiefern die in Kapitel 18 festgelegten Abweichungsmaße eingehalten wurden. Ökonometrische Auswertungen wurden mit der Einzelhandelsstatistik nicht vorgenommen. Jedoch können die aus den Simulationsexperimenten sowie den Fallbeispielen mit KSE und IAB-Betriebspanel herangezogen werden, um zu zeigen, dass sich mit der getrennten abstandsorientierten Mikroaggregation ebenso wie mit der multiplikativen Überlagerung nach dem Verfahren von Höhne grundsätzlich gute, beziehungsweise gut korrigierbare, Ergebnisse erzielen lassen.

37.5.2 Untersuchung der einzelnen Abweichungsmaße

In der Tabelle 37.19 sind die durchschnittlichen Abweichungen der wichtigsten Verteilungsmaße im Gesamtdatensatz dargestellt. In den Tabellen 37.20 und 37.21 wird gezeigt, wie häufig diese Verteilungsmaße über den in Kapitel 18 festgelegten Abweichungsschwellen liegen. Dies erfolgt analog für die Verteilungsmaße einzelner Teilgesamtheiten in den Tabellen 37.22 bis 37.24. In Tabelle 37.25 sind die Ergebnisse von t-Tests auf Mittelwertgleichheit nach diesen Teilgesamtheiten dargestellt.

Tabelle 37.19: Veränderung der univariaten Verteilungsmaße

	Durchschnittliche relative Abweichung der		Durchschnittliche absolute Abweichung der	
	arithmetischen Mittel (in %)	Standard-abweichungen (in %)	Korrelations-koeffizienten (× 100)	Rangkorre-lationen (× 100)
MA32G	0	6,7	3,4	0,1
Hoe_f11_s02	1,9	4,2	0,8	1,1
Hoe_f08_s02	1,5	3,1	0,6	0,1

Tabelle 37.20: Veränderungsraten der univariaten Verteilungsmaße

	Arithmetische Mittel (Anteil der Fälle mit einer Abweichung von über 10%) (in %)		Mediane (Anteil der Fälle mit einer Abweichung von über 10%) (in %)	Standardabweichungen (Anteil der Fälle mit einer Abweichung von über 10%) (in %)
	ungewichtet	gewichtet		
MA32G	0	3,1	0	*21,9*
Hoe_f11_s02	3,1	3,1	6,3	6,3
Hoe_f08_s02	3,1	0	3,1	3,1

Anteile bezogen auf die Anzahl der Merkmale

Tabelle 37.21: Veränderungsraten der Korrelationen

	Anteil der Korrelations-koeffizienten, die um mehr als 0,1 abweichen (in %)	Anteil der Korrelations-koeffizienten mit einem Vor-zeichenwechsel (in %)	Anteil der Rangkorre-lationen, die um mehr als 0,05 abweichen (in %)	Anteil der Rangkorre-lationen mit einem Vorzeichen-wechsel (in %)
MA32G	*11,1*	4,0	0	0
Hoe_f11_s02	0	0,0	0	0,2
Hoe_f08_s02	0	0	0	0,2

Anteile bezogen auf die Anzahl der berechneten Korrelationen

Tabelle 37.22: Veränderung der Verteilungsmaße einzelner Wirtschaftszweige

	Arithmetische Mittel (Anteil der Fälle mit einer Abweichung von über 10%) (in %)		Mediane (Anteil der Fälle mit einer Abweichung von über 10%) (in %)	Standardabweichungen (Anteil der Fälle mit einer Abweichung von über 10%) (in %)
	ungewichtet	gewichtet		
MA32G	6,9	5,1	0	*12,1*
Hoe_f11_s02	4,2	3,1	6,0	*15,2*
Hoe_f08_s02	1,3	2,0	3,1	2,0

Anteile bezogen auf die Anzahl der Merkmale multipliziert mit den betrachteten Teilgesamtheiten

Tabelle 37.23: Veränderung der Verteilungsmaße nach Ost/West

	Arithmetische Mittel (Anteil der Fälle mit einer Abweichung von über 10%) (in %)		Mediane (Anteil der Fälle mit einer Abweichung von über 10%) (in %)	Standardabweichungen (Anteil der Fälle mit einer Abweichung von über 10%) (in %)
	ungewichtet	gewichtet		
MA32G	0	0	0	*15,6*
Hoe_f11_s02	1,6	9,4	4,7	*4,7*
Hoe_f08_s02	4,7	6,3	3,1	4,7

Anteile bezogen auf die Anzahl der Merkmale multipliziert mit den betrachteten Teilgesamtheiten

Tabelle 37.24: Veränderung der Verteilungsmaße nach Wirtschaftszweigen und Ost/West

	Arithmetische Mittel (Anteil der Fälle mit einer Abweichung von über 10%) (in %)		Mediane (Anteil der Fälle mit einer Abweichung von über 10%) (in %)	Standardabweichungen (Anteil der Fälle mit einer Abweichung von über 10%) (in %)
	ungewichtet	gewichtet		
MA32G	3,8	3,2	0,1	6,9
Hoe_f11_s02	6,3	6,4	5,4	*19,6*
Hoe_f08_s02	2,9	3,2	3,3	4,0

Anteile bezogen auf die Anzahl der Merkmale multipliziert mit den betrachteten Teilgesamtheiten

Tabelle 37.25: Ergebnisse von t-Tests auf Mittelwertgleichheit für Teilgesamtheiten

	Anteil der signifikanten Abweichungen der gewichteten arithmetischen Mittel zum Signifikanzniveau von 10%		
	nach Wirtschaftszweigen (in %)	nach Ost/West (in %)	nach Wirtschaftszweigen und Ost/West (in %)
MA32G	0,2	1,7	0,3
Hoe_f11_s02	1,1	3,1	0,6
Hoe_f08_s02	1,1	4,7	1,0

Anteile bezogen auf die Anzahl der Merkmale multipliziert mit den betrachteten Teilgesamtheiten

37.5.3 Schlussfolgerungen für die Anonymisierung der Einzelhandelsstatistik aus Sicht des Analysepotenzials

Es zeigt sich, dass die multiplikativen stochastischen Überlagerungen bei der Einzelhandelsstatistik nicht schlechter abschneiden als die getrennte abstandsorientierte Mikroaggregation. Die Variante mit einem Abstand der Mittelwerte gegenüber 1 von 0,08 und einer Standardabweichung der beiden Normalverteilungen von 0,02 erzeugt sogar im Vergleich zur Mikroaggregation die besseren Ergebnisse. Hier wird in keinem Fall die Toleranzgrenze von maximal 10% Überschreitungen der Abweichungsschwelle verletzt. Es bietet sich daher an, auf Basis dieses Verfahrens ein Scientific-Use-File zu produzieren.

37.6 Zwei Scientific-Use-Files der Einzelhandelsstatistik 1999

In Absprache mit den Fachstatistikern wurde beschlossen, dass anonymisierte Daten der Einzelhandelsstatistik als faktisch anonym deklariert werden, falls die entsprechenden Enthüllungsrisiken von simulierten Massenfischzugsexperimenten in den Zellen der Merkmalskombination *Wirtschaftszweig / Regionalinformation / Beschäftigtengrößenklasse* jeweils unter dem Grenzwert von 0,5 liegen. Das bedeutet, dass die Wahrscheinlichkeit für einen Datenangreifer, in einer bestimmten Zelle (beispielsweise alle ostdeutschen Unternehmen mit mindestens 1.000 Beschäftigten im Wirtschaftszweig '523') Mikrodaten zu enthüllen, die weniger als 10% von ihrem Originalwert abweichen, weniger als 50% beträgt. Demnach ergibt sich ein Gesamtrisiko über alle Zellen, das deutlich unter 50% liegt.

Weiterhin müssen die Daten simulierten Einzelangriffen in ausgemachten Gefährdungsbereichen (beispielsweise kleine Besetzungszahlen) standhalten.

37.6.1 Erzeugung eines Scientific-Use-Files für die kleinen Unternehmen der Einzelhandelsstatistik 1999

Es wurde versucht, möglichst ohne datenverändernde Verfahren faktisch anonyme Daten für die Einzelhandelsstatistik zu erzeugen. Insbesondere der Wissenschaftliche Begleitkreis des Projekts sprach sich dafür aus, die Behandlung kategorialer Merkmale einer Veränderung metrischer Merkmale vorzuziehen. Bei den Mikroaggregationsvarianten wurden MA9G und MA1G wegen mangelndem Analysepotenzials ausgeschlossen. Demnach blieb als datenveränderndes Mikroaggregationsverfahren die auf den Schutz schwach wirkende eindimensionale getrennte Mikroaggregation übrig. Es konnten jedoch keine zusätzlichen traditionellen Anonymisierungsmaßnahmen gefunden werden, die gemeinsam mit diesem Verfahren einen ausreichenden Datenschutz für alle Unternehmen der Einzelhandelsstatistik gewährleisten. Das Problem liegt vor allem an der bereits in Abschnitt 37.1 erwähnten äußerst geringen Dichte an mittleren und großen Unternehmen. Dagegen existiert eine

große Anzahl an Unternehmen mit weniger als 50 Beschäftigten, so genannte kleine Unternehmen, was die Erzeugung eines Scientific-Use-Files für diese Unternehmen begünstigt. Im Folgenden werden die Anonymisierungsmaßnahmen beschrieben, mit denen es gelang, für die kleinen Unternehmen mit einem Mindestjahresumsatz von 250.000 Euro nahezu ohne datenverändernde Verfahren faktisch anonyme Daten zu erzeugen.

a) Informationsreduzierende Maßnahmen

Die Daten wurden formal anonymisiert, d.h. direkte Identifikatoren wie Name und Anschrift wurden entfernt. Die Merkmale aus dem Bereich „Umsatz nach Tätigkeiten bzw. Produkten (in %)“ wurden entfernt, da die Information über die Nebentätigkeiten eines Unternehmens die Reidentifikationsgefahr stark ansteigen lässt. Die Information der Haupttätigkeit eines Unternehmens bleibt durch die Kenntnis des Wirtschaftszweigs erhalten. Das Merkmal *Anzahl der rechtlich unselbständigen örtlichen Einheiten des Unternehmens* (Filialen) kann von einem Datenangreifer als Überschneidungsmerkmal verwendet werden. Aus Geheimhaltungsgründen wurden daher für die besonders reidentifikationsgefährdeten größeren Unternehmen bestimmte Kategorien für die Anzahl ihrer Filialen gebildet.[90]. Die Regionalangabe wurde auf die Ost-West-Klassifizierung vergröbert. Die Wirtschaftszweigklassifikation wurde auf den Dreisteller vergröbert. Eine Ausnahme bildet der Dreisteller '524' [Sonstiger Facheinzelhandel (in Verkaufsräumen)] der aufgrund seiner großen Besetzungszahl auf der Vierstellerebene angegeben werden kann (siehe Tabelle 37.26). Dadurch bleiben dem Datennutzer inhaltlich wertvolle Informationen erhalten:

b) Zusätzliche Anonymisierungsmaßnahmen

Zusätzlich zu den oben beschriebenen Maßnahmen wurden 7 Unternehmen aus den beiden relativ dünn besetzten Wirtschaftszweigen '525' “Einzelhandel mit Antiquitäten und Gebrauchtwaren (in Verkaufsräumen)“ und '5241' „Einzelhandel mit Textilien“ durch weitere Anonymisierung geschützt. Ein ostdeutsches Unternehmen aus dem Wirtschaftszweig '525' wurde aus dem Datensatz herausgenommen. Bei drei westdeutschen Unternehmen aus dem Wirtschaftszweig '525' wurden die Merkmale *Gesamtumsatz* und *Beschäftigte* gemittelt (punktuelle Mikroaggregation). Bei drei ostdeutschen Unternehmen aus dem Wirtschaftszweig '5241' wurden die Ausprägungen im Merkmal *Beschäftigte* gemittelt.

c) Überprüfung der Schutzwirkung

Es wurden Matchingexperimente mit folgenden Überschneidungsmerkmalen durchgeführt:

90) Siehe Tabelle 37.4

Tabelle 37.26: Die Tiefe der WZ-Klassifikation bei den Scientific-Use-Files

Wirtschaftszweig	WZ 93
Einzelhandel mit Waren verschiedener Art (in Verkaufsräumen)	521
Facheinzelhandel mit Nahrungsmitteln, Getränken und Tabakwaren (in Verkaufsräumen)	522
Apotheken; Facheinzelhandel mit medizinischen, orthopädischen und kosmetischen Artikeln (in Verkaufsräumen)	523
Einzelhandel mit Textilien	5241
Einzelhandel mit Bekleidung	5242
Einzelhandel mit Schuhen und Lederwaren	5243
Einzelhandel mit Möbeln, Einrichtungsgegenständen und Hausrat	5244
Einzelhandel mit elektrischen Haushaltsgeräten, Geräten der Unterhaltungselektronik und Musikinstrumenten	5245
Einzelhandel mit Metallwaren, Anstrichmitteln, Bau- und Heimwerkerbedarf	5246
Einzelhandel mit Büchern, Zeitschriften, Zeitungen, Schreibwaren und Bürobedarf	5247
Facheinzelhandel, anderweitig nicht genannt (in Verkaufsräumen)	5248
Einzelhandel mit Antiquitäten und Gebrauchtwaren (in Verkaufsräumen)	525
Einzelhandel (nicht in Verkaufsräumen)	526
Reparatur von Gebrauchsgütern	527

- Wirtschaftszweigklassifikation (WZ93, Drei- bzw. Vierstellerebene)
- Ost-West-Klassifizierung
- Gesamtumsatz
- Anzahl der Beschäftigten

Tabelle 37.27 zeigt die aus den Simulationen berechneten Enthüllungsrisiken nach Beschäftigtengrößenklassen. Dabei flossen die Risiken unter Verwendung der MARKUS-Datenbank zu 95% und des „Worst-Case"-Szenarios zu 5% ein.

Tabelle 37.27: Enthüllungsrisiken für ost- und westdeutsche Unternehmen

	Größenklassen		
Ost-West	1-19	20-49	Insgesamt
Ost	0,15	0,22	0,16
West	0,12	0,14	0,13
Insgesamt	0,13	0,16	0,13

Ebenso genügen die Enthüllungsrisiken dem zu Beginn dieses Abschnitts erwähnten Kriterium zur Beurteilung der faktischen Anonymität.
In Tabelle 37.34 findet sich eine Auflistung der in der anonymisierten Datei enthaltenen Merkmale.

37.6.2 Erzeugung eines Scientific-Use-Files für alle Unternehmen der Einzelhandelsstatistik 1999

Mit den später im Projekt getesteten Verfahren der Zufallsüberlagerung ist es gelungen, einen weiteren Scientific-Use-File ohne Abschneidegrenze hinsichtlich der Anzahl an Beschäftigten in einem Unternehmen zu erzeugen. Wie bei dem Scientific-Use-File für die kleinen Unternehmen beträgt der Mindestjahresumsatz 250.000 Euro.

a) Informationsreduzierende Maßnahmen

Grundsätzlich wurden hier dieselben Maßnahmen wie in Unterabschnitt 37.6.1 vorgenommen, außer dass die Regionalinformation entfernt wurde.

b) Datenverändernde Verfahren

Es wurde die Variante Hoe_f08_s02 aus der Verfahrensgruppe der Zufallsüberlagerungen angewendet und folgende Schritte dabei durchgeführt:
Multiplikative Überlagerung der Einzelwerte nach dem Verfahren von Höhne mit den Parametern $f = 0,08$und$s = 0,02$. Damit werden die Angaben des Betriebes entweder einheitlich erhöht oder gesenkt (Wahrscheinlichkeit je 50% und im Mittel um 8%), wobei der genaue Faktor der Veränderung im Durchschnitt um 2 Prozentpunkte von den 8% abweicht.

c) Zusätzliche Anonymisierungsmaßnahmen

Da die bisher beschriebenen Maßnahmen nicht in allen Zellen einen ausreichenden Schutz garantieren, wurden folgende punktuellen Eingriffe durchgeführt:

Mikroaggregation der drei Marktführer in den Wirtschaftszweigen '522', '523', '525' und '5247' sowie Mikroaggregation der sechs Markführer im Wirtschaftszweig '527'. Dabei wurden neben den Merkmalen *Umsatz* und *Beschäftigte* auch die mit diesen beiden stark korrelierten Merkmale (Korrelation $> 0,75$) ebenfalls mikroaggregiert.[91]

Insgesamt sind demnach 18 Merkmalsträger durch die angegebenen zusätzlichen Maßnahmen betroffen.

d) Überprüfung der Schutzwirkung

Da die Regionalinformation entfernt wurde, steht sie einem Datenangreifer im Gegensatz zu den Massenfischzugsimulationen von Unterabschnitt 37.6.1 nicht mehr zur Verfügung.

Demnach wurden die Matchingexperimente mit folgenden Überschneidungsmerkmalen durchgeführt:

- Wirtschaftszweigklassifikation (WZ93, Drei- bzw. Vierstellerebene)
- Gesamtumsatz
- Anzahl der Beschäftigten

Tabelle 37.28 zeigt die aus den Simulationen berechneten Enthüllungsrisiken nach Beschäftigtengrößenklassen. Dabei flossen die Risiken unter Verwendung der MARKUS-Datenbank

91) Dabei handelt es sich um die Merkmale *Anfangs- und Endbestand an Handelsware, Sonstige betriebliche Erträge, Bezüge von Handelsware, Löhne und Gehälter, Sozialabgaben, Mieten und Pachten einschl. Kosten für Operate Leasing und Bezogene Leistungen und andere betriebliche Aufwendungen*

zu 95% und das „Worst-Case"-Szenario zu 5% ein. Bei den Zellen, die nicht durch die Markus-Datenbank repräsentiert werden, wurden die Risiken des „Worst-Case"-Falls zu 100% übernommen.

Tabelle 37.28: Enthüllungsrisiken nach Beschäftigtengrößenklassen

GRKL	1-19	20-49	50-99	100-249	250-499	500-999	$\geq$ 1.000	Total
Risiko	0,01	0,01	0,03	0,05	0,07	0,13	0,16	0,02

Schließlich ergibt eine Überprüfung der Risiken nach Beschäftigtengrößenklassen und Wirtschaftszweigklassifikation, dass eine Enthüllung der Daten nur mit einem unverhältnismäßig hohen Aufwand möglich ist, und daher eine Weitergabe der faktisch anonymisierten Daten an die Wissenschaft unbedenklich erscheint.

e) Das Analysepotenzial

Getestet wird der mit dem Verfahren der multiplikativen stochstischen Überlagerung nach dem Verfahren von Höhne erzeugte Scientific-Use-File für alle Unternehmen der Einzelhandelsstatistik. Die Tabellen 37.29 bis 37.33 zeigen, dass die vorgegebenen Toleranzgrenze in keinem Fall überschritten wird. Da das gewählte Verfahren auch grundsätzlich als brauchbar eingestuft wird, ist der vorgeschlagene Scientific-Use-File aus Sicht des Analysepotenzials positiv zu beurteilen.

Tabelle 37.29: Durchschnittliche Veränderung der Verteilungsmaße

Durchschnittliche relative Abweichung der		Durchschnittliche absolute Abweichung der	
arithmetischen Mittel (in %)	Standard-abweichungen (in %)	Korrelations-koeffizienten (mal 100)	Rangkorre-lationen (mal 100)
1,5	3,2	5,9	0,1

Tabelle 37.30: Veränderung der univariaten Verteilungsmaße

Anteil der arithmetischen Mittel, die um mehr als 10% abweichen		Anzahl der Mediane, die um mehr als 10% abweichen	Anzahl der Standardabweichungen, die um mehr als 10% abweichen
(in %)		(in %)	(in %)
ungewichtet	gewichtet		
3,1	3,1	3,1	3,1

Anteile bezogen auf die Anzahl der Merkmale

Tabelle 37.31: Veränderung der Korrelationen

Anteil der Korrelationskoeffizienten, die um mehr als 0,1 abweichen (in %)	Anteil der Korrelationskoeffizienten mit einem Vorzeichenwechsel (in %)	Anteil der Rangkorrelationen, die um mehr als 0,05 abweichen (in %)	Anteil der Rangkorrelationen mit einem Vorzeichenwechsel (in %)
0	1,8	0	0,2

Anteile bezogen auf die Anzahl der berechneten Korrelationen

Tabelle 37.32: Veränderung der Verteilungsmaße für die Wirtschaftszweige

Arithmetische Mittel (Anteil der Fälle mit einer Abweichung von über 10%)		Mediane (Anteil der Fälle mit einer Abweichung von über 10%)	Standardabweichungen (Anteil der Fälle mit einer Abweichung von über 10%)
(in %)		(in %)	(in %)
ungewichtet	gewichtet		
1,3	2,0	3,1	6,9

Anteile bezogen auf die Anzahl der Merkmale multipliziert mit den betrachteten Wirtschaftszweigen

Tabelle 37.33: t-Tests auf Mittelwertgleichheit für die Wirtschaftszweige

Anteil der signifikanten Abweichungen zu einem Signifikanzniveau von 10% (in %)	
ungewichtet	gewichtet
2,4	4,2

Anteile bezogen auf die Anzahl der Merkmale multipliziert mit den betrachteten Wirtschaftszweigen

Tabelle 37.34: Merkmale der Scientific-Use-Files

1.	Wirtschaftszweig (WZ 93)
2.	Regionalbezug (Ost-West-Klassifizierung) 1)
3.	Gesamtumsatz
	Umsatzanteile in % aus
4.	Großhandel
5.	Einzelhandel, Reparatur von Gebrauchsgütern
6.	Sonstigen Dienstleistungstätigkeiten
7.	Herstellung, Verarbeitung, anderen industriellen Tätigkeiten oder aus Land- und
	Forstwirtschaft und Fischerei
8.	Sonstige betriebliche Erträge
	Einzelhandelsumsatz in % nach Absatzformen
9.	In Verkaufsräumen
10.	Aus Versandhandel
11.	An Verkaufsständen und auf Märkten
12.	Aus sonstigem Einzelhandel
13.	Anfangsbestand an Handelsware
14.	Endbestand an Handelsware
15.	Anfangsbestand an Roh-, Hilfs- und Betriebsstoffen
16.	Endbestand an Roh-, Hilfs- und Betriebsstoffen
17.	Anfangsbestand an selbsthergestellten und bearbeiteten Halb- und Fertigerzeugnissen
18.	Endbestand an selbsthergestellten und bearbeiteten Halb- und Fertigerzeugnissen
19.	Bezüge von Handelsware
20.	Bezüge von Roh-, Hilfs- und Betriebsstoffen
21.	Löhne und Gehälter
22.	Sozialabgaben
23.	Mieten und Pachten einschl. Kosten für Operate Leasing
24.	Betriebliche Steuern und Abgaben
25.	Bezogene Leistungen und andere betriebliche Aufwendungen
	Bruttoinvestitionen in
26.	Grundstücke
27.	Bestehende Gebäude
28.	Errichtung, Umbau und Erweiterung von Gebäuden
29.	Maschinen, Einrichtungen und Fahrzeuge
30.	Verkäufe von Sachanlagen
31.	Wert der im Geschäftsjahr über Finanzierungsleasing erworbenen Sachanlagen
32.	Zahl der rechtlich unselbständigen örtlichen Einheiten des Unternehmens am 31.12.
	Zahl der Beschäftigten am 30.9.
33.	Beschäftigte insgesamt
34.	Darunter Lohn- und Gehaltsempfänger
35.	Darunter Teilzeitbeschäftigte
36.	Hochrechnungsfaktor

1) Nur enthalten im Scientific-Use-File für die kleinen Unternehmen

Teil XII

Zusammenfassung: Handlungsempfehlungen für kommende Anonymisierungsprojekte

Ausgehend von den Ergebnissen der Projektarbeiten und den im Projekt gemachten Erfahrungen werden Handlungsempfehlungen für das Vorgehen bei zukünftigen Anonymisierungsvorhaben gegeben. Hierzu werden die einzelnen Arbeitsschritte formuliert und in dem folgenden Ablaufschema dargestellt.

Wesentlich ist eine enge Verzahnung der Analyseaspekte mit den Schutzaspekten. Es wird empfohlen, zunächst eine für die Wissenschaft (noch) vertretbare Einschränkung von Informationen vorzunehmen und erst danach, sofern notwendig, datenverändernde Anonymisierungsverfahren einzusetzen.

An datenverändernden Verfahren sind die abstandsorientierte getrennte Mikroaggregation und die multiplikative stochastische Überlagerung, insbesondere nach dem Verfahren von Höhne, zu empfehlen. Kommt letztere zum Einsatz, so ist auf die Startwertproblematik bei der Erzeugung von Zufallszahlen zu achten und eine nicht zu große Varianz zu wählen. Zudem muss den Nutzern dann auch ein zusätzliches File mit Instrumentvariablen zur Verfügung gestellt werden, bei dem die gleiche Anonymisierung (gleiches Verfahren, gleiche Parameter) angewendet wird, wie bei dem eigentlichen Scientific-Use-File.

Um die Akzeptanz datenverändernder Verfahren bei den Nutzern zu erhöhen, wird die projektbegleitende Einrichtung einer „Methodengruppe Anonymisierung“ empfohlen, der Experten der Forschungsdatenzentren und methodisch arbeitende empirische Wirtschaftsforscher angehören sollten. Diese soll den Prozess der Erstellung von Scientific-Use-Files mit datenverändernden Anonymisierungsverfahren begleiten, Nutzer und Datenproduzenten beraten, Analyseergebnisse mit den Daten sichten und diese für Empfehlungen zur Weiterentwicklung von Anonymisierungsverfahren nutzen.

Analyseseite (A)	Schutzseite (S)
A1. Erkundung des Nutzerkreises der Daten; Gewinnung von Nutzern zur Mitarbeit	S1. Recherche über das Zusatzwissen eines potenziellen Datenangreifers
A2. Feststellen der bevorzugten Forschungsinteressen, Festlegung des Merkmalskanons	S2. Aufbau einer Datenbank als Zusatzwissen für Datenangriff-Simulationen
A3. Festlegung der aus Analysesicht sinnvollen Tiefengliederung bei den kategorialen Merkmalen	S3. Vorläufige Festlegung der Teilgesamtheiten, für die das Enthüllungsrisiko bestimmt werden soll
A4. Festlegung von Richtgrößen für deskriptive Maße	S4. Festlegung von Schwellen γ_i für die Brauchbarkeit von Einzelangaben für potenziellen Datenangreifer
	S5. Festlegung einer oberen Risikoschwelle τ
AS1. Aufbereitung des aus Analysesicht gewünschten Datensatzes (formal anonymisiert)	
	S6. Überprüfung des Schutzbedarfs dieses Datensatzes durch Simulation von Datenangriffen (Massenfischzüge unter Verwendung der in S2. aufgebauten Datenbank; gegebenenfalls Einzelangriffe für Einheiten in besonders gefährdeten Bereichen)
AS2. Falls notwendig: Erzeugung probeweiser anonymisierter Daten durch: informationsreduzierende Maßnahmen, insbesondere Vergröberung kategorialer Merkmale so weit aus Schutzsicht erforderlich bzw. aus Nutzersicht vertretbar (diskursiver Prozess)	
	S7. Überprüfung der Schutzwirkung dieses Datensatzes durch Simulation von Datenangriffen (Massenfischzüge unter Verwendung der in S2. aufgebauten Datenbank; gegebenenfalls Einzelangriffe für Einheiten in besonders gefährdeten Bereichen)
AS3. Falls notwendig: Erzeugung probeweiser anonymisierter Daten durch datenverändernde Anonymisierungsverfahren (Empfehlung: getrennte abstandsorientierte, Mikroaggregation, multiplikative stochastische Überlagerung nach dem Verfahren von Höhne)	
A5. Test des Analysepotenzials der in AS3. erzeugten Daten: Durchführung der wichtigsten deskriptiven und inferenzstatistischen Auswertungen, Überprüfung der Abweichungsmaße	S8. Überprüfung der Schutzwirkung dieses Datensatzes durch Simulation von Datenangriffen (Massenfischzüge unter Verwendung der in S2. aufgebauten Datenbank; gegebenenfalls Einzelangriffe für Einheiten in besonders gefährdeten Bereichen)
AS4. Ausbalancieren der Parameter der Anonymisierung sowie der Gewichtung von informationsreduzierenden und datenverändernden Maßnahmen mit den Zielen: 1. Ausreichender Erhalt des Analysepotenzials 2. Unterschreiten der oberen Risikoschwelle	
AS5. Falls möglich, Erstellung von Scientific-Use-Files (mit Erstellung von Metadaten) Beachte: Bei multiplikativer Überlagerung Erstellung von Datensatz mit Instrumenten Falls nicht möglich, Prozess ab Stufe AS2. neu beginnen	

Anhang

Wesentlich für die Projektarbeiten war der Informationsaustausch und der ständige Diskurs des Projektteams mit den Datenproduzenten einerseits und den potenziellen Datennutzern – der Wissenschaft – andererseits. Deshalb wurden eigene Workshops zu wesentlichen Aspekten des Forschungsprojekts organisiert. Zudem haben die Projektmitarbeiter der Statistischen Ämter des Bundes und der Länder und des IAW das Projekt mit Vorträgen auf zahlreichen wissenschaftlichen Tagungen im In- und Ausland vorgestellt und die Projektarbeiten durch wissenschaftliche Publikationen dokumentiert. Dies wird im Folgenden dargestellt.

Anhang A

Eigene Workshops

Nutzer-Workshop: Anonymisierung wirtschaftsstatistischer Einzeldaten am 20. und 21. März 2003

Tagungsort: Eberhard Karls Universität, Neue Aula, Hörsaal 1
Wilhelmstraße 7, 72074 Tübingen

Dieser erste öffentliche Workshop im Rahmen des Forschungsvorhabens, der vom Statistischen Bundesamt und dem Institut für Angewandte Wirtschaftsforschung (IAW) gemeinsam in Tübingen veranstaltet wurde, diente dazu, interessierten Fachleuten aus dem Kreise der amtlichen Statistik sowie der empirischen Sozial- und Wirtschaftsforschung die Ziele und Inhalte des Projekts näher vorzustellen, sowie Anregungen aus dem Teilnehmerkreis für die weitere Arbeit aufzunehmen.

Die Veranstaltung, deren Beiträge in der Reihe „Forum der Bundesstatistik" als Band 42 veröffentlicht wurden, zeigte, dass trotz deutlicher Fortschritte bei der Entwicklung von Methoden und Verfahren zur Minderung des Reidentifikationsrisikos aus Sicht der Datennutzer beim Thema „Faktische Anonymisierung" noch Pionierarbeit geleistet werden musste. Die Resonanz der rund 70 Teilnehmer zeugte von großem Interesse an der weiteren Arbeit des Projekts und es gab zudem wichtige Denkanstöße für die Erstellung von Scientific-Use-Files aus dem Unternehmensbereich.

Programm am Donnerstag, 20. März 2003

13:00

Begrüßung und Einführung in den Workshop
Professor Dr. Gerd Ronning (Institut für Angewandte Wirtschaftsforschung, Universität Tübingen) Professor Dr. Eberhard Schaich (Rektor der Universität Tübingen) Reinhold Friedrich (Bundesministerium für Bildung und Forschung)

13:20

Einführung – Das Projekt der statistischen Ämter: Ziele, Ablauf, Beteiligte – Problematik des Datenschutzes
Dr. Roland Gnoss (Statistisches Bundesamt, Gruppe I A, Projektleiter)

13:30

Die Reidentifikationsproblematik bei wirtschaftsstatistischen Einzeldaten

a) Angriffsszenarien auf wirtschaftsstatistische Einzeldaten – ein Überblick
Dr. Heike Wirth (ZUMA)

b) Matching-Verfahren der Statistik und die Reidentifikation faktisch anonymisierter Einzeldaten
Dr. Rolf Wiegert (Institut für Angewandte Wirtschaftsforschung)

c) Reidentifikationsrisiken und Reidentifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios
Dr. Daniel Vorgrimler (Statistisches Bundesamt)

Ko-Referat: Dr. Uwe Blien (IAB)

anschließend Diskussion

15:45

Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten
Jörg Höhne (Statistisches Landesamt Berlin)

Ko-Referat: Dr. Ruth Brand (Statistisches Bundesamt)

anschließend Diskussion

17:30

Konzepte zur Bewertung von anonymisierten Datensätzen
Professor Josep Domingo-Ferrer (Universität Rovira y Virgili, CASC-Projekt)

Ko-Referat: Sarah Gießing (Statistisches Bundesamt)

anschließend Diskussion

18:30

Ende des Programms

Programm am Freitag, 21. März 2003

9:00

Anonymisierungsmethoden und ökonometrische Modelle
Professor Dr. Winfried Pohlmeier (Universität Konstanz)

Ko-Referat: Professor Dr. Gerd Ronning (Institut für Angewandte Wirtschaftsforschung, Eberhard Karls Universität Tübingen)

anschließend Diskussion

10:30

Arbeiten mit Einzeldaten der amtlichen Statistik am Beispiel des Monatsberichts im Verarbeitenden Gewerbe
Professor Dr. Joachim Wagner (Universität Lüneburg)

Ko-Referat: Dr. Harald Strotmann (Institut für Angewandte Wirtschaftsforschung)

anschließend Diskussion

11:30

Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik
Martin Rosemann (Institut für Angewandte Wirtschaftsforschung)

Ko-Referat: Dr. Georg Licht (ZEW)

anschließend Diskussion

12:30

Ende des Workshops

Wissenschaftlicher Workshop: „Ökonometrie der Anonymisierung von Firmendaten“/ „Econometric Analysis of Anonymised Firm Data“ am 18./19. März 2004 in Tübingen

Auf diesem vom IAW veranstalteten Workshop wurden Zwischenergebnisse bezüglich der Auswirkungen der Anonymisierung auf die Schätzung mikroökonometrischer Modelle vorgestellt. Diese Veranstaltung, an der rund 30 Wissenschaftler (hauptsächlich Anwender ökonometrischer Verfahren) teilnahmen, hatte zum Ziel, sich über die Wirkungen verschiedener datenverändernder Anonymisierungsverfahren auf das Analysepotenzial der Daten auszutauschen, sowie Anregungen für die weitere Projektarbeit zu gewinnen. Die meisten Beiträge dieses Workshops erscheinen demnächst in den Jahrbüchern für Nationalökonomie und Statistik (Heft 5 Band 225 (2005)) unter dem Titel „Econometrics of Anonymized Micro Data“.

Die wissenschaftliche Organisation dieser Veranstaltung lag bei Prof. Dr. Gerd Ronning (Institut für Angewandte Wirtschaftsforschung, Universität Tübingen), Prof. Dr. Winfried Pohlmeier (Universität Konstanz) und Prof. Dr. Joachim Wagner (Universität Lüneburg).

Programm am Donnerstag, 18. März 2004

14:00

Begrüßung: Prof. Dr. Gerd Ronning (Institut für Angewandte Wirtschaftsforschung, Universität Tübingen)

14:15

Anonymized Firm Data Under Test: Evidence from a Replication Study
Prof. Dr. Joachim Wagner (Universität Lüneburg)

15:30

Impacts of different versions of micro aggregation on the results of linear and non-linear estimations
Dipl.-Volksw. Martin Rosemann (Institut für Angewandte Wirtschaftsforschung)

17:00

Microdata Disclosure by Resampling – Empirical Findings for Business Survey Data
Dipl.-Volksw. Sandra Gottschalk (ZEW, Mannheim)

Programm am Freitag, 19. März 2004

9:00

A Comparison of the Effects of Disclosure Limitation Methods on Nonlinear Regression Estimators
Dipl.-Ökonomin Sandra Lechner und Prof. Dr. Winfried Pohlmeier (Universität Konstanz)

10:15

Estimation of the Probit Model from Anonymized Data
Prof. Dr. Gerd Ronning und Dipl.-Volksw. Martin Rosemann (Institut für Angewandte Wirtschaftsforschung, Eberhard Karls Universität Tübingen)

11:30

The Impact of Anonymization on Binary Choice Models – Empirical Evidence from the IAB-Establishment-Panel Baden-Württemberg
Dr. Harald Strotmann (Institut für Angewandte Wirtschaftsforschung)

12:30

Organisatorisches

13:00

Ende der Veranstaltung

Workshop „Faktische Anonymität von Unternehmens- und Betriebsdaten“ am 23. April 2004 im Statstischen Bundesamt in Wiesbaden

Während die intensive Diskussion mit den künftigen Nutzern von Einzeldaten bereits in den beiden vorigen Veranstaltungen in Tübingen geführt wurde, widmete sich der Workshop für die statistischen Ämter in Wiesbaden der Schutzwirkung von Anonymisierungsmaßnahmen.

Auf dem Workshop wurden zunächst das Schutzwirkungskonzept und Anonymisierungsmethoden vorgestellt. Im Anschluss wurden die Projektarbeiten zur Abschätzung des realistischen Schutzbedürfnisses vorgestellt: Für zwei amtliche Erhebungen (Kostenstrukturerhebung im Verarbeitenden Gewerbe und Umsatzsteuerstatistik) wurden Enthüllungsrisiken berechnet, die die Wirkung verschiedener Anonymisierungsmaßnahmen erkennbar machen.

In der Diskussion mit Fachstatistikern und Mitarbeitern der Forschungsdatenzentren in den Ämtern wurde das im Projekt erarbeitete Schutzwirkungskonzept unterstützt. Es bestand

Konsens, mit Hilfe von festzusetzenden Schwellenwerten das Vorhandensein von faktischer Anonymität bei wirtschaftsstatistischen Mikrodaten festzulegen. Die Entscheidung über die faktische Anonymität einer Datei wird – wie bei Haushalts- und Personendaten bereits seit langem praktiziert – durch Abstimmungsprozesse in den statistischen Ämtern und unter maßgeblicher Mitarbeit der für die jeweilige Erhebung zuständigen Fachstatistiker getroffen.

Die Veranstaltung konnte damit zur einheitlichen Haltung der Ämter bei der Handhabung der faktischen Anonymisierung von Unternehmens- und Betriebsdaten beitragen. Die Projektmitglieder sahen sich in ihrer Arbeit bestätigt.

Programm

9:00

Begrüßung und Einführung in den Workshop
Dr. Roland Gnoss (Statistisches Bundesamt)

9:10

Ein Konzept zur Schutzwirkung
Roland Sturm (Statistisches Bundesamt)

9:40

Methoden der Anonymisierung
Jörg Höhne (Statistisches Landesamt Berlin)

10:40

Anwendung des Schutzwirkungskonzeptes am Beispiel der Kostenstrukturerhebung im Verarbeitenden Gewerbe
Dr. Rainer Lenz (Statistisches Bundesamt)

11:25

Anwendung des Schutzwirkungskonzeptes am Beispiel der Umsatzsteuerstatistik
Dr. Daniel Vorgrimler (Statistisches Bundesamt)

12:00

Diskussion

Workshop „Mikrodaten über Unternehmen und Betriebe – Neue Datenangebote für die Wissenschaft“ am 22./23. September 2005 in Wiesbaden

Auf dem abschließenden Workshop stellte das Projektteam seine Ergebnisse der Öffentlichkeit vor. Dabei wurden die Auswirkungen von Anonymisierung auf Vertraulichkeit und Analysepotenzial beleuchtet, Empfehlungen zur Anonymisierung von Unternehmensdaten gegeben sowie erste anonymisierte Unternehmensdaten für die Wissenschaft präsentiert.

Programm

Donnerstag, 22. September 2005

14:00

Begrüßung
Johann Hahlen (Präsident des Statistischen Bundesamtes)

14:10 Einführung in das Thema
Jürgen Chlumsky (Statistisches Bundesamt)

14:20

Vorgehensweise bei der Anonymisierung von Unternehmensdaten
Roland Sturm (Statistisches Bundesamt)

15:30

Anonymisierungsmethoden und deren Auswirkungen
Jörg Höhne (Statistisches Landesamt Berlin)

16:15

Bewertung der Anonymisierung aus Nutzersicht
Martin Rosemann (Institut für Angewandte Wirtschaftsforschung)

17:00

Diskussion

Freitag, 23. September 2005

09:00

Ein Scientific-Use-File der Einzelhandelsstatistik
Michael Scheffler (Statistisches Bundesamt)

09:30

Ein Scientific-Use-File der Kostenstrukturerhebung
Dr. Rainer Lenz (Statistisches Bundesamt)

10:30

Ein Scientific-Use-File der Umsatzsteuerstatistik
Dr. Daniel Vorgrimler (Statistisches Bundesamt)

11:00

Diskussion

11:30

Ausblick: Anonymisierung von Paneldaten
Prof. Dr. Gerd Ronning (Institut für Angewandte Wirtschaftsforschung, Universität Tübingen)

Anhang B

Beteiligung an Konferenzen und Tagungen

Vorträge 2002

10. Oktober 2002

Im Rahmen des DFG-Rundgesprächs über Individualdaten zum Arbeitsmarkt beim ZEW in Mannheim referierte zum Thema

Anonymisierung von Mikrodaten aus dem Unternehmensbereich

Prof. Dr. Gerd Ronning [Universität Tübingen/IAW].

Vorträge 2003

6./7. März 2003

Im Rahmen des Workshops „Firmendaten aus der amtlichen Statistik (FiDaSt)“ an der FHTW Berlin referierte zum Thema

Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten – Strategien, Vorgehen, erste Ergebnisse

Dipl.-Volksw. Martin Rosemann [IAW].

7.-9. April 2003

Im Rahmen der „UN-ECE/Eurostat Work session on statistical data confidentiality“ in Luxembourg referierten zum Thema

SAFE – A Method for Statistical Disclosure Limitation of Microdata

Dipl.-Ökonom Jörg Höhne [Statistisches Landesamt Berlin]

und zum Thema

A graph theoretical approach to record linkage

Dr. Rainer Lenz [Statistisches Bundesamt].

9. Mai 2003

Im Rahmen des „Statistischen Kolloquiums 2003" des Statistischen Landesamtes Baden-Württemberg und der Universität Tübingen referierte zum Thema

Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten – Strategie, Vorgehen und erste Ergebnisse

Dipl.-Volksw. Martin Rosemann [IAW].

13.-20. August 2003

Im Rahmen der „54^{th} Session of the International Statistical Institute" in Berlin referierten zum Thema

Disclosure Risk of Anonymized Business Microdata Files – Illustrated with Empirical Key Variables

Dr. Daniel Vorgrimler; Dr. Rainer Lenz [Statistisches Bundesamt]

und zum Thema

Anonymization of Business Micro Data: A Glimpse of Work in Progress

Dipl.-Volksw. Roland Sturm [Statistisches Bundesamt].

20.-22. August 2003

Im Rahmen des „Workshop on Microdata" in Stockholm referierten zum Thema

A way to combine probabilistic with distance based record linkage

Dr. Rainer Lenz [Statistisches Bundesamt]

und zum Thema

Effects of anonymization on analytical validity and reidentification risk

Dr. Daniel Vorgrimler [Statistisches Bundesamt].

23. August 2003

Im Rahmen der „Statistischen Woche 2003“ in Potsdam referierte zum Thema

Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten

Dipl.-Volksw. Martin Rosemann [IAW].

15.-16. September 2003

Im Rahmen der „Comparative Analysis of (micro) Enterprise Data Conference (CAED 2003)“ in London referierte zum Thema

Disclosure of confidential information by means of multi-objective optimisation

Dr. Rainer Lenz [Statistisches Bundesamt].

Wintersemester 2003/2004

Für das Wahlpflichtfach Statistik im Bereich der Wirtschaftswissenschaftlichen Fakultät der Universität Tübingen las zum Thema

Statistische Verfahren zur Anonymisierung von Mikrodaten

Prof. Dr. Gerd Ronning [Universität Tübingen/IAW].

16. Dezember 2003

Am Fachbereich Mathematik und Informatik der Johannes-Gutenberg-Universität Mainz referierte zum Thema

Simulation eines Angriffs auf geheimhaltungspflichtige Unternehmensdaten

Dr. Rainer Lenz [Statistisches Bundesamt].

Vorträge 2004

5. März 2004

Im Ökonometrischen Ausschuss des Vereins für Socialpolitik in Rauischholzhausen referierte zum Thema

Estimation of the Probit Model and the ANOVA Model in Case of Misclassification

Prof. Dr. Gerd Ronning [Universität Tübingen/IAW].

27. März 2004

Im Rahmen des Statistisch-Ökonometrischen Kolloquiums an der Universität Würzburg referierte zum Thema

Estimation of the Probit Model and the ANOVA Model in Case of Misclassification

Prof. Dr. Gerd Ronning [Universität Tübingen/IAW].

26. April 2004

Am Rheinisch-Westfälischen Institut für Wirtschaftsforschung – RWI in Essen referierte zum Thema

Estimation of the Probit Model and the ANOVA Model in Case of Misclassification

Prof. Dr. Gerd Ronning [Universität Tübingen/IAW].

7. Mai 2004

An der Universität St. Gallen referierte zum Thema

Estimation of the Probit Model and the ANOVA Model in Case of Misclassification

Prof. Dr. Gerd Ronning [Universität Tübingen/IAW].

24.-26. Mai 2004

Im Rahmen der „European Conference on Quality and Methodology in Official Statistics (Q2004)“, veranstaltet vom Statistischen Bundesamt in Kooperation mit Eurostat u.a. europäischen Statistischen Ämtern in Mainz, referierte zum Thema

Estimation of the Probit Model Using Anonymized Micro Data

Dipl.-Volksw. Martin Rosemann [IAW]

und zum Thema

Simulation of a database cross match – as applied to the German structure of costs survey

Dr. Rainer Lenz [Statistisches Bundesamt].

4./5. Juni 2004

Im Rahmen der „2. Konferenz für Sozial- und Wirtschaftsdaten" in Wiesbaden gab eine

Einführung in das Forum Anonymisierungsfragen

Dr. Ruth Brand [Statistisches Bundesamt]

und es referierten zum Thema

Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten im Spannungsfeld von Datensicherheit und Analysepotenzial

Dipl.-Ökonom Jörg Höhne [Statistisches Landesamt Berlin]; Dipl.-Volksw. Martin Rosemann [IAW].

24.-26. Juni 2004

Im Rahmen des „Interdisciplinary Workshop on Duration and Survival Time Models – ZIF" an der Universität Bielefeld referierte zum Thema

Estimation of Models with Censoring or Truncation From Masked Data

Prof. Dr. Gerd Ronning [Universität Tübingen/IAW].

7.-9. Juli 2004

Im Rahmen der „Privacy in Statistical Databases Conference (PSD'2004)" in Barcelona referierte zum Thema

Matching German turnover tax statistics

Dr. Rainer Lenz [Statistisches Bundesamt].

6. August 2004

Im Rahmen einer Veranstaltung der Latvian Statistical Association in Riga (Lettland) referierte zum Thema

Scientific-Use-Files From Official Statistics. A German Project

Prof. Dr. Gerd Ronning [Universität Tübingen/IAW].

12./13. August 2004

Im Rahmen des ersten Treffens der Mitarbeiterinnen und Mitarbeiter der regionalen Standorte des Forschungsdatenzentrums der Länder mit dem Schwerpunkt „Einführung in die Anonymisierung von Mikrodaten“ referierten zum Thema

Methoden der Anonymisierung

Dipl.-Ökonom Jörg Höhne [Statistisches Landesamt Berlin]

und zum Thema

Konzept zur Beurteilung der Schutzwirkung von Anonymisierungsmaßnahmen

Dr. Daniel Vorgrimler [Statistisches Bundesamt].

24. November 2004

Im Rahmen einer Veranstaltung zur Anonymisierung an der Universität München referierten zum Thema

Auswirkungen von datenverändernden Anonymisierungsverfahren bei metrischen Variablen auf das Analysepotenzial wirtschaftsstatistischer Einzeldaten

Dipl.-Volksw. Martin Rosemann [IAW]

und zum Thema

Mikroökonometrische Aspekte der Anonymisierung von diskreten Merkmalen durch PRAM

Prof. Dr. Gerd Ronning [Universität Tübingen/IAW].

26. November 2004

Im Interdisziplinären Doktorandenkolloquium (IDK) der Ruprecht-Karls-Universität Heidel-

berg referierte zum Thema

Neuere Entwicklungen in der statistischen Geheimhaltung

Dr. Rainer Lenz [Statistisches Bundesamt].

Vorträge 2005

4.-13. April 2005

Im Rahmen der „55^{th} Session of the International Statistical Institute“ in Sydney referierte zum Thema

Anonymization of Business Microdata: Findings and Recommendations

Dipl.-Volksw. Roland Sturm [Statistisches Bundesamt].

21. April 2005

Auf der Regionalen Nutzerkonferenz „Amtliche Mikrodaten für die wissenschaftliche Forschung“ der Forschungsdatenzentren der Länder (FDZ) in Berlin referierten zum Thema

Mikroaggregationsverfahren und stochastische Überlagerungen als Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten

Dipl.-Volksw. Martin Rosemann [IAW]

und zum Thema

Erste Scientific-Use-Files aus den Wirtschaftsstatistiken

Dipl.-Volksw. Roland Sturm [Statistisches Bundesamt].

19. Mai 2005

Auf der Regionalen Nutzerkonferenz „Amtliche Mikrodaten für die wissenschaftliche Forschung“ der Forschungsdatenzentren der Länder (FDZ) am Institut für Weltwirtschaft in Kiel referierte zum Thema

Erste Scientific-Use-Files aus den Wirtschaftsstatistiken

Dr. Rainer Lenz [Statistisches Bundesamt].

7.-9. September 2005

Auf der Business Data Linking/Analysis of Enterprise Microdata (BDL/CAED 2005) Conference in Cardiff referierte zum Thema

Measuring the disclosure risk of masked enterprise microdata

Dr. Rainer Lenz [Statistisches Bundesamt].

9.-11. November 2005

Im Rahmen der „UN-ECE/Eurostat Work session on statistical data confidentiality“ in Genf referieren zum Thema

Estimation of the Probit Model From Anonymized Micro Data

Dipl.-Volksw. Martin Rosemann [IAW]; Prof. Dr. Gerd Ronning [Universität Tübingen/IAW]

und zum Thema

A standard for the Release of Business Microdata

Dr. Rainer Lenz [Statistisches Bundesamt].

Anhang C

Fachveröffentlichungen

Blien, U. (2003): „Die Reidentifikationsproblematik bei wirtschaftsstatistischen Einzeldaten". Einige allgemeine Gesichtspunkte in der Diskussion der Beiträge von Daniel Vorgrimler, Rolf Wiegert und Heike Wirth. In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

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.

Brand, R. (2003): Koreferat zum Beitrag „Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten". In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

Brand, R. und Sturm, R. (2002): „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten – Anonymisierungsverfahren: Entwicklungen und Praxistests". In: Methoden – Verfahren – Entwicklungen 2/2002 (Statistisches Bundesamt, Hrsg.), Wiesbaden, S. 6f.

Domingo-Ferrer, J., Mateo, M., Torres, A. (2003): „Concepts for the Evaluation of Anonymized Data". In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

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

Gießing, S. (2001): „The CASC Project: Integrating Best Practice Methods for Statistical Confidentiality". In: Proceedings der NTTS&ETK Konferenz, Hersonissos (Kreta) 18.-22. Juni 2001.

Gießing, S. (2003): Kommentar zum Beitrag „Concepts for the Evaluation of Anonymized Data". In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

Gnoss, R. (2003): „Das Projekt Faktische Anonymisierung wirtschaftsstatistischer Einzel-

daten – Ziele, Ablauf, Beteiligte“. In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

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

Gottschalk, S. (2003): „Microdata Disclosure by Resampling – Empirical Findings for Business Survey Data“. Joint UN-ECE/Eurostat work session on statistical dada confidentiality, Luxembourg.

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

Höhne, J. (2003b): „Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten“. In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

Höhne, J. (2003c): „SAFE – a method for statistical disclosure limitation of microdata“. Monographs of Official Statistics – Research in Official Statistics (Paper at the Joint UN-ECE/Eurostat work session on statistical dada confidentiality, Luxembourg.)

Höhne, J., Sturm, R. und Vorgrimler, D. (2003): „Konzept zur Beurteilung der Schutzwirkung von faktischer Anonymisierung“. In: Wirtschaft und Statistik, Heft 4, S.287-292.

Lechner, S. und W. Pohlmeier (2003): „Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten“. In: Forum der Bundesstatistik, Band 42, Wiesbaden.

Lechner, S. und W. Pohlmeier (2004): „Data Masking by Noise Addition and the Estimation of Nonlinear Regression Models“. Beitrag zum Workshop: Econometric Analyses of Anonymised Firm Data, 18./19. März 2004 in Tübingen.

Lechner, S. und W. Pohlmeier (2005): „Data Masking by Noise Addition and the Estimation of Nonparametric Regression Models“. Erscheint in: Jahrbücher für Nationalökonomie und Statistik, Band 225, Heft 5.

Lenz, R. (2003a): „A graph theoretical approach to record linkage“. Monographs of Official Statistics – Research in Official Statistics, Eurostat, S.324-334. (Paper at the Joint UN-ECE/Eurostat work session on statistical data confidentiality, Luxembourg.)

Lenz, R. (2003b): „A way to combine probabilistic with deterministic record linkage“. Erscheint in: Proceedings of the Workshop on Microdata, Stockholm.

Lenz, R. (2003c): „Disclosure of confidential information by means of multi-objective optimisation“. Proceedings of the Comparative Analysis of Enterprise (micro) Data Conference (CAED), London. Siehe: http://www.statistics.gov.uk/events/caed/abstracts/lenz.asp

Lenz, R. und Sturm, R. (2003): „Entwicklung eines leistungsfähigen Record Linkage Verfahrens“. In: Methoden – Verfahren – Entwicklungen 2/2003 (Statistisches Bundesamt, Hrsg.), Wiesbaden, S. 8f.

Lenz, R. , Sturm, R. und Vorgrimler, D. (2004): „Maße für die faktische Anonymität von Mikrodaten“. In: Wirtschaft und Statistik, Heft 6, S.621-638.

Lenz, R. und Vorgrimler, D. (2004): „Geheimhaltungsmethoden auf dem Prüfstand – eine Analyse anhand der Umsatzsteuerstatistik“. In: Wirtschaft und Statistik, Heft 6, S.639-648.

Lenz, R. und Vorgrimler, D. (2005): „Matching German turnover tax statistics“. In: Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, Arbeitspapier Nr. 4. Siehe: http://www.forschungsdatenzentrum.de/publikationen/arbeitspapiere/04.asp

Lenz, R. und Sturm, R. (2004): „Development of a powerful record linkage algorithm“. In: Methods – Approaches – Developments 1/2004.

Lenz, R., Vorgrimler, D. und Rosemann, M. (2005): „Ein Scientific-Use-File der Kostenstrukturerhebung im Verarbeitenden Gewerbe“. In: Wirtschaft und Statistik, Heft 2, S.91-96.

Lenz, R., Doherr, T. und Vorgrimler, M. (2004): „Simulation of a database cross match – as applied to the German structure of costs survey“. In: Proceedings of the European Conference on Quality and Methodology in Official Statistics (Q2004), CD-ROM publication, ISBN 3-8246-0733-6.

Lenz, R. (2005): „A standard for de facto anonymity of business microdata“. In: Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder, Arbeitspapier Nr. 6. Siehe: http://www.forschungsdatenzentrum.de/publikationen/arbeitspapiere/06.asp

Licht, G. (2003): Koreferat zum Beitrag „Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik“. In: Forum der Bundesstatistik, Band 42, Wiesbaden.

Ronning, G. (2003): „Was hat die Veranstaltung gebracht? Ein Resumee der Tagung“. In: Forum der Bundesstatistik, Band 42, Wiesbaden.

Ronning, G. (2003): Ökonometrie und Anonymisierung von Mikrodaten: Koreferat zum Beitrag „Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten“. In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

Ronning, G. (2005): „Randomized response and the binary probit model“, Economics letters, 86, S. 221-228.

Ronning, G. und M. Rosemann, M. (2004): „Estimation of the Probit Model from Anonymized Data“. Beitrag zum Workshop 'Econometric Analysis of anonymised firm data',

Tübingen, März 2004.

Ronning, G., M. Rosemann und H. Strotmann (2005): „Post-Randomization Under Test: Estimation of the Probit Model". Erscheint in: Jahrbücher für Nationalökonomie und Statistik, Band 225, Heft 5.

Rosemann, M (2003): „Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten und nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung und der Umsatzsteuerstatistik". In: Forum der Bundesstatistik, Band 42, Wiesbaden.

Rosemann, M. und Vorgrimler, D. (2003): „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten – Strategien, Verfahren, erste Ergebnisse". Beitrag zur Statistischen Woche 2003.

Rosemann, M. und Vorgrimler, D. (2003): „Effects of anonymization on analytical validity and reidentification risk". Workshop on Microdata, Statistik Schweden, Stockholm.

Rosemann, M. (2003): „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten – Strategien, Vorgehen, erste Ergebnisse". In: Pohl, R.; Fischer, J.; Rockmann, U. und Semlinger, K. (Hrsg.): Analysen zur regionalen Industrieentwicklung – Sonderauswertungen einzelbetrieblicher Daten der Amtlichen Statistik. Statistisches Landesamt Berlin, Seiten 157-184.

Rosemann, M., Vorgrimler, D. und Lenz, R. (2004): „Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten". In: Allgemeines Statistisches Archiv 88 (1), Seiten 73-99.

Rosemann, M. (2004): „Impacts of different versions of micro aggregation on the results of linear estimations". Beitrag zum Workshop Econometric Analysis of anonymised firm data, Tübingen, März 2004.

Scheffler, M. (2005): „Ein Scientific-Use-File der Einzelhandelstatistik 1999". In: Wirtschaft und Statistik, Heft 3, S.197-200.

Strotmann, H. (2003): Zu den Erwartungen der Datennutzer an das Anonymisierungsprojekt: Koreferat zum Beitrag „Arbeiten mit Einzeldaten der amtlichen Statistik am Beispiel des Monatsberichts im Verarbeitenden Gewerbe". In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

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

Sturm, R. (2002): „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten". In: Allgemeines Statistisches Archiv 86, S. 468 - 477.

Sturm, R. (2003): „Anonymization of business micro data: a glimpse of work in progress". In: ISI 2003 Contributed Papers Book 2.

Sturm, R. (2003).: „Anonymisierung von wirtschaftsstatistischen Einzeldaten – Ein Werkstattbericht." In: Wirtschaft und Statistik, Sonderheft, S.78-80.

Vorgrimler, D. (2003): „Reidentifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios". In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

Vorgrimler, D. und Lenz, R. (2003): „Über das Risiko der Reidentifikation in wirtschaftsstatistischen Einzeldaten". In: Wirtschaft und Statistik, Sonderheft, S.81-82.

Vorgrimler, D. und Lenz, R. (2003): „Disclosure Risk of Anonymized Business Microdata Files – Illustrated with Empirical Key Variables". In: Bulletin of the 54th International Statistical Institute (ISI), book 2, S.594-595.

Vorgrimler, D., Dittrich, S., Lenz, R. und Rosemann, M. (2005): „Ein Scientific-Use-File der Umsatzsteuerstatistik 2000". In: Wirtschaft und Statistik, Heft 3, S.201-210.

Vorgrimler, D., Dittrich, S., Lenz, R. und Rosemann, M. (2005): „Wissenschaftliche Analysen mit Hilfe der amtlichen Umsatzsteuerstatistik". In: Wirtschaftswissenschaftliches Studium, Heft 10, S.527-532.

Wagner, J. (2003): „Arbeiten mit Einzeldaten der amtlichen Statistik am Beispiel des Monatsberichts im Verarbeitenden Gewerbe". In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

Wiegert, R. (2003): „Matching Verfahren und die Reidentifikation faktisch anonymisierter Einzeldaten". In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

Wirth, H. (2003): „Szenarien für Angriffe auf wirtschaftsstatistische Einzeldaten - Ein Überblick". In: Forum der Bundesstatistik, Band 42, Wiesbaden 2003.

Literaturverzeichnis

Abowd, J. und Woodcock, S. (2002): Disclosure Limitation in Longitudinal Linked Data. In: Doyle, P., Lane, J., Theeuwes, J. und Zayatz, L. (Hrsg.), *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies*, S. 215–278

Baeyens, Y. und Defays, D. (1999): Estimation of Variance Loss Following Microaggregation by the Individual Ranking Method. *Proceedings of Statistical Data Protection*, Bd. 98, S. 101–108

Baltagi, B. (2001): *Randomisierung der Datenverarbeitung*. Addison-Wesley

Bellmann, L., Kohaut, S. und Schnabel, C. (1999): Flächentarifverträge im Zeichen von Abwanderung und Widerspruch: Geltungsbereich, Einflussfaktoren und Öffnungstendenzen. In: Bellmann, L. und Steiner, V. (Hrsg.), *Panelanalysen zu Lohnstruktur, Qualifikation und Beschäftigungsdynamik*, Bd. 229 von *Beiträge zur Arbeitsmarkt- und Berufsforschung*, S. 11–40

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

Brand, R. (2002): Masking through Noise Addition. In: Domingo-Ferrer, J. (Hrsg.), *Inference Control in Statistical Databases – From Theory to Practice, Lecture Notes in Computer Science*

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

Carroll, R., Küchenhoff, H., Lombard, F. und Stefanski, L. (1996): Asymptotics for the SIMEX Estimator in Nonlinear Measurement Error Models. *Journal of the American Statistical Association*, Bd. 91(433), S. 242–250

Carroll, R., Ruppert, D. und Stefanski, L. (1995): *Measurement Error in Nonlinear Models*. Chapman and Hall, London

Cook, J. und Stefanski, L. (1994): Simulation-Extrapolation Estimation in Parametric Measurement Error Models. *Journal of the American Statistical Association*, Bd. 89(428), S. 1314–1328

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

Dandekar, R., Cohen, M. und Kirkendall, N. (2001): Applicability of Latin Hypercube Sampling to Create Multivariate Synthetic Micro Data. In: *Proceedings of ETK-NTTS*, S. 839–847. Eurostat, Luxemburg

Dandekar, R., Domingo-Ferrer, J. und Sebé, F. (2002): LHS-Based Hybrid Microdata vs Rank Swapping and Micoraggregation for Numeric Microdata Protection. In: Domingo-Ferrer, J. (Hrsg.), *Inference Control in Statistical Databases – From Theory to Practice*. Springer, Berlin

Devroye, L. und Györfi, L. (1985): *Nonparametric Density Estimation*. New York

Dittrich, S. (2004): Umsätze und ihre Besteuerung. *Wirtschaft und Statistik*, Bd. 10, S. 1195–1200

Domingo-Ferrer, J., Mateo, J. und Torres, A. (2003): Concepts for the Evaluation of Anonymized Data. In: Gnoss, R. und Ronning, G. (Hrsg.), *Anonymisierung wirtschaftsstatistischer Einzeldaten*, Bd. 42, S. 100–110. Metzler-Poeschel, Wiesbaden

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. *Second Eurostat-UN/ECE Joint Work Session on Statistical Data Confidentiality, Skopje, Macedonia*

Duncan, G. und Lambert, D. (1989): The Risk of Disclosure for Microdata. *Journal of Business & Economic Statistic*, Bd. 7(2), S. 207–217

Elliot, M. und Dale, A. (1999): Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk. *Netherlands Official Statistics*, S. 6–10

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

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

Fienberg, S., Steele, R. J. und Makov, U. (1996): Statistical Notions of Data Disclosure Avoidance and Their Relationship to Traditional Statistical Methodology: Data Swapping and Loglinear Models. *Proceedings of the US Bureau of the Census 1996 Annual Research Conference*, S. 87–105

Frazis, H. und Loewenstein, M. (2003): Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables. *Journal of Econometrics*, Bd. 117, S. 151–178

Fritsch, M. und Stephan, A. (2003): Die Heterogenität der technischen Effizienz innerhalb von Wirtschaftszweigen – Auswertungen auf Grundlage der Kostenstrukturstatistik des Statistischen Bandesamtes. In: Pohl, R., Fischer, J., Rockmann, U. und Semlinger, K. (Hrsg.), *Analysen zur regionalen Industrieentwicklung – Sonderauswertungen einzelbetrieblicher Daten der Amtlichen Statistik*, S. 143–156. Statistisches Landesamt Berlin

Fürnrohr, M., Rimmelspacher, B. und von Roncador, T. (2002): Zumsammenführung von Datenbeständen ohne numerischen Identifikatoren. *Bayern in Zahlen*, S. 308–321

Fuller, W. (1980): Properties of Some Estimators for the Errors-in-Variables Model. *The Annals of Statistics*, Bd. 8, S. 407–422

Fuller, W. (1984): Measurement Error Models with Heterogeneous Error Variances. In: Chaubey, Y. und Dwivedi, T. (Hrsg.), *Topics in Applied Statistics*, S. 257–289. Concordia University

Fuller, W. (1987): *Measurement Error Models*. Wiley, New York

Gottschalk, S. (2004): Microdata Disclosure Control by Resampling – Empirical Findings for Business Survey Data. *Allgmeines Statistisches Archiv*, Bd. 88(3), S. 279–302

Gottschalk, S. (2005): *Unternehmensdaten zwischen Datenschutz und Analysepotenzial*. NOMOS-Verlag

Gräb, C. und Zwick, M. (2002): Die Umsatzsteuerstatistik. In: Fritsch, M. und Grotz, R. (Hrsg.), *Das Gründungsgeschehen in Deutschland – Darstellung und Vergleich der Datenquellen*, S. 129–140. Physica-Verlag, Heidelberg

Greene, W. (2003): *Econometric Analysis*. Prentice Hall, Upper Saddle River, 5. Aufl.

Hausman, J., Abrevaya, J. und Scott-Morton, F. (1998): Misclassification of the Dependent Variable in a Discrete-Response Setting. *Journal of Econometrics*, Bd. 87, S. 239–269

Helmcke, T. und Knoche, P. (1992): Projekt zur faktischen Anonymität von Mikrodaten – Bericht über ein Forschungsprojekt. *Wirtschaft und Statistik*, S. 139–144

Höhne, J. (2002): Messung der Qualität einer anonymen Datei. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Höhne, J. (2003a): Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten. In: Gnoss, R. und Ronning, G. (Hrsg.), *Anonymisierung wirtschaftsstatistischer Einzeldaten*, S. 69–94. Metzler-Poeschel, Wiesbaden

Höhne, J. (2003b): SAFE – A Method for Statistical Disclosure Limitation of Micro Data. *Contributed Paper for ECE/Eurostat Work Session On Statistical Data Confidentiality, April 7th to 9th 2003, Luxemburg*

Höhne, J. (2003c): SAFE – Ein Verfahren zur Anonymisierung statistischer Einzelangaben. *Statistisches Landesamt Berlin (Hg.): Sonderdruck, Statistische Monatsschrift*. Nr. 3

Höhne, J. (2004a): Varianten von Zufallsüberlagerungen. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Höhne, J. (2004b): Weiterentwicklung von Mikroaggregationsverfahren. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Höhne, J., Sturm, R. und Vorgrimler, D. (2003): Konzept zur Beurteilung der Schutzwirkung von faktischer Anonymisierung. *Wirtschaft and Statistik*, Bd. 4, S. 287–299

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

Iman, R. L. und Conover, W. J. (1982): A Distribution-Free Approach to Inducing Rank Correlation among Input Variables. *Communications in Statistics*, Bd. 11(3), S. 311–334

Janz, N., Ebling, G., Gottschalk, S. und Niggemann, H. (2001): The Mannheim Innovation Panels (MIP and MIP-S) of the Centre for European Economic Research (ZEW). *Schmollers Jahrbuch*, Bd. 121, S. 123–129

Kane, T., Rouse, C. und Staiger, D. (1999): Estimating Returns to Schooling When Schooling is Misreported. *NBER Working Paper 7235*

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

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

Kim, J. und Winkler, W. (1997): Masking Microdata Files. *Statistical Research Division RR97/03, U.S. Bureau of the Census,Washington, D.C.*, S. 114–119

Kim, J. und Winkler, W. (2001): Multiplicative Noise for Masking Continuous Data. In: *Proceedings of the Section on Survey Research Methods*. American Statistical Association

Kohaut, S. und Schnabel, C. (2003): Tarifverträge – nein danke!? – Ausmaß und Einflussfaktoren der Tarifbindung west- und ostdeutscher Betriebe. *Jahrbücher für Nationalökonomie und Statistik*, Bd. 23(3), S. 312–331

Kooiman, P., Willenborg, L. und Gouweleeuw, J. (1997): PRAM: A Method for Disclosure Limitation of Micro Data. *Department of Statistical Methods, Statistical Netherlands, Voorburg*

Kuhn, H. W. (1955): The Hungarian Method for the Assignment Problem. *Naval Research Logistics Quarterly*, Bd. 2, S. 83–97

KVI (2001): *Wege zu einer besseren informationellen Infrastruktur – Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik*. Nomos, Baden-Baden

Lechner, S. und Pohlmeier, W. (2003): Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten. In: Gnoss, R. und Ronning, G. (Hrsg.), *Anonymisierung wirtschaftsstatistischer Einzeldaten*, Bd. 42 von *Forum der Bundesstatistik*, S. 115–137. Wiesbaden

Lechner, S. und Pohlmeier, W. (2004): Data Masking by Noise Addition and the Estimation of Nonlinear Regression Models. *Beitrag zum Workshop: 'Econometric Analyses of Anonymised Firm Data', 18./19. März 2004 in Tübingen*

Lechner, S. und Pohlmeier, W. (2005): Data Masking by Noise Addition and the Estimation of Nonparametric Regression Models. *Erscheint in: Jahrbücher für Nationalökonomie und Statistik*, Bd. 225(5)

Lehmann, K. (2002): *Stabilität und Veränderung der Flächentarifbindung von Arbeitgebern in Deutschland – eine theoretische und empirische Analyse*. Lit Verlage, Münster

Lenz, R. (2003a): Disclosure of Confidential Information by Means of Multi Objective Optimisation. *Proceedings of the Comparative Analysis of (Micro) Enterprise Data Conference (CAED), London (published on CD-ROM)*

Lenz, R. (2003b): A Graph Theoretical Approach to Record Linkage. *Monographs of Official Statistics – Research in Official Statistics*, S. 324–334

Lenz, R. (2003c): A Way To Combine Probabilistic With Deterministic Record Linkage. *Erscheint in: Proceedings of the Workshop on Microdata, Stockholm*

Lenz, R., Doherr, T. und Vorgrimler, D. (2004a): Simulation of A Database Cross Match – As Applied to the German Structure of Costs Survey. *Proceedings of the European Conference on Quality and Methodology in Official Statistics (Q2004), Mainz (published on CD-ROM)*

Lenz, R., Sturm, R. und Vorgrimler, D. (2004b): Maße für die faktische Anonymität von Mikrodaten. *Wirtschaft und Statistik*, Bd. 6, S. 621–638

Lenz, R. und Vorgrimler, D. (2004): Geheimhaltungsmethoden auf dem Prüfstand – eine Analyse anhand der Umsatzsteuerstatistik. *Wirtschaft und Statistik*, S. 639–648

Lenz, R. und Vorgrimler, D. (2005): Matching German Turnover Tax Statistics. *Forschungsdatenzentren des Bundes und der Länder – Arbeitspapier Nr. 4*

Lin, A. (1989): Estimation of Multiplicative Measurement Error Models and Some Simulation Results. *Economics letters*, Bd. 31, S. 13–20

Little, R. (1993): Statistical Analysis of Masked Data. *Journal of Official Statistics*, Bd. 9, S. 407–426

Lütkepohl, H. (1997): *Handbook of Matrices*. John Wiley & Sons LtD

Mateo-Sanz, J. und Domingo-Ferrer, J. (1998): A Method For Data-Oriented Multivariate Microaggregation. *Statistical data protection, Proceedings of the conference Eurostat 1999*

McKay, M. D., Conover, W. J. und Beckman, R. J. (1979): A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics*, Bd. 21(2), S. 239–245

McLachlan, G. und Peel, D. (2000): *Finite Mixture Models*. Wiley, New York

Müller, W., Blien, U., Knoche, P. und Wirth, H. (1991): *Die faktische Anonymität von Mikrodaten*. Statistisches Bundesamt, Wiesbaden

Parzen, E. (1962): On Estimation of a Probability Density Function and Mode. *Annals of Mathematical Statistics*, Bd. 32, S. 1065–1076

Pollettini, S., Franconi, L. und Stander, J. (2002): Model Based Disclosure Protection. In: Domingo-Ferrer, J. (Hrsg.), *Inference Control in Statistical Data Bases – From Theory to Practice*. Springer, Berlin

Raghunathan, T., Reiter, J. und Rubin, D. (2003): Multiple Imputation for Statistical Disclosure Limitation. *Journal of Official Statistics*, Bd. 19, S. 1–16

Ronning, G. (1991): *Mikroökonometrie*. Springer, Berlin

Ronning, G. (2003): Neuere Entwicklungen in der Mikroökonometrie. In: Franz, W., Stadler, M. und Ramser, H. (Hrsg.), *Empirische Wirtschaftsforschung: Methoden und Anwendungen*, S. 41–50. Mohr-Siebeck, Tübingen. (Korreferat)

Ronning, G. (2004a): Fehlklassifikation im Modell der Varianzanalyse. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Ronning, G. (2004b): Mischung von Verteilungen und Anonymisierung. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Ronning, G. (2004c): Stochastische Überlagerung – einige grundsätzliche Überlgegungen. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Ronning, G. (2005): Randomized Response and the Binary Probit Model. *Economics letters*, Bd. 86, S. 221–228

Ronning, G., Brand, R., Höhne, J., Rosemann, M. und Wiegert, R. (2002): Anonymisierungsverfahren – Überblick and erste Bewertung. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Ronning, G. und Rosemann, M. (2003): Ansätze zur Operationalisierung des Analysepotenzials bei anonymisiertenDaten. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Ronning, G. und Rosemann, M. (2004): Estimation of the Probit Model from Anonymised Data. *Beitrag zum Workshop 'Econometric Analysis of Anonymised Firm Data',Tübingen, März 2004*

Ronning, G., Rosemann, M. und Strotmann, H. (2005): Post-Randomization Under Test: Estimation of the Probit Model. *Erscheint in: Jahrbücher für Nationalökonomie und Statistik*, Bd. 225(5)

Roque, G. (2000): *Masking Microdata Files with Mixtures of Multivariate Normal Distributions*. Dissertation, University of California, Riverside

Rosemann, M. (2003): Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisiertenand nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebungand der Umsatzsteuerstatistik. In: Gnoss, R. und Ronning, G. (Hrsg.), *Anonymisierung wirtschaftsstatistischer Einzeldaten*, Bd. 42 von *Forum der Bundesstatistik*, S. 154–183. Wiesbaden

Rosemann, M. (2004): Impacts Of Different Versions of Micro Aggregation on the Results of Linear Estimations. *Beitrag zum Workshop 'Econometric Analysis of anonymised firm data', Tübingen, März 2004*

Rosemann, M. (2005): Auswirkungen datenverändernder Anonymisierungsverfahren auf das Analysepotenzial wirtschaftsstatistischer Einzeldaten. Bisher unveröffentlichte Dissertationsschrift

Rosemann, M. und Vorgrimler, D. (2004): Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten – Strategien, Vorgehen and erste Ergebnisse. *Statistische Analysen*

Rosemann, M., Vorgrimler, D. und Lenz, R. (2004): Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischerEinzeldaten. *Allgemeines Statistisches Archiv*, Bd. 88, S. 73–99

Rubin, D. (1993): Discussion. Statistical Disclosure Limitation. *Journal of Official Statistics*, Bd. 9(2), S. 461–468

Rubin, D. und Schenker, N. (1991): Multiple Imputation in Health-Care Databases: An Overview and Some Applications. *Statistics in Medicine*, Bd. 10, S. 585–598

Schmid, M., Schneeweiß, H. und Küchenhoff, H. (2005): Consistent Estimation of a Simple Linear Model Under Microaggregation. *SFB Discussion Paper No. 415*

Sebé, F., Domingo-Ferrer, J., Mateo-Sanz, J. und Torra, V. (2002): Post-Masking Optimization of the Tradeoff Between Information Loss and Disclosure Risk in Masked Microdata Sets. In: Domingo-Ferrer, J. (Hrsg.), *Inference Control in Statistical Data Bases – From Theory to Practice*, S. 163–171. Springer, Berlin

Silverman, B. (1986): Density Estimation for Statistics and Data Analysis. *Monographs on Statistics and Applied Probability*, Bd. 26

Särndal, C.-E., Swensson, B. und Wretman, J. (1992): *Model Assisted Survey Sampling*. Springer, New York

Statistische Ämter des Bundes und der Länder und IAW (2003): *Forschungsprojekt: 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten' – Zwischenbericht 2003 an das BMBF*. Statistisches Bundesamt, Wiesbaden

Stefanski, L. und Carroll, R. (1985): Covariate Measurement Error in Logistic Regression. *The Annals of Statistics*, Bd. 13, S. 1335–1351

Stefanski, L. und Cook, J. (1996): Simulation Extrapolation: The Measurement Error Jackknife. *Journal of the American Statistical Association*, Bd. 90, S. 1247–1256

Strotmann, H. (2001): *Arbeitsplatzdynamik in der baden-württembergischen Industrie - Eine Analyse mit amtlichen Betriebspaneldaten*. Peter Lang, Frankfurt am Main

Strotmann, H. (2002): Tarifbindung in Baden-Württemberg im Jahr 2000 – ist der Flächentarifvertrag ein Auslaufmodell? *IAW-Mitteilungen*, Bd. 1/2002, S. 4–14

Strotmann, H. (2004): The Impact of Anonymisation on Binary Choice Models – Empirical Evidence from Monte Carlo Simulations Using the IAB Establishment Panel Baden-Wuerttemberg. *Beitrag zum Workshop 'Econometric Analysis of anonymised firm data', Tübingen, März 2004*

Sturm, R. (2002a): Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten. *Allgemeines Statistisches Archiv*, Bd. 86, S. 468–477

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

Sullivan, G. (1989): The Use of Added Error to Avoid Disclosure in Microdata Releases. Unpublished PhD Thesis, Iowa State University

Van den Hout, A. und van der Heijden, P. (2002): Randomized Response, Statistical Disclosure Control and Misclassification: A Review. *International Statistical Review*, Bd. 70, S. 269–288

Vorgrimler, D. (2002): Aspekte faktischer Anonymisierung. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Vorgrimler, D. (2003a): Probe-Anonymisierung der Umatzsteuerstatistik mit traditionellen Methoden. *Arbeitspapier des Projekts 'Faktische Anonymsisierung wirtschaftsstatistischer Einzeldaten'*

Vorgrimler, D. (2003b): Probe-Anonymisierung der Umsatzsteuerstatistik. *Arbeitspapier des Projekts 'Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten'*

Vorgrimler, D. (2003c): Reidentifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios. In: Gnoss, R. und Ronning, G. (Hrsg.), *Anonymisierung wirtschaftsstatistischer Einzeldaten*, Bd. 42 von *Forum der Bundesstatistik*, S. 40–59. Wiesbaden

Vorgrimler, D., Dittrich, S., Lenz, R. und Rosemann, M. (2005): Ein Scientific-Use-File der Umsatzsteuerstatistik 2000. *Wirtschaft und Statistik*, S. 201–210

Wagner, J. (1992): Firm Size, Firm Growth, and Persistence of Chance: Testing Gibrat´s Law with Establishment Data from Lower Saxony, 1978-1989. *Small Business Economics*, Bd. 4, S. 125–131

Wagner, J. (1994a): The Post-Entry Performance of New Small Firms in Manufacturing Industries. *Journal of Industrial Economics*, Bd. 42(2), S. 141–154

Wagner, J. (1994b): Small Firm Entry in Manufacturing Industries: Lower Saxony, 1979-1989. *Small Business Economics*, Bd. 6, S. 211–223

Warner, S. (1965): Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association*, Bd. 57, S. 622–627

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

Wirth, H. (2003): Szenarien für Angriffe auf wirtschaftsstatistische Einzeldaten – Ein Überblick. In: Gnoss, R. und Ronning, G. (Hrsg.), *Anonymisierung wirtschaftsstatistischer Einzeldaten*, Bd. 42 von *Forum der Bundesstatistik*, S. 11–24. Metzler-Poeschel, Wiesbaden

Yancey, W., Winkler, W. und Creezy, R. (2002): Disclosure Risk Assessment in Perturbative Micro Data Protection. In: Domingo-Ferrer, J. (Hrsg.), *Inference Control in Statistical Databases*, S. 135–152. Springer, Berlin

Index

Danksagung der Autoren

An dieser Stelle möchten wir einer Reihe von Personen danken, die zum Gelingen des vorliegenden Handbuchs beigetragen haben.

Für ihre inhaltlich aktive Mitarbeit danken wir herzlich der Projektgruppe „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“, die sich aus Mitarbeitern der Fachgruppen der Statistischen Ämter des Bundes und der Länder, des Instituts für Angewandte Wirtschaftsforschung (IAW), des Zentrums für Europäische Wirtschaftsforschung (ZEW) und des Instituts für Arbeitsmarkt- und Berufsforschung (IAB) zusammensetzte, Herrn Christian Arndt, Herrn Dr. Vladislav Bajaja, Herrn Stefan Bender, Herrn Stefan Dittrich, Herrn Thorsten Doherr, Herrn Dr. Sigurd Duckwitz, Herrn Joachim Engel, Frau Sarah Gießing, Frau Dr. Sandra Gottschalk, Herrn Christopher Gräb, Herrn Gustav Grillmaier, Herrn Ottmar Hennchen, Herrn Günther Klee, Frau Dr. Antje Krüger, Herrn Rainer Opfermann, Frau Dr. Anke Saebetzki, Herrn Dietmar Schrödter, Herrn Dr. Gerhard Stock, Herrn Dr. Harald Strotmann, Herrn Mario Walter, Herrn Erwin Wartenberg und Herrn Dr. Rolf Wiegert.

Für die projektbegleitende wissenschaftliche Beratung gebührt dem wissenschaftlichen Begleitkreis des Projektes, dem Herr PD Dr. Uwe Blien, Herr Prof. Dr. Reinhard Hujer, Herr Dr. Georg Licht, Herr Prof. Dr. Winfried Pohlmeier, Herr Prof. Dr. Gerhard Wagenhals, Herr Prof. Dr. Joachim Wagner und Frau Dr. Heike Wirth angehörten, unser Dank. Herr Prof. Dr. Josep Domingo Ferrer, Universität Tarragona (Spanien),Herr Ramesh A. Dandekar, U.S. Department of Energy, Frau Sandra Lechner, Universität Konstanz und Herr Kersten Magg, Universität Tübingen, haben für die Projektarbeiten wesentliche Software zur Verfügung gestellt. Zu einigen im Projekt entwickelten Verfahren waren die Anregungen von Herrn Dr. William E. Winkler, U.S. Census Bureau, sehr hilfreich.

Weiterer Dank gebührt den vielen Personen, die bei den vielfältigen Projektarbeiten in den beteiligten Institutionen geholfen haben: Frau Joyce Agbonkhese, Frau Petra Czarkowski, Frau Mareile Drechsler, Frau Anke Fink, Frau Hana Fischer, Herr Dr. Hans-Peter Hafner, Herr Holger Herrmann, Frau Irena Juznic, Frau Marija Kurtschanowa, Herr Muhammad Mughal, Herr Christoph Müller, Frau Anja Münch, Frau Franziska Peter, Herr Rémi Piatek, Frau Ramona Pohl, Herr Heiko Rüger, Herr Haymo Schneider, Herr Dieter Schukmann, Frau Simone Schüßler, Frau Birgit Ullrich und Herr Christian Wingerter. Jeder von ihnen hat auf seine eigene Weise zu einem erfolgreichen Abschluss des Projektes „Faktische Ano-

nymisierung wirtschaftsstatistischer Einzeldaten“ und damit auch zu der Entstehung des Handbuchs beigetragen. Herr Martin Weiß hat für die Unterstützung der Autoren bei der technischen Umsetzung und Arbeiten am Layout dieses Handbuchs besonderen Dank verdient.