DESTATIS
wissen. nutzen.

STATISTIK UND WISSENSCHAFT

Rainer Lenz

Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung

Band 18

Statistisches Bundesamt

DESTATIS
wissen. nutzen.

STATISTIK UND WISSENSCHAFT

Rainer Lenz

Methoden der Geheimhaltung wirtschaftsstatistischer Einzeldaten und ihre Schutzwirkung

Band 18

Statistisches Bundesamt

Bibliographische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliographie; detaillierte bibliographische Daten sind im Internet über www.d-nb.de abrufbar.

Herausgeber: Statistisches Bundesamt, Wiesbaden

Internet: www.destatis.de

Ihr Kontakt zu uns:

www.destatis.de/kontakt

Informationen zu dieser Publikation unter

Tel.: +49 (0) 681 / 5 86 72 44
rainer.lenz@htw-saarland.de

Statistischer Informationsservice

Tel.: +49 (0) 611 / 75 24 05
Fax: +49 (0) 611 / 75 33 30

Erschienen im Dezember 2010

Print
Preis: EUR 24,80 [D]
Bestellnummer: 1030818-10900-1
ISBN: 978-3-8246-0906-2

Kostenfreier Download (PDF)
Artikelnummer: 1030818-10900-4
ISBN: 978-3-8246-0907-9

Vertriebspartner: HGV Hanseatische Gesellschaft für Verlagsservice mbH
Servicecenter Fachverlage
Postfach 11 64
72125 Kusterdingen
Tel.: +49 (0) 70 71 / 93 53 50
Fax: +49 (0) 70 71 / 93 53 35
destatis@s-f-g.com

Vorwort

Anonymisierte Einzeldaten müssen zwei gleichrangigen Ansprüchen genügen. Das ist zum einen der wirkungsvolle Schutz der Befragten vor der möglichen Aufdeckung vertraulicher Informationen, zum anderen der Erhalt ihrer analytischen Aussagekraft. Beide Ziele sind sowohl für die Datennutzer als auch für die statistischen Ämter von großer Bedeutung.

Skeptiker hielten noch bis zum Jahre 2005 die sogenannte faktische Anonymisierung, verankert in § 16 Absatz 6 des Bundesstatistikgesetzes, im Falle wirtschaftsstatistischer Einzeldaten für aussichtslos. Die vorliegende Habilitationsschrift, die bei Herrn Prof. Dr. Walter Krämer an der Technischen Universität Dortmund erbracht wurde, liefert nun sowohl die theoretischen Grundlagen zur Operationalisierung der faktischen Anonymität als auch deren direkte empirische Umsetzung. Die Verfahren werden erstmalig in realistischen Datenangriffsszenarien getestet. Dabei zeigt sich erfreulicherweise, dass die Anonymisierungsverfahren einen sicheren Datenschutz gewährleisten. Die vorliegende Arbeit leistet damit Grundlagenforschung zur Verbesserung der Arbeitsbedingungen der empirischen Wirtschafts- und Sozialforschung insgesamt. Die statistischen Ämter können nun mit vertretbarem Aufwand der Wissenschaft eine Forschung auf Basis von Unternehmens- und Betriebsdaten ermöglichen.

Die Generierung faktisch anonymer Daten im Querschnitt gehört in den statistischen Ämtern nunmehr zur alltäglichen Arbeit. Auch für Längsschnittsdaten und insbesondere für Paneldaten, bei denen einem potentiellen Datenangreifer lückenlose Informationen über alle interessierenden bzw. verfügbaren Wellen hinweg vorliegen könnten, liegen seit Ende des Jahres 2008 erste Datenangebote zu den Monatsberichten und der Kostenstrukturerhebung im Verarbeitenden Gewerbe sowie zur Umsatzsteuerstatistik vor.

Die Schrift ist durch die Mitwirkung des Autors in verschiedenen Projekten im Statistischen Bundesamt und mit Fördermitteln des Bundesministeriums für Bildung und Forschung (BMBF) entstanden. Als Würdigung seiner Leistungen für die amtliche Statistik habe ich dem Autor die Möglichkeit gegeben, die Ergebnisse seiner Forschungstätigkeit in dieser Schriftenreihe zu veröffentlichen.

Wiesbaden, im November 2010

Roderich Egeler

Präsident des Statistischen Bundesamtes

Inhalt

Abbildungsverzeichnis

Tabellenverzeichnis

Einleitung

„Δημόκριτος (καὶ) τὸν αέρα φησὶν εις ομοιοσχήμονα θρύπτεσθαι σώματα . . ."[1]

„νόμω γλυκύ, νόμω πικρόν, νόμω θερμόν, νόμω ψυχρόν, νόμω χροιή, ετεῇ δὲ άτομα καὶ κενόν."[2]

(Demokrit)

Empirisch arbeitende Wirtschafts- und Sozialwissenschaftler haben ein besonderes Interesse an einer ganz bestimmten Form von Informationen, den sogenannten Einzel- oder Mikrodaten. Die statistischen Ämter als größter Datenproduzent veröffentlichen traditionell nur spezielle auf den Nutzer zugeschnittene Auswertungen, Verdichtungen und Zusammenfassungen von Einzeldaten zu Gesamtgrößen, Durchschnitts- und Veränderungsmaßen. Die Nutzungsmöglichkeiten der in den Ämtern vorhandenen Daten gehen in der Regel aber weit über die traditionellen Veröffentlichungsangebote hinaus. Insbesondere der Fortschritt der Informationstechnik hat die Voraussetzungen für die Datennutzer geschaffen, sehr große Datenmengen verarbeiten und auswerten zu können.

Viele der für die Wirtschafts- und Sozialforschung interessanten Fragestellungen erfordern die gleichzeitige Berücksichtigung verschiedener Einzelangaben über Individuen oder Einheiten, beispielsweise zur Bildung komplexer Indikatoren. Durch die steigende Diversifikation in der Wirtschaft hat sich damit der praktische Nutzen von Makroanalysen verringert und man ist bei wirtschaftspolitischen Fragen immer stärker auf die Analyse der Entwicklungen und ihrer Ursachen auch im Mikrobereich angewiesen. Für Themenkreise wie die Untersuchung der relativen Effizienz von Firmen, Gründungsgeschehen und Arbeitsplatzdynamiken, Analysen von Marktstrategien, Wirkungsanalysen und wirtschaftspolitischen Maßnahmen, um

1 „Demokrit sagt (auch), daß die Luft in ähnlichgestaltete Körper zerfällt . . ." (Aetios über Demokrit)

2 „Allein durch menschliche Übereinkunft wird Begriffen wie süß, bitter, warm, kalt und Farbe Gültigkeit verliehen; in Wahrheit jedoch sind es Atome und das Leere." (Demokrit, frg. 9)

nur wenige zu nennen, versprechen Einzeldaten Erkenntnisgewinne. Solche Fragestellungen können sektoral, regional, temporal und größenklassenspezifisch untersucht werden.

Da die Vielzahl der in den statistischen Ämtern verfügbaren Daten mit Auskunftspflicht erhoben wurde und wird, finden spezielle Regelungen zur statistischen Geheimhaltung Anwendung. Die heute praktizierten Regelungen beruhen im Wesentlichen auf den Diskussionen um die geplante Volkszählung im Jahre 1983, dem vom Bundesverfassungsgericht formulierten „Recht auf informationelle Selbstbestimmung", durch welches die Begriffe Datenschutz und Statistikgeheimnis besondere Bedeutung erlangt haben, und der Anpassung des Rechts der amtlichen Statistik an die Anforderungen des Volkszählungsurteils durch das Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG) vom 22. Januar 1987.[3]

Zuvor konnten der empirisch arbeitenden Wissenschaft nur „absolut anonymisierte" Daten übermittelt werden. Eine mögliche Enthüllung weitergegebener Einzelangaben durch den (wissenschaftlich und nicht-wissenschaftlich motivierten) Datennutzer musste zweifelsfrei ausgeschlossen sein. „Einzelangaben, die so anonymisiert werden, dass sie Auskunftspflichtigen oder Betroffenen nicht mehr zuzuordnen sind, dürfen vom Statistischen Bundesamt und von den Statistischen Ämtern der Länder übermittelt werden" (Statistisches Bundesamt 1981, S. 404). Mit In-Kraft-Treten des BStatG von 1987 wurden der Wissenschaft neue Chancen eröffnet. Nunmehr steht es Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung offen, neben absolut anonymisierten Einzeldaten auch „faktisch anonymisierte" Einzeldaten, die nur mit einem „unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft" zugeordnet werden können, zu nutzen. Dieses Unverhältnismäßigkeitsgebot ist gemeint, wenn von faktischer anstelle absoluter Anonymität gesprochen wird. Da bei solchen Daten ein „Restrisiko" der Deanonymisierung erlaubt ist, wird der Wissenschaft damit gegenüber den übrigen Datennutzern ein privilegierter Zugang zu amtlichen Einzeldaten gewährt.

Die Einhaltung von Regeln zur statistischen Geheimhaltung hat neben den gesetzlichen Pflichten auch eine unmittelbare positive Auswirkung auf die Analysequalität der amtlichen Einzeldaten. Es ist davon auszugehen, dass der amtlichen Statistik – im Gegensatz zu den befragten Personen, Haushalten oder Unternehmen – nicht erst durch einen tatsächlichen Missbrauch der Daten ein Schaden entsteht. Bereits der Verdacht eines leichtfertigen Umgangs mit Einzeldaten seitens der statistischen Ämter kann die Auskunftsbereitschaft der Befragten senken. Ein erkennbar verantwortungsvoller Umgang mit der statistischen Geheimhaltung ist daher wichtig für die Qualität der Statistik insgesamt.

Um der Wissenschaft einen möglichst reibungslosen und raschen Zugang zu Einzeldaten aus den Wirtschaftsstatistiken zu ermöglichen, ist es nötig, automatisierte Methoden zu entwickeln, welche den Grad der Anonymität einer Datei angemessen schätzen. Manuelle Entscheidungsprozesse sind hierbei soweit wie möglich zu reduzieren, da diese in der Regel

3 Zuletzt geändert durch Artikel 2 des Gesetzes vom 9. Juni 2005 (BGBl. I S. 1534).

sehr zeitraubend und arbeitsintensiv sind. Solche Methoden sind bei der Generierung von auf den Nutzer zugeschnittenen sogenannten „On-Site" anonymisierten Daten für einen Gastwissenschaftler (wo kurzfristig eine Beurteilung der Anonymität der vorgesehenen Datei nötig ist), bei der Datenfernverarbeitung und bei sogenannten „Scientific-Use-Files" von großer Bedeutung. Scientific-Use-Files sind standardisierte Dateien, die für einen breiteren Nutzerkreis verwendbar sind. Diese werden einem Wissenschaftler an seinem eigenen Arbeitsplatz zur Verfügung gestellt. Da ein potentieller Datenangreifer vorhandene Zusatzkenntnisse über die Daten, die er z.B. aus kommerziellen Unternehmensdatenbanken gewinnen kann, DV-technisch zuspielen könnte, stellt die Anonymisierung solcher Daten eine besondere Herausforderung für die Datenanbieter dar. Um das Wissenschaftsprivileg des BStatG beim Datenzugang nicht unnötig einzuschränken, darf eine faktische Anonymisierung andererseits nur auf realistische Gefährdungsszenarien abstellen und nicht dazu führen, dass unnötig hohe Schutzmaßnahmen aufgebaut werden. Den Nutzern faktisch anonymisierter Daten wird von Seiten der statistischen Ämter generell keine Missbrauchsabsicht unterstellt.

Die Anonymisierung von Personen- und Haushaltsdaten ist bereits seit Anfang der 1990er Jahre geübte Praxis (siehe Müller et al. 1991). Im Unterschied hierzu werden bei Unternehmens- und Betriebsdaten besondere Eigenheiten beobachtet. Bei solchen Erhebungen liegen in der Regel wesentlich kleinere Grundgesamtheiten vor[4] und damit vergleichsweise hohe Stichprobenauswahlsätze, wobei es vorkommt, dass bestimmte Schichten voll erhoben werden (z.B. dünn besetzte Wirtschaftszweige und obere Beschäftigten- oder Umsatzgrößenklassen). Da es nur wenige Unternehmen ab einer gewissen Größe gibt, sind diese oftmals zur Teilnahme an mehreren Erhebungen verpflichtet. Nicht zuletzt die Schiefe der Verteilung der Überschneidungsmerkmale mit kommerziellen Datenquellen sorgt dafür, dass ein beachtlicher Teil der Merkmalsträger einzigartige Kombinationen in diesen Merkmalen aufweist und daher mit großer Wahrscheinlichkeit reidentifiziert werden kann. Hinzu kommt, dass einem potentiellen Datenangreifer mit den oben angesprochenen im Handel erhältlichen Unternehmensdatenbanken ein breites, detailliertes und gut aufbereitetes Zusatzwissen zur Verfügung steht. Außerdem kann der Nutzen aus der Kenntnis über Unternehmens- und Betriebsdaten wesentlich höher eingestuft werden als bei Personen- und Haushaltserhebungen (z.B. vertrauliche Informationen über konkurrierende Unternehmen).

Die genannten Schwierigkeiten bei der Anonymisierung wirtschaftsstatistischer Einzeldaten haben dazu geführt, dass in der Begründung zu §16 Abs. 6 des BStatG von 1987 aus einer Sitzung des Ausschusses für Wirtschaft des Deutschen Bundestages, Arbeitsgruppe „Statistik", zitiert wird: „Wirtschaftsstatistische Daten eignen sich zumindest generell nicht für eine Anonymisierung."[5] Der Gesetzgeber hat zudem nicht näher erläutert, wie die faktische

4 Die wohl bekannteste deutsche Bevölkerungsstichprobe, der Mikrozensus, kann bei einem Auswahlsatz von einem Prozent als sehr klein eingestuft werden. Dennoch gehen einzelne Personen in den etwa 800.000 Befragten unter.

5 Vgl. Protokoll der Sitzung des Ausschusses für Wirtschaft des Deutschen Bundestages, Arbeitsgruppe „Statistik", vom 17. September 1979, S.77.

Anonymität eines Datenbestandes festgestellt werden kann. Das Ziel der vorliegenden Arbeit besteht daher im Wesentlichen in einer Operationalisierung des Begriffes der faktischen Anonymität und der empirischen Umsetzung.

Im Vordergrund der Arbeit steht neben den Methoden zur Messung der Datensicherheit vor allem deren empirische Umsetzung. Lange Zeit mussten sich die Datenanbieter mit Recht den Vorwurf gefallen lassen, vor ihrer Weitergabe zu starke Veränderungen an den Daten vorzunehmen. Um datenschutzrechtlich auf der sicheren Seite zu sein, wurden die Daten so verändert, dass eindeutige Zuordnungen von theoretischer Seite nicht möglich waren (z.B. die Verhinderung eindeutiger Kombinationen kategorialer Merkmale, ganz unabhängig von der Qualität der gewonnenen Informationen).

Nicht zuletzt durch die im Rahmen der Machbarkeitsstudie KombiFiD untersuchten Verknüpfungsmöglichkeiten von Erhebungen verschiedener Datenanbieter in der Breite (Zuwachs an Merkmalen) und in der Länge (Zuwachs an Merkmalsträgern), d.h. durch die Konfrontation mit großen Datenmengen, und durch den Fortschritt der Rechentechnik werden mittel- bis langfristig Simulationsprogramme gebraucht, die zum einen die komplexe Struktur der Daten berücksichtigen und zum anderen sehr effizient arbeiten.

Kapitel 1

Anonymisierungsmethoden

Im Allgemeinen kann eine Anonymisierungsmaßnahme oder ein Bündel solcher Maßnahmen als Abbildung f verstanden werden, welche die Originaldaten $\mathcal{O}$ auf anonymisierte, d.h. geeignet modifizierte Daten $\mathcal{A}$ abbildet:

$$\mathcal{A} = f(\mathcal{O}).$$

Die Abbildung f kann zum einen deterministische Elemente wie z.B. Vergröberung von Merkmalen (siehe Unterabschnitt 1.1.2) oder deterministische Mikroaggregation (siehe Unterabschnitt 1.2.2) enthalten. Zum anderen sind auch stochastische Methoden wie die Ziehung einer Zufallsstichprobe (siehe Unterabschnitt 1.1.1), die Methode der Post-Randomisierung (siehe Unterabschnitt 1.2.1) oder Varianten der additiven und multiplikativen Zufallsüberlagerung (siehe Unterabschnitt 1.2.2) auf die Daten anwendbar. Wenn aus Geheimhaltungssicht nichts dagegen spricht, dann ist die Abbildung f bzw. der Katalog an Anonymisierungsmaßnahmen an den Datennutzer weiterzugeben, damit dieser seine Analysemodelle der gegebenen Datengrundlage anpassen kann. In diesem Falle darf die Abbildung f allerdings nicht umkehrbar sein. Diese Mindestanforderung reicht jedoch im Allgemeinen nicht aus, da für einen erfolgreichen Reidentifikationsversuch die Anonymisierung nicht eins zu eins rückgängig gemacht werden muss. Einem Datenangreifer kann bereits die Information nützlich sein, dass die zur Reidentifikation verwendeten Merkmale nur schwach verfremdet wurden und/oder ein gesuchter Einzelwert mit großer Wahrscheinlichkeit in einem bekannten reellen Intervall liegt.

Die in diesem Kapitel aufgeführten Methoden der Anonymisierung wurden im Wesentlichen im Forschungsprojekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten *(FAWE)*" (siehe Lenz et al. 2006b) unter der wissenschaftlichen Leitung von Prof. Dr. Gerd Ronning zusammengestellt und zum Teil neu erarbeitet. Eine ausführliche Darstellung der für Querschnittsdaten empfohlenen Methoden findet sich in dem durch das Projektteam veröffentlichten Handbuch der Anonymisierung wirtschaftsstatistischer Einzeldaten (siehe Ronning et al. 2005). Einige der vorgestellten datenverändernden Methoden wurden zum Zwecke der Anwendung auf Längsschnittsdaten im Nachfolgeprojekt „Wirtschaftsstati-

stische Paneldaten und faktische Anonymisierung *(FAWE-Panel)*" weiterentwickelt (siehe hierzu Ronning et al. 2009). In Höhne (2008) werden die Möglichkeiten der Weiterentwicklung dieser Methoden zur Anonymisierung von Paneldaten diskutiert. Elaboriertere Methoden der Anonymisierung, insbesondere der datenverändernden Verfahren zur stochastischen Überlagerung, finden sich ebenfalls in Höhne (2008) und ausführlich in Höhne (2010). Darin wird grundsätzlich zwischen „Traditionellen" bzw. „Informationsreduzierenden Methoden" (*Verschweigen*) und „Datenverändernden Methoden" (*Notlüge*) unterschieden. Der wesentliche Unterschied besteht darin, dass durch Anwendung der informationsreduzierenden Methoden zwar die Analysemöglichkeiten eingeschränkt werden und im schlimmsten Falle sogar Analysen nicht mehr durchführbar sind, auf der anderen Seite aber die Ergebnisse durchführbarer Analysen keine Unterschiede zu den entsprechenden Analyseergebnissen mit den Originaldaten aufweisen. Bei den datenverändernden Methoden sind infolge der Veränderung der Einzelwerte zumindest kleine Abweichungen in den Ergebnissen mit Originaldaten und anonymisierten Daten nicht zu verhindern.

Abbildung 1.1 gibt einen unvollständigen Überblick über gängige Anonymisierungsmethoden (Ronning et al. 2005). Auf diese Methoden wird in den nachfolgenden Abschnitten eingegangen.

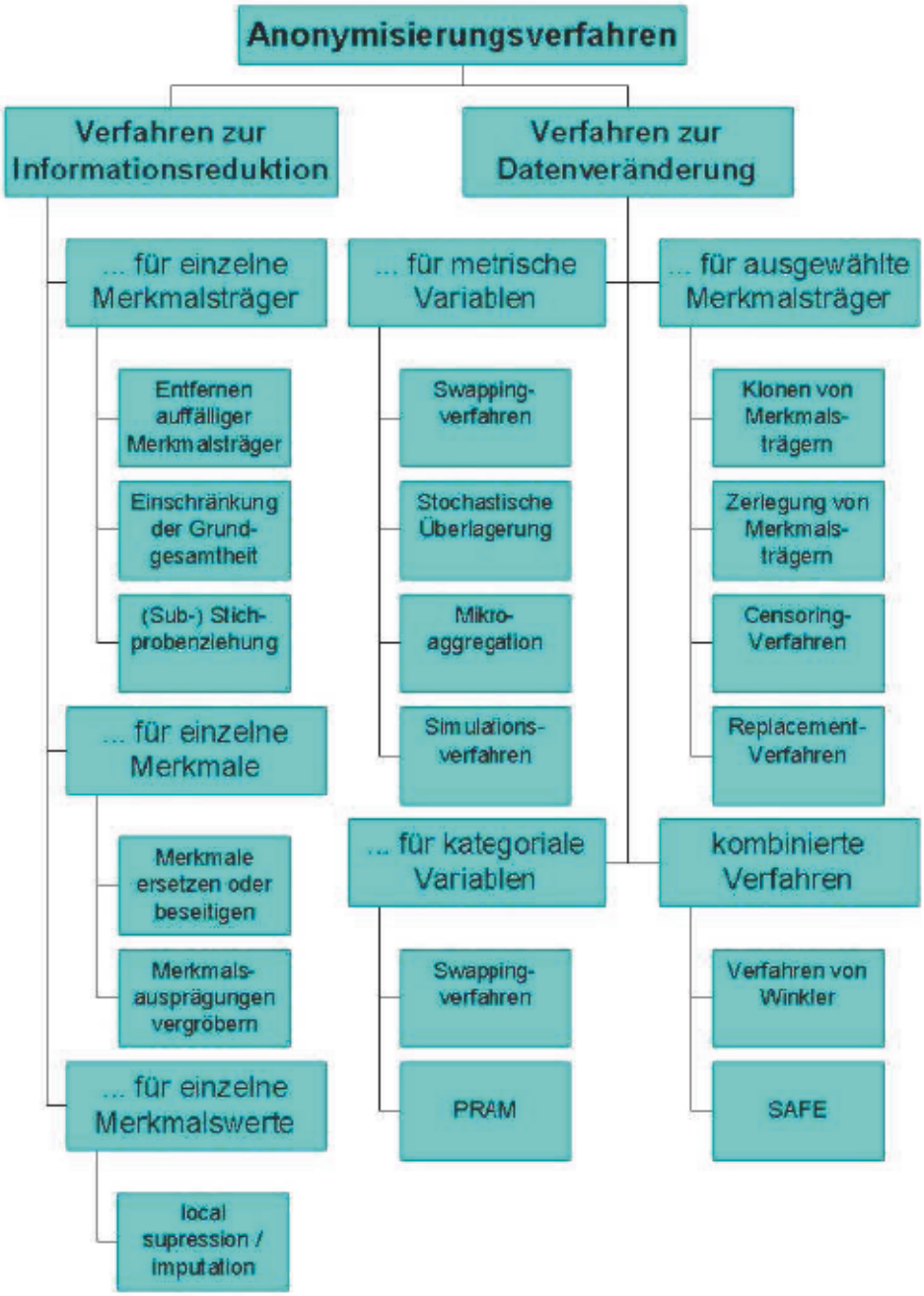
1.1 Informationsreduzierende Methoden

Diese Methoden wurden und 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 und Statistische Ämter des Bundes und der Länder 2003). Informationsreduktionen werden in der Regel realisiert, indem Informationen unterdrückt oder vergröbert werden. Sie können grundsätzlich an einzelnen oder Gruppen von Merkmalsträgern (Unterabschnitt 1.1.1), an einzelnen oder mehreren Merkmalen (Unterabschnitt 1.1.2) oder an einzelnen Ausprägungen (Unterabschnitt 1.1.3) ansetzen.

1.1.1 Merkmalsträgerbezogene Methoden zur Informationsreduktion

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

Abbildung 1.1
Übersicht der Anonymisierungsverfahren



Entfernen auffälliger Merkmalsträger: Ausreißer, d.h. besonders auffällige und daher reidentifikationsgefährdete Merkmalsträger werden entfernt. Dies sind Merkmalsträger, die einzigartige oder seltene Merkmalskombinationen aufweisen. 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.

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

Systematische Einschränkung der Grundgesamtheit: Beispielsweise werden alle publizitätspflichtigen Unternehmen, eine komplette Branche oder die Großunternehmen ab einer bestimmten Beschäftigtenzahl oder einem bestimmten Gesamtumsatz aus dem Datenbestand entfernt. Sofern dem Nutzer die Verkleinerung der Grundgesamtheit bekannt ist und die entfernten Teilgesamtheiten keinen wichtigen Beitrag zur empirischen Beurteilung der sozio-ökonomischen Fragestellungen liefern, bestehen für die Analyse der Restgesamtheit keinerlei Probleme. Das Verfahren wird bei Personendaten zum Beispiel zum Schutz von Abgeordneten eingesetzt. Wichtiger erscheint die Anwendung des Verfahrens bei Wirtschaftsdaten, weil insbesondere Großbetriebe und Großunternehmen einem besonders hohen Reidentifikationsrisiko unterliegen.

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

Ziehung einer Zufallsstichprobe: Durch die Ziehung einer Zufallsstichprobe wird die Wahrscheinlichkeit der Teilnahme eines Merkmalsträgers an der betrachteten Erhebung verringert.

Diese Stichprobe kann auch mit verschiedenen Auswahlwahrscheinlichkeiten (beispielsweise geschichtet nach Beschäftigtengrößenklassen und Wirtschaftszweigen) gezogen werden. 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 mithilfe 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 Wahrschein-

lichkeit sich der gesuchte Merkmalsträger in der Stichprobe befindet. Diese Unsicherheit wird im Falle einer Sub-Stichprobenziehung, wenn die Daten nach einem bestimmten Stichprobenauswahlsatz oder in einem eingeschränkten Berichtskreis erhoben wurden und damit bereits im Original als Stichprobe vorliegen, noch einmal verstärkt.

Stichprobenziehungen haben sich für die Anonymisierung von Personen- und Haushaltsdaten bewährt (z.B. Mikrozensus). Ihre Anwendbarkeit für die Anonymisierung von Unternehmensdaten muss als deutlich geringer eingeschätzt werden, weil bei Unternehmensdaten die Grundgesamtheiten kleiner und die Stichprobenauswahlsätze größer sind. Man beachte, dass in manchen Erhebungen, wie zum Beispiel der Kostenstrukturerhebung im Verarbeitenden Gewerbe, ab einer bestimmten Beschäftigtenanzahl alle Unternehmen einbezogen werden.

1.1.2 Merkmalsbezogene Methoden zur Informationsreduktion

Die merkmalsbezogenen Anonymisierungsmaßnahmen behandeln im Gegensatz zu merkmalsträgerbezogenen Maßnahmen (anzuwenden auf die Zeilen der Datenmatrix) einzelne oder mehrere Merkmale (anzuwenden auf die Spalten der Datenmatrix). Sie werden in der Regel bei den sogenannten Überschneidungsmerkmalen[6] eingesetzt, um eine korrekte Zuordnung zu verhindern, oder bei besonders sensiblen Merkmalen, um die originalen Einzelwerte vor Enthüllung zu bewahren. Dabei können Merkmale ersetzt oder ihre Ausprägungen zu Kategorien vergröbert werden.

Ersetzung von Merkmalen: Die Merkmale werden durch adäquate Linearkombinationen, Kennziffern oder Indizes ersetzt. Für die Ersetzung von Merkmalen bestehen die folgenden Möglichkeiten:

- Konstruktion von neuen Merkmalen aus mehreren ursprünglichen Merkmalen beispielsweise durch die Bildung von Linearkombinationen (z.B. Bildung der Summe aus Inlands- und Auslandsumsatz)
- Bildung von statistisch interpretierbaren Beziehungs- und 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. Gesamtumsatz des Jahres 1999 bezogen auf den Gesamtumsatz des Jahres 1980)

6 Zur Definition des Begriffes Überschneidungsmerkmal siehe Unterabschnitt 2.4.3.

Während die vollständige oder partielle Merkmalsunterdrückung gleichermaßen für metrische wie kategoriale Merkmale anwendbar ist, lässt sich die Ersetzung von Merkmalen durch Linearkombinationen, Beziehungs- und Verhältniszahlen sowie Indizes nur für metrische Merkmale realisieren.

Die Schutzwirkung dieser Verfahren beruht allein auf der Verringerung der Informationen im Datensatz. Werden durch die Anwendung dieser Verfahren Überschneidungsmerkmale entfernt, so sinkt auch die Zuordnungswahrscheinlichkeit. Werden hingegen sensible Merkmale aus dem Datensatz entfernt, so werden die Anreize verringert, eine Enthüllung vorzunehmen, da der Nutzen einer Reidentifikation für den potentiellen Angreifer sinkt.

Vergröberung von Merkmalsausprägungen: Bei der Vergröberung von Merkmalsausprägungen existieren in Abhängigkeit von der Skalierung der Wertebereiche der Merkmale unterschiedliche Ansätze:

- Gruppierung von metrischen Merkmalen zu Kategorien (z.B. Bildung von Beschäftigtengrößenklassen oder Umsatzgrößenklassen)
- Rundung der Werte metrischer Merkmale (z.B. Rundung von Umsatzangaben auf ganze Tausenderbeträge)
- Weitere Zusammenfassung bereits existierender Kategorien (z.B. Vereinigung benachbarter Beschäftigtengrößenklassen oder Wirtschaftszweige)

Durch die Anwendung dieser Methoden reduziert sich die Anzahl der möglichen Kombinationen von Ausprägungen der Überschneidungsmerkmale. Bei einer gleichbleibenden Anzahl von Merkmalsträgern führt dies dazu, dass die Wahrscheinlichkeit des Auftretens einzigartiger Ausprägungskombinationen sinkt und zugleich das Auftreten identischer Kombinationen zunimmt. Für einen Datenangreifer entsteht demnach eine erhöhte Unsicherheit, da die Wahrscheinlichkeit von Falschzuordnungen steigt. Außerdem sinkt der Nutzen durch eine Enthüllung, weil mit der Vergröberung ein Informationsverlust verbunden ist. Bedauerlicherweise ist dieser Informationsverlust auch für den wissenschaftlichen Datennutzer zu verzeichnen. Insbesondere die Umwandlung metrischer in kategoriale Merkmale kann komplette Analysen ausschließen, vor allem wenn zur Modellspezifikation das Merkmal in metrischer Form vorliegen muss. Zudem können auch die deskriptiven Statistiken des Datensatzes erheblich beeinflusst werden. Der Informationsverlust hängt dabei entscheidend vom Grad der Vergröberung ab.

In dem aus Forschersicht ungünstigsten Spezialfall der Vergröberung eines Merkmals werden alle Ausprägungen des Merkmals zu einer Kategorie zusammengefasst, was einer Entfernung des Merkmals aus der Datei gleich kommt.

Hat sich im Laufe des diskursiven Prozesses zur Erzeugung faktisch anonymen Datenmaterials zwischen Datenanbieter und Datennutzer in beidseitigem Einvernehmen herausgestellt, dass ein bestimmtes Merkmal wenig Relevanz für die wissenschaftlichen Fragestellungen des Datennutzers besitzt, so kann dieses Merkmal zugunsten eines besseren Erhaltes anderer interessierender Merkmale entfernt werden.

Liegt ein kategoriales, hierarchisches Merkmal vor, so ergeben sich die Vergröberungsmöglichkeiten unmittelbar aus der Struktur des hierarchisch partiell geordneten Wertebereiches. Man betrachte z.B. das in Wirtschaftsstatistiken oftmals enthaltene Merkmal *Wirtschaftszweigklassifikation*, das fünf Gliederungsebenen besitzt. Eine Vergröberung wird hier sehr einfach durch die Entfernung einer oder mehrerer Einheiten des fünfstelligen Zahlencodes erreicht. Abhängig von den Besetzungszahlen einzelner Branchen ist es durchaus empfehlenswert, verschiedene Gliederungstiefen zu verwenden (vgl. hierzu Abschnitt 4.2).

Zu den Verfahren der Merkmalsvergröberung sind auch die sogenannten „Abschneideverfahren" zu zählen. Diese können auf ordinale Merkmale (d.h., metrische oder kategoriale Merkmale mit linear geordnetem Wertebereich) angewendet werden. Nach Festlegung einer oberen (unteren) Schwelle werden alle Ausprägungen oberhalb (unterhalb) dieser Schwelle zu einer Kategorie zusammengefasst und jeweils durch einen vorgegebenen gemeinsamen Wert ersetzt. Bei der Anonymisierung von Wirtschaftsstatistiken eignet sich dieses Verfahren vor allem zum Schutz großer Unternehmen, die besonders reidentifikationsgefährdet aber für Analysen in der Regel unentbehrlich sind. Aus diesem Grunde wird das Verfahren (engl. top/bottom-coding bzw. replacement) in Abbildung 1.1 unter den Verfahren zur Veränderung ausgewählter Merkmalsträger geführt.

Beispielsweise wurden bei der Anonymisierung der amtlichen Umsatzsteuerstatistik und der Gehalts- und Lohnstrukturerhebung Abschneideverfahren auf spezielle Merkmale über Umsatz und Beschäftigtenzahl der einbezogenen Unternehmen angewendet. Bei den meisten Unternehmens- und Betriebserhebungen wird eine untere Abschneidegrenze bereits im Stichprobenplan festgelegt. So wird beispielsweise in der Kostenstrukturerhebung im Verarbeitenden Gewerbe eine Stichprobe aus den Unternehmen mit wenigstens 20 Beschäftigten gezogen. In der Umsatzsteuerstatistik werden alle Unternehmen ab einem Mindestjahresumsatz von 16 617 Euro erfasst.

1.1.3 Ausprägungsbezogene Methoden zur Informationsreduktion

Bei dem ausprägungsbezogenen Vorgehen zur Informationsreduktion handelt es sich in der Regel um die **Unterdrückung** einzelner Werte (engl. local suppression). Dies geschieht

meist bei Beobachtungen mit Ausprägungen oder Ausprägungskombinationen, die in der Stichprobe sehr selten oder einzigartig sind. Durch die Unterdrückung entstehen fehlende Werte. Damit ist keine Neukodierung des gesamten Merkmals erforderlich, vielmehr bleibt die Merkmalsdefinition des Ausgangsdatensatzes erhalten. Dies kann insbesondere dann sinnvoll sein, wenn Probleme bei der Anonymisierung von Werten für einzelne Großbetriebe und Großunternehmen bestehen. Merkmalswerte können sowohl bei metrischen als auch bei kategorialen Merkmalen unterdrückt werden.

Die Schutzfunktion dieses Verfahrens besteht zum einen in der direkten Verminderung der Anzahl möglicher Kombinationen zwischen den Ausprägungen der Überschneidungsmerkmale, insbesondere bei diskreten bzw. kategorialen Merkmalen. Durch die Unterdrückung sind vorher seltene oder einmalige Kombinationen nicht mehr aufdeckbar. Zum anderen können sensible Informationen für einzelne Beobachtungen unterdrückt werden, sofern es sich um seltene Ausprägungen handelt (Statistische Ämter des Bundes und der Länder 2003).

Eine Alternative zur Unterdrückung einzelner Werte stellt neben der Anwendung der im nächsten Abschnitt diskutierten datenverändernden Verfahren die vor allem bei kategorialen Merkmalen sinnvolle **Pseudonymisierung** dar. Durch eine Um- bzw. Neubenennung der Kategorien können einzelne Merkmalsträger einerseits nicht mehr der originalen Kategorie zugeordnet werden, andererseits sind Modelle, in denen dieses Merkmal als erklärendes Merkmal auftritt, möglicherweise weiterhin betrachtbar.

1.2 Datenverändernde Methoden

In diesem Abschnitt werden oft berücksichtigte datenverändernde Methoden für kategoriale und metrische Merkmale besprochen. Da solche Methoden im Allgemeinen auf wenig Akzeptanz von Nutzerseite stoßen, werden sie von den Datenanbietern mit Bedacht und viel Fingerspitzengefühl eingesetzt.

1.2.1 Datenverändernde Methoden für kategoriale Merkmale

1.2.1.1 Vertauschungsverfahren für kategoriale Merkmale

Die Vertauschungsverfahren (engl. Data Swapping) werden in der Regel auf einzelne Merkmale angewendet. Sie basieren auf einer systematischen oder zufälligen Vertauschung von Einzelwerten verschiedener Merkmalsträger und sind sowohl für metrische als auch für kategoriale Merkmale durchführbar (siehe daher auch 1.2.2). Die Vertauschungen der Einzelwerte zwischen den Merkmalsträgern erfolgen für alle zu anonymisierenden Merkmale getrennt. Dabei werden bestimmte Bereiche (z.B. benachbarte Regionstypen bei einer

nicht-administrativen Regionalangabe[7]) festgelegt, innerhalb derer Vertauschungen zulässig sind. Außerdem können Vertauschungen auch so ausgestaltet werden, dass sie nur innerhalb bestimmter Ausprägungskombinationen bezüglich nicht veränderter kategorialer Merkmale (z.B. Wirtschaftsbereich/Beschäftigtengrößenklasse) vorgenommen werden. Dadurch wird für den potentiellen Datenangreifer sowohl eine Reidentifikation von Merkmalsträgern erschwert als auch die hierbei enthüllte Information unbrauchbar.

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

1.2.1.2 Post-Randomisierung

Beim Verfahren der Post-Randomisierung (kurz PRAM) werden diskrete Merkmale durch die Definition von Übergangswahrscheinlichkeiten randomisiert (Kooiman et al. 1997; Willenborg und de Waal 2001). Dabei werden die Merkmalswerte mit bei der Anwendung festzulegenden Übergangswahrscheinlichkeiten in andere Ausprägungen transformiert.[8] Für den Fall eines dichotomen Merkmals ist die Matrix der Übergangswahrscheinlichkeiten in Gleichung (1.1) dargestellt.

$$P = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix} \tag{1.1}$$

Im allgemeinen Falle kann diese Matrix beliebige Struktur aufweisen, mit der Einschränkung, dass sämtliche Einträge nicht-negativ sein und sich die Zeilenwerte zu eins aufsummieren müssen.

Soll das Verfahren auf hierarchische Merkmale derart angewendet werden, dass innerhalb bestehender Kategorien auf tieferer Ebene Unterkategorien transformiert werden, dann entsteht eine Blockmatrix. Werden beispielsweise die Dreisteller der Wirtschaftszweigklassifikation unter Beibehaltung der Zweistellerebene verändert, so ist jedem Zweisteller ein Block in der Matrix zugewiesen.

7 Zur Definition der nicht-administrativen Gebietsschlüssel siehe unter http://www.bbr.bund.de auf der Internetseite des Bundesamtes für Bauwesen und Raumordnung.

8 Hierbei ist es durchaus zulässig, wenn auch unüblich, die fehlenden Werte zu dem Wertebereich des Merkmals zu zählen und damit einzelne Originalwerte in fehlende Werte zu transformieren.

Bei der Anwendung des Verfahrens auf ordinale Merkmale empfiehlt sich die Implementierung mittels einer n-diagonalen Matrix. Man betrachte zum Beispiel den nichtadministrativen Regionstyp BBR9 (sogenannter Neunerschlüssel) mit den Kategorien $1, \ldots, 9$, wobei ein Einzelwert mit vorgegebenen Wahrscheinlichkeiten unverändert bleibe oder durch eine der beiden Nachbarausprägungen ersetzt werde. Die resultierende Matrix ist dann nur auf der Hauptdiagonalen und den beiden Nebendiagonalen besetzt.

PRAM führt dazu, dass die veröffentlichten Werte nur noch mit einer durch das Verfahren festgelegten Wahrscheinlichkeit den Werten im Originaldatensatz entsprechen. Da kategoriale Merkmale häufig als Zusatzwissen eines potentiellen Datenangreifers auftreten, senkt das Verfahren vorrangig die Zuordnungswahrscheinlichkeit bei Datenangriffen. Auffällig ist, dass das Verfahren der Post-Randomisierung die einzige Anonymisierungsmethode für diskrete bzw. kategoriale Merkmale darstellt, der ein stochastisches Modell zugrunde liegt.

1.2.2 Datenverändernde Methoden für metrische Merkmale

1.2.2.1 Vertauschungsverfahren für metrische Merkmale

Wie bereits in Unterabschnitt 1.2.1 erwähnt, können die Vertauschungsverfahren grundsätzlich gleichermaßen für kategoriale wie für metrische Merkmale verwendet werden. Auch bei den metrischen Merkmalen können die Vertauschungen so beschränkt werden, dass sie nur innerhalb bestimmter Ausprägungskombinationen für unveränderte kategoriale Merkmale vorgenommen werden. Sortiert man die Merkmalswerte für jedes einzelne zu anonymisierende metrische Merkmal auf- oder absteigend und definiert danach Nachbarschaftsbereiche, auf die der Tausch beschränkt wird (z.B. Tausch innerhalb einer Beschäftigtengrößenklasse oder beschränkt auf benachbarte Einzelwerte), so kann der annähernde Erhalt der Rangstatistiken garantiert werden. Diese Verfahrensvariante wird daher im englischen Sprachraum mit „Rank Swapping" bezeichnet.

Werden metrische Überschneidungsmerkmale, also im Zusatzwissen eines potentiellen Datenangreifers vorhandene Merkmale, mit Vertauschungsverfahren behandelt, so reduziert sich in erster Linie das Risiko der korrekten Zuordnung. Werden auch andere sensible metrische Merkmale mit diesen Verfahren anonymisiert, so geht damit auch die Reduzierung des Nutzens enthüllter Einzelwerte für den Datenangreifer einher.

Die univariaten Verteilungen werden durch das Verfahren erfreulicherweise erhalten, während die multivariate Verteilung deutlich verändert wird. Offenbar ist die zu erwartende Veränderung dabei umso stärker, je größer die Bandbreite ist, innerhalb derer die Vertauschung stattfindet. Eine Formalisierung findet sich in Rosemann (2003).

1.2.2.2 Imputationsverfahren

Imputationsverfahren bestehen in einer Ersetzung von Angaben durch eingeschätzte Werte. Diese Idee wurde zuerst von Rubin (1993) vorgeschlagen und baut auf den Imputationsverfahren im Falle von fehlenden Antworten auf, die im Rahmen der „Nonresponse"-Forschung entwickelt wurden (Fienberg 1997). Im Unterschied zur klassischen Anwendung der Imputationsmethoden werden hier nicht fehlende Angaben, sondern besonders sensible Merkmalswerte und/oder Merkmalswerte von Überschneidungsmerkmalen durch neu einzuschätzende Werte ersetzt. Dabei können einzelne Merkmalswerte, die Merkmalswerte besonders gefährdeter Merkmalsträger oder alle Merkmalswerte eines Merkmals imputiert werden.

Für die Imputation stehen parametrische und nicht-parametrische Regressionsmodelle zur Verfügung (Pollettini et al. 2002). Grundsätzlich wird zwischen einfacher Imputation und multipler 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 Falle werden den Datennutzern mehrere anonymisierte Dateien bereitgestellt. Dies kann sich allerdings negativ auf das Reidentifikationsrisiko auswirken, da einem Datenangreifer durch ein Zusammenspielen dieser Dateien zuverlässigere Werte für die Überschneidungsmerkmale zur Verfügung stehen.

1.2.2.3 Stochastische Überlagerung

Stochastische Überlagerungen beziehungsweise Überlagerungen mit Zufallsfehlern stellen eine umfangreiche Verfahrensgruppe zur Anonymisierung von Einzeldaten dar. Die Grundidee besteht darin, dass zu den quantitativen Merkmalen eines Datensatzes Zufallszahlen addiert oder die Merkmalswerte mit Zufallszahlen multipliziert werden.

Die stochastischen Überlagerungsverfahren unterteilen sich grundsätzlich in additive und multiplikative Verfahren. 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. Selbstverständlich muss dann ein Erwartungswert von Eins gewählt werden. Alternativ kann der multiplikative Zufallsfehler auch gleichverteilt sein, einer Lognormalverteilung oder einer gestutzten Normalverteilung entstammen.

Bei einer stochastischen Überlagerung entsprechen die veröffentlichten Werte mit der Wahrscheinlichkeit Null den Originalwerten.[9] Für jeden Originalwert können jedoch Intervalle um den überlagerten Wert ermittelt werden, die diesen mit vorgegebener Wahrscheinlichkeit enthalten.

a) Additive stochastische Überlagerung

a1) Additive Überlagerung mit einer Normalverteilung

Bei der 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 X wie folgt darstellen (vgl. u.a. Höhne 2004a):

$$X^a = X + W \tag{1.2}$$

mit

$$W \sim N(0, \Sigma_{ww}), \tag{1.3}$$

wobei X^a und W dieselbe Dimension wie X aufweisen.

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 bei 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 schlägt Roque (2002) vor, 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.

9 Aufgrund der Atomlosigkeit stetiger Verteilungen können die Erwartungswerte Null (bei additiver Überlagerung) und Eins (bei multiplikativer Überlagerung) in Simulationen nicht beobachtet werden.

Es seien V_1 und V_2 zwei stetige Zufallsvariablen mit Dichtefunktionen $f_1(v_1)$ und $f_2(v_2)$ 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

$$g(w) = \alpha f_1(w) + (1 - \alpha) f_2(w) \tag{1.4}$$

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

Damit kann man sich eine Mischungsverteilung gedanklich auch wie folgt vorstellen: „Es gibt zwei (bzw. allgemeiner k) Zustände, die mit Wahrscheinlichkeit α und $1 - \alpha$ auftreten und 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 2004a).

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 2003; Statistische Ämter des Bundes und der Länder 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 Originalmerkmalen erzeugt.

Für die Störvariable gilt damit:

$$W = f(X). \tag{1.5}$$

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

$$W = VX \tag{1.6}$$

mit

$$E[V] = 0. \tag{1.7}$$

Daraus ergibt sich aber

$$X^a = X + W = X + VX = (1 + V)X = W^*X \tag{1.8}$$

mit

$$E[W^*] = 1. \tag{1.9}$$

Damit ist die heteroskedastische additive Überlagerung in dieser Form identisch mit der multiplikativen Überlagerung (Höhne 2004a). Allerdings gilt dies nur für den Fall des in Gleichung 1.6 unterstellten linearen Zusammenhanges zwischen der Varianz der Störgröße und den Originalwerten.

b) Multiplikative stochastische Überlagerung

Wie bereits in den vorangegangenen Unterabschnitten 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 2004b):

$$X^a = WX. \tag{1.10}$$

Dabei ist W eine stetige Zufallsvariable mit Erwartungswert 1 und Varianz $\sigma_u^2 > 0$.

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. Die erste Vorgehensweise hat aus Nutzersicht den Vorteil, dass die relativen Beziehungen zwischen den Merkmalen eines Merkmalsträgers erhalten bleiben. Dies kann allerdings auch zu einem höheren Reidentifikationsrisiko führen, da Quotienten von Überschneidungsmerkmalen unverändert vorliegen.

Wie bereits erwähnt wurde, sollte es sich bei der Zufallsvariable W in Gleichung (1.10) um eine positive Variable handeln, damit die Vorzeichen der Merkmalsträger nicht systematisch

verändert werden. Aus diesem Grunde werden bei multiplikativen Überlagerungen für die Störgrößen in der Regel Verteilungen verwendet, die lediglich positive Ausprägungen der Merkmalswerte zulassen. Gottschalk (2004) verwendet eine Gleichverteilung über dem Intervall $[0,5; 1,5]$. Das größte denkbare Intervall bei einer Gleichverteilung ist durch $[0; 2]$ gegeben, weil nur so der Wertebereich von W positiv ist und der Erwartungswert Eins beträgt. Um lediglich positive Werte für die Zufallsfehler zu erhalten, kann auch, wie von Kim und Winkler (2001) vorgeschlagen, bei der direkten multiplikativen Überlagerung der Originalwerte eine gestutzte Normalverteilung verwendet werden. Um die Symmetrie der Normalverteilung zu erhalten, schlägt Höhne (2004a) vor, die Varianzen der Zufallsfehler so gering zu wählen, dass negative Werte auch bei einer einfachen Normalverteilung nur mit sehr kleiner Wahrscheinlichkeit auftreten. Treten sie in der Praxis dennoch auf, werden die Zufallsfehler erneut gezogen.

Eine spezielle Form einer unechten Mischungsverteilung wird ebenfalls von Höhne (2004a) vorgeschlagen: Bei diesem in Kennerkreisen auch als Höhne-Verfahren bezeichneten Vorgehen wird zunächst mit Wahrscheinlichkeit von 0,5 entschieden, ob die Merkmalswerte eines Merkmalsträgers verkleinert oder vergrößert werden. Hierzu werden die Grundüberlagerungsfaktoren $1 - f$ und $1 + f$ verwendet. Jedem Merkmalsträger wird ein Grundüberlagerungsfaktor zugewiesen. Diese Faktoren werden anschließend für jeden Merkmalswert unabhängig additiv mit einer Normalverteilung mit Erwartungswert Null und Standardabweichung s (mit $s << f/2$) überlagert. Somit wird jeder Merkmalsträger zwar in die gleiche Richtung verzerrt, es erhält jedoch jeder einzelne Merkmalswert einen eigenen Überlagerungsfaktor.

Insbesondere bei sehr schief verteilten Originalvariablen hängt die Stärke der Abweichungen durch Überlagerung von der Konstellation der Zufallszahlen bei wenigen großen Merkmalsträgern ab. So kann es passieren, dass trotz der asymptotischen Erwartungstreue und qualitativ hochwertig generierten Zufallszahlen Mittelwerte und Summen nur sehr schlecht reproduziert werden. Höhne (2004a) entwickelte deshalb verschiedene Algorithmen für sogenannte „kontrollierte" multiplikative Überlagerungen. Spezielle Varianten werden in Kapitel 7 auf die Kostenstrukturerhebung im Verarbeitenden Gewerbe angewendet.

1.2.2.4 Mikroaggregationsverfahren

Die Grundidee von Mikroaggregationsverfahren besteht darin, „ähnliche Merkmalsträger" zu Gruppen zusammenzufassen und die Ursprungswerte durch die arithmetischen Mittel der Merkmalswerte aller Merkmalsträger innerhalb einer Gruppe zu ersetzen (Mateo-Sanz und Domingo-Ferrer 1998a; Höhne 2003a; Ronning et al. 2005). Die in der Literatur diskutierten Varianten der Mikroaggregation unterscheiden sich im Wesentlichen in der Definition des verwendeten Ähnlichkeitsbegriffes.

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):

a) Die deterministische bzw. abstandsorientierte Mikroaggregation, bei der möglichst ähnliche Einheiten zusammengefasst werden.

b) Die stochastische Mikroaggregation, bei der die Gruppierung der Einheiten rein zufällig erfolgt.

Zudem erfolgt eine Unterscheidung danach, ob die Mikroaggregation für alle Merkmale gemeinsam erfolgt – für die Durchschnittsbildung bei den verschiedenen Merkmalen folglich die gleichen Gruppen gebildet werden – oder die Gruppenbildung für jedes Merkmal getrennt erfolgt.

a) Deterministische 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. Zum anderen unterscheiden sich die Mikroaggregationsverfahren hinsichtlich der Bestimmung des Abstandes zwischen den einzelnen Objekten. Hierzu können ein (eindimensionale Mikroaggregation) oder mehrere (mehrdimensionale Mikroaggregation) Merkmal(e) herangezogen werden. Die einzelnen Verfahrensvarianten unterscheiden sich zudem danach, ob die Gruppen für alle metrischen Merkmale – oder auch Gruppen von Merkmalen – gemeinsam erfolgt oder die Merkmale getrennt mikroaggregiert werden.

Wird ein für n Merkmalsträger beobachtetes Merkmal mikroaggregiert, so gilt für den anonymisierten Merkmalsvektor

$$x^a = Dx. \tag{1.11}$$

Dabei hat die $n \times n$-Matrix D eine blockdiagonale Gestalt, da die Merkmalsträger vor der Gruppierung und anschließenden Anwendung einer Aggregationstechnik sortiert werden. Die Blöcke sind mit $\frac{1}{k}$, der restliche Teil der Matrix mit Nullen besetzt.

a1) Varianten der eindimensionalen Mikroaggregation

- **Sortierung und Gruppenbildung nach einem Merkmal**: Es wird ein dominierendes Merkmal herausgesucht und der Datenbestand danach sortiert. Danach werden absteigend immer k (oftmals wird $k = 3$ gesetzt) benachbarte Merkmalsträger in einer Gruppe zusammengefasst und alle ihre stetigen Merkmalswerte durch den Durchschnitt der Werte ersetzt. (Das dominierende Merkmal sollte dabei aus Nutzersicht mit möglichst vielen weiteren Merkmalen hoch korreliert sein.)

- **Sortierung und Gruppenbildung für jedes Merkmal getrennt**: Der Datenbestand wird jeweils nach dem zu anonymisierenden Merkmal 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 Merkmale wiederholt.

a2) Varianten der mehrdimensionalen Mikroaggregation

- **Sortierung und Gruppenbildung nach einem Hilfsmerkmal**: Die Sortierung erfolgt anhand eines Hilfsmerkmals. Als Hilfsmerkmale sind zum Beispiel die Hauptkomponente (als ein durch Transformation gebildetes Merkmal mit möglichst hoher Korrelation zu den anderen Merkmalen) oder die sogenannten „Z-Scores" (als Summe der standardisierten Originalmerkmale) denkbar.

- **Distanzbildung zwischen den Merkmalsträgern:** 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). Ein neuer Ansatz geht auf Domingo-Ferrer et al. (2008) zurück. Hier wird das Kriterium zur Gruppenbildung graphentheoretisch festgelegt. Eine erste Anwendung des Verfahrens auf Paneldaten der amtlichen Statistik findet sich in Abschnitt 7.4.

a3) Mikroaggregation für Gruppen von Merkmalen

Bei dieser Verfahrensvariante werden die Merkmale 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 Merkmale erfolgt nach den Korrelationen zwischen den Merkmalen (Statistische Ämter des Bundes und der Länder und IAW 2003; Lenz 2003a; Rosemann 2004). Für die einzelnen Merkmalsgruppen kann die Gruppenbildung analog zu den in a) und b) beschriebenen Varianten nach einem bestimmten Merkmal, einem Hilfsmerkmal oder via Distanzberechnungen vorgenommen werden.

b) Stochastische Mikroaggregation

Stochastische Mikroaggregationsverfahren wurden erstmals von Lechner und Pohlmeier (2003) vorgeschlagen. Dort werden zwei Möglichkeiten der stochastischen Mikroaggregation beschrieben, 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 Merkmale – beziehungsweise für Gruppen von Merkmalen – gemeinsam oder für alle Merkmale 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 –, sodass 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. Somit weicht dieser Ansatz wesentlich von der ursprünglichen Idee der Mikroaggregationsverfahren ab und könnte mit guten Argumenten auch unter den Imputationsmethoden aufgeführt werden.

Auch im Falle der zufälligen Mikroaggregation lässt sich die Matrixgleichung (1.11) aufstellen. Die Blockstruktur der Matrix ist ebenfalls erreichbar durch geeignete Umsortierung der

Merkmalsträger entsprechend der Gruppenbildung. In beiden Fällen ist die Mikroaggregationsmatrix D symmetrisch und idempotent, d.h. es gilt:

$$D^2 = D \quad \text{und} \quad D^T = D. \tag{1.12}$$

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

$$D_B = \frac{1}{k}\left(I_n + S_1 + S_2 + \cdots + S_{k-1}\right). \tag{1.13}$$

Dabei ist k die vorgegebene Blockgröße. Die Zeilen der $n \times n$ - „Selektionsmatrizen" S_i stellen beliebige Koordinateneinheitsvektoren dar, die an zufälliger Position eine Eins und an allen anderen Positionen Nullen enthalten. Somit ist die Aggregationsmatrix D_B eine Zufallsmatrix von blockdiagonaler Struktur. Wie bei der deterministischen und der zufälligen Mikroaggregation setzt sich ein anonymisierter Wert aus k Originalwerten zusammen, die aber nicht notwendigerweise verschieden sein müssen. Der wesentliche Unterschied zu den anderen Mikroaggregationsverfahren besteht darin, dass hier keine Gruppen identischer Merkmalsausprägungen entstehen, da die Gruppenbildung für jeden Merkmalsträger separat und zufällig erfolgt, was für einen näherungsweisen Varianzerhalt des behandelten Merkmals sorgt. Das arithmetische Mittel des originalen Merkmals wird zwar nicht erhalten, kann aber erwartungstreu geschätzt werden.

1.2.3 Methoden zum Schutz besonders gefährdeter Merkmalsträger

Bisher wurden datenverändernde Verfahren vorgestellt, die auf ganze Datenbestände oder zumindest für einzelne Merkmale anwendbar sind. In diesem Unterabschnitt 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.

- **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 dieselben Ausprägungen in den diskreten Merkmalen und ähnliche stetige Merkmalswerte, wenn zum Beispiel eine Störgröße mit Erwartungswert Null und kleiner Varianz dazu addiert wird. Das Verfahren führt zu einer Reduzierung der eindeu-

tigen Zuordnungen und damit zu einer steigenden Unsicherheit beim Datenangreifer. Allerdings werden die uni- und multivariaten Verteilungscharakteristika systematisch verzerrt.

- **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 dem Datennutzer unbekannten Schlüssel verteilt werden (Statistische Ämter des Bundes und der Länder 2003). Das Verfahren führt sowohl zu einer Reduzierung der eindeutigen Zuordnungen als auch zu einer Verringerung der Brauchbarkeit der enthüllten Einzelwerte. 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.

1.2.4 Kombination der Methoden für metrische und kategoriale Merkmale

Eine besondere Bedeutung kommt der Kombination unterschiedlicher Verfahren und Verfahrensgruppen zu, insbesondere wenn sowohl metrische als auch kategoriale Merkmale als Überschneidungsmerkmale auftreten. In der Praxis wird so vorgegangen, dass zunächst die kategorialen Merkmale „bis an die Schmerzgrenze" behandelt werden (in der Regel durch informationsreduzierende Methoden wie Vergröberung) und im Anschluss geeignete Verfahren auf die metrischen Merkmale Anwendung finden (etwa Abschneide- oder Mikroaggregationsverfahren). Beispiele hierzu finden sich zahlreich in Kapitel 4, weshalb an dieser Stelle darauf verzichtet wird. In diesem Abschnitt werden zwei spezielle, auf datenverändernde Methoden begrenzte Kombinationsmöglichkeiten vorgestellt.

Das Verfahren von Winkler

Das Verfahren von Winkler (Kim und Winkler 1995) stellt eine Kombination aus den stochastischen Verfahren der Zufallsüberlagerung und der Zufallsvertauschung dar. Merkmalsträger, die durch Zufallsüberlagerung nicht genügend anonymisiert werden können, werden nachträglich einem Vertauschungsverfahren unterzogen. D.h., es werden nur bei den Problemfällen Einzelwerte untereinander getauscht. Hiermit werden die Vorteile beider Verfahren ausgenutzt: Die höhere Datensicherheit durch das Vertauschungsverfahren einerseits und der bessere Erhalt der Datenqualität durch die Zufallsüberlagerung andererseits.

Die Arbeiten des Projektes FAWE haben gezeigt, dass sowohl mit dem Verfahren der additiven Zufallsüberlagerung als auch mit den verschiedenen Varianten der Vertauschungsverfahren nur geringe Erfolgsaussichten verbunden sind. Aus diesem Grunde wurde auch der Kombination beider Verfahrensgruppen wenig Bedeutung beigemessen. Hinzu kommt, dass das Verfahren von Winkler mit einem zusätzlichen manuellen Aufwand verbunden ist und damit für die alltägliche Arbeit in den statistischen Ämtern als unpraktikabel erscheint.

SAFE - ein Verfahren des Statistischen Landesamtes Berlin

Die Idee des Verfahrens SAFE besteht darin, einen Datenbestand zu erzeugen, in dem jeder Merkmalsträger mit mindestens zwei weiteren identisch ist (Evers und Höhne 1999; Höhne 2003a; Höhne 2003b; Höhne 2003c). Das im Statistischen Landesamt Berlin[10] entwickelte Verfahren SAFE stellt eine Kombination aus dem Vertauschen beziehungsweise Verändern diskreter Merkmalswerte mit einer Mikroaggregation bei den metrischen Merkmalen 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 die Mikroaggregationsverfahren. Es werden folgende Schritte durchgeführt:

- Geheimhaltung auf Basis diskreter Merkmale: Es wird nach dem Kriterium minimaler Fehler in den Randsummen eine diskrete Basisdatei erstellt, in der alle Ausprägungskombinationen diskreter Merkmale mit mindestens drei Einheiten besetzt sind.
- Zuordnung der Originalwerte der stetigen Merkmale zu den Merkmalskombinationen der zuvor veränderten diskreten Merkmale. Ziele sind die Erhaltung der größten Ähnlichkeit in den diskreten Merkmalen und das Verschieben der kleinsten Merkmalswerte der stetigen Merkmale.
- Bearbeitung der Dominanzen und des Problems der merkmalsbezogenen Fallzahlen unter zwei.
- Optimierung und Qualitätssicherung der Ergebnisse.
- Mikroaggregation der stetigen Merkmale innerhalb der gebildeten Gruppen.

SAFE reduziert das Risiko der korrekten und vor allem der eindeutigen Zuordnung von Merkmalsträgern, da immer mehrere identische und damit nicht unterscheidbare Einheiten

10 Am 1. Januar 2007 ist aus der Zusammenführung des Statistischen Landesamtes Berlin und dem Statistikteil des Landesbetriebes für Datenverarbeitung und Statistik Brandenburg das Amt für Statistik Berlin-Brandenburg hervorgegangen.

vorliegen. Außerdem wird durch die Mittelwertbildung eine Unsicherheit für den Datenangreifer erzeugt und damit eine weitere Schutzwirkung entfaltet.

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. 2005).

Kapitel 2

Konzept zur Messung der Datensicherheit anonymisierter Einzeldaten

In diesem Kapitel werden die Grundlagen für eine anwendungsfähige Operationalisierung des Begriffes der faktischen Anonymität gelegt (Höhne et al. 2003; Lenz et al. 2004b; Lenz et al. 2005b). Zunächst wird in Abschnitt 2.1 der Begriff der faktischen Anonymität erläutert. In den Folgeabschnitten 2.2 und 2.3 werden anschließend die für die Messung der Datensicherheit relevanten Szenarien möglicher Datenangriffe und das bei wirtschaftsstatistischen Einzeldaten zu beachtende Zusatzwissen eines potentiellen Datenangreifers dargestellt. In Abschnitt 2.4 werden die bei der Messung der Schutzwirkung von Anonymisierungsmaßnahmen zu berücksichtigenden Elemente aufgeführt und danach im abschließenden Abschnitt 2.5 geeignet zu einem Gesamtmaß für die faktische Anonymität zusammengeführt.

2.1 Der Begriff der faktischen Anonymität

Grundsätzlich dürfen die statistischen Ämter Einzeldaten an Dritte nur weitergeben, wenn es für die Empfänger keine Möglichkeit gibt, die Identität eines Merkmalsträgers zu ermitteln (d.h. diesen 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 hundertprozentig auszuschließen ist, dass ein Merkmalsträger (z.B. ein Unternehmen) identifiziert werden kann. 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. 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 Analyse-

zwecken übermittelt werden. Ausgehend vom Wortlaut des zitierten § 16 Abs. 6 (BStatG) gilt eine Datei als faktisch anonym, wenn der potentielle Datenangreifer aus rationalem Kalkül die Kosten der Deanonymisierung höher einschätzt als den Nutzen, den er aus einem erfolgreichen Versuch der Deanonymisierung 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 in einem Forschungsprojekt zur Anonymisierung von Personen- und Haushaltsdaten verwendet (Müller et al. 1991; Helmcke und Knoche 1992).

Eine Datei kann ebenfalls als faktisch anonym gelten, wenn der potentielle Datenangreifer die gewünschten Informationen aus alternativen Quellen kostengünstiger als durch eine Deanonymisierung der vertraulichen Daten beschaffen kann. Er wird auch in diesem Falle aus rationalem Kalkül heraus auf einen Deanonymisierungsversuch verzichten, da der Aufwand für ihn unverhältnismäßig hoch sein wird (Sturm 2002b). Es könnte demnach die Konstellation entstehen, dass für einen Datenangreifer nach Abwägung des Nutzens und der Kosten der Deanonymisierung zwar ein Nutzengewinn verbleibt, er aber den Datenangriff unterlässt, da er auf einem alternativen Wege einen höheren Nutzengewinn erzielen kann.

Nach dem Willen des Gesetzgebers muss also die Reidentifikation von faktisch anonymisierten Merkmalsträgern bzw. die Enthüllung von bestimmten Einzelinformationen nicht mit absoluter Sicherheit ausgeschlossen werden. Insbesondere muss dem Fall des „Datenschutzidealisten" im Konzept der faktischen Anonymität keine Relevanz beigemessen werden. Er stellt vielmehr einen Sonderfall dar, bei dem davon ausgegangen wird, dass es dem Datenangreifer 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 Situation soll bei der Beurteilung der faktischen Anonymität einer Unternehmensdatei nicht berücksichtigt 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 unmittelbar zwei Fragestellungen bezüglich der Schutzwirkung von Anonymisierungsmethoden:

- Inwieweit wird die Kosten-Nutzen-Relation des potentiellen Datenangreifers durch die Anonymisierung beeinflusst? Sei es, weil sich die Möglichkeit einer Reidentifikation oder der Nutzen der gewonnenen Informationen durch die Anonymisierung verringert oder weil sich die Kosten der Deanonymisierung erhöhen. Beides führt für den Datenangreifer zu einer negativen Beeinflussung seiner Kosten-Nutzen-Relation, wodurch ein Datenangriff unwahrscheinlicher wird.

- Welchen Einfluss haben alternative Quellen der Informationsbeschaffung (anstelle der vertraulichen Daten) auf das notwendige Anonymisierungsniveau? Je kostengünstiger diese Quellen für einen Datenangreifer zu erschließen sind, desto unwahrscheinlicher ist ein Datenangriff und niedriger das notwendige Anonymisierungsniveau. Allerdings können solche Quellen auch zur Erweiterung des Zusatzwissens (siehe Abschnitt 2.3) eines Datenangreifers dienen und damit die Chancen für eine Reidentifikation der den Datenangreifer interessierenden Einheiten erhöhen.

2.2 Szenarien des Datenangriffes

Zu einer Einschätzung des Reidentifikationsrisikos von Merkmalsträgern bzw. des Enthüllungsrisikos für einzelne Werte in den Zieldaten können Simulationen verschiedener Datenangriffsszenarien dienen (Elliot und Dale 1999; Vorgrimler und Lenz 2003a; Vorgrimler und Lenz 2003b), ausführlich zu Angriffsszenarien bei wirtschaftsstatistischen Einzeldaten vgl. Wirth (2003). Die relevantesten Szenarien sind ohne Zweifel der sogenannte 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 Informationen 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 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 potentieller Einzelangreifer die möglichen Kandidaten für eine richtige Zuordnung leichter überblicken und weitere individuelle Kenntnisse über das gesuchte Un-

ternehmen 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 (siehe auch Abschnitt 2.3) und vorhandener Simulationssoftware im Wesentlichen durch die i.d.R. überschaubare Rechenzeit begrenzt. Bei den Simulationen wird darüber hinaus zur Reduzierung des Aufwandes in den Analysen empfohlen, lediglich solche Merkmale zu verwenden, die für die Mehrzahl der Merkmalsträger vorliegen. Dies hat insbesondere den Vorteil, dass die geschätzten Reidentifikationsrisiken über mehrere Bereiche (etwa Größenklassen) hinweg verglichen und damit grundsätzlich gefährdete Bereiche aufgedeckt werden können.

Hierbei wird allerdings vernachlässigt, dass vor allem größere Unternehmen der Publizitätspflicht unterliegen und weitergehende Informationen leicht verfügbar wären. Da für solche Unternehmen im Rahmen der Simulation von Massenfischzügen ein unvollständiges Zusatzwissen unterstellt wird, sollten in einem zweiten Schritt für besonders reidentifikationsgefährdete Bereiche in den Zieldaten (wie z.B. die Klasse der Unternehmen mit wenigstens 500 Beschäftigten oder die Branchenführer in dünn besetzten Wirtschaftszweigen), 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 zeitaufwendig werden kann.

2.3 Struktur des Zusatzwissens eines Datenangreifers

Die Abgrenzung des möglichen Zusatzwissens eines Datenangreifers ist sehr schwierig. Die Ergebnisse dieses Abschnittes stützen sich im Wesentlichen auf die überaus gut recherchierte Arbeit von Vorgrimler (2003). Ohne die darin mühselig zusammengetragenen Quellen möglichen Zusatzwissens wäre eine Durchführung der in Abschnitt 2.2 beschriebenen Datenangriffsszenarien nicht möglich gewesen.

Das Risiko der Reidentifikation von Merkmalsträgern einer Erhebung hängt nicht allein von der zugrunde liegenden Zieldatei und den darauf angewendeten Anonymisierungsmaßnahmen ab. Es ist ebenfalls von Bedeutung, welche Vorabkenntnisse ein potentieller Datenangreifer mitbringt. Dieses „Zusatzwissen" 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. Die Recherche des Zusatzwissens ist daher nur eine Momentaufnahme, weshalb ein Datenanbieter bei der faktischen Anonymisierung einer Unternehmenserhebung eine solche Recherche für die spezielle Erhebung und das zugrunde liegende Betrichtsjahr 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 ändert. Daher wird in den nachfolgenden Abschnitten versucht, solche zeitlich stabilen Strukturmerkmale aufzuzeigen. Dies kann bei einer aktuellen Erstellung eines Scientific-Use-Files als Hilfestellung dafür dienen, das jeweilige Zusatzwissen zu analysieren. Bei den nachfolgenden konkreten Beispielen möglichen Zusatzwissens wird kein Anspruch auf Vollständigkeit erhoben. Sie illustrieren aber, wie künftig die Abgrenzung des jeweils relevanten Zusatzwissens vorgenommen werden kann.

Die Arten des Zusatzwissens können danach unterschieden werden, ob ein Datenangreifer dieses Wissen aus kommerziellen Datenbanken, nichtkommerziellen (öffentlichen) Informationsquellen (wie Telefonbücher, Handelsregister, Internet usw.) oder aus persönlichen Quellen bezieht (Elliot und Dale 1999).

2.3.1 Kommerzielle Unternehmensdatenbanken

Als kommerzielle Unternehmensdatenbanken werden solche verstanden, welche einem Nutzer gegen Entgelt Informationen über ein bestimmtes oder eine Auswahl von Unternehmen bereitstellen. Als prominente Beispiele sind hier die Datenbanken von Hoppenstedt und MARKUS-Datenbank zu nennen. Ein Nutzer kann generell zwischen Datenbanken für bestimmte Branchen und allgemeinen Unternehmensdatenbanken wählen. So kann er z.B. mithilfe 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 Unternehmensdatenbanken und teilweise oder sogar vollständig in den amtlichen Wirtschaftsstatistiken zu finden sind, sind die *Anzahl der Beschäftigten*, der *Gesamtumsatz*, 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 für den Massenfischzug nützlichen kommerziellen Datenbanken geführt wird und einem Einzelangreifer ohnehin bekannt ist.

Die Vorteile kommerzieller Unternehmensdatenbanken liegen in der hohen Anzahl an erfassten Unternehmen und der klaren Struktur der darin enthaltenen Informationen. Daher ermöglichen diese Datenbanken prinzipiell, wie in späteren Abschnitten dargelegt wird, die Durchführung eines Massenfischzuges. Allerdings muss für den Erwerb einer kompletten Datenbank mit erheblichen Kosten gerechnet werden, wohingegen die Informationen einzelner Unternehmen relativ kostengünstig erhältlich sind. Dagegen ist der bei einer Unternehmensrecherche auch berücksichtigende Zeitaufwand bei der Nutzung kommerzieller Datenbanken im Verhältnis zu anderen möglichen Quellen des Zusatzwissens recht gering.

Die kommerziellen Datenbankanbieter sind auf eine hohe Datenqualität angewiesen. Daher werden Maßnahmen zur Kontrolle und zur Verbesserung der Datenqualität ergriffen, die sich positiv auf die Eignung der Datenbank als Zusatzwissen auswirken. Die in den Datenbanken verfügbaren Überschneidungsmerkmale können qualitativ in zwei Gruppen eingeteilt werden:

- Merkmale mit vielen Ausprägungen, die einen Datenbestand sehr stark differenzieren, die aber im Zeitablauf eine hohe Variabilität aufweisen. Diese sind zum Beispiel die metrischen Merkmale *Gesamtumsatz* und *Anzahl der Beschäftigten*.
- Merkmale mit wenigen Ausprägungen, die einen Datenbestand weniger stark differenzieren, die jedoch im Zeitablauf nur eine geringe Variabilität aufweisen. Hierzu gehören zum Beispiel die kategorialen Merkmale *Regionalkennung, Branchenzugehörigkeit* und *Rechtsform*.

Aufgrund der starken Veränderung (vor allem der metrischen) Überschneidungsmerkmale über die Zeit sollte aus Sicht des Datenangreifers die kommerzielle Datenbank denselben 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 *Gesamtumsatz* und *Anzahl der Beschäftigten* als Überschneidungsmerkmale, zumindest dann, wenn der Erhebungszeitraum des veröffentlichten Zieldatensatzes einige Jahre zurückliegt.[11] Kategoriale Merkmale wie *Regionalkennung, Branchenzugehörigkeit* und *Rechtsform* eines Unternehmens sind im Zeitablauf zwar relativ stabil und nur gering fehlerbehaftet, differenzieren aber

11 In Abschnitt 7.2 wird beispielhaft die mangelhafte Qualität der MARKUS-Datenbank in Bezug auf die zeitliche Veränderung ihrer Merkmale aufgezeigt.

die Unternehmen weniger stark als metrische Merkmale wie *Gesamtumsatz* und *Anzahl der Beschäftigten*. Eindeutige Zuordnungen mittels der kategorialen Überschneidungsmerkmale werden die seltene Ausnahme sein, für einen Datenangreifer bieten sie aber eine sichere Möglichkeit, die Unternehmen in einem anonymisierten Datensatz zu „blocken" und den Datenangriff blockweise durchzuführen. Es wird also eine Vorauswahl möglicher Kandidatenpaare getroffen für eine spätere mittels der metrischen Überschneidungsmerkmale zu realisierende Zuordnung.

Abschließend werden in Tabelle 2.1 beispielhaft einige bekannte kommerzielle Unternehmensdatenbanken mit ihren Eigenschaften aufgelistet. Als gemeinsame Merkmale, die in Unternehmensdatenbanken und den vertraulichen Zieldaten als Überschneidungsmerkmale auftauchen können, sind die *Anzahl der Beschäftigten*, der *Gesamtumsatz*, die *Rechtsform* und die *Regionalkennung* zu nennen (Vorgrimler 2003 und Ronning et al. 2005).

Tabelle 2.1: Auswahl an Unternehmensdatenbanken

Bezeichnung	Schwerpunkt	Anzahl der Unternehmen	Preis
Hoppenstedt Top 7 000	die 7 000 größten Unternehmen	7 000	13 500 € + MWSt Teilinformationen zu niedrigeren Preisen
Hoppenstedt Firmen-datenbank	allgemeine Unternehmens-datenbank	ca. 152 000	1 500 € + MWSt oder z.B. 20 Profile für 155 €
M&M Deutsche Handels-datenbank	Handels-unternehmen („Top-Firmen")	ca. 8 500 (350)	1 500 €
BvD MARKUS	allgemeine Unternehmens-datenbank	ca. 850 000 aus Deutschland und Österreich	VhB Mindestumsatz: 5 000 Dollar
Zollner Unternehmens-profile	allgemeine Unternehmens-datenbank	ca. 70 000	keine Angabe

2.3.2 Nichtkommerzielle Informationsquellen

Mit nichtkommerziellen öffentlichen Informationsquellen sind Quellen gemeint, die allgemein zugänglich sind und deren Zweck nicht die kommerzielle Übermittlung von Unternehmensinformationen ist. Die Übermittlung der Informationen dient bei diesen Informationsquellen einem Nebenzweck (Vorgrimler 2003). Bei einem Telefonbuch etwa 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 können nichtkommerzielle Informationsquellen als solche definiert werden, für deren Abruf der Nutzer kein Entgeld zahlen muss. Nennenswert sind in erster Linie Datenbanken von Verbänden und Industrie- und Handelskammern und die Internetauftritte der Unternehmen. Am Beispiel der IHK-Datenbanken sollen im Folgenden diese Informationsquellen bewertet werden.[12]

Der wichtigste Vorteil ist die (kosten)freie Verfügbarkeit der Datenbanken über das Internet. Daneben ist weiterhin positiv, dass sie sehr kleine Unternehmen enthalten. Diesen Vorteilen stehen aber schwerwiegende Nachteile gegenüber. So werden nicht von allen Industrie- und Handelskammern solche Datenbanken angeboten. Diejenigen, die eine Datenbank anbieten, bieten i.d.R. lediglich eine Suchmöglichkeit innerhalb des eigenen Kammerbezirks. Eine Ausnahme bilden die Industrie- und Handelskammern 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 (Vorgrimler 2003).

Die Nutzung der IHK-Datenbanken für einen Datenangriff ist mit einer großen Unsicherheit behaftet. Zum einen, weil die Einträge auf freiwilliger Basis erhoben werden. Zum anderen ist die Datenqualität nicht abschätzbar (d.h. ob und mit welcher Genauigkeit die Unternehmen die angeforderten Merkmale angeben); dies gilt selbst bei dem als stabil einzuschätzenden Merkmal der Branchenzugehörigkeit. Aufgrund der genannten 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. Die Datenqualität ist eher fraglich. 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 Zeit für die Suche investieren. Im Rahmen eines Einzelangriffsszenarios wurde in den IHK-Datenbanken nach Zusatzwissen recherchiert, zu den Ergebnissen

12 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 Gesamtheit der einzelnen regionalen Datenbanken verstanden wird. Neben frei zugänglichen bieten die IHKs auch kostenpflichtige Datenbanken an. Diese werden zu den kommerziellen Datenbanken gezählt und daher in diesem Kontext nicht betrachtet.

siehe Vorgrimler (2003), Scheffler (2005) und Lenz et al. (2005a). 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 geführt wurden, war die Datenqualität nur selten befriedigend. Aus diesen Gründen haben sich die IHK-Datenbanken nur in wenigen Fällen als geeignete Informationsquellen zur Reidentifikation eines Unternehmens erwiesen.

Das Internet, das bisher als Medium zur Verwendung der IHK-Datenbanken bereits in die Recherchen 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. Neben dem Internet stehen auch Fachzeitschriften, Tagespresse, Fernsehen und weitere denkbare Medien zur Verfügung (Vorgrimler 2003).

2.3.3 Persönliche Informationsquellen

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

Das persönliche Wissen steht einem Datenangreifer zum Nulltarif zur Verfügung. Man kann davon ausgehen, dass er es bereits in der Vergangenheit zu einem anderen Zweck generiert hat und nun für den Nebenzweck des Datenangriffes abruft. Hierzu können auch private und berufliche Erfahrungen gezählt werden. Zum Beispiel hat ein Experte der Automobilbranche einen hohen Wissensstand über diesen Markt und kann diesen als Zusatzwissen bei der gezielten Suche nach einzelnen Unternehmen einsetzen.

Da nicht davon ausgegangen werden kann, dass eine oder mehrere Personen über eine Vielzahl von Unternehmen Informationen 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 Datenangreifer verfügt möglicherweise bereits vorab über Spezialkenntnisse über ein bestimmtes Unternehmen und kann diese durch gezielte Recherchen relativ kostengünstig vergrößern und so genügend Wissen sammeln, um einen Reidentifikationsversuch erfolgreich durchführen zu können.

Die verschiedenen denkbaren Datenangreifer bringen jeweils unterschiedliches Expertenwissen mit. So hat ein Marktexperte in seinem speziellen Bereich bereits vorab eine große Kenntnis über die betreffenden Unternehmen, während andere mögliche Datenangreifer hier weit weniger Vorabwissen mitbringen. Möglicherweise ist er selbst ein Auskunftgebender der betreffenden Erhebung und interessiert sich für Informationen konkurrierender Unterneh-

men. Ein aus Sicht des Autors sehr kritisch zu beurteilendes Szenario besteht, wenn der Datenanbieter (etwa ein nationales statistisches Amt, vertreten durch einen Mitarbeiter oder Fachstatistiker) im Rahmen der Datensicherheitsprüfungen bewusst oder unbewusst vertiefte über viele Jahre erworbene Spezialkenntnisse über die Erhebung dazu verwendet, erfolgreich Einzelinformationen herauszufiltern mit dem Ergebnis, daraus die Notwendigkeit strenger Anonymisierungsmaßnahmen abzuleiten. Dieses Vorgehen kann durchaus als Datenschutzidealismus interpretiert werden, welcher bei der Erzeugung faktisch anonymer Daten außen vor bleiben sollte (siehe hierzu auch die entsprechenden Ausführungen in Abschnitt 2.1).

Die Problematik des uneinheitlichen Zusatz- bzw. Expertenwissens 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 wissenschaftliche Forschungsinstitute und Hochschulen beschränkt, d.h. auf Einrichtungen und nicht auf Personen. Die Personen, die letztlich Zugang zu den 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 auf der einen Seite bereits vorab prüfen, inwieweit es eventuell zu Interessenkonflikten bei den Wissenschaftlern kommen kann, die Zugang zu den Daten erhalten. Die Wissenschaftler, welche die Einzeldaten in faktisch anonymer Form erhalten, müssen sich auf der anderen Seite vertraglich verpflichten und dabei erklären, dass kein Interessenskonflikt besteht. Die rechtlichen Rahmenbedingungen und vertraglichen Ausgestaltungen müssen also das Problem des schwierig abzugrenzenden Expertenwissens lösen. Nur auf diese Weise kann eine sinnvolle Anonymisierung der Daten gewährleistet werden. Würde man auch das Expertenwissen mit in die Anonymisierungsüberlegungen einbeziehen, so führten die daraus resultierenden Maßnahmen zu einem sehr starken und von wissenschaftlicher Seite nicht tolerierbaren Eingriff in das Analysepotential der Daten.

2.4 Elemente des Schutzes von Einzeldaten

Um zu einer Operationalisierung der faktischen Anonymität, d.h. zu einem Maßstab, was bei der Reidentifikation von Einzeldaten ein „unverhältnismäßig hoher Aufwand" ist, zu gelangen, werden realitätskonforme Annahmen für einen Datenangriff zugrunde gelegt. Aus dem Rationalkalkül eines potentiellen Datenangreifers kann zunächst abgeleitet werden: Eine Verletzung der faktischen Anonymität wird nur dann erfolgen, wenn eine für einen Datenangreifer als nutzbringend einzustufende Reidentifikation möglich erscheint.

Die Herstellung einer nutzbringenden Reidentifikation wäre für einen Datenangreifer erwägenswert, wenn er abschätzen könnte, dass der (erwartete) Ertrag eines Reidentifikationsversuchs dessen (erwarteten) Aufwand übersteigt. Kosten und Nutzen sind also davon abhängig, wie sicher ein Datenangreifer brauchbare Informationen enthüllen kann.

In der Praxis wird sich ein Datenangreifer mit zahlreichen Problemen konfrontiert sehen, hierzu eine kleine Auswahl:

- das Auftreten von Dateninkompatibilitäten zwischen Zusatzwissen und vertraulichen Zieldaten,
- die Unkenntnis darüber, ob eine gesuchte Einheit in den Zieldaten enthalten ist (im Falle von Stichprobenerhebungen),
- Unsicherheiten über die Korrektkeit von Zuordnungsversuchen und
- Unsicherheiten hinsichtlich der Qualität aufgedeckter Informationen.

2.4.1 Dateninkompatibilitäten zwischen Zusatzwissen und Zieldaten

Ein entscheidender Einflussfaktor für die Datenqualität ist der Weg, auf welchem die Daten generiert werden. Die Abhängigkeit der Daten vom datengenerierenden Prozess führt dazu, dass Merkmale zum gleichen Sachverhalt, die auf unterschiedlichen Wegen gewonnen werden, sehr unterschiedliche Werte annehmen können. Dies kann zum Beispiel darin begründet liegen, dass ein Sachverhalt zwar gleich deklariert wird, sich die exakten Definitionen hinter den Begriffen jedoch unterscheiden, oder darin, dass verschiedene Personen, die an den unterschiedlichen Erhebungen beteiligt sind, manche Begrifflichkeiten jeweils verschieden auffassen und damit unterschiedlich ermitteln. Je unterschiedlicher der Prozess, desto größere „Dateninkompatibilitäten" können auftreten. Bei den in späteren Kapiteln behandelten Daten handelt es sich um primärstatistische Daten, also um solche, die speziell für die statistische Fragestellung durch eine direkte Erhebung gewonnen werden (zum Beispiel die Kostenstrukturerhebung im Verarbeitenden Gewerbe), um sekundärstatistisch aus Verwaltungsquellen übernommene Daten (zum Beispiel die Umsatzsteuerstatistik) und um Mischformen aus primär- und sekundärstatistischen Daten (zum Beispiel die kommerzielle MARKUS-Datenbank). Die datengenerierenden Prozesse unterscheiden sich demnach erheblich, weshalb Dateninkompatibilitäten unvermeidlich sind. Solche Dateninkompatibilitäten wirken als natürlicher Schutz gegenüber Datenangriffen, da eine Reidentifikation bereits ohne zusätzliche Anonymisierung (abgesehen von der formalen Anonymisierung) der vertraulichen Zieldaten erschwert wird.

2.4.2 Unsicherheit über die Möglichkeit der Zuordnung

Um einen Merkmalsträger einwandfrei in einer Erhebung reidentifizieren zu können, muss ein Datenangreifer die Kenntnis darüber besitzen, dass der gesuchte Merkmalsträger an der Erhebung teilgenommen hat (Teilnahmekenntnis). Ohne diese Kenntnis ist jede erfolgte eindeutige Zuordnung mit der zusätzlichen Unsicherheit behaftet, ob diese Zuordnung

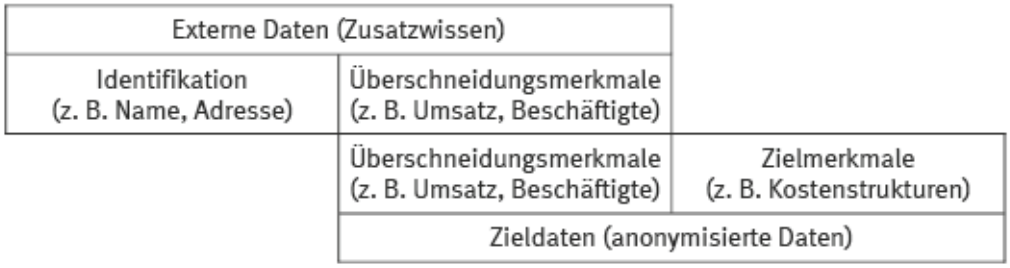
zustande kam, obwohl der gesuchte Merkmalsträger nicht in der Erhebung enthalten ist. Ist dies der Fall, kann es sich nur um eine Fehlzuordnung handeln. Bei der Kostenstrukturerhebung im Verarbeitenden Gewerbe 1999 werden alle Unternehmen mit wenigstens 500 Beschäftigten erfasst. Die mittleren Unternehmen (250-499 Beschäftigte) werden zu 73%, die kleineren Unternehmen (20-249 Beschäftigte) zu 38% und die kleinsten Unternehmen (weniger als 20 Beschäftigte) überhaupt nicht erfasst. Demnach hat ein Datenangreifer bei der Suche nach großen Unternehmen Kenntnis über die Teilnahme des gesuchten Unternehmens an der Erhebung. Dagegen muss er bei kleinen und mittleren Unternehmen die Unsicherheit in Kauf nehmen, dass sein gesuchtes Unternehmen mit großer Wahrscheinlichkeit nicht an der Erhebung teilgenommen hat.

Bei der Umsatzsteuerstatistik 2000 werden alle Unternehmen, die Umsatzsteuer-Voranmeldungen abgegeben haben, mit einem Jahresumsatz (ohne Umsatzsteuer) von über 32 500 DM (16 617 €) erfasst. Damit sind rund 2,9 Millionen Unternehmen in der Erhebung enthalten. Die Umsatzsteuerstatistik ist also naheuzu eine Vollerhebung und ein Datenangreifer besitzt daher nahezu gesicherte Kenntnis über die Teilnahme des gesuchten Unternehmens an der Erhebung.

2.4.3 Unsicherheit über die Korrektheit der Zuordnung

Das Zusatzwissen eines potentiellen Datenangreifers 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 2.1). Unter einer Reidentifikation wird schließlich verstanden, ein bestimmtes Zusatzwissen über einen gesuchten Mewrkmalsträger mithilfe der Überschneidungsmerkmale mit dem Zielmerkmalsträger eindeutig und richtig zu verknüpfen. Ein ernst zu nehmender Zuordnungsversuch kann nur mittels der angesprochenen Überschneidungsmerkmale erfolgen.

Abbildung 2.1
Überschneidungsmerkmale zwischen Datenbanken



Die Reidentifikation (richtige und eindeutige Zuordnung) eines gesuchten Merkmalsträgers eröffnet einem Datenangreifer dabei Kenntnisse über sensible Sachverhalte (Zielmerkmale) von ihn interessierenden Merkmalsträgern (Zielmerkmalsträger, Zielobjekte). Eine Information, die nach einer erfolgreichen Zuordnung enthüllt werden könnte, ist zu verstehen als ein Einzelwert y_i eines vor der Weitergabe der vertraulichen Daten nicht bekannten Zielmerkmals v_i für den gefundenen Merkmalsträger.

Merkmale werden im internationalen Sprachgebrauch „sensitive" bzw. „sensible Merkmale" genannt, wenn sie Informationen enthalten, die im Sinne der Auskunftgebenden nicht aufgedeckt werden dürfen. Da es den Datenanbietern aus Kapazitätsmangel im Allgemeinen nicht möglich ist, jeden einzelnen Teilnehmer einer Erhebung um die Zustimmung zur Weitergabe bestimmter Merkmale an die Wissenschaft zu bitten, müssen von juristischer Seite[13] sämtliche Merkmale (Überschneidungs- und Zielmerkmale) einer zu anonymisierenden Datei als sensitiv eingestuft werden.

2.4.4 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 und abgeleiteten Schlussfolgerungen 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 zuverlässig hält. Wird im Folgenden daher von richtig oder korrekt zugeordneten Informationen gesprochen, so sind alle Merkmalsausprägungen eines reidentifizierten Merkmalsträgers 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

13 Die Statistischen Ämter des Bundes und der Länder müssen dem Bundesstatistikgesetz, dem Datenschutzgesetz und in Einzelfällen dem für die jeweilige Erhebung formulierten Gesetz Folge leisten.

"wahren" Wert[14] entspricht oder diesem in einem bestimmten Maße ähnelt. Ab einer gewissen Abweichung wird ein Datenangreifer keinen Nutzen mehr aus einer richtig zugeordneten Information erzielen können. Die Abweichungsgrenze, ab welcher der Wert eines Merkmals i als unbrauchbar gilt, wird im Folgenden „Nützlichkeitsschwelle" γ_i genannt. Im Rahmen der Anonymisierung von Tabellen wurden vergleichbare Schwellen bereits früher festgelegt.

Auch an dieser Stelle ist es wieder wesentlich zu beachten: Ob eine 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, so wird das Risiko, eine unbrauchbare Information mit einer brauchbaren Information zu verwechseln, als zu hoch eingeschätzt. Daraus folgt, dass für einen Datenangreifer eine Information wertlos werden kann unabhängig davon, ob die gefundene Information innerhalb oder außerhalb einer bestimmten Abweichungsumgebung liegt.

2.5 Zusammenführung zu einem Maß für faktische Anonymität

In die nachfolgende Definition eines Maßes für die faktische Anonymität fließen nun die im vorherigen Abschnitt besprochenen Elemente ein. Nach einer durchgeführten Simulation eines Datenangriffes wird zum einen der Anteil richtiger Zuordnungen und zum anderen der Anteil nützlicher Informationen innerhalb der richtigen Zuordnungen berechnet. Auch eine richtige Zuordnung eines Merkmalsträgers kann einen für den potentiellen Datenangreifer erfolglosen Angriffsversuch darstellen, nämlich wenn der ihn interessierende Einzelwert (bzw. die ihn interessierende Information) relativ um wenigstens γ von dem tatsächlichen Originalwert, zum Beispiel $\gamma = 0.05$, abweicht[15]. Wir führen daher Schwellen γ_i für jedes Merkmal v_i ein. Sei $r^{(i)}$ der Wert des Merkmals v_i für einen bestimmten Merkmalsträger r der anonymisierten Daten und $o^{(i)}$ der zugehörige Originalwert. Der Wert $r^{(i)}$ wird als für den Datenangreifer „brauchbarer Einzelwert" verstanden, wenn

$$\tilde{r}^{(i)} := \frac{|o^{(i)} - r^{(i)}|}{|o^{(i)}|} < \gamma_i \tag{2.1}$$

für ein vorgegebenes $\gamma_i > 0$ gilt. In der Praxis hat sich bewährt, aus Gründen der einfachen Handhabbarkeit eine gemeinsame Schwelle γ für alle Merkmale zu setzen.[16]

14 Als wahrer Wert wird hier die entsprechende Ausprägung in den Originaldaten verstanden.

15 Strukturelle Nullen und fehlende Werte, die in der Regel bei der Anonymisierung erhalten bleiben, werden von dieser Betrachtung ausgenommen.

16 Bei der Untersuchung von Paneldaten könnte es sinnvoll sein, die Schwellen in Abhängigkeit der Aktualität der Merkmale zu setzen. Beispielsweise stufenweise $\gamma = 0.05, 0.10, 0.15, 0.2$ für die Merkmale *Gesamtumsatz 2005* bis *2008*.

Die Wahrscheinlichkeit, dass ein Datenangreifer einen Einzelwert $r^{(i)}$, welcher um weniger als γ von seinem Originalwert $o^{(i)}$ relativ abweicht, richtig zuordnet, werde mit $P_\gamma(o^{(i)}$ enthüllt) bezeichnet und im Folgenden Enthüllungsrisiko genannt.

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. Sei hierzu R die Menge der bereits reidentifizierten Einheiten, $r = (r^{(1)}, \ldots, r^{(n)})$ ein beliebiger Merkmalsträger der Zieldaten und $o = (o^{(1)}, \ldots, o^{(n)})$ der zugehörige Merkmalsträger in den Originaldaten. Wir haben folglich

$$P_\gamma(o^{(i)} \text{ enthüllt}) := P(\tilde{r}^{(i)} < \gamma \text{ und } r \in R) \tag{2.2}$$

zu bestimmen. Als Schätzer für $P(r \in R)$ verwenden wir die relative Häufigkeit $\hat{P}(r \in R)$ der durch den Datenangreifer korrekt zugeordneten Merkmalsträger, die nun auf ihre Brauchbarkeit hin untersucht werden. Unter Berücksichtigung aller korrekten Zuordnungen und aller metrischen Merkmale (inklusive Überschneidungsmerkmale) ist der Anteil $\hat{P}(\tilde{r}^{(i)} < \gamma \,|\, r \in R)$ der Einzelwerte, welche die Ungleichung (2.1) erfüllen, ein Schätzer für die bedingte Wahrscheinlichkeit $P(\tilde{r}^{(i)} < \gamma \,|\, r \in R)$, die sich liest „eine brauchbare Information (auf Merkmalsebene) zu finden, gegeben eine erfolgreiche Zuordnung (auf Merkmalsträgerebene)". Somit erhalten wir als Schätzer für das mit einer probeweise anonymisierten Datei verbundene Enthüllungsrisiko

$$\hat{P}_\gamma(o^{(i)} \text{ enthüllt}) := \hat{P}(r \in R) \cdot \hat{P}(\tilde{r}^{(i)} < \gamma \,|\, r \in R). \tag{2.3}$$

Ein risikoaverser Datenangreifer wird ab einer bestimmten Wahrscheinlichkeit, aus einem Enthüllungsversuch 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 nun, dass eine vorgegebene obere Risikoschwelle τ für das Enthüllungsrisiko nicht überschritten werden darf, d.h. es muss

$$\hat{P}_\gamma(o^{(i)} \text{ enthüllt}) < \tau \tag{2.4}$$

gelten. Wird diese von den Datenhaltern gewissenhaft a priori festzulegende Risikoschwelle nicht überschritten, so kann die zugrundeliegende Datei als faktisch anonym eingestuft werden. Je nach betrachteter Statistik kann diese Schwelle mit unterschiedlichen Werten angesetzt werden, wovon aber abgeraten wird. Setzt man die Risikoschwelle etwa mit $\tau = 0.5$ an, so würde erreicht, dass ein potentieller Datenangreifer bei der Suche nach einem bestimmten Einzelwert mit einer Wahrscheinlichkeit von über 50% eine unbrauchbare

Information aufdeckt. Aus diesem Grunde kann die Festlegung einer Risikoschelle $\tau \leq 0.5$ aus Sicht der Datenanbieter gut begründet werden.

In Kapitel 4 wird diese Vorgehensweise mit Beispielen für eine geeignete Schwellenwahl bei den amtlichen Wirtschaftsstatistiken Kostenstrukturerhebung im Verarbeitenden Gewerbe, Umsatzsteuerstatistik und Einzelhandelsstatistik vorgestellt.

Kapitel 3

Modellierung von Szenarien mit anonymisierten Querschnittsdaten

Zur Bewertung der Schutzwirkung von Anonymisierungsmaßnahmen muss gemäß Kapitel 2 dass Enthüllungsrisiko für Einzelwerte adäquat geschätzt werden. Nach Klärung der Grundbegriffe in Abschnitt 3.1 steht in diesem Kapitel das in Abschnitt 2.2 beschriebene Szenario des Massenfischzugs im Vordergrund. Mit dem Ziel, seine externe Datenbank qualitativ zu verbessern oder um weitere Informationen anzureichern, versucht ein Datenangreifer, möglichst viele Merkmalsträger der externen Datenbank den zugehörigen Merkmalsträgern in den Zieldaten eindeutig und korrekt zuzuordnen. Hierzu verwendet er wie bereits erwähnt Überschneidungsmerkmale. Nach der Definition geeigneter Abweichungs- bzw. Distanzmaße für diese Merkmale in Abschnitt 3.2 wird in Abschnitt 3.3 mathematisch die Möglichkeit der Sortierung der Überschneidungsmerkmale nach dem Grad ihrer Verlässlichkeit und Qualität mittels sogenannter Präferenzfunktionen modelliert. Die Aufgabe, die Anzahl korrekter Zuordnungen zu maximieren, wird in Abschnitt 3.4 als multikriterielles lineares Zuordnungsproblem modelliert und anschließend via geeignete Parametrisierung in ein lineares Zuordnungsproblem mit einer Zielfunktion transformiert. Zur Lösung dieses Problems kommen sowohl klassische Verfahren wie der weit verbreitete Simplex-Algorithmus als auch Näherungsheuristiken in Betracht. Danach wird versucht, auch eine geeignete Modellierung des in Abschnitt 2.2 beschriebenen Einzelangriffes vorzunehmen, um in Abschnitt 3.5 beide Strategien gegenüberstellen zu können. Abschließend wird an einem kleinen Beispiel der Einfluss verschiedener Parametersetzungen bei der in Abschnitt 3.3 vorgenommenen Gewichtung der Überschneidungsmerkmale auf das spätere Zuordnungsergebnis diskutiert.

3.1 Grundlagen

In diesem Abschnitt werden zunächst die in den Abschnitten 3.3, 3.4 und 7.4 verwendeten Bezeichnungen aus der Graphentheorie eingeführt. Danach beschäftigen wir uns mit den

verschiedenen Merkmalstypen und der Möglichkeit, geeignete Distanzmaße für ein Merkmal gegebenen Typs zu definieren. Dem Leser wird empfohlen, die in Unterabschnitt 3.1.1 eingeführten Begriffe und Definitionen zunächst zu übergehen und erst während der Lektüre der späteren Abschnitte 3.3, 3.4 und 7.4 nachzuschlagen.

3.1.1 Grundbegriffe der Graphentheorie

Ein (endlicher) *Graph* $\mathcal{G} = (V(\mathcal{G}), E(\mathcal{G}))$ ist eine relationale Struktur, bestehend aus einer (endlichen) Menge $V(\mathcal{G})$, deren Elemente *Ecken* (oder Punkte) genannt werden, und einer Menge $E(\mathcal{G}) \subseteq V(\mathcal{G})^2$ ungeordneter Paare von Ecken, den *Kanten* (oder Linien) von $\mathcal{G}$.

Wenn aus dem Kontext heraus klar ist, welcher Graph untersucht wird, schreiben wir abkürzend V oder E. Wir betrachten stets ungerichtete Graphen, welche die Implikation $(a, b) \in E \Longrightarrow (b, a) \in E$ erfüllen. D.h., E bestimmt eine symmetrische binäre Relation. Eine Kante $(a, b) \in E$ heißt *inzident* mit den Ecken a und b, und a, b heißen in diesem Falle *adjazent*. Ein Graph $\mathcal{T} = (V(\mathcal{T}), E(\mathcal{T}))$ heißt *Teilgraph* von $\mathcal{G}$, wenn $V(\mathcal{T}) \subseteq V(\mathcal{G})$ und $E(\mathcal{T}) \subseteq E(\mathcal{G})$ gilt.

Ein Graph $\mathcal{G}$ heißt *bipartiter Graph* mit Bipartition (X, Y), wenn $V(\mathcal{G})$ eine disjunkte Vereinigung $V = X \dot{\cup} Y$ ist, sodass jede Kante e mit einem $x \in X$ und einem $y \in Y$ inzident ist. Darüber hinaus heißt ein bipartiter Graph *vollständig*, wenn jedes $x \in X$ mit jedem $y \in Y$ verbunden ist und umgekehrt, d.h., wenn $E(\mathcal{G}) = X \times Y$ gilt. Eine *Zuordnung* (engl. matching) $\mathcal{M} \subseteq \mathcal{G}$ ist ein Teilgraph mit der Eigenschaft, dass je zwei verschiedene Kanten von $\mathcal{M}$ keine gemeinsame Ecke besitzen. Eine Zuordnung $\mathcal{M}$ wird *perfekte Zuordnung* genannt, wenn keine Zuordnung $\mathcal{M}'$ mit $\mathcal{M} \subset \mathcal{M}'$ existiert. Ist v eine Ecke von $\mathcal{M}$, d.h., $v \in V(\mathcal{M})$, dann wird v *$\mathcal{M}$-gesättigt* oder kurz *gesättigt* genannt. Ist jedes $v \in V$ $\mathcal{M}$-gesättigt, so ist $\mathcal{M}$ eine *perfekte Zuordnung*.

Ein *vektorgewichteter* Graph $\mathcal{G}$ ist ein Graph, versehen mit einer Gewichtsfunktion

$$\begin{aligned} w : E(\mathcal{G}) &\longrightarrow \mathbb{R}^k, \\ e &\longmapsto (w_1(e), \ldots, w_k(e)), \end{aligned}$$

die jeder Kante e ein k-Tupel reeller Zahlen zuordnet. Im Falle $k = 1$ wird der Kraph $\mathcal{G}$ *gewichteter Graph* genannt.

Ein *minimal spannender Wald* $\mathcal{F}$ ist ein zykelfreier Teilgraph eines gewichteten Graphen $\mathcal{G}$, wobei jede Ecke gesättigt ist und $\mathcal{F}$ mit dieser Eigenschaft von minimalem Gesamtgewicht

$$w(\mathcal{F}) := \sum_{e \in E(\mathcal{F})} w(e)$$

ist.

Zykelfrei bedeutet, dass je zwei Ecken höchstens über einen Kantenweg miteinander verbunden sind.

3.1.2 Merkmalstypen

Wie bereits in Abschnitt 2.4.3 geschildert wurde, sind zur Gegenüberstellung zweier Datenquellen sogenannte Überschneidungsmerkmale nötig, die in beiden Quellen enthalten sind. Offenbar ist die Qualität solcher Merkmale wesentlich für den Erfolg eines Reidentifikationsexperimentes. Wir unterscheiden grundsätzlich zwischen zwei Typen von Überschneidungsmerkmalen, den metrischen und kategorialen Merkmalen, welche nachfolgend zusammen mit geeigneten Distanzmaßen beschrieben werden.

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.

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).

Bei großen Datenmengen ist es aus Sicht eines potentiellen 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. Hierdurch entsteht eine Partitionierung des Datenbestandes in Blöcke bestehend aus $n : m$-Zuordnungen (d.h., n Objekte auf der einen werden m Objekten auf der anderen Seite zugeordnet), welche als Vorauswahl für eine spätere $1 : 1$-(jedes Objekt auf der einen wird genau einem Objekt auf der anderen Seite zugeordnet) bzw. $n : 1$-Zuordnung (n Objekte auf der einen werden ein und demselben Objekt auf der anderen Seite zugeordnet) interpretiert werden können.

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, welche zu derselben zugrundeliegenden Einheit gehören, landen möglicherweise nicht in demselben Block), sind Fehlzuordnungen besonders in großen Blöcken mit vielen ähnlichen Merkmalsträgern zu erwarten.[17] 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 Datenan-

17 Eine umfangreiche empirische Untersuchung hierzu findet sich in Lenz und Vorgrimler (2005).

greifer schwierig auszumachen. Im (für den Datenangreifer) schlimmsten Falle sind die in den beiden Quellen gebildeten zueinander gehörigen Blöcke disjunkt bzw. durchschnittsleer, im besten Falle sind die Blockmerkmale charakteristisch und dienen als direkte Identifikatoren, sodass 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 zerlegen.

In der Praxis sind kategoriale Merkmale wie *Wirtschaftszweigklassifikation, Rechtsform* oder *Regionalkennung* sehr gut als Blockmerkmale geeignet, auch wenn diese gemäß ihrer hierarchischen Struktur systematisch vergröbert wurden (siehe Unterabschnitt 3.2.1). Kleinere Probleme können entstehen, wenn die kategorialen Merkmale aus metrischen Merkmalen konstruiert werden, die möglicherweise zuvor mit datenverändernden Methoden behandelt wurden. Zum Beispiel kann es passieren, dass ein Unternehmen die Beschäftigtengrößenklasse nach vorheriger Veränderung der Beschäftigtenangabe wechselt. Größere Probleme sind bei Anwendung der zufallsorientierten Methode PRAM (siehe Unterabschnitt 1.2.1) zu erwarten. Allerdings sollte diese Methode mit Bedacht und damit nur unter gewissen Restriktionen bei der Veränderung einzelnder Werte eingesetzt werden. Zum Beispiel könnte bei Modifikation des siedlungsstrukturellen Kreistyps (sogenannter Neunerschlüssel BBR9: abhängig vom Grad der Urbanisierung der Umgebung des Hauptstandortes des Unternehmens wird eine der neun Ausprägungen $1 < 2 < \ldots < 9$ zugewiesen) gefordert werden, dass eine Ausprägung entweder unverändert oder mit gleicher Wahrscheinlichkeit nur durch ihre direkt benachbarten Ausprägungen (etwa „3" durch „2"„3" oder „4") ersetzt wird. Hierdurch entstünde als Übergangsmatrix eine Tridiagonalmatrix, die auf der Hauptdiagonalen mit 0,5 und den beiden Nebendiagonalen mit 0,25 besetzt ist. Für die Veränderungsregeln bei den Werten „1" und „9" könnten die Wahrscheinlichkeiten 0,75 und 0,25 sinnvoll sein. Bei der Wirtschaftszweigklassifikation, etwa auf der Zweistellerebene (sogenannte Wirtschaftsabteilungen), könnte eine Veränderung nur innerhalb der Abteilung, d.h. maximal auf der Dreistellerebene, gefordert werden. Die zugehörige Übergangsmatrix hätte dann eine Blockstruktur. In beiden Beispielen wäre eine geeignete Blockung des Datenbestandes möglich als Vorauswahl von Kandidatenpaaren für eine mögliche spätere Zuordnung.

3.2 Distanzmaße

Im Folgenden wird die Menge der Überschneidungsmerkmale zwischen den externen Daten A und den Zieldaten B mit $\{v_1, \ldots, v_k\}$ bezeichnet. Wir führen nun für jede Variable v_r

eine Distanzfunktion

$$d_i : A \times B \longrightarrow \mathbb{R}^+ \cup \{0\}, \; i = 1, \ldots, k$$

ein, welche Distanzen zwischen den Werten von v_r für a und b misst. In diesem Sinne kann ein Paar (a, b) von Merkmalsträgern intuitiv als „guter" Kandidat für eine spätere Zuordnung gelten, wenn seine Distanzen $d_r(a, b)$ in allen Komponenten $r = 1, \ldots, k$ verhältnismäßig klein gegenüber den Distanzen der anderen möglichen Paarungen sind.

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

$$\widetilde{d}_r(a, b) := \frac{d_r(a, b) - \min_{(\alpha,\beta)\in A\times B} d_r(\alpha, \beta)}{\max_{(\alpha,\beta)\in A\times B} d_r(\alpha, \beta) - \min_{(\alpha,\beta)\in A\times B} d_r(\alpha, \beta)}. \tag{3.1}$$

Setzt sich der Wertebereich eines Merkmals aus endlichen und (abzählbar) unendlichen Mengen zusammen, so kann er in geeignete Teilmengen zerlegt werden, um danach eine individuell auf diese Teilmengen zugeschnittene Distanzfunktion anzuwenden (vgl. Bacher 1994).

3.2.1 Distanzmaße für kategoriale Merkmale

Bei den kategorialen Merkmalen kann zwischen *nominalen* und *ordinalen* Merkmalen unterschieden werden. Bei den ordinalen Merkmalen liegt ein linear geordneter Wertebereich vor. D.h., je zwei Ausprägungen x und y sind vergleichbar (es gilt $x < y$ oder $y \leq x$), wobei nicht gefordert wird, dass eine Differenzbildung zwischen Ausprägungen möglich oder sinnvoll ist. Bei nominalen Merkmalen wird hingegen keinerlei Ordnung auf dem Wertebereich unterstellt. Hier sei angemerkt, dass auch Merkmale mit einem teilweise geordneten Wertebereich, wie beispielsweise hierarchisch geordnete Merkmale, zu den nominalen Merkmalen gezählt werden.

Nachfolgend werden zwei Ansätze zur Bestimmung von Distanzmaßen für kategoriale Merkmale vorgestellt: der probabilistische und der deterministische Ansatz. Die praktische Erfahrung bei der Anonymisierung wirtschaftsstatistischer Einzeldaten hat gezeigt, dass bei kategorialen Merkmalen aufgrund der geringeren bzw. durch den Datennutzer einschätzbaren Einschränkungen des Analysegehalts stets informationsreduzierende Anonymisierungsmethoden angewendet werden. Der probabilistische Ansatz ist daher nur in Ausnahmefällen zu empfehlen, etwa bei einer Anwendung des datenverändernden Verfahrens PRAM. Theoretisch denkbar ist auch eine Kombination beider Ansätze (Lenz 2003c), die sich aber in der Praxis nicht bewährt hat.

Probabilistischer Ansatz

Wir definieren zwei komplementäre Mengen $M, U \subseteq A \times B$ derart, dass ein Paar (a, b) von Merkmalsträgern zu M gezählt wird, wenn a und b zu demselben Individuum gehören, d.h., wenn sie korrekt einander zugeordnet sind. Andernfalls gehört das Paar (a, b) zu U. In Jaro (1989) werden merkmalsweise probabilistische Gewichte eingeführt mittels bedingter Wahrscheinlichkeiten. Für jedes Merkmal v_i werden für ein zufälliges Paar (a, b) von Merkmalsträgern

$$m_i := P(a \text{ und } b \text{ stimmen in Merkmal } v_i \text{ überein} \mid (a, b) \in M)$$

und $$u_i := P(a \text{ und } b \text{ stimmen in Merkmal } v_i \text{ überein} \mid (a, b) \in U)$$

definiert. Für jedes Überschneidungsmerkmal, dass in dem betrachteten Paar (a, b) übereinstimmt, müssen nun die bedingten Wahrscheinlichkeiten m_i und u_i geschätzt und danach miteinander verglichen werden.

Mit a_i und b_i werden im Folgenden die Werte von v_i für die beiden Merkmalsträger abgekürzt. In Jaro (1989) wird im Falle $a_i = b_i$ das Ähnlichkeitsmaß

$$w_i(a, b) = log_2(\frac{m_i}{u_i}) \tag{3.2}$$

und im Falle $a_i \neq b_i$ das Ähnlichkeitsmaß

$$w_i(a, b) = log_2(\frac{1 - m_i}{1 - u_i}) \tag{3.3}$$

vorgeschlagen. Die komplementären Wahrscheinlichkeiten geben nun ein Maß für die in unserem Kontext benötigte Distanz zweier Merkmalsträger in diesem Merkmal. Im Falle $a_i = b_i$ gilt

$$\begin{aligned} d_i(a, b) &:= 1 - w_i(a, b) \\ &= 1 - log_2(\frac{m_i}{u_i}) \\ &= log_2 2 - log_2(\frac{m_i}{u_i}) \\ &= log_2(\frac{2 \cdot u_i}{m_i}). \end{aligned} \tag{3.4}$$

Analog ergibt sich im Falle $a_i \neq b_i$ die logarithmische Tranformation

$$d_i(a, b) := log_2(\frac{2 \cdot (1 - u_i)}{1 - m_i}), \tag{3.5}$$

wobei in beiden Fällen der Faktor 2 im Argument bedeutungslos ist.

Da in den meisten Fällen $m_i > u_i$ zu erwarten ist, liefern Merkmale, die in beiden betrachteten Merkmalsträgern übereinstimmen, eine kleinere Komponentendistanz $d_i(a, b)$

als solche, bei denen keine Übereinstimmung vorliegt. Dies ist insbesondere dann der Fall, wenn die kategorialen Überschneidungsmerkmale unverändert oder nur informationsreduzierenden Anonymisierungsmethoden unterworfen wurden.

Da die Menge M sehr klein gegenüber ihrer Komplementärmenge $U = (A \times B) \setminus M$ ist, stimmt die Mächtigkeit von U näherungsweise mit der von $A \times B$ überein und es gilt

$$u_i \approx P(a_i = b_i).$$

Daher kann u_i durch die relative Häufigkeit aller im Merkmal v_i übereinstimmenden Paare von Merkmalsträgern geschätzt werden. In Jaro (1998) werden zudem verschiedene Ansätze zur Schätzung von m_i vorgestellt (darunter der erstmals in Dempster et al. (1971) eingeführte EM -Algorithmus), die sich als wesentlich schwieriger darstellt und nur mittels geeigneter Unabhängigkeitsannahmen möglich ist.

Deterministischer Ansatz

Die Ausprägungen nominaler (nicht-ordinaler) Merkmale v_i können nur auf Gleichheit untersucht werden. Es wird also mit

$$d_i(a, b) = \begin{cases} 0, & \text{wenn } a_i = b_i, \\ 1 & \text{sonst} \end{cases} \tag{3.6}$$

ein geeignetes Distanzmaß für nominale Merkmale definiert.

Auch die Realisierung der Blockung von Daten mittels ausgewählter Blockmerkmale kann in die Distanzberechnung integriert werden. In der Regel werden nur kategoriale Merkmale zur Blockung verwendet. Ist dennoch ein metrisches Merkmal zur Blockung vorgesehen, dann wird strengstens empfohlen, dieses zunächst in Intervalle bzw. Kategorien zu vergröbern, sodass jede Ausprägung eindeutig einer Kategorie zuzuordnen ist. Danach wird die Distanz in diesem Merkmal genau dann auf Null gesetzt, wenn beide Merkmalsträger in dieselbe Kategorie fallen.

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

$$d_i(a, b) = \frac{|\{c_j \mid \min(a_i, b_i) \leq_i c_j <_i \max(a_i, b_i)\}|}{r} \tag{3.7}$$

als Distanzmaß. Da die Differenz zwischen zwei Kategorien bei ordinalen Merkmalen in der Regel bedeutungslos ist, werden in obigem Maß lediglich die dazwischen liegenden Kategorien gezählt.

Bei wirtschaftsstatistischen Einzeldaten kommt es oftmals vor, dass der Wertebereich W eines kategorialen Merkmals zwar partiell, aber nicht linear geordnet ist (d.h., es gibt wenigstens zwei Ausprägungen c_i und c_j mit $c_i \not\leq c_j$ und $c_j \not\leq c_i$). Hier sollte obige Formel (3.7) zur Distanzberechnung nicht verwendet werden. In solchen Fällen wird empfohlen, die partiell geordnete Menge zu einer verbandsgeordneten Menge zu erweitern, sodass für alle Paare sowohl Supremum $\sup\{c_i, c_j\}$ als auch Infimum $\inf\{c_i, c_j\}$ existieren.[18] Unter Verwendung einer geeigneten ordnungserhaltenden bzw. monotonen Abbildung $f : \{c_1, \ldots, c_r\} \longrightarrow \mathbf{R}$ kann nun die Distanz im Merkmal v_i definiert werden durch

$$d_i(a, b) := |f(\sup\{a_i, b_i\}) - f(\inf\{a_i, b_i\})|. \tag{3.8}$$

Ein Beispiel ist durch den für Zeichenketten oftmals verwendeten n-gram –Ansatz gegeben (vgl. Efelky et al. 2002). Die Distanz zwischen zwei Zeichenketten wird hierin definiert durch

$$d_i(a, b) = \sqrt{\sum_{\forall s} |f_a(s) - f_b(s)|}, \tag{3.9}$$

wobei $f_a(s)$ und $f_b(s)$ die Anzahl des Auftretens der Teilzeichenkette s der Länge n in den beiden Zeichenketten a und b bestimmen. Sei beispielsweise $n = 3$, a =SCHMIDT und b =SCHMITT. Wir erhalten in diesem Falle d_i(SCHMIDT,SCHMITT) $= \sqrt{4}$, da es vier nicht gemeinsame Teilzeichenketten der Länge 3 in beiden Zeichenketten gibt: MID, IDT, MIT und ITT.

In unserem Kontext können hierarchische Merkmale auch dort auftauchen, wo ein kategoriales Merkmal infolge einer informationsreduzierenden Anonymisierungsmaßnahme in verschiedene Gliederungstiefen vergröbert wurde. Zum Beispiel kann in einer Wirtschaftsstatistik das Merkmal *Wirtschaftszweigklassifikation* für einige Merkmalsträger auf Dreistellerebene (etwa mit der Ausprägung „101“) und für andere auf Zweistellerebene (etwa mit der Ausprägung „10“) ausgewiesen sein, siehe Abbildung 3.1.

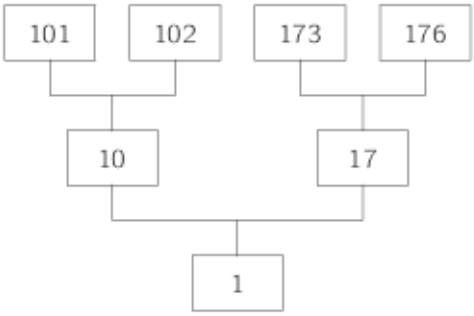
In solchen Beispielen ist die via Gleichung (3.8) berechnete Distanz der einfachen $(0-1)$-Distanz in (3.6) vorzuziehen, da letztere hier zu stark separierend wirkte.

3.2.2 Distanzmaße für metrische Merkmale

In diesem Abschnitt werden diskrete oder stetige Merkmale betrachtet, bei denen eine Differenzbildung von Ausprägungen möglich ist, zum Beispiel *Anzahl der Beschäftigten* oder *Gesamtumsatz* eines Unternehmens.

18 Die durch Erweiterung entstandene partiell geordnete Menge $(W; \leq)$ kann nun als algebraischer Verband $(W; \sup, \inf)$ aufgefasst werden, da die binären Operationen *sup* und *inf* die für einen Verband formulierten Axiome erfüllen (Birkhoff 1940).

Abbildung 3.1
Hierarchische Gliederung der Wirtschaftszweigklassifikation



Das für einen Datenangreifer naheliegende Distanzmaß ist durch $d_i(a,b) = |a_i - b_i|$ gegeben, sodass eine Aufsummierung aller komponentenweise definierten Distanzen zur *city-block* -Distanz

$$d^1(a,b) = \sum_{i=1}^{k} |a_i - b_i|, \tag{3.10}$$

führt, die einen Spezialfall der bekannten *Minkowski-Distanz* L_q darstellt, definiert durch

$$d^q(a,b) = \sqrt[q]{\sum_{i=1}^{k} |a_i - b_i|^q} \ . \tag{3.11}$$

Setzt man $q = 2$ und individuelle Gewichte $\lambda_1, \ldots, \lambda_k$, so erhält man eine leichte Modifikation der *euklidischen Distanz*

$$d(a,b) = \sqrt{\sum_{i=1}^{k} \lambda_i (a_i - b_i)^2}, \tag{3.12}$$

die sich über die Diagonalmatrix $D = diag(\lambda_1, \ldots, \lambda_k)$ und die Spaltenvektoren $a = (a_1, \ldots, a_k)^T$, $b = (b_1, \ldots, b_k)^T$ auch wie folgt schreiben lässt:

$$d(a,b) = \sqrt{(a-b)^T D (a-b)}. \tag{3.13}$$

In der Clusteranalyse findet die *Allgemeine Matrix Distanz*

$$d(a,b) = \sqrt{(a-b)^T C (a-b)} \tag{3.14}$$

große Beachtung, wobei C eine beliebige symmetrische, positiv definite Matrix ist. Allerdings kann diese Distanz bei großen Datenmengen einen inakzeptabel hohen Rechenaufwand bedeuten; insbesondere, wenn für C die Inverse S^{-1} der empirischen Kovarianzmatrix gewählt wird.

Eine mögliche Formalisierung der Gewichtung der komponentenweise definierten Distanzen wird im nachfolgenden Abschnitt vorgestellt.

3.3 Präferenzfunktionen und -zuordnungen

Im Folgenden werden vektorgewichtete bipartite Graphen $\mathcal{G}$ (zur Definition siehe Unterabschnitt 3.1.1) mit Bipartition $V(\mathcal{G}) = A \cup B$ betrachtet. Die Kanten werden vektoriell gewichtet mit den im vorherigen Abschnitt 3.2 besprochenen Distanzen d_i für die Überschneidungsmerkmale v_i, $i = 1 \ldots, k$. Es wird also für jede Kante $e \in E(\mathcal{G})$ der Distanzvektor

$$d(e) = (d_1(e), \ldots, d_k(e))$$

berechnet. Im Allgemeinen gibt es keine bzgl. aller Komponentendistanzen „optimale" perfekte Zuordnung, da keine lineare Ordnung auf der Menge aller reellen k-Tupel vorliegt. Sei nun $\mathcal{M}$ eine perfekte Zuordnung und $E(\mathcal{M}) \subseteq A \times B$ die zugehörige Kantenmenge. Wir definieren komponentenweise

$$d_i(\mathcal{M}) := \sum_{(a,b)\in E(\mathcal{M})} d_i(a, b), \; i = 1, \ldots, k \tag{3.15}$$

und weiterhin

$$d(\mathcal{M}) := \big(d_1(\mathcal{M}), \ldots, d_k(\mathcal{M})\big). \tag{3.16}$$

Wir bezeichnen mit $\prec_k$ die lexikographische Ordnung auf dem Vektorraum $\mathbb{R}^k$ und gelangen zu nachfolgender

Definition 1 *(Schweigert 1995) Eine perfekte Zuordnung $\mathcal{M}$ eines vektorgewichteten bipartiten Graphen wird* `effizient` *genannt, wenn keine weitere Zuordnung $\mathcal{M}'$ mit*

$$d(\mathcal{M}') \prec_k d(\mathcal{M})$$

existiert.

Im Sinne obiger Definition wird eine perfekte Zuordnung zwischen A und B demnach als effizient verstanden, wenn es keine perfekte Zuordnung gibt, die in allen Komponentendistanzen gleichwertig oder vorzuziehen ist. Solche Lösungen werden in der Literatur auch *Pareto-optimal* genannt. Bedauerlicherweise sind bereits in kleinen Beispielen mehrere effiziente perfekte Zuordnungen $\mathcal{M}$ mit paarweise nicht vergleichbaren Vektorgewichten $d(\mathcal{M})$ zu erwarten. Aus diesem Grunde ist der Entscheidungsträger bzw. in unserem Falle der Datengreifer gezwungen, Präferenzen unter den k Merkmalen zu setzen. Da der Erfolg der Zuordnungen im Wesentlichen von der Qualität der einzelnen Überschneidungsmerkmale abhängt, wird der Datenangreifer einige Merkmale den anderen vorziehen. Dieses Vorgehen kann mathematisch durch den Einsatz sogenannter Präferenzfunktionen realisiert werden:

Definition 2 *Sei $\Lambda = (\lambda_1, \ldots, \lambda_k) \in (\mathbb{R}^+)^k$ ein k-Tupel nicht-negativer reeller Zahlen. Für einen Merkmalsträger $r = (r^{(1)}, \ldots, r^{(s)})$, wobei s die Anzahl aller Merkmale beschreibe, seien o.B.d.A. die Einträge $r^{(1)}, \ldots, r^{(k)}$ die Werte der k Überschneidungsmerkmale. Wir definieren eine k-stellige* `lineare Präferenzfunktion` *$f_\Lambda : \mathbb{R}^k \longrightarrow \mathbb{R}$ durch*

$$f_\Lambda(x_1, \ldots, x_k) = \sum_{i=1}^{k} \lambda_i x_i. \tag{3.17}$$

Sei nun $(E(\mathcal{G}); \preceq_k)$ die teilweise geordnete Menge der vektorgewichteten Kanten von $\mathcal{G}$, versehen mit der lexikographischen Ordnungsrelation $\preceq_k$. Jede lineare Präferenzfunktion f_Λ induziert eine *Lineare Präferenzerweiterung* $(L_\Lambda(E(\mathcal{G})); \leq)$ durch $e_i \leq e_j$ genau dann, wenn

$$f_\Lambda(w_1(e_i), \ldots, w_k(e_i)) \leq f_\Lambda(w_1(e_j), \ldots, w_k(e_j))$$

gilt. Die so entstehende geordnete Menge $(L_\Lambda(E(\mathcal{G})); \leq)$ ist wohlgeordnet, da für alle $e_i, e_j \in E(\mathcal{G})$ entweder $e_i \leq e_j$ oder $e_j < e_i$ gilt.

Eine durch

$$\lambda_{\tau(1)} > \lambda_{\tau(2)} > \ldots > \lambda_{\tau(k)}$$

definierte Permutation τ kann als individuelle Priorisierung der Merkmale durch den Datenangreifer verstanden werden. Fordert man zusätzlich $\sum_{i=1}^{k} \lambda_i = 1$ und damit $\lambda_k = 1 - \sum_{i=1}^{k-1} \lambda_i$, so reduziert sich die Menge der Parameter um ein Element.

Definition 3 *Eine perfekte Zuordnung $\mathcal{M}$ wird* `Präferenzzuordnung` *genannt, wenn eine Präferenzfunktion f_Λ existiert mit $f_\Lambda(d(\mathcal{M})) \leq f_\Lambda(d(\mathcal{M}'))$ für jede perfekte Zuordnung $\mathcal{M}'$.*

Bereits ein einzelnes Paar (a, b) kann als (nicht perfekte) Zuordnung verstanden werden. Präferenzfunktionen bestimmen für jedes Paar $(a, b) \in A \times B$ eine Distanz

$$d(a, b) := f_\Lambda(\Delta(a, b)) = \sum_{i=1}^{k} \lambda_i \cdot d_i(a, b), \tag{3.18}$$

die weiterhin bzgl. ihrer mit kategorialen und metrischen Merkmalen verbundenen Kompenten aufgesplittet werden kann:

$$\begin{aligned} d(a, b) &= \sum_{i=1}^{k} \lambda_i \cdot d_i(a, b) \\ &= \tau \cdot \sum_{i \in CV} \tilde{\lambda}_i \cdot d_i(a, b) + \sum_{i \in NV} \lambda_i \cdot d_i(a, b), \end{aligned} \tag{3.19}$$

wobei CV die Indexmenge der kategorialen und NV die Indexmenge der metrischen Merkmale beschreibe. Mit dem Kontrollparameter $\tau = \lambda_i / \tilde{\lambda}_i$ kann der Einfluss der kategorialen Überschneidungsmerkmale auf das Gesamtmaß gesteuert werden.

Da die Bestimmung der Gewichte selten objektiv erfolgen kann, wird im Zweifel dazu geraten, eine Gleichgewichtung vorzunehmen, um ungewollte Verzerrungen zu vermeiden. Zur Beurteilung der Qualität von Überschneidungsmerkmalen müsste der Datenangreifer Abweichungsanalysen zwischen externer Datei und Zieldatei durchführen, was ohne das Vorhandensein direkter Identifikatoren wie Name und Adresse nicht möglicht ist. Allerdings könnte der Datenangreifer testen, ob einige Überschneidungsmerkmale in der Zieldatei stärker von den bei einem Datenangebot üblicherweise in den zugehörigen Metadaten beschriebenen Anonymisierungsmaßnahmen betroffen sind als andere (vgl. hierzu die späteren Ausführungen in den Abschnitten 3.6 und 6.3).

Wie bereits in Unterabschnitt 3.1.2 erwähnt wurde, kann die Blockung von Merkmalen in die Distanzberechnung integriert werden (Lenz 2003a). Hierzu werden die Distanzen $d(a, b)$ aufgesplittet nach den durch die Blockmerkmale (BV) und die übrigen Überschneidungsmerkmale (MV) gegebenen Komponentendistanzen:

$$d(a,b) = \sum_{i \in BV} d_i(a,b) + \sum_{i \in MV} \lambda_i \cdot d_i(a,b). \tag{3.20}$$

Die Blockmerkmale v_i werden dabei mit $\lambda_i = 1$ gewichtet. Stimmen zwei Merkmalsträger a und b in allen Blockmerkmalen überein, so ergibt die erste Summe den Wert Null. Ersetzt man nun für jedes der übrigen Überschneidungsmerkmale v_i den Parameter λ_i durch $\lambda_i / \sum_{j \in MV} \lambda_j$, so erhält man eine Konvexkombination der Komponentendistanzen, bei der die zweite Summe einen Wert kleiner oder gleich Eins aufweist. Mit anderen Worten wird zwei Merkmalsträgern a und b eine Distanz $d(a, b) \leq 1$ genau dann zugewiesen, wenn sie zu demselben Block gehören. Bei großen Datenmengen wird zusätzlich empfohlen, die errechneten Distanzen $d(a, b)$ mit einer vorgegebenen Schwelle $c < 1$ zu vergleichen und Kandidatenpaare oberhalb der Schwelle a priori zu verwerfen, d.h. von dem späteren Zuordnungsverfahren auszuschließen.

Eine alternative Realisierung der Blockung kann über blockweises Einlesen der Daten nach vorheriger Sortierung der Daten nach den Blockmerkmalen erfolgen. Die Erfahrung hat gezeigt, dass Letzteres bei Daten von moderater Größenordnung gut möglich ist. Es muss allerdings erwähnt werden, dass das durchaus zeitaufwendige blockweise Einlesen bei den üblichen theorieorientierten Komplexitätsanalysen für Algorithmen vernachlässigt wird.

3.4 Formulierung als lineares Zuordnungsproblem

Das Ziel eines Datenangreifers kann darin bestehen, seine bereits vorhandene externe Datenbank aufzubessern. Ein simultanes Zuspielen von Informationen aus den vertraulichen Daten zu möglichst vielen Merkmalsträgern der externen Daten ist durch einen Massenfischzug realisierbar.

3.4.1 Parametrisierung

Auf Basis der zuvor berechneten 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. Aus technischen Gründen nehmen wir im Folgenden $n = |A| = |B| = m$ an.
[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 Distanzvektoren auf

$$\left(\max_{(a,b)\in A\times B} d_1(a_i, b_j), \max_{(a,b)\in A\times B} d_2(a_i, b_j), \ldots, \max_{(a,b)\in A\times B} d_k(a_i, b_j) \right)$$

gesetzt werden.]

Die entsprechende Vektoroptimierungsaufgabe (Lenz 2003b; Lenz 2006a) ist hier von der Form

$$\text{Minimiere} \quad \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{n} d_1(a_i, b_j)x_{ij} \\ \sum_{i=1}^{n}\sum_{j=1}^{n} d_2(a_i, b_j)x_{ij} \\ \qquad\vdots \\ \sum_{i=1}^{n}\sum_{j=1}^{n} d_k(a_i, b_j)x_{ij} \end{cases} \tag{3.21}$$

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

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.

Unter Verwendung der über Präferenzfunktionen ermittelten Distanzen wird das Vektoroptimierungsproblem (3.21) in ein Zuordnungsproblem (3.22) mit einer Zielfunktion transformiert. Die wesentliche Idee besteht darin, alle Zielfunktionen zu einer einzigen zu minimierenden Funktion zu aggregieren. Dieser lineare Ansatz ist im Allgemeinen mit einem beachtlichen Informationsverlust verbunden. Die Problematik einer geeigneten Parametrisierung wird oftmals unterschätzt, ist aber von theoretischer Seite sehr schwierig lösbar. In Schweigert (1999) wird gezeigt, dass es unter gewissen Bedingungen für den Entscheidungsträger ausreicht, Intervalle für die einzelnen Parameter bzw. Gewichte anzugeben.

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

$$\text{Minimiere} \quad \sum_{i=1}^{n}\sum_{j=1}^{n} d_{ij}x_{ij} \tag{3.22}$$

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

Die graphentheoretische Formulierung des Problems lautet: Finde eine Präferenzzuordnung auf einem vektorgewichteten bipartiten Graphen! Die Aufgabe besteht also mit anderen Worten darin, bei bekannten Präferenzen eine Permutation π über der Menge $\{1,\ldots,n\}$ zu finden, welche die Summe $\sum_{i=1}^{n} d_{i,\pi(i)}$ (d.h., die Gesamtdistanz aller getroffenen Zuordnungen) minimiert.

An dieser Stelle sei erwähnt, dass jede Präferenzzuordnung effizient ist und umgekehrt (Isermann 1974). Anders formuliert bedeutet dies: Zu jeder Pareto-optimalen Lösung existiert ein Gewichtsvektor, sodass diese ebenfalls eine Lösung des zugehörigen parametrischen Optimierungsproblems (3.22) ist; umgekehrt ist bei beliebiger Wahl der Gewichte die Lösung von (3.22) gleichzeitig Pareto-optimal und damit in gewissem Sinne bestmöglich. Im Allgemeinen muss aber festgestellt werden, dass durch lineare Präferenzerweiterungen möglicherweise wertvolle Informationen verloren gehen und die korrekte Zuordnung, bei der die Merkmalsträger der externen Daten allesamt korrekt den Merkmalsträgern der Zieldaten zugeordnet werden, über den Ansatz der Priorisierung der Merkmale von theoretischer Seite nicht garantiert werden kann. Dies hängt aber nicht allein von der Wahl der Gewichte, sondern vielmehr von der Qualität der berechneten Distanzen bzw. der der Distanzberechnung zugrundeliegenden Überschneidungsmerkmale ab. Theoretisch ist es zudem möglich, dass verschiedene Präferenzzuordnungen $\mathcal{M}$ und $\mathcal{M}'$ dieselbe Gesamtdistanz $f_\Lambda(d(\mathcal{M})) = f_\Lambda(d(\mathcal{M}'))$ aufweisen. Man spricht dann von *äquivalenten Zuordnungen*. Bei zufälliger Parametersetzung ist dieser Fall allerdings von maßtheoretischer Seite ausgeschlossen.

Der naivste und rechentechnisch aufwendigste Weg zu einer Lösung von (3.22) wäre sicher, alle $n!$ möglichen perfekten Zuordnungen durchzuprobieren und diejenige mit der geringsten Gesamtdistanz auszuwählen. Obgleich es klassische Verfahren gibt, wie z.B. die bekannte Simplex-Methode (vgl. Papadimitriou und Steiglitz 1998), traten in dem vorliegenden Falle bereits Probleme bei der Arbeit mit (für Wirtschaftsstatistiken) verhältnismäßig kleinen Datenmengen der Größenordnung von 20 000 Einheiten auf.

Die in (3.21) und (3.22) formulierten Nebenbedingungen stellen ein lineares Gleichungssystem mit folgender Matrix dar:

$$A = \begin{pmatrix} J_1 & J_2 & \cdots & J_n \\ I_n & I_n & \cdots & I_n \end{pmatrix},$$

wobei $I_n = diag(1, 1, \ldots, 1)$ die Einheitsmatrix der Dimension $n \times n$ definiere und J_i, $i = 1, \ldots, n$, eine Matrix derselben Dimension ist, deren i-ter Zeilenvektor durch $(1\,1 \cdots 1)$ gegeben ist und deren verbleibende Einträge gleich Null sind. Allein diese Koeffizientenmatrix, bestehend aus $2n$ Zeilen und n^2 Spalten, kann zum Überschreiten des Arbeitsspeichers eines handelsüblichen PC führen. Die Matrix A ist total unimodular, d.h., es gilt $det\tilde{A} \in \{-1, 1\}$ für eine maximale Teilmatrix $\tilde{A}$ und $detA' \in \{-1, 0, 1\}$ für eine beliebige quadratische Teilmatrix A' von A. Definiert man mit $D = (d_{11}\, d_{12} \cdots d_{nn})$ und $\mathbf{x} = (x_{11}\, x_{12} \cdots x_{nn})^T$ zwei Vektoren mit n^2 sowie mit $\mathbf{1} = (1\,1 \cdots 1)^T$ einen Vektor mit $2n$ Komponenten, so ist es möglich, das Problem in allgemeiner Form

$$\min\{D\mathbf{x} \mid A\mathbf{x} = \mathbf{1},\ \mathbf{x} \geq 0\}$$

zu formulieren. Aufgrund der totalen Unimodularität von A sind ganzzahlige Lösungen dieses Problems garantiert und der weit verbreitete Simplex-Algorithmus ist damit anwendbar. Hierzu folgt ein kleiner Beweis.

[**Beweis:**
Aus der linearen Algebra ist bekannt, dass sich $A\mathbf{x} = \mathbf{1}$ auf ein (kleineres) lineares Gleichungssystem $B\mathbf{x} = \mathbf{1}$ reduzieren lässt, wobei B eine quadratische Teilmatrix vollen Ranges von A ist. Es folgt

$$x = B^{-1}\mathbf{1} = \frac{1}{detB} \cdot \mathcal{B}_{adj} \cdot \mathbf{1}.$$

Aus der totalen Unimodularität von A und der Kenntnis, dass sämtliche Einträge der adjungierten Matrix $\mathcal{B}_{adj}$ durch Determinanten von Teilmatrizen von B und damit von A gebildet werden, ergeben sich mit $detB \in \{-1, 1\}$ und $(\mathcal{B}_{adj})_{ij} \in \{-1, 0, 1\}$, $i, j = 1, \ldots, n$, für $\mathbf{x}$ nur ganzzahlige Komponenten.]

Obwohl der Simplex-Algorithmus im schlechtesten Falle exponentielle Laufzeit aufweist, hat er sich bei der empirischen Arbeit mit wirtschaftsstatistischen Einzeldaten als sehr effizient erwiesen. Dies liegt vermutlich daran, dass die erwartete Laufzeit des Algorithmus polynomial ist (Borgwardt 1982; Borgwardt 1987). Eine Verbesserung der Komplexität kann durch die Anwendung der sogenannten *Ungarischen Methode* erreicht werden, 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. Im Folgenden wird eine Kurzbeschreibung der Methode gegeben, die ursprünglich für perfekte Zuordnungen maximalen Gewichts formuliert wurde (Kuhn 1955; Munkres 1957). Sei hierzu $\mathcal{G} = (V, E)$ ein vollständiger, gewichteter bipartiter Graph. Eine *zulässige Eckennumerierung* l ist eine Abbildung von der Menge V in die reellen Zahlen mit der Eigenschaft

$$l(a) + l(b) \leq d(a, b).$$

Die Zahl $l(v)$ wird dann *Label* von v genannt. Der *Gleichheitsteilgraph* $\mathcal{G}_l$ ist ein Teilgraph von $\mathcal{G}$, der alle Ecken von $\mathcal{G}$ enthält aber nur solche Kanten (a, b), welche

$$l(a) + l(b) = d(a, b)$$

erfüllen. Eine Verbindung zwischen Gleichheitsteilgraphen und perfekten Zuordnungen minimalen Gewichts stellt nachfolgender Satz her.

Satz *Sei l eine zulässige Eckennumerierung von $\mathcal{G}$. Besitzt der Gleichheitsteilgraph $\mathcal{G}_l$ eine perfekte Zuordnung $\mathcal{M}$, so ist $\mathcal{M}$ von minimalem Gewicht.*

[**Beweis:**
Sei $\mathcal{M}$ eine perfekte Zuordnung von $\mathcal{G}_l$ und $\mathcal{M}'$ eine beliebige perfekte Zuordnung von $\mathcal{G}$. Dann folgt

$$\begin{aligned} d(\mathcal{M}') := \sum_{(a,b)\in\mathcal{M}'} d(a,b) &\geq \sum_{v\in V(\mathcal{G})} l(v) \quad \text{(da in } \mathcal{M}' \text{ alle Ecken gesättigt sind)} \\ &= \sum_{(a,b)\in\mathcal{M}} d(a,b) \qquad \text{(nach Definition von } \mathcal{M}) \\ &=: d(\mathcal{M}). \end{aligned}$$

Somit ist $\mathcal{M}$ eine perfekte Zuordnung minimalen Gewichts.]

Bei der Anwendung des Algorithmus werden die beiden Vektoren $(l(a_1), \ldots, l(a_n))$ und $(l(b_1), \ldots, l(b_n))$ von Eckennumerierungen verwendet mit dem Ziel, zulässige Kanten auszuwählen. Zu Beginn werden

$$\begin{aligned} & l(a_i) = 0 && \text{für } i = 1, \ldots, n \\ \text{und} \quad & l(b_j) = \min_{1\leq i\leq n} d(a_i, b_j) && \text{für } j = 1, \ldots, n \end{aligned}$$

gesetzt. Durch die Konstruktion sogenannter $\mathcal{M}$-Verbesserungswege (engl. augmenting paths) findet man eine Zuordnung $\mathcal{M}$ auf dem Graphen $\mathcal{G}_l$, die so viele Ecken wie möglich sättigt. Ist $\mathcal{M}$ perfekt, so hat $\mathcal{M}$ nach obigem Satz minimales Gewicht und der Agorithmus endet hier. $\mathcal{M}$ ist dann eindeutig bestimmt bis auf Äquivalenz. Andernfalls, wenn $\mathcal{M}$ nicht perfekt ist, wird die Eckennumerierung für einige Werte $l(a)$ und $l(b)$ erneuert, sodass neue Kanten zulässig werden. Ein konkurrierender Ansatz (engl. auction algorithm) zur Lösung des Zuordnungsproblems wurde in Bertsekas (1979) entwickelt und später in Bertsekas und Castanon (1989) auf allgemeine Transportprobleme erweitert.

3.4.2 Zuordnungsalgorithmus

In diesem Abschnitt wird basierend auf den bisherigen Überlegungen ein Algorithmus zur Simulation von Massenfischzügen vorgestellt. Zunächst werden zwei Näherungsheuristiken

betrachtet, die nach vorheriger Distanzberechnung eine perfekte Zuordnung von möglichst geringer Gesamtdistanz zwischen der externen Datei und der Zieldatei bestimmen. Sogenannte Greedy-Heuristiken werden in der Praxis oftmals aufgrund ihrer einfachen Implementierung und des gegenüber Verfahren zur Bestimmung einer optimalen Lösung geringeren Rechenaufwandes vorgezogen (vgl. Cormen et al. 1990).

Obwohl die beiden unten aufgeführten Greedy-Heuristiken nur Näherungslösungen für das lineare Programm (3.22) liefern, wurden sie aufgrund ihres zweifellosen Vorteils der guten Effizienz herangezogen und ihr Ergebnis mit der optimalen Lösung des linearen Programmes verglichen; eine umfangreiche empirische Vergleichsstudie findet sich in Lenz (2003a). 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.[19]

Nachfolgende Näherungsheuristik „Prozedur I" wird in der Literatur oft verwendet (vgl. Pagliuca und Seri 1976; Domingo-Ferrer und Mateo-Sanz 2001; Brand 2002):

Prozedur I:

`Beginn` {PROZ I}

$\mathcal{M} := \emptyset$

$i := 1$

`Solange` ($i \leq n$ `und` $B \neq \emptyset$) `führe aus:`

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

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

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

$i := i + 1$

`Ende` {PROZ 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 der externe Merkmalsträger a_{r+1} dem Merkmalsträger b in den Zieldaten mit minimalem Abstand zu a_{r+1} zugeordnet. Dabei ist b einer der verbliebenen Merkmalsträger in B, welche 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

19 D.h., es darf von einem „pseudo-quadratischen" Aufwand gesprochen werden.

Zuordnung nicht mehr eindeutig wäre. Eine solche Heuristik könnte als Simulation für einen Einzelangriff dienen, allerdings mit der Einschränkung, dass im Zusatzwissen einige, allein dem Einzelangreifer individuell verfügbare Überschneidungsmerkmale fehlen.

Das in Testsimulationen gegenüber der optimalen Lösung des linearen Programmes (3.22) 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: `Beginn` {PROZ II}

Sortiere die Distanzen aufsteigend in die Liste $\mathcal{L}$

`Solange` $\mathcal{L}$ nichtleer ist `führe aus:`

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

Entferne alle Elemente d_{rs} mit $r = i$ oder $s = j$ aus $\mathcal{L}$.

`Ende` {PROZ II}

Die größte Schwäche obiger Prozeduren ist zweifelsohne, dass nicht sämtliche Merkmalsträger simultan zugeordnet werden, wie dies bei der optimalen Lösung des Zuordnungsproblems der Fall ist. Die Erfahrung bei der Anwendungen der Prozeduren auf reale Wirtschaftsdaten hat jedoch gezeigt, dass Prozedur II Ergebnisse nahe der optimalen Lösung liefert (Lenz 2003a). Darüber hinaus ist es nahezu unmöglich, für große Dateien (wie z.B. die amtliche Umsatzsteuerstatistik bestehend aus fast drei Millionen Merkmalsträgern) in angemessener Zeit die optimale Lösung zu bestimmen. Hinzu kommt, dass im Rahmen des diskursiven Prozesses zwischen Wissenschaft und amtlicher Statistik in der Regel sehr viele Testläufe mit probeweise anonymisierten Dateien erforderlich sind (siehe Kapitel 8).

Zusammenfassend wird folgender Algorithmus zur Durchführung von Massenfischzügen vorgeschlagen:

Zuordnungsalgorithmus

1) EINGABE: 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) Zerlegung des Problems in Teilprobleme unter Verwendung der Blockmerkmale $BV \subseteq V$.

3) Berechnung der Komponentendistanzen $d_r(a_i, b_j)$ für $r = 1, \ldots, k$ zur Konstruktion eines vektorgewichteten bipartiten Graphen $\mathcal{G}$.

4) Parametrisierung des Vektoroptimierungsproblemes durch Setzung individueller Gewichte $\Lambda = (\lambda_1, \ldots, \lambda_k)$.

5) Alternative Anwendung

- der Ungarischen Methode oder
- der Näherungsheuristik Prozedur II.

6) AUSGABE: Injektive Abbildung $\varphi : A \longrightarrow B$ (eindeutige Zuordnungen).

3.4.3 Illustratives Beispiel

Zur Verdeutlichung des Algorithmus betrachten wir ein kleines durch untenstehende Tabelle 3.1 gegebenes Beispiel. Wir versuchen, vier Merkmalsträger in $A = \{a_1, \ldots, a_4\}$ mit vier Merkmalsträgern in $B = \{b_1, \ldots, b_4\}$ korrekt zu verknüpfen. Die korrekte, dem Datenangreifer unbekannte Zuordnung laute

$$\mathcal{M} = \{(a_1, b_3), (a_2, b_2), (a_3, b_1), (a_4, b_4)\}.$$

Als Überschneidungsmerkmale stehen dem Datenangreifer fünf metrische Merkmale $v_1, \ldots, v_5$ zur Verfügung.

Tabelle 3.1: Merkmalsträger und Merkmale am Beispiel

M-träger/ Merkmale	v_1	v_2	v_3	v_4	v_5
a_1	14 008 906	755 187	907 264	6 582 133	4 794 809
a_2	14 309 437	673 189	1 179 713	8 111 720	5 407 676
a_3	14 330 083	567 300	920 065	4 871 720	1 667 078
a_4	14 780 637	567 553	1 026 861	5 313 029	3 654 241
b_1	14 825 332	563 928	913 631	4 978 410	1 711 353
b_2	14 045 802	724 071	1 040 229	7 064 023	5 078 378
b_3	13 945 802	682 110	973 631	7 378 984	508 494
b_4	14 996 199	563 928	1 050 673	5 252 164	3 871 084

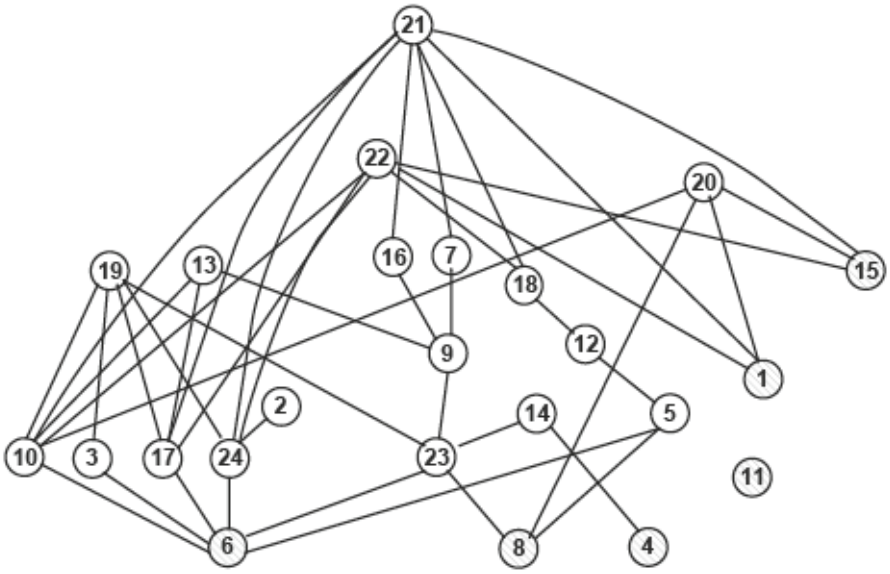
Eine Aufzählung aller 24 möglichen perfekten Zuordnungen zwischen A und B liefert Abbildung 3.3 auf Seite 75.

Ordnet man die 24 möglichen Zuordnungen nach ihren vektoriellen Gewichten – zu jeder Zuordnung gehört der Vektor der fünf Komponentendistanzen –, so ergibt sich die in Abbildung 3.2 abgebildete teilweise geordnete Menge perfekter Zuordnungen, lexikographisch nach den Gewichtsvektoren

$$\Delta(\mathcal{M}) = (d_1(\mathcal{M}), \ldots, d_5(\mathcal{M})) = \left(\sum_{(a_i,b_j)\in\mathcal{M}} d_1(a_i, b_j), \ldots, \sum_{(a_i,b_j)\in\mathcal{M}} d_5(a_i, b_j) \right)$$

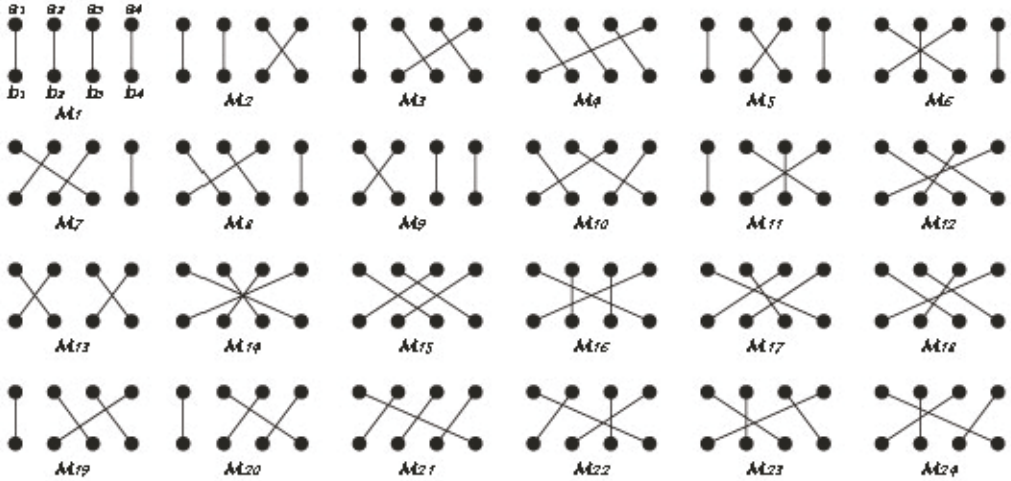
geordnet.

Abbildung 3.2: Teilweise geordnete Menge vektorgerwichteter perfekter Zuordnungen



Die minimalen Elemente $\mathcal{M}_1, \mathcal{M}_4, \mathcal{M}_6, \mathcal{M}_8, \mathcal{M}_{11}$ und $\mathcal{M}_{15}$ dieser teilweise geordneten Menge stellen die effizienten Lösungen dar. Nähme man Abbildung 3.4.3 als Basis für die Entscheidung des Datenangreifers, so würde dieser intuitiv $\mathcal{M}_6$ auswählen, da diese Zuordnung von allen nicht-effizienten Zuordnungen dominiert wird. In der Praxis ist es aber in der Regel nicht möglich, alle effizienten Zuordnungen (geschweige denn die vollständige teilweise geordnete Menge) anzuschauen. Daher wird die Entscheidung durch den Übergang von Problem (3.21) zu (3.22) wie in Abschnitt 3.4.1 beschrieben durch den Einsatz einer Präferenzfunktion f_Λ realisiert. Eine Gleichgewichtung aller Merkmale würde beispielsweise durch $\Lambda := (\frac{1}{5}, \ldots, \frac{1}{5})$ erreicht. In diesem Falle würde keine Zielfunktion in (3.21) gegenüber den anderen bevorzugt.

Abbildung 3.3
Perfekte Zuordnungen am Beispiel



In nachfolgender Tabelle werden die Ergebnisse der beiden Näherungsheuristiken mit der **U**ngarischen **M**ethode, d.h. der optimalen Lösung des linearen Programmes, verglichen. Der erste Eintrag einer jeden Zelle steht für die Anzahl korrekt gebildeter Paare, der zweite für die zugehörige Gesamtdistanz.

PROZ I	PROZ II	UM
2; 2,88	4; 2,38	4; 2,38

Unter Anwendung der Ungarischen Methode führen nahezu alle Gewichtungen zur Zuordnung $\mathcal{M}_6$. Allerdings können auch die restlichen effizienten Zuordnungen durch geeignete Gewichtswahl erreicht werden; z.B. führen $\Lambda_1 = (\frac{1}{3}, 0, \frac{2}{3}, 0, 0), \Lambda_4 = (0, 1, 0, 0, 0), \Lambda_8 = (0, 0, 0, 1, 0), \Lambda_{11} = (0, 0, 1, 0, 0)$ und $\Lambda_{15} = (0, 0, \frac{2}{3}, 0, \frac{1}{3})$ zu den effizienten Zuordnungen $\mathcal{M}_1, \mathcal{M}_4, \mathcal{M}_8, \mathcal{M}_{11}$ und $\mathcal{M}_{15}$.

Die Wahl der Parameter erfolgt individuell durch den Datenangreifer. Gibt es große Abweichungen zwischen externen Daten und Zieldaten im Überschneidungsmerkmal v_i, so wird er λ_i schwächer gewichten. Eine Einschätzung über solche Dateninkompatibilitäten ist sehr schwierig; nützliche Informationen können möglicherweise aus veröffentlichten Aggregaten beider Erhebungen – im Falle der Zieldaten zum Beispiel aus einer Fachserie des Statistischen Bundesamtes stammend – entnommen werden.

3.5 Massenfischzug versus Einzelangriff

Ein Datenangreifer kann verschiedene Beweggründe für den Datenangriff haben. Interessiert er sich für einen speziellen Merkmalsträger bzw. ein bestimmtes Unternehmen, so hat er das alleinige Ziel, zusätzliche Informationen über diesen zu sammeln. Das hiermit verbundene Szenario ist der in Abschnitt 2.2 beschriebene Einzelangriff. Ist $a = (a^{(1)}, \ldots, a^{(n)})$ der interessierende Merkmalsträger in den externen Daten und sind

$$b_j = (b_j^{(1)}, \ldots, b_j^{(n)}),\ j = 1, \ldots, m,$$

die Kandidaten für eine Zuordnung in den Zieldaten, sowie $d(a, b_j)$ ein Distanzmaß zwischen a und b_j, so ist die Optimierungsaufgabe

$$\min d(a, b_j),\ j = 1, \ldots, m \tag{3.23}$$

zu lösen. Betrachtet man die komponentenweisen Distanzen, so entsteht nachfolgendes multikriterielles Optimierungsproblem (3.24), welches mit derselben Argumentation wie

beim Massenfischzug (3.21) im Allgemeinen zu Zielkonflikten führt.

$$\text{Minimiere} \quad \begin{cases} \sum_{j=1}^{m} d_1(a, b_j)x_j \\ \sum_{j=1}^{m} d_2(a, b_j)x_j \\ \vdots \\ \sum_{j=1}^{m} d_k(a, b_j)x_j \end{cases} \tag{3.24}$$

$$\begin{aligned} \text{unter} \quad & x_j \in \{0, 1\} \quad \text{für} \quad j = 1, \ldots, m, \\ & \sum_{j=1}^{m} x_j = 1. \end{aligned}$$

D.h., der gesuchte Merkmalsträger a wird genau einem Zielmerkmalsträger b_j zugeordnet und in diesem Falle gilt $x_j = 1$.

Soll, etwa im Rahmen der Testrechnungen mit probeweise anonymisierten Zieldaten, für mehrere oder alle gesuchten Merkmalsträger a_i ein Einzelangriff simuliert werden, so stellt sich folgende Minimierungsaufgabe:

$$\text{Minimiere} \quad \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{m} d_1(a_i, b_j)x_{ij} \\ \sum_{i=1}^{n}\sum_{j=1}^{m} d_2(a_i, b_j)x_{ij} \\ \vdots \\ \sum_{i=1}^{n}\sum_{j=1}^{m} d_k(a_i, b_j)x_{ij} \end{cases} \tag{3.25}$$

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

Der wesentliche Unterschied zwischen (3.21) und (3.25) besteht in einer Reduktion der Nebenbedingungen. Im Gegensatz zum Massenfischzug sind hier auch Mehrfachzuordnungen erlaubt: es dürfen mehrere a_i ein und demselben b_j zugeordnet werden. Anders ausgedrückt sind bei der Simulation mehrerer Einzelangriffe auch Zuordnungen der Form $n : 1$ erlaubt.

Zudem ist es beim Einzelangriff gegenüber dem Massenfischzug technisch nicht nötig, auf einer Seite künstliche Merkmalsträger hinzuzufügen, um auf beiden Seite Parität herzustellen; es darf also $n \neq m$ gelten. Entsprechend zu Abschnitt 3.4.1 kann auch das multikriterielle Zuordnungsproblem (3.25) wieder parametrisiert werden und nachfolgende Näherungsheuristik Prozedur III, die durch eine leichte Modifikation von Prozedur I erhältlich ist, zum Einsatz kommen:

Prozedur III: `Beginn` {PROZ III}

$\mathcal{M} := \emptyset$

$i := 1$

`Solange` $i \leq n$ `führe aus:`

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

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

$i := i + 1$

`Ende` {PROZ III}

Obige Prozedur III realisiert wiederholte Simulationen von Einzelangriffen. Der Anteil korrekter Zuordnungen ist danach ein Schätzer für das mit einem einzelnen Merkmalsträger verbundene Reidentifikationsrisiko.

Die praktische Erfahrung hat gezeigt, dass dem Datenangreifer bei der manuellen Suche nach einem bestimmten Objekt oder Individuum nur die Möglichkeit bleibt, (nicht notwendigerweise symmetrische) Intervalle um die Ausprägungen der Überschneidungsmerkmale im Zusatzwissen zu legen und im Zieldatenbestand diesen k-dimensionalen Quader herauszufiltern. Prinzipiell stehen nach der Filterung dieselben Distanzmaße wie bei einem Massenfischzug zur Verfügung.

3.6 Sensitivität bei der Parametersetzung

Nach dem Aufbau der externen Datenbank im Falle eines Massenfischzugs oder der Recherche des Zusatzwissens über einen bestimmten Merkmalsträger im Falle eines Einzelangriffs besteht der einzig individuelle Einfluss des Datenangreifers in unseren Modellen in der Setzung der Koeffizienten der Zielfunktion des linearen Programmes, sprich der Gewichte für die einzelnen Überschneidungsmerkmale.

Die Problematik soll an einem kleinen Beispiel verdeutlicht werden. Die externe Datenbank und die Zieldatei enthalten jeweils zwei Merkmalsträger, d.h. $A = \{a_1, a_2\}$ und $B = \{b_1, b_2\}$. Desweiteren stehen zwei metrische Überschneidungsmerkmale v_1 und v_2 zur Verfügung. Die zu minimierende Zielfunktion

$$\begin{aligned} \mathcal{Z} &= \sum_{i=1}^{n}\sum_{j=1}^{m}\sum_{l=1}^{k}\lambda_l d_{ij}^{(l)} \\ &= \sum_{l=1}^{k}\lambda_l \sum_{i=1}^{n}\sum_{j=1}^{m} d_{ij}^{(l)} \end{aligned} \tag{3.26}$$

reduziert sich im vorliegenden Beispiel auf

$$\lambda_1(d_{11}^{(1)}x_{11} + d_{12}^{(1)}x_{12} + d_{21}^{(1)}x_{21} + d_{22}^{(1)}x_{22}) + \lambda_2(d_{11}^{(2)}x_{11} + d_{12}^{(2)}x_{12} + d_{21}^{(2)}x_{21} + d_{22}^{(2)}x_{22}).$$

Wir können ohne Beschränkung der Allgemeinheit eine Konvexkombination unterstellen, also $\lambda_1 + \lambda_2 = 1$, $\lambda_i \geq 0$ für $i = 1, 2$, und haben somit die Zielfunktion

$$\lambda(d_{11}^{(1)}x_{11} + d_{12}^{(1)}x_{12} + d_{21}^{(1)}x_{21} + d_{22}^{(1)}x_{22}) + (1-\lambda)(d_{11}^{(2)}x_{11} + d_{12}^{(2)}x_{12} + d_{21}^{(2)}x_{21} + d_{22}^{(2)}x_{22})$$

zu minimieren. Im Falle $\lambda = 1$ würde allein das Merkmal v_1 bei der Zuordnung berücksichtigt, im Falle $\lambda = 0$ allein das Merkmal v_2. Eine Gleichgewichtung würde durch $\lambda = \frac{1}{2}$ erreicht.

In Abhängigkeit von den Distanzen und dem Parameter λ muss sich der Datenangreifer für eine der Zuordnungsmengen $\mathcal{Z}_1 = \{(a_1, b_1), (a_2, b_2)\}$ und $\mathcal{Z}_2 = \{(a_1, b_2), (a_2, b_1)\}$ entscheiden. Diese sind mit den komplementären Lösungen $(x_{11}, x_{12}, x_{21}, x_{22}) = (1, 0, 0, 1)$ und $(x_{11}, x_{12}, x_{21}, x_{22}) = (0, 1, 1, 0)$ des linearen Programmes verbunden. Indifferenz besteht, wenn die Zielfunktion in beiden Situationen denselben Wert ermittelt:

$$\lambda_0(d_{11}^{(1)} + d_{22}^{(1)}) + (1-\lambda_0)(d_{11}^{(2)} + d_{22}^{(2)}) = \lambda_0(d_{12}^{(1)} + d_{21}^{(1)}) + (1-\lambda_0)(d_{12}^{(2)} + d_{21}^{(2)})$$

$$\Longleftrightarrow \lambda_0 = \frac{d_{12}^{(2)} + d_{21}^{(2)} - d_{11}^{(2)} - d_{21}^{(2)}}{d_{12}^{(2)} + d_{21}^{(2)} - d_{11}^{(2)} - d_{21}^{(2)} + d_{11}^{(1)} + d_{22}^{(1)} - d_{12}^{(1)} - d_{21}^{(1)}}. \tag{3.27}$$

Die Entscheidung ist also unsicher, wenn λ „nahe" bei λ_0 liegt. Wird $\lambda > \lambda_0$ gewählt, so fällt die Entscheidung für $\mathcal{Z}_1$, im anderen Falle für $\mathcal{Z}_2$ aus. Der Spezialfall $\lambda_0 = 1$ tritt u.a. dann auf, wenn die Distanzen in der ersten Komponente in beiden Alternativen identisch sind, wenn also $d_{11}^{(1)} + d_{22}^{(1)} - d_{12}^{(1)} - d_{21}^{(1)} = 0$ gilt, da in diesem Falle die zweiten Komponenten bei Indifferenz nicht zu berücksichtigen sind. Stimmen umgekehrt die Distanzen in der zweiten Komponente überein, so ergibt sich entsprechend $\lambda_0 = 0$. Stimmen beide Komponentendistanzen in beiden zulässigen Lösungen überein, so besteht unabhängig von der Wahl des Parameters λ Gleichgültigkeit bei der Entscheidungsfindung.

Die hier beschriebene Situation ist in der Praxis bereits entstanden bei vorheriger Partitionierung bzw. Blockung beider Datenquellen via kategoriale Merkmale wie zum Beispiel *Wirtschaftszweigklassifikation, Regionszugehörigkeit* und *Beschäftigtengrößenklasse* der untersuchten Unternehmen (insbesondere in der oberen Beschäftigtengrößenklasse). Eine Sensitivitätsanalyse und individuelle Parametersetzung wäre also für das Szenario des Einzelangriffes durchaus denkbar.

Führt der Datenangreifer einen Massenfischzug durch, so müssten diese Bemühungen für jeden Block separat erfolgen, was sich als nicht realisierbar darstellt. Bereits bei einer geringfügigen Erhöhung der gesuchten Unernehmen und der Zielunternehmen auf jeweils vier bei gleichbleibender Anzahl an Überschneidungsmerkmalen wird das Intervall $[0, 1]$ je nach Anzahl der effizienten Lösungen in bis zu 24 Teilintervalle zerlegt; d.h., die Schwellwerte λ_i

können sehr dicht beieinander liegen. Stehen zudem $k = 5$ Überschneidungsmerkmale zur Verfügung, so liegt die Situation aus dem ebenfalls kleinen Beispiel in Abschnitt 3.4.3 mit immerhin sechs effizienten Lösungen vor. Fällt der Parametervektor $\Lambda \in \mathbb{R}_{+}^{k-1}$ in einen der sechs konvexen Teilbereiche, so wird sich für die zugehörige effiziente Lösung entschieden. Die aus maßtheoretischer Sicht unmögliche Indifferenz besteht allgemein, wenn der Vektor Λ genau auf der Grenzlinie zweier (oder im Knotenpunkt mehrerer) Teilbereiche liegt.

3.7 Komplexitätsbetrachtung

Bekanntermaßen ist das lineare Zuordnungsproblem nach erfolgter Berechnung der Koeffizienten bei N zuzuordnenden Objekten von der Ordnung $O(N^3)$. Obwohl dieser Aufwand durchaus tolerabel ist, hat es sich in der Praxis als sehr sinnvoll erwiesen, auf dem Wege zu einer faktisch anonymen Datei in den zahlreichen zeitaufwendigen Vorläufen auf die effizientere Näherungsheuristik Prozedur II aus Unterabschnitt 3.4.2 zurück zu greifen und damit je Durchlauf wertvolle Zeit zu sparen. Tatsächlich liegen die Ergebnisse dieser Heuristik sehr nahe an der optimalen Lösung und die Zeitersparnis bei mehreren Stunden.[20]

Die innerhalb der Heuristik angewendete Sortierung, die für N Objekte bekanntermaßen mit einem Aufwand von $O(N \log N)$ realisierbar ist, erfolgt erfreulicherweise sehr schnell. Bezeichnet man mit n_a die Anzahl der Merkmalsträger in den externen Daten, mit n_b die Anzahl der Merkmalsträger in den Zieldaten und mit k die Anzahl der Überschneidungsmerkmale, so entsteht für die Berechnung der Distanzen, also der Koeffizienten des linearen Programmes, ein Aufwand der Ordnung $O(n_a n_b k)$.

Es sind via Prozedur II insgesamt $n_a n_b$ Distanzen zu sortieren. D.h., in unserem Falle ist der Sortieraufwand von der Ordnung $O(n_a n_b \log(n_a n_b))$. Nun sei mit $n = max(n_a, n_b)$ die maximale Anzahl an Merkmalsträgern einer Datei bezeichnet. Der Sortieraufwand kann danach durch $O(n^2 \log n)$ abgeschätzt werden. Entsprechendes gilt für den Aufwand bei der Koeffizientenberechnung.

Da im Allgemeinen $k << n$, ist die Komplexität des Gesamtverfahrens von der Ordnung

$$max(O(n^2 \cdot \log n), O(n^2 \cdot k)) = O(n^2 \log n),$$

was eine leichte Verbesserung gegenüber der (optimalen) Lösung des linearen Programmes darstellt und die praktische Erfahrung bestätigt. Grundsätzlich kann dieser Aufwand erheblich durch geeignete Zerlegung beider Datenquellen in Blöcke reduziert werden. Hierbei muss allerdings erwähnt werden, dass sich in den Simulationen das in theoretischen Komplexitätsbetrachtungen üblicherweise unberücksichtigte blockweise Ein- und Auslesen der Dateien als sehr zeitaufwendig erwiesen hat.

20 Detaillierte empirische Untersuchungen sowohl zur Näherung als auch zur Komplexität der Näherungsheuristiken finden sich in Lenz (2003a).

Kapitel 4

Beispiele der Anonymisierung wirtschaftsstatistischer Querschnittsdaten

Die Schätzung des mit einer probeweise anonymisierten Datei verbundenen Reidentifikationsrisikos hängt wesentlich von der Wahl des Zusatzwissens ab. Hierzu unterscheiden wir zwischen drei denkbaren Szenarien:

A *Zuordnung zwischen Originaldaten und anonymisierten Zieldaten.*
(Schätzung einer Obergrenze für das Reidentifikationsrisiko.)

B1 *Zuordnung zwischen externen Daten und formal anonymisierten Zieldaten (d.h. Originaldaten ohne direkte Identifikatoren).*
(Schätzung des natürlichen Schutzes in den Daten.)

B2 *Zuordnung zwischen externen Daten und anonymisierten Zieldaten.*
(Schätzung des realen Reidentifikationsrisikos.)

Als Zusatzwissen wurden in ersten Simulationen die Originaldaten herangezogen (Szenario A, im Folgenden „Worst-Case" –Situation genannt). Zum einen, weil im frühen Stadium der Arbeit noch keine Datenbank für die Durchführung realer Szenarien zur Verfügung stand. Zum anderen wäre bei erfolglosen Simulationen mit diesem bestmöglichen Zusatzwissen (über bessere Informationen als die Originaldaten kann kein Datenangreifer verfügen) die faktische Anonymität der untersuchten Daten zweifelsfrei nachgewiesen. Bedauerlicherweise waren diese Simulationen außerordentlich erfolgreich (Lenz 2003a), weshalb hierauf in der Folge verzichtet und der Fokus auf die realistischen Szenarien B1 und B2 gerichtet wurde.

Obwohl die Ergebnisse von Szenario A nicht die Realität widerspiegeln, wurde dieses Szenario vereinzelt dennoch in die Gesamtbeurteilung über die faktische Anonymität einer Datei einbezogen, da das tatsächlich verfügbare Zusatzwissen eines Datenangreifers schwierig abgrenzbar ist und die Ergebnisse verschiedener Simulationen der Szenarien B1 und B2 stark variieren können. Es kann niemals ausgeschlossen werden, dass vergleichbare Simulationen mit anderen Quellen des Zusatzwissens erfolgreicher wären. Die Ergebnisse der als

realistisch einzustufenden Szenarien B1 und B2 zeigen aber in jedem Falle auf, wie deutlich unter der Annahme theoretisch bestmöglichen Zusatzwissens in Szenario A das reale Reidentifikationsrisiko überschätzt wird.

In diesem Kapitel werden die genannten Szenarien mit realen Unternehmensdaten der amtlichen Statistik durchgeführt. Vor der Simulation der Szenarien B1 und B2 war es in allen Fällen notwendig, eine externe Datenbank aufzubauen. Hierzu wurden in Zusammenarbeit mit dem Zentrum für Europäische Wirtschaftsforschung (ZEW, Mannheim) zeit- und kostenaufwendige Adressabgleiche zwischen der kommerziellen MARKUS-Datenbank (siehe Unterabschnitt 4.1.1 und Creditreform 2009) und den amtlichen Mikrodaten durchgeführt, wobei moderne Techniken des „Pattern-Matching" zum Vergleich von Zeichenketten zum Einsatz kamen (Lenz et al. 2004a). Es hat sich in empirischen Arbeiten herausgestellt, dass sich zum Vergleich von Zeichenketten Kombinationen der klassischen Levensthein-Metrik, der Soundex-Methode und der *n*-gram -Methode gut eignen. Detaillierte Beschreibungen finden sich in Porter und Winkler (1999). Neben dem automatisierten Adressabgleich waren dennoch zusätzlich zeitaufwendige manuelle Nacharbeiten erforderlich, um eine hochwertige und repräsentative Referenzdatenbank für Datenangriffssimulationen zu erhalten.
Als Zieldaten werden im Folgenden die Kostenstrukturerhebung im Verarbeitenden Gewerbe (Abschnitt 4.1), die Umsatzsteuerstatistik (Abschnitt 4.2) und die Einzelhandelsstatistik (Abschnitt 4.3) untersucht. Eine ausführliche Dokumentation der Herangehensweise bei der Anonymisierung dieser Erhebungen unter besonderer Berücksichtigung eines bestmöglichen Erhaltes an Analysepotential findet sich in Ronning et al. (2005).

4.1 Anonymisierung der Kostenstrukturerhebung im Verarbeitenden Gewerbe

Bei der Kostenstrukturerhbebung im Verarbeitenden Gewerbe (KSE) des Jahres 1999 handelt es sich um eine Stichprobe von ungefährem Umfang 43%. Aufgrund der vergleichsweise geringen Anzahl von etwa 17 000 Einheiten – zum Beispiel besitzt die in Abschnitt 4.2 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 der Wirtschaftszweige 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 des folgenden Ausschnitts der Kostenstrukturerhebung in Tabelle 4.1 deutlich. Hier werden die Wirtschaftszweigklassifikation und der siedlungsstrukturelle Kreistyp BBR9[21] verwendet.

21 Dieser Schlüssel dient dem intraregionalen Vergleich. Die Typisierung der Kreise und Kreisregionen erfolgt außerhalb der Kernstädte nach der Bevölkerungsdichte. Insgesamt ergeben sich neun Kategorien (vgl. Abbildung 5.1 auf Seite 150).

Tabelle 4.1
Auszug der Kostenstrukturerhebung im Verarbeitenden Gewerbe

Wirtschafts-zweig (WZ 93)	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	1 894	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

Aus dem Auszug der KSE 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 4.2 angedeutet wird, welche die Verteilung der Daten auf Beschäftigtengrößenklassen enthält.

Tabelle 4.2: Verteilung der KSE-Unternehmen auf Beschäftigtengrößenklassen

Größenklasse	Abs. Häufigkeit	Rel. Häufigkeit	Kum. rel. 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 Arbeiten des Projektes „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten (FAWE)" besonders große Aufmerksamkeit geschenkt. Die Erfahrungen konnten später unmittelbar bei der Anonymisierung weiterer amtlicher Unternehmenserhebungen eingebracht werden (siehe Unterabschnitt 8.5). Nachfolgend wird die typische Vorgehensweise zur Generierung einer faktisch anonymen Datei mit bestmöglichem Erhalt an Potential für wissenschaftliche Analysen am Beispiel der Verfahrensgruppe Mikroaggregation (zur Beschreibung siehe Unterabschnitt 1.2.2) illustriert.

4.1.1 Verfügbares Zusatzwissen und Überschneidungsmerkmale

Wie in Abschnitt 2.3 erläutert wurde, kann das Zusatzwissen eines potentiellen 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 für die Massenfischzugsimulationen verwendeten Quellen des Zusatzwissens werden im Folgenden näher beschrieben. Diese sind im Einzelnen 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.

Umsatzsteuerstatistik als Zusatzwissen

Zunächst wurde die Umsatzsteuerstatistik verwendet, da diese von Beginn an für das Projekt FAWE verfügbar war. Die Ergebnisse von Angriffsszenarien mithilfe der Umsatzsteuerstatistik dienten einer ersten Einschätzung des Reidentifikationsrisikos. Als Überschneidungsmerkmale standen hier zur Verfügung:

- Gesamtumsatz,
- Regionalkennung,
- Wirtschaftszweigklassifikation.

Die Tabelle 4.3 enthält die Verteilung der 9 283 überprüfbaren Unternehmen (d.h. solche Einheiten, für welche nach einem Reidentifikationsversuch die Korrektheit der Zuordnung mittels Unternehmensidentifikationsnummer überprüft werden konnte) auf Beschäftigtengrößenklassen. Es sei angemerkt, dass diese Struktur nicht allein aus den Daten der Umsatzsteuerstatistik abgelesen werden kann, da das Merkmal *Anzahl der Beschäftigten* nicht in diese Daten enthalten ist. Alternativ wäre auch eine Tabellierung nach Umsatzgrößenklassen möglich und sinnvoll.

Tabelle 4.3: 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

MARKUS-Datenbank als Zusatzwissen

Für ein realistisches Szenario wurde die sogenannte 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,
- Wirtschaftszweigklassifikation.

Darüber hinaus enthält die MARKUS-Datenbank folgende Informationen:

- Firmenname und Adresse,
- Bilanzangaben,
- Stammkapital,
- Tätigkeitsbeschreibung,
- Beteiligungsstruktur,
- Angaben zur Geschäftsführung.

Die folgende Tabelle 4.4 zeigt die Verteilung der 9 394 überprüfbaren Einheiten aus der MARKUS-Datenbank auf Beschäftigtengrößenklassen:

Tabelle 4.4: Verteilung der MARKUS-Unternehmen auf Beschäftigtengrößenklassen

Größenklasse	Abs. Häufigkeit	Rel. Häufigkeit	Kum. rel. 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

Qualität des verwendeten Zusatzwissens

Abweichungsanalysen ergeben, dass sich bei Unternehmen des Verarbeitenden Gewerbes mit mindestens 20 Beschäftigten die Klassifikation der Wirtschaftszweige auf Zweistellerebene um 25% zwischen der KSE und der Umsatzsteuerstatistik unterscheidet, um fast 29% zwischen der Umsatzsteuerstatistik und der MARKUS-Datenbank und um 23% zwischen der KSE und MARKUS-Datenbank. Die Angaben des Umsatzes zwischen Umsatzsteuerstatistik und KSE stimmen verhältnismäßig gut überein (lediglich 18,8% der Unternehmen weisen Abweichungen von mehr als 10% auf). Dagegen sind starke Abweichungen zwischen Umsatzsteuerstatistik und MARKUS-Datenbank zu verzeichnen (über 60% der Unternehmen weisen Abweichungen von mehr als 10% auf). Zwischen der KSE und der MARKUS-Datenbank weichen bei fast 50% der Unternehmen die Ausprägungen in den Merkmalen *Gesamtumsatz* und *Anzahl der Beschäftigten* um mehr als 10% ab. Selbst bei der Regionalkennung treten zwischen jeweils zwei betrachteten Erhebungen gut zwei Prozent unterschiedliche Ausprägungen in der Regionalangabe, bezogen auf den siedlungsstrukturellen Regionstyp, auf.

4.1.2 Anonymisierungsmaßnahmen

Bei den Daten der Kostenstrukturerhebung wurden zahlreiche Anonymisierungsvarianten getestet. Hervorzuheben sind die Varianten der Mikroaggregation, der Zufallsüberlagerung sowie deren Verknüpfungen mit traditionellen informationsreduzierenden, auf die kategorialen Merkmale angewendeten Methoden.

Zum Verständnis der generellen Vorgehensweise bei der Anonymisierung wird das Beispiel der Mikroaggregation (zur Methode siehe Unterabschnitt 1.2.2) detailliert beschrieben. Es werden fünf Anonymisierungsvarianten betrachtet:

FORMAL: Diese Variante, die sogenannte formale Anonymisierung, entsteht allein aus der Herausnahme direkter Identifikatoren wie Name und Adresse.

MA33G: Dies ist die schwächste Form der eindimensionalen 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 mehrdimensionalen Mikroaggregation, bei der sämtliche metrische Merkmale zusammen gruppiert werden und somit nach den Mittelbildungen Tripel von Einheiten entstehen, welche sich höchstens durch die Ausprägungskombinationen in ihren kategorialen Merkmalen unterscheiden können.

4.1.3 Überprüfung der Schutzwirkung

Im Folgenden werden die Ergebnisse von Massenfischzügen mittels der in Abschnitt 4.1.1 vorgestellten Quellen des Zusatzwissens, der Umsatzsteuerstatistik und der kommerziell verfügbaren MARKUS-Datenbank, dargestellt. Anschließend werden die Ergebnisse dieser Szenarien mit dem sogenannten 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", welche in realistischen Szenarien in der Regel zur Verfügung stehen. Dies ermöglicht eine sinnvollere Gegenüberstellung der verschiedenen Simulationen.[22]

4.1.3.1 Realistische Szenarien

In diesem Abschnitt werden Szenarien mit der Umsatzsteuerstatistik (ca. 9 400 überprüfbare Einheiten) und der MARKUS-Datenbank (ca. 9 300 überprüfbare Einheiten) als Zusatzwissen durchgeführt. Obwohl die Umsatzsteuerstatistik einem potentiellen 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[23] verwendet:

22 Einem Datenangreifer ist eine Abschätzung des Risikos kaum möglich. Zwar könnte er Simulationen mit verschiedenen externen Quellen laufen lassen, ein Vergleich verschiedener Simulationen kann jedoch bestenfalls auf Basis der berechneten, wenig zuverlässigen Gesamtdistanzen für alle Zuordnungen angestellt werden.

23 Hier sind auch andere Gliederungstiefen bei den beiden kategorialen Merkmalen denkbar (vgl. hierzu spätere Ausführungen in Kapitel 6).

- Gesamtumsatz,
- Siedlungsstruktureller Kreistyp (BBR9),
- Wirtschaftszweigklassifikation (WZ93), Zweistellerebene.

Die nachfolgende Tabelle 4.5 enthält die Verteilung der Reidentifikation(srisik)en auf Beschäftigtengrößenklassen.

Tabelle 4.5: Reidentifikationen (Umsatzsteuerstatistik) nach Beschäftigtengrößenklassen

Varianten	Gesamt	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

Die Tabelle 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 4.5) annähert. Die Schutzwirkung formaler Anonymisierung, welche allein 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 4.5, dass besonders die Großunternehmen in der Klasse mit wenigstens 1000 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, welche 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 (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.

Die nachfolgende Tabelle 4.6 enthält dual zu Tabelle 4.5 die Verteilung der Reidentifikation(srisik)en auf Beschäftigtengrößenklassen.

Tabelle 4.6: Reidentifikationen (MARKUS) nach Beschäftigtengrößenklassen

Varianten	Gesamt	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

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 1 000 Beschäftigten. Bei den schwächeren Varianten MA11G, MA33G und FORMAL fällt der Anteil an Reidentifikationen für kleinere und mittlere Unternhemen (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.

Einzelangriffe

Es wurden Einzelangriffe auf 15 kleine und mittlere Unternehmen (< 250 Beschäftigte) sowie 26 große Unternehmen ($\geq$ 250 Beschäftigte) durchgeführt (Vorgrimler 2003; Vorgrimler und Lenz 2003a). Als Zieldaten dienten die formal anonymiserten Daten der Kostenstrukturerhebung. Durch Internetrecherchen konnte umfangreiches Zusatzwissen über Standort, Branchenzugehörigkeit, Anzahl der Beschäftigten und Gesamtumsatz generiert

werden. Vereinzelt waren auch Angaben über die Aufwendungen in Forschung und Entwicklung, Handelsaktivitäten und die Anzahl der tätigen Inhaber auffindbar und hilfreich. Insgesamt konnten 19 der 41 gesuchten Unternehmen korrekt zugeordnet werden, wobei lediglich eines der kleinen und mittleren Unternehmen in den Zieldaten auffindbar war. Für einige der gesuchten Unternehmen war keine eindeutige Zuordnung (weder richtig noch falsch) zu einem Zielunternehmen möglich.

4.1.3.2 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 4.1.3 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 4.1.3 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 4.1.3 daher nicht zur Verfügung. Durchgängig wurden die kategorialen Merkmale *Siedlungsstruktureller Kreistyp (BBR9)* und *Wirtschaftszweigklassifikation (WZ93)* zur Blockung der Daten verwendet.

Untenstehende Tabelle 4.7 enthält die absoluten und relativen Häufigkeiten der Reidentifikationen unter Verwendung von 1, 2, 3 und 33 metrischen Überschneidungsmerkmalen.

Tabelle 4.7: 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

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 dieses im Falle der Variante MA11G sogar leicht sinkt. Bestätigt wird die verhältnismäßig schwache Schutzwirkung der Variante MA11G, die bereits im vorhergehenden Abschnitt 4.1.3 beobachtet wurde.

Analog zu den Tabellen 4.5 und 4.6 wird in Tabelle 4.8 die Verteilung der Reidentifikationen auf Beschäftigtengrößenklassen betrachtet, beginnend mit der Simulation unter Verwendung eines Überschneidungsmerkmals:

Tabelle 4.8: Reidentifikationen (Worst-Case) mit einem Überschneidungsmerkmal nach Beschäftigtengrößenklassen

Variante	Gesamt	20-49	50-99	100-249	250-499	500-999	≥ 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

Tabelle 4.8 kann der Tabelle 4.5 (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 (siehe 4.1.3).

Nicht verwunderlich ist der Anstieg an Reidentifikationen gegenüber der vorherigen Simulation, da hier ein zusätzliches Überschneidungsmerkmal von bester Qualität für den Datenangreifer zur Verfügung steht.

Tabelle 4.9 kann der Tabelle 4.6 (Simulation mit der MARKUS-Datenbank) gegenüber gestellt werden, da dort dieselben Überschneidungsmerkmale im Zusatzwissen vorhanden waren. 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.

Tabelle 4.9: Reidentifikationen (Worst-Case) mit zwei Überschneidungsmerkmalen nach Beschäftigtengrößenklassen

Variante	Gesamt	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

4.1.3.3 Zusammenführung zu einem Gesamtrisikomaß

Sogar einer erfolgreiche Zuordnung eines Merkmalsträgers zu den Zieldaten kann vergeblich sein, wenn der den Datenangreifer interessierende Einzelwert von dem zugehörigen Originalwert relativ um mehr als eine vorgegebene Nutzenschwelle γ abweicht. Das auf diese Weise reduzierte Reidentifikationsrisiko wurde in Abschnitt 2.5 Enthüllungsrisiko genannt.

In Tabelle 4.10 sind die Enthüllungsrisiken für das zuvor simulierte Worst-Case Szenario mit zwei Überschneidungsmerkmalen dargestellt. Der erste Eintrag einer jeden Zelle bezieht sich auf den Schätzer $\hat{P}_\gamma(o^{(i)}$ enthüllt), der zweite auf $\hat{P}(r \in R)$ und der dritte auf $\hat{P}(\tilde{r}^{(i)} < \gamma \mid r \in R)$. Wurde keine Nutzenschwelle γ definiert, so berechnet sich der Schätzer für das Risiko der Enthüllung eines Einzelwertes durch den Anteil der korrekt zugeordneten Unternehmen innerhalb des ausgewählten Bereiches (siehe Spalte $\gamma = \infty$). Es wird in diesem Falle also jeder gefundene Einzelwert als brauchbar eingestuft, Reidentifikationsrisiko und Enthüllungsrisiko stimmen überein.

Beispielhaft werden in Tabelle 4.10 die mit den verschiedenen Anonymisierungsvarianten verbundenen Enthüllungsrisiken auf dem $\gamma = 0.05$-Niveau hervorgehoben. Insbesondere bei den stärkeren Varianten MA8G und MA1G wirkt sich die Anonymisierung nicht nur auf die Zuordnungsmöglichkeit der Unternehmen, sondern auch auf die Brauchbarkeit der gefundenen Einzelwerte aus. Durch alleinige Betrachtung des in der Literatur häufig verwendeten Reidentifikationsrisikos würde das eigentliche Gefährdungspotential dieser Dateien deutlich überschätzt. Zum Beispiel steht bei Variante MA8G der Trefferquote von 65,8% das Enthüllungsrisiko von 37,9% gegenüber, bei Variante MA1G reduziert sich das Reidentifikationsrisko von etwa 12% infolge der starken Veränderung der originalen Einzelwerte auf ein Enthüllungsrisiko von weniger als 4%. Da die Variante MA33G die Originaldaten nur geringfügig verändert, ist der Einfluss der Datenveränderung auf die Brauchbarkeit der

Einzelwerte und damit auf das endgültige Reidentifikationsrisiko vernachlässigbar.

Man beachte, dass aus Sicht eines risikoaversen Datenangreifers die zu (2.2) komplementäre Wahrscheinlichkeit maßgeblich ist:

$$\begin{aligned} P_\gamma(o^{(i)} \text{ nicht enthüllt}) &= 1 - P_\gamma(o^{(i)} \text{ enthüllt}) \\ &= P(r \notin R) + P(\tilde{r}^{(i)} \geq \gamma \wedge r \in R). \end{aligned}$$

Zum Beispiel würde ein Datenangreifer von gefundenen Einzelwerten von der Datei MA8G bei Kenntnis der Wahrscheinlichkeit von 62,1%, eine unbrauchbare Information aufzudecken, Abstand nehmen.

Tabelle 4.10: Enthüllungsrisiken (Worst-Case) auf dem γ-Niveau mit zwei Überschneidungsmerkmalen

target data\γ	∞	0,001	0,005	0,01	0,05	0,1	0.2
MA1G	0,1227	0,0246	0,0254	0,0266	**0,0352**	0,0446	0,0603
		0,1227	0,1227	0,1227	0,1227	0,1227	0,1227
		0,2002	0,2072	0,2168	0,2865	0,3636	0,4913
MA8G	0,6577	0,1858	0,2088	0,2326	**0,3790**	0,4851	0,5683
		0,6577	0,6577	0,6577	0,6577	0,6577	0,6577
		0,2825	0,3175	0,3536	0,5762	0,7376	0,8642
MA11G	0,9910	0,3226	0,4546	0,5565	**0,8094**	0,8815	0,9291
		0,9910	0,9910	0,9910	0,9910	0,9910	0,9910
		0,3255	0,4587	0,5616	0,8167	0,8895	0,9376
MA33G	0,9996	0,8908	0,9811	0,9904	**0,9976**	0,9985	0,9990
		0,9996	0,9996	0,9996	0,9996	0,9996	0,9996
		0,8911	0,9815	0,9908	0,9980	0,9989	0,9994

Setzt man nun im Sinne der Zusammenführung der verschiedenen Simulationsergebnisse die Nutzenschwelle beispielhaft mit $\gamma = 0,05$ an, so erhält man zunächst die globalen, mit den drei in Unterabschnitt 4.1.3 durchgeführten Experimenten verbundenen Enthüllungsrisiken. In nachfolgender Tabelle 4.11 und Abbildung 4.1 werden die Simulationen mit den formal anonymisierten Daten der Kostenstrukturerhebung (FORMAL, Szenario A) den Simulationen mit der Umsatzsteuerstatistik (Szenario B, TTS) und der MARKUS-Datenbank (Szenario B, MARKUS) als Zusatzwissen gegenüber gestellt.

Tabelle 4.11: Enthüllungsrisiken auf dem $\gamma = 0.05$-Niveau

Zieldaten \ Zusatzwissen	FORMAL	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

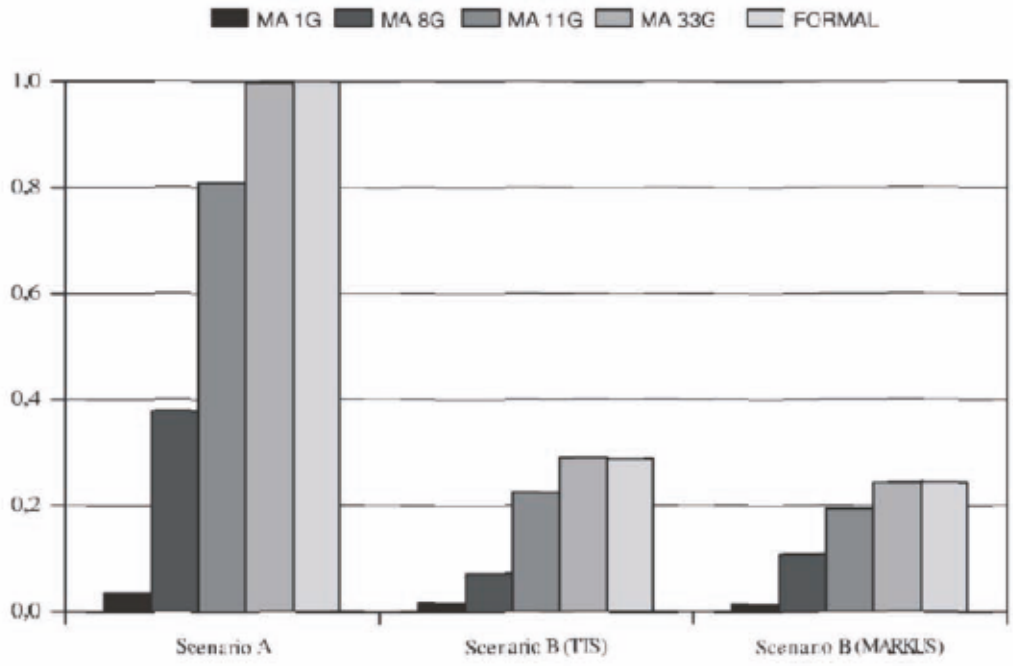
Zur Zusammenführung der in obiger Tabelle enthaltenen Risiken zu einem Gesamtrisikomaß kann z.B. eine Konvexkombination dienen (vgl. hierzu auch die Ausführungen in Abschnitt 2.5). Es bezeichne $\hat{P}_{A/\gamma}(w \text{ enthüllt})$ das mit dem Worst-Case Szenario geschätzte Risiko der Enthüllung des 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

$$\hat{P}_{\gamma}(w \text{ enthüllt}) := \lambda \cdot \hat{P}_{A/\gamma}(w \text{ enthüllt}) + (1-\lambda) \cdot \hat{P}_{B/\gamma}(w \text{ enthüllt}) \tag{4.1}$$

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 letztgenannten Falle sicher, dass einem potentiellen Datenangreifer kein besseres als das in den Simulationen verwendete Zusatzwissen zur Verfügung stehen könnte. Um einen vernünftigten 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 ans Herz gelegt, das Ergebnis des Worst-Case Szenarios nur geringfügig in das Gesamtrisikomaß einfließen zu lassen, da hier in der Regel sogar bei starker Anonymisierung sehr hohe Risken 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.

Abbildung 4.1
Enthüllungsrisiken bei der Kostenstrukturerhebung



Es reicht 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 4.12 und Abbildung 4.2 enthalten die Verteilung des Enthüllungsrisikos auf Beschäftigtengrößenklassen. Zur Zusammenführung der drei Simulationen wurde in obiger Gleichung als Stellparameter $\lambda = 0.2$ gewählt, wobei als Schätzer $\hat{P}_{B/\gamma}(w \text{ enthüllt})$ für die betrachteten Probeanonymisierungen das arithmetische Mittel der jeweiligen in Abschnitt 4.1.3 berechneten Enthüllungsrisiken (mit den MARKUS-Daten und der Umsatzsteuerstatistik als Zusatzwissen) bestimmt wurde.

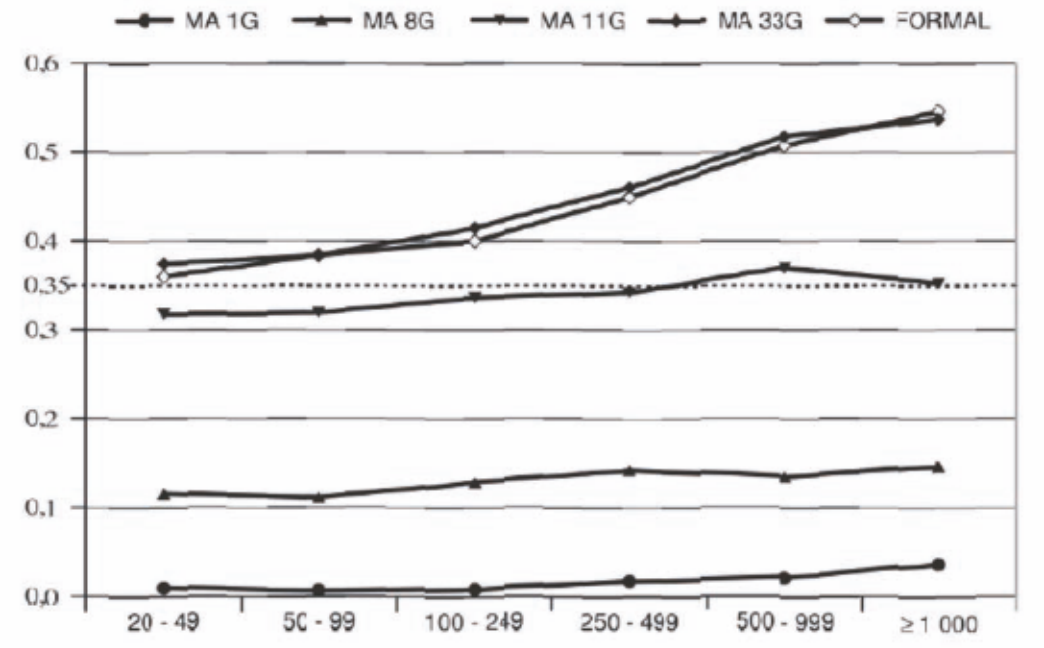
Tabelle 4.12: Enthüllungsrisiken auf dem $\gamma = 0.05$-Niveau nach Beschäftigtengrößenklassen

Zieldaten	Gesamt	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

Während auf der einen Seite mit zunehmender Beschäftigtenanzahl der Anteil korrekt zugeordneter 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 gemäß (2.4) in Abschnitt 2.5 unterhalb einer vorgegebenen Risikoschwelle τ, so kann die untersuchte Datei als faktisch anonym eingestuft werden. An dieser Stelle muss der Datenanbieter unter den Dateien, welche dieser Bedingung genügen, diejenige mit dem besten Erhalt des Analysepotentials 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 Lenz et al. 2005a) und auf das Worst-Case Szenario verzichtet wurde.

Abbildung 4.2
Vergleich der Schutzwirkung unterschiedlicher Mikroaggregationsverfahren bei der Kostenstrukturerhebung



4.1.4 Vergleich der Verfahrensgruppen Mikroaggregation und multiplikative Zufallsüberlagerung

In diesem Abschnitt stellen wir den oben betrachteten Mikroaggregationsvarianten fünf Varianten der multiplikativen Zufallsüberlagerung (zur Methodenbeschreibung siehe 1.2.2) gegenüber.

Bei den hier betrachteten Varianten wird mit f der Verschiebungsfaktor der beiden Mischungsverteilungen gegenüber 1 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.

Tabelle 4.13: Varianten der multiplikativen Zufallsüberlagerung

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

Nachfolgende Tabellen 4.14 und 4.15 enthalten die Reidentifikations- und Enthüllungsrisiken der oben beschriebenen Varianten der multiplikativen Zufallsüberlagerung unter Verwendung der MARKUS-Datenbank als Zusatzwissen. Als Nutzenschwelle wurde durchgehend $\gamma = 0.1$ gewählt. Zum Vergleich sind die in den vorherigen Abschnitten untersuchten Varianten der Mikroaggregation ebenfalls aufgeführt. Die Tabellen enthalten je Anonymisierungsvariante die jeweilige Risikoverteilung auf Beschäftigtengrößenklassen sowie das mit den Dateien verbundene Gesamtrisiko. Um eine lineare Ordnung zu erhalten, wurden die zehn betrachteten Anonymisierungsvarianten aufsteigend nach ihrem Reidentifikationsrisiko sortiert.

Nach abschließender Einbeziehung der Brauchbarkeit der durch den Datenangreifer gefundenen Werte ergibt sich in Tabelle 4.15 eine deutliche Reduktion der Risiken sowie eine leichte Verschiebung der vorherigen Sortierung.

Ausführlichere Berechnungen finden sich in (Lenz 2003a). Insbesondere werden dort die in Abschnitt 3.4.2 aufgeführten Näherungsheuristiken auf verschiedene Anonymisierungsvarianten angewendet und danach sowohl in ihren größenklassenabhängigen Ergebnissen für verschiedene Schwellenwerte γ als auch in der Rechenzeit miteinander verglichen.

Tabelle 4.14
Reidentifikationsrisiken aller Anonymisierungsvarianten

Variante	Beschäftigtengrößenklasse						
	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 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

Tabelle 4.15
Enthüllungsrisiken aller Anonymisierungsvarianten

Variante	Beschäftigtengrößenklasse						
	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 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

4.2 Anonymisierung der Umsatzsteuerstatistik

Die Besonderheit der Umsatzsteuerstatistik besteht in ihrem großen Berichtskreis. Sie umfasst mit wenigen Ausnahmen sämtliche Wirtschaftsbereiche, die entsprechend bei einer Anonymisierung berücksichtigt werden müssen. Zudem stellt sie eine Vollerhebung mit sehr geringer Abschneidegrenze dar. Die Unterscheidung zwischen einem Datenangreifer mit Teilnahmekenntnis und einem Datenangreifer ohne Teilnahmekenntnis entfällt bei der Umsatzsteuerstatistik, da diese aus den genannten Gründen als gegeben angenommen werden kann. Damit entfällt einerseits zwar der Schutz, der bei anderen Erhebungen durch das Stichprobendesign a priori gegeben ist, andererseits steht der Datenangreifer insbesondere bei der Suche nach einem kleinen oder mittleren Unternehmen nun vor einem großen Problem, da er die richtige unter einer Vielzahl ähnlicher Einheiten finden muss.

Ein zweites wesentliches Unterscheidungsmerkmal gegenüber anderen Unternehmenserhebungen der amtlichen Statistik ist die Datenstruktur. Wie im vorherigen Abschnitt bereits erwähnt wurde, besteht die Umsatzsteuerstatistik bei den metrischen Merkmalen im Wesentlichen aus dem Merkmal *Gesamtumsatz*. Die weiteren Merkmale, wie etwa Umsatz zu 16% Umsatzsteuer, sind Untermerkmale. Für einen Datenangreifer bedeutet dies zweierlei: Einerseits ist die Anzahl der potentiell verfügbaren metrischen Überschneidungsmerkmale sehr gering (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 Gesamtumsatz bereits als Überschneidungsmerkmal eingesetzt wird, da die zusätzlich enthüllbaren Merkmale nur eingeschränkt zusätzliche Informationen vermitteln. Dies erleichtert die Anonymisierung im Vergleich zu anderen Erhebungen wesentlich, obgleich in Betracht gezogen werden muss, dass mit der *Rechtsform* ein kategoriales Merkmal vorhanden ist, das weder in der Kostenstrukturerhebung im Verarbeitenden Gewerbe (Abschnitt 4.1) noch in der Einzelhandelsstatistik (Abschnitt 4.3) enthalten ist.

Aufgrund der geschilderten hohen Abhängigkeit der metrischen Merkmale untereinander macht es keinen Sinn, eine Anonymisierung auf die Überschneidungsmerkale zu beschränken. Zum einen würde dadurch die innere Plausibilität der einzelnen Datensätze negativ beeinflusst und zum anderen könnten die mit dem Umsatz hoch korrelierten Merkmale als Ersatzüberschneidungsmerkmale von einem Datenangreifer verwendet werden, womit die Anonymisierung ausgehebelt würde.

4.2.1 Verfügbares Zusatzwissen und Überschneidungsmerkmale

Als Überschneidungsmerkmale aus dem Zusatzwissen, das aus kommerziellen Datenbanken generiert werden kann, sind die *Rechtsform*, die *Regionalkennung*, die *Wirtschaftszweigklassifikation* und der *Gesamtumsatz* zu nennen. Die Umsatzsteuerstatistik unterscheidet sich damit von den anderen Projekterhebungen dahingehend, dass mit der Rechtsform ein zusätzliches kategoriales Überschneidungsmerkmal vorhanden ist, mit dem Gesamtum-

satz allerdings nur ein einziges metrisches Merkmal. Die Anzahl der Beschäftigten liegt als Überschneidungsmerkmal nicht vor. Für einen potentiellen 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, die entsprechend fehleranfällig und für Reidentifikationsversuche nur in Ausnahmefällen geeignet sind. Wie bereits mehrfach dargelegt wurde, sind die meisten metrischen Merkmale sehr stark mit dem Umsatz korreliert, wodurch sich deren Eignung als zusätzliches Überschneidungsmerkmal in engen Grenzen hält.

Insgesamt muss also festgestellt werden, dass ein Einzelangriff in der Regel mittels derselben Überschneidungsmerkmale simuliert werden sollte wie ein Massenfischzug. Dies hat zur Folge, dass die Ergebnisse von Massenfischzugsimulationen bereits gute Schätzer für die mit den zugehörigen Einzelangriffen verbundenen Ergebnisse darstellen.

Qualität des verwendeten Zusatzwissens

In den nachfolgenden Abweichungsanalysen werden nur die Merkmale *Wirtschaftszweigklassifikation* und *Gesamtumsatz* untersucht. Bei den beiden anderen Überschneidungsmerkmalen *Rechtsform* und *Gebietsschlüssel* wird 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.[24] Als mögliches Zusatzwissen stehen

24 Bei früheren Analysen mit tiefergehendem Regionalschlüssel konnten aber durchaus Abweichungen zwischen Zieldaten und Zusatzwissen festgestellt werden, allerdings waren diese mit 2% sehr gering, sodass es gerechtfertigt erscheint, diese im Folgenden zu vernachlässigen.

die Kostenstrukturerhebung des Verarbeitenden Gewerbes (KSE), die bereits im vorherigen Abschnitt verwendete kommerziell erhältliche MARKUS-Datenbank (beschränkt auf Unternehmen des Verarbeitenden Gewerbes) und die Einzelhandelsstatistik zur Verfügung.

Tabelle 4.16 enthält die Abweichungen zwischen der Umsatzsteuerstatistik und der Kostenstrukturerhebung. Die Abweichungen werden gemessen am Merkmal *Wirtschaftszweigklassifikation* (in unterschiedlicher Tiefengliederung) und am Anteil derjenigen Merkmalsträger, deren Umsätze in den beiden Erhebungen um weniger als x% voneinander differieren. In die Untersuchung gehen hierbei 9 283 gemeinsame Unternehmen beider Erhebungen ein.

Tabelle 4.16: 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	9 283	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 4.16 ersichtlich ist, hängt das 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 Gesamtumsatz von mindestens 10% und greift er mit Kenntnis der zweistelligen Wirtschaftszweigklassifikation an, so wird er knapp 40% der Unternehmen allein 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), wird erwartungsgemäß ein noch höheres natürliches Schutzniveau beobachtet. Bei rund 6 000 untersuchten Unternehmen, die gleichzeitg in der Umsatzsteuerstatistik und in der MARKUS-Datenbank enthalten und überprüfbar sind, liegt der Anteil an korrekten Wirtschaftszweigklassifizierungen zwar nur leicht unter dem entsprechenden mit der KSE erzielten Anteil (auf Zweistellerebene stimmen etwa 71% der Ausprägungen überein), die Abweichungen im Merkmal *Gesamtumsatz*

sind hier aber deutlich schlechter. Lediglich die Hälfte der untersuchten Unternehmen weisen Abweichungen von weniger als 10% bei diesem Überschneidungsmerkmal auf.

Beim Vergleich der gemeinsamen Unternehmen des Einzelhandels und der Umsatzsteuerstatistik ist das natürliche Schutzniveau etwas niedriger. Zum einen spielt die Wirtschaftszweigklassifikation keine große Rolle, da der Einzelhandel erst auf Dreistellerebene (sieben) unterschiedliche Ausprägungen aufweist, zum anderen wird eine näherungsweise Übereinstimmung der Ausprägungen des Merkmals *Gesamtumsatz* beobachtet. Bei etwa 73,5% der Unternehmen weichen die Umsatzangaben um weniger als 5% relativ voneinander ab. Allerdings zeigen sich besonders bei den größeren Unternehmen teilweise größere Abweichungen, was an leicht unterschiedlichen Umsatzdefinitionen in beiden Erhebungen liegt.

4.2.2 Anonymisierungsmaßnahmen

Im Folgenden werden die Anonymisierungsmaßnahmen beschrieben, die letztlich zum Scientific-Use-File der Umsatzsteuerstatistik 2000 geführt haben. Eine ausführliche Beschreibung des Scientific-Use-Files findet sich in Vorgrimler et al. (2005a).

Maßnahmen für kategoriale Merkmale

Zu Beginn des Projektes FAWE galt die Umsatzsteuerstatistik als diejenige der Projekterhebungen, bei der eine erfolgreiche faktische Anonymisierung unter Verzicht auf datenverändernde Verfahren am Leichtesten möglich erschien. Entsprechend konzentrierten sich die ersten Anonymisierungsversuche auf den Einsatz informationsreduzierender Methoden, die wiederum in erster Linie bei den kategorialen Merkmalen ansetzten:

- Das Merkmal *amtlicher Gemeindeschlüssel* wurde auf „neue/alte Bundesländer" vergröbert, wobei die neuen Bundesländer inklusive Berlin verstanden werden. Alternativ war eine regionale Aufteilung nach dem nichtadministrativen Gebietsschlüssel des Bundesamtes für Bauwesen und Raumordnung in der Diskussion. Dieser teilt die Regionen nach unterschiedlichen Siedlungsstrukturen auf. Für den im Projekt FAWE entwickelten Scientific-Use-File wurde ein Schlüssel mit den drei Ausprägungen „Agglomerationsraum", „verstädterter Raum" und „ländlicher Raum" konstruiert. Die in dem für das Projekt eigens eingerichteten Wissenschaflichen Begleitkreis vertretenen Wissenschaftler sprachen sich mehrheitlich für eine Verschlüsselung nach „neue/alte Bundesländer" und gegen die nichtadministrativen Varianten aus. Da die Sicherheitsprüfungen für die beiden Alternativen keine nennenswerten Unterschiede brachten, ist die Entscheidung zugunsten der administrativen Einteilung gefallen.
- Das Merkmal *Wirtschaftszweigklassifikation* wird in unterschiedlicher Tiefengliederung abhängig von den Besetzungszahlen in die Daten aufgenommen. Dabei wer-

den 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. In den Tabellen 4.17 und 4.18 sind die einzelnen ausgewiesenen Wirtschaftsbereiche des Scienitific-Use-Files aufgelistet.

- Aus dem Merkmal *Dauer der Steuerpflicht* wurde das Merkmal *Neugründung* mit den Ausprägungen „1 = ja" und „0 = nein" gebildet. Bei Unternehmen mit mehr als 100 Millionen Euro Umsatz wird das Merkmal generell auf Null gesetzt. Von über 150 000 als Neugründungen gekennzeichneten Unternehmen haben 118 Unternehmen einen Umsatz von über 100 Millionen Euro. Bei diesen wird dieses Merkmal auf 0 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* wurde sowohl aus Sicherheits- als auch aus Plausibilitätsgründen entfernt.

- Das Merkmal *Rechtsform* wurde zu folgenden Ausprägungen vergröbert:

  - Personengesellschaften,
  - Kapitalgesellschaften,
  - Erwerbs- und Wirtschaftsgenossenschaften sowie Betriebe gewerblicher Art von Körperschaften des öffentlichen Rechts und
  - Sonstige Rechtsform.

Tabelle 4.17
Ausgewiesene Wirtschaftsbereiche im Scientific-Use-File Umsatzsteuerstatistik 2000, Teil I

Wirtschafts-bereich (WZ 93)	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	Herstellung 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)
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

Tabelle 4.18
Ausgewiesene Wirtschaftsbereiche im Scientific-Use-File Umsatzsteuerstatistik 2000, Teil II

Wirtschafts-bereich (WZ 93)	Bezeichnung
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

Maßnahmen für metrische Merkmale

In einem ersten Schritt wurde versucht, mit traditionellen Abschneideverfahren besonders die Merkmalsausprägungen der großen Unternehmen zu verändern. Dies hatte jedoch eine nichtakzeptable Einschränkung des Analysepotentials 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 Analysepotential 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 Analysepotentials 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 Analysepotential 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 Mikroaggregation für alle Unternehmen angewendet.

- Zum Schutz der regionalen Branchenmarktführer wird 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* (Gesamtumsatz) das dominierende Merkmal ist[25]. Da dieses Verfahren getrennt nach den beiden Regionen angewandt wird, sind insgesamt 408 Merkmalsträger 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 verhältnismäßig wenigen großen Unternehmen in den neuen Bundesländern Rechnung getragen.

Tabelle 4.19 auf Seite 113 enthällt die Datensatzbeschreibung der Umsatzsteuerstatistik nach den beschriebenen Anonymisierungsmaßnahmen, die letztendlich zu einem Scientific-Use-File geführt haben.

25 Als regionale Branchenmarktführer gelten die drei Unternehmen, die in einer Branche (abgegrenzt nach der im Scientific-Use-File ausgewiesenen Wirtschaftsbereiche) und einer Region (Ost/West) die höchsten Umsätze aufweisen.

4.2.3 Überprüfung der Schutzwirkung

Die nachfolgenden Ausführungen konzentrieren sich auf die Schutzwirkung der Anonymisierung, die zur Erstellung eines Scientific-Use-Files geführt hat (siehe Vorgrimler et al. 2005a).

Zum Test der Schutzwirkung wurden Massenfischzug- und Einzelangriffszenarien durchgeführt. Zunächst wurde mithilfe der KSE[26] versucht, Merkmalsträger dieses Wirtschaftsabschnittes zu reidentifizieren. Im zweiten Szenario wurden 12 500 Unternehmen der Einzelhandelsstatistik als Zusatzwissen[27] verwendet und ebenfalls ein Massenfischzug durchgeführt.

Im Unterschied zu Abschnitt 4.1.3 werden im Folgenden nur realistische Szenarien simuliert. Die Worst-Case Simulationen sind hier aus Gründen der unangemessenen Rechenzeit nicht durchführbar. Man beachte, dass in diesem Falle bei einem Massenfischzug etwa 2,9 Millionen Merkmalsträger der Originaldaten den 2,9 Millionen Merkmalsträgern der anonymisierten Zieldaten zuzuordnen wären. Hinzu kommt, dass die zu erwartenden Ergebnisse mit denen aus Abschnitt 4.1.3 vergleichbar wären.

Auch kann im vorliegenden Falle im Gegensatz zu den entsprechenden Abschnitten zur Kostenstrukturerhebung und Einzelhandelsstatistik auf den Verfahrensvergleich zwischen Varianten der Mikroaggregation und der multiplikativen Zufallsüberlagerung verzichtet werden, da die hinsichtlich eines bestmöglichen Erhaltes an Analysepotential bewährte Variante der einfachen für jedes Merkmal separat durchgeführten Mikroaggregation bereits eine ausreichende Schutzwirkung entfaltet (siehe nachfolgende Ausführungen).

Kostenstrukturerhebung versus anonymisierte Umsatzsteuerstatistik

Bei den nachfolgenden Simulationsexperimenten wird versucht, knapp 9 300 Unternehmen der KSE (Zusatzwissen) den Merkmalsträgern der Umsatzsteuerstatistik (Zieldaten) zuzuordnen. In der Kostenstrukturerhebung des Verarbeitenden Gewerbes sind Unternehmen mit wenigstens 20 Beschäftigten als Stichprobe enthalten. Von den 2,9 Millionen Unternehmen der Umsatzsteuerstatistik kamen daher ca. 37 000 Unternehmen als Zielunternehmen in Frage. Da es sich bei der Umsatzsteuerstatistik bestehend aus Unternehmen mit Lieferungen und Leistungen von über 16 617 Euro prinzipiell um eine Vollerhebung handelt, konnte hier von der Kenntnis des potentiellen Datenangreifers über die Teilnahme des gesuchten Unternehmens an der Zielerhebung ausgegangen werden. Tabelle 4.20 enthält die Anteile

26 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 *Gesamtumsatz*. Darüberhinaus hat es den Vorteil, Anonymisierungsmaßnahmen, die bei der Rechtsform ansetzen, in ihrer Schutzwirkung bewerten zu können.

27 Als zusätzliches Merkmal wurde den Merkmalsträgern wiederum die Rechtsform hinzugespielt.

korrekter Zuordnungen, die entsprechenden Anteile an brauchbaren Informationen und die sich aus beiden ergebenden Enthüllungsrisiken. Die Tabelle enthält sowohl die Gesamtquote als auch die auf Umsatzgrößenklassen verteilten Quoten. Als Nützlichkeitsschwelle wurde $\gamma = 0.1$ gesetzt.

Das durchschnittliche Enthüllungsrisiko aller Unternehmen liegt bei 16,5%, wobei in keiner Umsatzgrößenklasse 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 für den Datenangreifer brauchbaren Informationen (von 78 auf 61%). Bei den mittleren Größenklassen ist zwar der Anteil an unbrauchbaren Informationen bei den zugeordneten Unternehmen recht gering, allerdings gilt dies auch für die zugehörigen Zuordnungsquoten.

MARKUS-Datenbank versus anonymisierte Umsatzsteuerstatistik

Zu einem früheren Zeitpunkt des Projektes FAWE wurde ein Massenfischzug mithilfe der MARKUS-Datenbank simuliert. Hierbei wurde versucht, 6 300 Unternehmen der MARKUS-Datenbank, die zum Verarbeitenden Gewerbe zu zählen sind, den 37 000 Zielunternehmen zuzuordnen. Auch wenn der Massenfischzug nicht bei dem eigentlichen Scientific-Use-File durchgeführt wurde, sondern lediglich bei einer schwächer anonymisierten Datei, zeigen die Ergebnisse eindrucksvoll die in den Daten vorhandene natürliche Schutzwirkung. 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 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.

Einzelhandelsstatistik versus anonymisierte Umsatzsteuerstatistik

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 (EHS) als Zusatzwissen verwendet wurden.

Tabelle 4.19
Datensatzbeschreibung des Scientific-Use-Files der Umsatzsteuerstatistik 2000

Lfd. 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 öffentl. 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	Nachrichtlich: innergemeinschaftl. Erwerbe		
25	LuL 1999		
26	Vorauszahlungssoll 1999		
27	Punktuell mikroaggregiert	1 = ja	

Tabelle 4.20: Enthüllungsrisiken (KSE) nach Umsatzgrößenklassen

Umsatzklasse (in Euro)	Anteil korrekter Zuordnungen (in %)	Anteil brauchbarer Informationen (in %)	Enthüllungsrisiko (in %)
bis 1 Mill.	9,1	100	9,1
1-10 Mill.	14,2	100	14,2
10-100 Mill.	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

Strukturell unterscheidet sich dieses dritte Szenario von den ersten beiden in

- der geringeren Anzahl an unterschiedlichen Kategorien der Wirtschaftszweigklassifikation (7 Kategorien gegenüber 22 Kategorien im vorherigen Experiment) und
- in der nicht vorhandenen Abschneidegrenze für kleine Unternehmen.

Beide Punkte sprechen für wesentlich schlechtere Voraussetzungen für einen erfolgreichen Massenfischzug. Sie führen dazu, dass die gesuchten Unternehmen innerhalb einer wesentlich höheren Anzahl an Zielunternehmen gesucht werden, welche sich zudem auf wenige disjunkte Kategorien verteilen.

Insgesamt werden 12 500 Unternehmen der Einzelhandelstatistik innerhalb von über 300 000 Unternehmen aus der Umsatzsteuerstatistik gesucht. Die riesigen Datenmengen können nur in angemessener Zeit verarbeitet werden, indem die Unternehmen in beiden Dateien in unterschiedliche Umsatzgrößenklassen geblockt werden. Durch Blockungsfehler können zwar a priori korrekte Zuordnungen verhindert werden. Andererseits wurde bereits *mit* der Blockung zur Generierung der in nachfolgender Tabelle 4.21 aufgeführten Ergebnisse eine Rechenlaufzeit von 19 Stunden (in CPU-Zeit) benötigt.

Tabelle 4.21: Enthüllungsrisiken (EHS) nach Umsatzgrößenklassen

Umsatzklasse (in Euro)	Anteil korrekter Zuordnungen (in %)	Anteil brauchbarer Informationen (in %)	Enthüllungsrisiko (in %)
bis 1 Mill.	7,0	99,9	7,0
1-10 Mill.	9,9	99,8	9,9
10-100 Mill.	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 4.21 werden wiederum die Zuordnungsquoten, die Anteile brauchbarer Einzelinformationen und die entsprechenden Enthüllungsrisiken abgetragen. Insgesamt ergibt sich ein deutlich geringeres Enthüllungsrisiko als bei dem mit der KSE als Zusatzwissen durchgeführten Szenario. Dies liegt an der unterschiedlichen Größenstruktur der Unternehmen, die mit der fehlenden Abschneidegrenze zusammenhängt. Bei den Einzelhandelsunternehmen dominieren die kleinen Unternehmen, was sich auf das Gesamtmaß für das Enthüllungsrisiko auswirkt. Da das Reidentifikationsrisiko für kleine bzw. umsatzschwache Unternehmen äußerst gering ist, ergibt sich auch für die Gesamtheit der gesuchten Unternehmen ein geringes Enthüllungsrisiko.

Geht man von einer Gesamtbetrachtung zu einer größenabhängigen Betrachtung über, so wird deutlich, dass sich die Ergebnisse der beiden vorherigen 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 im Merkmal *Gesamtumsatz* kompensiert. Auffallend ist allerdings der deutlich bessere Schutz von Großunternehmen der Einzelhandelsstatistik gegenüber den Großunternehmen der Kostenstrukturerhebung. Eine Erklärung hierfür könnte sein, dass bei den Einzelhandelsunternehmen weniger Großunternehmen existieren und damit hier ein größerer Anteil in dieser Größenklasse von der zusätzlichen Anonymisierung der Zieldaten via punktuelle Mikroaggregation betroffen ist. Dies äußert sich sowohl durch eine geringere Trefferquote (23 gegenüber 43%) als auch durch einen geringeren Anteil brauchbarer Informationen (55 gegenüber 61%).

Einzelangriffe

Bereits frühzeitig wurden im Projekt FAWE 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 gesuchten Unternehmen aus vier verschiedenen Branchen konnten tatsächlich alle korrekt zugeordnet werden.

Aus diesem Grunde wurde die bereits beschriebene punktuelle Mikroaggregation angewendet. Diese führte dazu, dass sich die drei regionalen Branchenmarktführer bestenfalls im Merkmal *Rechtsform* unterscheiden ließen und zudem keine Originalwerte, sondern lediglich Durchschnittswerte der drei regional führenden Unternehmen in den Zieldaten veröffentlicht wurden. Eine eindeutige Zuordnung wäre also nur in den wenigen Fällen der

Übereinstimmung der drei Unternehmen in ihrer Rechtsform möglich. Bei den 136 regionalen Branchen (jeweils 68 Wirtschaftszweige in Ost- und Westdeutschland) unterscheidet sich in 95 mindestens ein Unternehmen von den anderen und könnte somit korrekt zugeordnet werden. Eine solche korrekte Zuordnung würde aber nicht gegen die faktische Anonymität verstoßen, da ein Datenangreifer infolge der Datenveränderung keine brauchbaren Informationen gewinnen würde. So würde er weiterhin lediglich den durchschnittlichen Gesamtumsatz der drei regionalen Branchenmarktführer kennen und nicht den exakten Umsatzwert seines gesuchten Unternehmens. Diese durchschnittlichen Werte sind in mehr als der Hälfte der Fälle wenigstens 50% von den Originalwerten entfernt. Die Gruppe der regionalen Branchenmarktführer kann daher aufgrund der Maßnahme der punktuellen Mikroaggregation als faktisch anonymisiert eingestuft werden.

4.3 Anonymisierung der Einzelhandelsstatistik

Die Stichprobe der Einzelhandelsstatistik (EHS) umfasst 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 Millionen 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 % der Unternehmen mit mindestens 50 Beschäftigten schon um eine Vollerhebung handelt. Das bedeutet, dass bei diesen Unternehmen im Allgemeinen von einer Teilnahmekenntnis eines potentiellen Datenangreifers ausgegangen werden muss.

Im Gegensatz zu den anderen im laufenden Kapitel 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 MARKUS-Datenbank gewinnen. Daher wurden diese Merkmale aus Datenschutzgründen entfernt.

4.3.1 Verfügbares Zusatzwissen und Überschneidungsmerkmale

In diesem Abschnitt werden die verfügbaren Informationsquellen an Zusatzwissen und mögliche 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. Hinsichtlich der Durchführung von Einzelangriffen stellt die Informationsbeschaffung über das Internet hier eine besondere, nicht zu vernachlässigende Möglichkeit dar. Insbesondere können auf diesem Wege in kommerziellen Unternehmensdatenbanken in der Regel nicht enthaltene Informationen zum Merkmal *Anzahl der Filialen* gewonnen werden.

M&M Deutsche Handelsdatenbank als Zusatzwissen

Die M&M Deutsche Handelsdatenbank enthält rund 350 Firmenporträts von allen wichtigen Firmen und Organisationen des Lebensmittelhandels. Dabei sind folgende Überschneidungsmerkmale vorhanden:

- Gesamtumsatz,
- Filialen,
- Regionalinformation.

Daneben gibt es Informationen zu folgenden Sachverhalten:

- Firmenname und Adresse,
- Entscheidungsträger,
- Aufteilung Food/Nonfood,
- Nationale/internationale Kooperationen,
- Verkaufsflächen,
- Zusatzinformationen (Einkaufs-, Sortiments- und Preispolitik, Eigenmarken, Unternehmensstruktur usw.).

Während in der Einzelhandelsstatistik rechtlich selbstständige Einheiten erfasst werden (Unternehmensprinzip), ist bei der M&M Datenbank jedoch „nur" die jeweilige Firma vorhanden (Gruppenprinzip). Während beispielsweise eine Firma „Muster GmbH" in der M&M

Datenbank auch als solche genau einmal geführt wird, können dagegen in der Einzelhandelsstatistik zehn rechtlich selbstständige Einheiten dieser Firma erfasst worden sein. Daher wird eine mögliche Reidentifikation mit diesem Zusatzwissen enorm erschwert. Sowohl durchgeführte Massenfischzugs- als auch Einzelangriffsexperimente haben gezeigt, dass eine korrekte Zuordnung nahezu ausgeschlossen werden kann.

MARKUS-Datenbank als Zusatzwissen

Die MARKUS-Datenbank liefert Geschäftsinformationen zu ausgewählten Unternehmen der Creditreform. Folgende Überschneidungsmerkmale sind hier vorhanden:[28]

- Wirtschaftszweigklassifikation,
- Regionalinformation,
- Gesamtumsatz,
- Beschäftigte,
- Umsatz nach Tätigkeiten bzw. Produkten (in %).

Darüber hinaus lassen sich noch folgende Informationen gewinnen:

- Firmenname und Adresse,
- Geschäftsführer und Vorstände,
- Bilanzangaben,
- Stammkapital,
- Beteiligungsstruktur,
- Tätigkeitsbeschreibung.

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.

28 In manchen Fällen lässt sich auch die Anzahl der Filialen eines Unternehmens feststellen.

Tabelle 4.22 zeigt die Anteile gleich und verschieden klassifizierter Merkmalsträger in den kategorialen Überschneidungsmerkmalen *Regionalinformation* (BBR3, BBR9) und *Wirtschaftszweigklassifikation* (Dreisteller- und Viersstellerebene). 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 4.22: Abweichungen in den kategorialen Überschneidungsmerkmalen zwischen MARKUS-Datenbank und Einzelhandelsstatistik

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 BBR3 und BBR9 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 gehen die metrischen Merkmale *Gesamtumsatz* und *Beschäftigte* ein. Die folgende Tabelle 4.23 zeigt die relativen Abweichungen in diesen Merkmalen zwischen Zusatzwissen (MARKUS) und Zieldaten (EHS):

Tabelle 4.23: Abweichungen in den metrischen Überschneidungsmerkmalen zwischen MARKUS-Datenbank und Einzelhandelsstatistik

Merkmal	Gesamtumsatz	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%)

Tabelle 4.23 zeigt, dass beim Merkmal *Gesamtumsatz* nur gut ein Fünftel der Unternehmen eine relative Abweichung unter 10% von ihrem entsprechenden Originalwert vorweist. 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.

Durch die deutlichen Abweichungen in den Merkmalen *Wirtschaftszweigklassifikation*, *Gesamtumsatz* und *Beschäftigte* besteht bereits ein nicht zu vernachlässigender natürlicher Schutz der formal anonymisierten Daten gegenüber diesem Zusatzwissen.

Unter Berücksichtigung der Recherchearbeiten können insgesamt folgende Überschneidungsmerkmale in externen, realistischen Quellen des Zusatzwissens ausgemacht werden:

- Regionalinformation,
- Wirtschaftszweigklassifikation,
- Gesamtumsatz,
- Anzahl der Beschäftigten,
- Anzahl der Filialen,
- Teilumsätze nach Tätigkeiten bzw. Produkten (in %).

4.3.2 Anonymisierungsmaßnahmen

Eine geeignete Anonymisierungsstrategie kann aus informationsreduzierenden und/oder datenverändernden Methoden bestehen.

Getestet wurde in den Untersuchungen die Wirkung folgender datenverändernder Anonymisierungsverfahren, wobei bei der Mikroaggregation durchgehend $k = 3$ als Gruppengröße gesetzt wurde:[29]

- Formale Anonymisierung (FORMAL).
- Eindimensionale getrennte Mikroaggregation über 32 Merkmalsgruppen (MA32G).[30]
- Blockweise Anwendung der mehrdimensionalen Mikroaggregation über 9 Merkmalsgruppen (MA9G)[31].
- Mehrdimensionale gemeinsame Mikroaggregation über eine Gruppe (MA1G).[32]

29 Andere datenverändernde Verfahren, die sich bereits in früheren Simulationen bei der Kostenstrukturerhebung und der Umsatzsteuerstatistik als nicht aussichtsreich erwiesen haben, wurden nicht verfolgt.

30 Die Bezeichnung rührt daher, dass die Einzelhandelsstatistik 32 metrische Merkmale aufweist. Jedes Merkmal definiert demnach seine eigene Gruppe.

31 Die Einteilung der Gruppen wurde auf Basis der Ergebnisse von Clusteranalysen vorgenommen. Dabei enthält eine einzelne Gruppe zwischen drei und fünf paarweise möglichst hoch korrelierten Merkmalen. Sortiert wurde anschließend nach der Summe der normierten Einzelwerte der Merkmale einer Gruppe.

32 Die Merkmalsträger sind in dieser Variante – wenn überhaupt – allein durch charakteristische Kombinationen in den kategorialen Merkmalen unterscheidbar.

- Zü_Wink: Additive Überlagerung der logarithmierten Werte mit mehrdimensionaler Normalverteilung nach dem Verfahren von Winkler (siehe Kim und Winkler 2001). Die Varianz-Kovarianz-Matrix der Überlagerung entspricht dem 0.0005-fachen der Kovarianz-Matrix der logarithmierten Werte.

- Mult_f_s: Multiplikative Überlagerung mit Mischungsverteilung $W \sim \mathcal{N}(1 \pm f, s)$. Ein gleichbleibender Überlagerungsfaktor f für die metrischen Merkmale eines Merkmalträgers; Erwartungswert $1+f$ bzw. $1-f$ (je nach Mischungskomponente) und Standardabweichung s.

- Hoe2_f_s: Multiplikative Überlagerung der Einzelwerte mit einer Summe aus einer zweigipfligen Mischungsverteilung ($1+f$ bzw. $1-f$ mit je 50% Wahrscheinlichkeit) einheitlich für die metrischen Merkmale eines Merkmalsträgers und einer Normalverteilung $\mathcal{N}(0, s)$ für die Einzelwerte. Damit werden die Angaben des Betriebes entweder einheitlich erhöht oder gesenkt (Wahrscheinlichkeit je 50% und im Mittel um f%), wobei der Faktor der Veränderung über die Beobachtungen eines Merkmalsträgers mit s leicht variiert.

Die Verfahren der Mikroaggregation werden in Unterabschnitt 1.2.2 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 1.2.2 vorgenommen.

4.3.3 Überprüfung der Schutzwirkung

Zunächst wird die Schutzwirkung in Abhängigkeit zu den angewendeten Anonymisierungsvarianten der Mikroaggregation untersucht. Dabei werden Simulationsexperimente mit verschiedenen Quellen des Zusatzwissens und Variation der zur Verfügung stehenden Überschneidungsmerkmale durchgeführt. Aufgrund der zahlreichen Quellen möglichen Zusatzwissens erschien es notwendig, die Einzelangriffe hier auf breiterer Basis durchzuführen.[33] In den Simulationsexperimenten werden jeweils die formale Anonymisierung und die drei Mikroaggregationsverfahren mit unterschiedlichem Zusatzwissen angegriffen. Dabei werden in einem ersten Schritt die Reidentifikationsrisiken geschätzt. Im zweiten Schritt wird die Brauchbarkeit der enthüllten Einzelinformationen untersucht. Schließlich werden die sich aus den beiden Sachverhalten ergebenden Enthüllungsrisiken tabelliert.

33 An dieser Stelle gebührt ein Dank des Autors an den Freund und ehemaligen Kollegen Michael Scheffler für die Anwendung der Simulationsprogramme auf zahlreiche Anonymisierungsvarianten der Einzelhandelsstatistik; dies geschah zu einem Zeitpunkt, als es noch keine komfortable Programmoberfläche gab und die Simulationen nur nach intensiver vorheriger Einarbeitung angewendet werden konnten. Auch die Dokumentation der Simulationsergebnisse und die Tabellierung der Programmausgaben waren damals sehr zeitaufwendig.

Bei der Durchführung von Worst-Case Simulationen wird zusätzlich der Einfluss des in Einzelfällen charakteristischen Merkmals *Anzahl der Filialen* auf das Reidentifikationsrisiko untersucht.

Die Untersuchungen werden mit einem Vergleich der Schutzwirkung zwischen den Mikroaggregationsvarianten und den verschiedenen Varianten der Zufallsüberlagerung.

4.3.3.1 Realistische Szenarien

In diesem Abschnitt werden zunächst realitätsnahe Massenfischzüge unter Verwendung der MARKUS-Datenbank und der Umsatzsteuerstatistik als Zusatzwissen simuliert. Danach folgen unter Variation des möglichen Zusatzwissens einige Einzelangriffsversuche.

MARKUS-Datenbank versus anonymisierte Einzelhandelsstatistik

Mit der Annahme, dass der Datenangreifer über die Unternehmensinformationen der kommerziell erwerblichen MARKUS-Datenbank verfügt, wird ein realistisches Szenario abgebildet. Bei diesem Simulationsexperiment konnten 8 199 Unternehmen für den Massenfischzug verwendet werden. Dabei entspricht die Verteilung dieser Unternehmen nach Beschäftigtengrößenklassen weitgehend der Originalverteilung. Daher sind die in Tabelle 4.24 dargestellten Ergebnisse als aussagekräftig einzustufen. Mit der formalen Anonymisierung (FORMAL) lässt sich der natürliche Schutz der Originaldaten feststellen.

Bereits der natürliche Schutz der Daten der Einzelhandelsstatistik ist bemerkenswert: Nicht einmal jedes dritte Unternehmen aus der obersten Beschäftigtengrößenklasse wird reidentifiziert.

Neben dem Reidentifikationsrisiko spielt die Brauchbarkeit der gewonnenen Informationen eine entscheidende Rolle bei der Beurteilung der Schutzwirkung von Anonymisierungsverfahren. Dabei wird ein gefundener Einzelwert im Folgenden als brauchbar angesehen, wenn er weniger als 10% von seinem Originalwert abweicht (d.h. $\gamma = 0.1$). Die Tabelle 4.25 enthält die Anteile brauchbarer Informationen innerhalb der reidentifizierten Unternehmen aus dem vorherigen Szenario, verteilt auf Beschäftigtengrößenklassen.

Tabelle 4.24
Reidentifikationen (MARKUS) nach Beschäftigtengrößenklassen

Variante	Beschäftigtengrößenklasse							
	1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥1 000	Gesamt
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

Tabelle 4.25
Zellenweise Brauchbarkeit (MARKUS) nach Beschäftigtengrößenklassen

Variante	Beschäftigtengrößenklasse							
	1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥1 000	Gesamt
FORMAL	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
MA32G	1,00	1,00	1,00	1,00	1,00	1,00	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

Umsatzsteuerstatistik versus anonymisierte Einzelhandelsstatistik

Bei diesem Szenario wird unterstellt, dass dem Datenangreifer Zusatzwissen von der Qualität der amtlichen Einzeldaten aus der Umsatzsteuerstatistik zur Verfügung stehen. Das zuvor verwendete Überschneidungsmerkmal *Anzahl der Beschäftigten* ist in der Umsatzsteuerstatistik nicht vorhanden. Bei diesem Experiment werden 12 102 Unternehmen der Umsatzsteuerstatistik in den Zieldaten der Einzelhandelsstatistik gesucht.

Tabelle 4.26 zeigt, dass die Trefferquoten in den oberen Beschäftigtengrößenklassen gegenüber dem Experiment mit der MARKUS-Datenbank angestiegen sind. Andererseits sind die Trefferquoten bei den kleineren und mittleren Unternehmen in beiden Experimenten vergleichbar und fallen zudem erfreulich niedrig aus. Die Anteile brauchbarer Informationen innerhalb der reidentifizierten Unternehmen aus dem vorherigen Szenario, verteilt auf Beschäftigtengrößenklassen, ist in Tabelle 4.27 dargestellt.

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.

Einzelangriffe

Zur Einschätzung der von Einzelangriffen ausgehenden Gefahr wurde aus den Originaldaten der Einzelhandelsstatistik eine zufällige Stichprobe von 20 Unternehmen mit 1 000 und mehr Beschäftigten entnommen, da diese „großen" Unternehmen als besonders gefährdet einzustufen sind. Da sich bei der Recherchearbeit herausgestellt hat, dass es einige für einen potentiellen Datenangreifer interessante Quellen des Zusatzwissens gibt, wurden in diesem Abschnitt verhältnismäßig viele Einzelangriffe simuliert. Hierdurch soll insbesondere mit dem Vorurteil aufgeräumt werden, dass mittels Durchführung eines Massenfischzuges das reale Reidentifikationsrisiko unterschätzt wird.

Tabelle 4.26
Reidentifikationen (Umsatzsteuerstatistik) nach Beschäftigtengrößenklassen

Variante	Beschäftigtengrößenklasse							
	1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥1 000	Gesamt
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 4.27
Zellenweise Brauchbarkeit (Umsatzsteuerstatistik) nach Beschäftigtengrößenklassen

Variante	Beschäftigtengrößenklasse							
	1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥1 000	Gesamt
FORMAL	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
MA32G	1,00	1,00	1,00	1,00	1,00	1,00	0,97	0,99
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

a) Einzelangriffe mithilfe 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 *Gesamtumsatz* verwendet (vgl. Unterabschnitt 4.3.1). 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. Tabelle 4.28 zeigt die Ergebnisse der simulierten Einzelangriffe.

Tabelle 4.28: Einzelangriffe mit dem Internet als Quelle des Zusatzwissens

	Unternehmen insgesamt	Unternehmen mit 1 000 - 4 999 Beschäftigten	Unternehmen mit wenigstens 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	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.

Die zuvor beschriebenen Einzelangriffe wurden unter denselben Rahmenbedingungen bei gleichzeitiger Vergröberung des Merkmals *Anzahl der Filialen* in neun Kategorien wiederholt

(siehe Tabelle 4.33). Durch diese Anonymisierungsmaßnahme halbierte sich die Anzahl der reidentifizierten Unternehmen (ausführlich siehe Lenz und Scheffler 2004).

b) Einzelangriffe mithilfe der MARKUS-Datenbank als Zusatzwissen

Bei diesen Reidentifikationsversuchen wurden die Überschneidungsmerkmale *WZ-Viersteller, Regionalkennung Ost-West, Beschäftigte* und *Gesamtumsatz* 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 4.29 wiedergegeben.

Tabelle 4.29: Einzelangriffe mit der MARKUS-Datenbank als Quelle des Zusatzwissens

	Unternehmen insgesamt	Unternehmen mit 1 000 - 4 999 Beschäftigten	Unternehmen mit wenigstens 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	5 (45,50%)	2 (28,60%)	3 (75%)

Demnach konnten sechs Unternehmen den originalen Merkmalsträgern richtig zugeordnet werden. Die Wahrscheinlichkeit, dass ein Datenangreifer ein Unternehmen reidentifizieren kann, liegt somit bei dieser Simulation bei 54,5%. Er müsste also mit einer immer noch beachtlichen Wahrscheinlichkeit von 45,5% damit rechnen, falsch zugeordnet zu haben.

c) Einzelangriffe mit kombiniertem Zusatzwissen

Schließlich wurden in einer weiteren Simulation sowohl die Informationen der Internetrecherche als auch der MARKUS-Datenbank verwendet. Dabei wurde für jedes Überschneidungsmerkmal die qualitativ bessere Information eingebracht. Mit dieser Vorgehensweise erhält man die bestmögliche 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 4.30 dargestellt wird.

Dieses idealtypisch kombinierte Zusatzwissen führt demnach zu einer beachtlichen Trefferquote von 73%. Ein Datenangreifer wäre demzufolge in der Lage, ca. drei von vier

Unternehmen reidentifizieren zu können. Die hohe Trefferquote verdeutlicht, dass Unternehmen mit mindestens 1000 Beschäftigten besonders schutzbedürftig sind. Ausführliche Berechnungen finden sich in Scheffler und Lenz (2004).

Tabelle 4.30: Einzelangriffe mit dem Internet und der MARKUS-Datenbank als Quellen des Zusatzwissens

	Gesuchte Unternehmen	Reidentifikationen	Trefferquote
Internetrecherche	11	4	36%
MARKUS-Datenbank	11	6	55%
Internetrecherche & MARKUS-Datenbank	11	8	73%

4.3.3.2 Worst-Case Szenario

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*, *Gesamtumsatz* 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 4.31 stellt die entsprechenden Trefferquoten nach Beschäftigtengrößenklassen dar.

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.

Die Tabelle 4.32 enthält die Anteile brauchbarer Informationen innerhalb der reidentifizierten Unternehmen aus dem vorherigen Szenario, verteilt auf Beschäftigtengrößenklassen.

Da das Merkmal *Anzahl der Filialen* gut über Internetrecherchen gewonnen werden kann, sollte es unbedingt als verfügbares Überschneidungsmerkmal angesehen werden. Bei den in diesem Unterabschnitt simulierten Massenfischzügen wurden die Originaldaten als Zusatzwissen verwendet. Damit ist es möglich, den Einfluss des Merkmals *Anzahl der Filialen* zu untersuchen, obwohl dieses Merkmal nicht in den anderen beiden verwendeten Unternehmensdatenbanken vorhanden ist.

Tabelle 4.31
Reidentifikationen (Worst Case) nach Beschäftigtengrößenklassen

Variante	Beschäftigtengrößenklasse							
	1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥1 000	Gesamt
FORMAL	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
MA32G	0,99	1,00	1,00	1,00	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

Tabelle 4.32
Zellenweise Brauchbarkeit (Worst Case) nach Beschäftigtengrößenklassen

Variante	Beschäftigtengrößenklasse							
	1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥1 000	Gesamt
FORMAL	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
MA32G	1,00	1,00	1,00	1,00	1,00	1,00	0,97	1,00
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

Im Folgenden wird untersucht, welche Auswirkungen die Ergänzung des Zusatzwissens um das Merkmal *Anzahl der Filialen* auf das Reidentifikationsrisiko hätte. Ebenso wird geprüft, inwieweit sich mit einer Vergröberung dieses Merkmals zu neun Kategorien (siehe Tabelle 4.33) das Reidentifikationsrisiko verringert. Bei der Regionalinformation wurde auf den siedlungsstrukturellen Kreistyp BBR9 und bei der Wirtschaftszweigklassifikation auf die Zweistellerebene zurückgegriffen. Getestet wurden die drei Mikroaggregationsvarianten MA1G, MA9G und MA32G. Tabelle 4.34 enthält die Trefferquoten der Massenfischzugsimulation nach Beschäftigtengrößenklassen unter Verwendung folgender Kanons an Überschneidungsmerkmalen:

Variante 1: WZ-Dreisteller, Regionalbezug BBR9, Gesamtumsatz, Beschäftigte
Variante 2: WZ-Dreisteller, Regionalbezug BBR9, Gesamtumsatz, Beschäftigte, Filialkategorien
Variante 3: WZ-Dreisteller, Regionalbezug BBR9, Gesamtumsatz, Beschäftigte, Anzahl Filialen

Da mit dem Verfahren der einfachen für alle metrischen Merkmale separaten durchgeführten Mikroaggregation (MA32G) in früheren Experimenten bereits mit vier Überschneidungsmerkmalen Trefferquoten nahe bei 100 % erzielt wurden, ist eine Betrachtung der anderen beiden Mikroaggregationsvarianten MA1G und MA9G für diese Untersuchung aufschlussreicher.

Erwartungsgemäß steigt das Reidentifikationsrisiko mit zunehmender Detaillierung des Zusatzwissens an.[34] Es lässt sich feststellen, dass die Reidentifikationsrisiken vor allem in den oberen Beschäftigtengrößenklassen beachtlich ansteigen. Geringere Unterschiede treten in der untersten Beschäftigtengrößenklasse auf. Da diese Klasse sehr dicht besetzt ist und gut 93% der dort vorhandenen Unternehmen nur eine Filiale aufweisen, trägt das vergröberte Merkmal bei diesen Unternehmen nicht zu einer Erhöhung der Reidentifikationsgefahr bei.

Bei Betrachtung der obersten Beschäftigtengrößenklasse ist festzustellen, dass die Bildung von Filialkategorien bei dem Verfahren MA9G zu knapp 10 Prozentpunkten und mit MA1G zu rund 12,5 Prozentpunkten das Reidentifikationsrisiko reduziert.

34 Variante 2 besitzt gegenüber Variante 1 das zusätzliche Überschneidungsmerkmal *Anzahl der Filialen* in vergröberter Form; in Variante 3 wurde dieses Merkmal nicht verändert.

Tabelle 4.33
Kategorisierung der Merkmals „Anzahl der Filialen“

Anzahl der Filialen	1	2	3	4	5	6 – 10	11 – 50	51 – 100	>100
Kategorie	1	2	3	4	5	6	7	8	9
Häufigkeit	11 618	1 095	455	455	155	360	362	98	79

Tabelle 4.34
Reidentifikation der drei Varianten nach Beschäftigtengrößenklassen

Variante	Beschäftigtengrößenklasse							
	1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥1 000	Gesamt
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

4.3.3.3 Zusammenführung zu einem Gesamtrisikomaß

Das Risiko der Enthüllung brauchbarer Werte wurde gemäß der Beschreibung in Abschnitt 2.5 berechnet. Die Enthüllungsrisiken nach Beschäftigtengrößenklassen für die drei vorher beschriebenen Experimente sind in Tabelle 4.35 dargestellt. Das Gesamtrisiko ergibt sich wie in 4.1 als Konvexkombination zwischen dem Enthüllungsrisiko für das Worst-Case Szenario und dem Enthüllungsrisiko für die beiden realistischen Szenarien:[35]

$$\hat{P}_\gamma = \lambda \hat{P}_{\gamma,wc} + (1-\lambda)\hat{P}_{\gamma,real}$$

für ein geeignetes $\lambda \in [0,1]$.Hier wurde für λ der Wert $0,2$ gewählt. Da die Brauchbarkeit aufgedeckter Informationen unabhängig von der Wahl des Zusatzwissens ist, reduzieren sich die in den Tabellen 4.31 bis 4.34 dargestellten Risiken simultan in allen Experimenten.

4.3.3.4 Auswirkungen der Vergröberung kategorialer Überschneidungsmerkmale auf die Datensicherheit

In diesem Unterabschnitt wird untersucht, wie sich verschiedene Vergröberungen bei den beiden kategorialen Überschneidungsmerkmalen *Regionalinformation* und *Wirtschaftszweigklassifikation* auf das im vorherigen Unterabschnitt 4.3.3.3 kombinierte Enthüllungsrisiko auswirken. Die Ergebnisse finden sich in den Tabellen 4.35 bis 4.38.

Hinsichtlich der Wahl einer geeigneten Mischung aus informationsreduzierenden und datenverändernden Verfahren lassen sich beispielsweise die beiden folgenden Beobachtungen machen. Mit der Merkmalskombination „WZ-Viersteller / BBR9" entfaltet die Variante MA1G eine ähnliche Schutzwirkung wie Variante MA9G, wenn bei letzterer nur der *WZ-Dreisteller* und *BBR3* als Überschneidungsmerkmale verfügbar sind.[36] Als Ausgangspunkt für das zweite Beispiel sollen die nutzerfreundlichere datenverändernde Variante MA32G und die Merkmalskombination „WZ-Dreisteller / BBR9" dienen. Danach lässt sich durch die Vergröberung der Regionalinformation auf den BBR3-Schlüssel – abgesehen vom Worst-Case Fall – mit der Herausgabe der formal anonymisierten Daten in allen Zellenwerten eine Verbesserung des Datenschutzes beobachten.[37]

35 Dabei wurde $\hat{P}_{\gamma,real}$ als Mittelwert der beiden Risiken für die realistischen Experimente berechnet.

36 Dazu vergleiche man die Zeilen zu MA1G in Tabelle 4.36 mit den Zeilen zu MA9G in Tabelle 4.38, und hier insbesondere die Klasse der Unternehmen mit wenigstens 250 Beschäftigten. Außerdem sollten die Ergebnisse in der Spalte „Gesamt" nicht überbewertet werden, da hier größtenteils die überproportional vorhandenen kleinen Unternehmen einfließen.

37 Vgl. die Zeilen zu MA32G in Tabelle 4.35 mit den Zeilen zu FORMAL in Tabelle 4.38.

4.3.4 Vergleich der Verfahrensgruppen Mikroaggregation und multiplikative Zufallsüberlagerung

Im Folgenden werden die Varianten der Zufallsüberlagerung (Mult_f04_s02, Mult_f08_s02, Mult_f11_s03, Mult_f11_s03_k, ZÜ_Wink, Hoe2_f04_s02, Hoe2_f08_s02, Hoe2_f11_s02 und Hoe2_f11_s02_k)[38] mit der formalen Anonymisierung und den drei Mikroaggregationsvarianten MA1G, MA9G und MA32G hinsichtlich ihrer Schutzwirkung verglichen.

Dabei wurden in den Simulationen folgende Überschneidungsmerkmale zwischen der Einzelhandelsstatistik und der kommerziell erhältlichen MARKUS-Datenbank verwendet:

- WZ-Viersteller,
- Regionalbezug BBR9,
- Gesamtumsatz,
- Beschäftigte.

Die abschließende Tabelle 4.39 gibt Auskunft über die Verteilung der Enthüllungsrisiken auf die Beschäftigtengrößenklassen. Die Sortierung der Varianten erfolgte absteigend nach den Enthüllungsrisiken in der obersten Beschäftigtengrößenklasse (1000 und mehr Beschäftigte).

Erwartungsgemäß fällt die Rangfolge innerhalb der Varianten der Zufallsüberlagerung aus. Die Varianten Mult_f04_s02, Mult_f08_s02, Hoe2_f04_s02 und ZÜ_Wink weisen ähnliche Enthüllungsrisiken wie die einfache Mikroaggregation (MA32G) auf. Demnach entfalten sie nur eine geringe Schutzwirkung. Betrachtet man die Verteilung der Enthüllungsrisiken auf die Kombinationen von "Wirtschaftszweig / Regionalinformation / Beschäftigtengrößenklasse", so steigen obige Risiken noch einmal deutlich an und liegen häufig über der kritischen Marke von 50%, die als maximal zulässige obere Risikoschwelle angesehen werden kann.

Die Diskussionen innerhalb des Projektes FAWE haben ergeben, dass Analysepotential und Datensicherheit bei der Variante Hoe2_f08_s02 ausreichen, um faktische Anonymität und ein Mindestmaß an wissenschaftlicher Tauglichkeit gewährleisten zu können, sodass ein alle Unternehmen beinhaltender Scientific-Use-File für die Einzelhandelsstatistik 1999 erzeugt werden konnte (Scheffler 2005).

38 Der Buchstabe k in der Variantenbezeichnung bedeutet, dass zusätzlich eine Korrektur der ersten und zweiten Momente vorgenommen wurde.

Tabelle 4.35
Enthüllungsrisiko mit WZ 93-Dreisteller und BBR9

Variante	Zusatz-wissen	Beschäftigtengrößenklasse							
		1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1 000	Gesamt
MA1G	Worst Case	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	Gesamt	0,02	0,02	0,03	0,04	0,06	0,08	0,12	0,02
MA9G	Worst Case	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	Gesamt	0,04	0,06	0,09	0,11	0,18	0,17	0,23	0,05
MA32G	Worst Case	0,99	1,00	1,00	1,00	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	Gesamt	0,29	0,30	0,34	0,36	0,42	0,41	0,48	0,30
FORMAL	Worst Case	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
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	Gesamt	0,29	0,30	0,34	0,35	0,43	0,40	0,52	0,30

Tabelle 4.36
Enthüllungsrisiken mit WZ 93-Viersteller und BBR9

Variante	Zusatzwissen	Beschäftigtengrößenklasse							
		1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1 000	Gesamt
MA1G	Worst Case	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	Gesamt	0,06	0,07	0,08	0,10	0,12	0,15	0,18	0,06
MA9G	Worst Case	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	Gesamt	0,10	0,13	0,17	0,21	0,28	0,24	0,30	0,11
MA32G	Worst Case	1,00	1,00	1,00	1,00	1,00	0,99	0,97	1,00
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	Gesamt	0,33	0,339	0,45	0,44	0,51	0,49	0,50	0,35
FORMAL	Worst Case	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
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	Gesamt	0,33	0,39	0,45	0,44	0,51	0,49	0,53	0,35

Tabelle 4.37
Enthüllungsrisiken mit WZ 93-Viersteller und BBR3

Variante	Zusatz-wissen	Beschäftigtengrößenklasse							
		1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1 000	Gesamt
MA1G	Worst Case	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	Gesamt	0,03	0,04	0,04	0,06	0,07	0,10	0,15	0,03
MA9G	Worst Case	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	Gesamt	0,05	0,07	0,11	0,12	0,22	0,17	0,26	0,06
MA32G	Worst Case	0,99	1,00	1,00	1,00	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	Gesamt	0,28	0,34	0,37	0,39	0,47	0,44	0,49	0,30
FORMAL	Worst Case	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
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	Gesamt	0,29	0,34	0,37	0,39	0,47	0,45	0,51	0,30

Tabelle 4.38
Enthüllungsrisiken mit WZ 93-Dreisteller und BBR3

Variante	Zusatzwissen	Beschäftigtengrößenklasse							
		1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1 000	Gesamt
MA1G	Worst Case	0,04	0,04	0,05	0,06	0,04	0,12	0,13	0,04
MA1G	MARKUS	0,00	0,01	0,01	0,02	0,01	0,05	0,05	0,00
MA1G	USt	0,01	0,01	0,01	0,01	0,01	0,05	0,08	0,01
MA1G	Gesamt	0,01	0,01	0,02	0,02	0,02	0,06	0,08	0,01
MA9G	Worst Case	0,08	0,11	0,15	0,16	0,26	0,21	0,37	0,09
MA9G	MARKUS	0,00	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	Gesamt	0,02	0,03	0,05	0,05	0,10	0,11	0,20	0,03
MA32G	Worst Case	0,98	1,00	1,00	1,00	1,00	0,99	0,97	0,98
MA32G	MARKUS	0,00	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	Gesamt	0,25	0,26	0,28	0,29	0,32	0,34	0,46	026
FORMAL	Worst Case	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
FORMAL	MARKUS	0,00	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	Gesamt	0,26	0,27	0,28	0,30	0,33	0,35	0,48	0,26

Tabelle 4.39
Enthüllungsrisiken aller Anonymisierungsvarianten nach Beschäftigtengrößenklassen

Variante	Beschäftigtengrößenklasse							
	1 – 19	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1 000	Gesamt
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
Hoe2_f04_s02	0,04	0,07	0,14	0,17	0,23	0,20	0,31	0,07
ZÜ_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
Hoe2_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	0,10	0,12	0,20	0,20	0,18	0,04
Hoe2_f11_s02_k	0,02	0,04	0,08	0,09	0,13	0,10	0,17	0,03
Mult_f11_s03_k	0,01	0,03	0,08	0,10	0,13	0,09	0,17	0,03
Hoe2_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

Kapitel 5

Empirische Untersuchungen zur Schutzwirkung informationsreduzierender Methoden

In diesem Kapitel soll am Beispiel der bereits in Abschnitt 4.2 untersuchten Umsatzsteuerstatistik empirisch die besondere Wirkung informationsreduzierender Verfahren auf die Datensicherheit vorgestellt werden. Es werden hierzu Vergröberungen der Merkmale *Rechtsform, Wirtschaftszweigklassifikation* und der *Regionalangabe* getestet.

Die Umsatzsteuerstatistik eignet sich aus folgenden Gründen hervorragend für diesen Zweck:

- Es werden nahezu alle Wirtschaftsbereiche in beliebig tiefer Gliederung in der Erhebung abgebildet, weshalb der Einfluss einer Vergröberung des Merkmals *Wirtschaftszweigklassifikation* gut ermittelt werden kann.

- Selbst die unterste Beschäftigtengrößenklasse (1 – 24 Beschäftigte) ist im Gegensatz zu den anderen verfügbaren Unternehmenserhebungen dicht besetzt.

- Gegenüber Erhebungen wie der Kostenstrukturerhebung oder der Einzelhandelsstatistik kommt mit der *Rechtsform* ein zusätzliches kategoriales Überschneidungsmerkmal hinzu.

Ebenso werden „bereinigte Trefferquoten" betrachtet: Wie würden sich die Risiken verändern, wenn die kategorialen Überschneidungsmerkmale fehlerfrei wären bzw. in externen Daten und Zieldaten übereinstimmten? Die bereinigten Trefferquoten zeigen nicht nur den Einfluss von Abweichungen in den kategorialen Merkmalen auf, sondern erlauben auch eine besssere Einschätzung der Wirkung datenverändernder Methoden wie im vorliegenden Falle der Mikroaggregation.

5.1 Verwendetes Datenmaterial

Für die Durchführung der Reidentifikationsexperimente wurden die Merkmale *Wirtschaftszweigklassifikation, Gesamtumsatz, Rechtsform* und *Regionalkennung*[39] als Überschneidungsmerkmale verwendet. Das verwendete Zusatzwissen enthielt knapp 9 300 Unternehmen mit 20 oder mehr Beschäftigten aus den Wirtschaftsabteilungen 10-37 (Verarbeitendes Gewerbe). Rund 37 000 Unternehmen wiesen seitens der Umsatzsteuerstatistik diese Charakteristika auf, sodass diese als Zielunternehmen in Frage kamen. Daraus ergab sich die Aufgabe, die 9 300 Unternehmen des Zusatzwissens innerhalb der 37 000 Unternehmen der Zieldaten zu reidentifizieren. Für die Massenfischzüge wurden die kategorialen Merkmale unterschiedlich vergröbert. So wurden Massenfischzüge auf der Ebene des WZ-Vier- bis Einstellers durchgeführt und es wurden Massenfischzüge simuliert, bei denen vollständig auf die Wirtschaftszweigangabe verzichtet wurde. Die Rechtsform wurde einmal mit acht und einmal mit vier Ausprägungen verwendet. Der Regionalschlüssel wies in den Simulationen neun, sieben oder drei Ausprägungen auf.

5.2 Anonymisierungsvarianten

Die Daten der Umsatzsteuerstatistik wurden darüber hinaus auf drei unterschiedliche Weisen anonymisiert:

- Die formale Anonymisierung als schwächste Anonymisierungsform. Dabei werden lediglich die direkten Identifikatoren Name und Anschrift gelöscht.
- Die zweite Variante ist die schwächste Form der Anonymisierung durch (eindimensionale) Mikroaggregation. Dabei wird jedes stetige Merkmal separat mikroaggregiert.
- Die dritte Variante beinhaltet dagegen die stärkste Form der (mehrdimensionalen) Mikroaggregation, bei der sämtliche metrische Merkmale gemeinsam aggregiert werden.

39 Als Regionalkennung wurden die siedlungsstrukturellen Kreistypen verwendet. Diese nichtadministrativen Schlüssel dienen dem intraregionalen Vergleich. Es wird nach „Kernstädten" und sonstigen Kreisen bzw. Kreisregionen unterschieden. Als Kernstädte werden kreisfreie Städte mit mehr als 100 000 Einwohnern ausgewiesen. Kreisfreie Städte unterhalb dieser Größe werden mit ihrem Umland zu Kreisregionen zusammengefasst. Die Typisierung der Kreise und Kreisregionen erfolgt außerhalb der Kernstädte nach der Bevölkerungsdichte. Um den großräumigen Kontext zu berücksichtigen, wird dann weiter nach der Lage im siedlungsstrukturellen Regionstyp differenziert. Mit dieser Einordnung wird der Überlegung Rechnung getragen, dass die Lebensbedingungen in den Kreisen sowie ihre Entwicklung wesentlich auch von der Entwicklung und der Struktur der jeweiligen Region bzw. des Regionstyps abhängig sind. Insgesamt ergeben sich neun Kreistypen, die unter dem Merkmal BBR9 angegeben werden. Nach einer Zusammenfassung auf regionale Grundtypen ergeben sich drei Ausprägungen, die unter dem Merkmal BBR3 ausgewiesen werden. Vgl. Bundesamt für Bauordnung und Raumwesen (www.bbr.bund.de) unter raumordnung/raumbeobachtung

Um eine Ober- und Untergrenze für das mit der Mikroaggregation verbundene Reidentifikationsrisiko bestimmen zu können, werden die schwächste und die stärkste Variante der Mikroaggregation auf die Zieldaten angewendet. Während bei der schwächsten Variante (MA21G) jedes metrische Merkmal seine eigene Gruppe definiert, werden bei der stärksten Variante (MA1G) alle metrischen Merkmale zusammen gruppiert, sodass nach der Mikroaggregation Tripel von Merkmalsträgern entstehen, welche in allen metrischen Merkmalen übereinstimmen und sich lediglich in den kategorialen Merkmalen (wie z.B. *Wirtschaftszweigklassifikation* oder *Regionalkennung*) unterscheiden können. Tatsächlich kann man bei der Variante MA1G sehr große Abweichungen der anonymisierten von den Originaldaten beobachten. Mehr als 60 % der veränderten Einzelwerte weichen um mehr als 10 % von ihrem zugehörigen Originalwert ab. Bei der Variante MA21G hingegen werden die Originaldaten nur sehr geringfügig modifiziert. Hier weichen mehr als 99,9 % der veränderten Werte um weniger als 5 % von ihren zugehörigen Originalwerten ab. Dies gilt sogar für beinahe 90 % der Unternehmen mit mehr als 500 Beschäftigten, welche bekanntermaßen besonders reidentifikationsgefährdet sind.

5.3 Überprüfung der Schutzwirkung

Bei den nun folgenden Ergebnissen der Massenfischzüge werden zwei Arten von Trefferquoten unterschieden. Bei der ersten werden die Treffer in Beziehung zu allen gesuchten Unternehmen gesetzt. Bei der zweiten dagegen werden die Treffer lediglich im Verhältnis zu den Unternehmen betrachtet, die nicht aufgrund des „natürlichen Schutzes" per se sicher sind (vgl. Tabelle 4.16 auf Seite 111). Diese Trefferquote wird im Folgenden als „korrigierte Trefferquote" bezeichnet. Als Blockmerkmale werden die kategorialen Merkmale *Rechtsform, Wirtschaftszweigklassifikation* und *Regionalkennung* verwendet. Als metrisches Überschneidungsmerkmal steht das Merkmal *Gesamtumsatz* zur Verfügung.

5.3.1 Effekte durch Variation der Tiefe der wirtschaftlichen Gliederung

Die Tabelle 5.1 zeigt die Reidentifikationen in Abhängigkeit zur Tiefe der Wirtschaftszweigklassifikation, bis hin zum kompletten Verzicht auf die Branchenangabe („Nullsteller"). Letzteres kommt der Annahme gleich, dass ein Datenangreifer keinerlei Wissen über die ökonomischen Aktivitäten des gesuchten Unternehmens hat, außer dass es sich um ein Unternehmen des Verarbeitenden Gewerbes handelt.

Es überrascht nicht, dass die schwächste Form der Mikroaggregation (MA 21G) die Trefferquoten nicht reduzieren kann. Die Unterschiede zwischen den Erhebungen erschweren einen Datenangriff bereits so stark, dass eine zusätzliche minimale Veränderung der Einzelwerte durch MA21G bei einem Massenfischzug nicht ins Gewicht fällt. Betrachtet man die

Ergebnisse ohne Verwendung der Wirtschaftszweigklassifikation, so kann man sogar einen enthüllenden Effekt der Mikroaggregation beobachten: Es wurden mehr Unternehmen nach der Anonymisierung gefunden als vor der Anonymisierung (1 270 gegenüber 1 259). Die Variante MA1G dagegen generiert nahezu sichere Daten. Allerdings sei an dieser Stelle darauf hingewiesen, dass diese Mikroaggregationsvariante das Analysepotential zu stark einschränkt.

Die Trefferquoten bei der Verwendung einer vierstelligen und einer dreistelligen Wirtschaftszweigklassifikation ähneln sich bei den formal anonymisierten und den durch MA21G anonymisierten Daten sehr. Beim Übergang von vier auf drei Stellen werden zwar die Blöcke, innerhalb derer versucht wird, Unternehmen zu reidentifizieren, größer (und damit auch das Risiko der Falschzuordnung, vgl. Tabelle 5.2). Andererseits nimmt die Anzahl derjenigen Unternehmen ab, die bereits auf natürliche Weise geschützt sind (vgl. Tabelle 4.16). Diese beiden gegenläufigen Effekte heben sich in diesem Fallbeispiel auf. Wird die Wirtschaftszweigklassifikation jedoch weiter gekürzt, so kann ein Anstieg der Schutzwirkung beobachtet werden.

Tabelle 5.2 enthält die korrigierten Trefferquoten, d.h., es wird die reine Schutzwirkung aufgrund der getroffenen Anonymisierungsmaßnahmen für die metrischen Merkmale dargestellt. Es zeigt sich wiederum, dass durch die Variante MA21G die Reidentifikationen nicht wesentlich reduziert werden. Zusätzlich wird in der Tabelle die Schutzwirkung durch Vergröberung der Wirtschaftszweigklassifikation deutlich, da dieser Effekt nun nicht mehr durch einen gleichzeitigen Verlust an natürlichem Schutz untergraben wird. Daher ist erkennbar, dass die Trefferquoten bei einer Kürzung der Wirtschaftszweigklassifikation von vier auf drei Stellen zurückgehen. Die erste Zeile der Tabelle zeigt die Effektivität des verwendeten Matchingverfahrens. Ohne Inkompatibilitäten bei den Blockvariablen und ohne zusätzliche Anonymisierung kann der Algorithmus drei von vier Unternehmen richtig zuordnen. Bei den größten wurden sogar sämtliche Unternehmen reidentifiziert.

Tabelle 5.1
Reidentifikationen in Abhängigkeit zur Tiefe der wirtschaftlichen Gliederung (WZ 93)

Zieldaten	WZ 93	Gesamt	Beschäftigtengrößenklasse					
			1 – 24	25 – 99	100 – 999	1 000 – 4 999	5 000 – 14 999	≥ 15 000
Formal anonymisiert	4-Steller	0,40	0,35	0,36	0,46	0,55	0,58	0,70
Formal anonymisiert	3-Steller	0,40	0,36	0,36	0,45	0,53	0,65	0,60
Formal anonymisiert	2-Steller	0,35	0,32	0,32	0,40	0,54	0,62	0,80
Formal anonymisiert	1-Steller	0,21	0,18	0,19	0,23	0,43	0,35	0,40
Formal anonymisiert	0-Steller	0,14	0,12	0,12	0,15	0,29	0,42	0,40
MA21G	4-Steller	0,40	0,35	0,36	0,46	0,55	0,58	0,70
MA21G	3-Steller	0,40	0,36	0,36	0,45	0,53	0,58	0,80
MA21G	2-Steller	0,35	0,32	0,32	0,39	0,55	0,62	0,80
MA21G	1-Steller	0,21	0,18	0,18	0,23	0,42	0,35	0,50
MA21G	0-Steller	0,14	0,11	0,12	0,15	0,33	0,31	0,40
MA1G	4-Steller	0,28	0,21	0,22	0,35	0,54	0,65	0,60
MA1G	3-Steller	0,23	0,16	0,17	0,30	0,53	0,73	0,80
MA1G	2-Steller	0,14	0,07	0,10	0,19	0,47	0,69	0,60
MA1G	1-Steller	0,05	0,02	0,03	0,07	0,21	0,35	0,30
MA1G	0-Steller	0,03	0,01	0,02	0,03	0,12	0,27	0,30

Tabelle 5.2
Korrigierte Trefferquoten in Abhängigkeit zur Tiefe der wirtschaftlichen Gliederung (WZ 93)

Zieldaten	WZ 93	Gesamt	Beschäftigtengrößenklasse					
			1 – 24	25 – 99	100 – 999	1 000 – 4 999	5 000 – 14 999	≥ 15 000
Formal anonymisiert	4-Steller	0,73	0,71	0,70	0,75	0,85	0,83	1,00
Formal anonymisiert	3-Steller	0,64	0,62	0,61	0,67	0,76	0,81	0,75
Formal anonymisiert	2-Steller	0,48	0,46	0,44	0,50	0,68	0,73	1,00
Formal anonymisiert	1-Steller	0,25	0,23	0,23	0,27	0,49	0,39	0,50
Formal anonymisiert	0-Steller	0,14	0,12	0,12	0,15	0,30	0,42	0,44
MA21G	4-Steller	0,73	0,71	0,70	0,75	0,85	0,83	1,00
MA21G	3-Steller	0,64	0,62	0,61	0,66	0,77	0,71	1,00
MA21G	2-Steller	0,48	0,46	0,44	0,50	0,70	0,73	1,00
MA21G	1-Steller	0,25	0,23	0,23	0,26	0,49	0,36	0,63
MA21G	0-Steller	0,14	0,12	0,12	0,15	0,33	0,31	0,44
MA1G	4-Steller	0,51	0,43	0,43	0,58	0,83	0,94	0,86
MA1G	3-Steller	0,37	0,28	0,29	0,44	0,76	0,91	1,00
MA1G	2-Steller	0,19	0,10	0,13	0,24	0,59	0,82	0,75
MA1G	1-Steller	0,07	0,03	0,04	0,08	0,24	0,39	0,38
MA1G	0-Steller	0,03	0,01	0,02	0,04	0,12	0,27	0,33

Je größer einzelne Datenblöcke sind (d.h., je mehr Unternehmen eine bestimmte Kombination aus *Rechtsform, Regionalkennung* und *Wirtschaftszweigklassifikation* aufweisen), desto geringer ist die Wahrscheinlichkeit der korrekten Zuordnung eines Unternehmens (siehe Tabelle 5.3). Dies spricht dafür, eine Mindestanzahl von Unternehmen je Wirtschaftszweig nicht zu unterschreiten.

Tabelle 5.3: Korrelation zwischen Reidentifikationsrate und Besetzungszahlen der Wirtschaftszweige

	WZ 93		
Zieldaten	4-Steller	3-Steller	2-Steller
Formal anonymisiert	-0,504	-0,683	-0,777
MA21G	-0,499	-0,655	-0,779
MA1G	-0,474	-0,644	-0,661

Die Tabelle 5.4 enthält ausgewählte beschreibende Statistiken der korrekt zugeordneten Unternehmen. Zum Vergleich zeigt die erste Zeile die Statistik aller Unternehmen des Zusatzwissens. Man erkennt, dass mit steigendem Anonymisierungsgrad die durchschnittliche Beschäftigung innerhalb der jeweils reidentifizierten Unternehmen ansteigt. Betrachtet man allerdings die Größe des jeweils kleinsten gefundenen Unternehmens, so erkennt man, dass auch kleine Unternehmen korrekt zugeordnet werden können.

5.3.2 Effekte durch Vergröberung der Rechtsform

Im nächsten Schritt wird die Rechtsform zu vier Kategorien vergröbert und die Massenfischzüge erneut bei den formal anonymisierten und den via Mikroaggregation veränderten Daten wiederholt. Im Folgenden wird die Schutzwirkung dieser Vergröberung beschrieben.

Der Schutzeffekt, der mit einer Vergröberung der Rechtsform einhergeht, fällt deutlich schwächer aus als erwartet, da die Rechtsform in beiden Erhebungen für zueinander gehörige Merkmalsträger identisch ausgewiesen wird. Die Vergröberung konnte daher ihre volle Schutzwirkung entfalten und wurde nicht – wie bei der Vergröberung der Wirtschaftszweigklassifikation – durch einen geringeren „natürlichen Schutz" unterbunden. Ein Vergleich der Tabellen 5.1 und 5.5 zeigt, dass eine Vergröberung der Wirtschaftszweigklassifikation eine höhere Schutzwirkung als eine Vergröberung der Rechtsform bewirkt. Die durch Vergröberung der Rechtsform einhergehende Anonymisierung hat wiederum zur Folge, dass die durchschnittliche Größe der korrekt zugeordneten Unternehmen ansteigt. Auch hier werden demnach besonders kleinere Unternehmen geschützt.

Tabelle 5.4
Beschreibende Statistik der reidentifizierten Unternehmen

Zieldaten	WZ 93	Reidentifikation des größten Unternehmens	Durchschnittl. Größe	Kleinstes Unternehmen	Standardabweichung	Anzahl der Unternehmen
			Anzahl der Beschäftigten			
Formal anonymisiert	4-Steller	ja	431,3	20	4 326,3	3 726
Formal anonymisiert	3-Steller	ja	421,8	20	4 279,4	3 720
Formal anonymisiert	2-Steller	ja	488,6	20	4 673,4	3 287
Formal anonymisiert	1-Steller	nein	445,5	20	5 469,0	1 951
Formal anonymisiert	0-Steller	nein	626,0	20	6 897,7	1 259
MA21G	4-Steller	ja	433,2	20	4 328,9	3 723
MA21G	3-Steller	ja	443,3	20	4 396,3	3 709
MA21G	2-Steller	ja	491,5	20	4 677,9	3 282
MA21G	1-Steller	ja	532,4	20	8 924,5	1 934
MA21G	0-Steller	ja	644,5	20	5 883,1	1 270
MA1G	4-Steller	ja	561,3	20	5 117,5	2 593
MA1G	3-Steller	ja	688,4	20	5 743,1	2 169
MA1G	2-Steller	ja	936,6	20	7 126,3	1 332
MA1G	1-Steller	ja	1 200,6	21	8 924,5	497
MA1G	0-Steller	ja	1 946,6	20	12 750,1	241
Alle Unternehmen	–	–	295,8	20	2 888,7	9 283

Tabelle 5.5
Reidentifikationen nach Vergröberung der Rechtsform zu vier Kategorien

Zieldaten	WZ 93	Gesamt	Beschäftigtengrößenklasse					
			1 – 24	25 – 99	100 – 999	1 000 – 4 999	5 000 – 14 999	≥ 15 000
Formal anonymisiert	4-Steller	0,38	0,33	0,33	0,44	0,53	0,50	0,60
Formal anonymisiert	3-Steller	0,38	0,33	0,34	0,43	0,51	0,70	0,70
Formal anonymisiert	2-Steller	0,32	0,27	0,28	0,36	0,49	0,46	0,30
Formal anonymisiert	1-Steller	0,17	0,44	0,15	0,19	0,37	0,31	0,20
Formal anonymisiert	0-Steller	0,11	0,10	0,10	0,13	0,13	0,27	0,20
MA21G	4-Steller	0,38	0,33	0,33	0,44	0,53	0,50	0,60
MA21G	3-Steller	0,38	0,33	0,34	0,43	0,50	0,58	0,70
MA21G	2-Steller	0,32	0,27	0,28	0,36	0,49	0,42	0,40
MA21G	1-Steller	0,17	0,14	0,14	0,19	0,35	0,31	0,50
MA21G	0-Steller	0,11	0,10	0,10	0,12	0,27	0,27	0,50
MA1G	4-Steller	0,25	0,19	0,19	0,33	0,51	0,58	0,50
MA1G	3-Steller	0,20	0,13	0,14	0,28	0,46	0,65	0,70
MA1G	2-Steller	0,11	0,04	0,15	0,15	0,38	0,42	0,70
MA1G	1-Steller	0,03	0,01	0,02	0,05	0,16	0,23	0,60
MA1G	0-Steller	0,02	0,01	0,02	0,03	0,09	0,12	0,40

5.3.3 Effekte durch Vergröberung der Regionalkennung

Neben der Vergröberung der Rechtsform stellt die Vergröberung der Regionalkennung eine weitere Anonymisierungsmaßnahme dar. Als Regionalkennung wurde wie erwähnt der siedlungsstrukturelle Kreistyp verwendet. Dieser besteht aus einer Hauptstufe mit drei Ausprägungen und einer Unterstufe, durch die der BBR9 seine insgesamt neun Ausprägungen erhält. Durch den Verzicht auf die Unterstufe reduziert sich die Anzahl der Ausprägungen auf drei. Im Folgenden wird daher vom BBR3 gesprochen, wenn auf die Unterstufe verzichtet wird. Alternativ könnte auch der regionsstrukturelle Kreistyp mit sieben Ausprägungen (BBR7) verwendet werden. Der Zusammenhang zwischen den verschiedenen Regionalschlüsseln des Bundesamtes für Bauordnung und Raumwesen mit drei, sieben und neun Ausprägungen ist in Abbildung 5.1 dargestellt. Die Schutzwirkung, die durch Verwendung des BBR7 entsteht, ist allerdings vernachlässigbar, sodass auf eine weitere Diskussion auf Basis des BBR7 verzichtet werden kann. Tabelle 5.6 enthält die Ergebnisse der Massenfischzüge.

5.4 Bewertung der Ergebnisse

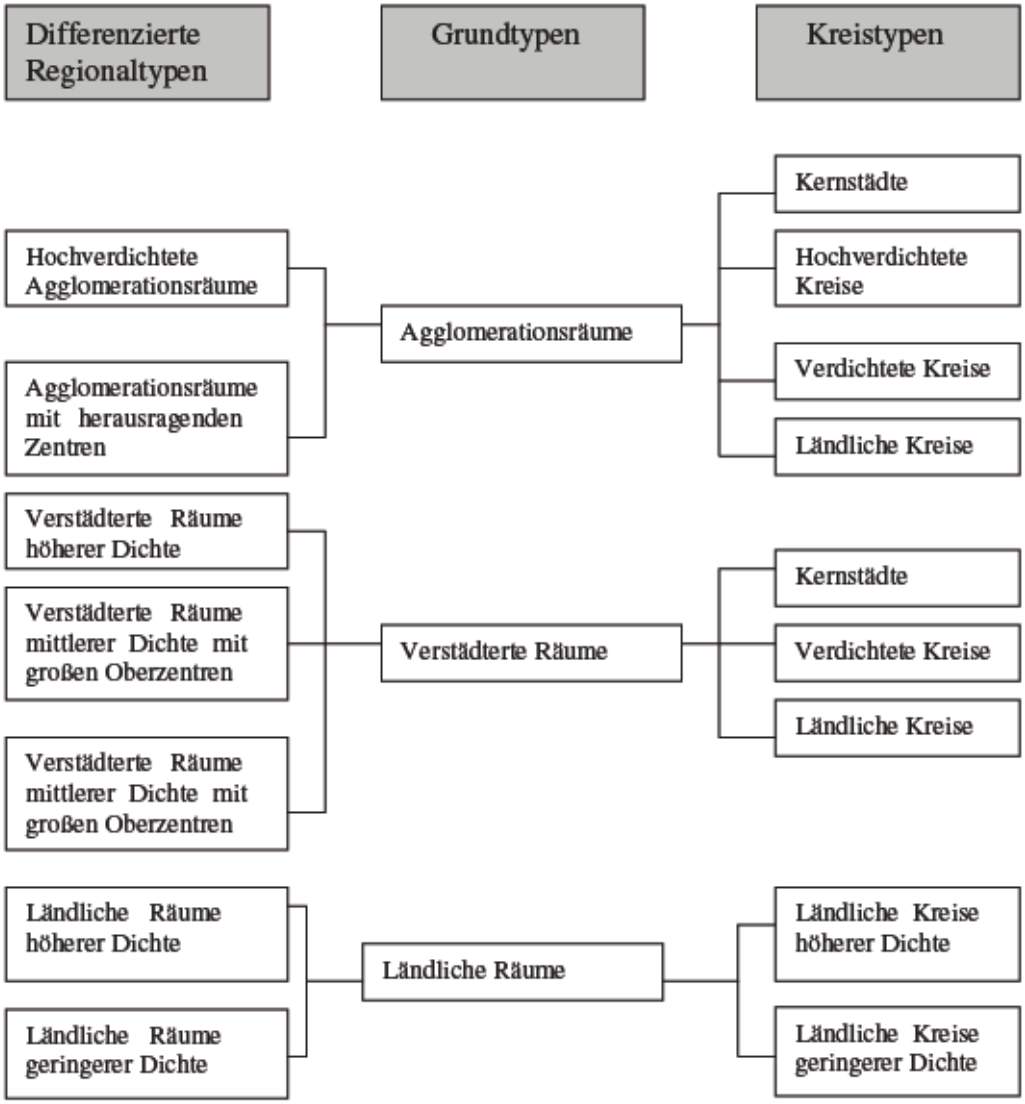
Mit der Vergröberung der Regionalkennung wird eine größere Anonymisierungswirkung erzielt als mit der vorhergehenden Vergröberung der Rechtsform. Dies verdeutlicht die Bedeutung von Regionalangaben bei einer versuchten Deanonymisierung. Mit Blick auf Tabelle 5.6 sieht man, dass eine Anonymisierung durch Vergröberung der Wirtschaftszweigklassifikation von vier auf zwei Stellen sowie durch Vergröberung der Rechtsform und der Regionalkennung zu einem Rückgang der Reidentifikationsrate von über 40% führt, was zeigt, dass informationsreduzierende Maßnahmen auch bei der Anonymisierung wirtschaftsstatistischer Einzeldaten eine wichtige Rolle spielen.

Es zeigt sich insgesamt, dass die Methoden der Informationsreduktion eine beachtenswerte Schutzwirkung entfalten. In Abstimmung mit den Datennutzern sollten die Datenanbieter daher zunächst informationsreduzierende Methoden einsetzen. In einem ersten Schritt sollten durch den Nutzer nicht benötigte Merkmale aus der Datei entfernt werden. In einem zweiten Schritt wird eine (weitere) Vergröberung der kategorialen Überschneidungsmerkmale empfohlen sowie eine eventuelle Kategorisierung metrischer Überschneidungsmerkmale wie *Gesamtumsatz, Anzahl der Beschäftigten* (z.B. ausgewiesen in Größenklassen) oder *Aufwendungen in Forschung und Entwicklung* (ja/nein). Erst danach sollten mit Bedacht ausgewählte datenverändernde Methoden eingesetzt werden.

Tabelle 5.6
Reidentifikationen nach Vergröberung der Rechtsform und der Regionalkennung

Zieldaten	WZ 93	Gesamt	Beschäftigtengrößenklasse					
			1 – 24	25 – 99	100 – 999	1 000 – 4 999	5 000 – 14 999	≥ 15 000
Formal anonymisiert	4-Steller	0,34	0,28	0,29	0,39	0,50	0,54	0,70
Formal anonymisiert	3-Steller	0,31	0,27	0,27	0,36	0,49	0,46	0,60
Formal anonymisiert	2-Steller	0,23	0,18	0,20	0,26	0,38	0,39	0,20
Formal anonymisiert	1-Steller	0,10	0,08	0,08	0,11	0,24	0,27	0,20
Formal anonymisiert	0-Steller	0,06	0,06	0,05	0,06	0,15	0,19	0,10
MA21G	4-Steller	0,34	0,28	0,29	0,39	0,50	0,54	0,70
MA21G	3-Steller	0,31	0,27	0,27	0,36	0,49	0,46	0,60
MA21G	2-Steller	0,23	0,18	0,20	0,26	0,37	0,39	0,50
MA21G	1-Steller	0,10	0,08	0,08	0,11	0,23	0,35	0,60
MA21G	0-Steller	0,06	0,06	0,05	0,06	0,13	0,23	0,30
MA1G	4-Steller	0,18	0,12	0,12	0,25	0,46	0,54	0,60
MA1G	3-Steller	0,14	0,08	0,08	0,19	0,42	0,58	0,80
MA1G	2-Steller	0,06	0,02	0,04	0,08	0,24	0,42	0,70
MA1G	1-Steller	0,02	0,01	0,01	0,02	0,08	0,15	0,60
MA1G	0-Steller	0,01	0,01	0,01	0,01	0,04	0,19	0,30

Abbildung 5.1
Typisierung der Regionen



Kapitel 6

Modellierung von Datenangriffsszenarien mit anonymisierten Paneldaten

Mit einem „Panel" bezeichnet man ganz allgemein eine Erhebung, in der dieselben Untersuchungseinheiten (Personen, Haushalte, Betriebe, Unternehmen) in mehreren aufeinanderfolgenden Zeitpunkten bezüglich verschiedener Erhebungsmerkmale beobachtet oder befragt werden. In der modernen Zeitreihenanalyse wird der Begriff Panel unter dem Stichwort „Kointegration in Paneldaten" auch für Volkswirtschaften insgesamt verwendet. Allerdings ist damit die ursprüngliche Bedeutung, die sich nur auf Mikrodaten bezog, aufgeweicht worden.

Paneldaten werden von der Wissenschaft zunehmend nachgefragt. Das Interesse an der Analyse von Paneldaten ist vor allem deshalb gewachsen, weil damit die folgenden zwei bedeutsamen Erweiterungen der Analysemöglichkeiten gegenüber der Verwendung von Mikrodaten im Querschnitt ermöglicht werden (Ronning et al. 2009):

- 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 nahezu 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. Zum Beispiel könnte 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 in diesem Beispiel ist, dass diese Eigenschaften (z.B. das Geschlecht bei Personen

oder die Branchenzugehörigkeit bei Unternehmen) in der Regel über die Zeit hin konstant sind. Dies erlaubt die Separierung dieser Einflussgrößen von anderen, zeitlich variierenden Regressoren, etwa dem Einkommen oder dem Umsatz. Im Falle unbeobachtbarer Einflüsse greift man zu speziellen Techniken, die unter dem Begriff der unbeobachteten Heterogenität bekannt sind.

Während die Arbeiten des Projektes „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" (FAWE, 2002 – 2005) erfreulicherweise gezeigt haben, dass eine faktische Anonymisierung von wirtschaftsstatistischen Einzeldaten im Querschnitt gelingen kann, musste dieser Nachweis für im Längsschnitt verknüpfte Daten noch erbracht werden. Aus diesem Grunde wurde in den Folgejahren 2006 – 2008 das Forschungsprojekt „Wirtschaftsstatistische Paneldaten und faktische Anonymisierung (FAWE-Panel)" unter der wissenschaftlichen Leitung von Prof. Dr. Gerd Ronning und der Projektleitung durch den Autor durchgeführt. In diesem Kapitel werden die Untersuchungen hinsichtlich der faktischen Anonymisierbarkeit von Daten solcher Struktur besprochen. Die aufwendigen Arbeiten zur Verknüpfung und Aufbereitung der in großen Teilen nur im Querschnitt verfügbaren Erhebungen sowie zur Systematisierung des bei Paneldaten zu erhaltenden Analysepotentials werden in Ronning et al. (2009) sowie in einer Vielzahl von Einzelveröffentlichungen zum Thema beschrieben. Die Methoden zur Anonymisierung von Daten solcher Struktur werden systematisch in Höhne (2008) und Höhne (2010) entwickelt.

Da ein gut vorbereiteter Datenangreifer nunmehr Merkmalsinformationen über mehrere Zeitpunkte in das Datenangriffsszenario einbringen kann, ist es naheliegend, diese komplexere Struktur auch in den Koeffizienten der Zielfunktion des linearen Programmes (3.22) abzubilden. Hierzu werden in Abschnitt 6.1 verschiedene Ansätze ins Auge gefasst und in Abschnitt 6.2 miteinander kombiniert. Abschließend folgen in Abschnitt 6.3 ein paar Bemerkungen zur Sensitivität hinsichtlich der vorgenommenen Parametrisierung sowie in Abschnitt 6.4 eine Abschätzung der bei den Simulationen theoretisch zu erwartenden Rechenzeit.

6.1 Ansätze zur Koeffizientenberechnung

Die Hauptaufgabe besteht nun darin, die Koeffizienten der Zielfunktion des linearen Programmes der Struktur von Paneldaten angemessen zu wählen. Bei der Behandlung von Querschnittsdaten hat sich die Unterscheidung zwischen metrischen und kategorialen Merkmalen bewährt. Bei Paneldaten jedoch stellt sich hinsichtlich der Zuweisung geeigneter Distanzmaße eine Unterscheidung zwischen nominalen und ordinalen Merkmalen als sinnvoller heraus.

Zunächst wird in Unterabschnitt 6.1.1 auf mögliche Distanzmaße für nominale Merkmale eingegangen. Die in den darauf folgenden Unterabschnitten 6.1.2 bis 6.1.4 vorgestellten Ansätze für ordinale Merkmale sind sowohl auf metrische als auch auf kategoriale Merkmale

mit linear geordnetem Wertebereich anwendbar. Nach Berechnung der (zuvor standardisierten) Maße $d_{nom}(a,b)$ für nominale und $d_{ord}(a,b)$ für ordinale Merkmale werden diese konvex kombiniert:

$$d(a,b) = \lambda d_{nom}(a,b) + (1-\lambda) d_{ord}(a,b),\ \lambda \in [0,1]. \tag{6.1}$$

Da wirtschaftsstatistische Einzeldaten überwiegend aus ordinalen Merkmalen bestehen, liegt der Schwerpunkt auf der Beschreibung von Distanzmaßen für solche Merkmale. Während sich bei der Untersuchung von Querschnittsdaten ein Distanzmaß, gebildet durch alle Überschneidungsmerkmale beider Datenquellen, bei der Koeffizientenbestimmung bewährt hat, erscheint es nun nötig, für die Behandlung von Paneldaten weitere Zusammenhangsmaße heranzuziehen.

In den nachfolgenden Unterabschnitten werden stets mit $\tilde{a} = (a_1, \ldots, a_k)$ und $\tilde{b} = (b_1, \ldots, b_k)$ die k-Tupel der Ausprägungen eines externen Merkmalsträgers a und eines Merkmalsträgers b der Zieldaten in k Überschneidungsmerkmalen bezeichnet.

6.1.1 Behandlung nominaler Merkmale

Grundsätzlich können die Distanzmaße für nominale Merkmale aus Unterabschnitt 3.2.1 auch bei Paneldaten eingesetzt werden. Dies gilt insbesondere für solche Maße, die auf Merkmale mit teilweise geordnetem Wertebereich abstellen. Zusätzlich kann es aber durchaus sinnvoll sein, Merkmale, die bei Querschnittsdaten ausschließlich zur Blockung geeignet erschienen, nun ebenfalls in Ähnlichkeits- bzw. Distanzmessungen einzubeziehen. Man betrachte zum Beispiel das hierarchisch gegliederte Merkmal *Wirtschaftszweigklassifikation*, beobachtet über mehrere Jahre auf der Vier- oder Fünfstellerebene. Ein zwischenzeitlicher Wechsel der Branchenzugehörigkeit auf diesen Ebenen seitens der Zieldaten passiert nicht selten, wohingegen diese Veränderung im Zusatzwissen erfahrungsgemäß nicht abgebildet wird (vgl. Abschnitt 7.2). Bei einer Blockung über dieses Merkmal würden daher viele Paare zueinander gehörender Einheiten in verschiedene Blöcke klassifiziert und damit a priori von dem späteren Zuordnungsverfahren ausgeschlossen. Desweiteren können solche Fehlklassifikationen durch den Einsatz datenverändernder Anonymisierungsverfahren für nominale Merkmale wie beispielsweise PRAM entstehen.

Wir richten im Folgenden ohne Beschränkung der Allgemeinheit den Fokus auf dichotome Merkmale (mit den beiden Ausprägungen 0 und 1), da jedes nominale Merkmal v mit endlichem Wertebereich, etwa $|v| = k$, in k binäre Merkmale $v_1, \ldots, v_k$ zerlegt werden kann. Sei hierzu a ein beliebiger Merkmalsträger. Wir definieren

$$v_i(a) \quad := \quad \begin{cases} 1, & \text{wenn } v(a) = i, \\ 0 & \text{sonst.} \end{cases} \tag{6.2}$$

Für ein dichotomes Überschneidungsmerkmal v_i nimmt das Tupel (a_i, b_i) eine der Kombinationen $(0,0), (0,1), (1,0)$ oder $(1,1)$ an. Das Auftreten einer jeden Kombination

wird nun über alle Überschneidungsmerkmale hinweg gezählt:

$\alpha = \sum_{i=1}^{k} \min\{a_i, b_i\}$ (Anzahl der Positionen, an denen $\tilde{a}$ und $\tilde{b}$ den Wert 1 annehmen).

$\beta = (\sum_{i=1}^{k} a_i) - \alpha$ (Anzahl der Positionen, an denen $\tilde{a}$ den Wert 1 und $\tilde{b}$ den Wert 0 annimmt).

$\gamma = (\sum_{i=1}^{k} b_i) - \alpha$ (Anzahl der Positionen, an denen $\tilde{a}$ den Wert 0 und $\tilde{b}$ den Wert 1 annimmt).

$\delta = k - (\alpha + \beta + \gamma)$ (Anzahl der Positionen, an denen $\tilde{a}$ und $\tilde{b}$ den Wert 0 annehmen).

Nun kann eine Distanzfunktion wie folgt gebildet werden (vgl. Schweigert 1999):

$$d(\tilde{a}, \tilde{b}) = 1 - \frac{\alpha + s\delta}{\alpha + s\delta + t(\beta + \gamma)} \tag{6.3}$$

mit den beiden Kontrollparametern $0 \leq s \leq 1$ und $t \geq 0$, durch welche man die Distanzfunktion gezielt beinflussen kann. Mit $s = t = 1$ ergibt sich für den Quotienten die relative Häufigkeit der Übereinstimmungen beider Merkmalsträger; dieser naheliegende Ansatz wird in der englischsprachigen Literatur oftmals mit „Simple Matching" bezeichnet. Als weitere Spezialfälle oder verwandte Maße tauchen in der Literatur auf:

$$\begin{aligned}
d_{nom}(\tilde{a}, \tilde{b}) &= 1 - \frac{\alpha}{k} && \text{(Russel, Rao)} \\
d_{nom}(\tilde{a}, \tilde{b}) &= 1 - \frac{\alpha}{\alpha + \beta + \gamma} && \text{(Jaccard, Tanimoto)} \\
d_{nom}(\tilde{a}, \tilde{b}) &= 1 - \frac{2\alpha}{2\alpha + \beta + \gamma} && \text{(Dice)} \\
d_{nom}(\tilde{a}, \tilde{b}) &= 1 - \frac{\alpha}{\alpha + 2\beta + 2\gamma} && \text{(Sneath, Sokal)} \\
d_{nom}(\tilde{a}, \tilde{b}) &= 1 - \frac{\alpha + \delta}{k + \beta + \gamma} && \text{(Tanimoto)}
\end{aligned} \tag{6.4}$$

Beispiel: Die Regionalangabe des Hauptstandortes von Unternehmen werde über acht Jahre mit den dazugehörigen Merkmalen $v_1, \ldots, v_8$ erhoben, wobei der Standort über die Jahre zwischen den alten (0) und neuen (1) Bundesländern wechseln kann. Nachfolgende Tabelle enthalte die Ausprägungen der acht Merkmale für je einen Merkmalsträger a der externen Daten und b der Zieldaten.

	v_1	v_2	v_3	v_4	v_5	v_6	v_7	v_8
a_i	1	1	0	0	1	1	1	1
b_i	1	1	1	0	0	0	1	1

Durch Berechnung der in (6.4) aufgelisteten Maße entsteht folgende Tabelle 6.1:

Tabelle 6.1: Distanzmaße für nominale Merkmale am Beispiel

	s	t	$d_{nom}(a,b)$
Simple Matching	1	1	$\frac{3}{8}$
Russel, Rao	−−	−−	$\frac{1}{2}$
Jaccard, Tanimoto	0	1	$\frac{3}{7}$
Dice	0	0.5	$\frac{3}{11}$
Sneath, Sokal	0	2	$\frac{3}{5}$
Tanimoto	1	2	$\frac{6}{11}$

Das kleine Beispiel zeigt, dass die Wahl eines geeigneten Distanzmaßes auch bei nominalen Merkmalen wesentlich ist.

6.1.2 Konventioneller Ansatz

Bei der Untersuchung von Querschnittsdaten haben sich die in Unterabschnitt 3.2.2 vorgestellten konventionellen Distanzmaße, darunter insbesondere die bei Gleichgewichtung aller Überschneidungsmerkmale enstehende euklidische Distanz, durchgesetzt.

Der konventionelle Ansatz weist bei der Betrachtung von Paneldaten bereits dann Schwächen auf, wenn die Definition eines Überschneidungsmerkmales in den beiden Datenquellen systematische Abweichungen hervorruft. Etwa wenn bei dem Merkmal *Anzahl der Beschäftigten* auf der einen Seite die Absolutzahl aller Beschäftigten ausgewiesen und auf der anderen Seite Teil- in Vollzeitbeschäftigte umgerechnet wurden. Ein weiteres Beispiel ist gegeben, wenn auf der einen Seite der *Gesamtumsatz* und auf der anderen Seite der Gesamtumsatz abzüglich eines bestimmten Teilumsatzes abgefragt wurde.

Ein weiterer Nachteil besteht darin, dass dieser Ansatz die multivariaten Zusammenhänge zwischen Überschneidungsmerkmalen nicht ausreichend berücksichtigt.

6.1.3 Korrelationsbasierter Ansatz

Seien $v_1^e, \ldots, v_k^e$ und $v_1^t, \ldots, v_k^t$ ordinale Überschneidungsmerkmale der externen Daten und der vertraulichen Zieldaten.

Wir interpretieren im Folgenden (v^e, v^t) als zweidimensionales Merkmal, das k-mal realisiert wurde und berechnen den empirischen Korrelationskoeffizienten nach Spearman. Dieser Koeffizient kann sowohl im Falle metrischer als auch kategorialer ordinaler Merkmale

eingesetzt werden. Seien weiterhin $\tilde{a} = (a_1, \ldots, a_k)$ und $\tilde{b} = (b_1, \ldots, b_k)$ die zu den Merkmalen v^e und v^t gehörigen Ausprägungen zweier Merkmalsträger a und b. Je kleiner

$$d_{ord}(a, b) = 1 - corr(\tilde{a}, \tilde{b}) \tag{6.5}$$

ist, desto wahrscheinlicher erscheint eine korrekte Zuordnung.

Beispiel: Seien $\tilde{a} = (a_1, \ldots, a_4)$ und $\tilde{b} = (b_1, \ldots, b_4)$ die zu den Merkmalsträgern a der externen Daten und b der Zieldaten gehörigen Quadrupel des an vier Zeitpunkten beobachteten Überschneidungsmerkmals *Anzahl der Beschäftigten*. Die i-te Komponente enthalte die Mitarbeiterzahl zum Zeitpunkt i:

a_i	570	542	512	477
b_i	626	523	565	454

Das Paar (a, b) weist gemäß Gleichung (6.5) eine Distanz von $d_{ord}(a, b) = 0.2$ in diesem Merkmal auf.

6.1.4 Verteilungsbasierter Ansatz

Bei der Anonymisierung von Paneldaten muss davon ausgegangen werden, dass einem Datenangreifer zu jedem Überschneidungsmerkmal Informationen über mehrere Jahre vorliegen, z.B. der Gesamtumsatz eines Unternehmens des Verarbeitenden Gewerbes in den Jahren 1999 bis 2002. Wie bereits in Unterabschnitt 6.1.2 angesprochen wurde, sind systematische Abweichungen zwischen den beiden verschiedenen Datenquellen A und B möglich. Diesem Problem kann begegnet werden, indem man die Entwicklung des Merkmals – über die Jahre beobachtet – als relative Häufigkeitsverteilung interpretiert. Nun kann mit statistischen Methoden der Grad an „Ähnlichkeit" der beiden Verteilungen auf Seiten der externen Daten und der Zieldaten geschätzt werden.

Im Folgenden sei $m := n \cdot k$ und $\tilde{v}^t = (v_1^t, \ldots, v_m^t)$, wobei n für die Anzahl der Jahre und k für die Anzahl der ordinalen über alle Jahre verfügbaren Überschneidungsmerkmale steht. Im Falle ordinaler und nicht-metrischer Merkmale werden die Merkmalswerte durch ihre Rangplätze analog zur Berechnung der Rangkorrelationen nach Spearman ersetzt. Somit können $\tilde{v}^t, \tilde{v}^e$ als kategoriale Merkmale interpretiert werden, deren Beobachtungen sich auf m Klassen $c_1, \ldots, c_m$ verteilen. Hierbei steht v_i^t für die (möglicherweise nicht ganzzahlige) Häufigkeit der Beobachtungen der Zieldaten, die in Klasse c_i landen. Unter Verwendung derselben Notationen für $\tilde{a} = (a_1, \ldots, a_m)$ und $\tilde{b} = (b_1, \ldots, b_m)$ wie in den vorangegangenen Unterabschnitten wird mit

$$\chi^2_*(\tilde{a}, \tilde{b}) \quad = \quad \sum_{i=1}^{m} \frac{(a_i - b_i)^2}{b_i} \tag{6.6}$$

die bekannte Statistik des Chi-Quadrat Anpassungstestes berechnet mit dem Ziel, für jedes Paar $(a, b) \in A \times B$ von Merkmalsträgern die Abweichung der Verteilung von a auf die Klassen c_1 bis c_m von der entsprechenden Verteilung von b zu messen.

Betrachtet man das Beipiel aus Unterabschnitt 6.1.3, so ergibt sich für das Paar (a, b) das Maß $d_{ord}(a, b) \approx 11.84$.

Der Wertebereich sollte hinreichend groß sein, um vernünftig differenzieren zu können. Probleme können beim verteilungsbasierten Ansatz insbesondere dann entstehen, wenn die Daten fehlende Werte aufweisen, sodass bei der Berechnung des χ^2-Maßes Nullen im Nenner auftauchen. Hier wird empfohlen, auf eine andere Statistik auszuweichen (z.B. Kolmogorov-Smirnov), basierend auf der empirischen Verteilungsfunktion F_N. Diese Funktion setzt sich aus den Werten v_i^t, die weiterhin als Häufigkeiten interpretiert werden, zusammen.

Der verteilungsbasierte Ansatz weist weitere Schwächen auf, wenn die Überschneidungsmerkmale hohe Varianzen besitzen. In diesem Falle ist von dem Ansatz Abstand zu nehmen oder auf ein vereinfachtes Maß, das allein auf das Monotonieverhalten der Merkmale über die Jahre abstellt, zurückzugreifen. Etwa könnte man Vektoren vergleichen, die an erster Position eine 0 und an den weiteren $k-1$ Positionen die Einträge 0, 1 oder 2 besitzen, wobei der Eintrag 1 (bzw. 2) an Position i bedeutet, dass der Merkmalswert von v_i gegenüber v_{i-1} (d.h., von Zeitpunkt $i-1$ zu Zeitpunkt i) gewachsen (bzw. gefallen) ist. Bleibt der Wert unverändert (was zum Beispiel bei Beschäftigtenangaben wahrscheinlicher ist als bei Umsatzangaben), so wird der Eintrag 0 gesetzt. Besonders gut anwendbar ist der Ansatz, wenn die Anonymisierungsstrategie so ausgelegt wurde, dass sie kurzfristige Trends in den Überschneidungsmerkmalen erhält. In diesem Falle kann eine Abweichung zweier Merkmalsträger in dem vereinfachten Maß sogar als Ausschlusskriterium verwendet werden.

Beispiel: In einer Unternehmenserhebung enthalten die Merkmale $v_k, v_{k+1}, \ldots, v_{k+4}$ den Gesamtumsatz der erfassten Unternehmen für die Jahre $1998, 1999, \ldots, 2002$ in Mill. Euro. Für ein spezielles Unternehmen a der externen Daten werde der zu den Merkmalen gehörige Vektor $(2, 2.5, 2.2, 2.2, 3)$, für ein spezielles Unternehmen b der Zieldaten der Vektor $(1.3, 2.6, 1.8, 1.8, 2)$ beobachtet. Offenbar entsteht bei der Betrachtung des Monotonieverhaltens in beiden Fällen dasselbe Quadrupel $(1, 2, 0, 1)$. Somit ergibt sich für das vereinfachte Maß der Wert $d_{ord} = 0$.

6.1.5 Kollinearitätsbasierter Ansatz

Nehmen wir zunächst an, einem Datenangreifer würden Informationen zweier Überschneidungsmerkmale über einen vorgegebenen Zeitraum, z.B. Gesamtumsätze $(u_1, \ldots, u_N)$ und Anzahl der Beschäftigten $(b_1, \ldots, b_N)$ eines Unternehmens in N Jahren, vorliegen. Interpretiert man die Wertepaare (u_i, b_i) als Realisierungen zweier zufälliger Variablen u und b, so ist zu erwarten, dass zueinander gehörige Merkmalsträger aus verschiedenen Datenquellen ähnliche empirische Korrelationskoeffizienten aufweisen werden.

Seien $(u_i^t(a), b_i^t(a)), i = 1, \ldots, N$, die Ausprägungen der beiden Merkmale u^t und b^t in den Zieldaten (für einen speziellen Merkmalsträger a) und $(u_i^e(b), b_i^e(b)), i = 1, \ldots, N$, die entsprechenden Ausprägungen der Merkmale u^e und b^e in den externen Daten (für einen speziellen Merkmalsträger b). Dann ist im Falle

$$corr(u^e(a), b^e(a)) \approx corr(u^t(b), b^t(b)) \tag{6.7}$$

eine korrekte Zuordnung wahrscheinlich.

Durch die empirische Korrelation wird nur ein linearer Zusammenhang geschätzt. Eine Korrelation nahe bei Null schließt aber nicht aus, dass es einen anderen Zusammenhang zwischen den Merkmalen geben kann. In Spezialfällen können die beiden geschätzten Korrelationskoeffizienten sehr deutlich voneinander abweichen, auch wenn es einen direkten nichtlinearen funktionalen Zusammenhang zwischen den Merkmalen gibt.

Eine Verallgemeinerung des Ansatzes ergibt sich, wenn mehr als zwei Gruppen ordinaler Überschneidungsmerkmale vorliegen. Das k-Tupel $(\alpha_1^t(a), \ldots, \alpha_k^t(a))$ bezeichne die Ausprägungen der k Gruppen von Überschneidungsmerkmalen im Merkmalsträger a der Zieldaten und $(\alpha_1^e(b), \ldots, \alpha_k^e(b))$ die entsprechenden Ausprägungen im Merkmalsträger b der externen Daten. Damit kann α_i^t als ein zu N Zeitpunkten $t = 1, \ldots, N$ beobachtetes Merkmal interpretiert werden; d.h., $\alpha_{i1}^t, \ldots, \alpha_{iN}^t$ sind Realisierungen von α_i^t. Dann ist im Falle

$$coll(\alpha_1^e(a), \ldots, \alpha_k^e(a)) \approx coll(\alpha_1^t(b), \ldots, \alpha_k^t(b)), \tag{6.8}$$

wobei $coll$ ein geeignetes Maß für die lineare Abhängigkeit der Spaltenvektoren bezeichne, eine korrekte Zuordnung wahrscheinlich.

In der Literatur findet man verschiedene Ansätze zur Messung des Grades an linearer Abhängigkeit, etwa über kanonische Korrelationen oder aufwendige Determinantenberechnungen (siehe Lenz 2008). Die Situation des Vorliegens von mehr als zwei Gruppen ordinaler Überschneidungsmerkmale ist jedoch in der bisherigen Praxis noch nicht aufgetreten. Daher gibt es hierzu keine empirischen Ergebnisse.

6.2 Kombinierte Zuordnungsverfahren

Die in den vorherigen Unterabschnitten aufgeführten Ansätze haben den Nachteil, dass sie nur einen Teil der in den Daten verfügbaren Informationen nutzen. Daher erscheint es sinnvoll, verschiedene Ansätze derart zu kombinieren, dass die Sub-Ansätze ihre Schwächen untereinander kompensieren, möglichst bei gleichzeitigem Erhalt ihrer Stärken. Zur Kombination der Ansätze können sowohl sogenannte *hybride* als auch *zusammengesetzte Zuordnungsverfahren* dienen.

6.2.1 Hybride Zuordnungsverfahren

Bei der hybriden Zuordnung werden mehrere vielversprechende Kriterien simultan erfüllt. Zum Beispiel können verschiedene Distanzmaße zu einem Gesamtmaß kombiniert und im Anschluss via Lösung des Zuordnungsproblems (3.22) Paare gebildet werden. Es seien (a, b) ein Paar von Merkmalsträgern und $d_{conv}(a, b)$, $d_{chi2}(a, b)$, $d_{corr}(a, b)$ und $d_{coll}(a, b)$ die zu den in 6.1.2 bis 6.1.5 vorgestellten Maßen gehörigen Koeffizienten. Zunächst sollten diese Maße auf eine gemeinsame Skala standardisiert werden. Danach besteht der einfachste Weg einer Zusammenführung der Maße in einer Konvexkombination (d.h., einer Gewichtung wie beispielsweise beim arithmetischen Mittel)

$$d_{ord}(a, b) = \lambda_1 d_{conv}(a, b) + \lambda_2 d_{chi2}(a, b) + \lambda_3 d_{corr}(a, b) + \lambda_4 d_{coll}(a, b) \tag{6.9}$$

mit $\lambda_i \geq 0$, $i = 1, \ldots 4$, und $\sum_{j=1}^{4} \lambda_j = 1$. Denkbar sind auch Kombinationen wie

$$\begin{aligned} d_1(a, b) &:= \min(d_{conv}(a, b), d_{chi2}(a, b), d_{corr}(a, b), d_{coll}(a, b)) \quad \text{oder} \\ d_2(a, b) &:= \max(1_{[\gamma, \infty)}(d_{conv}(a, b)), d_{chi2}(a, b), d_{corr}(a, b)), \end{aligned}$$

wobei

$$1_{\{x|x>\gamma\}}(x) \quad = \quad \begin{cases} 1 & \text{wenn } x > \gamma \\ 0 & \text{sonst} \end{cases} \tag{6.10}$$

die charakteristische Funktion auf der Menge $\{x \mid x > \gamma\}$ bezeichne. Bei dem Maß d_2 erhält ein Paar (a, b) den größtmöglichen Wert Eins und wird als mögliches Kandidatenpaar ausgeschlossen, wenn die Distanz d_{conv} eine vorgegebene Schwelle γ überschreitet.

Die Effektivität hybrider Zuordnungsverfahren ist beachtlich, da in der Regel viele Programmdurchläufe gespart werden und zudem Kandidatenpaare für mögliche Zuordnungen, die keinem oder nur einem Kriterium genügen, frühzeitig von den Berechungen ausgeschlossen werden können. Als nachteilig bei der hybriden Zuordnung kann sich erweisen, dass das resultierende Progamm nicht modular aufgebaut und damit sehr unflexibel gegenüber gewünschten Modifikationen ist.

6.2.2 Zusammengesetzte Zuordnungsverfahren

Die Grundidee bei der zusammengesetzten Zuordnung besteht darin, Ergebnisse verschiedener Simulationen, die bereits eine Vorauswahl möglicher Zuordnungen darstellen, mittels eines geeigneten Algorithmus zu kombinieren. Man kann dieses Vorgehen als Anwendung einer Funktion f verstehen, deren Eingabe Ergebnismengen verschiedener Zuordnungsalgorithmen und deren Ausgabe die endgültigen Zuordnungen sind:

$$\mathcal{M}_{comp,f} := f(\mathcal{M}_{conv}, \mathcal{M}_{chi2}, \mathcal{M}_{corr}, \mathcal{M}_{coll}), \tag{6.11}$$

wobei $\mathcal{M}_{conv}, \mathcal{M}_{chi2}, \mathcal{M}_{corr}$ und $\mathcal{M}_{coll}$ die durch die verschiedenen Ansätze erhaltenen Zuordnungsmengen bezeichnen. Erste Beispiele für die oben angesprochene Funktion entstehen durch eine Komposition von Mengenoperationen wie z.B. $\cap$ (Durchschnitt), $\cup$ (Vereinigung) und $\neg$ (relatives Komplement) auf den Zuordnungsmengen:

$$\begin{aligned}
f_1(\mathcal{M}_{conv}, \mathcal{M}_{chi2}, \mathcal{M}_{corr}, \mathcal{M}_{coll}) &:= \mathcal{M}_{conv} \cap \mathcal{M}_{chi2} \cap \mathcal{M}_{corr} \cap \mathcal{M}_{coll}, \\
f_2(\mathcal{M}_{conv}, \mathcal{M}_{chi2}, \mathcal{M}_{corr}, \mathcal{M}_{coll}) &:= (\mathcal{M}_{conv} \cap \mathcal{M}_{chi2}) \cup \mathcal{M}_{corr} \text{ oder} \\
f_3(\mathcal{M}_{conv}, \mathcal{M}_{chi2}, \mathcal{M}_{corr}, \mathcal{M}_{coll}) &:= \mathcal{M}_{conv} \cap \neg\mathcal{M}_{chi2}.
\end{aligned}$$

Zum Beispiel werden durch die Funktion f_1 nur solche Paare ausgewählt, die durch alle vier Maße vorgeschlagen wurden. Dies würde bedeuten, dass der Datenangreifer nur eine kleine Teilmenge der getroffenen Zuordnungen akzeptiert, diese aber mit sehr großer Wahrscheinlichkeit korrekt sind. Diese Vorgehensweise ist durchaus realistisch einzustufen. Ein potentieller Datenangreifer könnte versuchen, „einzelne interessant erscheinende Unternehmen im Mikrodatenfile zu reidentifizieren und für diese Information interessierte Konkurrenten zu suchen" (Wirth 2003, S. 20). Bei der Funktion f_2 werden auch Mehrfachzuordnungen erlaubt und es muss in diesen Fällen noch eine abschließende individuelle Entscheidung des Datenangreifers für eine eindeutige Zuordnung erfolgen. Mittels f_3 werden Paare ausgesucht, die durch das eine Maß präferiert und durch das andere abgelehnt würden.

Beispiel: Es seien $a_1, \ldots, a_8$ die Merkmalsträger einer externen Datei und $b_1, \ldots, b_8$ die Merkmalsträger der Zieldaten, wobei die korrekten Zuordnungen durch (a_i, b_i) für $i = 1, \ldots, 8$ gegeben sind. Durch Anwendung der in 6.1.2 bis 6.1.5 vorgestellten Maße seien folgende Lösungen entstanden:

$$\begin{aligned}
\mathcal{M}_{conv} &= \{(a_1,b_1),(a_2,b_6),(a_3,b_3),(a_4,b_5),(a_5,b_4),(a_6,b_8),(a_7,b_7),(a_8,b_2)\} \\
\mathcal{M}_{chi2} &= \{(a_1,b_1),(a_2,b_7),(a_3,b_3),(a_4,b_5),(a_5,b_4),(a_6,b_8),(a_7,b_6),(a_8,b_2)\} \\
\mathcal{M}_{corr} &= \{(a_1,b_1),(a_2,b_6),(a_3,b_5),(a_4,b_2),(a_5,b_4),(a_6,b_3),(a_7,b_7),(a_8,b_8)\} \\
\mathcal{M}_{coll} &= \{(a_1,b_1),(a_2,b_3),(a_3,b_5),(a_4,b_6),(a_5,b_4),(a_6,b_8),(a_7,b_7),(a_8,b_2)\}
\end{aligned}$$

Damit ergibt sich

$$
\begin{aligned}
f_1(\mathcal{M}_{conv}, \mathcal{M}_{chi2}, \mathcal{M}_{corr}, \mathcal{M}_{coll}) &= \{(a_1, b_1), (a_5, b_4)\}, \\
f_2(\mathcal{M}_{conv}, \mathcal{M}_{chi2}, \mathcal{M}_{corr}, \mathcal{M}_{coll}) &= \{(a_1, b_1), (a_3, b_3), (a_4, b_5), (a_5, b_4), (a_6, b_8), (a_8, b_2), \\
&\quad (a_2, b_6), (a_3, b_5), (a_4, b_2), (a_6, b_3), (a_7, b_7), (a_8, b_8)\}, \\
f_3(\mathcal{M}_{conv}, \mathcal{M}_{chi2}, \mathcal{M}_{corr}, \mathcal{M}_{coll}) &= \{(a_2, b_6), (a_7, b_7)\}.
\end{aligned}
$$

Offenbar ist diese Art, verschiedene Ansätze zu kombinieren, weitaus flexibler als die zuvor in Unterabschnitt 6.2.1 beschriebene, da die Kombination (über eine Funktion f) auf fertige Module für den Erhalt der Eingabemengen $\mathcal{M}_{conv}, \mathcal{M}_{chi2}, \mathcal{M}_{corr}, \mathcal{M}_{coll}$ zugreift. Diese Module müssen im Gegensatz zur hybriden Zuordnung nicht mehr verändert oder individuell angepasst werden.

6.2.3 Zweistufiges kombiniertes Zuordnungsverfahren

Die empirischen Ergebnisse der zuvor beschriebenen kombinierten Zuordnungsverfahren haben zu kontroversen Diskussionen im Kontext des Projektes FAWE-Panel geführt. Dies betrifft vor allem die zusammengesetzten Verfahren und die damit verbundenen verhältnismäßig hohen Reidentifikationsrisiken. Der Kompromissvorschlag einer zweistufigen Vorgehensweise wurde schließlich von allen am Projekt beteiligten wissenschaftlichen Bearbeitern und Beratern angenommen und wird seither erfolgreich auf verschiedene Paneldatenangebote der amtlichen Statistik angewendet.

Zunächst wird via zusammengesetzte Zuordnungsverfahren ein Teil der gesuchten Unternehmen den Zieldaten zugeordnet (*Phase 1*). Die so entstandenen richtig und falsch gebildeten Paare von Merkmalsträgern werden nun auf beiden Seiten (Zusatzwissen und Zieldatei) herausgenommen. Die restlichen gesuchten und zu diesem Zeitpunkt noch nicht zugeordneten Unternehmen werden nun via hybride Zuordnungsverfahren innerhalb der verbleibenden Zieldaten gesucht (*Phase 2*). Mit dieser Vorgehensweise wird erreicht, dass jedem gesuchten Unternehmen eindeutig ein Unternehmen der Zieldatei zugeordnet wird. Damit gehen im Gegensatz zur zusammengesetzten Zuordnung alle gesuchten Unternehmen direkt[40] in die Gesamtbewertung der Datensicherheit der anonymisierten Zieldatei ein.

6.3 Sensitivität bei der Parametersetzung

Hinsichtlich der Durchführbarkeit einer Sensitivitätsanalyse gelten a fortiori die in Abschnitt 3.6 für Querschnittsdaten formulierten Probleme. Da im Falle von Paneldaten

40 Bei der zusammengesetzten Zuordnung wird zwar nur eine Teilmenge aller gesuchten Unternehmen den Zieldaten zugeordnet, die nicht zugeordneten Unternehmen sorgen aber dennoch für Falschzuordnungen und gehen daher *indirekt* in die Gesamtbewertung der Datensicherheit der Zieldatei ein.

weit mehr Überschneidungsmerkmale auftreten, sind entsprechend mehr Parameter zu berücksichtigen. Darüber hinaus kommen die zu den verschiedenen Ansätzen gehörenden Gewichtungsparameter hinzu.

Wenig denkbar, aber nicht ausgeschlossen, ist allerdings folgendes Szenario: Ein Datenangreifer verfolgt die Strategie der zusammengesetzten Zuordnung, um eine möglicherweise kleine Teilmenge der gesuchten Unternehmen mit großer Wahrscheinlichkeit korrekt zuzuordnen. Danach führt er auf Basis dieser zuverlässigen Zuordnungen Abweichungsanalysen zwischen den beiden auf die Teilmenge reduzierten Datenquellen (vergleichbar mit den Analysen in Abschnitt 7.2) durch. Die Gewichte für die einzelnen Überschneidungsmerkmale können nun dem Grad an Übereinstimmung beider Datenquellen angepasst werden. Ist die angesprochene Teilmenge verhältnismäßig klein, wovon in der Realität auszugehen ist, so können diese Gewichte zwar nicht für einzelne Blöcke in beliebig tiefer Gliederung, zumindest aber global gesetzt werden.

6.4 Komplexitätsbetrachtung

Es sei wie in Abschnitt 3.7 mit $n = max(n_a, n_b)$ die maximale Anzahl an Merkmalsträgern einer Datei (externe Daten und Zieldaten) und mit k die Anzahl der metrischen Überschneidungsmerkmale bezeichnet. Da zum einen die in Abschnitt 6.1 zur Distanzberechnung diskutierten Maße mit quadratischem Aufwand berechnet werden können und zum anderen $k << n$ gilt – auch bei Paneldaten sind weit weniger Überschneidungsmerkmale als Merkmalsträger zu erwarten –, ist die Komplexität des Gesamtverfahrens unter Verwendung einer Näherungsheuristik mit derselben Argumentation wie in Abschnitt 3.7 von der Ordnung $O(n^2 \log n)$.

Allerdings muss berücksichtigt werden, dass der Aufwand der Simulationen mit Paneldaten gegenüber dem Aufwand mit Querschnittsdaten um ein Vielfaches angestiegen ist, was aber die theoretische Komplexitätsbetrachtung unberührt lässt. Der Mehraufwand liegt in den hybriden und zusammengesetzten Kombinationsmöglichkeiten begründet. Man beachte, dass bei der zusammengesetzten Zuordnungsstrategie mehrere Programmdurchläufe notwendig sind. Bei der hybriden Zuordnung müssen mehrere Maße zur Koeffizientenberechnung kombiniert werden. In der Praxis hat sich dennoch gezeigt, dass die Algorithmen in angemessener Zeit durchlaufen und daher in vollem Maße praxistauglich sind.

Aus der Perspektive des reinen Mathematikers ist jeder polynomiale Algorithmus – d.h., die Anzahl der durchzuführenden elementaren Operationen ist durch ein Polynom endlichen Grades in der Länge n des Eingabevektors abschätzbar – uneingeschränkt empfehlenswert. Aus der Perspektive des angewandten Mathematikers kann ein zusätzlicher Faktor n bzw. die Erhöhung des Polynomgrades wesentlich sein und aus wenigen Minuten je Durchlauf mehrere Tage machen. Dabei nutzt es in diesem Moment wenig, dass sich die Rechenzeit bei polynomialen Algorithmen im Zuge des Fortschritts der Rechentechnik schon mittelfristig auf ein Mindestmaß reduzieren wird.

Kapitel 7

Beispielsimulationen mit wirtschaftsstatistischen Paneldaten

In diesem Kapitel werden Beispielsimulationen mit verschiedenen Anonymisierungsvarianten der Kostenstrukturerhebung im Verarbeitenden Gewerbe, beobachtet über einen Zeitraum von vier Jahren, als Zieldaten vorgestellt. Für die Durchführung realitätsnaher Simulationen mit wirtschaftsstatistischen Paneldaten ist es notwendig, eine externe Datenbank von derselben Struktur aufzubauen (Abschnitt 7.1). Hierzu wurde wie in Kapitel 4 auf die MARKUS-Datenbank zurück gegriffen, nun mit Unternehmensinformationen über mehrere Jahre. Da die späteren Simulationsergebnisse wesentlich von der Qualität dieser Datenbank und den Abweichungen zu den Zieldaten abhängen, erscheinen zunächst einige Auswertungen und Abweichungsanalysen zwischen beiden Datenmaterialien notwendig (Abschnitt 7.2).

Bereits anhand der Beispielsimulationen wird deutlich, dass die theoretischen Überlegungen aus dem vorherigen Kapitel schwierig umsetzbar sind, was im Wesentlichen daran liegt, dass die Informationen im Zeitverlauf nicht in vergleichbarer Qualität in dem verfügbaren Zusatzwissen des Datenangreifers vorliegen. Disputabel ist auch der Gedanke, dass die Abschätzung des Nutzens einer Reidentifikation von der Aktualität des gefundenen Einzelwertes abhängig gemacht werden könnte. So würden die Abweichungsschwellen für frühere Wellen höher als am aktuellen Rand, beispielsweise in äquidistanten Abstufungen, angesetzt. Aus diesen Gründen kann die Frage nach der faktischen Anonymisierbarkeit wirtschaftsstatistischer Paneldaten grundsätzlich mit „Ja" beantwortet werden. Die in den abschließenden Abschnitten 7.4 und 7.5 durchgeführten Simulationen zeigen, dass hierzu Varianten der multiplikativen Zufallsüberlagerung besser als Varianten der Mikroaggregation geeignet sind.

7.1 Verwendetes Datenmaterial

Als Zieldaten werden die Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe (KSE) der Jahre 1999 bis 2002 verwendet. Diese Daten enthalten etwa 13 200 Merkmalsträger, über die in allen vier Wellen Informationen vorliegen. Durch einen aufwändigen Namens- und Adressabgleich (hierzu wurden moderne Methoden der Mustererkennung verwendet) konnten etwa 10 000 Unternehmen der externen Daten in den Zieldaten der KSE gefunden werden. Aus diesen wurden nur diejenigen ausgewählt, die im gesamten Zeitraum von 1999 bis 2002 an der KSE-Erhebung teilgenommmen haben. Ferner wurden Unternehmen entfernt, die im Jahre 1999 weniger als 20 Beschäftigte (untere Abschneidegrenze der KSE) oder in einem der Jahre 1999 bis 2002 weniger als 10 000 Euro Umsatz aufwiesen. Somit verbleiben noch 8 941 Unternehmen in der externen Datei. Auf diese beziehen sich die in diesem Abschnitt vorgestellten Analysen. Für vergleichende Betrachtungen werden aus den KSE-Daten ebenfalls nur diese Unternehmen herangezogen.

Als Überschneidungsmerkmale stehen dem Datenangreifer zur Verfügung:

- Gesamtumsatz 1999-2002,
- Anzahl der Beschäftigten 1999-2002,
- Wirtschaftsabteilungen (im Wesentlichen Zweisteller),
- Regionalinformation (Ost/West).

Dabei wird das Zusatzwissen als „lückenlos" verstanden, wenn sich die Umsatzangaben der enthaltenen Unternehmen in je zwei aufeinander folgenden Wellen ändern und gleichzeitig über alle Wellen hinweg wenigstens zwei verschiedene Beschäftigtenangaben vorliegen, als „lückenhaft" anderenfalls.

Bei den kategorialen Merkmalen wurden die Vergröberungen gewählt, die bereits bei der Anonymisierung der Querschnittsdaten der Kostenstrukturerhebung (siehe Abschnitt 4.1) verwendet wurden.

Bei den folgenden Untersuchungen werden nur solche Unternehmen berücksichtigt, die durchgängig über vier Jahre boebachtet wurden. Untenstehende Tabelle 7.1 enthält die Verteilung der ca. 13 300 Unternehmen der Zieldaten auf Beschäftigtengrößenklassen:

Tabelle 7.1: Beschäftigtengrößenklassen (BGK) der Zieldaten für die Wellen 1999 bis 2002

BGK	Abs. Häufigkeit	Rel. Häufigkeit	Kum. abs. Häufigkeit	Kum. rel. Häufigkeit
20-49	3898	29,32	3898	29,32
50-99	3364	25,30	7262	54,63
100-249	3130	23,54	10392	78,17
250-499	1394	10,49	11786	88,66
500-999	866	6,51	12652	95,17
$\geq$ 1 000	642	4,83	13294	100,00

Als externe Datenquelle wird wie vorab erwähnt auf die MARKUS-Datenbank zurück gegriffen. In Kooperation mit dem Zentrum für Europäische Wirtschaftsforschung (ZEW) konnte via aufwendige Adressabgleiche ein verwendbarer Auszug von ca. 9 000 Unternehmen erzeugt werden.[41] Nachfolgende Tabelle 7.2 enthält die Verteilung der Unternehmen der lückenhaften externen Daten auf Beschäftigtengrößenklassen:

Tabelle 7.2: Beschäftigtengrößenklassen (BGK) der externen Daten für die Wellen 1999 bis 2002, lückenhaft

BGK	Abs. Häufigkeit	Rel. Häufigkeit	Kum. abs. Häufigkeit	Kum. rel. Häufigkeit
20-49	2542	28,43	2542	28,43
50-99	2250	25,16	4792	53,60
100-249	2181	24,39	6973	77,99
250-499	993	11,11	7966	89,10
500-999	555	6,21	8521	95,30
$\geq$ 1 000	420	4,70	8941	100,00

In Tabelle 7.3 ist die entsprechende Auswertung für das lückenlose Zusatzwissen, das etwa 3 500 Einheiten enthält, dargestellt.

41 An dieser Stelle gebührt Herrn Thorsten Doherr ein herzlicher Dank für die gute Zusammenarbeit in mehreren Projekten.

Tabelle 7.3: Beschäftigtengrößenklassen (BGK) der externen Daten für die Wellen 1999 bis 2002, lückenlos

BGK	Abs. Häufigkeit	Rel. Häufigkeit	Kum. abs. Häufigkeit	Kum. rel. Häufigkeit
20-49	806	23,09	806	23,09
50-99	740	21,20	1546	44,30
100-249	916	26,28	2463	70,57
250-499	485	13,90	2948	84,47
500-999	291	8,34	3239	92,81
$\geq$ 1 000	251	7,19	3490	100,00

7.2 Zur Datenqualität des Zusatzwissens

In den folgenden Auswertungen wird das lückenhafte Zusatzwissen, also die oben angesprochenen 8 941 Unternehmen der MARKUS-Datenbank, herangezogen. Bei der bisherigen Arbeit mit diesen Daten hat sich gezeigt, dass sie weit schlechtere Qualität als die Zieldaten der KSE besitzen. So weisen 238 Unternehmen im Merkmal Umsatz in den vier betrachteten Jahren identische Angaben auf; bei 1 663 Unternehmen gibt es nur einmal, bei 3 177 Unternehmen zweimal eine Umsatzveränderung. Insgesamt beobachtet man bei 5 078 der 8 941 MARKUS-Unternehmen, also bei etwa 57%, in wenigstens zwei aufeinanderfolgenden Jahren identische Umsätze. Dieses unwahrscheinliche Phänomen tritt bei den Zieldaten der KSE bei keinem Merkmalsträger auf.

Die Qualität der Beschäftigtenangaben in der externen Datei ist ähnlich schlecht: Es weisen 2 516 Unternehmen im Merkmal Beschäftigte in den vier betrachteten Jahren identische Angaben auf; bei 2 230 Unternehmen gibt es nur einmal, bei 2 373 Unternehmen zweimal eine Beschäftigtenveränderung. Insgesamt beobachtet man bei 8 119 der 8 941 MARKUS-Unternehmen, also bei etwa 91%, in wenigstens zwei aufeinanderfolgenden Jahren dieselbe Beschäftigtenzahl. Auf der anderen Seite weisen nur 65 Unternehmen der KSE in den vier betrachteten Jahren identische Beschäftigtenangaben auf; bei 316 Unternehmen gibt es nur einmal, bei 1 842 Unternehmen zweimal eine Beschäftigtenveränderung. Insgesamt beobachtet man bei nur 2 223 der 13 227 KSE-Unternehmen, also bei etwa 17%, in wenigstens zwei aufeinanderfolgenden Jahren identische Beschäftigtenangaben.

7.2.1 Konstante Werte im Zeitverlauf

Auf den ersten Blick fällt auf, dass sich für viele Unternehmen die Anzahl der Beschäftigten und/oder der Gesamtumsatz in den MARKUS-Daten nicht ändert. Dies deutet darauf

hin, dass die Daten nicht jedes Jahr aktualisiert werden. Für 2 516 Unternehmen bleibt die Mitarbeiterzahl über die betrachteten vier Jahre hinweg konstant. Der Gesamtumsatz ändert sich für 238 Unternehmen in diesem Zeitraum nicht. Für 177 Unternehmen ändern sich im ganzen Zeitraum weder die Anzahl der Beschäftigten noch der Gesamtumsatz. Betrachtet man Fälle, in denen die Werte in zwei aufeinanderfolgenden Jahren konstant sind, so trifft dies bei den Mitarbeiterzahlen für etwa 91% (8 119) der Unternehmen zu und auch beim Gesamtumsatz ist über die Hälfte (5 078) betroffen. Sowohl die Anzahl der Mitarbeiter als auch der Umsatz sind für 4 741 Unternehmen (53%) mindestens einmal in zwei aufeinanderfolgenden Jahren konstant.

Die entsprechende Anzahl an Unternehmen mit über die Zeit konstanten Einzelwerten ist in der KSE deutlich kleiner: Gleichbleibende Mitarbeiterzahlen über vier Jahre werden nur für 42 Unternehmen, gleichbleibende Umsatzangaben überhaupt nicht beobachtet. Unveränderte Mitarbeiterzahlen in zwei aufeinanderfolgenden Jahren treten hier für 2 127 Unternehmen auf, während vernachlässigbare 16 Unternehmen gleichbleibende Umsatzzahlen aufweisen. Dass beide Merkmale (*Anzahl der Beschäftigten* und *Gesamtumsatz* gleichzeitig) in zwei aufeinanderfolgenden Jahren unverändert bleiben, wird in der KSE nicht beobachtet.

7.2.2 Deskriptive Maße im Vergleich

Die durchschnittliche Absolutanzahl der Mitarbeiter liegt bei den MARKUS-Daten für alle Jahre mit 299 – 305 um ca. 10 – 20 höher als in der KSE (283 – 290). Die maximalen Mitarbeiterzahlen sind in den MARKUS-Daten mit bis zu 194 000 deutlich höher als in der KSE mit bis zu 138 000.

Die durchschnittlichen Gesamtumsätze sind 1999 und 2000 bei den MARKUS-Daten niedriger (96,9 und 95,2% der KSE), 2001 und 2002 hingegen höher (104,7 und 106,3% der KSE).

Die im Vergleich zur KSE höheren durchschnittlichen Gesamtumsätze für 2001 und 2002 dürften hauptsächlich damit zusammenhängen, dass die Währung in den MARKUS-Daten nicht immer korrekt angegeben wird, obwohl nach der Datensatzbeschreibung alle Umsatzangaben in DM erhoben werden. Rechnet man die Werte testweise in Euro um, so fällt auf, dass für knapp 300 Unternehmen der Gesamtumsatz für 2001 zwischen dem 1,8-fachen und dem 2,2-fachen des Gesamtumsatzes für 2000 liegt. Dies legt die Vermutung nahe, dass diese Angaben bereits in Euro vorlagen. Sicher nachweisen lässt sich dies allerdings nicht.

Die maximal auftauchenden Gesamtumsätze sind in beiden Datenquellen in allen Jahren nahezu identisch.

7.2.3 Sprünge in aufeinanderfolgenden Jahren

Die Änderungsraten zwischen aufeinanderfolgenden Jahren sind in den MARKUS-Daten wesentlich höher als in den KSE-Daten. Bei den Mitarbeiterzahlen liegen die Änderungsraten zwischen 0,06 und 8 000%. In der KSE bewegen sich diese zwischen 3,2 und 1 079%.

Es folgen zwei aus Geheimhaltungsgründen geringfügig modifizierte Beispiele von Unternehmen, die in den MARKUS-Daten einen sehr großen Sprung aufweisen, während die zugehörigen Veränderungen in den KSE-Daten deutlich schwächer sind (siehe Tabellen 7.4 und 7.5).

Tabelle 7.4: Sprünge in den Beschäftigtenangaben I

Jahr\Datenquelle	MARKUS	KSE
1999	64	348
2000	64	371
2001	1 800	394
2002	1 800	413

Tabelle 7.5: Sprünge in den Beschäftigtenangaben II

Jahr\Datenquelle	MARKUS	KSE
1999	3 400	2 379
2000	2	2 876
2001	2	3 060
2002	2	2 848

Noch häufiger sind solche Sprünge bei den Umsatzangaben in der MARKUS-Datenbank zu beobachten. Die Änderungen bewegen sich bei den MARKUS-Daten zwischen 0,08 und über 291 000%. In der KSE liegen die entsprechenden Änderungsraten zwischen 14,2 und fast 1 000%. Auch hierzu folgen zwei in den Tabellen 7.6 und 7.7 dargestellte Beispiele.

Tabelle 7.6: Sprünge in den Umsatzangaben I

Jahr\Datenquelle	MARKUS	KSE
1999	12 782 297	70 114 990
2000	26 499 997	75 281 594
2001	78 593	69 740 000
2002	229 064 029	65 904 000

Tabelle 7.7: Sprünge in den Umsatzangaben II

Jahr	Datenquelle	
	MARKUS	KSE
1999	25 564 594	27 927 662
2000	25 564 594	33 841 523
2001	33 000 000	37 556 318
2002	106 085	40 771 550

7.2.4 Vergleich der Verlaufsmuster von MARKUS und KSE

Wir betrachten nun den Verlauf der Mitarbeiter- bzw. Umsatzentwicklung von 1999 bis 2002 und unterscheiden zwei Verlaufsmuster:

1) Nimmt zu oder bleibt gleich.
2) Geht zurück.

Aufgrund der vielen über die Zeit unveränderten Merkmalswerte in den MARKUS-Daten erscheint es wenig sinnvoll, „bleibt unverändert" als eigene Kategorie zu definieren.

Zunächst wird die Veränderung von 1999 auf 2002 (Vierjahrestrend) untersucht. In Bezug auf die Mitarbeiterzahl zeigen gut 55% der in beiden Datenquellen vorhandenen Unternehmen (4 974) einen unterschiedlichen Trend. Etwas günstiger fällt dieser Vergleich bei den Umsatzangaben aus. Hier zeigen etwa 72% der Unternehmen (6 451) in beiden Quellen den gleichen Trend. D.h., für diese Unternehmen ist der Gesamtumsatz in beiden Datenquellen in 2002 entweder höher oder gleich hoch wie in 1999 oder er fällt in beiden Datenquellen niedriger aus. Andererseits ist er für gut 28% der Unternehmen in einer Quelle in 2002 höher oder gleich hoch wie in 1999, während er in der anderen in diesem Zeitraum abgenommen hat.

Vergleicht man alle Veränderungen zwischen den einzelnen Jahren (1999 auf 2000, 2000 auf 2001, 2001 auf 2002), so ergibt sich folgendes Bild: Beim Gesamtumsatz gibt es in 63% der Fälle in beiden Quellen den gleichen Verlauf, bei der Mitarbeiterzahl trifft dies für 60% zu. Besonders beim Übergang von 2001 auf 2002 ist die Übereinstimmung mit 55% recht gering. Diese Auswertungen lassen den Schluss zu, dass das im Projekt FAWE-Panel diskutierte Kriterium des Trenderhalts bei Anonymisierungsverfahren keine Erhöhung der Reidentifikationsrisiken mit sich führen muss.

7.2.5 Abweichungen zwischen Wirtschaftszweig- und Regionalangaben

Für etwa ein Viertel der Unternehmen (je nach Jahr zwischen 24,4 und 25,2%) unterscheiden sich die Wirtschaftszweigangaben auf der Ebene der Zweisteller in den beiden Datenquellen.

Die Wirtschaftszweigzugehörigkeit hingegen bleibt im Zeitverlauf in beiden Quellen für die Mehrzahl der Unternehmen unverändert. So beobachtet man in den MARKUS-Daten für 231 Unternehmen (2,6%) und in der KSE für nur 179 Unternehmen (2,0%) wenigstens einen Wechsel auf der Zweistellerebene.

Die Regionalangaben passen in beiden Quellen sehr gut zueinander. Bei der Unterteilung in neue/alte Bundesländer treten vernachlässigbare Unterschiede auf. Bezogen auf den nichtadministrativen siedlungsstrukturellen Regionstyp werden zwischen den beiden Datenquellen in nur 188 Fällen (ca. 2,1%) geringfügige Abweichungen beobachtet.

7.3 Natürliche Schutzwirkung

Aufgrund der mitunter großen Abweichungen zwischen externen Daten und Zieldaten kann für einen Teil der gesuchten Unternehmen bereits allein durch die Entfernung direkter Identifikatoren wie Name und Adresse ein ausreichender Schutz erreicht werden. Dieser ohne die Entwicklung weiterer Anonymisierungsstrategien beobachtbare Effekt wurde in Kapitel 4 „natürliche Schutzwirkung" genannt. Etwas weiter gefasst kann auch die Wirkung informationsreduzierender Methoden bei gleichzeitigem Verzicht auf Datenveränderung als natürliche Schutzwirkung verstanden werden. Hierzu wurden in Kapitel 5 bereits umfangreiche empirische Analysen zur Anwendung informationsreduzierender Methoden auf verschiedene Kombinationen der kategorialen Überschneidungsmerkmale *Rechtsform, Wirtschaftszweigklassifikation* und *Regionalkennung* durchgeführt.

In diesem Abschnitt werden die Wirtschaftszweige in den Zieldaten, und damit notwendigerweise auch in den externen Daten, auf die Zweistellerebene vergröbert. Als Regionalangabe wird der siedlungsstrukturelle Regionstyp, dessen Wertebereich aus drei Kategorien entsprechend dem Grad der Urbanisierung einer Region besteht, verwendet. Aus Sicht des Datenangreifers ist es erfreulich, dass vernachlässigbare Veränderungen der Beobachtungen in diesen Merkmalen über die Jahre festzustellen sind.

Es werden nun die verschiedenen Datenangriffsstrategien aus dem vorherigen Kapitel angewendet. Gemäß den Ausführungen in den Abschnitten 6.1.2 bis 6.1.5 werden die resultierenden Zuordnungsmengen mit $\mathcal{M}_{conv}$, $\mathcal{M}_{corr}$, $\mathcal{M}_{chi2}$ und $\mathcal{M}_{coll}$ bezeichnet. Dabei werden bei jeder Strategie sämtliche gesuchte Unternehmen einem Zielunternehmen zugeordnet. Die folgende Tabelle 7.8 enthält die (globalen) Reidentifikationsrisiken

tabelliert nach Beschäftigtengrößenklassen der gesuchten Unternehmen und nach Ansatz für die Koeffizienten der Zielfunktion.

Tabelle 7.8: Reidentifikationsrisiken nach Beschäftigtengrößenklasse und Strategie für die Zuordnung

class\Strategie	$\mathcal{M}_{conv}$	$\mathcal{M}_{chi2}$	$\mathcal{M}_{corr}$	$\mathcal{M}_{coll}$
$20-49$	0.25	0.08	0.08	0.06
$50-99$	0.26	0.13	0.11	0.07
$100-249$	0.34	0.11	0.10	0.06
$250-499$	0.61	0.23	0.18	0.15
$500-999$	0.72	0.63	0.52	0.28
≥ 1000	0.85	0.45	0.37	0.30

Etwas überraschend ist, dass bei einigen Strategien das Reidentifikationsrisiko beim Übergang von der zweitobersten zur obersten Beschäftigtengrößenklasse abnimmt.

In Tabelle 7.9 werden globale Reidentifikationsrisiken unter Verwendung der symmetrischen binären Mengenoperation $f(A,B) := A \cap B$ ausgewiesen (binäre zusammengesetzte Zuordnung).[42]

Tabelle 7.9: Reidentifikationsrisiken bei der binären zusammengesetzten Zuordnung

$\cap$	$\mathcal{M}_{conv}$	$\mathcal{M}_{chi2}$	$\mathcal{M}_{corr}$	$\mathcal{M}_{coll}$
$\mathcal{M}_{conv}$	0.30	0.72	0.57	0.54
$\mathcal{M}_{chi2}$		0.13	0.29	0.27
$\mathcal{M}_{corr}$			0.11	0.25
$\mathcal{M}_{coll}$				0.07

Die Diagonale obiger Tabelle enthält die durch Verfolgung einer einzigen Stragie erzielten Ergebnisse. Man beachte, dass bei der zusammengesetzten Zuordnung nur eine kleine Teilmenge aller gesuchten Unternehmen einem Zielunternehmen (richtig oder falsch) zugeordnet wird. Beispielsweise enthält die Menge $\mathcal{M}_{conv} \cap \mathcal{M}_{chi2}$ nur 476 Elemente bzw. Unternehmen, von denen 344 dem zugehörigen Zielunternehmen zugeordnet wurden. Insgesamt werden je nach Kombination zwischen 421 ($= \mathcal{M}_{corr} \cap \mathcal{M}_{coll}$) und 476 ($= \mathcal{M}_{conv} \cap \mathcal{M}_{chi2}$) gesuchte Unternehmen eindeutig einem Zielunternehmen zugeordnet. Somit sind die in

42 Die Tabelle enthält nicht die Mengendurchschnitte, sondern die relativen Häufigkeiten der Durchschnitte gemessen an allen zugeordneten Unternehmen.

Tabelle 7.9 aufgeführten Risiken repräsentativ für die Gesamtheit aller Unternehmen der externen Datei. Man beachte hierzu auch die Ausführungen in Abschnitt 6.2.3.

Weiter steigende Reidentifikationsrisiken bei gleichzeitig sinkender Repräsentativität für die Gesamtdatei werden erwartungsgemäß nach der Anwendung der ternären Operation $f(A, B, C) := A \cap B \cap C$ (ternäre zusammengesetzte Zuordnung) beobachtet:

Tabelle 7.10: Reidentifikationsrisiken bei der ternären zusammengesetzten Zuordnung

	Reidentifikationsrisiko
$\mathcal{M}_{conv} \cap \mathcal{M}_{chi2} \cap \mathcal{M}_{corr}$	0.88
$\mathcal{M}_{conv} \cap \mathcal{M}_{chi2} \cap \mathcal{M}_{coll}$	0.82
$\mathcal{M}_{conv} \cap \mathcal{M}_{corr} \cap \mathcal{M}_{coll}$	0.77
$\mathcal{M}_{chi2} \cap \mathcal{M}_{corr} \cap \mathcal{M}_{coll}$	0.52

Insgesamt werden je nach betrachtetem Tripel zwischen 349 ($= \mathcal{M}_{chi2} \cap \mathcal{M}_{corr} \cap \mathcal{M}_{coll}$) und 380 ($= \mathcal{M}_{conv} \cap \mathcal{M}_{chi2} \cap \mathcal{M}_{corr}$) gesuchte Unternehmen eindeutig einem Zielunternehmen zugeordnet, weshalb die tabellierten relativen Häufigkeiten erneut nicht das mit der Gesamtdatei verbundene Reidentifikationsrisiko schätzen.

Offensichtlich steigen die Reidentifikationsrisiken unter Anwendung der quaternären Operation $f(A, B, C, D) := A \cap B \cap C \cap D$ weiter an. In diesem Falle werden 98 %, also fast alle Zuordnungen richtig getroffen. Allerdings werden dabei nur 311 (von knapp 9 000 gesuchten) Unternehmen einem Zielunternehmen zugeordnet, wovon 306 Treffer zu verzeichnen sind. Die Informationen der verbleibenden Unternehmen der externen Daten können mit dieser gemischten Strategie nicht aufgedeckt werden.

Während der Laufzeit des Projektes „Wirtschaftsstatistische Einzeldaten und faktische Anonymisierung" wurde der nachfolgend simulierten Strategie der hybriden Zuordnung größere Aufmerksamkeit geschenkt, da hier jedem Unternehmen der externen Daten ein Zielunternehmen zugeordnet wird. D.h., Tabelle 7.11 enthält globale Reidentifikationsrisiken, die in diesem Beispiel durch Gleichgewichtung zweier Parameter λ_i und Setzen der restlichen Parameter auf Null entstanden sind (zweifache hybride Zuordnung). Die für die Maße d_{conv}, d_{chi2}, d_{corr} und d_{coll} gesetzten Gewichte werden in dem Vektor $\Lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4)$ zusammengefasst.

Tabelle 7.11: Globale Reidentifikationsrisiken bei der zweifachen hybriden Zuordnung

Λ	Reidentifikationsrisiko
$(\frac{1}{2}, \frac{1}{2}, 0, 0)$	0.30
$(\frac{1}{2}, 0, \frac{1}{2}, 0)$	0.29
$(\frac{1}{2}, 0, 0, \frac{1}{2})$	0.23
$(0, \frac{1}{2}, \frac{1}{2}, 0)$	0.12
$(0, \frac{1}{2}, 0, \frac{1}{2})$	0.08
$(0, 0, \frac{1}{2}, \frac{1}{2})$	0.11

Abschließende Tabelle 7.12 enthält globale Reidentifikationsrisiken verbunden mit der hybriden Strategie und Gleichgewichtung jeweils dreier Maße und Nichtberücksichtigung des verbleibenden Maßes (dreifache hybride Zuordnung).

Tabelle 7.12: Globale Reidentifikationsrisiken bei der dreifachen hybriden Zuordnung

Λ	Reidentifikationsrisiko
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}, 0)$	0.31
$(\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3})$	0.25
$(\frac{1}{3}, \frac{1}{3}, 0, \frac{1}{3})$	0.23
$(0, \frac{1}{3}, \frac{1}{3}, \frac{1}{3})$	0.11

Im Vergleich zu Tabelle 7.11 wird keine Steigerung der Reidentifikationsrisiken beobachtet. Daher empfiehlt sich für einen potentiellen Datenangreifers, das einfachere und weniger rechenintensive zweiparametrige Modell dem dreiparametrigen Modell vorzuziehen. Auch durch eine Gleichgewichtung aller vier Maße wird hier mit 31-prozentiger Trefferquote kein nennenswerter Mehrwert erzielt.

Bei der Diskussion der Ergebnisse sollte beachtet werden, dass bei einigen Strategien nur ein Teil der Merkmalsträger zugeordnet wird, bei anderen hingegen für jeden gesuchten Merkmalsträger ein Kandidat gefunden wird. Beide gegenläufigen Herangehensweisen haben aus Sicht des Datenangreifers Vor- und Nachteile. Bei der Auswahl einer kleinen Teilmenge von Zuordnungen besteht für den Datenangreifer eine große Sicherheit, richtig zugeordnet zu haben; allerdings muss er sich dann mit einer kleinen Absolutanzahl korrekter Zuordnungen zufrieden geben. Umgekehrt geht eine hohe absolute Trefferanzahl einher mit einer großen Unsicherheit bei der Zuordnung.

Hauptgrund für das teilweise schlechte Abschneiden der vorgestellten Strategien bzw. der darin berechneten Zusammenhangsmaße ist nach Ansicht des Autors die schlechte Qualität des verfügbaren Zusatzwissens. Etwa 4 000 von 9 000 Unternehmen der externen Daten weisen jährliche Veränderungen im Merkmal *Gesamtumsatz*, etwa 900 Unternehmen im Merkmal *Anzahl der Beschäftigten*.

Da in den vorangegangenen Simulationsexperimenten die 8 941 Unternehmen der externen Daten eine Teilmenge der 13 294 Zielunternehmen darstellten, wurde stillschweigend von der Kenntnis über die Teilnahme der gesuchten Unternehmen an der Zielerhebung ausgegangen. In einer (dem Autor bislang noch nicht zu Ohren gekommenen) realen Datenangriffssituation verfügt ein Datenangreifer aber keineswegs über diese Information, da die Kostenstrukturerhebung im Verarbeitenden Gewerbe in den meisten Bereichen keine Vollerhebung darstellt. Das Reidentifikations- bzw. Enthüllungsrisiko wurde demnach systematisch überschätzt. Aus diesem Grunde wird in der Praxis das geschätzte Enthüllungsrisiko zellenweise mit dem Stichprobenauswahlsatz multipliziert. Beachtet man, dass die Kostenstrukturerhebung im Verarbeitenden Gewerbe knapp 40% aller Unternehmen mit weniger als 250 Beschäftigten erfasst, so werden die ohnehin geringen Risiken in den unteren Beschäftigtengrößenklassen vernachlässigbar klein. Unglücklicherweise kann ein solcher Effekt bei der erfolgversprechenden Suche nach größeren Unternehmen der Kostenstrukturerhebung nicht beobachtet werden, da von den Unternehmen mit 250 bis 499 Beschäftigten ca. 80%, die Unternehmen mit wenigstens 500 Beschäftigen sogar voll erhoben werden. Letzteres gilt auch für sehr dünn besetzte Wirtschaftszweige.

7.4 Varianten der Mikroaggregation

Wir untersuchen im Folgenden die Schutzwirkung vier spezieller Varianten der deterministischen Mikroaggregation auf die Daten der Kostenstrukturerhebung.

Aus Geheimhaltungsgründen werden in den Tabellen (*nicht* in den Simulationen) die Wirtschaftszweigangaben auf der Zweistellerebene (sogenannte Abteilungen) pseudonymisiert. Dies bedeutet, dass die ursprünglichen Zweisteller neu bezeichnet bzw. umcodiert werden, um das Gefährdungspotential einzelner Wirtschaftsabteilungen nicht offen zu legen.

Im Folgenden werden zunächst für die Variante 1 (einfache Mikroaggregation mit 3er Gruppenbildung) Enthüllungsrisiken detailliert nach Wirtschaftszweigen und Beschäftigtengrößenklassen ausgewiesen. Desweiteren werden zwei Strategien (je eine spezielle hybride und eine zusammengesetzte) verfolgt und bzgl. des in Abschnitt 7.1 beschriebenen „lückenhaften" und „lückenlosen" Zusatzwissens unterschieden.

Die mit den Varianten 2 bis 4 verbundenen Risiken werden danach nur noch in aggregierter Form dargestellt, um eine Übersichtlichkeit und einen direkten Vergleich der verschiedenen Varianten untereinander zu gewährleisten.

Variante 1: einfache Mikroaggregation mit 3er Gruppenbildung

Bei dieser Variante wird die Mikroaggregation für jedes Merkmal getrennt durchgeführt. Das jeweils zu anonymisierende Merkmal bzw. die zugehörige Spalte der Datenmatrix wird zunächst sortiert. Danach werden absteigend immer drei benachbarte Merkmalsträger in

einer Gruppe zusammengefasst und alle ihre stetigen Merkmalswerte durch den Durchschnitt der drei Einzelwerte ersetzt. Sollten nach der Gruppenbildung am unteren Ende der Spalte maximal zwei Werte übrig bleiben, so werde diese zum dem darüberliegenden Tripel gruppiert, weshalb im Einzelfall auch einer Gruppe mit vier oder fünf zu mittelnden Einzelwerten entstehen kann. In unserem Falle formiert sich wegen $13294 \, mod \, 3 = 1$ eine 4er Gruppe „am Ende des Feldes".

Wir beginnen mit der hybriden Strategie, bei der bekanntermaßen alle gesuchten externen Unternehmen einem Zielunternehmen zugeordnet werden. Die so gewonnene Schätzung der Risiken darf somit als „Globales Enthüllungsrisiko" bezeichnet werden. Wie in der Vergangenheit werden gefundene Einzelwerte als „nicht enthüllt" verstanden, wenn sie (nach der Anonymisierung) um wenigstens 10% von ihrem zugehörigen Originalwert abweichen. Die Gewichtung der Maße $d_{conv}, d_{chi2}, d_{corr}$ und d_{coll} wird bei lückenhaftem Zusatzwissen mit $\Lambda = (4, 1, 1, 1)$ und bei lückenlosem Zusatzwissen mit $\Lambda = (2, 1, 1, 1)$ gesetzt. Damit wird das konventionelle Distanzmaß höher gewichtet als die übrigen drei Maße. Diese Entscheidung basiert auf der schwachen Qualität vor allem des lückenhaften Zusatzwissens, in dem Merkmalsverläufe gar nicht oder nur sehr schlecht abgebildet werden (vgl. Abschnitt 7.2). Weitere Testläufe mit verschiedenen Parametern werden vermieden, da ein potentieller Datenangreifer diese Vergleichsmöglichkeiten nicht hätte und ein realitätsfernes „Lernen" der bestmöglichen Parameterkonstellation, was eine unvermeidbare Erhöhung der Risiken mit der Zeit mit sich bringen würde, vermieden werden soll.

Die Tabellen 7.13 bis 7.16 enthalten globale Enthüllungsrisiken bezogen auf Kombinationen von Beschäftigtengrößenklasse und Wirtschaftsabteilung, wobei in den Simulationen sowohl das lückenhafte als auch das lückenlose Zusatzwissen verwendet wurde. Bei der Regionalangabe wurde zwischen neuen (Ost) und alten (West) Bundesländern unterschieden.

Tabelle 7.13
Enthüllungsrisiken nach Beschäftigtengrößenklassen und Wirtschaftszweigen (WZ 93) – hybrid, lückenlos, Ost –

Beschäftigtengrößenklasse	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
20 – 49	0,50	0,36	0,25	0,27	0,86	0,26	0,13	0,63	0,71
50 – 99	0,50	0,25	0,71	0,36	0,38	0,24	0,24	0,25	0,50
100 – 249	0,71	0,60	0,20	0,60	0,80	0,41	0,32	1,00	0,86
250 – 499	–	0,75	1,00	0,57	0,67	1,00	0,29	0,67	–
500 – 999	–	1,00	0,99	0,99	1,00	0,00	1,00	–	0,00
≥ 1 000	–	–	–	1,00	0,99	–	0,99	1,00	1,00

Beschäftigtengrößenklasse	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
20 – 49	0,18	1,00	0,33	0,22	0,67	0,46	0,18	0,20
50 – 99	0,25	1,00	0,23	0,40	0,00	0,47	0,05	0,50
100 – 249	0,40	–	0,71	0,13	1,00	0,13	0,52	0,50
250 – 499	0,00	–	0,33	0,33	0,99	0,50	0,33	1,00
500 – 999	–	–	–	1,00	–	0,00	1,00	–
≥ 1 000	–	–	0,00	1,00	0,99	–	0,00	–

Tabelle 7.14
Enthüllungsrisiken nach Beschäftigtengrößenklassen und Wirtschaftszweigen (WZ 93) – hybrid, lückenhaft, Ost –

Beschäftigtengrößenklasse	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
20 – 49	0,27	0,13	0,28	0,19	0,43	0,23	0,21	0,47	0,48
50 – 99	0,60	0,19	0,61	0,28	0,41	0,17	0,16	0,57	0,36
100 – 249	0,70	0,70	0,31	0,56	0,77	0,31	0,32	1,00	0,75
250 – 499	1,00	0,67	0,67	0,43	0,60	0,60	0,21	0,50	–
500 – 999	–	1,00	0,50	0,67	1,00	0,50	1,00	0,99	0,00
≥ 1 000	–	0,00	–	1,00	0,99	0,00	0,99	1,00	1,00

Beschäftigtengrößenklasse	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
20 – 49	0,34	0,50	0,27	0,13	1,00	0,29	0,13	0,22
50 – 99	0,23	0,50	0,26	0,24	0,33	0,50	0,14	0,29
100 – 249	0,63	–	0,63	0,25	0,50	0,27	0,36	0,52
250 – 499	0,33	–	0,40	0,43	1,00	0,33	0,29	0,67
500 – 999	–	–	–	1,00	–	0,33	0,80	0,50
≥ 1 000	–	0,99	0,00	0,40	1,00	–	–	0,99

Tabelle 7.15
Enthüllungsrisiken nach Beschäftigtengrößenklassen und Wirtschaftszweigen (WZ 93)
– hybrid, lückenlos, West –

Beschäftigtengrößenklasse	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
20 – 49	0,19	0,14	0,19	0,07	0,21	0,03	0,06	0,38	0,07
50 – 99	0,14	0,27	0,19	0,25	0,20	0,07	0,14	0,40	0,04
100 – 249	0,17	0,30	0,30	0,35	0,40	0,18	0,20	0,30	0,48
250 – 499	0,50	0,29	0,12	0,36	0,45	0,20	0,25	0,42	0,20
500 – 999	0,50	0,35	0,37	0,67	0,53	0,39	0,40	0,40	0,40
≥ 1 000	0,99	0,65	0,56	0,70	0,69	0,33	0,53	0,75	0,50

Beschäftigtengrößenklasse	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
20 – 49	0,17	0,44	0,05	0,06	0,17	0,14	0,06	0,11
50 – 99	0,13	0,33	0,08	0,20	0,33	0,24	0,14	0,23
100 – 249	0,32	0,33	0,13	0,22	0,67	0,30	0,11	0,22
250 – 499	0,31	0,17	0,38	0,35	0,20	0,57	0,24	0,21
500 – 999	0,36	1,00	0,52	0,43	1,00	0,30	0,32	0,19
≥ 1 000	0,67	0,00	0,43	0,64	–	0,50	0,33	0,72

Tabelle 7.16
Enthüllungsrisiken nach Beschäftigtengrößenklassen und Wirtschaftszweigen (WZ 93)
– hybrid, lückenhaft, West –

Beschäftigtengrößenklasse	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
20 – 49	0,14	0,08	0,08	0,11	0,19	0,04	0,06	0,38	0,04
50 – 99	0,10	0,15	0,17	0,13	0,19	0,07	0,09	0,19	0,11
100 – 249	0,25	0,21	0,20	0,30	0,30	0,13	0,17	0,36	0,25
250 – 499	0,40	0,28	0,18	0,27	0,48	0,22	0,19	0,28	0,21
500 – 999	0,40	0,33	0,34	0,54	0,45	0,28	0,35	0,43	0,14
≥ 1 000	0,99	0,57	0,28	0,65	0,59	0,32	0,45	0,90	0,50

Beschäftigtengrößenklasse	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
20 – 49	0,10	0,22	0,05	0,07	0,11	0,07	0,02	0,10
50 – 99	0,07	0,26	0,05	0,09	0,15	0,17	0,06	0,17
100 – 249	0,23	0,32	0,09	0,19	0,45	0,24	0,06	0,18
250 – 499	0,30	0,56	0,39	0,21	0,22	0,35	0,14	0,26
500 – 999	0,38	0,57	0,47	0,38	0,99	0,35	0,15	0,29
≥ 1 000	0,31	0,00	0,39	0,52	0,33	0,78	0,22	0,68

Bei der zusammengesetzten Zuordnung werden im Folgenden nur die Maße d_{conv} und d_{chi2} verwendet, da auf diese Weise eine größere Teilmenge der externen Daten in die Schätzung der Risiken einbezogen wird. In den Tabellen 7.17 bis 7.19 werden die mit beiden Strategien (hybride und zusammengesetzte Zuordnung) verbundenen Enthüllungsrisiken, verteilt auf Beschäftigtengrößenklassen, gegenüber gestellt.

Tabelle 7.17: Enthüllungsrisiken nach BGK und Strategie für die Zuordnung (lückenhaft, Ost)

Strategie\BGK	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1000
hybrid	0.25	0.28	0.45	0.43	0.72	0.65
zusammengesetzt	0.67	0.68	0.89	0.84	0.87	0.73

Tabelle 7.18: Enthüllungsrisiken nach BGK und Strategie (lückenhaft, West)

Strategie\BGK	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1000
hybrid	0.08	0.11	0.19	0.26	0.36	0.47
zusammengesetzt	0.43	0.63	0.89	0.90	0.84	0.87

Tabelle 7.19: Enthüllungsrisiken nach BGK und Strategie (lückenhaft)

Strategie\BGK	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1000
hybrid	0.11	0.14	0.23	0.28	0.37	0.48
zusammengesetzt	0.57	0.65	0.89	0.89	0.84	0.84

In den Tabellen 7.20 bis 7.22 werden die mit beiden Strategien (hybride und zusammengesetzte Zuordnung) verbundenen Enthüllungsrisiken, verteilt auf Zweisteller der Wirtschaftszweigklassifikation (im Wesentlichen die sogenannten Abteilungen), gegenübergestellt. Diese wurden wie eingangs bemerkt pseudonymisiert, d.h. einer zufälligen Nummerierung unterzogen, um die vorliegende Veröffentlichung zu ermöglichen.

Tabelle 7.20
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, West –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,18	0,22	18,00	0,23	0,31	0,11	0,16	0,36	0,15
Zusammeng.	0,60	0,95	0,73	0,87	0,88	0,69	0,84	0,75	0,73

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,19	0,32	0,17	0,16	0,22	0,20	0,07	0,24
Zusammeng.	0,67	0,83	0,93	0,91	0,86	0,74	0,81	0,61

Tabelle 7.21
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, Ost –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,52	0,32	0,41	0,35	0,56	0,25	0,23	0,59	0,51
Zusammeng.	1,00	0,90	0,50	0,69	0,89	0,59	0,94	1,00	0,83

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,37	0,50	0,36	0,24	0,82	0,35	0,23	0,39
Zusammeng.	0,25	1,00	0,80	0,63	1,00	0,75	0,82	0,80

Tabelle 7.22
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, Gesamt –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,22	0,23	0,21	0,25	0,35	0,13	0,17	0,42	0,20
Zusammeng.	0,64	0,93	0,67	0,81	0,88	0,64	0,87	0,87	0,76

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,22	0,34	0,20	0,16	0,32	0,23	0,10	0,27
Zusammeng.	0,56	0,88	0,89	0,79	0,92	0,74	0,81	0,68

Variante 2: einfache Mikroaggregation mit 6er Gruppenbildung

Nun wird analog zu Variante 1 die Mikroaggregation für jedes Merkmal getrennt durchgeführt. Erneut wird das jeweils zu anonymisierende Merkmal bzw. die zugehörige Spalte der Datenmatrix separat sortiert. Danach werden absteigend immer sechs benachbarte Merkmalsträger in einer Gruppe zusammengefasst und alle ihre stetigen Merkmalswerte durch den Durchschnitt der drei Einzelwerte ersetzt. Am unteren Ende einer Spalte formiert sich in diesem Falle wegen $13294\, mod\, 6 = 4$ eine 4-er Gruppe, die nicht mit der darüberliegenden Gruppe vereinigt wird.

In Tabelle 7.23 werden die mit beiden Strategien (hybride und zusammengesetzte Zuordnung) verbundenen Enthüllungsrisiken, verteilt auf Zweisteller der Wirtschaftszweigklassifikation (im Wesentlichen die sogenannten Abteilungen), gegenübergestellt. Tabelle 7.26 enthält die entsprechenden Enthüllungsrisiken verteilt auf die pseudonymisierten Wirtschaftsabteilungen.

Tabelle 7.23: Enthüllungsrisiken nach BGK und Strategie (lückenhaft), V2

Strategie\BGK	$20-49$	$50-99$	$100-249$	$250-499$	$500-999$	≥ 1000
hybrid	0.10	0.14	0.23	0.28	0.37	0.45
zusammengesetzt	0.57	0.63	0.88	0.89	0.86	0.79

Vegleicht man die Tabellen 7.23 und 7.26 mit den entsprechenden Tabellen 7.19 und 7.22, so erkennt man, dass die erhoffte Steigerung der Schutzwirkung beim Übergang von Variante 1 zu Variante 2 offenbar ausgeblieben ist.

Variante 3: einfache Mikroaggregation mit Varianzerhalt und Extremwertbehandlung

Der Vorteil der Mikroaggregationsverfahren besteht darin, dass die arithmetischen Mittel für einzelne Merkmale erhalten bleiben, Standardabweichungen und Varianzen werden jedoch systematisch reduziert. Die Eigenschaft von systematisch unterschätzten Varianzen führt zudem zu verzerrten Ergebnissen von Parametertests in ökonometrischen Modellen (Brand 2000; Lechner und Pohlmeier 2003).

Bei Variante 3 wird nach Höhne (2004b) eine Modifikation der einfachen Mikroaggregation vorgenommen mit dem Ziele des näherungsweisen Varianzerhaltes. Es werden spaltenweise Gruppen von jeweils vier Merkmalsträgern gebildet, von denen zwei Merkmalsträgern der um die Standardabweichung innerhalb der Gruppe verminderte Gruppendurchschnitt als neuer Merkmalswert zugewiesen wird, den beiden anderen der um die Standardabweichung erhöhte Gruppendurchschnitt. Im vorliegenden Falle von 13 294 Einheiten besteht (für beliebiges Merkmal) die unterste Gruppe aus zwei Elementen. Allgemein bleiben für jede Gruppe die Durchschnitte zwar erhalten, die beiden Teilgruppen sind aber nicht mehr als zu demselben Mikroaggregat gehörig erkennbar.

Die Entscheidung darüber, welche Merkmalsträger den um die Standardabweichung reduzierten Wert zugewiesen erhalten und welche den um die Standardabweichung erhöhten, wird im Sinne eines bestmöglichen Erhaltes der Korrelationsmatrix gefällt. Bei extrem schief verteilten Merkmalen kann es vorkommen, dass der Gruppenmittelwert kleiner als die gruppeninterne Varianz ist und somit bei der Differenzbildung negative anonymisierte Werte entstehen. Da solche Werte oftmals unerwünscht sind (zum Beispiel negative Beschäftigtenangaben), wird in Höhne (2008) eine Anpassung der Gruppenaufteilungen via sogenannter Extremwertbehandlung vorgeschlagen.

Die Tabellen 7.24 und 7.27 enthalten die mit beiden Strategien (hybride und zusammengesetzte Zuordnung) verbundenen Enthüllungsrisiken, verteilt auf Beschäftigtengrößenklassen und pseudonymisierte Wirtschaftsabteilungen. Dabei zeigt sich im Vergleich mit den Tabellen 7.23 und 7.26, dass erneut keine nennenswerte Verbesserung der Datensicherheit gegenüber der einfachen Mikroaggregation mit 6er Gruppenbildung eingetreten ist.

Tabelle 7.24: Enthüllungsrisiken nach BGK und Strategie (lückenhaft), V3

Strategie\BGK	$20-49$	$50-99$	$100-249$	$250-499$	$500-999$	≥ 1000
hybrid	0.10	0.13	0.22	0.28	0.37	0.46
zusammengesetzt	60	65	88	87	84	86

Variante 4: Mehrdimensionale abstandsorientierte Mikroaggregation

Die Grundidee des in Domingo-Ferrer et al. (2008) vorgestellten Verfahrens besteht wie bei den eindimensionalen Varianten darin, möglichst ähnliche Merkmalsträger zu gruppieren. Dabei wird die Menge der Merkmalsträger zusammen mit der Menge der euklidischen Distanzen zwischen diesen Objekten als „Graph" mit „Ecken" und „Kanten" interpretiert. In einem ersten Schritt wird ein „minimal spannender Wald" erzeugt (zur Definition der graphentheoretischen Begriffe siehe Unterabschnitt 3.1.1), wozu ein mit dem bekannten Algorithmus von Kruskal vergleichbarer Algorithmus auf die Menge der Merkmalsträger Anwendung findet. Danach wird in einem zweiten Schritt via geeigneter Zerlegungsprozedur die Eckenmenge derart in Gruppen partitioniert, dass jede Gruppe (sogenannter „Baum") wenigstens k Elemente besitzt. Im dritten Schritt wird durch eine weitere Zerlegung aller Bäume T mit wenigstens $2k$ Ecken die zusätzliche Restriktion erreicht, dass die Implikation

$$(u, v) \in T \Longrightarrow v \text{ ist einer der } k-1 \text{ nächsten Nachbarn von } u$$

erfüllt ist. Es entsteht also insgesamt eine Zerlegung in Bäume T mit $k \leq |T| \leq 2k-1$ Ecken. Abschließend werden die Überschneidungsmerkmale der Merkmalsträger einer Gruppe (bzw. Ecken eines Baumes) simultan gemäß einer vorgegebenen Aggregationstechnik verändert, in unserem Falle aus Gründen der Vergleichbarkeit mit den vorhergehenden Varianten durch Mittelwertbildung. Gäbe es keine kategorialen Überschneidungsmerkmale in den Daten, so wäre das Reidentifikationsrisiko für einen beliebigen Merkmalsträger nach

oben durch $\frac{1}{k}$ beschränkt, da es dann wenigstens k (im vorliegenden Falle $k = 3$) mittels der metrischen Überschneidungsmerkmale nicht unterscheidbare Einheiten in der Datei gäbe.

Die Tabellen 7.25 und 7.28 enthalten die mit beiden Strategien (hybride und zusammengesetzte Zuordnung) verbundenen Enthüllungsrisiken, verteilt auf Beschäftigtengrößenklassen und pseudonymisierte Wirtschaftsabteilungen.[43] Dabei zeigt sich im Vergleich der Tabellen 7.23 und 7.26, dass eine bemerkenswerte Verbesserung der Datensicherheit gegenüber den Varianten der einfachen Mikroaggregation mit fester und variabler Gruppengröße eingetreten ist. Bedauerlicherweise geht im Allgemeinen bei Varianten der mehrdimensionalen Mikroaggregation ein von wissenschaftlicher Seite nicht tolerierbarer Einschnitt in das Analysepotential einher. Eine Eignung solchen Materials für spezielle Nutzerwünsche oder für den Einsatz zu Lehrzwecken ist aber nicht ausgeschlossen.

Tabelle 7.25: Enthüllungsrisiken nach BGK und Strategie (lückenhaft), V4

Strategie\BGK	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1000
hybrid	0.06	0.08	0.09	0.11	0.13	0.16
zusammengesetzt	0.16	0.18	0.26	0.27	0.26	0.30

7.5 Varianten der multiplikativen Zufallsüberlagerung

Wir betrachten im Folgenden vier spezielle Varianten der multiplikativen Zufallsüberlagerung, deren detaillierte Beschreibung in Höhne (2008) zu finden ist. Man beachte hierzu auch die Ausführungen in Abschnitt 1.2.

Aus Geheimhaltungsgründen werden in den Tabellen die Wirtschaftszweigangaben auf der Zweistellerebene wie im vorherigen Abschnitt pseudonymisiert.

Die mit den Varianten 5 bis 8 verbundenen Risiken werden wie zuvor bei den Mikroaggregationsvarianten 2 bis 4 in aggregierter Form dargestellt, um eine Übersichtlichkeit und einen direkten Vergleich der verschiedenen Varianten untereinander zu gewährleisten. D.h., auch in diesem Abschnitt beschränken wir uns auf die Tabellierung der Risiken nach Strategie und den Ausprägungen in den beiden kategorialen Überschneidungsmerkmalen *Beschäftigtengrößenklasse* und *Wirtschaftszweigklassifikation*, in Verbindung mit lückenhaftem Zusatzwissen.

43 An dieser Stelle gebührt dem Kollegen Josep Domingo-Ferrer ein herzlicher Dank für die Bereitstellung des Programmes zur Erzeugung mehrdimensional mikroaggregierter Daten.

Tabelle 7.26
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, V2 –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,22	0,23	0,20	0,25	0,35	0,12	0,16	0,40	0,20
Zusammeng.	0,64	0,86	0,62	0,79	0,86	0,67	0,87	0,88	0,76

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,21	0,35	0,19	0,16	0,32	0,23	0,10	0,26
Zusammeng.	0,60	0,71	0,92	0,74	0,92	0,74	0,77	0,67

Tabelle 7.27
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, V3 –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,22	0,23	0,20	0,25	0,35	0,13	0,16	0,41	0,20
Zusammeng.	0,64	0,88	0,63	0,82	0,89	0,68	0,87	0,87	0,76

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,21	0,34	0,19	0,16	0,32	0,23	0,10	0,27
Zusammeng.	0,56	0,88	0,89	0,79	0,92	0,77	0,79	0,69

Tabelle 7.28
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, V4 –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,10	0,09	0,09	0,10	0,11	0,07	0,08	0,14	0,10
Zusammeng.	0,19	0,24	0,25	0,21	0,30	0,21	0,26	0,29	0,20

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,09	0,11	0,08	0,08	0,10	0,12	0,05	0,11
Zusammeng.	0,17	0,23	0,29	0,25	0,22	0,21	0,18	0,22

Variante 5: Elementweise multiplikative Zufallsüberlagerung

Bei dieser Variante wird jeder Originalwert x_{ij}^{o}, beobachtet für den Merkmalsträger i und das Merkmal j, multiplikativ überlagert mit dem Faktor $1+f\cdot z_{ij}+\varepsilon_{ij}$, wobei z_{ij} mit jeweils 50% Wahrscheinlichkeit den Wert -1 oder $+1$ annimmt und ε_{ij} eine normalverteilte Zufallsgröße mit Erwartungswert Null und Standarabweichung σ bestimmt. Im vorliegenden Falle wird $\sigma = 0.03$ und der sogenannte Verschiebungsparameter $f = 0.1$ gesetzt.

Die Tabellen 7.29 und 7.33 enthalten die mit beiden Strategien (hybride und zusammengesetzte Zuordnung) verbundenen Enthüllungsrisiken, verteilt auf Beschäftigtengrößenklassen und pseudonymisierte Wirtschaftsabteilungen.

Tabelle 7.29: Enthüllungsrisiken nach BGK und Strategie (lückenhaft),V5

Strategie\BGK	$20-49$	$50-99$	$100-249$	$250-499$	$500-999$	≥ 1000
hybrid	0.05	0.06	0.09	0.11	0.13	0.26
zusammengesetzt	0.14	0.14	0.29	0.33	0.44	0.37

Es fällt auf, dass diese Variante eine deutlich höhere Schutzwirkung entfaltet als die eindimensionalen Mikroaggregationsvarianten 1 bis 3, wobei das Gefährdungspotential mit dem der mehrdimensionalen Mikroaggregationsvariante 4 durchaus vergleichbar ist.

Variante 6: Zeilenweise multiplikative Zufallsüberlagerung

Im Unterschied zu Variante 5 wird mit $1+f\cdot z_i+\varepsilon_{ij}$ die Zufallszahl z_i nun zeilenweise gezogen. D.h., die Beobachtungen zu einem bestimmten Merkmalsträger werden alle in derselben Richtung (mit $1+f$ oder $1-f$) verzerrt, wobei sich die Überlagerungsfaktoren durch die weiterhin zellenweise zum Faktor addierte Störgröße ε_{ij} unterscheiden. Dies ist aus Gründen der Datensicherheit unerlässlich, da ein Datenangreifer ansonsten über Quotientenbildung zweier sensibler metrischer Merkmale den zugehörigen Quotienten der beiden Originalmerkmale erzeugen könnte.

Die Tabellen 7.30 und 7.34 enthalten die mit beiden Strategien (hybride und zusammengesetzte Zuordnung) verbundenen Enthüllungsrisiken, verteilt auf Beschäftigtengrößenklassen und pseudonymisierte Wirtschaftsabteilungen.

Tabelle 7.30: Enthüllungsrisiken nach BGK und Strategie (lückenhaft),V6

Strategie\BGK	$20-49$	$50-99$	$100-249$	$250-499$	$500-999$	≥ 1000
hybrid	0.05	0.06	0.07	0.10	0.11	0.22
zusammengesetzt	0.37	0.29	0.42	0.54	0.49	0.63

Der erwartete Anstieg der Enthüllungsrisiken gegenüber der vorherigen Variante 5 der elementweisen multiplikativen Überlagerung blieb im Falle der hybriden Zuordnungsstrategie überraschenderweise aus. Dies gilt sowohl für die Risikoverteilung auf Beschäftigtengrößenklassen (vgl. die oberen Zeilen der Tabellen 7.29 und 7.30), also auch für die entsprechende Verteilung auf Wirtschaftsabteilungen (vgl. die oberen Zeilen der Tabellen 7.33 und 7.34). Möglicherweise wurden hier die erhöhten Reidentifikationsraten, die bei einer Verzerrung aller Beobachtungen eines Merkmalsträgers in dieselbe Richtung vor allem durch den Einsatz von Zusammenhangsmaßen bei der Distanzbildung entstehen, durch ein gleichzeitiges Abnehmen der Anteile brauchbarer Einzelwerte kompensiert.

Verfolgt man jedoch die Strategie der zusammengesetzten Zuordnung, so steigen die Enthüllungsrisiken beachtlich an (vgl. die unteren Zeilen der Tabellen). Da die Zusammenhangsmaße bei Verzerrung in dieselbe Richtung zuverlässiger werden, wird es wahrscheinlicher, dass Paare von Merkmalsträgern, die durch mehrere Maße zur Zuordnung vorgeschlagen werden, tatsächlich zu demselben Individuum bzw. Unternehmen gehören.

Variante 7: Eindimensionale kontrollierte multiplikative Zufallsüberlagerung

Die Überlagerung erfolgt wie in Variante 6 zeilenweise. Allerdings wird zunächst der Datenbestand nach einem bestimmten Merkmal, im vorliegenden Falle nach dem Merkmal *Gesamtumsatz 1999* absteigend sortiert.

Die Auswahl des Faktors z_i erfolgt nun zeilenweise nach dem aktuellen Fehler in der Summe der Abweichungen zwischen anonymisierten Werten x_{ij}^a und originalen Werten x_{ij}^o in diesem Merkmal:

$$z_i = -1 \iff A_i := \sum_{l=1}^{i-1}(x_{lj}^a - x_{lj}^o) \geq 0, \qquad (7.1)$$

$$z_i = +1 \iff A_i := \sum_{l=1}^{i-1}(x_{lj}^a - x_{lj}^o) < 0.$$

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

Die Tabellen 7.31 und 7.35 enthalten die mit beiden Strategien (hybride und zusammengesetzte Zuordnung) verbundenen Enthüllungsrisiken, verteilt auf Beschäftigtengrößenklassen und pseudonymisierte Wirtschaftsabteilungen.

Erwartungsgemäß entfalten die Varianten 6 und 7 vergleichbare Schutzwirkung. Erwartungsgemäß deshalb, weil die wesentliche Neuerung von Variante 7 gegenüber Variante 6 lediglich in dem näherungsweisen Erhalt der univariaten Verteilungen liegt. Dabei spielt es hinsichtlich der Datensicherheit auch keine Rolle, dass bei Variante 6 „echte" Zufallszahlen

erzeugt werden, wohingegen bei Variante 7 massiv in den Zufallsprozess (zumindest bei der Festlegung der tendentiellen Veränderung in eine bestimmte Richtung) eingegriffen wird.

Tabelle 7.31: Enthüllungsrisiken nach BGK und Strategie (lückenhaft),V7

Strategie\BGK	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1000
hybrid	0.05	0.06	0.08	0.10	0.12	0.22
zusammengesetzt	0.40	0.37	0.45	0.51	0.52	0.59

Variante 8: Mehrdimensionale kontrollierte multiplikative Zufallsüberlagerung

Der wesentliche Unterschied zu Variante 7 besteht nun darin, dass der Datenbestand nicht nach einem bestimmten Merkmal, sondern nach dem Durchschnitt der normierten Beträge aller metrischen Merkmale absteigend sortiert wird.

Die Auswahl des Faktors z_i erfolgt wiederum nach dem aktuellen Fehler in der Summe der Abweichungen zwischen anonymisierten und originalen Werten, wobei entsprechend dem mehrdimensionalen Sortierkriterium die Summe der normierten Abweichungen, summiert über alle k metrischen Merkmale, als Referenz dient. Als Abweichungsmaß wird hierzu

$$A_i \quad := \quad \sum_{j=1}^{k} \frac{|\sum_{l=1}^{i-1}(x_{lj}^a - x_{lj}^o)|}{x_{\cdot j}^o} \tag{7.2}$$

verwendet.

Die Tabellen 7.32 und 7.36 enthalten erneut die mit beiden Strategien (hybride und zusammengesetzte Zuordnung) verbundenen Enthüllungsrisiken, verteilt auf Beschäftigtengrößenklassen und pseudonymisierte Wirtschaftsabteilungen.

Tabelle 7.32: Enthüllungsrisiken nach BGK und Strategie (lückenhaft),V8

Strategie\BGK	20 – 49	50 – 99	100 – 249	250 – 499	500 – 999	≥ 1000
hybrid	0.04	0.04	0.06	0.09	0.10	0.18
zusammengesetzt	0.28	0.23	0.38	0.41	0.38	0.53

Die deutliche Reduktion der Enthüllungsrisiken gegenüber der zuvor betrachteten Variante 7 ist nicht leicht zu erklären. Hinsichtlich der Brauchbarkeit aufgedeckter Informationen dürfte es keinen systematischen Unterschied zwischen den beiden Varianten geben. Möglicherweise sind Reidentifikationen bei Variante 7 etwas wahrscheinlicher, da der Datenbestand im Zuge der Anonymisierung nach dem Merkmal *Gesamtumsatz*

1999 sortiert wurde, der hoch korreliert mit den sieben weiteren für die Zuordnung entscheidenden Überschneidungsmerkmalen *Gesamtumsatz 2000 – 2002* und *Anzahl der Beschäftigten 1999 – 2002* ist. Im Gegensatz hierzu gingen bei der Erzeugung von Variante 8 alle circa 130 metrischen Merkmale in die Sortierung ein, wodurch der Zusammenhang zwischen den Überschneidungsmerkmalen etwas mehr zerstört und damit die Qualität der in der Simulation berechneten Zusammenhangsmaße beeinträchtigt wurde.

Obwohl alle in diesem Kapitel getesteten Anonymisierungsvarianten so konzipiert wurden, dass sie gerade bei großen Unternehmen besonders wirken, steigen die Enthüllungsrisiken mit wachsender Mitarbeiterzahl der Unternehmen teilweise beträchtlich. Dennoch erweisen sich die einfachen Mikroaggregationsvarianten gerade bei den großen Unternehmen nicht nur gegenüber dem klassischen Distanzmaß, sondern auch gegenüber den Zusammenhangsmaßen als schützend, da hier oftmals dieselben Gruppen bei verschiedenen Merkmalen gebildet wurden. Bei den Modellen der Zufallsüberlagerung sinken die Enthüllungsrisiken gegenüber den Mikroaggregationsvarianten noch einmal deutlich. Die stärkere Veränderung der Einzelwerte bei der Überlagerung bewirkt zum einen eine Reduktion der Reidentifikationen, zum anderen sinkt der Anteil der für den Datenangreifer brauchbaren Einzelwerte erheblich. Die Varianten der multiplikativen Zufallsüberlagerung sind aus diesen Gründen grundsätzlich besser für eine faktische Anonymisierung geeignet als die einfachen Mikroaggregationsvarianten. Die mehrdimensionalen Mikroaggregationsvarianten wie beispielsweise Variante 4 sind mangels Erhalt des Analysepotentials sowohl für Quer- als auch für Längsschnittsdaten nicht oder nur eingeschränkt verwertbar (Rosemann 2005; Ronning et al. 2005; Ronning et al. 2009).

Abschließend muss erwähnt werden, dass mit den in Beispielsimulationen geschätzten Enthüllungsrisiken vorsichtig umgegangen werden sollte. Es ist offensichtlich, dass mit wachsender Anzahl von Kombinationen der kategorialen Überschneidungsmerkmale auch der Maximalwert der in den Tabellen auftauchenden Risiken ansteigen wird. Um realistisch zu bleiben, sollten daher nur einige ausgewählte Kombinationen, die möglichst der Suchstrategie eines realen Datenangreifers anzupassen sind, in die Beurteilung der faktischen Anonymität einer Datei einfließen.

Tabelle 7.33
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, V5 –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,12	0,09	0,09	0,11	0,14	0,05	0,05	0,24	0,11
Zusammeng.	0,25	0,27	0,07	0,25	0,56	0,28	0,29	0,40	0,47

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,10	0,21	0,09	0,07	0,20	0,08	0,04	0,11
Zusammeng.	0,15	0,60	0,38	0,28	0,38	0,27	0,21	0,23

Tabelle 7.34
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, V6 –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,10	0,08	0,08	0,08	0,13	0,05	0,04	0,21	0,09
Zusammeng.	0,68	0,60	0,47	0,23	0,45	0,35	0,47	0,72	0,46

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,08	0,18	0,08	0,06	0,18	0,08	0,05	0,10
Zusammeng.	0,29	0,51	0,40	0,30	0,55	0,36	0,46	0,43

Tabelle 7.35
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, V7 –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,09	0,08	0,09	0,08	0,12	0,04	0,05	0,23	0,09
Zusammeng.	0,48	0,53	0,32	0,31	0,59	0,26	0,54	0,65	0,64

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,09	0,20	0,07	0,06	0,15	0,09	0,05	0,13
Zusammeng.	0,36	0,67	0,55	0,35	0,48	0,59	0,41	0,51

Tabelle 7.36
Enthüllungsrisiken nach Wirtschaftszweigen (WZ 93) und Strategie
– lückenhaft, V8 –

Strategie	Wirtschaftszweig								
	01	02	03	04	05	06	07	08	09
Hybrid	0,10	0,07	0,08	0,06	0,09	0,04	0,04	0,16	0,09
Zusammeng.	0,38	0,64	0,23	0,24	0,35	0,31	0,29	0,60	0,41

Strategie	Wirtschaftszweig							
	10	11	12	13	14	15	16	17
Hybrid	0,05	0,12	0,07	0,04	0,20	0,06	0,04	0,13
Zusammeng.	0,25	0,46	0,44	0,23	0,37	0,39	0,23	0,43

Kapitel 8

Entstehungsprozess faktisch anonymer Daten für die Wissenschaft

Bei der Generierung faktisch anonymer Daten für die Wissenschaft sind folgende zumeist gegenläufige Ziele zu verfolgen: Einerseits muss die faktische Anonymität gewährleistet sein, andererseits muss ein großes Analysepotential in den Daten für möglichst viele Nutzerprofile erhalten bleiben. Von theoretischer Seite optimal ist es daher, die Anonymisierungsmaßnahmen so zu wählen, dass das Analysepotential „maximiert" wird unter der Nebenbedingung, dass die faktische Anonymität gerade noch gegeben ist. Während das Enthüllungsrisiko für Einzelwerte einer wirtschaftsstatistischen Erhebung als eindimensionales Zielkriterium darstellbar ist, ist das Analysepotential aufgrund der Vielfältigkeit der zu beachtenden Aspekte 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 Analysepotential häufig nicht quantifizierbar oder es bestehen gar widersprüchliche Bewertungen aus der Sicht unterschiedlicher Gruppen potentieller Datennutzer.

Dies wird bereits durch folgendes Beispiel deutlich: 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 der industrieökonomisch orientierte Forscher eher die erste Variante präferieren. Damit wird bereits angedeutet, dass es in der Praxis „den" optimalen Weg zur Erzeugung faktisch anonymen Datenmaterials für die Wissenschaft nicht gibt, selbst wenn man sich auf eine bestimmte Erhebung beschränkt. Bei der Weitergabe von Scientific-Use-Files, d.h. von standardisierten für einen breiten Nutzerkreis brauchbaren Daten, würde man in dem geschilderten Beispiel versuchen, sowohl Regional- als auch Wirtschaftszweiginformationen in den Daten zu belassen.

Ferner soll die erzeugte Datei für unterschiedliche 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 überhaupt nicht beschäftigen, wenn die Daten beispielsweise mittels Vertauschungsverfahren 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 betrachtet wird, beispielsweise bei der Schätzung von stochastischen Modellen, ist diese Methode aus Nutzersicht nicht mehr akzeptabel, weil die Zusammenhänge zwischen den Merkmalen stark verändert werden.

Es ist a priori nicht klar, ob und in welcher Form ein Merkmal in späteren Analysen Verwendung finden wird. Der Erhalt der univariaten Verteilung dieses Merkmals ist verhältnismäßig leicht zu realisieren. Taucht das Merkmal aber beispielsweise in einem Regressionsmodell auf, so ist zu prüfen, ob die Koeffizientenschätzung nach der Anonymisierung genauso funktioniert wie vorher, und zwar unabhängig davon, ob das anonymisierte Merkmal als erklärende oder abhängige im Modell auftritt. Auch die Situation, in welcher die Merkmale auf beiden Seiten anonymisiert wurden, muss Beachtung finden.

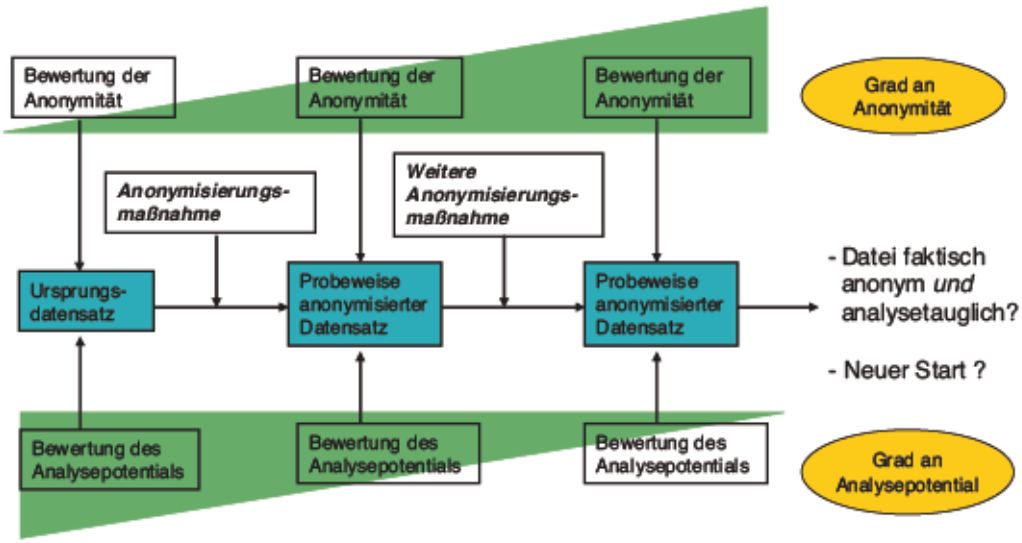
Es kommt hinzu, dass durch die Möglichkeit, verschiedene Anonymisierungsverfahren miteinander zu kombinieren bzw. die Notwendigkeit, bei gemeinsamer Anonymisierung von metrischen und kategorialen Merkmalen einen Methodenmix einzusetzen, die Festlegung eines optimalen Verfahrens ein hochdimensionales Entscheidungsproblem ist. Somit ist auch die Formulierung eines effizienten Algorithmus zur systematischen Suche nach dem optimalen Anonymisierungsverfahren in der Praxis nicht möglich. Naturgemäß ist die Bewertung aus Sicht der Datenanbieter anders als aus Sicht der Datennutzer. Gleichwohl müssen sich beide Seiten in einem diskursiven, interaktiven Prozess auf eine Datei verständigen, welche die Schutzwirkungsanforderungen – gerade – erfüllt und gleichzeitig ein möglichst großes Analysepotential aufweist. Das Vorgehen bei einem solchen Prozess ist in Abbildung 8.1 dargestellt.

Auf dem Wege zu einer faktischen Anonymisierung einer gegebenen Unternehmenserhebung besteht der erste Schritt in einer gründlichen Recherche des möglichen Zusatzwissens (Abschnitt 8.1). Danach ist in Abstimmung mit potentiellen Datennutzern eine Vorauswahl zu testender informationsreduzierender und, falls es aus Gründen der Datensicherheit nicht vermeidbar sein sollte, datenverändernder Anonymisierungsmethoden (Abschnitt 8.2) zu treffen. Die in Abschnitt 8.3 an einem kleinen Beispiel skizzierte Ausbalancierung der Parameter dieser Methoden geschieht ebenfalls in einem diskursiven Prozess zwischen Datenanbieter und Datennutzer. Um die einzelnen Tests auf faktische Anonymität effizient durchführen zu können, empfiehlt sich die Anwendung eines nutzerfreundlichen Simulationsprogrammes, etwa des in Abschnitt 8.4 vorgestellten Programmes „Destatis-Anonymeter". Das Kapitel schließt ab mit einigen Beispielen von bereits über beliebige Datenzugänge bereitgestellten amtlichen Unternehmens- und Betriebserhebungen (Abschnitt 8.5).

8.1 Recherche über das Zusatzwissen

Zu Beginn der Untersuchung sollte eine ausführliche Recherche über das mögliche Zusatzwissen eines potentiellen Datenangreifers durchgeführt werden (zum Vergleich siehe Abschnitt 2.3). Bereits hier können kritische, bei Datenangriffen besonders gefährdete Bereiche in den Daten aufgedeckt werden. Diese sind zum Teil den Fachleuten schon vor der Recherche bekannt (z.B. dünne Besetzungszahlen in Tabellen einer Fachserie des Statistischen Bundesamtes). Des Weiteren sollten sich die Fachleute der Erhebung und der Anonymisierung zusammensetzen und sich auf Schwellen für die Brauchbarkeit (z.B. $\gamma_i = 0.05$ für alle Merkmale) von Einzelangaben sowie eine obere Risikoschwelle (z.B. $\tau = 0.5$) für das Enthüllungsrisiko verständigen.

Abbildung 8.1
Prozess zur Erstellung faktisch anonymen Datenmaterials



In einem nächsten Schritt sollte versucht werden, eine repräsentative 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. In den meisten Fällen liegen in beiden Dateien Merkmale über Namen und Adressen vor, welche aber deutlich voneinander abweichen können. Die Problematik des Zusammenspielens von Adressdaten wird in Lenz et al. (2004a) anhand empirischer Untersuchungen mit Daten der kommerziellen MARKUS-Datenbank und der Kostenstrukturerhebung im Verarbeitenden Gewerbe dargestellt.

Ist das Zusatzwissen in Form einer Datenbank vorhanden, so kann der in Abschnitt 2.2 definierte und in Kapitel 3 formalisierte Massenfischzug simuliert werden. Im Gegensatz zu den Beispielsimulationen in Kapitel 4 hat jedoch ein Datenangreifer keine Möglichkeit, die durch das Programm getroffenen Zuordnungen auf Korrektheit zu überprüfen.

Es ist sinnvoll, den Massenfischzug zunächst bei formal anonymisierten Daten, die aus den Originaldaten durch Wegnahme 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 aber sehr zeitaufwendig sein (vgl. Unterabschnitte 4.1.3, 4.2.3 und vor allem 4.3.3).

8.2 Vorauswahl von Anonymisierungsmethoden

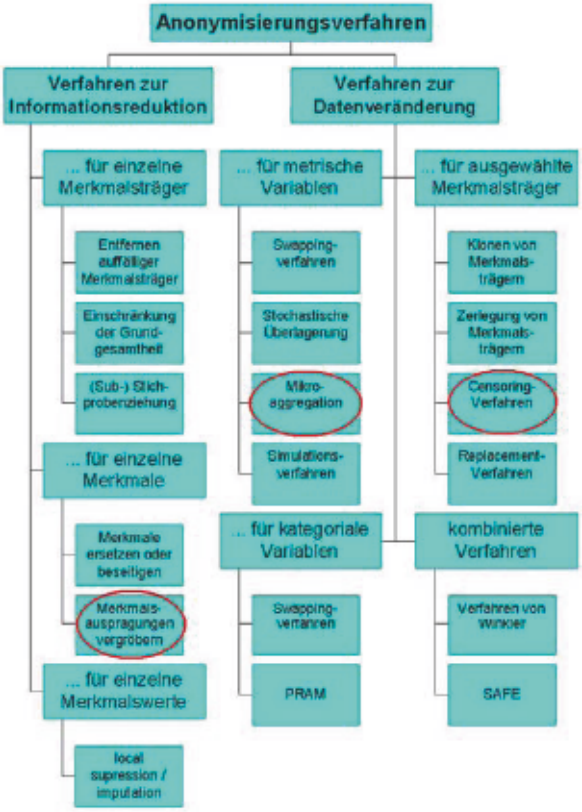
Bei der Entwicklung eines Anonymisierungskonzeptes – mit besonderem Fokus auf die zuvor aufgedeckten gefährdeten Bereiche – erscheint oftmals eine Mischung aus informationsreduzierenden und datenverändernden Methoden als beste Lösung (siehe Abbildung 8.2). Informationsreduzierende Methoden wie die Entfernung von Merkmalen, welche 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 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 Abschnitt 4.3) oder mit geringfügiger Datenveränderung (z.B. bei Unternehmen der Kostenstrukturerhebung im Verarbeitenden Gewerbe mit maximal 49 Beschäftigten, siehe Abschnitt 4.1) 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 Abschnitt 4.2).

8.3 Geheimhaltung versus Analysepotential

Hat man sich bei den datenverändernden Anonymisierungsverfahren auf eine Verfahrensgruppe verständigt, so müssen im Folgenden die Parameter des Verfahrens ausbalanciert und Datenangriffe solange simuliert werden, bis der gewünschte Anonymisierungsgrad erreicht und die vorgegebene obere Risikoschwelle (siehe Abschnitt 2.5) unterschritten wird.

Abbildung 8.2
Vorauswahl der Anonymisierungsverfahren

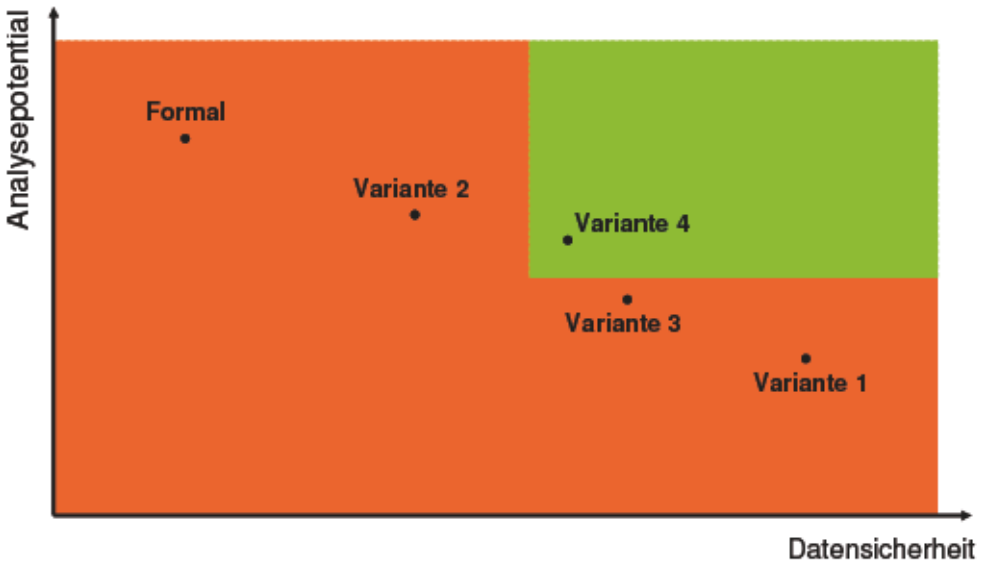


Wie in der Vorbemerkung bereits beschrieben wurde, sollten diese Untersuchungen nicht isoliert von der Bewertung des Analysepotentials durchgeführt werden.

In der Abbildung 8.3 wird der klassische Zielkonflikt am Beispiel von fünf Anonymisierungsvarianten einer Datei verdeutlicht. Die Abbildung stellt die Auflösung des Zielkonfliktes sehr vereinfachend dar, da wie eingangs erwähnt der Grad des Analysepotentials einer Datei im Gegensatz zur Datensicherheit nicht wie abgebildet auf einer linearen Skala messbar ist. Die Variante FORMAL bezeichne die formal anonymisierten Daten, die bestmögliches Analysepotential aufweisen. Sollten diese Daten kein ausreichendes Analysepotential entfalten, so muss geprüft werden, ob Merkmale anderer Erhebungen mit demselben Berichtskreis (z.B. Kostenstrukturerhebung im Verarbeitenden Gewerbe verknüpft mit der Umsatzsteuerstatistik oder dem amtlichen Unternehmensregister) hinzugespielt werden können. Da der natürliche Schutz einer Datei in der Regel nicht ausreicht, werden im nächsten Schritt zur Erreichung der notwendigen Schutzwirkung informationsreduzierende Anonymisierungsmaßnahmen (z.B. Ausweisung der Branchenzugehörigkeit auf der WZ-Zweistellerebene und Entfernung der Regionalangabe) angewendet. Die so entstehende Variante 1 weist nun Schwächen bei der Auswertung von Teilmassen auf (z.B. sind branchenspezifische Analysen auf WZ-Dreistellerebene und solche mit Regionalbezug nicht mehr durchführbar), weshalb

eine schwächere Variante 2 der Informationsreduktion (z.B. Ausweisung der WZ-Dreisteller und/oder Hinzunahme des administrativen Regionsschlüssels) bei gleichzeitiger Anwendung datenverändernder Maßnahmen (z.B. eindimensionale Mikroaggregation) untersucht wird. Da diese Variante gemäß Abbildung die vorgegebene zulässige Risikoschwelle über- bzw. den Grad an Datensicherheit unterschreitet, wird eine neue Variante 3 mit stärkeren datenverändernden Maßnahmen entwickelt (z.B. mehrdimensionale Mikroaggregation). Diese Variante erfüllt zwar wieder die Anforderungen an den Datenschutz, erhält aber das Analysepotential der Daten nicht ausreichend (z.B. Zerstörung multivariater Zusammenhänge zwischen den Analysevariablen). Erfolgreich beendet wird der diskursive Prozess mit der Entwicklung der Variante 4 (z.B. Vergröberung des Regionalmerkmals in ein binäres Merkmal „neue/alte Bundesländer" plus eindimensionale Mikroaggregation und Abschneideverfahren bei den drei Branchenführern auf WZ-Dreistellerebene), die sowohl die zulässige Risikoschwelle unterschreitet als auch das Analysepotential für den ausgewählten Nutzerkreis ausreichend erhält.

Abbildung 8.3
Diagramm Analysepotential-Datensicherheit

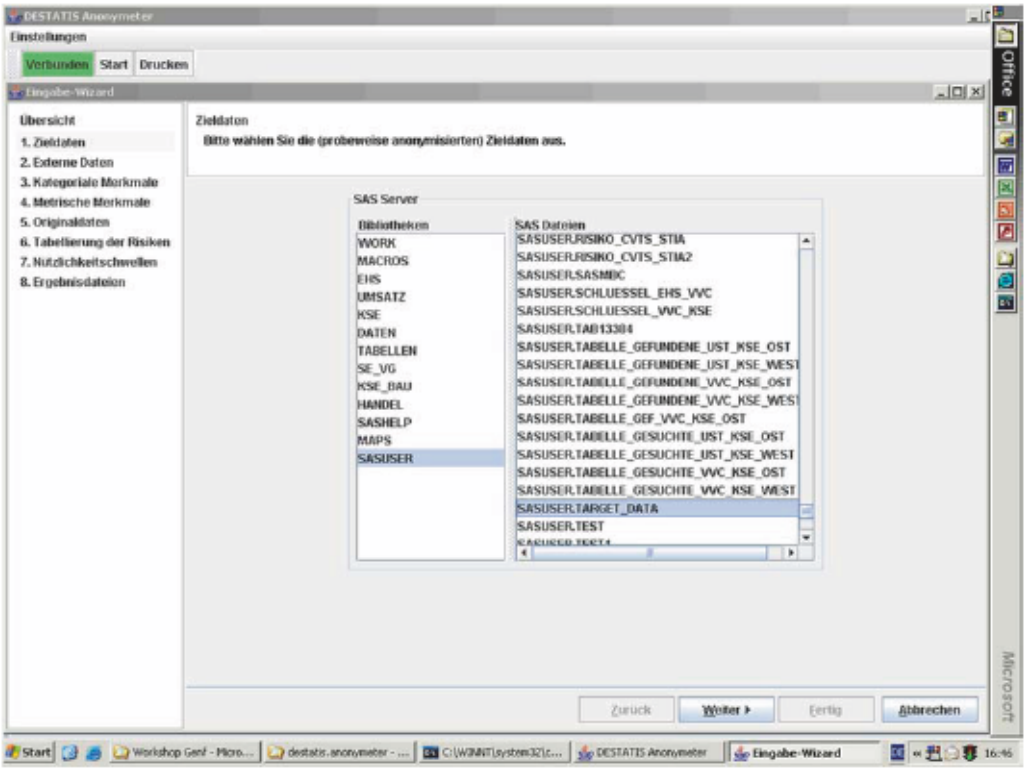


8.4 Simulationsprogramm Destatis – Anonymeter

In diesem Abschnitt wird ein Programm zur Simulation von Datenangriffen vorgestellt, basierend auf den Überlegungen in Kapitel 3. Das Programm ist im Jahre 2005 im Rahmen des Projektes FAWE entstanden und wird seither im Statistischen Bundesamt zur Überprüfung der Datensicherheit bei zuvor probeweise anonymisierten amtlichen Unternehmens- und Be-

triebsdaten eingesetzt. Unter anderem konnten die in nachfolgendem Abschnitt 8.5 vorgestellten Erhebungen nach vorheriger (aus Sicht des Datenangreifers) erfolgloser Anwendung des Programmes der Wissenschaft zur Verfügung gestellt werden.

Abbildung 8.4
Destatis–Anonymeter: Programmstart, Einlesen der Zieldaten



Die Programmierung erfolgte im Wesentlichen via Statistikanalysesoftware SAS/IML.[44] Aus diesem Grunde muss vor der ersten Anwendung des Programmes entweder SAS lokal auf dem Rechner installiert sein oder innerhalb eines Netzwerkes eine Verknüpfung mit einem SAS-Server hergestellt werden.[45] Vor der Anwendung des Programmes müssen drei Dateien im SAS-Format vorliegen. Diese sind die Originaldaten, die (anonymisierten) Zieldaten und die externen Daten (Zusatzwissen). Die erste Spalte muss jeweils das Identifikatormerkmal (z.B. die in der amtlichen Statistik vorhandene Unternehmensregisternummer)

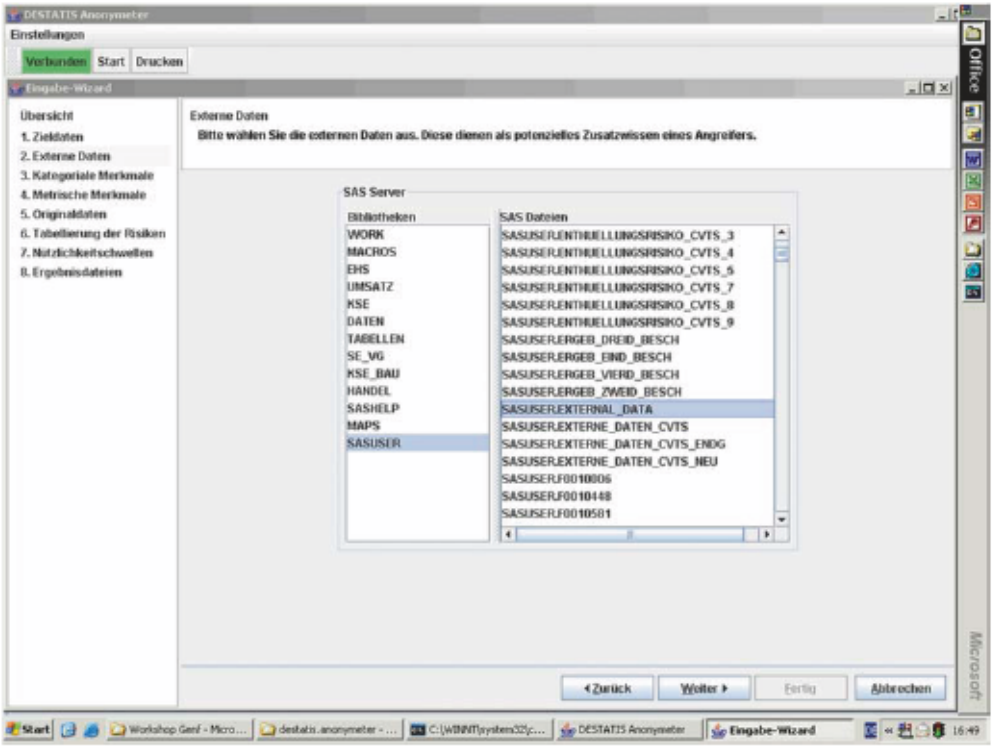
44 Eine Einführung in SAS und in das Modul SAS/IML (Interactive Matrix Laboratory) geben Krämer et al. (2008)

45 Eine in C programmierte Version liegt ebenfalls vor. Diese wurde ebenfalls vom Autor im Rahmen des EU-Methodenprojektes CASC (Computational Aspects of Statistical Confidentiality) für das frei erhältliche μ-Argus Paket entwickelt.

aus technischen Gründen in dem für metrische Merkmale üblichen numerischen Format enthalten. Zudem werden auch die Originaldaten in derselben Struktur wie die anonymisierten Zieldaten benötigt.

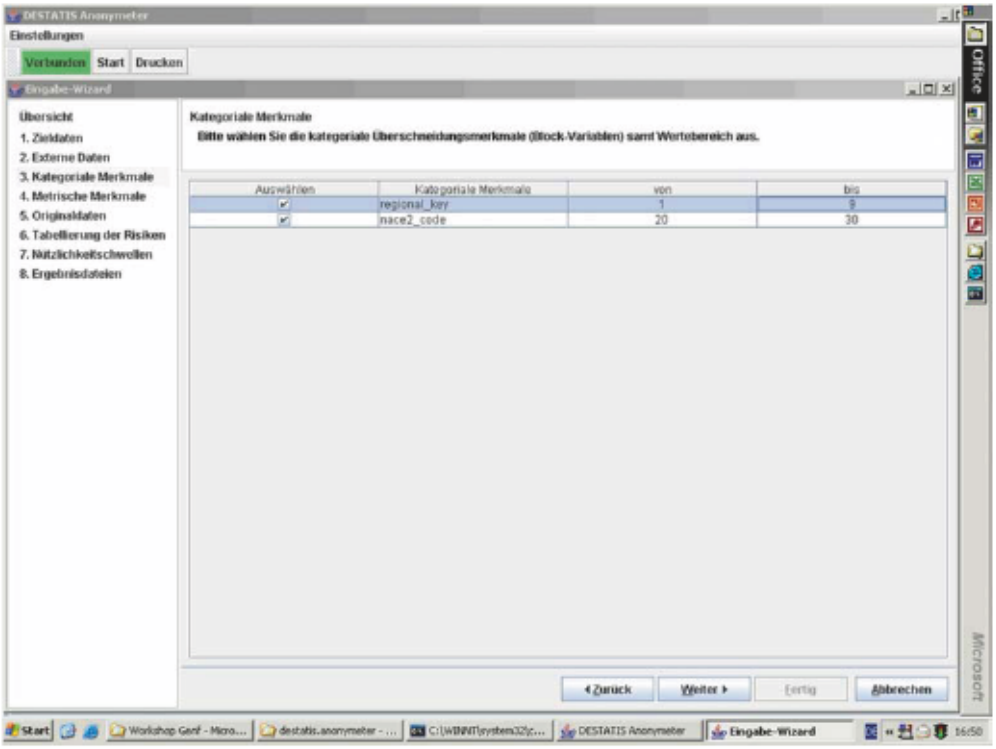
Nach dem Start des Programmes wird durch Klick auf die Schaltfläche „Verbinden" (links oben in der Menüleiste) eine Verbindung zu SAS hergestellt. Danach wird der Nutzer Schritt für Schritt durch das Programm geführt. Zunächst wird er aufgefordert, die Zieldaten auszuwählen (siehe Abbildung 8.4). Hierzu kann der Nutzer in einem kleinen Browser die entsprechende Datei im SAS-Verzeichnis suchen. Nach dem Klick auf die Schaltfläche „weiter" werden die externen Daten auf dieselbe Weise eingelesen (siehe Abbildung 8.5). Auf der linken Seite der Programmoberfläche kann der Nutzer zu jeder Zeit den Ablauf sowie den aktuellen Stand (grau unterlegt) erkennen. Mittels der Schaltfläche „zurück" ist es zudem jederzeit möglich, vorherige Einstellungen zu überprüfen und gegebenenfalls zu ändern.

Abbildung 8.5
Destatis–Anonymeter: Einlesen der externen Daten



Im nächsten Schritt werden die kategorialen Überschneidungsmerkmale festgelegt (siehe Abbildung 8.6). Diese Merkmale werden im späteren Zuordnungsprozess als Blockmerkmale verwendet. Es ist also nötig, dass diese Merkmale in beiden Datenquellen (externe Daten und Zieldaten) denselben Wertebereich besitzen und im Falle hierarchischer Merkmale in derselben Gliederungstiefe vorliegen.

Abbildung 8.6
Destatis–Anonymeter: Auswahl der kategorialen Überschneidungsmerkmale



Bei der nachfolgenden Auswahl der metrischen Überschneidungsmerkmale wird durch das Programm automatisch ein Häkchen an das Identifikatormerkmal gesetzt (siehe Abbildung 8.7). Hierdurch wird der Nutzer daran erinnert, dass die eingelesenen SAS-Dateien in der ersten Spalte das Identifikatormerkmal enthalten müssen. Dies ist deshalb wichtig, da ansonsten ein anderes metrisches Überschneidungsmerkmal als Identifikatormerkmal interpretiert würde. Dieser Fehler wäre von semantischer Art und würde durch das Programm bedauerlicherweise nicht erkannt, obwohl er spätestens bei der Ausgabe niedriger Risiken zu Programmende auffallen müsste. Das Identifikatormerkmal wird in der letzten Phase des Programmes, nach erfolgter Zuordnung sämtlicher Merkmalsträger der externen Daten zu geeigneten Merkmalsträgern der Zieldaten, zur Überprüfung der Zuordnungen auf Korrektheit benötigt.

Das Einlesen der Originaldaten funktioniert genauso wie das anfängliche Einlesen der anderen beiden Datenquellen (siehe Abbildung 8.8). Diese Daten werden benötigt, um nach erfolgter Zuordnung und Überprüfung derselben auf Korrektheit den Anteil der durch den Datenangriff enthüllten und brauchbaren Informationen zu bestimmen. Aus diesem Grunde ist es erforderlich, dass Original- und Zieldaten dieselbe Struktur besitzen, um paarweise die relativen Abweichungen zwischen zueinander gehörigen metrischen Merkmalen berechnen zu können.

Abbildung 8.7
Destatis–Anonymeter: Auswahl der metrischen Überschneidungsmerkmale

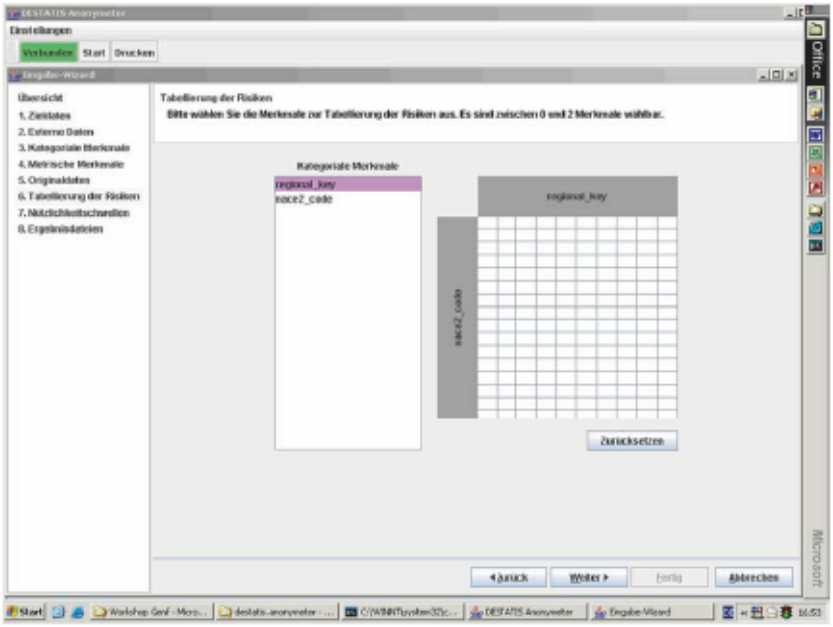
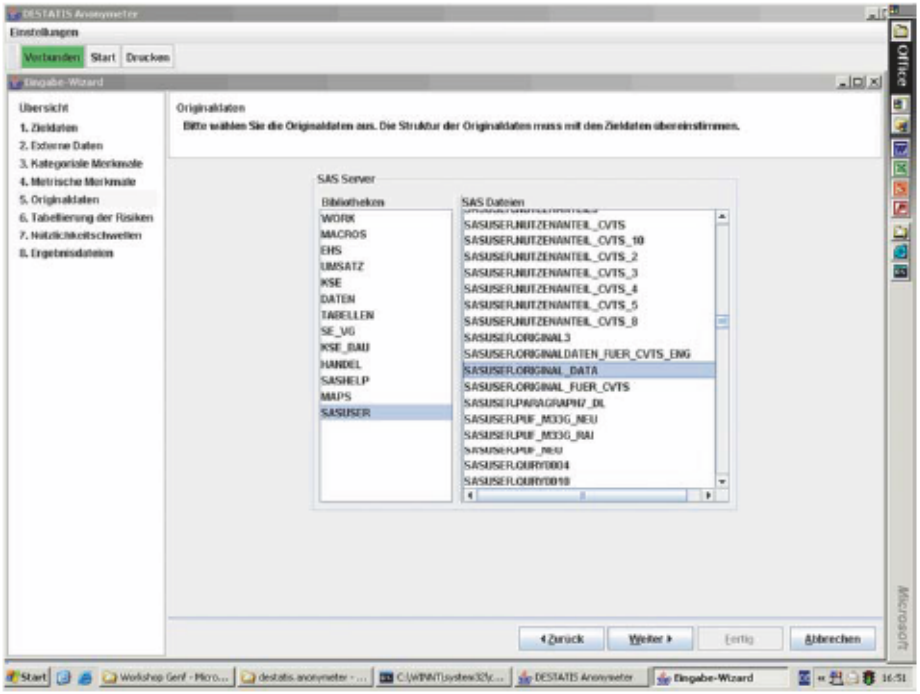


Abbildung 8.8
Destatis–Anonymeter: Einlesen der Originaldaten



Anschließend gelangt der Nutzer zur Auswahl der kategorialen Merkmale zwecks späterer Tabellierung der Risiken (siehe Abbildungen 8.9 und 8.10).

Wählt er kein Merkmal aus, so wird ein globales Risiko (d.h., eine Zahl zwischen 0 und 1 als Schätzer für das mit den Zieldaten verbundene Risiko der Enthüllung von Einzelwerten) berechnet. Wählt der Nutzer nur ein Merkmal aus, beispielsweise das Merkmal *Wirtschaftszweigklassifikation*, so werden die Risiken branchenspezifisch ausgewiesen. Maximal ist die Auswahl zweier Merkmale zur Tabellierung der Risiken möglich. Hierzu müssen die beiden Merkmale nach dem „Drag and Drop"-Prinzip mit festgehaltener linker Maustaste an die gewünschte Position in der Tabelle (Zeile oder Spalte) gezogen werden.

Die Einschränkung auf maximal zwei kategoriale Merkmale ist nicht wesentlich, da in der Praxis davon abgeraten wird, mehr als zwei kategoriale Merkmale zur Tabellierung zu verwenden. Andernfalls kann der Datenangriff dennoch simuliert werden, indem die Daten für mehrere Programmdurchläufe blockweise eingelesen werden. Zum Beispiel wurden zum Testen der Sicherheit der anonymisierten Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe die drei kategorialen Merkmale *Wirtschaftszweigklassifikation, Regionalkennung* und *Beschäftigtengrößenklasse* zur Tabellierung verwendet. Letzteres Merkmal ist durch geeignete Vergröberung des metrischen Überschneidungsmerkmals *Anzahl der Beschäftigten* entstanden. Da das Merkmal *Regionalkennung* den kleinsten Wertebereich besaß (neue und alte Bundesländer), mussten lediglich zwei Programmdurchläufe, jeweils für die Teilgesamtheiten „neue" und „alte Bundesländer", durchgeführt werden.

Abbildung 8.9
Destatis–Anonymeter: Tabellierung der Risiken I

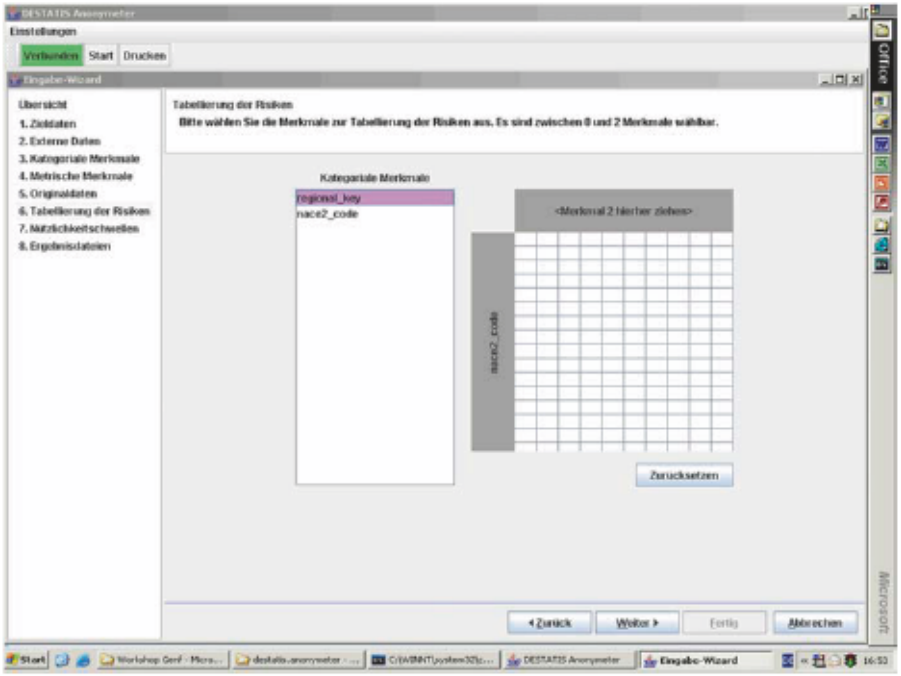
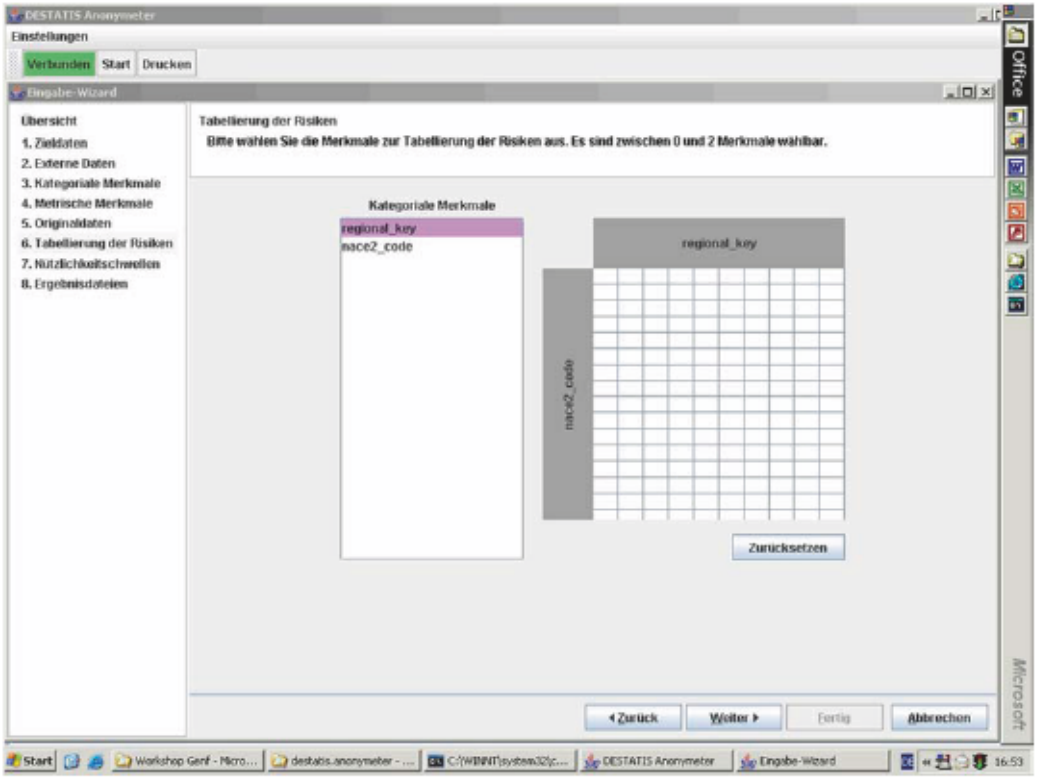


Abbildung 8.10
Destatis–Anonymeter: Tabellierung der Risiken II



Um beurteilen zu können, ob ein durch den Datenangreifer enthüllter Einzelwert für ihn auch brauchbar ist, müssen nun Nützlichkeitsschwellen γ_i festgelegt werden. Zum Beispiel bedeutet $\gamma_i = 0.1$, dass ein im Merkmal v_i enthüllter Einzelwert dann als für den Datenangreifer nützlich eingeschätzt wird, wenn er relativ um weniger als 10% von seinem zugehörigen Originalwert abweicht. Die Schwellen können global oder für jedes Merkmal einzeln gesetzt werden (siehe Abbildung 8.11).

Abschließend wird der Nutzer aufgefordert, Dateinamen einzugeben, unter denen die Programmausgaben später abgespeichert werden sollen (siehe Abbildung 8.12). Ausgegeben werden die reidentifizierten Merkmalsträger (eine Spalte, die die eindeutigen Identifikatoren wie etwa Unternehmensnummern der korrekt zugeordneten Merkmalsträger enthält), die zellenweisen Reidentifikationsrisiken, die zellenweisen Nutzenanteile und die zellenweisen Enthüllungsrisiken in der zuvor definierten Tabellierung. Desweiteren ist eine obere Risikoschwelle τ einzutragen, im Beispiel wird $\tau = 0.5$ gesetzt.

Abbildung 8.11
Destatis–Anonymeter: Schwellen der Brauchbarkeit enthüllter Einzelwerte

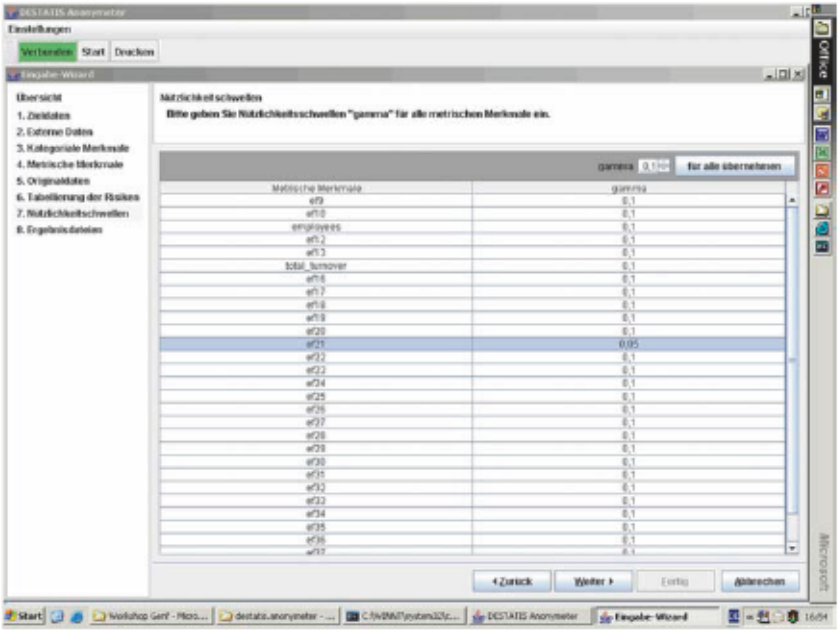
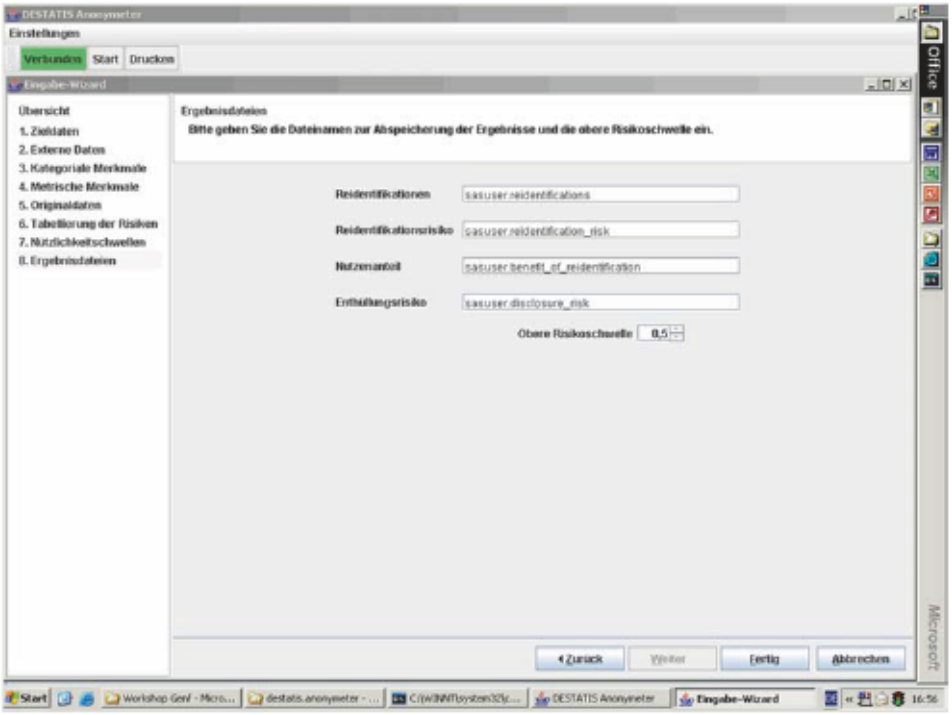
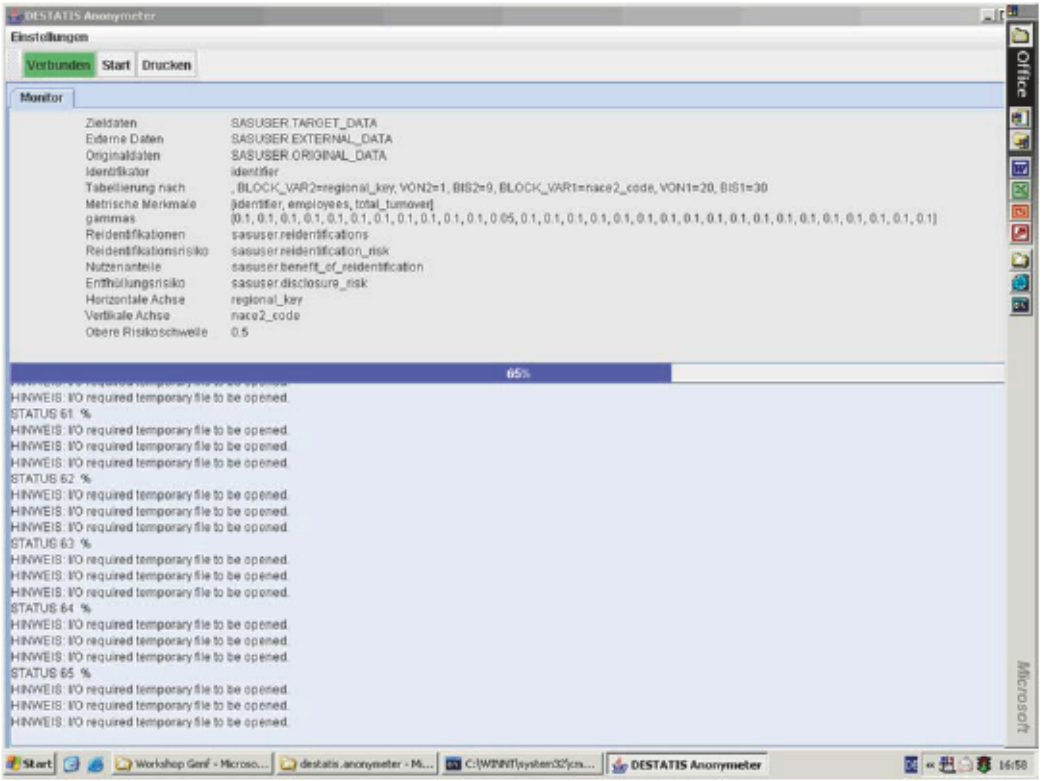


Abbildung 8.12
Destatis–Anonymeter: Speicherung der Programmausgabe



Nun sind alle notwendigen Parameter des Programmes eingestellt. Durch Klick auf die Schaltfläche „Fertig" wird das eigentliche Zuordnungsverfahren gestartet (siehe Abbildungen 8.13 und 8.14). Während der im Hintergrund laufenden Berechnungen kann der Nutzer anhand des dunkelblauen Balkens die Wartezeit abschätzen. Der dunkelblaue Balken teilt das Monitorfenster (oberer Bereich), welches noch einmal die Grundeinstellungen wiedergibt, von dem durch das SAS-System generierten Protokollfenster (unterer Bereich).

Abbildung 8.13
Destatis–Anonymeter: Durchführung des Zuordnungsverfahrens I



Zu Programmende werden die Ausgaben wie bereits erwähnt für den Fall einer weiteren Verarbeitung der Daten durch den Nutzer in vier Dateien abgespeichert. Beipielsweise könnte es interessieren, ob eine Abhängigkeit zwischen dem Reidentifikationsrisiko und der Unternehmensgröße oder dem Standort eines Unternehmens besteht.

Die Differenz zwischen realer und CPU-Rechenzeit kann in Einzelfällen beachtlich sein; insbesondere dann, wenn innerhalb eines Netzwerkes zahlreiche Nutzer auf denselben SAS-Server zugreifen müssen.

Abbildung 8.14
Destatis–Anonymeter: Durchführung des Zuordnungsverfahrens II

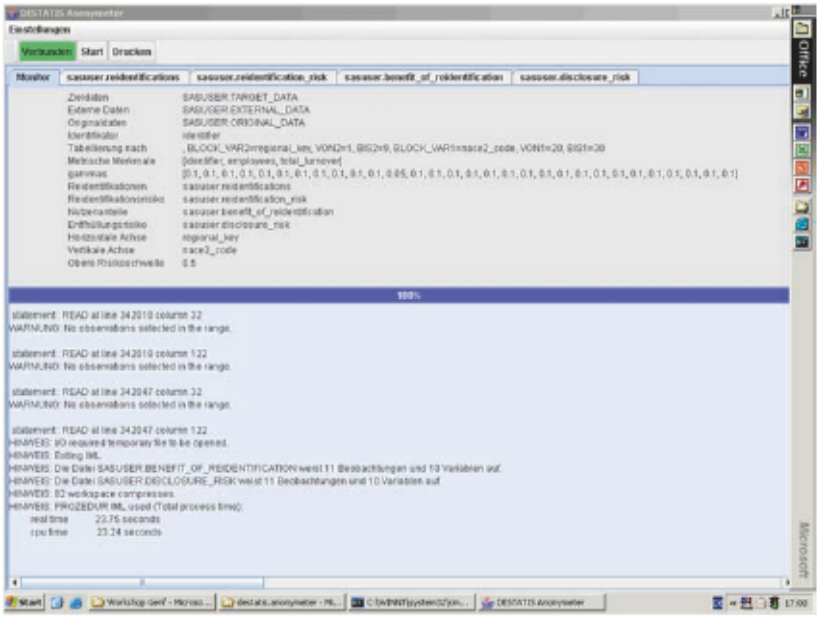
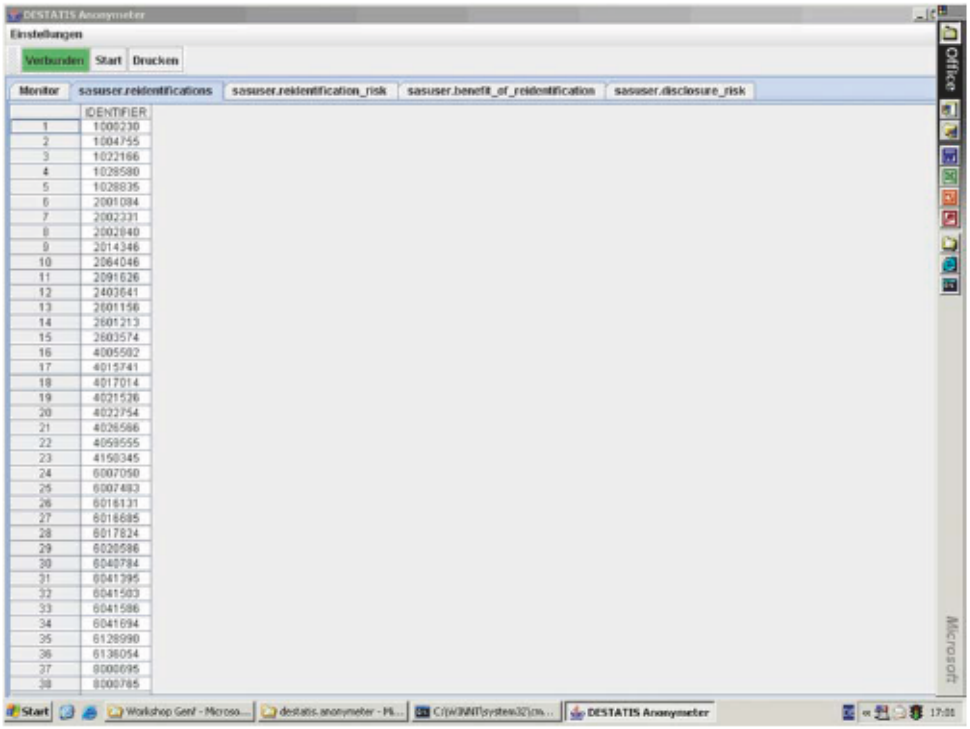


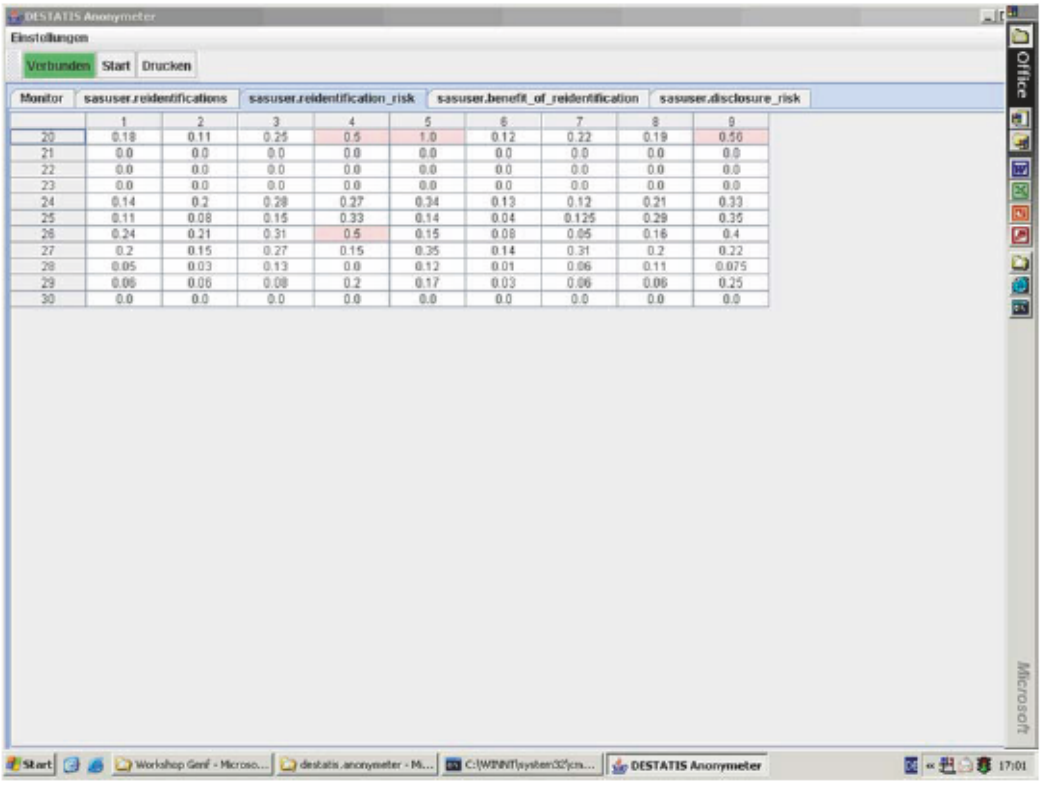
Abbildung 8.15
Destatis–Anonymeter: Ausgabe der Reidentifikationen



Zusätzlich werden die vier Ergebnisdateien auf dem Bildschirm angezeigt (siehe Abbildungen 8.15 bis 8.18). Die jeweilige Datei kann durch Klick auf den Dateinamen auf dem Reiter unterhalb der Menüzeile ausgewählt werden.

Zellen, deren Risiken die durch den Nutzer festgelegte obere Schwelle τ erreichen oder überschreiten, werden in der Ausgabe rot hervorgehoben.

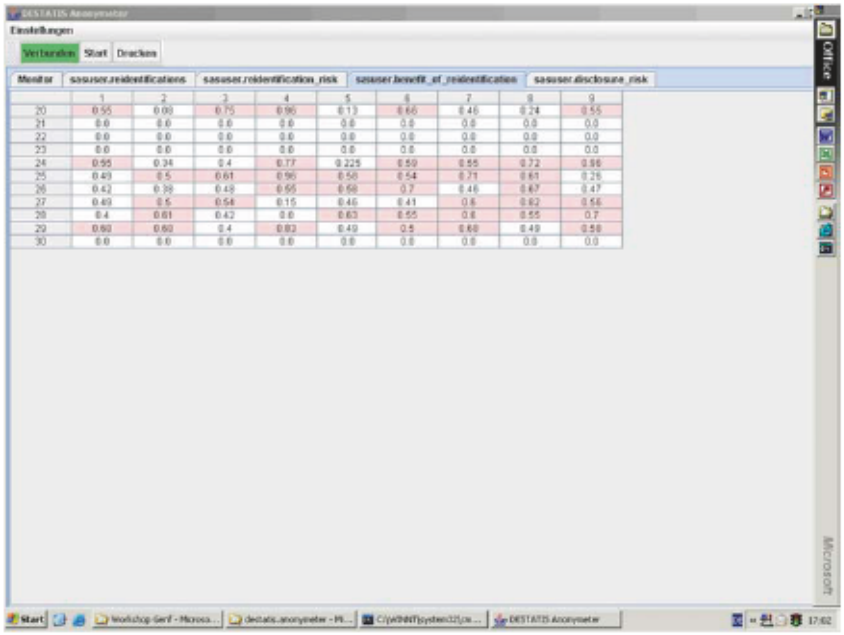
Abbildung 8.16
Destatis–Anonymeter: Ausgabe der Reidentifikationsrisiken



	1	2	3	4	5	6	7	8	9
20	0.18	0.11	0.25	0.5	1.0	0.12	0.22	0.19	0.50
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	0.14	0.2	0.28	0.27	0.34	0.13	0.12	0.21	0.33
25	0.11	0.08	0.15	0.33	0.14	0.04	0.125	0.29	0.35
26	0.24	0.21	0.31	0.5	0.15	0.08	0.05	0.16	0.4
27	0.2	0.15	0.27	0.15	0.35	0.14	0.31	0.2	0.22
28	0.05	0.03	0.13	0.0	0.12	0.01	0.06	0.11	0.075
29	0.06	0.06	0.08	0.2	0.17	0.03	0.06	0.06	0.25
30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

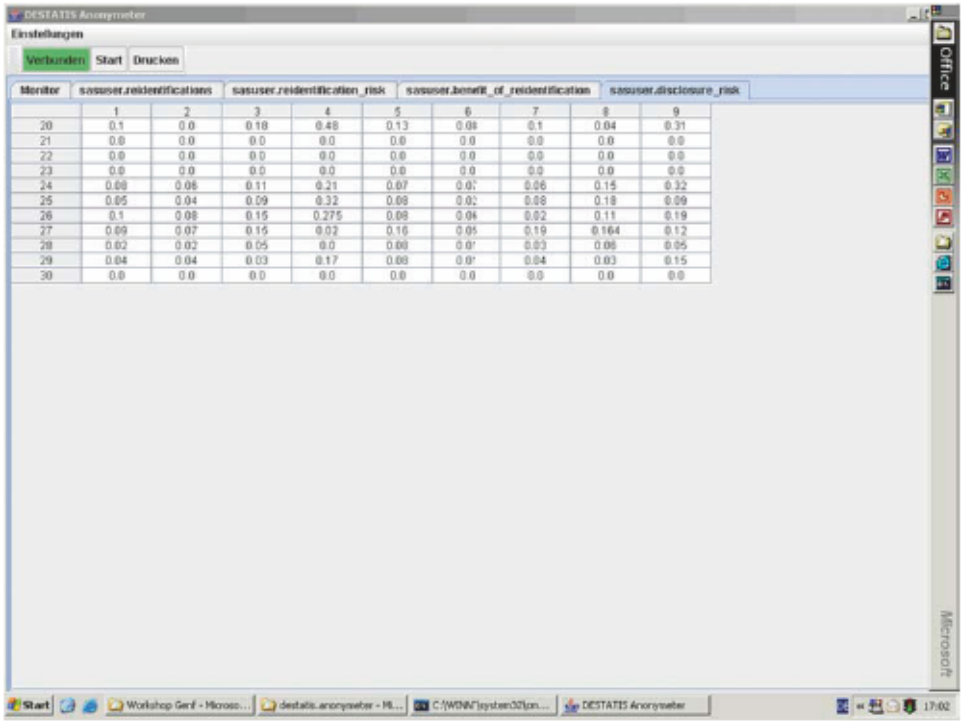
Zum Beispiel weisen in Abbildung 8.16 vier Zellen erhöhte Reidentifikationsrisiken auf, wobei die Anfälligkeit von Unternehmen des Wirtschaftszweiges 20 auffällig ist. Dagegen wird nur eine geringe Abhängigkeit zum regionalen Standort (Grad der Urbanisierung im Neunerschlüssel) beobachtet.

Abbildung 8.17
Destatis–Anonymeter: Ausgabe der Anteile brauchbarer Einzelinformationen



	1	2	3	4	5	6	7	8	9
20	0.55	0.00	0.75	0.90	0.13	0.66	0.46	0.24	0.55
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	0.55	0.34	0.4	0.77	0.225	0.59	0.55	0.72	0.56
25	0.49	0.5	0.61	0.90	0.58	0.54	0.71	0.61	0.25
26	0.42	0.38	0.48	0.65	0.58	0.7	0.46	0.67	0.47
27	0.49	0.5	0.58	0.15	0.46	0.41	0.6	0.62	0.56
28	0.4	0.61	0.42	0.0	0.63	0.55	0.6	0.55	0.7
29	0.60	0.60	0.4	0.82	0.49	0.5	0.60	0.49	0.58
30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Abbildung 8.18
Destatis–Anonymeter: Ausgabe der Enthüllungsrisiken

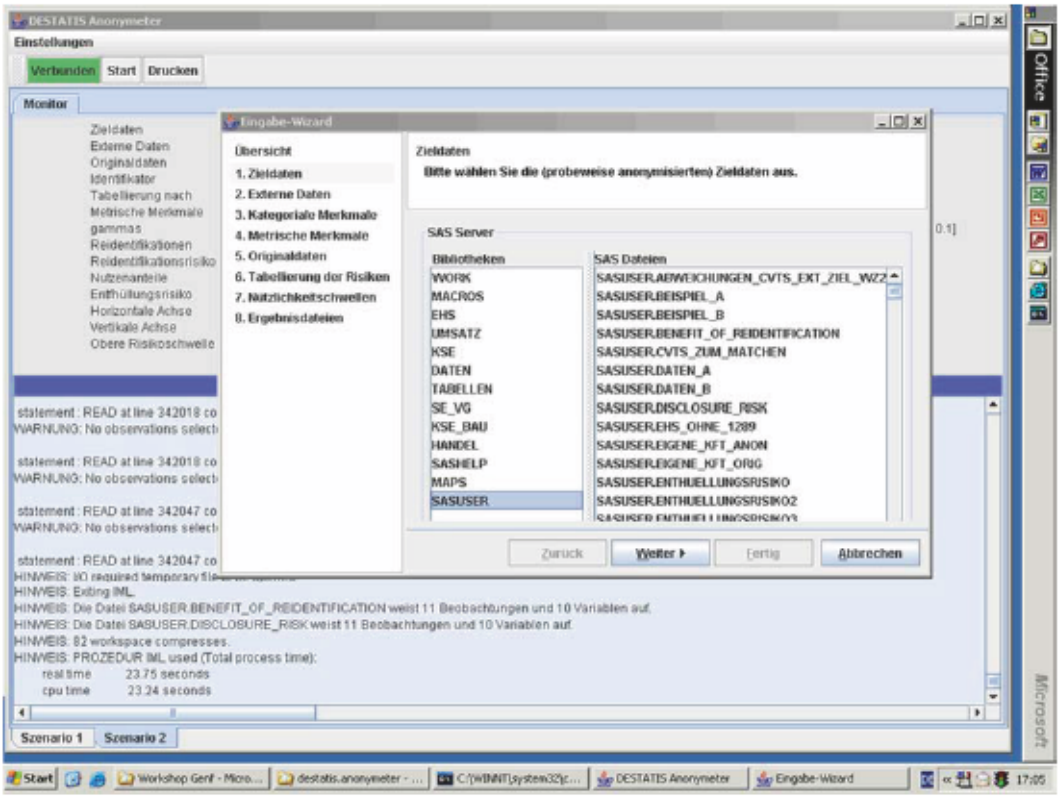


	1	2	3	4	5	6	7	8	9
20	0.1	0.0	0.18	0.48	0.13	0.08	0.1	0.04	0.31
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
22	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
24	0.08	0.08	0.11	0.21	0.07	0.07	0.06	0.15	0.32
25	0.05	0.04	0.09	0.32	0.08	0.02	0.08	0.18	0.09
26	0.1	0.08	0.15	0.275	0.08	0.04	0.02	0.11	0.19
27	0.09	0.07	0.15	0.02	0.16	0.05	0.19	0.164	0.12
28	0.02	0.02	0.05	0.0	0.08	0.0	0.03	0.08	0.05
29	0.04	0.04	0.03	0.17	0.08	0.0	0.04	0.03	0.15
30	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

In Abbildung 8.17 finden sich die zellenweisen Anteile brauchbarer Informationen. Zahlreiche rot unterlegte Zellen belegen hier, dass ein Großteil der veränderten Einzelwerte verhältnismäßig nah an den Originalwerten liegt.

Erfreulicherweise tauchen in Abbildung 8.18 keine rot unterlegten Zellen auf. Durch die zellenweise Produktbildung der Reidentifikationsraten mit den zugehörigen Anteilen brauchbarer Informationen entstehen zellenweise Enthüllungsrisiken, die sämtlich unter $\tau = 0.5$ liegen.[46] Dies bedeutet, dass die Datei als faktisch anonym eingestuft werden kann. Andernfalls müssten weitere Anonymisierungsmaßnahmen (etwa speziell für den Wirtschaftszweig 20) ergriffen und das Datenangriffsszenario wiederholt werden (siehe abschließende Abbildung 8.19).

Abbildung 8.19
Destatis–Anonymeter: Möglichkeit der Wiederholung der Simulation



46 Interpretiert man die Tabellen aus den Abbildungen 8.16 bis 8.18 als Matrizen, so ergibt sich die Matrix der Enthüllungsrisiken aus dem Hadamard-Produkt der beiden anderen Matrizen.

8.5 Auswahl anonymisierter Wirtschaftsstatistiken

In diesem Abschnitt wird eine kleine Auswahl bereits anonymisierter und der Wissenschaft zur Verfügung stehender Unternehmens- und Betriebserhebungen der amtlichen Statistik vorgestellt. Einige der Erhebungen liegen nicht nur im Quer-, sondern durch die im Projekt FAWE-Panel durchgeführten und teilweise sehr aufwendigen Verknüpfungsarbeiten auch im Längsschnitt vor. Dies betrifft die Kostenstrukturerhebung im Verarbeitenden Gewerbe (Unterabschnitt 8.5.1), den mit weiteren Erhebungen verknüpften Monatsbericht im Verarbeitenden Gewerbe (Unterabschnitt 8.5.2) und die Umsatzsteuerstatistik (Unterabschnitt 8.5.3). Detaillierte Ausführungen zu den im Längsschnitt verknüpften Daten finden sich in Brandt et al. (2008b), ein Gesamtüberblick über bislang anonymisierte amtliche Wirtschaftsstatistiken in Lenz und Zwick (2009b). Die aufgeführten Erhebungen sind grundsätzlich über verschiedene Formen des Datenzugangs für die Wissenschaft verfügar. Die Daten sind zugänglich an einem Gastwissenschaftlerarbeitsplatz an einem beliebigen Standort der Forschungsdatenzentren (sogenannter On-Site Zugang), als Scientific-Use-File am eigenen Arbeitsplatz (Off-Site) und via kontrollierte Datenfernverarbeitung. Die verschiedenen Arten des Zugangs zu Einzeldaten der amtlichen Statistik werden in Zühlke et al. (2004) beschrieben. Je nach gewählter Zugangsart muss der Forscher eine mehr oder weniger stark eingreifende Datenveränderung zwecks Erreichung der faktischen Anonymität in Kauf nehmen.

8.5.1 Kostenstrukturerhebung im Bergbau und Verarbeitenden Gewerbe

Die jährlich erhobene Kostenstrukturerhebung im Verarbeitenden Gewerbe, im Bergbau sowie der Gewinnung von Steinen und Erden (KSE) ist für vielfältige Strukturuntersuchungen geeignet und liefert umfassende Informationen über die Produktionsergebnisse, die eingesetzten Produktionsfaktoren sowie über die Wertschöpfung von Unternehmen mit wenigstens 20 Beschäftigten. Sie ist eine nach Wirtschaftszweigen und Beschäftigtengrößenklassen geschichtete Zufallsstichprobe von maximal 18 000 Unternehmen pro Jahr, wobei einige Schichten voll erhoben werden (obere Beschäftigtengrößenklassen und dünn besetzte Wirtschaftszweige). In der Regel wird die Stichprobe der KSE nur alle vier Jahre neu gezogen, um die auskunftspflichtigen Unternehmen zu entlasten. Durch diese Rotation können etwa 12 000 Unternehmen, die in der vorangegangenen Stichprobe erfasst waren, von der Erhebung freigestellt werden. Im Jahr 1995 wurde eine neue Wirtschaftszweigsystematik eingeführt. Durch diese Umstellung gab es viel Änderungsbedarf, sodass bereits nach zwei Jahren (1997) eine neue Stichprobe gezogen wurde. In den Jahren 1998 und 1999 war eine größere Insolvenzwelle zu verzeichnen und im Hinblick auf die Material- und Wareneingangserhebung wurde nach zwei weiteren Jahren (1999) eine neue Stichprobe gezogen. Aufgrund der veränderten Stichpro-

benziehung im Jahr 1997, 1999 und im Jahr 2003 steht die Mehrzahl der Unternehmen des Vorjahres in der Welle nach der Stichprobenziehung nicht mehr zur Verfügung.

Die Bereitstellung von Querschnittsdaten der KSE als Scientific-Use-File war bereits vor Beginn des Projektes FAWE-Panel möglich (siehe Lenz et al. 2005a). Im Projekt wurden die Daten für die Jahre 1995 bis 2005 über die Unternehmensidentifikationsnummer, die einen eindeutigen Identifikator darstellt, verknüpft. Der Längsschnittsdatensatz beinhaltet für die Jahre 1995 bis 2005 gut 43 000 Fälle. Die Form der Aufbereitung bietet sowohl die Möglichkeit für Analysen im Querschnitt des jeweiligen Berichtsjahres als auch im Längsschnitt. Von 1995 bis 2005 existieren gut 2 000 Unternehmen, die jedes Jahr befragt wurden. Von diesen Unternehmen stammt ein großer Teil aus den oben angesprochenen voll erhobenen Bereichen. Innerhalb des Berichtskreises für die Jahre 1999 bis 2002 existieren immerhin knapp 13 300 Unternehmen, die in jedem Jahr befragt wurden und somit ausreichend Potential für wissenschaftliche Analysen bieten.

Ein Beispiel für das Analysepotential und die Forschungsmöglichkeiten der Längsschnittsdaten der KSE wird in Görzig et al. (2005) dargestellt. Mit der Verknüpfung über die Zeit sind Analysen zum Zusammenhang von „Outsourcing" und Unternehmenserfolg möglich. Häufig werden Unternehmen nach Branchen oder Wirtschaftszweigen aggregiert ausgewertet. Mit der Analyse von Mikrodaten der KSE konnten Fritzsch und Stephan (2007) zeigen, dass selbst innerhalb dieser Branchen die Heterogenität der Unternehmen bezüglich ihrer jährlichen Beschäftigtenentwicklung und der jährlichen Ausgaben für Forschung und Entwicklung recht groß ist. Gleiches gilt für die Unternehmensgröße, da es in jeder Beschäftigtengrößenklasse effizient und ineffizient agierende Unternehmen geben kann.

8.5.2 Monatsbericht, Investitions- und Kleinbetriebserhebung im Verarbeitenden Gewerbe

In diesem Abschnitt werden drei Erhebungen aus dem Bereich des Verarbeitenden Gewerbes vorgestellt, die im Quer- und Längsschnitt zusammengespielt werden konnten. Insbesondere durch die „Verbreiterung" im Querschnitt eröffnen sich vielfältige Analysemöglichkeiten, da nunmehr dem interessierten Wissenschaftler sämtliche in den drei Statistiken erhobene Merkmale für Analysen zur Verfügung stehen. Als zeitlicher Ausgangspunkt für die Aufbereitung und spätere Verknüfung der Daten wurde das Jahr 1995 ausgewählt, da in diesem Jahr die oben angesprochene grundlegende Änderung in der Systematik der Wirtschaftszweigklassifikation erfolgte. Zunächst wurden die Daten bis einschließlich 2005 aufbereitet. Zukünftig ist geplant, die Daten weiterer Wellen sukzessiv zu ergänzen. Die drei einbezogenen Erhebungen werden im Folgenden kurz vorgestellt.

Monatsbericht im Verarbeitenden Gewerbe

Berichtspflichtig für den Monatsbericht sind sämtliche im Inland gelegenen Betriebe aus dem genannten Wirtschaftsbereich, die 20 oder mehr Beschäftigte aufweisen. Betriebe mit weniger als 20 Beschäftigten werden nur dann einbezogen, wenn sie zu einem Unternehmen des Verarbeitenden Gewerbes mit wenigstens 20 Beschäftigten gehören. Vereinfacht ausgedrückt stellt der Monatsbericht also eine Vollerhebung mit Abschneidegrenze dar. Die Zuordnung von Betrieben und Unternehmen zum Verarbeitenden Gewerbe erfolgt nach ihrem wirtschaftlichen Schwerpunkt.

Zu den Erhebungsmerkmalen des Monatsberichtes zählen unter anderem die *Anzahl der Beschäftigen* im Betrieb, der *Inlands-* und *Auslandsumsatz*, die gezahlten *Löhne und Gehälter* sowie die geleisteten *Arbeitsstunden*. In Kombination mit verfügbaren Basisinformationen, die Auskunft über den Wirtschaftszweig, den Standort und die Unternehmenszugehörigkeit geben, sind bereits differenzierte Analysen möglich. Im Prinzip wären solche Analysen auch auf Monatsbasis möglich. Da die anderen herangezogenen Erhebungen aus dem Verarbeitenden Gewerbe aber jährliche Erhebungen darstellen, wurde beim Monatsbericht auf die Jahresergebnisse zurückgegriffen, die Jahressummen und Jahresdurchschnittswerte ausweisen.

Investitionserhebung im Verarbeitenden Gewerbe

Die zweite einbezogene Erhebung ist die jährliche Investitionserhebung bei Betrieben im Bereich des Verarbeitenden Gewerbes, die im Wesentlichen den gleichen Berichtskreis wie der Monatsbericht abdeckt. Die Daten stellen eine wertvolle Ergänzung dar, da die im Verlauf eines Jahres getätigten Investitionen für Untersuchungen auf Betriebsebene häufig von Interesse sind. Verfügbar ist unter anderem der Wert der erworbenen und selbst erstellten Sachanlagen (Bruttoanlageinvestitionen), differenziert nach verschiedenen Bereichen. In den Daten finden sich außerdem wie beim Monatsbericht auch elementare Informationen wie Wirtschaftszweigzugehörigkeit und Standort des Betriebes.

Industrielle Kleinbetriebserhebung im Verarbeitenden Gewerbe

Bei der industriellen Kleinbetriebserhebung, der dritten einbezogenen Erhebung, wurden bis 2002 einmal jährlich Industriebetriebe mit Schwerpunkt im Bereich des Verarbeitenden Gewerbes befragt, die aufgrund ihrer geringen Beschäftigtenanzahl nicht monatsberichtspflichtig waren. Erhoben wurde die *Anzahl der Beschäftigten* im Betrieb zum Zeitpunkt Ende September, der *Gesamtumsatz* zum selben Zeitpunkt sowie der *Gesamtumsatz des Vorjahres*.

Die Daten ermöglichen eine Betrachtung kleiner Betriebe im Quer- und im Längsschnitt. Durch Kombination mit den in den Monatsberichten erfassten Betrieben liegt ein Bestand vor, der alle Betriebsgrößenklassen aus dem Bereich des Verarbeitenden Gewerbes abdeckt. Nicht zuletzt ist durch Informationen der Kleinbetriebserhebung die Möglichkeit einer Identifikation solcher Betriebe gegeben, die im Zeitverlauf die Erfassungsgrenze des Monatsberichts unter- oder überschritten haben, was für betriebsdemografische Analysen relevant sein kann (Brandt et al. 2008b).

8.5.3 Umsatzsteuerstatistik

Zur Beurteilung der Struktur und Wirkungsweise der Steuern hat der Gesetzgeber Bundesstatistiken über die wichtigsten Steuern angeordnet. Bei den Steuerstatistiken werden Daten der Finanzverwaltung ausgewertet, die im Rahmen des Besteuerungsverfahrens anfallen. Die Umsatzsteuerstatistik bildet die Umsätze und die steuerlichen Merkmale der Unternehmen ab, die im Berichtsjahr Umsatzsteuer-Voranmeldungen abgegeben haben und deren Umsatz über der Grenze für Kleinunternehmer gelegen hat. Nicht erfasst werden Unternehmen, die ausschließlich steuerfreie Umsätze tätigen bzw. bei denen keine Steuerzahllast entsteht (z.B. niedergelassene Ärzte und Zahnärzte ohne Labor, Behörden, Versicherungsvertreter und landwirtschaftliche Unternehmen).

Aus der Beobachtung der Umsätze ergeben sich wertvolle Informationen für die Haushaltsplanungen und Steuerschätzungen des Bundes, der Länder und der Gemeinden. Die Umsatzsteuerstatistik ist nicht allein ein Instrument der Fiskal- und Steuerpolitik; sie dient darüber hinaus auch der allgemeinen Wirtschaftsbeobachtung. Mit ihren Angaben über die Entwicklung der Umsätze in allen Bereichen der Volkswirtschaft liefert sie Informationen, die in dieser Vollständigkeit in keiner anderen Bundesstatistik enthalten sind. Die Ergebnisse der Umsatzsteuerstatistik sind eine wichtige Datenbasis für die Erstellung der Volkswirtschaftlichen Gesamtrechnungen. Aufgrund ihrer tiefen wirtschaftszweigsystematischen Gliederung lassen sich mithilfe der Umsatzsteuerstatistik auch branchenspezifische Analysen sowie Konzentrationsuntersuchungen durchführen (siehe Vorgrimler et al. 2005b). Die Umsatzsteuerstatistik ist eine sogenannte Sekundärstatistik, da sie auf Daten zurückgreift, die woanders, nämlich bei der Finanzverwaltung anfallen. Erfasst werden wie oben erwähnt alle Unternehmen, die Umsatzsteuer-Voranmeldungen abgeben, mit einem Jahresumsatz (ohne Umsatzsteuer) von über 32 500 DM (16 617 Euro). In der Erhebung des Jahres 2000 werden rund 2,9 Mill. Unternehmen erfasst.

Bereits vor Beginn des Projektes FAWE-Panel war eine Bereitstellung der Daten der Umsatzsteuerstatistik im Querschnitt als Scientific-Use-File möglich (Vorgrimler et al. 2005a; Sturm und Lenz 2005). Im Rahmen des Projektes wurden die Erhebungen der Jahre 2001 bis 2005 zu einem Panel zusammengeführt. Das Panel enthält insgesamt 19 umsatzsteuerli-

che Merkmale wie beispielsweise die Lieferungen und Leistungen insgesamt (Umsatz) sowie die zu 16% und 7% versteuerten Umsätze, die steuerfreien Lieferungen und Leistungen sowie das Vorauszahlungssoll. Daneben sind wie in den jährlichen Erhebungen unter anderem Angaben zu Wirtschaftszweig, Rechtsform und zum Unternehmenssitz der Steuerpflichtigen enthalten. Erstmals wurden im Rahmen des Projektes Angaben zu den sozialversicherungspflichtig Beschäftigten aus dem amtlichen Unternehmensregister hinzugefügt. Das Panel der Umsatzsteuerstatistik 2001-2005 enthält insgesamt 4,3 Mill. Merkmalsträger. Davon enthalten beachtliche 1,9 Mill. Fälle (etwa 43%) Umsatzangaben über alle fünf Berichtszeiträume, 421 000 Fälle über vier Jahre (10%), 483 000 (11%) über drei Jahre, 642 000 (15%) über zwei Jahre und 909 000 (21%) für jeweils ein Jahr. Dabei weisen die erste und letzte Welle aufgrund der fehlenden Vor- bzw. Folgewelle mit 310 000 bzw. 370 000 Fällen erwartungsgemäß deutlich mehr „Einmalbeobachtungen" auf als die Zwischenjahre.

Mit dem beschriebenen Panel wird den Datennutzern ein Produkt zur Verfügung gestellt, dass es bisher in dieser Form nicht gegeben hat. Ergebnisse von Auswertungen der Umsatzsteuerstatistik im Längsschnitt liegen daher bislang noch nicht vor. Da die Umsatzsteuerstatistik bis auf die beschriebenen Ausnahmen nahezu eine Vollerhebung aller umsatzsteuerpflichtigen Unternehmen darstellt, sind zahlreiche Auswertungsfelder denkbar. So sind in Kombination mit den Regional- und Wirtschaftszweigangaben Untersuchungen zur Umsatzentwicklung der Unternehmen, zur Gründungs- und Schließungsdynamik (Zu- und Abgänge im Panel) oder auch zum Einfluss der Exporttätigkeit auf die Umsatzentwicklung denkbar.

8.5.4 Einzelhandelsstatistik

Die Ergebnisse der Jahreserhebung im Einzelhandel liefern wirtschaftspolitisch bedeutsame Informationen über die Struktur, Rentabilität und Produktivität der im Einzelhandel tätigen Unternehmen. Neben der Ermittlung des Rohertrages und der Bruttowertschöpfung sind qualitativ hochwertige Schätzungen für die Vorratsveränderungen in der Wirtschaft möglich. Aus konjunkturpolitischer Sicht können mit den Ergebnissen von Jahreserhebungen aufeinander folgender Jahre sowohl die Beschäftigtensituation als auch die Lohn- und Gehaltsstrukturen beobachtet und analysiert werden. Auf der betriebswirtschaftlichen Ebene lässt sich die Entwicklung von Arbeitsintensität und -produktivität überprüfen. Weiterhin sind die Investitionen und ihre Veränderungen wichtiger Indikator für die längerfristige Umsatzerwartung eines Unternehmens.

Die jährliche Einzelhandelsstatistik erfasst für das Jahr 1999 etwa 23 500 Unternehmen in der Stichprobe. Diese repräsentieren ca. 300 000 Unternehmen des Einzelhandels mit einem Umsatz von rund 300 Mrd. Euro. Betrachtet man nur die Monatsmelder, so repräsentieren etwa 14 500 Unternehmen in der Stichprobe über 110 000 Unternehmen der Einzelhandelsstatistik, die einen Umsatz von rund

285 Mrd. Euro tätigen. Aufgrund der relativ hohen Anzahl von Unternehmen mit weniger als 50 Beschäftigten konnte für diese ein Scientific-Use-File nahezu ohne den Einsatz datenverändernder Verfahren erreicht werden. Dagegen wurde für die Unternehmen mit 50 und mehr Beschäftigten bislang noch keine Anonymisierungsmethode gefunden, die sowohl ausreichende Vertraulichkeit als auch hinreichend gutes Analysepotential gewährleistet. Dies liegt insbesondere an den niedrigen Besetzungszahlen der oberen Unternehmensgrößenklassen. Eine Analyse der vollständigen Erhebung ist jedoch am Gastwissenschaftlerarbeitsplatz und mittels kontrollierter Datenfernverarbeitung möglich.

Das Scientific-Use-File für das Jahr 1999 enthält etwa 12 600 Unternehmen, welche 97,6% der Grundgesamtheit aller Unternehmen mit einem Mindestjahresumsatz von 250 000 Euro repräsentieren und zu gut einem Drittel zum Gesamtumsatz im Einzelhandel beitragen. Erfreulicherweise konnte bei der Generierung des Scientific-Use-Files nahezu auf datenverändernde Verfahren verzichtet werden. Begünstigt wurde dies insbesondere durch die große Anzahl an Unternehmen mit weniger als 50 Beschäftigten. Der Datennutzer bekommt hier die Möglichkeit, mit unverfälschten Daten an dem von ihm bevorzugten Ort arbeiten zu können.

Eine Auflistung aller Merkmale der Einzelhandelsstatistik und der zur Erzeugung des Scientific-Use-Files ausgewählten Anonymisierungsmethoden findet sich in Scheffler (2005).

8.5.5 Gehalts- und Lohnstrukturerhebung

Die Gehalts- und Lohnstrukturerhebung wird von den Statistischen Ämtern des Bundes und der Länder seit 1951 durchgeführt. Nachdem die Erhebung in der Vergangenheit in unregelmäßigen Abständen mit einer großen Lücke zwischen 1978 und 1990 stattfand, sollen die Daten zukünftig vierjährlich erhoben werden. Auf Grundlage einer Verordnung der Europäischen Gemeinschaft von 1999 findet die Erhebung in allen EU-Ländern statt. Somit liegen europaweit vergleichbare Daten vor. Wegen des europäischen Kontextes wird die Erhebung zumeist mit SES (Structure of Earnings Survey) abgekürzt.

Die Angaben zu Arbeitszeit und Verdienst beziehen sich immer auf den Berichtsmonat Oktober. Für Deutschland lieferten 2001 gut 22 000 Betriebe Angaben zu über 846 000 Beschäftigten. Zum Berichtskreis gehören Betriebe des Verarbeitenden Gewerbes und ausgewählte Teile des Dienstleistungsbereiches. Neben Handel und Kredit- und Versicherungsgewerbe gibt es seit 2001 Daten für Gastgewerbe, Verkehr und Nachrichtenübermittlung, Grundstücks- und Wohnungswesen, Vermietung beweglicher Sachen und Erbringung von Dienstleistungen überwiegend für Unternehmen. Auch bei den Beschäftigten erfolgte eine Ausweitung gegenüber früheren Erhebungen: Auszubildende, geringfügig Beschäftigte und Arbeitnehmer in Altersteilzeit sind erstmals 2001 erfasst worden.

Die Statistik enthält Informationen zur Person (Geschlecht, Alter, Ausbildung, Steuerklasse, Kinderfreibeträge), zur Tätigkeit (Berufsschlüssel der Sozialversicherung, Stellung im Beruf, Leistungsgruppe, Arbeitszeit, Dauer der Betriebszugehörigkeit) und zum Verdienst (Brutto, Netto, Zulagen für Schicht-/Nachtarbeit, Sonderzahlungen, Lohnsteuer, Sozialabgaben). Auf Betriebsebene gibt es zusätzlich Angaben darüber, ob die öffentliche Hand beteiligt ist, ob der Betrieb in der Handwerksrolle eingetragen ist und ob Tarifverträge gelten sowie Angaben zur Anzahl der Beschäftigten, differenziert nach Geschlecht sowie nach Arbeitern und Angestellten.

Eine Übersicht der bislang mit den Daten durchgeführten Analysen sowie weitergehende Analysemöglichkeiten finden sich in Hafner und Lenz (2008).

In den Forschungsdatenzentren des Statistischen Bundesamtes und der Statistischen Ämter der Länder stehen die Einzeldaten des SES 2006, des SES 2001 und des SES 1995 für ganz Deutschland zur Nutzung an den Gastwissenschaftlerarbeitsplätzen an allen Standorten der Forschungsdatenzentren des Bundes und der Länder und mittels kontrollierter Datenfernverarbeitung zur Verfügung. Darüber hinaus sind die Daten auch als Scientific-Use-Files zur Verwendung am eigenen Arbeitsplatz verfügbar. Weitere Informationen zu Methodik und Merkmalen des SES 2001 findet man in Frank-Bosch (2003), Details zur Anonymisierung in Hafner et al. (2007).

Die Anonymisierung der SES-Daten stellt für die Datenanbieter eine besondere Herausforderung dar, da ein Gleichgewicht zwischen dem Datenschutz einerseits und dem Erhalt des Analysepotentials andererseits schwieriger erreichbar ist als bei Daten, die nur Informationen entweder über Unternehmen oder über Beschäftigte enthalten. Hierzu wird in Lenz und Hafner (2006a) ein besonderes Schutzwirkungskonzept vorgestellt.

8.5.6 Daten zur betrieblichen Weiterbildung

Seit Ende 2005 können Wissenschaftler Daten zur beruflichen Weiterbildung in Unternehmen für eigene Analysen nutzen. In einem Kooperationsprojekt zwischen dem Statistischen Bundesamt und dem Hessischen Statistischen Landesamt konnten die Einzeldaten der Zweiten Europäischen Erhebung über die berufliche Weiterbildung (CVTS2, Second Continuing Vocational Training Survey) aus dem Jahre 2000 mit Berichtsjahr 1999 faktisch anonymisiert werden (Lenz et al. 2006a).

In der Erhebung liegen Angaben von 3 184 deutschen Unternehmen mit 10 und mehr Beschäftigten aus den Wirtschaftszweigen C – K und O zur Teilnahme von Mitarbeitern an Maßnahmen zur beruflichen Weiterbildung im Jahre 1999 vor. Die Daten enthalten Informationen zum Angebot von verschiedenen Formen beruflicher Weiterbildung, zu Teilnehmern an Lehrveranstaltungen, Teilnahmestunden in Lehrveranstaltungen und Kosten für Lehrveranstaltungen sowie qualitative Angaben zur Weiterbildungskonzeption und zum Stellenwert der Weiterbildung in Unternehmen. Bei der Anonymisierung ist es unter anderem gelungen,

eine wissenschaftliche Behandlung relevanter Fragestellungen nach Wirtschaftsbereichen und Beschäftigtengrößenklassen zu ermöglichen. Ausführliche Informationen zum Basismaterial sowie eine vollständige Merkmalsliste finden sich in Egner (2002), Informationen zu den europäischen Daten in Eurostat (2002) und Nestler und Kailis (2002).

Aufgrund des kleinen Stichprobenauswahlsatzes stellte die Anonymisierung eine verhältnismäßig leichte Aufgabe dar, verglichen beispielsweise mit der Anonymisierung des im vorherigen Unterabschnitt beschriebenen SES. Hier bestand die größere Herausforderung in der Harmonisierung der Anonymisierungsmaßnahmen mit der Forderung der Vergleichbarkeit von Analysen mit den parallel in den anderen europäischen Ländern erhobenen Daten.

Die Daten wurden rechtzeitig bereitgestellt, um im Rahmen des durch den Rat für Sozial- und Wirtschaftsdaten ausgerufenen Expertisenwettbewerbs eine erste Verwendung zu finden. Auf der Internetseite *www.ratswd.de* sind die 16 prämierten Forschungsarbeiten verfügbar. Weitere Forschungspotentiale werden in Lenz und Hafner (2006) aufgezeigt.

Mit dem demografischen Wandel und der zunehmenden Alterung der Bevölkerung ist eine Veränderung der bisherigen Rentenpraxis in der Bundesrepublik Deutschland unumgänglich. Um eine weitere Belastung Jüngerer und Rentenkürzungen zu vermeiden, wird eine Erhöhung des Renteneintrittsalters und somit eine Verlängerung der Lebensarbeitszeit notwendig. Durch die Forderung nach einem längeren Verbleib im Erwerbsleben werden die Ansprüche an ältere Arbeitnehmer steigen. Um den Anforderungen dieses immer dynamischeren Arbeitsmarktes gerecht zu werden, wird lebenslanges Lernen unerlässlich, und die berufliche Weiterbildung gewinnt für den Verbleib im Erwerbsleben an Bedeutung. Mit der Transformation zur sogenannten Wissensgesellschaft wird in Zukunft ein großer Bedarf an gut ausgebildeten und hoch qualifizierten Personen bestehen. In diesem Zusammenhang wird nicht nur für die Arbeitnehmer und Unternehmen das Thema der beruflichen Weiterbildung wichtiger, sondern auch in der Politik und Wissenschaft wächst die Nachfrage nach Daten zur beruflichen Weiterbildung (siehe Brandt et al. 2006).

Ende des Jahres 2008 konnte die entwickelte Anonymisierungsstrategie auch auf die Anonymisierung der Folgeerhebung CVTS3 übertragen werden, die in allen 27 Mitgliedstaaten der Europäischen Union und Norwegen im Jahre 2007 durchgeführt wurde, wobei sich die erhobenen Informationen auf das Jahr 2006 beziehen. Detaillierte Informationen zur CVTS3 finden sich in Statistisches Bundesamt (2007a), Statistisches Bundesamt (2007b) und Schmidt (2007).

Kapitel 9

Zusammenfassung und Ausblick

On ne finit pas un oeuvre, on l'abandonne *(Gustave Flaubert)*

In den letzten Jahren hat sich das Datenangebot der amtlichen Statistik für empirisch arbeitende Wirtschafts- und Sozialwissenschaftler deutlich verbessert. Die Grundlagen zur Bereitstellung von Unternehmens- und Betriebsdaten wurden in den Projekten „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" *(FAWE)* und „Wirtschaftsstatistische Paneldaten und faktische Anonymisierung" *(FAWE-Panel)* gelegt. Seitdem stehen der Wissenschaft grundsätzlich sämtliche wirtschaftsstatistischen Erhebungen der amtlichen Statistik sowohl im Quer- als auch im Längsschnitt über verschiedene Formen des Datenzugangs zur Verfügung. In der vorliegenden Arbeit wurde der Weg hin zu einer Operationalisierung des durch den Gesetzgeber formulierten Begriffes der faktischen Anonymität und zu dessen Anwendung auf reale Wirtschaftsdaten aufgezeigt. Hierzu dienten und dienen die in der vorliegenden Arbeit mathematisch modellierten und rechentechnisch umgesetzten Datenangriffsszenarien.

Bei den Simulationen von Datenangriffsszenarien mit dem Ziele der faktischen Anonymisierung wirtschaftsstatistischer Erhebungen ist wesentlich, dass stets realistische Rahmenbedingungen angenommen werden und anschließend auf Basis „realer" Datenangriffsszenarien über die faktische Anonymität einer Datei entschieden wird. Um einen interessierten Wissenschaftler, der über seine arbeitgebende wissenschaftliche Einrichtung einen entsprechenden Nutzungsantrag gestellt hat, möglichst schnell mit Einzeldaten versorgen zu können, mussten effiziente Programme zur zeitnahen Feststellung der faktischen Anonymität der angeforderten Daten entwickelt werden. Im Laufe der letzten Jahre haben zudem aufgrund der großen Nachfrage nach amtlichen Einzeldaten (sowohl im Unternehmens- und Betriebsbereich als auch im Bereich der Personen- und Haushaltserhebungen) die Forschungsdaten-

zentren des Bundes und der Länder[47] als Schnittstelle zwischen Wissenschaft und amtlicher Statistik viel Erfahrung sammeln können. Dies zeigt sich zum einen in der schnellen Entwicklung geeigneter auf die jeweilige Datennutzung angepasster Anonymisierungsmaßnahmen und zum anderen in einer deutlichen Beschleunigung des üblichen Abstimmungsprozesses zwischen Bund und Ländern vor der Weitergabe des Datenmaterials.

Die entwickelten Programme zur Messung der Datensicherheit einer zuvor probeweise anonymisierten Datei haben nachhaltig zu einer Verbesserung des Datenangebotes einerseits und zu einer effizienteren Arbeitsweise an der Schnittstelle zwischen Wissenschaft und amtlicher Statistik andererseits beigetragen. Allerdings haben sich die Forschungsdatenzentren in Zusammenarbeit mit verschiedenen Kooperationspartnern vorgenommen, beide Ziele (Verbesserung des Datenangebotes und Verbesserung der Schnittstelle) auch im Hinblick zukünftiger Nutzerinteressen weiter zu verfolgen.

Eine qualitative Verbesserung des Datenangebotes wird derzeit durch die Zusammenführung bestehender Datenbestände erreicht. Im Projekt *AFiD* (Amtliche Firmendaten für Deutschland; siehe Malchin und Pohl 2007 und Hafner 2009) konnten kürzlich bereits verschiedene Unternehmens- und Betriebsdaten der amtlichen Statistik erfolgreich zusammengespielt werden. Hierzu war die Unternehmensidentifikationsnummer des amtlichen Unternehmensregisters (oftmals abgekürzt mit URS) sehr dienlich. Des Weiteren werden in diesem Projekt Einzeldaten aus den Umweltstatistiken im Quer- und Längsschnitt integriert.

Bedauerlicherweise ist es technisch schwierig und in Deutschland rechtlich selten möglich, Erhebungen verschiedener Datenproduzenten – mangels einer einheitlichen Gesetzgebung – zu verknüpfen. In der Regel werden Erhebungen verschiedener Datenproduzenten auf Grundlage unterschiedlicher Bestimmungen oder Verordnungen erhoben. Insbesondere ist es nicht erlaubt, Unternehmens- und Betriebsdaten der Bundesagentur für Arbeit, der Deutschen Bundesbank und der Statistischen Ämter des Bundes und der Länder zusammenzuspielen. Bisher ist es den Datenanbietern nur möglich, die von ihrer Seite gesammelten Daten in den institutionseigenen Forschungsdatenzentren der Wissenschaft anzubieten. In der durch die genannten Institutionen in Kooperation mit der Fachhochschule Mainz und der Leuphana Universität Lüneburg durchgeführten ambitionierten Machbarkeitsstudie *KombiFiD* (Kombinierte Firmendaten für Deutschland; siehe Bender et al. 2007 und Konold und L'Assainato 2009) wurden hierzu ausgewählte Unternehmen angeschrieben und um ihre schriftliche Einwilligung zur Datenzusammenführung gebeten. Derzeit laufen die Auswertungen des überaus vielversprechenden Rücklaufs.

„Machbarkeit" bezieht sich im Projekt *KombiFiD* nicht nur auf die rechtlichen und technischen Möglichkeiten der Zusammenführung, sondern auch auf die Frage, ob sich mit einer solchen Zusammenführung sinnvolle Analysemöglichkeiten eröffnen. Nach Projektende sol-

47 Detaillierte Informationen zur Arbeit in den Forschungsdatenzentren und den verschiedenen Zugangsformen zu (Einzel-)Daten der amtlichen Statistik finden sich in Zühlke et al. (2004).

len Empfehlungen für eine dauerhafte Zusammenspielung von Daten über die Grenzen der Datenanbieter hinweg abgegeben werden. Dies würde allerdings eine Gesetzesänderung erfordern. Eine schöne Übersicht über das derzeitige Datenangebot an Unternehmens- und Betriebsdaten der amtlichen Statistik und der Bundesagentur für Arbeit wird in Brandt et al. (2008c) gegeben.

Eine Erleichterung des Zugangs zu amtlichen Einzeldaten und eine gleichzeitige Verbesserung der Qualität des für den Wissenschaftler verfügbaren Datenmaterials soll mit dem kürzlich gestarteten Forschungsprojekt „Eine informationelle Infrastruktur für das ‚E-Science Age' – Verbesserung der Angebote der kontrollierten Datenfernverarbeitung durch neue ‚Datenstrukturfiles' und automatisierte Ergebniskontrolle *(infinitE)*" methodisch vorbereitet werden (Brandt und Zwick (2009); Lenz 2009; Lenz und Zwick 2009a). In der Vergangenheit hat sich gezeigt, dass in den Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder die kontrollierte Datenfernverarbeitung mittlerweile die am meisten genutzte Zugangsform zu wirtschaftsstatistischen Einzeldaten darstellt. Ziel des Projektes *infinitE* ist es daher, diese Zugangsform zu verbessern und die theoretischen Grundlagen für eine voll automatisierte Datenfernverarbeitung zu schaffen. Um diese Art von Datenzugang für die unabhängige wissenschaftliche Forschung zu gewährleisten, müssen jedoch eine Reihe methodischer, technischer und rechtlicher Fragen geklärt werden. Der Schwerpunkt des Projektes liegt in der Bearbeitung der methodischen und rechtlichen Herausforderungen der automatisierten Datenfernverarbeitung.[48] Dem Forscher wäre es danach möglich, an seinem eigenen Arbeitsplatz entwickelte komplexe Programme oder formulierte einfache Datenabfragen an einen geschützten Rechner der amtlichen Statistik zu senden und in „Echtzeit" die entsprechenden Programmausgaben oder Abfrageergebnisse zurück zu erhalten. Die Programmausgaben können dabei sowohl aus Tabellen als auch aus Punktschätzern bestehen.

Eine besondere methodische Herausforderung des Projektes *infinitE* besteht in der Automatisierung der Datensicherheitsprüfungen, die bislang im Zuge der kontrollierten Datenfernverarbeitung nur manuell erfolgen und damit sehr zeitaufwendig sind, was Personalkapazitäten in den Forschungsdatenzentren bindet und für die Wissenschaftler unangenehme Wartezeiten verursacht. Neben der angesprochenen Zeitreduktion besteht ein weiterer Vorteil des Datenzuganges via automatisierte Datenfernverarbeitung darin, dass die Vernetzung der Wissenschaftler untereinander und die wissenschaftliche Transparenz gefördert werden, da die mit den Daten arbeitenden Forscher ihre Ergebnisse untereinander vergleichen oder replizieren können.

Begleitend zum Datenzugang des automatisierten Fernrechnens wird der empirisch arbeitende Wissenschaftler von morgen mit einer Datei versorgt, die dieselbe Struktur wie die Originaldaten und gleichzeitig ein Mindestmaß an Analysepotential aufweist. Daher sollen im Projekt *infinitE* auch Strategien zur Erstellung dieser sogenannten Datenstrukturfiles

48 Rein technisch wäre es heute bereits möglich, das automatisierte Fernrechnen als eine von Zeit und Ort flexible Bearbeitung der Daten seitens der Wissenschaftler zu ermöglichen.

entwickelt werden; d.h. von Daten, die es dem Wissenschaftler vor der Absendung seiner Analyseprogramme erlauben, ein Programm auf syntaktische und semantische Fehler zu überprüfen. Jetzige Datenstrukturfiles, welche die Datennutzer der kontrollierten Datenfernverarbeitung erhalten, erlauben nur eine syntaktische Überprüfung. Das Material enthält dabei dieselben Merkmale wie die Originaldaten, bestenfalls mit Erhalt des originalen Wertebereiches, aber ohne Berücksichtigung des Erhaltes einfacher deskriptiver Größen oder gar multivariater Zusammenhänge. Der Entwicklung höherwertiger Datenstrukturfiles könnten neben den in Unterabschnitt 1.2.4 vorgestellten kombinierten Verfahren nach Ansicht des Autors spezielle Varianten der multiplikativen Zufallsüberlagerung (Höhne 2008) und geeignete Modelle der partiellen und/oder multiplen Imputation (Reiter und Drechsler 2007) dienlich sein.

Anhang A

Metadaten zum Scientific-Use-File der Kostenstrukturerhebung im Verarbeitenden Gewerbe

Die folgenden Ausführungen geben einen Einblick in die typische Metadatenstruktur zu einem Scientific-Use-File. Hierzu wird das Beispiel der Kostenstrukturerhebung im Verarbeitenden Gewerbe mit dem Jahr 1999 als Berichtszeitraum vorgestellt.

Das grundsätzliche Vorgehen bei der Anonymisierung dieser Daten wurde bereits in Abschnitt 4.1 dargestellt. In den Metadaten werden insbesondere die Anonymisierungsmaßnahmen, die letztendlich zu einem Scientific-Use-File geführt haben, beschrieben.

Die Daten enthalten Informationen zu Umsatz (3 Merkmale), Anzahl der Beschäftigten (4 Merkmale), Lagerhaltung (6 Merkmale), Materialeinsatz (3 Merkmale), Lohnkosten (6 Merkmale), sonstigen Kosten (6 Merkmale), Wertschöpfung (2 Merkmale) und Forschung und Entwicklung (2 Merkmale) von über 13 000 Unternehmen, die zwischen 20 und 249 Mitarbeiter beschäftigen. Zusätzlich lassen sich die Unternehmen noch nach ihrem Hauptstandort (neue oder alte Bundesländer) unterscheiden. Insgesamt sind in der Datei 33 Merkmale enthalten.

Die Daten sind als Textdatei gespeichert, wobei Einleseroutinen für SAS und SPSS den Einstieg erleichtern. Die faktisch anonymisierte Kostenstrukturerhebung im Verarbeitenden Gewerbe 1999 kann für wissenschaftliche Fragestellungen über die Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder bezogen werden. Anträge zur Nutzung der Daten werden dort entgegengenommen. Weitere Angaben zum Datenbezug, Antragsformulare und zusätzliche Kontaktadressen finden sich im Internet unter

www.forschungsdatenzentrum.de.

Die Daten werden dem Wissenschaftler auf einer CD-Rom geliefert, deren Verzeichnis sich in die folgenden vier Teilbereiche untergliedert. In diesem Kapitel werden nur die in der Gliederung *kursiv* hervorgehobenen Abschnitte abgebildet.

Der erste Teil „Konzeption und Realisierung" beinhaltet die Beschreibung der Anonymisierungsmethoden.

1. Konzeption und Realisierung

   1.1 *Anonymisierungsbeschreibung*

Der zweite Teil „Metadaten" umfasst alle beschreibenden Informationen wie z.B. den Fragebogen, Erläuterungen und Randauszählungen.

2. Metadaten

   2.1 *Fragebogen zur Kostenstrukturerhebung*

   2.2 *Erläuterungen zur Kostenstrukturerhebung*

   2.3 *Beschreibung der Kostenstrukturerhebung*

   2.4 Literaturverzeichnis

   2.5 *Randauszählungen*

Der dritte Teil „Daten" enthält Einleseprozeduren in SPSS und SAS sowie die Datei der Kostenstrukturerhebung im Textformat. Die in Teil 3 befindlichen Dateien können über den Explorer geöffnet werden.

3. Daten

   3.1 Setup_SAS

   3.2 Setup_SPSS

   3.3 Textdatei zur Kostenstrukturerhebung

Der vierte Teil „Kontakt" beinhaltet alle Ansprechpartner zu dem faktisch anonymisierten Mikrodatenfile der Kostenstrukturerhebung im Verarbeitenden Gewerbe 1999.

4. Kontakt

   4.1 Ansprechpartner zu den Daten der Kostenstrukturerhebung

A.1 Anonymisierungsbeschreibung

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Anonymisierungsbeschreibung

1. Einstimmung

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

Durch die Arbeiten des Projektes „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" kann nun eine erste faktisch anonyme Datei für die Wissenschaft (ein so genannter Scientific-Use-File), generiert aus den Daten der Kostenstrukturerhebung im Verarbeitenden Gewerbe, angeboten werden.[2] Auf dem Weg dahin war ein klassischer Zielkonflikt zu lösen: Sicherstellung der faktischen Anonymität bei gleichzeitig bestmöglichem Erhalt des Potenzials für wissenschaftliche Analysen. Die Ergebnisse haben gezeigt, dass in der Regel eine Unterdrückung oder Vergröberung von Informationen bei den qualitativen Merkmalen wie z.B. die Zusammenfassung von Wirtschaftsabteilungen oder eine Vergröberung der Regionalangabe beachtlich zur Anonymisierung beiträgt und eine vergleichsweise schwache Modifikation der im Datensatz vorhandenen quantitativen Merkmale ermöglicht. Bei den für das Scientific-Use-File auf die Originaldaten angewendeten Anonymisierungsmaßnahmen wurde daher ein großes Gewicht auf die Behandlung der qualitativen Merkmale gelegt.

2. Anonymisierungsmaßnahmen

Wie bereits erwähnt, wurde den Anregungen der Nutzer folgend ein stärkeres Gewicht auf die Behandlung der qualitativen Merkmale gelegt.

Traditionelle Anonymisierungsverfahren

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.

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 Projektes „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" wurde die Möglichkeit ausgeschlossen, einen Scientific-Use-File zu erstellen, der administrative Gebietsangaben auf der Ebene

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

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 bei den quantitativen Merkmalen weitestgehend verzichten zu wollen. Als erste Möglichkeit wurde der administrative Gebietsschlüssel durch den nichtadministrativen siedlungsstrukturellen Kreistyp BBR9 und den siedlungsstrukturellen Regionstyp BBR3 ersetzt. Die sehr deutliche Verbesserung der Schutzwirkung durch diese Vergröberung der Regionalinformation wurde in (Lenz/Vorgrimler)[3] festgestellt. Allerdings sprach sich der Wissenschaftliche Begleitkreis dafür aus, anstelle dieser nicht-administrativen Schlüssel eine 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, Ausgabe 1993 (WZ93), auf der Vierstellerebene (Klasse) erhoben und aufbereitet. Diese Klassifikation ist von der europäischen Klassifikation NACE Rev.1 abgeleitet, die aufgrund der NACE-Verordnung des Rates der Europäischen Gemeinschaften seit 1995 in allen Mitgliedstaaten der Europäischen Union sowohl für die Erhebung als auch für die Darstellung der statistischen Daten anzuwenden ist.[4] 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 darauf verständigt, bei dem hierarchischen Merkmal WZ93 die Gliederungstiefe 2 (Zweistellerebene) nicht zu unterschreiten, da hierdurch zum einen eine beachtliche Schutzwirkung und zum anderen nach Einschätzung der beteiligten Wissenschaftler für einen Scientifc-Use-File eine ausreichende Breite an Analysemöglichkeiten erhalten wird.

In den Veröffentlichungen der statistischen Ämter werden aufgrund von Geheimhaltungsaspekten die Ergebnisse einiger Wirtschaftsabteilungen nicht veröffentlicht. Es handelt sich dabei um Unternehmen der Abteilungen *10, 11, 14, 16, 23, 30, 32, 35* und *37* der WZ93. Bei den im Projekt durchgeführten Simulationen hat sich bestätigt, dass diese Abteilungen neben den Abteilungen *15, 17, 18, 19* , 22 und *34* größerer Geheimhaltung bedürfen. Um diese kritischen Abteilungen im Scientific-Use-File belassen und weitgehend auf datenverändernde Verfahren bei den quantitativen Merkmalen verzichten zu können, werden 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

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

3 Siehe Lenz, R./Vorgrimler, D.: „Matching German Turnover Tax Statistics", erscheint in der Reihe der Diskussionspapiere des Forschungsdatenzentrums des Statistischen Bundesamtes, Wiesbaden.

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

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 enthält nachfolgende Tabelle:

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

Eindimensionale Mikroaggregation

Die im Datensatz verbleibenden 30 quantitativen Merkmale wurden eindimensional für jedes Merkmal separat mikroaggregiert.[5] Bei dieser Variante der Mikroaggregation werden zunächst die Merkmalsausprägungen je Merkmal absteigend sortiert. Dann werden tripelweise (aus den

5 Zur Methode der Mikroaggregation und anderen im Projekt untersuchten Methoden siehe Höhne, J.: „Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten" in Ronning, G./Gnoss, R.: „Anonymisierung wirtschaftsstatistischer Einzeldaten", Band 42 der Schriftenreihe „Forum der Bundesstatistik", Wiesbaden 2003, S. 69 ff.

Merkmalsausprägungen dreier benachbarter Merkmalsträger) die Durchschnittswerte ermittelt, die Originalwerte durch diese Durchschnittswerte ersetzt und wieder an die ursprüngliche Position zurücksortiert. 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 ist für das Ziel einer möglichst vielseitigen Datennutzung, sowohl für deskriptive als auch für ökonometrische Auswertungen, das schonendste Verfahren innerhalb der Klasse der Mikroaggregationsverfahren.

3. Stellungnahme des Wissenschaftlichen Begleitkreises

In den Rückmeldungen des für das Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" eingerichteten Wissenschaftlichen Begleitkreises wurden die methodischen Arbeiten zur Beurteilung des Analysepotenzials, welche sowohl deskriptive Maße als auch inferenzstatistische Auswertungen in Form linearer und nichtlinearer ökonometrischer Modellierung beinhalten, als überzeugend beurteilt. Die konkrete Anonymisierungsstrategie für die Kostenstrukturerhebung im Verarbeitenden Gewerbe wurde als sorgfältig und umfassend eingestuft und der Empfehlung zur Erstellung eines Scientific-Use-Files voll zugestimmt. Insgesamt wurden die Forschungsarbeiten des Projektes als sehr erfolgreich im Hinblick auf die Untersuchungsziele bewertet.

A.2 Fragebogen zur Kostenstrukturerhebung

Statistisches Bundesamt
Abteilung IV C
65180 Wiesbaden

Bei Rückfragen erreichen Sie uns:
Telefon 0611-75-2301 oder 2304
Telefax 0611-75-3940

Kostenstrukturerhebung für das Jahr 1999

Jahreserhebung bei Unternehmen des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden

Geschäftsleitung

Unternehmensnummer
(bei Schriftwechsel bitte unbedingt angeben)

Bitte berichtigen Sie Ihre Anschrift, falls sie sich geändert hat.

- **Rechtsgrundlagen, Geheimhaltung, Hilfsmerkmale, Trennen** und **Löschen** sowie Statistikregister siehe Erläuterungen, die Bestandteil des Erhebungsvordrucks sind.
- **Hinweise für das Ausfüllen:** Die Meldung ist für das **gesamte Unternehmen** als rechtlich selbständige Einheit einschließlich aller produzierenden und nichtproduzierenden Teile, jedoch ohne Zweigniederlassungen im Ausland abzugeben. **Nichteinzubeziehen** sind rechtlich selbständige Tochtergesellschaften. Berichtsjahr ist das Kalenderjahr. Deckt sich das Geschäftsjahr nicht mit dem Kalenderjahr, so ist das Geschäftsjahr zu Grunde zu legen, das im Laufe des Jahres 1999 zu Ende ging. **In das Geschäftsjahr sind höchstens 12 Monate einzubeziehen.** Wenn keine Angabe in Betracht kommt, bitten wir, bei der entsprechenden Position einen (-) einzusetzen. **Es ist unbedingt erforderlich, bei den mit ● gekennzeichneten Positionen die beigefügten Erläuterungen zu beachten.**
- **Meldetermin:** Bitte senden Sie ein Exemplar der Erhebungsvordrucke spätestens bis zu dem im Anschreiben genannten Termin ausgefüllt an das Statistische Bundesamt. Sollte der endgültige Jahresabschluss zu diesem Zeitpunkt noch nicht vorliegen, genügen vorläufige Werte aus den entsprechenden Konten oder sorgfältig geschätzte Angaben. Das zweite Exemplar der Erhebungsvordrucke ist für Ihre Unterlagen bestimmt.

I Allgemeine Fragen ❶

1. **Geschäftsjahr vom** ____ **bis** **1999**

2. **Wirtschaftlicher Schwerpunkt des Unternehmens sowie weitere produzierende Tätigkeiten**
Geben Sie bitte den genauen Wirtschaftszweig entsprechend der beigefügten Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ 93) an. Geben Sie bitte zuerst den Schwerpunkt an.

3. **Bitte geben Sie die in dieser Meldung verwendete Währung an.**
Es darf nur eine Währung verwendet werden. → DM od. EUR

II Tätige Personen Ende September 1999 ❷

		Anzahl
1 **Tätige Inhaber(-innen), tätige Mitinhaber(-innen) sowie unbezahlt mithelfende Familienangehörige**	21	
2 **Angestellte und Arbeiter(-innen), einschl. Heimarbeiter(-innen), Zusteller(-innen) einschl. Auszubildende und Teilzeitbeschäftigte**	22	
darunter: Auszubildende		23
darunter: Teilzeitbeschäftigte ❸		24
Anzahl der Teilzeitbeschäftigten in Vollzeiteinheiten (z.B. 3 Halbtagsbeschäftigte ergeben 1,5 Teilzeitbeschäftigte in Vollzeiteinheiten) ❹		25
Summe = (21 + 22)	27	
Außerdem Leiharbeitnehmer (von Dritten zur Verfügung gestelltes Personal) ❺	28	

III Gesamtleistung im Geschäftsjahr 1999 (6)

volle DM od. Euro

1 Umsatz (ohne Umsatzsteuer)		
a Umsatz aus eigenen Erzeugnissen sowie Wert der für Dritte geleistete Lohnarbeiten (einschl. Lohnveredelung) und Erlöse für Reparaturen, Instandhaltung und Installationen, Montagen u. ä. (einschl. Materialien) (7)		35
b Umsatz aus Handelsware (8)		37
c Provisionen aus der Handelsvermittlung (9)		38
d Umsatz aus sonstigen Tätigkeiten (10)		39
Gesamtumsatz = (35 + 37 + 38 + 39)		40
2 Bestände an unfertigen und fertigen Erzeugnissen aus eigener Produktion einschl. geleisteter und noch nicht abgerechneter Lohnarbeiten, Reparaturen, Instandhaltungen, Installationen, Montagen u. ä. (ohne Roh-, Hilfs- und Betriebsstoffe, ohne Handelsware) (11)		
a am Anfang des Geschäftsjahres 1999	J.	41
b am Ende des Geschäftsjahres 1999	+	42
Bestandsveränderung = (42 J. 41)		43
3 Selbsterstellte Anlagen (einschl. Gebäude und selbst durchgeführter Großreparaturen) zu Herstellungskosten, soweit aktiviert, im Geschäftsjahr 1999 (12)		44
Gesamtleistung = (40 + J. 43 + 44)		46

IV Rohstoffe und sonstige fremdbezogene Vorprodukte, Hilfs- und Betriebsstoffe (13)

(Fertigungsmaterial, Fremdbauteile, Energie und Wasser, Büro- und Werbematerial sowie nichtaktivierte geringwertige Wirtschaftsgüter, jedoch ohne Handelsware und ohne Kosten für durch andere Unternehmen ausgeführte Lohnarbeiten) zu Anschaffungskosten, ohne Umsatzsteuer, die als Vorsteuer abzugsfähig ist, im Geschäftsjahr 1999

1 Bestände (14)			
a am Anfang des Geschäftsjahres 1999	+	50	
b am Ende des Geschäftsjahres 1999	J.	51	
2 Eingänge (Einkäufe)		52	
Verbrauch = (50 J. 51 + 52)		53	
darunter Energieverbrauch (ohne Rohstoffe) – keine Mengenangaben – (Brenn- und Treibstoffe, Elektrizität, Gas, Wärme u.dgl.) (15)			55

V Handelsware zu Anschaffungskosten, ohne Umsatzsteuer, die als Vorsteuer abzugsfähig ist, im Geschäftsjahr 1999 (16)

1 Bestände		
a am Anfang des Geschäftsjahres 1999	+	56
b am Ende des Geschäftsjahres 1999	J.	57
2 Eingänge (Einkäufe)		58
Einsatz (vgl. auch 37) = (56 J. 57 + 58)		59

VI Kosten (ohne Materialverbrauch, ohne Einsatz an Handelsware) im Geschäftsjahr 1999

Bitte beachten Sie, dass alle Aufwendungen, die den nachstehenden Tatbeständen entsprechen, vollständig zugeordnet werden. Nicht zu melden sind hier Aufwendungen, die nicht unmittelbar aus der laufenden Produktion resultieren und betriebsfremde Aufwendungen.

Als Kosten sind die auf das Geschäftsjahr entfallenden Beträge anzugeben, nicht die in diesem Geschäftsjahr tatsächlich gezahlten. Nachzahlungen für vorhergehende Jahre und Vorauszahlungen für spätere dürfen daher in den Zahlenangaben nicht enthalten sein.

Wenn Kosten mit Umsatzsteuer belastet sind, die als Vorsteuer abzugsfähig ist, sind die Beträge ohne Umsatzsteuer anzugeben.

	volle DM od. Euro	
1 Bruttogehaltsumme und Bruttolohnsumme (17) (einschl. Arbeitnehmeranteile zur Kranken-, Pflege-, Renten- und Arbeitslosenversicherung, jedoch ohne Arbeitgeberanteile)	60	
2 Sozialkosten		
a Gesetzlich vorgeschriebene Sozialkosten (18) (nur Arbeitgeberanteile zur Kranken-, Pflege-, Renten- und Arbeitslosenversicherung, Berufsgenossenschaftsbeiträge u. ä.)	61	
b Sonstige Sozialkosten (19) (z.B. Beihilfen und Zuschüsse im Krankheitsfalle, Aufwendungen für die betriebliche Altersversorgung, Beiträge zur Aus- und Fortbildung und dgl.)	62	
3 Kosten für Leiharbeitnehmer (durch Dritte zur Verfügung gestelltes Personal) (20)	63	
4 Kosten für durch andere Unternehmen ausgeführte Lohnarbeiten (21)	64	
5 Kosten für Reparaturen, Instandhaltungen, Installationen, Montagen u.ä. (nur fremde Leistungen)	65	
6 Mieten und Pachten (z.B. gemietete und gepachtete Produktionsmaschinen, Datenverarbeitungsanlagen, Fahrzeuge, Fabrikations- und Lagerräume einschl. Kosten für Leasing, jedoch ohne kalkulatorische Mieten)	66	
darunter: Kosten für langfristig gemietete und mit Operating-Leasing beschaffte Produktionsanlagen (22)		67
7 Sonstige Kosten (z.B. Werbekosten [Marketingagenturen usw.], Vertreterkosten, Reisekosten, Provisionen, Lizenzgebühren, Kosten für Grünen Punkt, Ausgangsfrachten und sonstige Kosten für den Abtransport von Gütern durch fremde Unternehmen, Porto- und Postgebühren, Ausgaben für durch Dritte durchgeführte Beförderung der Lohn- und Gehaltsempfänger zwischen Wohnsitz und Arbeitsplatz, Versicherungsbeiträge [einschl. Versicherungsteuer], Prüfungs-, Beratungs- und Rechtskosten, Bankspesen, Beiträge zur Industrie- und Handelskammer, zur Handwerkskammer, zu Wirtschaftsverbänden und dgl., jedoch ohne Kosten für Büro- und Werbematerial sowie Energieverbrauch [gehört zu Pos. IV] usw., ohne kalkulatorische Kosten) Nicht anzugeben sind Aufwendungen, die nicht unmittelbar aus der laufenden Produktion resultieren, und betriebsfremde Aufwendungen (23)	68	
darunter: gezahlte Versicherungsbeiträge		69
8 Steuern sowie öffentliche Gebühren und Beiträge (z.B. Grundsteuer, Gewerbesteuer, Kraftfahrzeugsteuer, Verbrauchsteuern; ohne Einkommen- und Körperschaftsteuer, ohne Lastenausgleichsabgaben, ohne Umsatzsteuer) (24)	71	
darunter: Verbrauchsteuern (nur auf selbst hergestellte Erzeugnisse) (25)		72
9 Steuerliche Abschreibungen auf Sachanlagen Die steuerlichen Abschreibungen sind ohne die in den Erläuterungen aufgeführten Sondervergünstigungen anzugeben. (26)	74	
10 Fremdkapitalzinsen (ohne Bankspesen) (27)	75	
Summe = (60 bis 66 + 68 + 71 + 74 + 75)	78	

volle DM od. Euro

VII Subventionen

Subventionen für die laufende Produktion im Geschäftsjahr 1999 (28)	80

VIII Umsatzsteuer im Geschäftsjahr 1999 (29)

1 Umsatzsteuer, die Kunden in Rechnung gestellt wurde	82	
2 Abzugsfähige Umsatzsteuer, die dem Unternehmen von seinen Lieferanten in Rechnung gestellt wurde, sowie abzugsfähige Erwerb- und Einfuhrumsatzsteuer (Vorsteuer)	83	
darunter: Abzugsfähige Vorsteuer auf den Käufen von Sachanlagen		84

IX Innerbetriebliche Forschung und Entwicklung im Geschäftsjahr 1999 (30)

1 Gesamtaufwendungen für innerbetriebliche Forschung und Entwicklung	86
2 Anzahl der für Forschung und Entwicklung eingesetzten Lohn- und Gehaltsempfänger	87

Bemerkungen (besondere Hinweise, falls außergewöhnliche Verhältnisse die Angaben beeinflusst haben):

Bitte teilen Sie uns mit, an wen wir uns bei Rückfragen wenden dürfen (Freiwillige Angaben).

Ort und Datum:
(Firmenstempel)

Frau / Herr

Telefonnummer

Vielen Dank für Ihre Mitarbeit!

A.3 Erläuterungen zur Kostenstrukturerhebung

Statistisches Bundesamt
65180 Wiesbaden

Bj 1999

Jahreserhebung bei Unternehmen des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden

Kostenstrukturerhebung

Erläuterungen zum Erhebungsvordruck

Zweck, Art und Umfang der Erhebung
Die Kostenstrukturerhebung wird jährlich als repräsentative Stichprobe bei höchstens 18 000 Unternehmen des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden durchgeführt. Ihre Ergebnisse liefern notwendige Informationen als Grundlage der Wirtschaftspolitik auf nationaler und internationaler Ebene. Grundlegende Bedeutung gewinnt die Erhebung mit der Vollendung des gemeinsamen Binnenmarktes auf europäischer Ebene. Darüber hinaus dient sie auch den Unternehmen und ihren Verbänden als wertvolle Informationsquelle.

Rechtsgrundlagen
Verordnung (EG, Euratom) Nr. 58/97 des Rates vom 20. Dezember 1996 über die strukturelle Unternehmensstatistik (ABl. EG Nr. L 14 S. 1).

Gesetz über die Statistik im Produzierenden Gewebe (ProdGewG) in der Fassung der Bekanntmachung vom 30. Mai 1980 (BGBl. I S. 641), zuletzt geändert durch Artikel 1 des Gesetzes vom 6. August 1998 (BGBl. I S. 2036) in Verbindung mit dem Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz - BStatG) vom 22. Januar 1987 (BGBl. I S. 462, 565), zuletzt geändert durch Artikel 2 des Gesetzes vom 16. Juni 1998 (BGBl. I S. 1300).

Erhoben werden die Tatbestände zu §3 Buchstabe B Ziffer II ProdGewG Anhang 2 Abschnitt 4 der Verordnung (EG, Euratom) Nr. 58/97.

Die Auskunftspflicht ergibt sich aus §9 ProdGewG, Artikel 6 Abs. 2 Verordnung (EG, Euratom) Nr. 58/97 in Verbindung mit §§15, 18, 26 Abs. 4 Satz 1 BStatG. Hiernach sind die Inhaber oder Leiter der Unternehmen auskunftspflichtig. Gem. §15 Abs. 6 BStatG haben Widerspruch und Anfechtungsklage gegen die Aufforderung zur Auskunftserteilung keine aufschiebende Wirkung.

Geheimhaltung
Die erhobenen Einzelangaben werden nach §16 BStatG grundsätzlich geheimgehalten. Nur in ausdrücklich gesetzlich geregelten Ausnahmefällen dürfen Einzelangaben übermittelt werden. Nach §16 Abs. 6 BStatG ist es möglich, den Hochschulen oder sonstigen Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung für die Durchführung wissenschaftlicher Vorhaben Einzelangaben dann zur Verfügung zu stellen, wenn diese so anonymisiert sind, dass sie nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft dem Befragten oder Betroffenen zugeordnet werden können. Nach §47 des Gesetzes gegen Wettbewerbsbeschränkungen in der Fassung der Bekanntmachung vom 26. August 1998 (BGBl. I S. 2546), das durch Artikel 9 des Gesetzes vom 22. Dezember 1999 (BGBl. I S. 2626) geändert worden ist, dürfen der Monopolkommission für die Begutachtung der Entwicklung der Unternehmenskonzentration zusammengefasste Einzelangaben über die Vomhundertanteile der größten Unternehmen, Betriebe oder fachlichen Teile von Unternehmen des jeweiligen Wirtschaftsbereichs übermittelt werden. Hierbei dürfen die zusammengefassten Einzelangaben nicht weniger als drei Einheiten betreffen und keine Rückschlüsse auf zusammengefasste Angaben von weniger als drei Einheiten ermöglichen. Die Pflicht zur Geheimhaltung besteht auch für Personen, die Empfänger von Einzelangaben sind.

Hilfsmerkmale, Trennen und Löschen, Statistikregister
Name und Anschrift des Unternehmens, Name und Telefonnummer der für eventuelle Rückfragen zur Verfügung stehenden Person, Ort, Datum und die Unterschrift sowie die Angaben zu Abschnitt I „Allgemeine Fragen" sind Hilfsmerkmale, die lediglich der technischen Durchführung der Erhebung dienen. Sie werden nach Abschluss der Prüfung der Angaben vom Erhebungsvordruck getrennt, gesondert aufbewahrt und spätestens nach Abschluss der nächsten Erhebung mit Ausnahme von Namen und Anschrift des Unternehmens vernichtet. Die verwendete Unternehmensnummer dient der Unterscheidung der in die Erhebung einbezogenen Unternehmen. Sie besteht aus einem Regionalschlüssel für das jeweilige Bundesland und aus einer laufenden, frei vergebenen Nummer. Hinzu kommen eine Nummer, die den wirtschaftlichen Schwerpunkt des Unternehmens darstellt (WZ93), sowie ein Schlüssel für die jeweilige Rechtsform des Unternehmens.

Die Angaben zu Name und Anschrift des Unternehmens, die Unternehmensnummer, WZ93 und Rechtsform werden zur Führung des Unternehmensregisters für statistische Verwendungszwecke (Statistikregister) verwendet. Rechtsgrundlagen hierfür sind §13 BStatG und die Verordnung (EWG) Nr. 2186/93 des Rates vom 22. Juli 1993 über die innergemeinschaftliche Koordinierung des Aufbaus von Unternehmensregistern für statistische Verwendungszwecke (ABl. EG Nr. L 196 S. 1).

Berichtskreisabgrenzung
Die Erhebung erstreckt sich auf Unternehmen des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden. Als Unternehmen gilt die kleinste Einheit, die aus handels- und/oder steuerrechtlichen Gründen Bücher führt und bilanziert.

Rechtlich selbständige Tochtergesellschaften, Arbeitsgemeinschaften, Betriebsführungsgesellschaften usw. müssen getrennt berichten.

Die Meldung ist grundsätzlich für das gesamte Unternehmen einschl. aller produzierenden und nichtproduzierenden Teile, jedoch ohne Zweigniederlassungen im Ausland, abzugeben. Zusammengefasste Meldungen für zwei oder mehrere rechtlich selbständige Unternehmen sind nicht zulässig.

Soweit die vorhandenen Unterlagen zur Beantwortung einzelner Fragen nicht ausreichen, genügen vorläufige Werte aus den entsprechenden Konten oder sorgfältig geschätzte Angaben.

1 I Allgemeine Fragen

Der wirtschaftliche Schwerpunkt und weitere produzierende Tätigkeiten des Unternehmens sind so anzugeben, wie sie durch die vierstelligen Positionen der beigefügten Klassifikation der Wirtschaftszweige, Ausgabe 1993 (WZ93), unterschieden werden.

Zum Beispiel:
Wirtschaftlicher Schwerpunkt des Unternehmens:
29.40 Herstellung von Werkzeugmaschinen
Weitere produzierende Tätigkeiten des Unternehmens:
29.71 Herstellung von elektrischen Haushaltsgeräten
28.62 Herstellung von Werkzeugen

Umstellung der Währung DM/Euro:
Sämtliche Wertangaben sind in der Währungseinheit anzugeben, die Sie unter „Allgemeine Angaben" des Erhebungsvordrucks eingetragen haben. Es darf nur eine Währung verwendet werden.

2 II Tätige Personen

Tätige Personen sind:
- tätige Inhaber und Mitinhaber (nur von Personengesellschaften),
- unbezahlt mithelfende Familienangehörige, soweit sie mindestens 1/3 der üblichen Arbeitszeit im Unternehmen tätig sind,
- Angestellte und Arbeiter
- Gesellschafter, Vorstandsmitglieder und andere leitende Kräfte, soweit sie vom befragten Unternehmen Bezüge erhalten, die steuerlich als Einkünfte aus nichtselbständiger Arbeit angesehen werden sowie Auszubildende, Volontäre, Praktikanten, Reisende im Angestelltenverhältnis, Aushilfsarbeiter, Heimarbeiter und Zusteller.

Voll als tätige Personen zuzählen sind:
- Erkrankte, Urlauber, Personen, die lediglich Übungen bei der Bundeswehr ableisten, im Mutterschutz oder Erziehungsurlaub (weniger als 1 Jahr) befindliche Personen und alle sonstigen vorübergehend Abwesenden,
- Streikende und von der Aussperrung Betroffene, solange das Arbeitsverhältnis nicht gelöst ist,
- Saison- und Aushilfsarbeiter, Teilzeitbeschäftigte und Kurzarbeiter,
- das Personal auf Bau- und Montagestellen, Fahrzeugen usw.,
- nur vorübergehend im Ausland Tätige (weniger als 1 Jahr).

Nicht zu melden sind:
- ständig im Ausland tätige Personen (mindestens 1 Jahr),
- zum Grundwehrdienst Einberufene, Zivildienstleistende,

- Arbeitskräfte, die von Arbeitsvermittlungsagenturen u.ä. Einrichtungen gegen Entgelt zur Arbeitsleistung gemäß dem Arbeitnehmerüberlassungsgesetz bereit gestellt werden (Leiharbeitnehmer wie Fremdlöhner, Zeitbeschäftigte für Bürotätigkeiten usw.),
- Arbeitskräfte, die als Beauftragte anderer Unternehmen im meldenden Unternehmen Montage- und Reparaturarbeiten durchführen,
- Arbeitskräfte, die 1 Jahr oder länger im Erziehungsurlaub sind,
- Strafgefangene,
- Empfänger von Vorruhestandsgeld,
- unbezahlt mithelfende Familienangehörige mit weniger als 1/3 der branchenüblichen Arbeitszeit.

3 **Teilzeitbeschäftigte** sind ständig Beschäftigte, deren normale Arbeitszeit kürzer als die reguläre Arbeitszeit ist. Dies betrifft alle Formen der Teilzeitarbeit (Halbtagsbeschäftigung, Beschäftigung an einem, zwei oder drei Tagen in der Woche usw.). Hierzu zählen auch Zusteller im Verlagsgewerbe, die in der Lohnliste geführt werden sowie Altersteilzeitbeschäftigte, dagegen nicht Zwischenmeister und Hausgewerbetreibende.

4 **Teilzeitbeschäftigte in Vollzeiteinheiten** sind Anzahl der durch alle Teilzeitbeschäftigten eines Unternehmens erarbeiteten Wochenarbeitsstunden geteilt durch die in diesem Unternehmen reguläre Wochenarbeitszeit eines Vollzeitbeschäftigten, z.B.:

Müller	19,25 Std./Woche
Maier	25,50 Std./Woche
Becker	15,00 Std./Woche
	59,75 Std./Woche
	: 37,5 Std./reguläre Wochenarbeitszeit (Beispiel)
	1,6 Vollzeiteinheiten

5 **Leiharbeitnehmer** sind Arbeitskräfte, die von Arbeitsvermittlungsagenturen u.ä. Einrichtungen gegen Entgelt zur Arbeitsleistung gemäß dem Arbeitnehmerüberlassungsgesetz überlassen wurden. Nicht anzugeben ist von Tochterunternehmen und verbundenen Unternehmen zur Verfügung gestelltes Personal.

6 **II Gesamtleistung**
Als Umsatz gilt, unabhängig vom Zahlungseingang, der Gesamtbetrag **(ohne Umsatzsteuer)** der abgerechneten Lieferungen und Leistungen an Dritte.

Einzubeziehen sind:
- Erlöse aus Lieferungen und Leistungen an mit dem Unternehmen verbundene rechtlich selbständige Konzern- und Verkaufsgesellschaften,
- auch etwa getrennt in Rechnung gestellte Kosten für Fracht, Porto und Verpackung.

Abzusetzen sind:
- Preisnachlässe (Rabatte, Boni, Skonti, Abzüge, die auf begründeten Beanstandungen beruhen und dgl.) sowie Retouren.

Nicht einzubeziehen sind:
- Erträge, die nicht unmittelbar aus laufender Produktionstätigkeit resultieren,
- Erlöse aus dem Verkauf von Sachanlagen,
- Erlöse aus der Verpachtung von Grundstücken,
- Zinserträge, Dividenden u.dgl.,
- Erzeugnisse und Leistungen, die für eigene Investitionen und Sachanlagen (Grundmittel) bestimmt sind (vgl. auch 12).

7 **Umsatz aus eigenen Erzeugnissen schließt ein:**
- Umsätze aus dem Verkauf von allen im Rahmen der Produktionstätigkeit des Unternehmens entstandenen Erzeugnissen,
- die vollen Erlöse aus dem Verkauf von eigenen Erzeugnissen, die unter Verwendung von Fremdbauteilen hergestellt wurden,
- Umsätze aus dem Verkauf von Waren, die in Lohnarbeit bei anderen Unternehmen hergestellt wurden,
- Umsätze aus dem Verkauf von Elektrizität, Fernwärme, Gas, Dampf, Wasser,
- Umsätze aus dem Verkauf von Nebenerzeugnissen,
- Erlöse für verkaufsfähige Produktionsrückstände (z.B. bei der Produktion anfallender Schrott, Gussbruch, Wollabfälle u.ä.),
- Erlöse für die Vermietung bzw. das Leasing von im Rahmen der Produktionstätigkeit des Unternehmens selbst hergestellten Erzeugnissen oder Anlagen,
- Erlöse aus Redaktions- und Verlagstätigkeit,
- Umsatz aus Recycling.

Bei den Erlösen für Reparaturen, Instandhaltungen, Installationen, Montagen u.ä. sind die Erlöse für die bei diesen Leistungen verbrauchten Materialien (z.B. Ersatzteile, Zubehör, Hilfs- und Betriebsstoffe) einzubeziehen.
Nicht einzubeziehen sind Erlöse für Reparaturen von Gebrauchsgütern, Instandhaltung und Reparatur von Kraftwagen, Krafträdern sowie Büromaschinen, Datenverarbeitungsgeräten und -einrichtungen (s. auch 10).

8 **Als Umsatz aus Handelsware gilt** der Umsatz von fremden Erzeugnissen, die im allgemeinen unbearbeitet und ohne fertigungstechnische Verbindung mit eigenen Erzeugnissen weiterverkauft werden. Die hier angegebenen Erlöse sind mit dem Einsatz an Handelsware zu Anschaffungskosten abzustimmen (vgl. auch 16).

9 **Provisionen aus der Handelsvermittlung** sind Vergütungen für den gewerbsmäßigen Kauf oder Verkauf im eigenen Namen von Waren für Rechnung eines anderen.

10 **Umsatz aus sonstigen Tätigkeiten:**
Hierzu zählen im wesentlichen:
- Umsätze aus der Vermietung und Verpachtung von Geräten, betrieblichen Anlagen und Einrichtungen, die nicht im Rahmen der Produktionstätigkeit des Unternehmens entstanden sind (einschl. Leasing),
- Erlöse aus Wohnungsvermietung (von betrieblich und nicht betrieblich genutzten Wohngebäuden), jedoch ohne Erlöse aus Grundstücksverpachtung,
- Erlöse aus der Veräußerung von Patenten und der Vergabe von Lizenzen,
- Erlöse aus Transportleistungen für Dritte,
- Erlöse aus Belegschaftseinrichtungen (z.B. Erlöse einer vom Unternehmen auf eigene Rechnung betriebenen Kantine),
- Erlöse aus dem Verkauf von eigenen landwirtschaftlichen Erzeugnissen,
- Erlöse aus Reparaturen von Gebrauchsgütern, Instandhaltung und Reparatur von Kraftwagen und Krafträdern,
- Erlöse aus Instandhaltung und Reparatur von Büromaschinen, Datenverarbeitungsgeräten und -einrichtungen,
- Erlöse aus Beratungs- und Planungstätigkeit,
- Provisionseinnahmen.

11 **Die Bestände an unfertigen und fertigen Erzeugnissen aus eigener Produktion** sind zu **Herstellungskosten** zu bewerten.
Bestände an Einzel-, Ersatz- und Einbauteilen aus eigener Produktion sind einzubeziehen.
Anzahlungen bzw. Abschlagszahlungen (z.B. im Stahlbau, Schiffbau, Großapparatebau) dürfen nicht abgesetzt werden.

12 **Es sollen die im Geschäftsjahr mit eigenen Arbeitskräften** (einschl. Leiharbeitnehmer) **selbsterstellten Anlagen** (einschl. im Bau befindlicher Anlagen) mit dem auf dem Anlagenkonto aktivierten Wert (Herstellungskosten) als Leistungen des eigenen Unternehmens angegeben werden, sofern die Kosten für die Erstellung in den Angaben unter IV und VI mitenthalten sind.

Zu den **selbsterstellten Anlagen** gehören auch selbsthergestellte Sachanlagen, die an Dritte vermietet oder verpachtet wurden, selbsthergestellte Maschinen, Werkzeuge, Modelle für das eigene Unternehmen, Versuche usw., soweit diese aktiviert wurden.

Abschreibungen auf die selbsterstellten Anlagen sind nicht abzusetzen.

13 **IV Rohstoffe und sonstige fremdbezogene Vorprodukte, Hilfs- und Betriebsstoffe**
Zu den Roh-, Hilfs- und Betriebsstoffen zählen alle Materialien und Fremdbauteile (ohne Handelsware), die entweder im Unternehmen be- oder verarbeitet oder verbraucht oder an Dritte zur Be- oder Verarbeitung weitergegeben werden. Es spielt auch keine Rolle, in welchem Bereich des Unternehmens diese Stoffe verwendet werden.

Mit anzugeben sind also z.B. auch **Energie** (Brenn- und Treibstoffe, Elektrizität, Gas, Wärme u.dgl.) und **Wasser, Ersatzteile**, Büro- und Werbematerial, Verpackungsmaterial und Waren, die in einer vom Unternehmen auf eigene Rechnung betriebenen Kantine u.dgl. verarbeitet oder verkauft werden. Einzubeziehen sind auch **Materialien, die für die Herstellung von selbsterstellten Anlagen** benötigt werden.

14 **Die Bestände** und **Eingänge** an Roh-, Hilfs- und Betriebsstoffen sind zu **Anschaffungskosten** (ohne als Vorsteuer abzugsfähige Umsatzsteuer) zu bewerten. Als Anschaffungskosten gelten die Anschaffungspreise zuzüglich Anschaffungsnebenkosten wie Fracht, Verpackung,

Zoll, Verbrauchsteuern u.dgl. abzüglich Preisnachlässe (Rabatte, Boni, Skonti, Abzüge, die auf begründeten Beanstandungen beruhen u.dgl.). Subventionen sind hier nicht abzusetzen (s. auch 28).

Als Eingänge ist der Wert aller von Dritten bezogenen Materialien und Fremdbauteile (ohne Handelsware) zu melden, gleichgültig, ob diese Eingänge über Bestandskosten oder unmittelbar als Aufwand verbucht wurden. Einzubeziehen sind auch nichtaktivierte geringwertige Wirtschaftsgüter.

15 **Als Energieverbrauch ist der Gesamtverbrauch an Brenn- und Treibstoffen, Elektrizität, Gas, Wärme u.dgl. anzugeben.**
Wasser - als Bestandteil der Roh-, Hilfs- und Betriebsstoffe - ist in die Position „Energieverbrauch" nicht einzubeziehen.

16 **V Handelsware**
Als Handelsware gelten Waren fremder Herkunft, die im allgemeinen unbearbeitet und ohne fertigungstechnische Verbindung mit eigenen Erzeugnissen weiterverkauft werden (vgl. auch 4 Umsatz aus Handelsware).
Die Bestände und Eingänge an Handelsware sind zu Anschaffungskosten (ohne als Vorsteuer abzugsfähige Umsatzsteuer) zu bewerten. Als Anschaffungskosten gelten die Anschaffungspreise zuzüglich Anschaffungsnebenkosten wie Fracht, Verpackung, Zoll, Verbrauchsteuern u.dgl. abzüglich Preisnachlässe (Rabatte, Boni, Skonti, Abzüge, die auf begründeten Beanstandungen beruhen u.dgl.).

17 **VI Kosten**
Bruttogehälter und Bruttolöhne
Bei den Bruttogehältern und Bruttolöhnen ist die Summe der Bruttobezüge (Bar- und Sachbezüge) ohne jeden Abzug anzugeben. Diese Beträge verstehen sich einschl. Arbeitnehmeranteile, jedoch ohne Arbeitgeberanteile zur Kranken-, Renten- und Arbeitslosenversicherung.

Zur Bruttogehalt- und Bruttolohnsumme gehören auch die an tätige Personen in eigenen Sozialeinrichtungen (z.B. Werksarzt) gezahlten Beträge und auch die Bezüge von Gesellschaftern, Vorstandsmitgliedern und anderen leitenden Kräften, soweit sie steuerlich als Einkünfte aus nichtselbständiger Arbeit anzusehen sind sowie die Entgelte für Heimarbeiter, Aushilfen und Zusteller.

In die Bruttogehalt- und Bruttolohnsumme einzubeziehen sind:
- sämtliche Zuschläge (z.B. für Akkord-, Band-, Montage-, Schicht- und Sonntagsarbeit sowie Leistungs-, Schmutz- und Lästigkeitszulagen),
- Vergütungen für Feiertage, Urlaub, Arbeitsausfälle u.dgl.,
- Gehalt- und Lohnfortzahlung im Krankheitsfall einschl. Zuschüsse zum Krankengeld,
- Gratifikationen, zusätzliche Monatsgehälter, Gewinnbeteiligungen, Urlaubsbeihilfen und sonstige einmalige Gehalt- und Lohnzahlungen,
- Entschädigungen für nicht gewährten Urlaub,
- Mietbeihilfen und Wohnungszuschüsse, tarif- oder einzelvertraglich vereinbarte Kindergelder und sonstige Familienzuschläge sowie Erziehungsbeihilfen,
- Essengeld, Wegezeitentschädigungen, Fahrtkostenersatz und -zuschüsse für Fahrten von und zur Arbeitsstätte, sofern hierfür Lohnsteuer entrichtet wurde,
- Auslösungen, sofern hierfür Lohnsteuer entrichtet wurde (Auslösungen, die als Spesenersatz gelten, sind bei den Sonstigen Kosten auszuweisen),
- Leistungen des Arbeitgebers im Sinne von § 3 des Fünften Gesetzes zur Förderung der Vermögensbildung der Arbeitnehmer,
- an Angestellte gezahlte Provisionen und Tantiemen,
- an Arbeitnehmer gezahlte Abfindungen.

Unternehmen, die in ihrer Gewinn- und Verlustrechnung „Löhne und Gehälter" entsprechend den handelsrechtlichen Bestimmungen ausweisen, geben hier diesen Wert an.

Abzüglich geleisteter Zuschüsse der Bundesanstalt für Arbeit (z.B. Kurzarbeitergeld, Leistungen nach dem Altersteilzeitgesetz).
Nicht einzubeziehen sind Beträge, die für Leiharbeitnehmer gezahlt werden und der kalkulatorische Unternehmerlohn.

18 **Zu den gesetzlich vorgeschriebenen Sozialkosten zählen:**
- Arbeitgeberanteile zur Kranken-, Pflege-, Renten- und Arbeitslosenversicherung,
- Berufsgenossenschaftsbeiträge,
- Aufwendungen und Zuschüsse zur Betriebskrankenkasse nach der RVO,
- gesetzlich vorgeschriebene Beiträge zur Krankenversicherung nichtversicherungspflichtiger Arbeitnehmer.

Nicht zu den gesetzlich vorgeschriebenen Sozialkosten zählen die im Rahmen von Vorruhestandsleistungen anfallenden Arbeitgeberbeiträge zu Renten- und Krankenversicherung.

19 **Zu den Sonstigen Sozialkosten zählen insbesondere:**
- direkte Zuwendungen an die Arbeitnehmer oder deren Familienangehörige bei besonderen Anlässen, wie z.B. Weihnachtsgeschenke, Jubiläumsgelder, Treueprämien, Zuwendungen aus Anlass von Familienereignissen, Baraufwendungen anlässlich von Betriebsfeiern, Belegschaftsausflügen usw.,
- Beihilfen und Zuschüsse im Krankheitsfalle, zu Erholungs- und Kuraufenthalten und für sonstige Zwecke,
- Aufwendungen für die betriebliche Altersversorgung (Alters-, Invaliditäts- und Hinterbliebenenversorgung) wie
- unmittelbare Versorgungszahlungen an frühere Arbeitnehmer oder deren Hinterbliebene, sofern sie nicht aus Pensionsrückstellungen geleistet werden,
- Rückstellungen für Pensionsverpflichtungen im Sinne von § 6a Einkommensteuergesetz,
- Zuwendungen an Pensions- und Unterstützungskassen, einmalige oder laufende Beiträge für die zur betrieblichen Altersversorgung abgeschlossenen Lebensversicherungen (Direktversicherungen),
- unmittelbare Zahlungen an Bezieher von Vorruhestandsgeld, sofern sie nicht aus Rückstellungen für Vorruhestandsleistungen getätigt werden, sowie Rückstellungen für Vorruhestandsleistungen (die Vorruhestandsleistungen verstehen sich einschl. der Arbeitgeberbeiträge zur Renten- und Krankenversicherung für den in Frage kommenden Personenkreis und abzüglich der im Rahmen der Vorruhestandsvereinbarungen geleisteten Zuschüsse der Bundesanstalt für Arbeit),
- periodische Zahlungen an ausgeschiedene Mitarbeiter,
- anstelle von laufenden Versorgungsleistungen gewährte Kapitalabfindungen,
- Beiträge an den Träger der Insolvenzsicherung gegen die Nichterfüllung von Versorgungsansprüchen,
- Beiträge oder Beitragsteile zu Weiter-, Über- bzw. Zusatzversicherungen und an private Krankenkassen, soweit die Leistung den gesetzlich vorgeschriebenen Betrag übersteigt,
- Beiträge zur Ausbildung und Fortbildung (Zahlung von Schulgeld, Umlagebeiträge für Berufs- und Fachschulen), Geldzuweisungen für Lehrlingsheime, Kantinen sowie für den Gesundheitsdienst, die Betriebsfürsorge u.dgl.).

Hierzu gehören nicht Kosten, die im Rahmen von betrieblichen Sozialeinrichtungen (wie Gesundheitsdienst, Betriebsfürsorge u.dgl.), für Löhne und Gehälter, Material usw. entstanden sind. Diese sind bei den anderen Kostenarten aufzuführen. Auszuschließen sind hier auch Kosten, die als Spesenersatz anzusehen und unter den Sonstigen Kosten (Pos. VI 7) auszuweisen sind.

20 **Kosten für Leiharbeitnehmer**
Hierzu zählen nur die Aufwendungen für Arbeitskräfte, die von Arbeitsvermittlungsagenturen u.ä. Einrichtungen gegen Entgelt zur Arbeitsleistung gemäß dem Arbeitnehmerüberlassungsgesetz überlassen wurden (siehe auch 5).

21 **Kosten für durch andere Unternehmen ausgeführte Lohnarbeiten**
sind Entgelte für die Be- oder Verarbeitung von eigenem (beigestelltem) Material durch fremde Unternehmen (auswärtige Bearbeitung). Hierzu zählen auch die Entgelte an Zwischenmeister und Hausgewerbetreibende, nicht dagegen Löhne für Heimarbeiter oder Zusteller.

22 **Aufwendungen für die langfristige Anmietung** sind jene für die Anmietung von Sachanlagen über einen Zeitraum von mehr als einem Jahr. **Beim Operating-Leasing handelt es sich um Mietverträge über bewegliche dauerhafte Sachanlagen.**

23 **Sonstige Kosten**
Es sind u.a. Kosten für den Abtransport von Gütern durch fremde Unternehmen aufzuführen. Transportkosten, die bei der Anlieferung von Roh-, Hilfs- und Betriebsstoffen usw. durch fremde Unternehmen entstanden sind, sind in den Material- und Wareneingängen und Material- und Warenbeständen enthalten und gehen damit in den ermittelten Materialverbrauch und Wareneinsatz (Pos. IV u. V) ein. Die Kosten für den eigenen Fuhrpark sind aufgegliedert bei den einzelnen Kostenpositionen anzugeben, z.B. Fahrerlöhne bei Pos. VI 1, Instandhaltungskosten bei Pos. VI 5, Kraftfahrzeugsteuer bei Pos. VI 8, Abschreibungen bei Pos. VI 9 und Versicherungsprämien bei Pos. VI 7.
Falls ein Sammelkonto (Kostenstelle Kfz-Kosten) besteht und dessen Aufgliederung besondere Schwierigkeiten bereitet, genügen sorgfältig geschätzte Angaben zu den einzelnen Positionen. Die eigenen Transportkosten bleiben also bei Selbstabholung von Roh-, Hilfs- und Be-

triebsstoffen u.dgl. bei den Material- und Wareneingängen und Material- und Warenbeständen unberücksichtigt und gehen deshalb nicht in den ermittelten Materialverbrauch und Wareneinsatz (Pos. IV u. V) ein.

Provisionen an Angestellte sind bei den Gehältern (Pos. VI 1a) auszuweisen; alle übrigen Provisionen hier bei den Sonstigen Kosten.

Zu den **Sonstigen Kosten** zählen z.B. nicht Einkommen-, Körperschaft- und Erbschaftsteuer sowie Lastenausgleichsabgaben, an Abnehmer gewährte Preisnachlässe (Rabatte, Boni, Skonti, Abzüge, die auf begründeten Beanstandungen beruhen u.dgl.).

(24) **Zu den Steuern, die als Kosten anzusehen sind, zählen u.a. die**
- Grundsteuer,
- Gewerbesteuer,
- Kraftfahrzeugsteuer,
- Verbrauchsteuern (s. auch (25)).

Es sind nur die auf das Geschäftsjahr tatsächlich entfallenden Beträge anzugeben.
Öffentliche Gebühren und Beiträge sind Abgaben, die für bestimmte Leistungen des Staates bezahlt werden, wie Eichgebühren usw. Beiträge zur Industrie- und Handelskammer und zur Handwerkskammer sind nicht hier, sondern bei den **Sonstigen Kosten** zu melden.

(25) **Es sind nur die Verbrauchsteuern** (Bier-, Mineralöl-, Schaumwein-, Tabaksteuer und Branntweinaufschlag) anzugeben, die das Unternehmen auf die **selbst hergestellten** verbrauchsteuerpflichtigen Erzeugnisse schuldet, unabhängig davon, ob eine Zahlung erfolgt.
Verbrauchsteuern auf bezogene Erzeugnisse gelten als **Anschaffungsnebenkosten** bei der Bewertung der Bestände und Eingänge an Roh-, Hilfs- und Betriebsstoffen (Pos. IV) bzw. an Handelsware (Pos. V).

Werden von Unternehmen der Spirituosenindustrie Alkohol oder Destillate zur Weiterverarbeitung von der Bundesmonopolverwaltung oder von in- oder ausländischen Unternehmen bezogen, so ist die hierauf entfallende Branntweinsteuer hier nicht anzugeben. Dies gilt auch, wenn das Vorprodukt im Wege des Begleitscheinverfahrens bezogen wird und die später fällige Branntweinsteuer an die Zollverwaltung abzuführen ist. Von Brennereien und Unternehmen mit eigener Brennerei (Vorprodukte: Wein, Getreide, Obst) ist der für das fertige Erzeugnis fällige Branntweinaufschlag hier auszuweisen).

(26) Zu den Sonderabschreibungen bzw. erhöhten Absetzungen, die nicht mit aufzuführen sind, gehören insbesondere Abschreibungen nach § 7d, § 7e EstG (Umweltschutzinvestitionen; Bewertungsfreiheit für Fabrikgebäude, Lagerhäuser und landwirtschaftliche Betriebsgebäude), § 81 EstDV (Bewertungsfreiheit für bestimmte Wirtschaftsgüter des Anlagevermögens im Kohlen- und Erzbergbau).

Dagegen sind **geringwertige Wirtschaftsgüter** im Sinne von § 6 Abs. 2 EstG, soweit sie nicht in einer anderen Kostenposition (z.B. IV 2) schon enthalten sind, einzubeziehen.

(27) Zu den **Fremdkapitalzinsen** gehören die Zinsen für langfristige Schulden, für Lieferanten- und Bankkredite, Zinsen für sonstige Schulden einschl. Diskont (ohne Wechselspesen) und Provisionen für Bankkredite (insbesondere Kredit- und Überziehungsprovision sowie Kreditbereitstellungsprovision). Fremdkapitalzinsen aufgrund reiner Finanzgeschäfte dürfen nicht enthalten sein. Bankspesen (z.B. Kontoführungsgebühren, Wechselspesen, Gebühren für Scheck- und Überweisungsvordrucke, Depotgebühren) sind unter den Sonstigen Kosten (Pos. VI 7) anzugeben. Die Fremdkapitalzinsen dürfen nicht mit Zinserträgen saldiert ausgewiesen werden.

(28) **VII Subventionen**
Unter Subventionen sind zu melden:
Zuwendungen, die Bund, Länder und Gemeinden oder Einrichtungen der Europäischen Gemeinschaften ohne Gegenleistung an das Unternehmen für Forschungs- und Entwicklungsvorhaben (soweit nicht spezielle Auftragsforschung für den Staat) oder für laufende Produktionszwecke gewähren, um
- die Produktionskosten zu verringern und/oder
- die Verkaufspreise der Erzeugnisse zu senken und/oder
- eine hinreichende Entlohnung der Produktionsfaktoren

zu ermöglichen.

Hierzu zählen z.B.:
- Zinszuschüsse, gleichgültig für welche Zwecke sie gewährt werden (auch dann, wenn sie an den Kreditgeber direkt gezahlt werden),
- Frachthilfen,
- Lohnkostenzuschüsse für ältere Arbeitnehmer nach § 97 AFG,
- Stützungsmaßnahmen für Magermilch.

Subventionen dürfen in den Umsatzerlösen nicht enthalten sein.

Nicht zu den Subventionen zählen:
Steuererleichterungen, Investitionszuschüsse, -zulagen sowie Ersatzleistungen für Katastrophenschäden und sonstige außerordentliche bzw. außerhalb des Verantwortungsbereiches des Unternehmens liegende Verluste.

(29) **VIII Umsatzsteuer**
Es ist nur die auf das Geschäftsjahr entfallende Umsatzsteuer anzugeben. Hierzu zählt auch die Umsatzsteuer auf geleistete und empfangene Anzahlungen.

Nicht anzugeben ist die Einfuhrumsatzsteuer für Materialien, die von ausländischen Auftraggebern zur Lohnveredlung beigestellt worden sind.

Von Tochtergesellschaften ist die Umsatzsteuer auf ihre Außenumsätze und -bezüge zu melden, obwohl sie von der Muttergesellschaft getragen bzw. in Anrechnung gebracht wird. Diese Beträge sind nicht von der Muttergesellschaft nachzuweisen.

Soweit entsprechende Unterlagen über die abzugsfähige Umsatzsteuer auf den Käufen von Sachanlagen nicht vorliegen, genügt eine sorgfältige Schätzung (16 % der Käufe von Sachanlagen).

(30) **IX Innerbetriebliche Forschung und Entwicklung**
Forschung und Entwicklung umfasst systematische schöpferische Arbeiten mit dem Ziel, das Wissenspotential zu erweitern sowie die Nutzung dieses Wissenspotentials zur Schaffung neuer Anwendungen.

Bei den innerbetrieblichen Aufwendungen handelt es sich um sämtliche Aufwendungen, die für die im Unternehmen selbst durchgeführten Forschungs- und Entwicklungsarbeiten anfallen, unabhängig von der Herkunft der Mittel (einschl. Investitionsaufwendungen).

Folgende Tätigkeiten zählen nicht zu innerbetrieblicher Forschung und Entwicklung:
- Tätigkeiten im Rahmen des Bildungswesens,
- Sonstige Tätigkeiten im wissenschaftlich-technischen Bereich (z.B. Informationsdienste, Prüfung und Standardisierung, Durchführbarkeitsstudien usw.),
- Sonstige industrielle Tätigkeiten (z.B. Produktionsvorbereitung, Erwerb externen Wissens, Mitarbeiterschulung, Marketing, auch wenn sie im Rahmen von Innovationen durchgeführt werden).

Für Forschung und Entwicklung eingesetzte Lohn- und Gehaltsempfänger:

Hierunter fallen alle direkt mit Forschungs- und Entwicklungsarbeiten befassten Mitarbeiter sowie das direkte Dienstleistungen erbringende Personal, wie Manager, Verwaltungs- und Büroangestellte. Mitarbeiter, die indirekte Dienstleistungen erbringen, wie Kantinenpersonal und Betriebsschutzmitarbeiter, fallen nicht unter diese Position, auch wenn ihre Löhne und Gehälter als Gemeinkosten in diese Aufwendungen eingehen.

Nicht erfasst werden:
- Mitarbeiter, die sich mit Weiterbildung und Schulungen befassen,
- Mitarbeiter, die sich mit anderen wissenschaftlichen und technischen Aktivitäten (z.B. Informationsdienste, Erprobung und Vereinheitlichung, Durchführbarkeitsstudien usw.) befassen,
- Mitarbeiter, die sich mit anderen industriellen Aktivitäten (z.B. industrielle Innovationen usw.) befassen.

Die Ergebnisse der Kostenstrukturerhebung des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden werden in der Fachserie 4, Reihe 4.3 veröffentlicht. Sie sind zu beziehen durch:

SFG-Servicecenter
Fachverlage GmbH
Postfach 43 43
72774 Reutlingen

Telefon (07071) 93 53 50
Telefax (07071) 3 36 53

A.4 Beschreibung der Kostenstrukturerhebung

STATISTISCHE ÄMTER DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Beschreibung der Kostenstrukturerhebung im Bergbau und Verarbeitenden Gewerbe – kleinere und mittlere Unternehmen

1. Ziele der Kostenstrukturerhebung

Inhaltlich liefern die Kostenstrukturerhebungen im Verarbeitenden Gewerbe (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. Die Kostenstrukturerhebung enthält zahlreiche Informationen zu Beschäftigtengrößen, Umsätzen und Kostenstruktur der Unternehmen (Die einzelnen in der Datengrundlage vorhandenen Variablen finden sich in Anhang A).

2. Stichprobenerhebung, Hochrechnung und Darstellung der Ergebnisse

Die Kostenstrukturerhebung im Verarbeitenden Gewerbe des Jahres 1999 – reduziert auf Unternehmen mit 20 bis einschließlich 249 Beschäftigten (kleinere und mittlere Unternehmen) – erfasst als hochrechnungsfähige Stichprobe von ca. 38% etwa 13 000 Unternehmen. Die Befragung erfolgt zentral durch das Statistische Bundesamt im Wege der Selbstausfüllung durch die Unternehmen. Die in der Stichprobe gewonnenen Ergebnisse werden auf die Gesamtheit der Unternehmen zwischen 20 und 249 Beschäftigten hochgerechnet. Die Stichprobe wird in der Regel alle 4 Jahre neu gezogen, so dass kleinere und mittlere Unternehmen durch Rotation entlastet werden können. Die Unternehmen der Kostenstrukturerhebung im Verarbeitenden Gewerbe des Jahres 1999 tragen in der Größenklasse „20-49 Beschäftigte" zu etwa 30.2% zur Gesamtzahl der Beschäftigten und zu 36.2% zum Gesamtumsatz aller Unternehmen dieser Größenklasse bei. Von der Größenklasse „50-99 (bzw. 100-249) Beschäftigte" werden etwa 43% (bzw. 56%) der Gesamtzahl der Beschäftigten und 49.2% (bzw. 60.7%) des Gesamtumsatzes erfasst.

Die Stichprobe lässt sich schematisch wie folgt darstellen:

Erhebungseinheit:	Unternehmen
Grundgesamtheit:	Unternehmen mit 20 bis einschließlich 249 Beschäftigten
Stichprobenumfang:	ca. 13 000 Unternehmen
Periodizität:	Jährlich
Art der Erhebung:	Geschichtete Zufallsstichprobe

Schichtungskriterien: Anzahl der Beschäftigten und wirtschaftliche Tätigkeit (Klassen der NACE Rev.1)

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 absolute Werte, und zwar so, dass die einzelnen Positionen von Jahr zu Jahr miteinander verglichen und die zwischenzeitlichen Veränderungen mit ausreichender Sicherheit festgestellt werden können. Die Ergebnisse werden nach der Klassifikation der Wirtschaftszweige (WZ 93) und Beschäftigtengrößenklassen dargestellt, wobei die Zuordnung der Unternehmen nach ihrem wirtschaftlichen Schwerpunkt erfolgt.

3. Inhalte der Erhebung

Zwischen den Leistungsgrößen, die jeweils als Restgrößen ermittelt werden, gelten folgende definitorische Beziehungen, die zugleich die Übergänge zu den Volkswirtschaftlichen Gesamtrechnungen erkennen lassen:

Gesamtumsatz

± Bestandsveränderungen an unfertigen und fertigen Erzeugnissen aus eigener Produktion

\+ Selbsterstellte Anlagen

= Bruttoproduktionswert (Gesamtleistung)

Bruttoproduktionswert

\- Materialverbrauch, Einsatz an Handelsware, Kosten für Lohnarbeiten

= Nettoproduktionswert

Nettoproduktionswert

\- Sonstige Vorleistungen

= Bruttowertschöpfung

Bruttowertschöpfung

- indirekte Steuern (ohne Umsatzsteuer) abz. Subventionen

= Bruttowertschöpfung zu Faktorkosten

Bruttowertschöpfung zu Faktorkosten

- Abschreibungen

= Nettowertschöpfung zu Faktorkosten

Der Nettoproduktionswert, eine in der Industriestatistik häufig verwendete Größe, entspricht dem Rohertrag in der betriebswirtschaftlichen Terminologie. Er unterscheidet sich vom Census Value Added der internationalen Industriestatistik insofern, als er noch die Kosten für Reparaturen, Instandhaltungen, Installationen, Montagen u.ä. enthält.

Die Bruttowertschöpfung umfasst – nach Abzug sämtlicher Vorleistungen – die insgesamt produzierten Güter und Dienstleistungen zu den am Markt erzielten Preisen und ist somit der Wert, der den Vorleistungen durch Bearbeitung hinzugefügt worden ist.

Die Nettowertschöpfung zu Faktorkosten dient zur Entlohnung der im Produktionsprozess eingesetzten Produktionsfaktoren. Sie stellt das Einkommen der Produktionsfaktoren nach Erhaltung des realen Vermögensbestandes, d.h. nach Abzug der Abschreibungen, nach Abführung der indirekten Steuern an den Staat und nach Berücksichtigung der vom Staat gewährten Subventionen dar. Sie verteilt sich auf:

- das Bruttoeinkommen aus unselbständiger Arbeit,
- die Fremdkapitalzinsen,
- die Grundrente,
- das Unternehmereinkommen.

Die Differenz zu den in der Kostenstrukturerhebung erfassten Positionen „Bruttoeinkommen aus unselbständiger Arbeit" und „Fremdkapitalzinsen" umfasst deshalb als Restgröße neben der Grundrente für den Produktionsfaktor Boden das Unternehmereinkommen.

Eine graphische Darstellung der Ableitung der oben beschriebenen Leistungsgrößen findet sich in der Datei „Leistungsgrößen.pdf". Die dort auftauchenden Zahlen beziehen sich auf alle in der Kostenstrukturerhebung 1999 erfassten Einheiten (inklusive Unternehmen mit 250 Beschäftigten und mehr).

Die Merkmale der Kostenstrukturerhebung sind:

Nr Laufende Nummer
ef1. Wirtschaftsabteilung (WZ93 auf Zweistellerebene)
ef2. Regionalbezug (Ost-West Klassifizierung)
ef3. Beschäftigtengrößenklasse
07 = 20 –49
08 = 50 – 99
11 = 100 – 249
14 = 250 – 499
17 = 500 – 999
22 = 1000 und mehr
ef4. Teilzeitbeschäftigte
ef5. Teilzeitbeschäftigte umgerechnet in Vollzeiteinheiten
ef6. Tätige Personen insgesamt
ef7. Umsatz aus eigenen Erzeugnissen
ef8. Umsatz aus Handelsware
ef9. Gesamtumsatz (entspricht nicht der Summe aus 7. und 8.)
ef10. Gesamtleistung/Bruttoproduktionswert
ef11. Anfangsbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen am Umsatz aus eigenen Erzeugnissen
ef12. Endbestand an unfertigen und fertigen Erzeugnissen aus eigener Produktion gemessen am Umsatz aus eigenen Erzeugnissen
ef13. Anfangsbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und Betriebsstoffen gemessen am Umsatz aus eigenen Erzeugnissen
ef14. Endbestand an Rohstoffen und sonstigen fremdbezogenen Vorprodukten, Hilfs- und Betriebsstoffen gemessen am Umsatz aus eigenen Erzeugnissen
ef15. Verbrauch an Rohstoffen
ef16. Energieverbrauch
ef17. Anfangsbestand an Handelsware gemessen am Umsatz aus Handelsware
ef18. Endbestand an Handelsware gemessen am Umsatz aus Handelsware
ef19. Einsatz an Handelsware
ef20. Bruttogehalts- und -lohnsumme
ef21. Gesetzliche Sozialkosten
ef22. Sonstige Sozialkosten
ef23. Kosten für Leiharbeitnehmer
ef24. Kosten für Lohnarbeiten
ef25. Kosten für Reparaturen
ef26. Mieten und Pachten
ef27. Sonstige Kosten
ef28. Fremdkapitalzinsen
ef29. Kosten insgesamt
ef30. Bruttowertschöpfung zu Faktorkosten
ef31. Nettowertschöpfung zu Faktorkosten
ef32. Gesamtaufwendungen für innerbetriebliche Forschung und Entwicklung
ef33. Anzahl der für Forschung und Entwicklung eingesetzten Lohn- und Gehaltsempfänger

A.5 Randauszählungen

STATISTISCHE ÄMTER
DES BUNDES UND DER LÄNDER
FORSCHUNGSDATENZENTREN

Randauszählungen

Die nachfolgende Tabelle 1 zeigt, wie sich die Unternehmen in den originalen, nicht anonymisierten Daten auf Wirtschaftsabteilungen und Beschäftigtengrößenklassen verteilen. Im Vergleich mit Tabelle 2 fällt auf, dass sich nur geringfügige Verschiebungen nach der Anonymisierung der Daten ergeben haben.

Tabelle 1: Verteilung der Unternehmen auf Wirtschaftsabteilungen und Beschäftigtengrößenklassen - Originaldaten

Wirtschaftsgliederung	WZ93-Angabe	20-49	50-99	100-249	Gesamtanzahl
Bergbau und Gew. v. Steinen u. Erden	C (10, 11 und 14)	104	52	26	182
Ernährungsgewerbe u. Tabakverarbeitung	DA (15 und 16)	662	539	503	1 704
Textil- u. Bekleidungsgewerbe	DB (17 und 18)	411	298	246	955
Holzgewerbe (oh. H. v. Möbeln)	20	223	141	84	448
Papier-, Verlags- u. Druckgewerbe	DE (21 und 22)	390	327	346	1 063
Chemische Industrie	DG (bzw. 24)	212	203	213	628
H. v. Gummi- u. Kunststoffwaren	DH (bzw. 25)	247	203	209	659
Glasgewerbe, Keramik, Ver. v. Steinen u. Erden	DI (bzw. 26)	350	228	205	783
Metallerz g. u. -bearbeitung	27	133	126	164	423
H. v. Metallerzeugnissen	28	659	510	469	1 638
Maschinenbau	DK (bzw. 29)	727	551	559	1 837
H. v. Büromasch., Dv-Geät. u. einr., Geät. d. Elektriz.erzg.- verteilung u. ä.	30 und 31	270	270	217	757
Rundfunk-, Fernseh- u. Nachrichtentechnik	32	83	74	62	219
Medizin-, Mess-, Steuer- u. Regelungstechnik, Optik	33	207	168	135	510
Fahrzeugbau	DM (34 und 35)	195	155	175	525
H.v.Möbeln,Schmuck,Musikinstr., Sportger. usw	36	301	201	229	731
Sonstige: Ledergewerbe; Kokerei, Mineralölverarbeitung, H. v. Brutstoffen	19 und 23	59	54	51	164

Tabelle 2: Verteilung der Unternehmen auf Wirtschaftsabteilungen und Beschäftigtengrößenklassen-Anonymisierte Daten

Wirtschaftsgliederung	WZ93-Angabe	20-49	50-99	100-249	Gesamtanzahl
Bergbau und Gew. v. Steinen u. Erden	C (10, 11 und 14)	104	52	26	182
Ernährungsgewerbe u. Tabakverarbeitung	DA (15 und 16)	662	539	503	1 704
Textil- u. Bekleidungsgewerbe	DB (17 und 18)	411	298	246	955
Holzgewerbe (oh. H. v. Möbeln)	20	223	141	84	448
Papier-, Verlags- u. Druckgewerbe	DE (21 und 22)	390	327	346	1 063
Chemische Industrie	DG (bzw. 24)	213	202	213	628
H. v. Gummi- u. Kunststoffwaren	DH (bzw. 25)	247	203	209	659
Glasgewerbe, Keramik, Ver. v. Steinen u. Erden	DI (bzw. 26)	350	228	205	783
Metallerz g. u. -bearbeitung	27	133	126	164	423
H. v. Metallerzeugnissen	28	659	510	469	1 638
Maschinenbau	DK (bzw. 29)	727	551	559	1 837
H. v. Büromasch., Dv-Geät. u. einr., Geät. d. Elektriz.erzg.- verteilung u. ä.	30 und 31	270	269	218	757
Rundfunk-, Fernseh- u. Nachrichtentechnik	32	83	74	62	219
Medizin-, Mess-, Steuer- u. Regelungstechnik, Optik	33	207	168	135	510
Fahrzeugbau	DM (34 und 35)	195	155	175	525
H.v.Möbeln,Schmuck,Musikinstr., Sportger. usw	36	301	201	229	731
Sonstige: Ledergewerbe; Kokerei, Mineralölverarbeitung, H. v. Brutstoffen	19 und 23	59	54	51	164

Die nachfolgende Tabelle 3 zeigt, wie sich Gesamtumsatz und Beschäftigte auf die Wirtschaftsabteilungen in den originalen, nicht anonymisierten Daten verteilen. Auch hier fallen im Vergleich mit Tabelle 4 nur geringfügige Änderungen infolge der Anonymisierung auf.

Tabelle 3: Verteilung von Gesamtumsatz und Beschäftigten auf die Wirtschaftsabteilungen-Originaldaten

Wirtschaftsgliederung	WZ93-Angabe	Anzahl	Beschäftigte[1]		Gesamtumsatz[2]	
			%	Anzahl	%	1 000 DM
Bergbau und Gew. v. Steinen u. Erden	C (10, 11 und 14)	182	*1.01*	10 804	*1.17*	3 683 861
Ernährungsgewerbe u. Tabakverarbeitung	DA (15 und 16)	1704	*12.90*	137 280	*20.70*	65 061 771
Textil- u. Bekleidungsgewerbe	DB (17 und 18)	955	*6.91*	73 528	*5.91*	18 585 323
Holzgewerbe (ohne H. v. Möbeln)	20	448	*2.81*	29 923	*2.61*	8 206 498
Papier-, Verlags- u. Druckgewerbe	DE (21 und 22)	1063	*8.49*	90 314	*8.40*	26 421 275
Chemische Industrie	DG (bzw. 24)	628	*5.06*	53 870	*7.84*	24 641 276
H. v. Gummi- u. Kunststoffwaren	DH (bzw. 25)	659	*5.12*	54 528	*4.24*	13 319 292
Glasgewerbe, Keramik, Verarb. v. Steinen u. Erden	DI (bzw. 26)	783	*5.47*	58 190	*5.45*	17 118 607
Metallerzeug. u. -bearbeitung	27	423	*3.79*	39 719	*4.46*	14 004 819
H. v. Metallerzeugnissen	28	1638	*11.94*	127 061	*9.12*	28 663 266
Maschinenbau	DK (bzw. 29)	1837	*14.07*	149 729	*11.57*	36 358 168
H. v. Büromasch., Dv-Geräten u. -einr., Geräten d. Elektriz.erzeug., -verteilung u. ä.	30 und 31	757	*5.88*	62 597	*5.05*	15 864 625
Rundfunk-, Fernseh- und Nachrichtentechnik	32	219	*1.69*	17 954	*1.39*	4 369 869
Medizin-, Mess-, Steuer- und Regelungstechnik, Optik	33	510	*3.71*	39 513	*2.76*	8 671 004
Fahrzeugbau	DM (34 und 35)	525	*4.24*	45 115	*3.85*	12 101 895
H.v. Möbeln, Schmuck, Musikinstr., Sportger. usw	36	731	*5.69*	60 515	*4.01*	12 610 642
Sonstige Ledergewerbe; Kokerei, Mineralölverarbeitung, H. v. Brutstoffen	19 und 23	164	*1.26*	13 364	*1.45*	4 574 230

1) Alle am Jahresende im Betrieb /im Unternehmen tätigen Personen einschl. tätiger Inhaber/Inhaberinnen und mithelfender Familienangehöriger (auch unbezahlt mithelfende Familienangehörige, soweit sie mindestens ein Drittel der üblichen Arbeitszeit im Betrieb tätig sind), aber ohne Heimarbeiter/Heimarbeiterinnen und Zusteller/Zustellerinnen im Verlagsgewerbe. Einbezogen werden u.a. auch Erkrankte, Urlauber/ -innen, Kurzarbeiter/Kurzarbeiterinnen, Streikende, von der Aussperrung Betroffene, Personen in Altersteilzeitregelungen, Auszubildende, Saison- und Aushilfsarbeiter/ -arbeiterinnen sowie Teilzeitbeschäftigte. Die Angestellten umfassen auch die kaufmännischen Auszubildenden (einschl. der Auszubildenden in den übrigen nichtgewerblichen Ausbildungsberufen), die Arbeiter/Arbeiterinnen auch die gewerblich Auszubildenden.

2) Umsatz aus eigener Erzeugung (einschl. Umsatz aus dem Verkauf von Energie und Nebenerzeugnissen und Abfällen sowie Entgelte für industrielle Dienstleistungen, wie Reparaturen, Instandhaltungen, Installationen und Montagen), Umsatz aus Handelsware und sonstigen nichtindustriellen/nichthandwerklichen Tätigkeiten (z.B. Erlöse aus Vermietung und Verpachtung sowie aus Lizenzverträgen, Provisionseinnahmen und Einnahmen aus der Veräußerung von Patenten).

Tabelle 4: Verteilung von Gesamtumsatz und Beschäftigten auf die Wirtschaftsabteilungen-Anonymisierte Daten

Wirtschaftsgliederung	WZ93-Angabe	Anzahl	Beschäftigte		Gesamtumsatz	
			%	Anzahl	%	1 000 DM
Bergbau und Gew. v. Steinen u. Erden	C (10, 11 und 14)	182	*1.02*	10 804	*1.17*	3 683 797
Ernährungsgewerbe u. Tabakverarbeitung	DA (15 und 16)	1704	*12.90*	137 284	*20.70*	65 056 450
Textil- u. Bekleidungsgewerbe	DB (17 und 18)	955	*6.91*	73 529	*5.91*	18 585 363
Holzgewerbe (ohne H. v. Möbeln)	20	448	*2.81*	29 925	*2.61*	8 206 696
Papier-, Verlags- u. Druckgewerbe	DE (21 und 22)	1063	*8.49*	90 311	*8.41*	26 421 392
Chemische Industrie	DG (bzw. 24)	628	*5.06*	53 873	*7.84*	24 640 473
H. v. Gummi- u. Kunststoffwaren	DH (bzw. 25)	659	*5.12*	54 525	*4.24*	13 318 912
Glasgewerbe, Keramik, Verarb. v. Steinen u. Erden	DI (bzw. 26)	783	*5.47*	58 188	*5.45*	17 118 619
Metallerzeug. u. -bearbeitung	27	423	*3.79*	39 720	*4.46*	14 004 740
H. v. Metallerzeugnissen	28	1638	*11.94*	127 055	*9.12*	28 663 369
Maschinenbau	DK (bzw. 29)	1837	*14.07*	149 732	*11.57*	36 357 908
H. v. Büromasch., Dv-Geräten u. -einr., Geräten d. Elektriz.erzeug., -verteilung u. ä.	30 und 31	757	*5.88*	62 596	*5.05*	15 865 903
Rundfunk-, Fernseh- und Nachrichtentechnik	32	219	*1.69*	17 955	*1.39*	4 370 507
Medizin-, Mess-, Steuer- und Regelungstechnik, Optik	33	510	*3.71*	39 513	*2.76*	8 671 100
Fahrzeugbau	DM (34 und 35)	525	*4.24*	45 114	*3.85*	12 101 547
H.v. Möbeln, Schmuck, Musikinstr., Sportger. usw	36	731	*5.69*	60 516	*4.01*	12 610 384
Sonstige Ledergewerbe; Kokerei, Mineralölverarbeitung, H. v. Brutstoffen	19 und 23	164	*1.26*	13 363	*1.46*	4 574 646

Literaturverzeichnis

Aëtios: ξυναγωγή περὶ τῶν αρεσκόντων (über die Sätze der Naturlehre). Ca. 100 n. Chr. IV, 19, 3. In: Kirk und Raven (1959).

Bacher, S. (1994): Clusteranalyse. *Oldenbourg-Verlag.*

Bender, S., Wagner, J. und Zwick, M. (2007): KombiFiD – Kombinierte Firmendaten für Deutschland. *Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Arbeitspapier Nr. 21.*

Bertsekas, D. P. (1979): A distributed algorithm for the assignment problem. *Lab. for Information and Decision Systems,* Working Paper, MIT Press.

Bertsekas, D. P. und Castanon, D. A. (1989): The auction algorithm for transportation problems. *Annals of Operations Research 20,* S. 67–96.

Birkhoff, G. (1940): Lattice theory. *American Mathematical Society, Colloq. Publ., vol. 25.*

Borgwardt, K. H. (1982): The Average Number of Pivot Steps Required by the Simplex-Method is Polynomial. *Zeitschrift für Operations Research 7 (3),* S. 157–177.

Borgwardt, K. H. (1987): The simplex method. A probabilistic Analysis. *Springer-Verlag, Heidelberg.*

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

Brandt, M. und Zwick, M. (2009): An informational infrastructure for the E-Science Age – On the way to remote data access. *Conference "New Techniques and Technologies for Statistics (NTTS 2009)"*, Brüssel.

Brandt, M., Lenz, R., Hafner, H.-P. und Schmidt, D. (2006): Scientific-Use-File und Analysen auf Basis der europäischen Erhebung zur betrieblichen Weiterbildung (CVTS2). *Erfahrungen und Perspektiven, DRV-Schriften Band 55/2006,* S. 116–132.

Brandt, M., Lenz, R. und Rosemann, M. (2007): Analytical validity and confidentiality protection of anonymised longitudinal enterprise microdata – Survey of a German project. *Eurostat/European Commission: Methodologies and Working Papers* , 2009 edition, S. 340–365 (Work session on statistical data confidentiality, Manchester 12/2007).

Brandt, M., Lenz, R. und Rosemann, M. (2008a): Anonymisation of panel enterprise microdata – Survey of a German Project. In: *Domingo-Ferrer, J., Saygin, Y. (Hrsg.): Privacy in Statistical Databases, Lecture Notes in Computer Science, vol. 5262, Springer-Verlag,* Heidelberg, S. 139–151.

Brandt, M., Dittrich, S. und Konold, M. (2008b): Wirtschaftsstatistische Längsschnittsdaten für die Wissenschaft. *Wirtschaft und Statistik 3/2008.*

Brandt, M., Oberschachtsiek, D. und Pohl, R. (2008c): Neue Datenangebote in den Forschungsdatenzentren – Betriebs- und Unternehmensdaten im Längsschnitt. *Wirtschafts- und Sozialstatistisches Archiv, Band 2, Heft 3,* S. 193–207.

Cohen, W. W., Ravikumar, P. und Fienberg, S. E. (2003): A comparison of string distance metrics for name-matching tasks. *Proceedings of the IJCAI-2003,* (siehe citeseer.ist.psu.edu).

Cormen, T. H., Leiserson, C. E., Rivest, R. L. and Stein, C. (1990): Introduction to Algorithms. *Cambridge, Mass., MIT Press.*

Creditreform (2009): MARKUS (Marketinguntersuchungen) – Die große Datenbank deutscher und österreichischer Unternehmen. *Bureau van Dijk, Electronic Publishing,* (siehe www.creditreform.com).

Dalenius, T. und Reiss, S. (1982): Data-swapping: A technique for disclosure control. *Journal of Statistical Planning and Inference 6,* S. 73–85.

Dandekar, R., Cohen, M. und Kirkendall, N. (2001): Applicability of latin hypercube sampling to create multi variate synthetic micro data. *Proceedings of the conference "New Techniques and Technologies for Statistics (NTTS 2001)"*, S. 839–847.

Dempster, A. P., Laird, N. M. und Rubin, D. B. (1971): Maximum Likelihood From Incomplete Data Via the EM Algorithm. *Journal of the Royal Statistical Society B, 39,* S. 1-38.

Dittrich, S. (2004): Umsätze und ihre Besteuerung. *Wirtschaft und Statistik 10/2004,* S. 1195–1200.

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.

Domingo-Ferrer, J. und Mateo-Sanz, J. M. (2002): Practical data-oriented microaggregation for statistical disclosure control. *IEEE Transactions on Knowledge and Data Engineering, 39 (1),* S. 189–201.

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. Forum der Bandesstatistik, Band 42,* Wiesbaden, S. 100–110.

Domingo-Ferrer, J., Sebé, F. und Solanas, A. (2008): A Polynomial-Time Approximation to Optimal Multivariate Microaggregation. Erscheint in: *Computers and Mathematics with Applications.*

Duncan, G. und Lambert, D. (1989): The risk of disclosure for microdata. *Journal of Business & Economic Statistics 7(2),* S. 207–217.

Efelky, M., Verykios, V. und Elmagarmid, A. (2002): TAILOR: A Record Linkage Toolbox. *Proc. of the 18th Int Conf. on Data Engineering,* San Jose, California.

Egner, U. (2002): Berufliche Weiterbildung in Unternehmen (CVTS2). Statistisches Bundesamt, Wiesbaden, *Projektbericht.*

Elamir, E. und Skinner, C. (2006): Record-level Measures of Disclosure Risk for Survey Microdata. *Journal of Official Statistics 22 (3),* S. 525–539.

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

Eurostat (2002): Europäische Sozialstatistik. Erhebung über die betriebliche Weiterbildung (CVTS2).

Evers, K. und Höhne, J. (1999): SAFE – ein Verfahren zur Anonymisierung und statistischen Geheimhaltung wirtschaftsstatistischer Einzeldaten. *Statistisches Bundesamt (Hrsg.): Spektrum der Bandesstatistik, Band 14, Wiesbaden,* S. 136–147.

Fellegi, I. P. und Sunter, A. P. (1969): A Theory for Record Linkage. *Journal of the American Statistical Association, 64,* S. 1183–1210.

Fienberg, S. E. (1997): Confidentiality and disclosure limitation methodology: Challenges for national statistics and statistical research. *Technical report No. 161,* Carnegie Mellon University, Pittsburgh.

Frank-Bosch, B. (2003): Verdienststrukturen in Deutschland: Methode und Ergebnisse der Gehalts- und Lohnstrukturerhebung 2001. *Wirtschaft und Statistik 12/2003,* S. 1137–1151.

Fritzsch, M. und Stephan, A. (2003): Die Heterogenität der technischen Effizienz innerhalb von Wirtschaftszweigen – Auswertungen auf Grundlage der Kostenstrukturstatistik des Statistischen Bundesamtes. *In: Pohl, Ramona; Joachim Fischer; Ulrike Rockmann und Klaus Semlinger (Hrsg.): Analysen zur regionalen Industrieentwicklung – Sonderauswertungen einzelbetrieblicher Daten der Amtlichen Statistik,* Statistisches Landesamt Berlin, S. 143–156.

Fritzsch, M. und Stephan, A. (2007): Die Heterogenität der Effizienz innerhalb von Branchen – Eine Auswertung von Unternehmensdaten der Kostenstrukturerhebung im Verarbeitenden Gewerbe. *Vierteljahreshefte zur Wirtschaftsforschung 3/2007,* S. 59–75.

Fürnrohr, M., Rimmelspacher, B. und Roncador v., T. (2002): Zusammenführung von Datenbeständen ohne numerische Identifikatoren. *Bayern in Zahlen,* S. 308–321.

Gesetz über die Statistik für Bundeszwecke (Bundesstatistikgesetz – BStatG). *Bundesgesetzblatt (BGBl.) I S. 1534.*

Gottschalk, S. (2004): Microdata disclosure control by resampling – empirical findings for business survey data. *Allgmeines Statistisches Archiv 88(3),* S. 279–302.

Görzig, B., Kaminiarz, A. und Stephan, A. (2005): Wie wirkt sich Outcourcing auf den Unternehmenserfolg aus? Neue Evidenz. *Journal of Applied Social Science Studies (Schmollers Jahrbuch) 125 (4),* S. 489–507.

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.

Hafner, H.-P. (2009): Integrated data sources and methodological issues. Case studies from the German AFID project. *Conference "New Techniques and Technologies for Statistics (NTTS 2009)"*, Brüssel.

Hafner, H.-P. und Lenz, R. (2007): Anonymisation of Linked Employer Employee Datasets using the example of the German Structure of Earnings Survey. *Eurostat/European Commission: Methodologies and Working Papers* , 2009 edition, S. 96–106 (Work session on statistical data confidentiality, Manchester 12/2007).

Hafner, H.-P. und Lenz, R. (2008): The German Structure of Earnings Survey: Methodology, data access and research potential. *Journal of Applied Social Science Studies (Schmollers Jahrbuch) 128 (3),* S. 489–500.

Hafner, H.-P. und Lenz, R. (2009): Synthetic Data Structure Files: Development and Disclosure Control. *Invited Paper, Joint UN/ECE Work session on statistical data confidentiality* , Bilbao 12/2009.

Hafner, H.-P. und Lenz, R. (2010): Erzeugung synthetischer Datensätze mit Methoden der multiplen Imputation. *14. Konferenz der SAS-Anwender in Forschung und Entwicklung,* Berlin 2/2010.

Hafner, H.-P., Lenz, R. und Mischler, F. (2007): Einzeldaten der Gehalts- und Lohnstrukturerhebung 2001 als Scientific-Use-File. *Wirtschaft und Statistik 2/2007,* S. 144–149.

Helmcke, T. und Knoche, P. (1992): Projekt zur faktischen Anonymität von Mikrodaten – Bericht über ein Forschungsprojekt. *Wirtschaft und Statistik*, S. 139–144.

Hernandez, M. und Stolfo, S. (1998): Real World Data is Dirty: Data Cleansing and the Merge/Purge Problem. *Journal of Data Mining and Knowledge Discovery 2 (1),* S. 9–37.

Höhne, J. (2003a): Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten. *In: Gnoss, R. und Ronning, G. (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten. Forum der Bundesstatistik, Band 42, Wiesbaden,* S. 69–94.

Höhne, J. (2003b): SAFE - a method for statistical disclosure limitation of micro data. *Contributed Paper for the UNECE/Eurostat work session on statistical data confidentiality,* Luxembourg.

Höhne, J. (2003c): SAFE - Ein Verfahren zur Anonymisierung statistischer Einzelangaben. *Statistisches Landesamt Berlin (Hrsg.): Sonderdruck, Statistische Monatsschrift, Nr. 3/2003,* Berliner Statistik.

Höhne, J. (2008): Anonymisierungsverfahren für Paneldaten. *Wirtschafts- und Sozialstatistisches Archiv, Band 2, Heft 3,* S. 259–275.

Höhne, J. (2010): Verfahren zur Anonymisierung von Einzeldaten. *Statistisches Bundesamt (Hrsg.): Statistik und Wissenschaft,* Band 16, Wiesbaden.

Höhne, J., Sturm, R. und Vorgrimler, D. (2003): Konzept zur Beurteilung der Schutzwirkung von faktischer Anonymisierung. *Wirtschaft und Statistik, Band 4,* S. 287–299.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Gießing, S., Lenz, R., Longhurst, J., Schulte Nordholt, E., Seri, G. und De Wolf, P. (2007): Handbook on Statistical Disclosure Control, Version 1.1.0 (siehe www.neon.vb.cbs.nl).

Iserman, H. (1974): Proper efficiency and the linear vector maximum problem. *Operations Research 22*, S. 189–191.

Jaro, M. A. (1989): Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association, 89,* S. 415–435.

Kadane, J. B. (2001): Some Statistical Problems in Merging Data Files. *Journal of Official Statistics, 17 (3),* S. 423–433.

Kim, J. (1986): A method for limiting disclosure in microdata based on random noise and transformation.*American Statistical Association (Hrsg.): Proceedings of the Section on Survey Research*, S. 370–374.

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

Kim, J. und Winkler, W. (2001): Multiplicative noise for masking continuous data. *American Statistical Association (Hrsg.): Proceedings of the Section on Survey Research Methods.*

Kirk, G. S. und Raven, J. E. (1959): The presocratic philosophers. *Cambridge university press.*

Konold, M. und L'Assainato, S. (2009): Matching business data from different sources: The case of the KombiFiD-project in Germany. *Conference "New Techniques and Technologies for Statistics (NTTS 2009)"*, Brüssel.

Kooiman, P., Willenborg, L. und Gouweleeuw, J. (1997): Pram: a method for disclosure limitation of micro data. *Department of Statistical Methods, Statistics Netherlands,* Voorburg.

Krämer, W., Schoffer, O. und Tschiersch, L. (2008): Datenanalyse mit SAS – Statistische Verfahren und ihre grafischen Aspekte. *Springer-Verlag Berlin Heidelberg, 2. überarbeitete Auflage.*

Kuhn, H. W. (1955): The hungarian method for the assignment problem. *Naval Res. Logist. Quart. 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 and Statistik, *Nomos Verlagsgesellschaft,* Baden-Baden, ISBN: 3-7890-7388-1.

Lechner, S. und Pohlmeier, W. (2003): Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten.*In: Gnoss, R. und G.Ronning (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten. Forum der Bundesstatistik, Band 42,* Wiesbaden, S. 115–137.

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, CD-ROM Publikation.

Lenz, R. (2003b): A graph theoretical approach to record linkage. *Monographs of Official Statistics – Research in Official Statistics,* S. 324–334 (vorgestellt auf der UNECE/Eurostat work session on statistical data confidentiality, Luxembourg.)

Lenz, R. (2003c): A way to combine probabilistic with deterministic record linkage. *Proceedings of the Workshop on Microdata,* Stockholm.

Lenz, R. (2005a): Measuring the disclosure risk of masked enterprise microdata. *Invited paper, vorgestellt auf der ONS/BDL Analysis of Enterprise Microdata Conference (CAED 2005),* Cardiff, CD-ROM Publikation.

Lenz, R. (2005b): Messung der Schutzwirkung zufällig überlagerter Einheiten der KSE. Arbeitspapier zur Projektgruppensitzung „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" vom 14.01.2005.

Lenz, R. (2006a): Measuring the disclosure protection of micro aggregated business microdata – an analysis taking as an example the German Structure of Costs Survey. *Journal of Official Statistics 22 (4),* S. 681–710.

Lenz, R. (2006b): Measuring the anonymity of longitudinal linked economic statistics microdata. *Methods ... Approaches ... Developments 2,* S. 3–6.

Lenz, R. (2008): Risk Assessment Methodology for Longitudinal Business Micro Data. *Wirtschafts- und Sozialstatistisches Archiv, Band 2, Heft 3,* S. 241–257.

Lenz, R. (2009): Défis méthodiques lors de la réalisation de l'accès aux données économiques allemandes par la téléinformatique automatisée. 41ème Journées de Statistique de la Société Française de Statistique (SFdS), Université Victor Segalen, Bordeaux, 23.-30. Mai 2009.

Lenz, R. und Hafner, H.-P. (2006a): Anonymisation of Linked Employer Employee Datasets. *Privacy in Statistical Databases Conference PSD 2006,* 13.-15. december 2006, Rome, ISBN 84-690-2100-1.

Lenz, R. und Hafner, H.-P. (2006b): Scientific analyses using the Continuing Vocational Training Survey 2000 (CVTS 2). *Proceedings of the European Conference on Quality and Methodology in Survey Statistics (Q2006),* Cardiff.

Lenz, R. und Scheffler, M. (2004): Datenangriffe auf die Einzelhandelsstatistik. Arbeitspapier zur Projektgruppensitzung „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" vom 12.07.2004.

Lenz, R. und Vorgrimler, D. (2004): Geheimhaltungsmethoden auf dem Prüfstand – eine Analyse anhand der Umsatzsteuerstatistik. *Wirtschaft und Statistik 6/2004,* S. 639–648.

Lenz, R. und Vorgrimler, D. (2005): Matching german turnover tax statistics. *Forschungsdatenzentren der Statistischen Ämter des Bundes und der Länder – Arbeitspapier Nr. 4* (vorgestellt auf der Privacy in Statistical Databases Conference PSD'2004, Barcelona.)

Lenz, R. und Zwick, M. (2005): Integrierte Mikrodatenfiles – Methoden zur Verknüpfung von Einzeldaten. *Statistik und Wissenschaft, Neue Wege statistischer Berichterstattung - Mikro- und Makrodaten als Grundlage sozioökonomischer Modellierungen, Band 10,* S. 97–103.

Lenz, R. und Zwick, M. (2009a): Methodological aspects assuring remote access to German business microdata. Erscheint in: *Bulletin of the* 60^{th} *International Statistical Institute (ISI 2009),* Durban.

Lenz, R. und Zwick, M. (2009b): German Business Microdata: Linkage and Anonymization. *Journal of Applied Social Science Studies (Schmollers Jahrbuch) 129 (4)* , S.645–653.

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, CD-ROM Publikation.

Lenz, R., Sturm, R. und Vorgrimler, D. (2004b): Maße für die faktische Anonymität von Mikrodaten. *Wirtschaft und Statistik 6/2004,* S. 621–638.

Lenz, R., Vorgrimler, D. und Rosemann, M. (2005a): Ein Scientific-Use-File der Kostenstrukturerhebung im Verarbeitenden Gewerbe. *Wirtschaft und Statistik 2/2005,* S. 91–96.

Lenz, R., Vorgrimler, D. und Scheffler, M. (2005b): A standard for the release of business microdata. *Monographs of Official Statistics – Research in Official Statistics.* (Invited paper, vorgestellt auf der UN-ECE/Eurostat work session on statistical data confidentiality, 9.-11. November, Genf 2005), S. 197–206.

Lenz, R., Hafner, H.-P. und Schmidt, D. (2006a): Daten zur betrieblichen Weiterbildung für die Wissenschaft. *Wirtschaft und Statistik 5/2006,* S. 201–210.

Lenz, R., Rosemann, M., Vorgrimler, D. und Sturm, R. (2006b): Anonymising business micro data – results of a German project. *Journal of Applied Social Science Studies (Schmollers Jahrbuch) 126 (4),* S. 635–651.

Lenz, R., Hafner, H.-P. und Schmidt, D. (2006c): Daten für wissenschaftliche Analysen zur betrieblichen Weiterbildung in Unternehmen. *Staat und Wirtschaft in Hessen, Heft 6/2006,* 61. Jahrgang, S. 158–163.

Little, R. (1993): Statistical Analysis of Masked Data. *Journal of Official Statistics vol. 9,* S. 407–426.

Malchin, A. und Pohl, R. (2007): Firmendaten der amtlichen Statistik – Datenzugang und neue Entwicklungen in den Forschungsdatenzentren. *Vierteljahreshefte zur Wirtschaftsforschung 3/2007.*

Mateo-Sanz, J. und Domingo-Ferrer, J. (1998a): A method for data-oriented multivariate microaggregation.*Proceedings of the conference on statistical data protection*, Eurostat 1999.

Munkres, J. (1957): Algorithms for the assignment and transportation problem. *L. Soc. Indust. Appl. Math., 5,* S. 32–38.

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

Nestler, K. und Kailis, E. (2002): Betriebliche Weiterbildung in der Europäischen Union und Norwegen (-CVTS2-). In: *Statistik kurz gefasst – Bevölkerung und Soziale Bedingungen, 3/2002,* S. 1–7.

Pagliuca, D. und Seri, G. (1976): Some results of Individual Ranking Method on the System of Enterprise Accounts Annual Survey. *Esprit SDC Project, Deliverable MI-3/D2.*

Papadimitriou, C. H. und Steiglitz, K. (1998): Combinatorial Optimization. *Dover Publ.,* Mineola, New York.

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, Lecture Notes in Computer Science.*

Porter, E. H. und Winkler, W. (1999): Approximate String Comparison and its Effect on an Advanced Record Linkage System. Record Linkage Techniques – 1997. *National Academy Press,* Washington DC, S. 190–199.

Reiter, J. und Drechsler, J. (2007): Releasing Multiply – Imputed Synthetic Data Generated in Two Steps to Protect Confidentiality. *IAB Discussion Paper No. 20/2007.*

Ronning, G. (2004a): Fehlklassifikation im Modell der Varianzanalyse. *Arbeitspapier im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten".*

Ronning, G. (2004b): Mischung von Verteilungen und Anonymisierung. *Arbeitspapier im Projekt „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten".*

Ronning, G. (2005): Randomized response and the binary probit model. *Economics letters 86 (2),* S. 221–228.

Ronning, G., Lenz, R., Bender, S., Biewen, E., Brandt, M., Drechsler, J., Höhne, J. und Rosemann, M. (2009): Wirtschaftsstatistische Paneldaten und faktische Anonymisierung. Abschlussbericht des gleichnamigen Forschungsprojektes, Statistisches Bundesamt, Wiesbaden.

Ronning, G. und 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., Rosemann, M. und Strotmann, H. (2005): Post-Randomization Under Test: Estimation of the Probit Model. *Jahrbücher der Nationalökonomie und Statistik 225 (5), Lucius & Lucius Verlag,* Stuttgart, S. 544–566.

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M. und Vorgrimler, D. (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten. *Statistik und Wissenschaft, Band 4,* 632 Seiten. (Buchbesprechung in: *Allgemeines Statistisches Archiv 90,* S. 487-489.)

Roque, G. (2000): Masking Microdata Files with Mixtures of Multivariate Normal Distributions. Dissertationsschrift, University of California, Riverside.

Rosemann, M. (2003): Erste Ergebnisse von vergleichenden Untersuchungen mit anonymisierten and nicht anonymisierten Einzeldaten am Beispiel der Kostenstrukturerhebung and der Umsatzsteuerstatistik. *In: Gnoss, R. und Ronning, G. (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten. Forum der Bundesstatistik, Band 42,* Wiesbaden, S. 154–183.

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.

Rosemann, M. (2005): Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten. *IAW Forschungsbericht Nr. 66,* Tübingen, Dissertationsschrift.

Rosemann, M., Vorgrimler, D. und Lenz, R. (2004): Erste Ergebnisse faktischer Anonymisierung wirtschaftsstatistischer Einzeldaten. *Allgemeines Statistisches Archiv 88,* S. 73–99.

Rubin, D. (1993): Discussion: statistical disclosure limitation. *Journal of Official Statistics 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,* S. 585–598.

Scheffler, M. (2005): Ein Scientific-Use-File der Einzelhandelsstatistik. *Wirtschaft und Statistik 3/2005,* S. 196–200.

Schmid, M. (2006): Estimation of a linear model under microaggregation by individual ranking. *Allgemeines Statistisches Archiv 90 (3),* S. 419–438.

Schmidt, D. (2007): Berufliche Weiterbildung in Unternehmen 2005 – Methodik und erste Ergebnisse. *Wirtschaft und Statistik 7/2007,* S. 699–711.

Schweigert, D. (1995): Vector Weighted Matchings. In: Colbourn, C. J. und Mahmoodian, E. S. (Hrsg.): Combinatorial Advances, *Kluwer,* S. 267–276.

Schweigert, D. (1999): Order and Clustering. *Human Centered Processes,* 10^{th} Mini Euro Conference, Brest, S. 397–401.

Sextos: $\pi\rho\grave{o}\varsigma \quad \mu\alpha\theta\eta\mu\alpha\tau\iota\kappa o\acute{\upsilon}\varsigma$. Ca. 180-200 n. Chr., VII, 135. Siehe hierzu: Kirk und Raven (1959).

Skinner, C. J. und Elliot, M. J. (2002): A measure of disclosure risk for microdata. *Journal of the Royal Statistical Society B, (4) 64,* S. 855–867.

Statistische Ämter des Bundes und der Länder und IAW (2003): Forschungsprojekt "Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" – Zwischenbericht 2003 an das BMBF, Statistisches Bundesamt, Wiesbaden.

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

Statistisches Bundesamt (2005a): Kostenstruktur im Verarbeitenden Gewerbe, im Bergbau sowie in der Gewinnung von Steinen und Erden. Qualitätsbericht.

Statistisches Bundesamt (2005b): Umsatzsteuerstatistik. Qualitätsbericht.

Statistisches Bundesamt (2007a): Berufliche Weiterbildung in Unternehmen 2007. Qualitätsbericht.

Statistisches Bundesamt (2007b): Dritte Europäische Erhebung über die berufliche Weiterbildung in Unternehmen (CVTS3). (Enthält ergänzende Tabellen zu Statistisches Bundesamt 2007a.)

Statistisches Bundesamt (2007c): Monatsbericht für Betriebe des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden. Qualitätsbericht.

Statistisches Bundesamt (2007d): Investitionserhebung bei Unternehmen und Betrieben des Verarbeitenden Gewerbes sowie des Bergbaus und der Gewinnung von Steinen und Erden. Qualitätsbericht.

Sturm, R. (2002a): Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten. *Allgemeines Statistisches Archiv 86,* S. 468–477.

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

Sturm, R. und Lenz, R. (2005): Erste Scientific-Use-Files aus den Wirtschaftsstatistiken. In: *Amtliche Mikrodaten für die Sozial- und Wirtschaftswissenschaften, Beiträge zu den Nutzerkonferenzen des FDZ der Statistischen Landesämter,* Institut für Weltwirtschaft an der Universität Kiel, S. 191–207.

Vorgrimler, D. (2003): Reidentifikationsmöglichkeiten am Beispiel eines konkreten Angriffsszenarios. *In: Gnoss, R. und Ronning, G. (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten. Forum der Bundesstatistik, Band 42,* Wiesbaden, S. 40–59.

Vorgrimler, D. und Lenz, R. (2003a): Über das Risiko der Reidentifikation in wirtschaftsstatistischen Einzeldaten. *Wirtschaft und Statistik,* Sonderausgabe, S. 81–82.

Vorgrimler, D. und Lenz, R. (2003b): Disclosure Risk of Anonymized Business Microdata Files - Illustrated with Empirical Key Variables. *Bulletin of the 54th International Statistical Institute (ISI), 2,* S. 594–595.

Vorgrimler, D., Dittrich, S., Lenz, R. und Rosemann, M. (2005a): Ein Scientific-Use-File der Umsatzsteuerstatistik 2000. *Wirtschaft und Statistik 3/2005,* S. 201–210.

Vorgrimler, D., Dittrich, S., Lenz, R. und Rosemann, M. (2005b): Wissenschaftliche Analysen anhand der Umsatzsteuerstatistik. *Wirtschaftswissenschaftliches Studium, Heft 10/2005,* S. 327–332.

Willenborg, L. und de Waal, T. (2001): Elements of statistical disclosure control. *Springer-Verlag, Lecture Notes in Statistics 155.*

Wirth, H. (2003): Szenarien für Angriffe auf wirtschaftsstatistische Einzeldaten – Ein Überblick. *In: Gnoss, R. und Ronning, G. (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten. Forum der Bundesstatistik, Band 42,* Wiesbaden, S. 11–24.

Zaslavsky, A. M. und Horton, N. J. (1998): Balancing disclosure risk against the loss of nonpublication. *Journal of Official Statistics, 14,* S. 411–419.

Zühlke, S., Zwick, M., Scharnhorst, S. und Wende, T. (2004): The research data centres of the Federal Statistical Office and the Statistical Offices of the Länder. *Journal of Applied Social Science Studies (Schmollers Jahrbuch) 124,* S. 567–578.

Zwick, M. und Lenz, R. (2009): Business Micro Data: Access and Anonymization. *In: Statistics in a Changing Society – 175 Years in Progress, Proceedings of the Royal Statistical Society.*

Zwick, M. und Lenz, R. (2009): Record Linkage in German Official Statistics. Erscheint in: *Bulletin of the 60th International Statistical Institute (ISI 2009),* Durban.

Nachwort

Nach Fertigstellung meiner Dissertation Ende 2001 über Funktionen- und Relationenalgebren an der TU Kaiserslautern schien der weitere Weg schon vorgezeichnet: Ich würde weitere Jahre im Elfenbeinturm mit universellen Algebren und relationalen Strukturen verbringen, die mich auch außerhalb des Büros nie losließen, gelegentlich unterbrochen durch Lehraufträge und Assistententätigkeit, um den ganzen Spaß auch finanzieren zu können.

Meine damalige Initiativbewerbung beim Statistischen Bundesamt in Wiesbaden kann daher durchaus als Versuch gesehen werden, aus dem Trott der klassischen wissenschaftlichen Laufbahn herauszubrechen und in einem Team interdisziplinär – das damalige Modewort für fachübergeifende Zusammenarbeit – zu agieren.

Mit der Zusage aus Wiesbaden und dem Weggang aus Kaiserslautern schien zugleich eine Habilitationsmöglichkeit in weite Ferne gerückt. In Wiesbaden durfte ich dann aber in einigen interessanten Forschungsprojekten rund um das Thema Statistische Geheimhaltung mitwirken. Inter„disziplin"arität war dann tatsächlich für den einzigen Mathematiker unter zahlreichen Wirtschaftswissenschaftlern und Soziologen erforderlich. Dem Thema der Geheimhaltung wirtschaftsstatistischer Mikrodaten bin ich auch nach späteren Rufannahmen an die Fachhochschule Mainz in 2006 und die Hochschule für Technik und Wirtschaft des Saarlandes in 2010, insbesondere durch die Leitung des Forschungsprojektes „Wirtschaftsstatistische Paneldaten und faktische Anonymisierung", treu geblieben.

Meiner Lebenspartnerin Astrid Alter danke ich von ganzem Herzen für ihre moralische Unterstützung und unendliche Toleranz während des Habilitationsverfahrens. Sie musste nicht nur auf gemeinsame Freizeitunternehmungen verzichten, sondern auch meine ständige geistige Abwesenheit ertragen.

Widmen möchte ich die vorliegende Arbeit in memoriam meinem Lehrer und Freund Prof. Dr. Dietmar Schweigert, der im Januar 2006 viel zu früh von uns gegangen ist. Von ihm habe ich die systematische Herangehensweise an (wissenschaftliche) Probleme erlernt sowie die in manchen Situationen notwendige Ruhe und Gelassenheit. Lernen durfte ich danach bei Herrn Prof. Dr. Gerd Ronning durch die intensive Zusammenarbeit in mehreren Forschungsprojekten; vor allem hoffe ich, dass die Fähigkeit, sich einerseits durch tiefe mathematische

Details zu wuseln und andererseits den Blick auf das Wesentliche nie aus den Augen zu verlieren, ein wenig auf mich abgefärbt hat.

Dank gebührt auch meinen Freunden Roland Sturm und Dr. Markus Zwick, bei denen ich während meiner Zeit im Statistischen Bundesamt reichlich Erfahrung sammeln durfte. Von Roland habe ich mir die Fähigkeit einer zielgerichteten Arbeitsweise abgeschaut, von Markus die Fähigkeit, auch einmal Wagnisse einzugehen frei nach dem Motto: Zwei Schritte vorwärts und einer zurück ergeben eine Netto-Vorwärtsbewegung. Beiden danke ich auch für die Gewährung flexibler Arbeitszeiten im Zusammenhang mit Lehraufträgen, die ich parallel zu meiner Tätigkeit im Statistischen Bundesamt variierend zwischen 4 und 13 Semesterwochenstunden an verschiedenen Hochschulen durchführen durfte.

Dem Präsidenten der Fachhochschule Mainz, Herrn Prof. Dr. Gerhard Muth, und dem Rektor Prof. Dr. Wolfgang Cornetz der Hochschule für Technik und Wirtschaft des Saarlandes danke ich sehr für die Unterstützung der Kooperationsprojekte mit dem Statistischen Bundesamt und anderen Institutionen durch den Erlass von Deputatsstunden sowie für die Ermöglichung einer Gastprofessur an der Leeds Metropolitan University im Sommersemester 2009.

Weiterhin danken möchte ich Herrn Präsident Roderich Egeler für die freundliche Aufnahme dieses Buches in das Veröffentlichungsprogramm des Statistischen Bundesamtes. Herrn Georg Schuck danke ich für die redaktionelle Unterstützung und die Bearbeitung der Tabellen und Grafiken.

Herrn Prof. Dr. Walter Krämer bin ich zu besonders großem Dank verpflichtet für seine Aufgeschlossenheit meinem nachträglichen Habilitationsvorhaben gegenüber. Er hat das Vorhaben, an der renommierten Fakultät Statistik der TU Dortmund zu habilitieren, seit unserem ersten Kontakt mit Rat und Tat voll unterstützt. Dies war bei meinem nichtlinearen wissenschaftlichen Werdegang keineswegs selbstverständlich. Dank gebührt an dieser Stelle auch den drei anonymen Gutachtern dieser Arbeit.

(Pecunia) nervus rerum: Ohne geeignete Finanzierung wäre ein Zustandekommen der vorliegenden Arbeit niemals möglich gewesen. Nicht zuletzt möchte ich daher dem Bundesministerium für Bildung und Forschung (BMBF) für die Finanzierung zahlreicher Projekte zur statistischen Geheimhaltung danken.