DESTATIS
wissen.nutzen.

STATISTIK UND WISSENSCHAFT

Jörg Höhne

Verfahren zur Anonymisierung von Einzeldaten

Band 16

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.

Zugel.: Eberhard-Karls-Universität Tübingen, Diss., 2009

Herausgeber: Statistisches Bundesamt, Wiesbaden

Internet: www.destatis.de

Ihr Kontakt zu uns:

www.destatis.de/kontakt

Statistischer Informationsservice

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

Informationen zu dieser Publikation unter

Tel.: +49 (0) 30 / 90 21 34 45
joerg.hoehne@statistik-bbb.de

Erschienen im September 2010

Print
Preis: EUR 24,80 [D]
Bestellnummer: 1030816-10900-1
ISBN: 978-3-8246-0901-7

Kostenfreier Download (PDF)
Artikelnummer: 1030816-10900-4
ISBN: 978-3-8246-0902-4

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

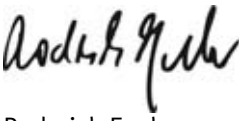
Die Einrichtung von Forschungsdatenzentren bei den Statistischen Ämtern des Bundes und der Länder, aber auch den anderen großen Datenproduzenten in Deutschland eröffnete die Möglichkeit, die erhobenen Mikrodaten der Forschung für wissenschaftliche Analysen bereitzustellen. Unabdingbare Voraussetzung für einen Zugang der Wissenschaft zu Mikrodaten ist die Wahrung der statistischen Geheimhaltung. Hierzu gibt es in der amtlichen Statistik zahlreiche Anonymisierungsverfahren.

Diese Methoden und Verfahren werden in dieser Publikation näher vorgestellt. Es ist das besondere Verdienst des Autors, die Anonymisierungsverfahren untersucht und Verfahrenserweiterungen entwickelt zu haben. Sie eröffnen der amtlichen Statistik neue Möglichkeiten, die Vertraulichkeit gegenüber den Auskunftgebenden bei der Verwendung statistischer Daten zu sichern.

Die vorliegende Arbeit ist als Dissertation bei Prof. Dr. Gerd Ronning an der Eberhard-Karls-Universität Tübingen entstanden. Jörg Höhne hat diese Verfahren im Rahmen der Projekte „Faktische Anonymisierung wirtschaftsstatischer Einzeldaten“ (von 2003 bis 2005) und „Faktische Anonymisierung von wirtschaftsstatistischen Paneldaten“ (von 2006 bis Ende 2008) entwickelt. Die beiden Projekte sind zwei vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Forschungsprojekte, bei denen in Zusammenarbeit zwischen den Statistischen Ämtern des Bundes und der Länder, dem Institut für Arbeitsmarkt- und Berufsforschung der Bundesanstalt für Arbeit sowie dem Institut für Angewandte Wirtschaftsforschung Tübingen die Möglichkeiten untersucht wurden, amtliche Mikrodaten für die Nutzung durch empirisch arbeitende Wissenschaftler über Scientific-Use-Files zu erschließen.

Das Statistische Bundesamt setzt mit der Veröffentlichung dieser Dissertation in seiner Reihe „Statistik und Wissenschaft“ die Tradition fort, die Zusammenarbeit der amtlichen Statistik mit der Wissenschaft zu fördern und zu dokumentieren.

Wiesbaden, im August 2010



Roderich Egeler

Präsident des Statistischen Bundesamtes

Inhalt

Abbildungsverzeichnis

Verwendete Variablen/Bezeichnungen

X	=	Matrix der statistischen Daten mit n Merkmalsträgern in den Zeilen und m Merkmalswerten in den Spalten $X_{n,m}$.(Bei der Untersuchung in Modellen enthält X ggf. nur die im Modell benötigte Auswahl an Daten.)
n	=	Anzahl Merkmalsträger/Einheiten im Datenbestand
m	=	Anzahl Merkmale/Variablen im Datenbestand
$x_{i,j}$	=	Wert des Merkmalsträgers i beim Merkmal j
$x_{i.}$	=	Zeile i der Matrix X
$x_{.j}$	=	Spalte j der Matrix X
X^o	=	Matrix enthält die Originalwerte
X^a	=	Matrix enthält die anonymisierten Werte

Die Kennungen ^{a} und ^{o} können auch für $x_{i,j}$, $x_{i.}$ und $x_{.j}$ verwendet werden.

Werden die Werte eines ausgewählten Merkmals unabhängig von den übrigen betrachtet, z. B. als abhängiges Merkmal bei Modellierungen, so werden sie als Vektor Y bezeichnet.

Y^o	=	Vektor der Originalwerte für ein ausgewähltes Merkmal
Y^a	=	Vektor der anonymisierten Werte für ein ausgewähltes Merkmal
y_i^o, y_i^a	=	Wert des Merkmalsträgers i beim ausgewählten Merkmal
$\odot$	=	Hadarmad-Produkt, d. h. für elementweise Multiplikation
$\otimes$	=	Kronecker-Produkt, d. h. jedes Element der Matrix A wird mit der Matrix B multipliziert. Ist A eine $\mathrm{m \times n}$ Matrix und B der Dimension $\mathrm{p \times r}$, so ist $\mathrm{C=A \otimes B}$ der Dimension $\mathrm{mp \times nr}$.

Einleitung und Begriffsbestimmung

Einleitung

Für die amtliche Statistik, aber auch für jede andere Institution, die mit Auswertungen beschäftigt ist, ist das Vertrauen in die Richtigkeit der von ihr bereitgestellten Analysen und Ergebnisse die grundlegende Voraussetzung für die Akzeptanz ihrer Arbeit. Das setzt nicht nur die fehlerfreie eigene Arbeit voraus, sondern bedingt auch, dass die erhobenen Daten wahr sind. Das Erheben wahrer Angaben gelingt jedoch nur, wenn den Berichtspflichtigen Vertraulichkeit zugesichert wird, d. h. dass gewährleistet wird, dass niemand die gelieferten Angaben zum Schaden des Berichtspflichtigen verwenden kann.

Aus diesem Grunde sind in den Gesetzen, die Institutionen zum Durchführen von Statistiken aber auch zum Pflegen von Registern (z. B. statistische Ämter und Rentenversicherungsträger) ermächtigen, gesetzliche Regelungen vorhanden, die die Geheimhaltung, aber auch eine eventuelle Nutzung der Einzelangaben regeln. Für Daten der amtlichen Statistik sind diese Regelungen im Bundesstatistikgesetz und in den Landesstatistikgesetzen enthalten.

Neben dem Schutzinteresse für Einzeldaten gibt es andererseits auch ein berechtigtes wissenschaftliches Interesse an den Einzeldaten. Das liegt darin begründet, dass diese Daten einerseits ein riesiges Analysepotential besitzen und sich andererseits gegenüber Eigenerhebungen von Wissenschaftlern durch einen bedeutend größeren Erhebungsumfang auszeichnen. Oftmals sind amtliche Daten sogar Totalerhebungen.[1] Damit haben sie oft eine Qualität, die mit den begrenzten Ressourcen einzelner wissenschaftlicher Einrichtungen nicht zu erreichen ist. Um den Schutz von Einzeldaten zu gewährleisten und gleichzeitig einen Zugang zu Einzeldaten für Analysezwecke zu ermöglichen, ist eine Anonymisierung erforderlich. Diese kann auf verschiedenen Wegen erfolgen (siehe Abschnitt „Begriffsbestimmung“).

In der vorliegenden Arbeit sollen Verfahren zur Anonymisierung von wirtschaftsstatistischen Einzeldaten vorgestellt werden. Dabei liegt der Schwerpunkt auf der näheren Vorstellung von Verfahrenserweiterungen und Methodenentwicklungen, die im Rahmen der Projekte „Faktische Anonymisierung wirtschaftsstatischer Einzeldaten“ (von 2003 bis 2005) und „Faktische Anonymisierung von wirtschaftsstatistischen Paneldaten“ (von 2006 bis Ende 2008) vorgenommen wurden. Die beiden Projekte sind zwei vom Bundesministerium für Bildung und Forschung (BMBF) geförderte Forschungsprojekte, bei denen in Zusammenarbeit von Datenproduzenten und Datennutzern die Möglichkeiten untersucht wurden, amtliche Mikrodaten für die Nutzung durch empirisch arbeitende Wissenschaftler zu erschließen. Für die Datenproduzenten beteiligten sich das Statistische Bundesamt sowie einzelne Statistische Landesämter über die Forschungsdatenzentren des Bundes und der Länder sowie die Bundesagentur für Arbeit (BA) über das Forschungsdatenzentrum der BA im Institut für Arbeitsmarkt und Berufsforschung am

1 Im Bereich der Wirtschaftsstatistiken haben selbst Stichprobenerhebungen im Bereich der großen Unternehmen oft einen Auswahlsatz von 100 %, so dass alle Unternehmen dieser Schicht in der Erhebung enthalten sind (Totalerhebung). Damit sind die Erhebungen häufig nur im Bereich der kleinen und mittleren Unternehmen echte Stichproben.

Projekt. Von Seiten der Datennutzer war das Institut für angewandte Wirtschaftsforschung Tübingen direkt im Projekt vertreten. Außerdem war eine Reihe weiterer Institutionen über einen Wissenschaftlichen Begleitkreis am Projekt beteiligt.

Der Autor arbeitete für das Amt für Statistik Berlin-Brandenburg (ehemals Statistisches Landesamt Berlin) an den Projekten mit. Im Rahmen der Projekte waren folgende Aufgabenteilungen/Arbeitsschwerpunkte vereinbart:

- Bereitstellung und Aufbereitung der untersuchten Daten (alle Datenproduzenten).
- Test von bekannten Anonymisierungsverfahren und Entwicklung von ggf. erforderlichen Verfahrenserweiterungen (Amt für Statistik Berlin-Brandenburg).
- Untersuchung der Schutzwirkung von getesteten Anonymisierungsverfahren (Statistisches Bundesamt).
- Untersuchung des Analysepotentials von anonymisierten Daten (Institut für Angewandte Wirtschaftsforschung).
- Untersuchung von Imputationsverfahren für die Anonymisierung von Daten der Bundesanstalt für Arbeit (Forschungsdatenzentrum der BA am Institut für Arbeitsmarkt und Berufsforschung). [2]

Die Ergebnisse des umfassenden Methodenvergleichs und die im Rahmen der Projekte bereitgestellten Mikrodaten sind bereits in den Veröffentlichungen der Projekte publiziert (siehe z. B. Gnoss et al. 2003 und Ronning et al. 2005). Deshalb wird auf diese Ergebnisse nur insofern eingegangen, wie sie für das Gesamtverständnis der Arbeit erforderlich sind.

Der Schwerpunkt der vorliegenden Arbeit liegt in der Darstellung der vom Autor im Rahmen des Projektes untersuchten Anonymisierungsverfahren, sowie der entwickelten Verfahrenserweiterungen.

Begriffsbestimmung

Einzeldaten (Mikrodaten) sind eine Sammlung von gleichartigen Angaben zu verschiedenen statistischen Objekten. Die gleichartigen Angaben betreffen dabei die konkreten Ausprägungen (Merkmalswerte) für verschiedene Eigenschaften (Merkmale) der statistischen Objekte (Merkmalsträger). Statistische Objekte sind die eigentlichen juristischen Einheiten, bei denen die Erhebung von Daten erfolgt. Das können einerseits Betriebe oder Unternehmen bei Wirtschaftsstatistiken sein, aber auch Personen, wie z. B. in der Gesundheitsstatistik oder der Einkommenssteuerstatistik. Liegen die Einzeldaten in maschinenlesbarer Form vor, so spricht man auch von einer Einzeldatendatei.

Das Gegenstück zu Mikrodaten sind **Makrodaten**. Dabei handelt es sich um Zusammenfassungen der Merkmalswerte für mehrere statistische Objekte, wie z. B. Summen- oder

2 Der unterschiedliche Arbeitsschnittt zwischen den statistischen Ämtern und der Bundesanstalt für Arbeit (BA) resultiert aus den anderen gesetzlichen Grundlagen, die eine institutionelle Streuung der Teilaufgaben bei der Untersuchung der Daten der BA verhinderte.

Durchschnittsangaben. Makrodaten sind somit Angaben, die üblicherweise nur einer Gruppe von statistischen Objekten zugeordnet werden können.[3]

Anonymität ist in Mikrodaten dann gegeben, wenn diese nicht zur Gewinnung von Informationen über die einzelnen statistischen Objekte dienen können. Eine Gewinnung von Informationen erfolgt üblicherweise in zwei Schritten. Der erste Schritt ist die eindeutige Zuordnung eines einzelnen Objektes zu einem Datensatz in der Mikrodatendatei. Danach können dann aus diesem Mikrodatensatz alle vorhanden Informationen abgelesen werden. Befinden sich darunter noch Informationen, deren Kenntnis nicht schon für die Zuordnung erforderlich war, hat man einen Informationsgewinn. Es existieren aber auch andere Möglichkeiten, einen Informationsgewinn über einzelne Objekte aus Datenbeständen zu erzielen (siehe Abschnitt 1.1).

Es existieren drei Stufen für die Anonymität von Mikrodaten. Das sind:

a) Formale Anonymität

Formale Anonymität ist gewährleistet, wenn die für diesen Datenbestand üblichen Zuordnungsmöglichkeiten der Datensätze zu den Merkmalsträgern nicht mehr möglich sind.

Für die im Erhebungs- und Plausibilisierungsprozess von statistischen Daten erforderlichen Zuordnungen werden im Datenbestand Identifikationsmerkmale geführt. Diese Merkmale sind für jeden Merkmalsträger eindeutig und ermöglichen somit eine schnelle Zuordnung der Datensätze. Solche Merkmale sind z. B. Namen und Adressangaben, Registernummern u. Ä. Werden diese Angaben aus dem Datenbestand entfernt, sind Zuordnungen nur noch durch das Auffinden von zufälligen eindeutigen Kombinationen von Merkmalsausprägungen möglich. In diesem Falle spricht man von „formaler Anonymität".

b) Absolute Anonymität

Absolute Anonymität ist dann gewährleistet, wenn es nicht möglich ist, trotz beliebig viel vorausgesetztem Zusatzwissen eine eindeutige und fehlerfreie Zuordnung vorzunehmen.

Dieser Begriff ist die härteste Form von Anonymität. Hierbei wird nicht unbedingt der Gewinn von Informationen aus dem Datenbestand, sondern bereits die erfolgreiche Zuordnung von Daten zu Objekten als Verletzung der Anonymität gesehen.[4]

3 In Einzelfällen kann es auch in Makrodaten vorkommen, dass eine eindeutige Zuordnung zu einem einzelnen statistischen Objekt möglich ist. Obwohl Makrodaten über eine entsprechende Aggregationsregel gebildet werden, kann es aber trotzdem sein, dass Objekte mit auffälligen einzigartigen Ausprägungskombinationen wieder eindeutig erkennbar sind, wenn sie durch die Aggregationsregel nicht mit anderen Objekten zusammengefasst werden.

4 Diese strenge Definition wird vor allem im Zusammenhang mit der Argumentation des „Vertrauensverlustes" beim Datenlieferanten gesehen. Dabei wird unterstellt, dass Datenlieferanten für die Statistik nicht unbedingt den Beweis der Verletzung der Anonymität durch Externe abwarten, sondern ggf. selber testen, ob ihre Angaben vertaulich verwendet werden. Für diesen Test besitzen sie natürlich das vollständige Wissen über ihr Unternehmen. Der Nutzen eines solchen Versuchs im Sinne des Informationsgewinns wird vernachlässigt und unterstellt, dass bereits bei erfolgreichem Auffinden der eigenen Angaben in statistischen Veröffentlichungen ein Vertrauensverlust für die amtliche Statistik entsteht. Deshalb ist es erforderlich, mit der verstärkten Herausgabe von Mikrodaten an die Wissenschaft auch die Stufen der Anonymität transparent mit darzustellen, um solchen Maximalforderungen von absoluter Anonymität entgegenzuwirken.

c) Faktische Anonymität

Faktische Anonymität ist dann gewährleistet, wenn der Aufwand für eine Zuordnung den Nutzen durch einen Informationsgewinn bei einer eventuellen erfolgreichen Zuordnung übersteigt.

Faktische Anonymität stellt somit eine Zwischenstufe zwischen der formalen und der absoluten Anonymität dar. Das Aufwands-/Nutzensverhältnis als Grundlage für die Bestimmung der Anonymität führt dazu, dass bei der Aufwandsbestimmung neben dem Deanonymisierungsaufwand auch der Aufwand z. B. für die Informationsbeschaffung aus alternativen Quellen berücksichtigt werden muss. Analog erfordert eine Nutzensbestimmung neben der Möglichkeit eines Informationsgewinns auch die Messung des Wertes der gewonnenen Informationen.

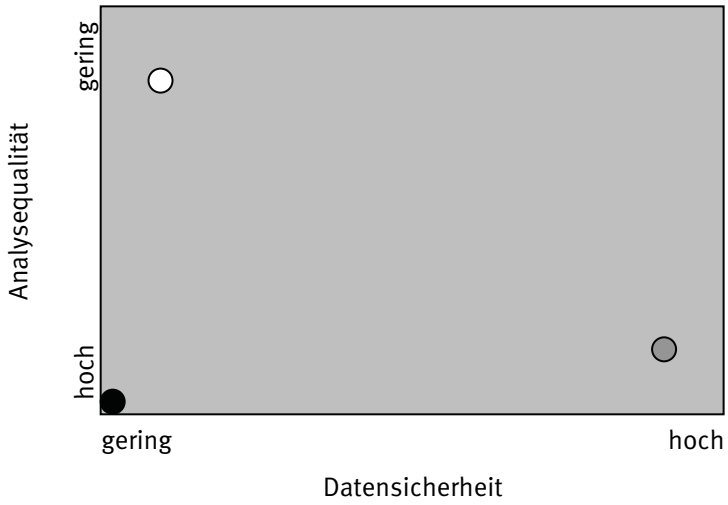
Das Bundesdatenschutzgesetz (§ 14 ff.) und das Bundesstatistikgesetz (BStatG § 16) regeln die Erhebung und mögliche Nutzung amtlich erhobener Einzeldaten (personenbezogener Daten). Das BStatG schreibt vor, dass absolut anonyme Einzeldaten an Einzeldatennutzer frei herausgegeben werden können (BStatG § 16 (1) Punkt 4). Bei faktisch anonymen Einzeldaten ist die Herausgabe der Mikrodaten an den Kreis der „unabhängigen Wissenschaft“ unter gewissen Voraussetzungen möglich (siehe BStatG § 16 (6)). Ein Zugriff auf formal anonymisierte Einzeldaten (im Bundesdatenschutzgesetz als pseudonymisierte Daten bezeichnet) ist nur unter sehr restriktiven Regeln innerhalb der Ämter z. B. im Rahmen von kontrollierter Datenfernverarbeitung in Datenzentren zulässig. Da die Art des ermöglichten Zugriffs auf Einzeldaten auch den Aufwand für Re-Identifikationsversuche bestimmt, haben sich für die faktische Anonymität verschiedene Stufen herausgebildet. So ist z. B. auch der Begriff „On-Site faktisch anonym“ entstanden, bei dem die organisatorischen Einschränkungen des Datenzugriffs an einem Gastwissenschaftlerarbeitsplatz und die daraus resultierende Begrenzung der technischen Möglichkeiten für einen Re-Identifikationsversuch (im Unterschied zu einem Scientific-Use-File am Arbeitsplatz des Wissenschaftlers) bei der Aufwandsbestimmung mit berücksichtigt werden.

Anonymisierungsziel

Ziel der Einzeldatenanonymisierung (Erstellung von Scientific-Use-Files) ist die Veränderung von originalen Einzeldaten (X^o) um eine Einzeldatendatei zu erzeugen, die einen Informationsgewinn über einzelne Datenlieferanten (statistische Einheiten) höchstens mit einem unverhältnismäßig großem Aufwand zulässt. Andererseits sollen bei gewünschten Auswertungen und Analysen mit den anonymen Einzeldaten möglichst zu den Originaldaten ähnliche Ergebnisse geliefert werden.

Beide Ziele scheinen sich zu widersprechen. Die Erzeugung möglichst ähnlicher Auswertungsergebnisse wird am besten durch eine sehr wenig (optimalerweise gar nicht) veränderte Einzeldatendatei erreicht. Eine Verhinderung von Informationsgewinn bei Datenangreifern ist am besten durch sehr starke Veränderungen der Einzeldaten möglich.

Abbildung 1
Datensicherheit und Analysequalität bei anonymisierten Einzeldaten



– Menge möglicher veränderter Einzeldaten X^a

– originale Einzeldaten X^o

– Lösung mit schlecht anonymisierten Einzeldaten X^a

– Lösung mit gut anonymisierten Einzeldaten X^a

Von einer nutzbaren anonymisierten Einzeldatendatei X^a wird eine hohe Analysequalität und Datensicherheit gefordert. Das bedeutet, dass bei der Anonymisierung eine Veränderung möglichst nur entlang der Achse der Datensicherheit gewünscht ist. Leider lassen sich bei Veränderung der Einzeldaten die beiden Wirkungen nicht voneinander trennen, so dass eine Kompromisslösung angestrebt wird. Dieser Kompromiss besteht darin, dass eine Lösung gesucht ist, die eine ausreichende Datensicherheit gewährleistet und gleichzeitig eine möglichst hohe Analysequalität erreicht. Die Daten sollen somit nur soweit verändert werden, wie es für die Erreichung der Anonymität erforderlich ist. Dabei sind Verfahren anzuwenden, die die Analysequalität möglichst wenig beeinflussen, bzw. dessen Auswirkungen gut abschätzbar und somit bei Analysen korrigierbar sind.

Bedeutung anonymisierter Einzeldaten

Für den Zugriff von Wissenschaftlern auf statistische Daten gibt es praktisch drei Möglichkeiten:

– Die erste Variante ist die Nutzung von Gastwissenschaftlerarbeitsplätzen in den Forschungsdatenzentren (FDZ) der statistischen Ämter. Hier besteht die Möglichkeit, in geschützten Bereichen mit für das Forschungsprojekt gezielt zusammengestellten

Datenbeständen zu arbeiten. Die erhaltenen Projektergebnisse werden auf Datensicherheitsaspekte geprüft, bevor sie weiterverwendet und z .B. publiziert werden dürfen.

– Die zweite Variante ist das Fernrechnen. Bei diesem Zugangsweg werden die Auswertungsprogramme der Wissenschaftler direkt mit den Originaldaten berechnet und die Ergebnisse durch die FDZ-Mitarbeiter auf Datenschutz geprüft, bevor sie dem Wissenschaftler übermittelt werden. Eine Anwesenheit des Wissenschaftlers am FDZ-Standort ist hier jedoch nicht erforderlich.
– Die dritte Variante stellen für die Wissenschaftler anonymisierte Einzeldaten (Scientific-Use-Files oder SUF) dar. Diese Datenfiles werden nur für die unabhängige wissenschaftliche Forschung bereitgestellt. Sie können direkt am Arbeitsplatz des Wissenschaftlers ausgewertet werden.

Da das Sicherheitsrisiko für die Daten bei Scientific-Use-Files gegenüber den ersten beiden Varianten jedoch höher ist, wurden diese Daten allerdings auch stärker anonymisiert. Daraus resultieren auch gewisse Vorbehalte für diese Daten, auf die im Laufe dieses Abschnittes näher eingegangen werden soll. In den Projekten zur Erstellung von Scientific-Use-Files wurde in der Regel auch die Datenbereitstellung für die beiden ersten Zugriffsvarianten mit geregelt, da die Aufbereitung und Dokumentation der Daten ein wichtiger vorbereitender Arbeitsschritt war. Außerdem gab es bei den Erstdaten projektbedingt auch immer eine zeitliche Verschiebung zwischen der Datenaufbereitung (und somit Bereitstellung für das Fernrechnen und Gastwissenschaftlerarbeitsplätze) und der Fertigstellung der ersten Scientific-Use-Files (SUFs), so dass ein großer Teil des aktuellen Forschungsinteresses bereits über die beiden anderen Zugangswege befriedigt wurde. Hier werden die SUFs erst bei späteren bereitzustellenden Zeiträumen ihre Vorteile zeigen, wenn sie möglichst zeitgleich mit der Datenaufbereitung zu Verfügung stehen können.

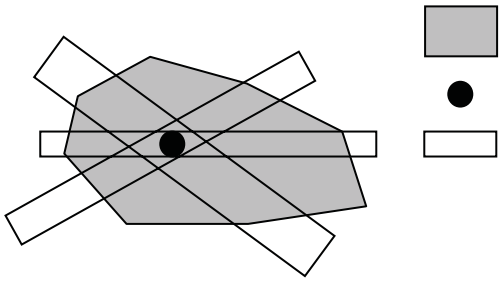
Generell darf der hohe manuelle Aufwand für die Gewährleistung des Datenschutzes am Gastwissenschaftlerarbeitsplatz und beim Fernrechnen nicht vernachlässigt werden. Je intensiver Daten ausgewertet und Ergebnisse publiziert werden, um so mehr vorhandenes Angriffswissen muss bei potentiellen Datenangreifern unterstellt werden, da durch Internettechnologien viele Veröffentlichungen sehr schnell und einfach zusammenzusuchen und zu recherchieren sind. Gerade Probleme der tabellenübergreifenden Geheimhaltung (siehe z. B. Giessing 2004) lassen sich so immer schwerer kontrollieren. Bei der Tabellengeheimhaltung werden neben den geheimzuhaltenden Informationen (Primärsperrungen) weitere Informationen aus der Tabelle entfernt (Sekundärsperrungen), die wegen der in Tabellen bestehenden linearen Abhängigkeiten sonst eine Berechnung der geheimen Werte ermöglichen würden. Da in Auswertungstabellen meistens auch Summen über Zeilen oder Spalten enthalten sind, bzw. aus anderen Quellen ermitteln lassen, sind Sekundärsperrungen erforderlich. Primärsperrungen sind das Ergebnis eines direkten Geheimhaltungsbedürfnisses einer Information. Sekundärsperrungen sind entfernte Informationen um nicht als „Hilfsmittel für die Rückrechnung“ zu dienen. Diese Notwendigkeit der Sekundärsperrung muss in anderen Tabellierungen nicht bestehen. Trotzdem darf diese Information auch dort nicht veröffentlicht werden, damit die Offenlegung von Informationen auch bei Auswertung mehrerer Tabellen verhindert wird. Mit dem zuneh-

menden Wunsch nach flexiblen tabellarischen Auswertungen wird die Geheimhaltungsprüfung somit immer komplexer. Aus diesem Grunde gibt es Bestrebungen, den Zugriff beim Fernrechnen zu automatisieren, so dass die Ansteuerung der Programme für den Wissenschaftler selber möglich wird und automatische Kontrollmechanismen den Datenschutz gewährleisten. [5]

Die meisten Ansätze gehen von einer Informationsreduktion in den bereitgestellten Datenbanken für den automatisierten Zugriff aus. Ein anderer Weg ist der automatische Schutz der Informationen durch nachträgliche Veränderungen der Ergebnisse z. B. die Jackknife-Methode (siehe Heitzig 2005). Bei diesen Ansätzen werden nur die zusammengefassten Ergebnisse mehrerer Bootstrap-Simulationen präsentiert und keine Originalergebnisse übermittelt. Diese Ansätze lösen jedoch nicht das eigentliche Anonymisierungsproblem, der auswertungsübergreifenden Anonymität. Dieses Problem soll an der folgenden Abbildung 2 veranschaulicht werden.

Auf Grund der allgemeinen inhaltlichen Definition der Merkmale einer Mikrodatei könnten in der Regel sehr viele Wertematrizen als potentielle Mikrodatendatei möglich sein. Jedes Auswertungsergebnis der Mikrodaten beschreibt aber Eigenschaften der Originaldatei, die nur noch von einem Teil dieser Mikrodatensätze erfüllt werden. Werden sehr verschiedenartige Auswertungen unabhängig voneinander durchgeführt, kann es vorkommen, dass die Menge der möglichen Mikrodatensätze so klein wird, dass bereits nicht mehr für alle Objekte Anonymität gesichert ist. Anonymitätsrisiken allein durch die Veröffentlichung von Korrelationskoeffizienten werden z. B. in Heitzig (2004) näher untersucht.

Abbildung 2
Erhöhung des Re-Identifikationsrisikos durch mehrfache Auswertungen der Originaldaten



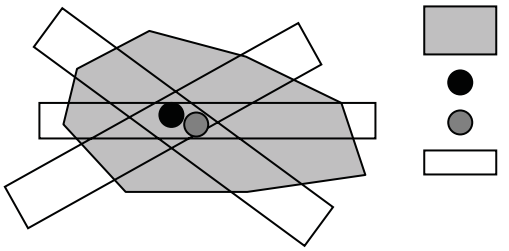
5 Eine umfassende Darstellung verschiedener nationaler Ansätze findet man z. B. in der Dokumentation der „Joint UNECE/Eurostat work session on statistical data confidentiality“ (vom 9. – 11. November 2005 in Genf). Auf dieser Tagung wurden z. B. die Vorgehensweisen in Schweden (Söderberg, L.-J. 2005), in Dänemark (Borchsenius, L. 2005), den Niederlanden (Hundepool, A. und de Wolf, P.-P. 2005) und den USA (Steel, P. und Reznek, A. 2005) näher vorgestellt.

Ein häufiges Beispiel dafür ist, die Ausführung gleicher Analysen für einen Datenbestand mit und ohne Ausreißer. Beide Analysen stellen für sich in der Regel kein Geheimhaltungsproblem dar, da sehr viele Einheiten in die Untersuchung eingehen. Werden jedoch nur sehr wenige Einheiten als Ausreißer klassifiziert, so bietet ggf. die Abweichung in den Ergebnissen die Möglichkeit, auf die Eigenschaften der einzelnen entfernten Einheiten rückzuschließen. Dieses Problem verschwindet auch dann nicht, wenn die Ergebnisse nach ihrer Berechnung mit einem zusätzlichen Unsicherheitsbereich versehen werden.

Hier könnten sich durch Scientific-Use-Files neue Möglichkeiten eröffnen: Werden die Auswertungen mit den anonymisierten Daten vorgenommen und die Sicherheitsintervalle der Ergebnisse dann so korrigiert, dass sie auch die Originaldaten mit einschließen, so kann unabhängig von der Anzahl der wissenschaftlichen Analysen das Intervall immer nur so klein sein, dass es eine Lösungsmenge beschreibt, die mindestens beide Lösungen (Original und Anonym) mit einschließt (siehe Abbildung 3). Dazu müssten nur im Ergebnis der Analysen die Abweichungen zu den Originalergebnissen mit angegeben werden. Dass muss aber ungerichtet erfolgen, damit kein direktes Rückrechnen möglich ist. Besitzt die anonyme Lösung einen ausreichenden Abstand zu den Originaldaten, so ist auch die Anonymität der bereitgestellten Auswertungsergebnisse immer sichergestellt.

Tabellenauswertungen aus anonymen Einzeldaten ermöglichen immer nur einen Rückschluss auf diese veränderten Einzeldaten. Diese kann um die Angabe eines Abweichungintervalls zum Original erweitert werden. Außerdem basieren alle Auswertungen auf dem gleichen Basismaterial (veränderte Einzeldaten). Damit sind die Auswertungen untereinander konsistent, was ein erheblicher Vorteil gegenüber Ansätzen der unabhängigen Auswertungsanonymisierung (siehe Heitzig 2005) darstellt, die eine Datenveränderung an den einzelnen zu schützenden Auswertungen vorschlagen. Ist die Anonymisierung der Einzeldaten an möglichst hochwertigen Qualitätskriterien optimiert worden, so stellt sie eine gute Basis für ein flexibles Auswertungssystem dar. Wichtig ist jedoch, dass den Datennutzern die Qualität der Daten bekannt sein muss. Zu erwartende Abweichungen sollten deshalb in allgemeiner Form dokumentiert werden.

Abbildung 3
Erhöhung des Re-Identifikationsrisikos durch mehrfache Auswertungen der anonymen Lösung



– mögliche Mikrodatensätze
– Originaldatensatz
– anonyme Lösung
– Menge an Mikrodatensätzen, die der bereitgestellten Lösung entsprechen

Damit wird die Suche nach optimal anonymisierten Mikrodatenbeständen (Scientific-Use-Files) auch weiterhin eine große Bedeutung haben, da anonymisierte Mikrodaten neue Möglichkeiten für die Datenschutzkontrolle beim automatisierten Fernrechnen eröffnen können.

Ein häufig anzutreffender Vorbehalt gegen anonymisierte Mikrodaten (SUFs) ist, dass diese wegen der Anonymisierung „die Realität nicht richtig abbilden können". Generell sind anonymisierte Daten keine Echtdaten. Aber auch die „Echtdaten" der amtlichen Statistik sind nicht die „absolute Wahrheit". Jede statistische Erhebung hat mit Erhebungsfehlern wie fehlende und fehlerhafte Rückläufe von Erhebungsunterlagen, aber auch Fehler in der Aufbereitung der Daten (Erfassungs- und Codierfehler) zu tun. Außerdem ist nicht gewährleistet, dass die Zeitpunkte/Zeiträume der Erhebung mit denen einer gewünschten Analyse völlig in Überstimmung gebracht werden können. Deshalb ist jede Statistik (amtliche wie auch nicht amtliche) als Ergebnis eines fehlerbehafteten Prozesses zu betrachten. Wird durch die Einzeldatenanonymisierung erreicht, dass durch eine begrenzte Fehlererhöhung eine flexiblere/schnellere Auswertung ermöglicht wird und das Schutzinteresse des Einzelnen gewahrt bleibt, so ist ein guter Kompromiss gefunden.

Andererseits verlangt Einzeldatenanonymisierung auch entsprechende Verantwortung von den Datennutzern. Eventuelle Analyseprobleme aus den Datenveränderungen sind auch für den Datennutzer oft vorhersehbar und können in Auswertungen berücksichtigt werden. Rückschlüsse auf Einzelinformationen müssen auch und gerade in anonymisierten Einzeldaten verhindert werden. Deshalb haben seltene Ausprägungskombinationen in Einzeldatendateien (und somit geringe Häufigkeiten in Auswertungstabellen) ein hohes Risiko, im Ergebnis der Anonymisierung stark verändert worden zu sein, bzw. bei Stichproben nur schlecht für Hochrechnungen genutzt werden zu können. Wurde der Datenbestand z. B. durch Stichprobenziehung anonymisiert, steht mit der Stichprobentheorie auch der methodische Apparat bereit, um sich die Unsicherheit in den Daten durch Konfidenzintervalle zu veranschaulichen. In der Regel kann mit der Bildung größerer Datengruppen, z. B. durch Entfernung oder Vergröbern einzelner kategorialer Merkmale, eine stärkere Zusammenfassung erreicht werden, womit die Auswertungsqualität steigt.

Bekannte Geheimhaltungsfälle der Tabellengeheimhaltung (z. B. Dominanzen von Monopolunternehmen) sollen sich auch mit anonymisierten Daten nicht offen legen lassen. Deshalb muss auch hier der Wissenschaftler damit rechnen, dass ihm „grobe Schätzungen" übergeben werden.

1 Messung der Sicherheit von anonymen Einzeldaten

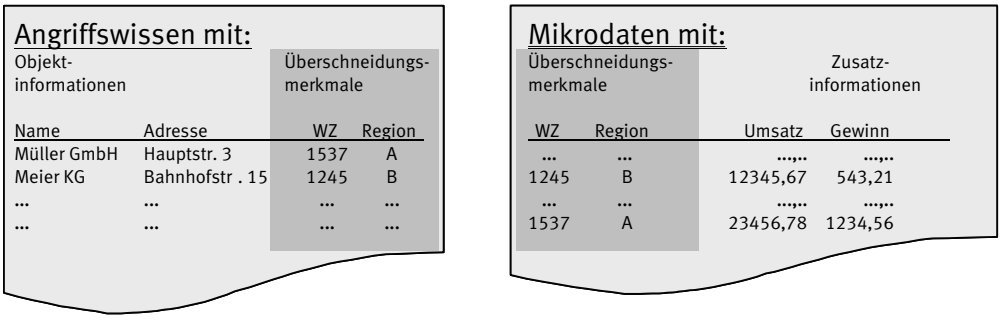
1.1 Risikoaspekte bei anonymisierten Einzeldaten

Um die Sicherheit von Einzeldaten messen zu können, ist zuerst eine Analyse der Risiken erforderlich. Risiken entstehen durch Personen, die versuchen, Zuordnungen von in den Mikrodaten enthaltenen Informationen zu den statistischen Objekten vorzunehmen (Datenangreifer). Eine umfassende Analyse über die Motivationen von Datenangreifern findet man bei Wirth oder Sturm (siehe Wirth 2003 und Sturm 2002). Hier sollen deshalb nur mögliche Techniken näher erläutert werden. Bei den Angriffen wird grundsätzlich zwischen „Einzelangriffen“ und einem „Massenfischzug“ unterschieden.

Bei Einzelangriffen wird versucht, ein einzelnes bekanntes Objekt im Mikrodatenbestand aufzufinden und neue Informationen über das Objekt zu gewinnen. Das kann auch nacheinander für mehrere Objekte erfolgen. Einzelangriffe sind mit einem sehr hohen Aufwand je gesuchtem Objekt verbunden. Damit können sie nicht in großem Umfang durchgeführt werden, um die Sicherheit des Datenbestandes zu testen. Einzelangriffe werden deshalb immer nur im begrenzten Umfang von Datentestern durchführbar sein, womit einerseits nur wenig über die Repräsentativität der Ergebnisse ausgesagt werden kann. Im Gegensatz zu Testern für die Sicherheit von Datenbeständen haben echte Datenangreifer andererseits das zusätzliche Problem, dass sie bei erfolgreichen Zuordnungen die richtigen von den fehlerhaften Zuordnungen nur schwer oder gar nicht unterscheiden können. Diese Möglichkeiten werden stark durch die Eigenschaften der Anonymisierungsverfahren und die Information über die durchgeführte Anonymisierung beeinflusst. In Anonymisierungsverfahren fehlerfrei erhaltene Merkmalswerte oder Zusammenhänge geben Datenangreifern eine höhere Sicherheit, wenn ihnen diese Eigenschaften bekannt sind.

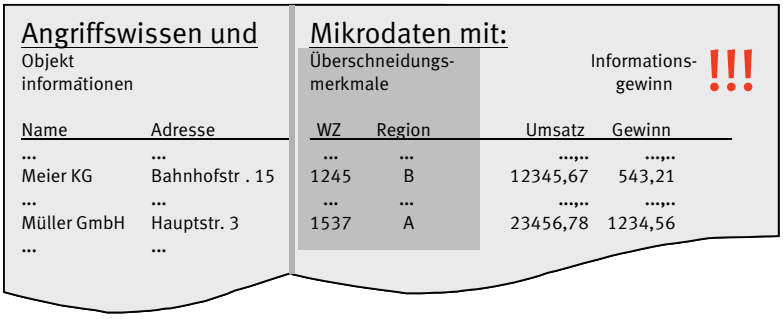
Bei Massenfischzügen besteht das Ziel darin, die eigenen Informationen über eine Reihe von Objekten (z. B. aus kommerziellen Datenbanken, eigenen Erhebungen) durch einen Abgleich mit der Mikrodatendatei anzureichern. Während beim Einzelangriff üblicherweise das ähnlichste Objekt gesucht wird, wird bei Massenfischzügen der Gesamtabstand zwischen den Paaren aus vorhandenen Objekten und den Objekten in der Mikrodatendatei minimiert, wobei ggf. auch zusätzlich unterstellt wird, dass jedes Objekt in der Mikrodatendatei nur einmal zugeordnet werden kann.

Abbildung 4
Die Überschneidung von Angriffswissen und Mikrodatendatei



noch: Abbildung 4
Die Überschneidung von Angriffswissen und Mikrodatendatei

Bei erfolgreicher Zuordnung der Überschneidungsmerkmale



Bei Datenangriffen hat man es als Angreifer mit dem Problem von Dateninkompatibilitäten zu tun, d.h. die Informationen in der Mikrodatendatei und im Angriffswissen sind in der Regel nicht identisch, obwohl es sich um die gleichen Objekte handelt. Das kann mehrere Ursachen haben:

– Die Bezugszeiträume sind nicht identisch.

  Wenn dem Datenangreifer z. B. nur Vorjahresangaben vorliegen, werden üblicherweise leichte Abweichungen vorhanden sein.

– Es sind beim Datenangreifer nur Informationen aus anderen Quellen vorhanden, bei denen das Merkmal etwas anders abgegrenzt definiert ist.

  Die „tätigen Personen“ und die „Beschäftigten“ variieren z. B. nur um die „tätigen Inhaber“ die meistens nur einen sehr kleinen Teil der „tätigen Personen“ ausmachen.

– Erhebungsfehler bei den Mikrodaten und/oder beim Datenangreifer

  Die Beschaffung der Informationen ist auf beiden Seiten mit gewissen Unsicherheiten behaftet, wie fehlerhafte Auskünfte, Erfassungs- oder Plausibilisierungsfehler. Dadurch können ebenfalls unterschiedliche Informationen zwischen Angriffswissen und Mikrodaten entstehen.

Datenangriffe werden üblicherweise durch einen Abgleich von Angriffswissen mit den Mikrodaten vorgenommen. Dazu ist es erforderlich, dass unter den Merkmalen im Angriffswissen und in den Mikrodaten Überschneidungen existieren. Bei diesen Überschneidungsmerkmalen (auch Schlüsselvariablen genannt) können Übereinstimmungen und Ähnlichkeiten dazu genutzt werden, einen Satz der Mikrodatendatei eindeutig zuzuordnen (siehe Abbildung 4).

Informationsgewinn muss nicht immer in zusätzlichen Informationen in der Mikrodatendatei bestehen. Es ist auch ein Informationsgewinn, wenn bereits bekanntes Wissen über ein Unternehmen durch amtliche Daten bestätigt und aktualisiert wird. Wird z. B. ein Unternehmen gesucht, dessen Umsatz nur aus Vorjahren bekannt ist, so kann dieses Merkmal einerseits für den Datenangriff genutzt werden und andererseits noch einen Informationsgewinn darstellen.

Neben der Kenntnis von konkreten Merkmalsausprägungen für die Zuordnung können in Spezialfällen auch andere Informationen ausreichen, um einen Informationsgewinn zu erzielen. Diese Geheimhaltungsprobleme sind bereits aus der Tabellengeheimhaltung bekannt und dort ebenfalls beschrieben (siehe z. B. Gießing 1999 und Gießing 2004).

a) Eine Zuordnung des Objekts ist auch bei der Kenntnis der Einzigartigkeit von Eigenschaften möglich. Ist z. B. bekannt, dass innerhalb einer bestimmten Menge von Objekten (z. B. Region oder Branche) nur das gesuchte Objekt über die Besonderheit verfügt (z. B. als einziges zu exportieren oder im betreffenden Jahr eine größere Investition getätigt zu haben), so reicht bereits diese Information, um das Unternehmen im Datenbestand zu lokalisieren. Solche strukturellen Besonderheiten haben selbst bei datenverändernden Anonymisierungsverfahren ein Zuordnungsrisiko. Hier wirkt sich der Fakt, dass jede Mikrodatendatei auch eine Basis für die Erzeugung von Makrodateien sein kann, negativ auf die Anonymität aus. Sichert ein Verfahren Eigenschaften, die auch Gegenstand von Makrodateien sind, fehlerfrei (wie z. B. Teilsummen oder Durchschnitte), so können innerhalb der Makrodateien wieder Anonymitätsprobleme auftreten. Es können Fallzahl- und Dominanzprobleme wie bei der Tabellengeheimhaltung auftreten.

   Beispiele:

   Würde z. B. der oben erwähnte Export auf verschiedene Objekte der gleichen Teilmenge aufgeteilt, würde die Berechnung der Summe ausreichen, um den Einzelwert wieder zu bestimmen (Einer-Fallzahlproblem [6]).

   Analog führt die Kenntnis von besonders großen Investitionen eines Einzelunternehmens dazu, dass die Berechnung der Summe der Investitionen in der Region eine bessere Schätzung ergeben würde als jeder eventuell zugeordnete Einzeldatensatz der Mikrodatendatei (Dominanzproblem).

b) Neben der Einzigartigkeit von Eigenschaften ist es auch kritisch, wenn nur zwei Objekte innerhalb von Teilmengen eindeutig lokalisierbar sind. Hier besteht das Risiko darin, dass diese Objekte sich gegenseitig eindeutig identifizieren können, weil sie in der Lage sind, nach dem Lokalisieren der zwei Objekte sich selbst wiederzuerkennen und so aus dem Datenbestand auszuschließen (Zweier-Fallzahlproblem der Tabellengeheimhaltung). Analog können sie aus Summenangaben für zwei Objekte ihren eigenen Beitrag abrechnen und so den Wert für den Anderen ermitteln. Da unter marktwirtschaftlichen Bedingungen zwei gleichartige Unternehmen in der Regel Konkurrenten darstellen, ist ihr Interesse an Informationen über den Anderen als besonderes Risiko für die Daten einzustufen. Dieses Datenschutzproblem tritt auch bei einer größeren Anzahl von Objekten auf, wenn von den n Objekten n-1 zusammen gehören. Sind bei einer Betriebserhebung z. B. alle Betriebe der Region bis auf einen Betrieb Teil eines Unternehmens, so muss ein Informationsaustausch zwischen diesen n-1 Betrieben unterstellt werden.

6 Die Begriffe Einer- und Zweier-Fallzahlproblem sowie Dominanzproblem sind Begriffe aus der Tabellengeheimhaltung. Bei der Tabellengeheimhaltung werden Risiken für die Bestimmung von Einzelwerten aus Auswertungstabellen betrachtet (siehe z. B. Gießing 1999 und Repsilber 1999).

c) Außerdem können durch tabellenübergreifende Datenangriffe Informationen gewonnen werden. Dabei werden Inkompatibilitäten zwischen Auswertungstabellen, die aus der Mikrodatendatei generiert wurden, und anderen verfügbaren Auswertungstabellen (z. B. Veröffentlichungen der amtlichen Statistik) genutzt, um Datenveränderungen durch Anonymisierungsmaßnahmen zu reproduzieren.

   Beispiele:

   Werden nur einzelne sehr große Einheiten aus dem Datenbestand entfernt, so ist es einfach möglich, deren Angaben zu reproduzieren, indem man den Fehlbetrag der Merkmalssummen in der anonymisierten Mikrodatendatei zu anderen Veröffentlichungen von Summenangaben der Merkmale bestimmt.

   Auf analoge Weise kann reproduziert werden, wenn einzelne sehr große Einheiten durch Veränderung von diskreten Merkmalen in andere Teilmengen verschoben wurden. Hier steht einem Fehlbetrag eine andere Teilmenge mit einer größeren Merkmalssumme als in den Veröffentlichungen gegenüber.

Diese Anonymitätsrisiken aus der Tabellengeheimhaltung werden in den Arbeiten zur Anonymisierung von Mikrodaten meist stark vernachlässigt (oder ignoriert), weil sie das Problem so stark erschweren würden, dass es ggf. nicht lösbar wäre. Deshalb wird darauf Wert gelegt, die Weitergabe der genauen Verfahrensbeschreibung von Anonymisierungsverfahren einzuschränken, wenn diese ggf. Auswirkungen auf die Sicherheit der anonymisierten Mikrodatendateien hat. Werden nicht alle Eigenschaften der anonymisierten Mikrodatendateien in Bezug auf Zusammenhänge zum originalen Datenbestand veröffentlicht, erzeugt dies eine Unsicherheit für den Datenangreifer, da er nicht mehr davon ausgehen kann, dass sein Datenangriff auf richtigen Annahmen beruht. Ein Beispiel ist die p%-Regel zum Schutz vor Dominanzproblemen in der Tabellengeheimhaltung durch Sperrung von Tabellenfeldern (siehe Gießing 2004). Grundidee dieser Regel ist es, jene Werte geheim zu halten, die es ermöglichen statistische Einheiten mit einem Fehler von weniger als p% zu schätzen. Ist dem Datenangreifer der Anwendungsparameter p bekannt, so ist es ihm möglich, das sichere Intervall für einen zu reidentifizierenden Wert stärker einzugrenzen als bei unbekanntem p. Hier erfordert die Bekanntgabe des Parameters eine Verwendung von größeren Parameterwerten, um eine gleichartige Schutzwirkung zu erhalten, so dass eine Unkenntnis des Parameters p einem noch größeren Informationsverlust durch weitere Löschungen von Tabellenfeldern vorzuziehen ist.

Das Bundesstatistikgesetz berücksichtigt diese Sicherheitsprobleme deshalb auch dadurch, dass es die Verantwortlichkeit für die statistische Geheimhaltung nicht nur einseitig bei den Produzenten der Daten verankert. Mit den §§ 21 und 22 BStatG (Verbot der Reidentifizierung und zugehörige Strafvorschrift) wird Re-Identifikation generell unter Strafe gestellt, unabhängig davon, ob der Zugang zu den Einzeldaten im Rahmen der Statistikerstellung, Nutzung im Verwaltungsprozess zur Planung und Gesetzgebung (BStatG § 16 (4)) oder zum Zwecke der wissenschaftlichen Forschung erfolgte (BStatG § 16 (6)).

1.2 Vorschläge zur Messung der Sicherheit

Die Messung der Sicherheit in anonymisierten Mikrodatendateien beschränkt sich somit auf die Bewertung der Risiken einer eindeutigen Zuordnung und der Qualität des Informationsgewinns. Die anderen in Abschnitt 1.1 erwähnten Anonymitätsrisiken werden ver-

nachlässigt, da die Vielzahl von Angriffsvarianten einerseits nur schwer abgeschätzt werden kann und andererseits auch stark vom Wissen über das Anonymisierungsverfahren und daraus resultierenden Eigenschaften der Mikrodatendatei abhängt. Das würde es erforderlich machen, jede anonymisierte Datei abhängig vom Anonymisierungsverfahren und den Konstellationen im Datenbestand (z. B. welche Betriebe gehören zu einem Unternehmen) anders zu bewerten.

Ein großer Teil von Verfahren zur Messung der Sicherheit beschränkt sich auf die Quantifizierung der Möglichkeit der eindeutigen und richtigen Zuordnung von originalen und anonymen Daten (z. B. Greedy-Matchingverfahren; siehe Lenz 2003). Neuere Verfahren (z. B. Lambert 1993) ermöglichen auch die Berücksichtigung verschiedener Qualitäten des Angriffswissens. Während z. B. bei Regionalangaben damit zu rechnen ist, dass es Datenangreifern gelingt, diese sehr exakt zu ermitteln, ist die Bestimmung der exakten in der Statistik verwendeten Wirtschaftszweigklassifikation für die Unternehmen selbst nicht so einfach. Analog haben vom Unternehmen selbst bereitgestellte Angaben (z. B. zur Anzahl der Beschäftigten im Internetangebot des Unternehmens) eine andere Qualität als geschätzte oder veraltete Umsatzwerte. Gelingt es, die Sicherheit der Angaben mit Hilfe von Verteilungen zu beschreiben, so kann sie im Ansatz von Lambert mit berücksichtigt werden und somit die Wahrscheinlichkeit der richtigen Zuordnung erhöhen. Hier liegt aber auch das Problem des Verfahrens. Vielen Datenangreifern wird es schwer möglich sein, die Qualitätsunterschiede zwischen den einzelnen Merkmalen des Angriffswissens hochwertig zu beschreiben. Gelingt das dem Datenangreifer jedoch nicht, so kann er nur stark vereinfachende Hypothesen (z. B. gleiche Qualität oder wenige Qualitätsstufen) unterstellen, womit diese Verfahren ihren Vorteil verlieren.

Reine Matchingverfahren haben jedoch den Nachteil, dass sie durch die fehlende Bewertung des potentiellen Informationsgewinns nicht geeignet sind, Verfahren zu bewerten, die die Merkmalswerte der gewonnenen Zusatzinformation verändern, also den Schwerpunkt der Anonymisierung auf die Reduzierung des Informationsgewinns legen. Eine eindeutige fehlerfreie Zuordnung eines Objektes zu einem Mikrodaten-Datensatz hat natürlich eine unterschiedliche Bedeutung, wenn einerseits davon ausgegangen werden kann, dass die dann gewonnenen Merkmalswerte fehlerfrei sind oder andererseits diese Merkmalswerte durch die Anonymisierung Veränderungen unterworfen wurden.

Deshalb wurde im Rahmen des Projektes „Faktische Anonymisierung wirtschaftsstatischer Einzeldaten“ ein Ansatz entwickelt, der sowohl das Zuordnungsrisiko als auch die Qualität zugeordneter Werte berücksichtigt (siehe Höhne, Sturm, Vorgrimmler 2003). Dabei wird das Deanonymisierungsrisiko durch ein zweistufiges Maß bestimmt. Nach der Bestimmung des Anteils an richtigen Zuordnungen bei Reidentifikationsversuchen wird dieser Anteil noch um die Brauchbarkeit der bei diesen zugeordneten Einheiten gewinnbaren Informationen korrigiert. Damit gibt das Maß an, wie viele nutzbare Informationen (mit Fehlern unterhalb einer Abweichungsschwelle) zu einem gesuchten Objekt im Datenbestand gefunden werden können. Da der Datenangreifer bei seinem Zuordnungsversuch die Richtigkeit seiner Zuordnung nicht verifizieren kann, ist es somit ein allgemeines Maß für das Deanonymisierungsrisiko des Datenbestandes.

2 Überblick über Anonymisierungsverfahren für Mikrodaten

Für die Anonymisierung von Einzeldaten existieren eine ganze Reihe von Möglichkeiten. Grundsätzlich gibt es drei verschiedene Anonymisierungseffekte (E1 – E3), mit denen Verfahren die Anonymität von Einzeldaten erreichen:

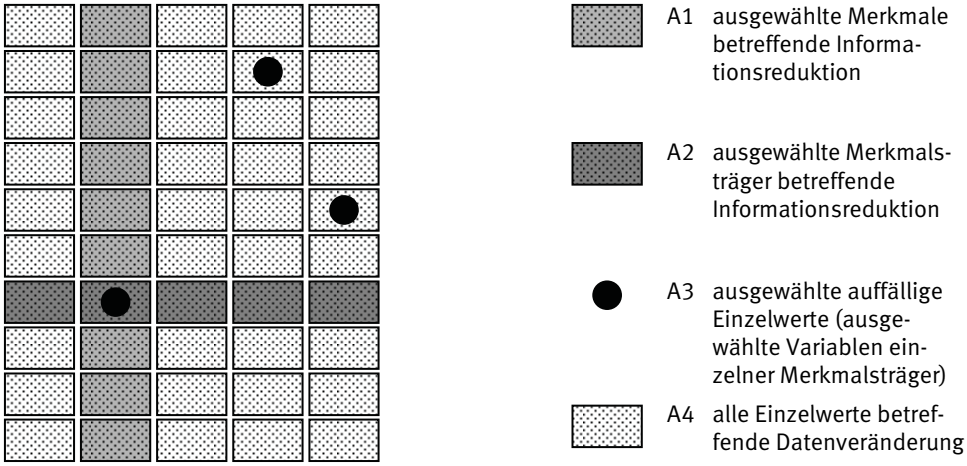
E1 Verhinderung der eindeutigen Zuordnung von Merkmalsträgern (durch Entfernung oder Veränderung von Überschneidungsmerkmalen).

E2 Verhinderung eines Informationsgewinns bei erfolgter Zuordnung (durch Entfernung von zusätzlichen Merkmalen).

E3 Reduzierung des Nutzens des Informationsgewinns (z. B. Unsicherheit der Information durch Veränderung der zusätzlichen Merkmale).

Alle Anonymisierungsverfahren führen Veränderungen an der Mikrodatendatei durch. Hierbei gibt es verschiedene Ansätze der Mikrodatenveränderung, die die verschiedenen Anonymisierungseffekte ausnutzen. Diese Ansätze (A1 – A4) sind in Abbildung 5 symbolisch dargestellt. Dabei symbolisieren die Spalten die Merkmale der Mikrodaten und die Zeilen die Merkmalsträger (Einheiten).

Abbildung 5
Ansätze zur Anonymisierung von Mikrodaten



A1 Bei der Informationsreduktion für ausgewählte Merkmale wird für alle Einheiten die Möglichkeit der eindeutigen Zuordnung reduziert, wenn Überschneidungsmerkmale bearbeitet werden. Gleichzeitig reduzieren diese Ansätze den potentiellen Informationsgewinn. Diese Verfahren empfehlen sich deshalb, wenn im Datenbestand durch den Umfang der Überschneidungsmerkmale ein hohes Re-Identifikationsrisiko besteht oder die zusätzlichen Merkmale einen zu hohen Informationsgewinn befürchten lassen.

A2 Besteht das hohe Deanonymisierungsrisiko nur für einen kleinen Teil der Merkmalsträger, so ist auch eine gezielte Informationsreduktion für diese Merkmalsträger möglich, um sowohl die Zuordnungsmöglichkeit als auch den Informationsgewinn für diese zu reduzieren.

A3 Manchmal ist es auch möglich, einzelne Merkmalsträger zu schützen, indem nur einzelne auffällige Merkmalswerte bearbeitet werden.

A4 Eine alle Einzelwerte betreffende Datenveränderung (z. B. Zufallsüberlagerung) zielt in erster Linie auf eine Reduzierung des potentiellen Informationsgewinns, aber auch auf eine Reduzierung der Möglichkeit der richtigen Zuordnung.

Anonymisierungsverfahren reduzieren oder verändern die Informationen einer Mikrodatendatei. Die verschiedenen Anonymisierungsverfahren werden von der Wissenschaft bislang in die traditionellen und die datenverändernden Verfahren eingeteilt. Diese Unterscheidung ist zum einen historisch gewachsen, da die traditionellen Verfahren bereits seit längerem in den statistischen Ämtern zum Einsatz kommen. Zum anderen ist ihr Anonymisierungsansatz üblicherweise eher inhaltlich orientiert und beruht auf einer Informationsunterdrückung, während die datenverändernden Verfahren eher algorithmisch vorgehen und formal Informationen verändern. Die Abgrenzung ist in Einzelfällen allerdings schwierig. Ebenso ist die Namensgebung teilweise verwirrend, da auch bei Verfahren, die zu den traditionellen gezählt werden, Veränderungen an den Werten vorgenommen werden (Höhne 2003). Im Folgenden wird versucht, die Anonymisierungsverfahren anhand ihrer Wirkung auf den Datenbestand und/oder ihrer Verfahrensähnlichkeit zu klassifizieren.

Neben einer kurzen Beschreibung der Verfahren und ihres Prinzips der Schutzwirkung soll hier auch dargestellt werden, welchen Einfluss die einzelnen Verfahren auf die Auswertbarkeit der Mikrodaten haben. Für die Beurteilung der Auswertbarkeit wurden folgende Eigenschaften gewählt:

- Erhalt der Mittelwerte.
- Erhalt der Verteilungen (Varianz/Standardabweichung/Spannweite).
- Auswertbarkeit von Teilmassen.
- Erwartungstreue oder Verzerrung von Koeffizienten in linearen Regressionsmodellen.

Zusätzlich werden die Restrisiken beurteilt und ggf. Hinweise für deren Beseitigung gegeben.

2.1 Traditionelle Anonymisierungsverfahren

2.1.1 Variablenunterdrückung

Die Unterdrückung von Variablen beinhaltet das Entfernen von kompletten Merkmalen aus dem Datenbestand. Das ist mit und ohne Bereitstellung von Ersatzinformationen möglich. Folgende Möglichkeiten existieren:

- **Variablenunterdrückung ohne Ersatzinformation**

  Existieren im Datenbestand ein oder mehrere Merkmale, die das Re-Identifikationsrisiko sehr stark erhöhen, so können diese entfernt werden.

Merkmale mit einem hohen Re-Identifikationsrisiko (wie z. B. Auslandsumsätze oder Investitionen) aus der Mikrodatendatei zu entfernen, würde dieses Risiko stark verringern.

– **Variablenunterdrückung mit Ersatzinformation über Variablenkonstruktion**

  Der Informationsverlust im Datenbestand kann durch die Bildung von Linearkombinationen aus den kritischen Merkmalen (z.B. Summen, Durchschnitte) reduziert werden.

  (Es wäre z. B. möglich an Stelle von Inlands- und Auslandsumsatz nur den Gesamtumsatz anzugeben.)

– **Variablenunterdrückung mit Ersatzinformation über Beziehungs- und Verhältniszahlen**

  Beziehungs- und Verhältniszahlen verhindern das Erkennen von stark dominierenden Unternehmen, während sie die strukturellen Informationen erhalten. Merkmalsträger mit auffälligen Strukturen bleiben dadurch aber ungeschützt.

  (Wenn statt des Inlands- und Auslandsumsatzes nur der Anteil des Auslandsumsatzes im Datenbestand enthalten ist, so sind große Unternehmen nicht mehr über den Umsatz identifizierbar. Unternehmen mit einem auffällig hohen Exportanteil können aber trotzdem erkannt werden.)

– **Variablenunterdrückung mit Ersatzinformation über Indexbildung bei Zeitreihen und Paneldaten**

  Indexzahlen verhindern wie Verhältniszahlen das Erkennen von stark dominierenden Unternehmen, während sie die zeitlichen Trendinformationen erhalten. Dabei wird für alle Werte nur die Veränderung zu einem Basiszeitraum ausgewiesen.

  (Das Ausweisen von Indexzahlen weist einen sehr hohen Schutz auf. Problematisch sind hier nur Unternehmen, die innerhalb der Zeitreihe eine markante einmalige Entwicklung aufweisen. Das können z. B. Konkurse oder Firmenzukäufe ins Unternehmen sein, die einerseits als Angriffsinformation leicht beschafft werden können und andererseits einen Bruch in den Zeitreihen bewirken.)

Mit der Variablenunterdrückung bei Schlüsselmerkmalen wird die Anzahl der eindeutigen Schlüsselvariablenkombinationen, d. h. der in der Mikrodatendatei nur einmal (oder auch zweimal) auftretenden Kombinationen von Merkmalsausprägungen der Schlüsselvariablen und damit die Möglichkeit der eindeutigen Zuordnung reduziert. Bei der Entfernung von sensiblen Merkmalen sinkt automatisch der Nutzen einer Re-Identifikation. Variablenunterdrückungen wirken auf alle Merkmalsträger in der gleichen Weise, d. h., die Informationsreduktion wird für alle Merkmalsträger vorgenommen (siehe Ansatz A1 in Abbildung 5).

Wichtig bei der Variablenunterdrückung ist, dass die unterdrückte Information nicht einfach aus den übrigen Merkmalen (z. B. mit Hilfe der bereitgestellten Ersatzinformationen) ermittelt werden kann. Es dürfen z. B. nicht die Ersatzinformationen Gesamtumsatz und Auslandsumsatzanteil zusammen bereitgestellt werden, da dann die angestrebte Informationsreduktion zum Auslandsumsatz nicht mehr gegeben ist, weil er einfach berechnet werden kann.

Für die Analysemöglichkeiten der Mikrodaten sind Variablenunterdrückungen zwiespältig zu betrachten. Da an den übrigen Merkmalen keine Veränderungen vorgenommen werden, sind somit auch alle Auswertungen und Modelle, die nur die verbliebenen Merkmale verwenden, fehlerfrei möglich. Für die nicht unterdrückten Variablen bleiben somit alle Eigenschaften erhalten. Sind die unterdrückten Merkmale jedoch für diese Auswertungen dringend erforderlich, so sind die Auswertungen nicht mehr möglich. Deshalb sollten Variablenunterdrückungen nur in sehr eingeschränktem Umfang Anwendung finden. Optimal anwendbar sind sie, wenn die anonymisierten Mikrodaten für ein konkretes Projekt erzeugt werden sollen. Dann besteht die Möglichkeit in Absprache mit den Wissenschaftlern, sich bei den Variablenunterdrückungen auf die Merkmale zu konzentrieren, die für die Auswertung nicht erforderlich sind, aber die Anonymität stark erhöhen.

2.1.2 Unterdrückung von Objekten oder Werten

Bei der Unterdrückung von Objekten (Merkmalsträgern) werden die betroffenen Merkmalsträger komplett aus dem Datenbestand entfernt. Damit wird die Möglichkeit der Re-Identifikation dadurch reduziert, dass die Objekte nicht mehr im Datenbestand sind. Sind die Objekte nur auf Grund einzelner Merkmalswerte auffällig im Datenbestand, besteht die Möglichkeit auch nur diese zu unterdrücken. Verfahrensvarianten sind:

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

Mit der **Stichprobenziehung** wird durch die Unsicherheit, dass die Merkmalsträger noch im Datenbestand sind, das Risiko einer falschen Re-Identifikation auch bei eindeutigen Zuordnungen erzeugt. Dieses Risiko ist um so höher, je geringer der Auswahlsatz ist. Wirkungslos ist die Stichprobenziehung gegenüber Angriffswissen, das auf einmaligen Besonderheiten beruht (z. B. das einzige Unternehmen mit Auslandsumsatz in einer Teil-

menge). Wenn diese Merkmalsträger mit den besonderen Eigenschaften in der Stichprobe enthalten sind, sind sie wieder leicht zuzuordnen. Sind die Elemente nicht enthalten, sind sie geschützt. Datenbestände, die durch Stichprobenziehung anonymisiert wurden, sind gut analysierbar, weil mit den Analyseverfahren für Stichprobenerhebungen ein umfangreicher Methodenapparat existiert, der auch hier genutzt werden kann. Sind bei geschichteter Stichprobenziehung die Auswahlsätze bekannt, so können auch Teilmassen gut ausgewertet werden.

In der amtlichen Statistik werden eine ganze Reihe von Erhebungen auf der Basis von Stichproben durchgeführt. Auch diese Datenbestände weisen jedoch noch ein Sicherheitsrisiko auf, wenn die Teilnahme an der Erhebung sich aus den methodischen Grundlagen der Statistik herleiten lässt. So existieren bei vielen Wirtschaftsstatistiken Bereiche, in denen Auswahlsätze von 100 % vorgegeben sind. Damit ist allein durch die Kenntnis des entsprechenden Wirtschaftszweiges und/oder der Größenklasse eines Unternehmens klar, dass es auch im Datenbestand der Stichprobenerhebung enthalten ist. Analog wissen Befragte des Mikrozensus, dass auch die Angaben ihrer Nachbarn in der Stichprobe existieren, da immer alle Haushalte eines Einzugsbereiches (Adresse) befragt werden. Deshalb ist auch bei Stichprobenerhebungen in der Regel eine zusätzliche Stichprobenziehung für die Anonymisierung der Mikrodaten erforderlich.

Mit durch Stichprobenziehung anonymisierten Daten können die Mittelwerte und Summen von Variablen erwartungstreu geschätzt werden. Die Spannweite wird in der Regel unterschätzt, da nur noch mit der Wahrscheinlichkeit von n(n-1)/N(N-1) (mit n - Anzahl der Einheiten in der Stichprobe bei N - Einheiten in der Gesamtheit) gewährleistet ist, dass sowohl das Minimum als auch das Maximum der jeweiligen Variablen in der Stichprobe enthalten sind.

Auswertbarkeit von Teilmengen

Die Analyse von Teilmengen ist ebenfalls erwartungstreu möglich. Problematisch kann für die Teilmengen ggf. der Umfang der Stichprobe sein, wenn auf Grund der kleinen Größe der Teilmenge in der Stichprobe keine gesicherten Auswertungen mehr möglich sind.

Die **Einschränkung der Grundgesamtheit** ist gegenüber der Stichprobenziehung ein Verfahren, mit dem gezielt Merkmalsträger geschützt werden können, für die ein besonderes Zusatzwissen existiert. Da die übrigen Merkmalsträger nicht verändert werden, sind sie ggf. weiter ungeschützt. Sollte ein weiterer Schutz erforderlich sein, empfiehlt es sich, nach diesem Verfahren z. B. noch eine Stichprobenziehung anzuwenden. Aussagen über die Gesamtheit können mit einem eingeschränkten Datenbestand nicht getroffen werden, da das Entfernten einer ganzen Gruppe in der Regel zu systematischen Verzerrungen der Ergebnisse führt. Bei der Einschränkung der Grundgesamtheit wird die Analysemöglichkeit dann nicht eingeschränkt, wenn die Analyse der verbliebenen Menge sinnvoll ist, d. h. auf die betroffenen Merkmalsträger verzichtet werden kann. Für die verbleibende Menge sind alle Auswertungen unverzerrt möglich. Wurde die Einschränkung vorgenommen, weil andere Quellen über diese Merkmalsträger ein zu hohes Re-Identifikationsrisiko erzeugen, kann geprüft werden, ob die Informationen aus diesen Quellen verwendet werden könnten, um separate Untersuchungen für die entfernte Teilgesamtheit durchzuführen.

Beim **Abschneideverfahren** besteht das Problem, dass in der Regel sehr große Unternehmen davon betroffen sind. Diese haben jedoch oft einen starken Einfluss auf das Gesamtergebnis. Die Aussagen, die mit einer so anonymisierten Mikrodatendatei erhalten werden, sind bezüglich Summen- und Durchschnittsangaben meistens nach unten verzerrt. Streuungsmaße werden ebenfalls nach unten verzerrt, da das Abschneideverfahren Objekte vom Rand des Datenbestandes (die größten oder kleinsten) entfernt. Wenn die zu analysierende Teilmenge vom Abschneideverfahren betroffen wurde, sind auch ihre Ergebnisse verzerrt. Nicht betroffene Teilmengen, z. B. Branchen ohne Großunternehmen, lassen sich fehlerfrei analysieren. Dass das Abschneideverfahren auch für besonders kleine Merkmalsträger angewandt werden kann, kompensiert diesen Effekt nicht, da für eine Kompensation viel mehr kleinere Merkmalsträger als große entfernt werden müssten. Bei einem Vorhandensein von mehreren kleinen Merkmalsträgern wären diese jedoch nicht mehr mit einem besonderen Re-Identifikationsrisiko behaftet. Werden die Regeln des Abschneideverfahrens bei der Interpretation berücksichtigt, liefert die Mikrodatendatei für die verbleibenden Merkmalsträger fehlerfreie Ergebnisse. So ist es z. B. möglich, für die Analyse von klein- und mittelständischen Unternehmen Ergebnisse eines Abschneideverfahrens zu nutzen.

Die Verfahren Stichprobenziehung, Einschränkung der Grundgesamtheit und die Abschneideverfahren entfernen die Information für ganze Merkmalsträger im Datenbestand (entspricht Ansatz A2 in Abbildung 5).

Bei der **Unterdrückung einzelner Werte** wird die Information nur für einzelne Merkmalswerte entfernt (siehe Ansatz A3 in Abbildung 5). Damit bleiben alle nicht zu schützenden Informationen eines Merkmalsträgers erhalten. Die Unterdrückung erfolgt einerseits, um den Informationsgewinn aus der Mikrodatendatei zu verhindern. Dieser Informationsgewinn ist jedoch nur möglich, wenn vorher eine Zuordnung des Datensatzes zu einem Merkmalsträger erfolgen konnte. War diese Zuordnung ohne Nutzung dieses Merkmalswertes möglich, so besteht für die anderen Merkmalswerte des Datensatzes immer noch ein Re-Identifikationsrisiko und somit noch keine Anonymität. Deshalb wird die Unterdrückung einzelner Werte andererseits immer auch zur Verringerung des Re-Identifikationsrisikos verwendet. Damit entsteht das Problem, dass eine einzelne Werteunterdrückung in der Teilmenge nicht ausreicht, um die Anonymität zu erreichen. Wird akzeptiert, dass als Angriffswissen die Existenz besonderer Merkmalswerte (wie z. B. ein einziges Unternehmen mit Auslandsumsatz oder hohen Investitionen) vorhanden ist, ermöglicht eine einzelne Werteunterdrückung wieder die Zuordnung des Merkmalsträgers zum Datensatz. Es muss also mindestens ein zweiter Wert mit unterdrückt werden, um die Anonymität herzustellen. Einen Ausweg bietet die Imputation eines neuen Wertes. Mit der Einschätzung eines neuen Wertes besteht nicht mehr die Notwendigkeit, die Entfernung kenntlich zu machen. Die besonderen Werte im Datenbestand (Ausreißer) werden dann durch Schätzungen ersetzt (z. B. aus Regressionsfunktionen). Das Verfahren funktioniert allerdings nicht, wenn der „Ausreißer“ nur dadurch entstanden ist, dass das Unternehmen eine auffällige Größe besitzt, während die Struktur jedoch im Datenbestand typisch ist. In diesem Fall liefern die Imputationen ggf. sehr gute Schätzungen des originalen Wertes und somit keine Informationsreduktion für den Datenangreifer. Die Werteunterdrückung ist somit nicht geeignet, um die besonders großen oder kleinen Unternehmen zu schützen. Hier kann die Werteunterdrückung keine zum Abschneideverfahren gleichwertige Sicherheit liefern.

Mittelwerte und Summen werden bei der Werteunterdrückung für die nicht behandelten Variablen fehlerfrei erhalten. Die behandelten Variablen werden in ihrem Mittelwerten üblicherweise unterschätzt, da in der Regel sehr große Werte die Variablenunterdrückung auslösten. Erfolgte eine Imputation der unterdrückten Werte, sind mit dem Datenbestand wieder erwartungstreue Schätzungen der Mittelwerte und Streuungsmaße möglich.

2.1.3 Informationsreduktion für Objekte

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

- **Gruppierung**

  Bei metrischen Merkmalen werden Intervalle gebildet, die dann als Intervallklassen immer einen Bereich von möglichen Werten beschreiben. Es wird dann der metrische Wert durch die entsprechende Klasse ersetzt (z. B. Beschäftigtengrößenklassen oder Umsatzgrößenklassen).

  Bei kategorialen Merkmalen werden neue Kategorien eingeführt, die dann mehrere Ausprägungen umfassen und an Stelle der originalen Kategorien verwendet werden (z. B. Zusammenfassungen der Klassifikation der Wirtschaftszweige (WZ) von 5-Steller auf 3- oder 2-Steller oder Zusammenfassung der Regionalschlüssel wie Kreisangabe statt Gemeinde).

- **Rundung**

  Metrische Merkmale werden so stark gerundet, dass dadurch sowohl die eindeutige Zuordnung als auch die Brauchbarkeit der Information für den Datenangreifer stark reduziert wird.

- **Censoring (Top-/Bottomcoding)**

  Merkmalsträger, die wegen ihrer Größe ein erhöhtes Re-Identifikationsrisiko bei einem Merkmal haben (besonders groß oder klein), werden verändert, indem die kritischen Merkmalswerte (oberhalb/unterhalb eines Grenzwertes) auf die festgelegten Grenzwerte gesetzt werden. Es muss beachtet werden, dass mehr als zwei Einheiten durch Censoring behandelt werden, wenn davon auszugehen ist, dass die Summenangaben über andere Veröffentlichungen verfügbar sind.

- **Replacement**

  Merkmalsträger, die wegen ihrer Größe ein erhöhtes Re-Identifikationsrisiko bei einem Merkmal haben (besonders groß oder klein), werden verändert, indem die kritischen Merkmalswerte (oberhalb/unterhalb eines Grenzwertes) auf den Durchschnitt des Merkmalswertes über alle betroffenen Merkmalsträger gesetzt werden.

  Um die Sicherheit zu gewährleisten, müssen mindestens drei Merkmalsträger von der Durchschnittsbildung betroffen sein (Fallzahlproblem). Außerdem ist die Dominanz innerhalb der Gruppe noch zu prüfen, da die Summe der drei Durchschnittsangaben dem Zweitgrößten einen guten Ansatz bietet, den Größten zu schätzen (siehe Fallzahl- und Dominanzgeheimhaltung im Abschnitt 1.1), wenn der Anteil des Drittgrößten an

der Summe nur sehr klein ist. Sind beide (Zweit- und Drittgrößter) gegenüber dem Größten sehr klein, so können auch Unbeteiligte die Summe der drei Angaben als gute Schätzung der Angaben des Größten nutzen. In Branchen, in denen die starke Dominanz durch ein Unternehmen als bekannt unterstellt werden kann, ist somit nur ein Censoring möglich, um die Daten zu schützen. Solche Fälle sind z. B. bei Post oder Bahn, aber auch bei regionalen Nahverkehrsanbietern in regional gegliederten Datenbeständen möglich.

– **Klonen**

  Einzelne kleine Merkmalsträger, die wegen ihrer seltenen Ausprägungskombinationen auffällig sind, werden anonymisiert, indem gleichartige künstliche Merkmalsträger erzeugt werden. Die künstlichen Merkmalsträger haben die gleichen kategorialen Merkmale, wenn diese als Angriffswissen unterstellt werden können. Dadurch wird die eindeutige Zuordnung verhindert. Stellen die Ausprägungen der kategorialen Merkmale bereits einen Informationsgewinn dar, sollten sie variieren. Die metrischen Merkmale werden durch ähnliche, aber nicht identische Werte belegt, wodurch ein Informationsgewinn verhindert wird.

– **Zerlegung**

  Einzelne große Merkmalsträger, die wegen der Größe ihrer metrischen Merkmalswerte auffällig sind, werden anonymisiert, indem ihre metrischen Merkmalswerte auf mehrere künstliche Merkmalsträger nach einem geheimen Verteilungsschlüssel verteilt werden. Dabei muss beachtet werden, dass die Merkmalsträger nicht über ihre kategorialen Merkmalswerte bereits eindeutig sind, da sie sonst leicht wieder durch Summenbildung reproduziert werden könnten.

Ziel dieser Verfahren ist es, die Möglichkeit der eindeutigen Zuordnung zu reduzieren und ggf. den Informationsgewinn zu verringern.

Gruppierung und Rundung stellen eine Vergröberung der Merkmalswerte dar, die sich bei der Analyse gut berücksichtigen lässt. Problematisch wird nur die Untersuchung von Teilmassen, da diese durch die Vergröberung nicht mehr ausgewählt werden können.

Censoring führt zu einer Verzerrung von Durchschnitten und Varianzen für die Mikrodatendatei. Wenn nicht ein ausschließliches Bottomcoding (Censoring der kleinen Werte) durchgeführt wurde, sind die Durchschnitte nach unten verzerrt (Begründung analog zu Abschneideverfahren im Abschnitt 2.1.2). Die Spannweite und Varianzen von mit Censoring behandelten Merkmalswerten sind in jedem Falle kleiner, da die Extremwerte im Datenbestand den durch das Verfahren vorgegebenen Grenzen entsprechen.

Im Gegensatz zum Censoring ist **Replacement** zwar bezüglich der Summen und Durchschnitte neutral, die Spannweite und Varianzen werden aber auch hier bei den betroffenen Merkmalswerten verringert.

Bei der Analyse von Teilmassen ist zu beachten, dass in den Teilmassen, die von den Maßnahmen nicht betroffen sind, die Analyse natürlich fehlerfrei möglich ist, während die beschriebenen Verzerrungen in den betroffenen Teilmassen stärker wirken, als auf den Gesamtbestand (wegen der kleineren Grundgesamtheit in der Teilmasse). Deshalb wäre es wichtig, die Regeln (Grenzen für Censoring und Replacement) zu kennen, um

gerade bei Teilmassenauswertungen die Aussagefähigkeit noch beurteilen zu können. Die Bekanntgabe der Grenzen ist möglich, wenn Censoring und Replacement so angewendet werden, dass nicht nur Einzelfälle in den Teilmassen betroffen sind. Diese Einzelfälle könnten sonst mit Hilfe externer Quellen korrigiert werden.

2.2 Datenverändernde Anonymisierungsverfahren

2.2.1 Allgemeine Bemerkungen zur Behandlung diskreter und kategorialer Merkmale

Die meisten datenverändernden Verfahren wurden ursprünglich ausschließlich für stetige metrische Merkmale entwickelt. Ausnahmen bilden nur die Zufallsüberlagerung nach Sullivan, PRAM, SAFE und das Resampling-Verfahren. Deshalb haben sich gleichzeitig einige Vorgehensweisen entwickelt, mit denen man in der Lage ist, diskrete aber auch kategoriale Merkmale zu behandeln. Obwohl viele Verfahren im Zusammenhang mit ganz speziellen datenverändernden Verfahren entwickelt wurden, ist ihre Vorgehensweise jedoch so allgemein und verfahrensunabhängig, dass sie auch zusammen mit anderen Anonymisierungsverfahren für stetige Werte angewendet werden könnten. Deshalb sollen die Möglichkeiten der Behandlung dieser Merkmale hier zusammen dargestellt werden.

Folgende Varianten werden für die Behandlung diskreter Merkmale empfohlen:

a) Behandlung als stetige Merkmale mit anschließender Rundung

Bei diskreten Merkmalen (z. B. Anzahl „tätiger Personen“) besteht die Möglichkeit, diese bei der Anonymisierung wie stetige Merkmale zu behandeln und die Werte nach der Anonymisierung ganzzahlig zu runden.

Eine solche Vorgehensweise hat einerseits den Vorteil, dass sie sehr einfach zu handhaben ist. Andererseits gehen durch das Runden ggf. einige Verteilungseigenschaften, wie z. B. der Erhalt der Summen- und Durchschnittswerte, verloren. Diese Werte werden zufällig durch die Fehler der Rundungsreste überlagert. Handelt es sich um größere Merkmalswerte (nicht nahe 1), können diese Fehler jedoch oft vernachlässigt werden.

b) Behandlung als stetige Merkmale ohne anschließende Rundung (Interpretation aus dem Verfahren)

Bei manchen Verfahren werden die anonymen Merkmalswerte über Berechnungen gebildet, die eine Interpretation zulassen, ohne dass unbedingt die Ganzzahligkeit der Werte gegeben sein muss. Hier findet eine kleine inhaltliche Änderung für die diskreten Merkmale statt.

Beispiel: Mikroaggregationen erzeugen ihre Merkmalswerte innerhalb der Gruppen durch Durchschnittsbildung. Wird z. B. die Anzahl „tätiger Personen“ behandelt, so ist es ohne Probleme auch möglich, die Angaben als Durchschnittswert der Gruppe zu interpretieren und somit auch mit Dezimalstellen zu verwenden. Diese Vorgehensweise hat gegenüber Variante a) den Vorteil, dass keine zusätzlichen Fehler durch Rundungen erzeugt werden.

Für kategoriale Merkmale werden folgende Verfahren empfohlen:

c) Teilung des Mikrodatenbestandes anhand der vorhandenen kategorialen Merkmalskombinationen

Die Anwendung der Verfahren erfolgt jetzt auf jede Teilmenge getrennt. Es müssen nur noch die metrischen Merkmale behandelt werden. Anschließend werden die Teildatenbestände mit den originalen kategorialen Merkmalen wieder zusammengefügt.

Diese Variante eignet sich besonders gut, wenn es sich um sehr große Mikrodatenbestände handelt. Durch die Teilung in mehrere Teildatenbestände wird in der Regel der Rechenzeitaufwand und der Ressourcenbedarf (z. B. Hauptspeicher des Rechners) stark reduziert. Gleichzeitig gewährleistet eine solche Vorgehensweise, dass die Eigenschaften des Verfahrens bezüglich des Erhalts von Verteilungseigenschaften (Mittelwerte, Varianzen usw.) jetzt auch für die Teilmengen gelten. Nachteilig ist dieser Ansatz, wenn bei kleineren Datenbeständen oder bei sehr schiefen Verteilungen einige Teilgruppen sehr klein sind. Dann kann die anonymisierende Wirkung der Verfahren stark eingeschränkt sein und ggf. die Erreichung von Anonymität verhindert werden. Werden z. B. Gruppen erzeugt, die aus weniger als 3 Objekten bestehen, ist es bei den meisten Verfahren nicht mehr möglich, Anonymität zu erzeugen.

d) Schaffung einer Entsprechung von diskreten numerischen Werten für die kategorialen Merkmale (durchnummerieren).

Werden die kategorialen Merkmale auf diese Weise in diskrete Werte codiert, können sie anschließend auch wie stetige Werte anonymisiert und anschließend gerundet werden (siehe Variante a)), z. B. empfohlen von Dandekar für das LHS-Verfahren (Dandekar, Cohen und Kirkendall 2001).

Eine solche Vorgehensweise hat mehrere Probleme: Einerseits kann ein solches Verfahren nicht verhindern, dass Ausprägungen generiert werden, die vorher in den bearbeiteten Werten nicht vorhanden waren. So kann z. B. bei einer Mikroaggregation aus 3 Kategorien durch die Durchschnittsbildung eine völlig neue Kategorie entstehen, nur weil sie zufällig in der Liste der kategorialen Merkmalsausprägungen zwischen den 3 existierenden Merkmalsausprägungen steht. Andererseits ist es mit dieser Vorgehensweise nicht möglich, hierarchische kategoriale Merkmale zu behandeln. Werden die hierarchischen Merkmale (z. B. bei Regionalschlüsselebenen die Ebenen Gemeinde, Kreis, Land) separat behandelt, ist nicht gewährleistet, dass sich die Ergebnisse wieder in einem inhaltlich richtigen Zusammenhang befinden. Wird nur die unterste Ebene im Verfahren berücksichtigt und werden anschließend die höheren Ebenen aus Referenzlisten neu erzeugt, sind zwar die Zusammenhänge richtig dargestellt, es ist im Verfahren aber keine Kontrolle möglich, ob sich die Auswirkungen der Anonymisierung in den höheren Merkmalsebenen (Kreis, Land) in Grenzen halten. Es kann dann nicht verhindert werden, dass die Veränderungen auf höherer Aggregationsebene zufällig zu starken Verschiebungen führen.

e) Definition von „Abstandsmaßen“

Bei der Definition von numerischen Entsprechungen für die Veränderung von Schlüsseln werden „Abstandsmaße“ definiert, mit denen die Veränderung von kategorialen Merkmalen bewertet werden kann. Damit können Anonymisierungsverfahren, die

Abstandsmaße für die Beurteilung von Veränderungen nutzen, auch kategoriale Merkmale behandeln. Das Verfahren wurde von Torra (Torra 2004) für Mikroaggregationen vorgeschlagen, bei denen so der Abstand zwischen zwei Merkmalsträgern einschließlich der kategorialen Merkmale bestimmt wurde.

Soll z. B. das Regionalmerkmal „Bundesland" mit anonymisiert werden, so könnte der Abstand zwischen zwei Einheiten folgendermaßen gesetzt werden:

0 – wenn beide Einheiten aus dem gleichen Bundesland sind,

0,5 – wenn beide Einheiten aus verschiedenen Bundesländern aber aus der gleichen Region (z.B. Ost/West) sind und

1 – wenn beide Einheiten aus verschiedenen Regionen sind.

Dieser Ansatz hat einerseits den Vorteil, dass keine Transformation der Merkmale erforderlich ist. Andererseits muss die Festlegung der „Abstände" und die Bestimmung der „Durchschnitte aus verschiedenen Ausprägungen" stets individuell für jeden Fall einzeln gelöst werden. Die Gewichtung der verschiedenen „Abstandsmaße" der kategorialen Merkmale untereinander, als auch im Verhältnis zu den metrischen Merkmalen beeinflussen natürlich das Ergebnis der Mikroaggregation. Außerdem erhält man durch das Verfahren zwar eine Bestimmung der Objekte, die in einer Gruppe zusammenzufassen sind, aber noch keine Empfehlung, welche kategoriale Ausprägung für die Gruppe als Repräsentant zu übernehmen ist.

f) Transformation in stetige Merkmale nach Sullivan

Diese Vorgehensweise ist die Grundlage der Behandlung kategorialer Merkmale im Verfahren von Sullivan (siehe Sullivan 1989; Brand 2000). Die kategorialen Merkmale werden über folgendes Verfahren in gleichverteilte stetige Merkmale transformiert:

1) Es wird für *a-1* Ausprägungen des kategorialen Merkmals eine neue Spalte der Datenmatrix hinzugefügt. Dabei erhält das *i*-te Element der Spalte *l* den Wert 1, wenn das Objekt *i* die Ausprägung *l* besitzt und 0 wenn es die Ausprägung nicht besitzt. Damit hat man im ersten Schritt die diskreten Ausprägungen (0 oder 1) erzeugt. Die letzte Merkmalsausprägung *a* wird nicht hinzugefügt, da sie sich leicht aus der linearen Abhängigkeit:

$$x_{ia} = 1 - \sum_{l=1}^{a-1} x_{il}$$ ermitteln lässt.

2) Für jede kategoriale Merkmalsausprägung *l* wird die Häufigkeit ihres Auftretens h_l bestimmt.

3) In der neuen Spalte *l* werden für die kategoriale Merkmalsausprägung *l* folgende stetigen Werte am Objekt *i* eingetragen:

$$D_{il} = \left\{ \begin{array}{ll} Z\,(1-h_l) & ;wenn\ Ausprägung\ l\ nicht\ vorhanden \\ 1-h_l(1-Z) & ;wenn\ Ausprägung\ l\ \ vorhanden \end{array} \right\} \quad ;l = 1,2,...,a-1$$

Dabei ist *Z* eine gleichverteilte Zufallszahl zwischen 0 und 1, die für jede Merkmalsausprägung und jedes Objekt neu erzeugt wird. Somit wird auch D_{il} eine gleichverteilte Zufallszahl zwischen 0 und 1, für die gilt $D_{il}<(1-h_l)$, wenn die Ausprägung nicht vorhanden ist, sonst $D_{il}\geq(1-h_l)$.

4) Die eigentliche Anonymisierung erfolgt durch die Anwendung von Verfahren für stetige Merkmale. Es werden nur die stetigen Merkmale und die neu eingeführten transformierten kategorialen Merkmale D_l (sind jetzt ebenfalls stetig) anonymisiert. Die originalen kategorialen Merkmale bleiben unberücksichtigt. Im Verfahren von Sullivan wird zur Anonymisierung eine stochastische Überlagerung angewendet. Diese Art der Behandlung kategorialer Merkmale ist aber auch mit jedem anderen Anonymisierungsverfahren für stetige Merkmale (z. B. Mikroaggregationen) kombinierbar. Eine Kombination mit Verfahren, die auch direkt auf die kategorialen Variablen anwendbar sind (z. B. Swappingverfahren), ist jedoch nicht sinnvoll.

5) Rücktransformation der Merkmale

   Sie erfolgt mit folgender Regel. Die maskierten stetigen Merkmale werden wieder in eine diskrete Verteilung (0,1) transformiert über:

$$x_{il} = \begin{cases} 0 & ;wenn\ D_{il} \in (0\,,1-h_l) \\ 1 & ;wenn\ D_{il} \in [1-h_l\,,1) \end{cases} \quad ;l = 1,2,...,a-1$$

   Erhalten nach der Rücktransformation für ein Merkmal mehrere Ausprägungen x_{il} den Wert 1, so wird die Ausprägung l verwendet, für die D_{il} maximal ist und $x_{il}=1$ gilt. Erhält kein x_{il} den Wert 1, so wird die a-te Ausprägung dem Merkmal zugeordnet.

Die Vorgehensweise von Sullivan scheint die eleganteste zu sein, da die Merkmale in echte stetige Merkmale transformiert werden. Nachteilig sind jedoch folgende zwei Aspekte:

- Vergrößerung des Geheimhaltungsproblems

  Da jede Merkmalsausprägung eine neue Variable erzeugt, wird die Anzahl der Variablen sehr stark vergrößert (um die Summe aller Ausprägungen kategorialer Merkmale). Das führt in der Regel bei großen Datenbeständen zu technischen Problemen bei der Berechnung, in Einzelfällen auch zu Lösbarkeitsproblemen in der Geheimhaltungsaufgabe (siehe Sullivan 1989).

- Erzeugung von vorher nicht vorhandenen Ausprägungskombinationen

  Es ist auch bei diesem Verfahren möglich, dass die originale Merkmalsausprägung durch eine neue Ausprägung ersetzt wird, die mit den anderen Merkmalen eine eigentlich unmögliche Ausprägungskombination darstellt. Das Verfahren erhält zwar möglichst gut die Abhängigkeiten zwischen den transformierten Merkmalen und somit indirekt auch zwischen den originalen Merkmalen, kann einzelne Ausreißer (exotische Kombinationen) jedoch nicht verhindern. Es können somit ebenso unmögliche Kombinationen sowohl in den Merkmalshierarchien als auch zwischen verschiedenen Merkmalen entstehen (siehe Erläuterungen zu d)).

Dadurch ist dieses Verfahren nur eingeschränkt nutzbar. Es ist meistens eine Beschränkung der mit dem Verfahren behandelten Merkmale erforderlich.

Abschließend kann festgestellt werden, dass es eine Reihe von Möglichkeiten gibt, kategoriale und diskrete Merkmale wie stetige Merkmale in die Anonymisierungsverfahren mit

einzubeziehen. Da alle Verfahren ihre Vor- und Nachteile besitzen, kann nur im konkreten Fall entschieden werden, welches Verfahren am besten geeignet ist, oder ob diese Merkmale durch spezielle Verfahren wie z. B. Gruppierung oder PRAM (siehe S. 37 ff.) behandelt werden können. Eine Vernachlässigung der Merkmale im Laufe der Anonymisierung sollte aber vermieden werden, da diese Merkmale meistens für ein hohes Reidentifikationsrisiko verantwortlich sind.

Im Folgenden werden die verschiedenen Gruppen von datenverändernden Anonymisierungsverfahren näher vorgestellt.

2.2.2 Zufallsüberlagerung

Eine besonders verbreitete Verfahrensgruppe zur Anonymisierung von Einzeldaten ist die Überlagerung mit Zufallszahlen und somit eine Erzeugung von Zufallsfehlern in den Daten. Grundprinzip der Zufallsüberlagerung ist die Erzeugung von neuen Merkmalswerten durch ein stochastisches Verfahren aus den originalen Werten (Brand 2000).

Für stetige Merkmale gibt es eine ganze Reihe von Verfahren, die mit dem Prinzip der Zufallsüberlagerung arbeiten (Kim/Winkler 2003 und Sullivan 1989). Bei stetigen Merkmalen wird der neue Merkmalswert in der Regel aus dem originalen und einer Zufallszahl durch Addition oder Multiplikation generiert. Das bedeutet, dass die anonyme Lösung entweder durch

$X^a = X^o + W$ (additive Überlagerung)

oder durch

$X^a = X^o \odot W$ (multiplikative Überlagerung)

mit $\odot$ – Hadarmard-Produkt für elementweise Multiplikation

erzeugt werden. Dabei ist W eine Matrix aus Zufallszahlen, die die gleiche Dimension wie die Matrix der Originalwerte X^o besitzt. Die Matrix der Zufallszahlen wird so gebildet, dass folgende Annahmen gelten:

– bei additiver Überlagerung
  $E(W) = \underline{0}$ ($\underline{0}$ – Nullmatrix der Dimension $n*m$)
– bei multiplikativer Überlagerung
  $E(W) = \underline{1}$ ($\underline{1}$ – Einsmatrix der Dimension $n*m$),

  sowie

  Nichtnegativität der Elemente von W, um Vorzeichenwechsel zu verhindern.

Diese Erwartungswerte von W sind erforderlich, damit bei beiden Überlagerungsarten $E(X^a) = E(X^o)$ erhalten bleibt. Als Form der Verteilung wird für die Zufallszahlen in der Regel eine Normalverteilung, in Einzelfällen auch Gleich- oder Lognormalverteilung verwendet. Die Größe der Schutzwirkung hängt von der Standardabweichung der Verteilung ab. Die Eigenschaft der Normalverteilung, dass sie ihre größte Wahrscheinlichkeitsdichte um den Erwartungswert selbst besitzt, ist aber für die Anonymisierung problematisch. Damit werden sehr viele Originalwerte nur wenig verändert und wenige Originalwerte stark verändert. Da bei der Anonymisierung aber immer ein Kompromiss zwischen der Datensicher-

heit (erfordert Veränderungen) und der Datenqualität (ist am besten mit wenig Veränderungen zu halten) gesucht wird, scheint die echte Normalverteilung nicht gut geeignet zu sein. Es gibt deshalb eine Reihe von Ansätzen, die diesen Konflikt zu lösen versuchen.

Eine dem Problem gut entsprechende Verteilung hätte möglichst keine extremen Ausreißer, aber auch wenige Werte in unmittelbarer Nähe des Erwartungswertes. Ein Ansatz besteht in der Anwendung „gestutzter" Normalverteilungen (Kim und Winkler 2003). Hier wird die Verteilung der Zufallszahlen als gestutzte Normalverteilung (Normalverteilung mit Ausschlussbereichen) definiert. Dabei sind Bereiche nahe dem Erwartungswert und extrem außerhalb des Erwartungswertes nicht zulässig. Sollte beim Generieren einer Zufallszahl ein solcher Wert gezogen werden, so findet das Ziehen der Zufallszahl einfach noch einmal statt. Auf diese Weise kann ein Mindestabstand zwischen den originalen und anonymen Werten gesichert werden und gleichzeitig sind extreme Verschiebungen ausgeschlossen.

Ein anderer Ansatz besteht in der Verwendung von Mischungsverteilungen aus mehreren Normalverteilungen. Diese einzelnen Elemente haben zwar nicht den gesuchten Erwartungswert von 0 bzw. 1. Es werden aber mehrere Normalverteilungen so miteinander kombiniert (gemischt), dass die gewünschten Eigenschaften bezüglich des Erwartungswertes und der geringen Häufigkeit im Bereich um den Erwartungswert wieder erreicht sind (siehe Roque 2000 und Yancey 2002).

Grundsätzlich haben sich additive und multiplikative Zufallsüberlagerung fast gleichberechtigt als Anonymisierungsverfahren entwickelt und daher wird für das jeweilige Problem die entsprechend günstige Variante genutzt. Die Ursachen dafür liegen darin, dass beide Verfahrensgruppen ihre Vor- aber auch Nachteile besitzen. Für die additive Zufallsüberlagerung spricht die Möglichkeit, die Zufallszahlen so zu generieren, dass sie vorgegebenen Mittelwerten, Varianzen aber auch Kovarianzen genügen. Man nutzt diese Möglichkeit und überlagert die Originalwerte mit Zufallszahlen, die im Erwartungswert 0 und in der Varianz-Kovarianz-Matrix ($\Sigma(W)$) proportional zu den Originaldaten sind

$\Sigma(W) = d\,\Sigma(X^o)$ $\quad$ d – Parameter zur Regelung der Stärke der Überlagerung

Dann gilt auch

$$E(X^a) = E(X^o)$$

und

$$\Sigma(X^a) = (1+d)\,\Sigma(X^o)$$

Somit sind mit den anonymen Daten auch die Mittelwerte und Korrelationen erwartungstreu und lineare Modelle fehlerfrei schätzbar. Für die Korrektur der erhöhten Varianz-Kovarianz-Matrix bestehen ebenfalls Möglichkeiten, z. B. mit Hilfe der Kim-Korrektur der Daten (Bemerkungen zur Varianzkorrektur nach Kim siehe S. 61 ff.). Eine Möglichkeit, für lineare Modelle gleichwertige anonyme Daten zu erstellen, ist für multiplikative Überlagerung nicht bekannt. Der gravierende Nachteil der additiv überlagerten Daten resultiert aber aus der gleichen Eigenschaft, dass die Varianz-Kovarianz-Matrix der Überlagerungen für alle Objekte gleich ist. Dadurch ist auch bedingt, dass alle Werte mit gleich starken Zufallsfehlern überlagert werden. Diese führen bei schiefen Datenbeständen dazu, dass

entweder die großen Daten unzureichend geschützt sind, oder bei entsprechend starken Zufallsfehlern der Schutz großer Daten zwar ausreicht, kleine Daten aber völlig unbrauchbar werden. Bei ihnen wird einerseits der Zufallsfehler im Verhältnis zum Originalwert unverhältnismäßig groß und andererseits die bei vielen Merkmalen inhaltlich bedingte Nichtnegativität der Daten nicht mehr gesichert (Vorzeichenwechsel). Trotz dieser gravierenden Mängel ist das Verfahren für viele Daten die bessere Wahl, wenn der Datenbestand sich durch relativ homogene Datenstrukturen auszeichnet (z. B. Mietspiegel, Firmendaten für kleine und mittlere Unternehmen, Lohn- und Einkommensdaten). Selbst bei einigen wenigen Ausreißern im Datenbestand erreicht das Verfahren in Kombination mit Top- bzw. Bottomcoding immer noch sehr gute Ergebnisse. Die Veränderung der Ausreißerwerte auf die vorher festgelegten Maximum- bzw. Minimumwerte erzeugt wieder die erforderliche Homogenität der Daten. Bei starker Anwendung von Top-/Bottomcoding sind diese Datenbestände für makroökonomische Rückschlüsse dann allerdings weniger geeignet, weil Top-/Bottomcoding in der Regel die Mittelwerte und Varianzen systematisch nach unten verzerrt (siehe Abschnitt 2.1.3).

Multiplikative Überlagerungen gewährleisten automatisch, dass einerseits die Veränderung der Werte größenabhängig ist. Bei Auswahl ausschließlich nichtnegativer Zufallszahlen wird andererseits ein Vorzeichenwechsel automatisch verhindert. Somit ist die multiplikative Zufallsüberlagerung der additiven Zufallsüberlagerung bei sehr schief verteilten Datenbeständen in Bezug auf die Datensicherheit und die Datenplausibilität überlegen. Abhängigkeiten zwischen den Variablen des Datenbestandes werden aber wegen der fehlenden Einflussmöglichkeit auf die Varianz-Kovarianz-Matrix schlechter reproduziert.

Eine Reihe von Verfahren haben auch eine Mischung der Eigenschaften versucht. Ein Ansatz ist z. B., schiefe Datenbestände zu logarithmieren, bevor man sie einer additiven Überlagerung unterzieht. Danach werden die Daten wieder zurückgewandelt (siehe Kim und Winkler 2003). Die Formel zur Anonymisierung lautet deshalb:

$$x^a = e^{\ln(x^o)+w}$$

Problematisch ist hier einerseits die Untauglichkeit dieses Ansatzes für die Behandlung von Null- und Missingwerten. Im Vorschlag von Kim und Winkler sollten diese Werte durch einen kleinen „Aufschlag“ logarithmierbar gemacht werden. Ein Erhalt der Struktur bezüglich der Anzahl an Nullen und Missings ist so allerdings nicht möglich. Außerdem gewährleistet der Ansatz, dass nur die logarithmierten Werte in ihrer Korrelationsstruktur erhalten bleiben. Die Standardabweichungen der nicht logarithmierten Werte können nur unzureichend erhalten werden (siehe Kim und Winkler 2003, S. 12). Selbst bei kleinen Überlagerungen (1 % der Kovarianzmatrix der logarithmierten Werte – Parameter $d=0{,}01$) sind Fehler in der Standardabweichung von ca. 30 % möglich. Damit ist dieser Ansatz auch nur dann gut geeignet, wenn die gewünschten Analysen und Modelle auf logarithmierten Werten aufsetzen. Vorteilhaft an der Überlagerung der logarithmierten Werte ist, dass nach der Rücktransformation für x^a die Nichtnegativität immer gewährleistet ist.

Post Randomisation Method (PRAM)

Bei der Post-Randomisierung (PRAM) (Willenborg und de Waal 2001) werden kategoriale Merkmale durch die Definition von Übergangswahrscheinlichkeiten randomisiert. Dabei werden die Merkmale mit bei der Anwendung festzulegender Übergangswahrscheinlichkeiten in andere Ausprägungen transformiert. Die kategorialen Merkmalswerte werden wie bei der in Erhebungen verwendeten Randomisierung von Antworten (sog. Randomised Response Technique) verändert. Die veröffentlichten Werte entsprechen nur noch mit einer im Verfahren festgelegten Wahrscheinlichkeit den Ausprägungen im Originaldatensatz. Die Richtigkeit der Zuordnung von Merkmalsträgern ist dadurch mit einer erhöhten Unsicherheit behaftet, wenn die Überschneidungsmerkmale mit PRAM behandelt wurden. Wurde das Verfahren auf Merkmale angewandt, die kein Angriffswissen darstellen, so ist der Informationsgewinn reduziert, da hier die Unsicherheit besteht, ob die erhaltene Aussage wahr ist.

Da bei PRAM die neuen Merkmalswerte nicht durch einen Tausch der Ausprägungen zwischen zwei Merkmalsträgern bestimmt werden, sondern durch eine auf Zufallswerten basierende Veränderung der Merkmalskategorien gebildet werden, wird es hier zu den Zufallsüberlagerungen und nicht zu den Zufallsvertauschungen zugeordnet.

Werden die Übergangswahrscheinlichkeiten in einer Matrix zusammengestellt, so erhält man die Markov-Matrix *P*. Wird diese Matrix dem Datennutzer mit übergeben, so bestehen für ihn Korrekturmöglichkeiten wenn die Inverse P^{-1} existiert. Bezeichnet nämlich T_o die Häufigkeitstabelle der unbearbeiten Werte x^o und T_a die Häufigkeitstabelle der anonymisierten Werte x^a, so gilt:

$$E(T_a \mid x_1^a, x_2^a, ..., x_n^a) = P^t T_o$$

Bei existierender Inverse P^{-1} kann ein unverzerrter Schätzer für T_o über:

$$\hat{T}_o = (P^{-1})^t T_a$$

berechnet werden (de Wolf u. a. 1998).

Werden beim PRAM-Verfahren die Erwartungswerte der Häufigkeiten der originalen Merkmalsausprägungen erhalten, so spricht man von invariantem PRAM. Diese PRAM-Varianten führen keine Verschiebungen in den Häufigkeiten der Ausprägungen durch. Eine Bestimmungsregel für invariante PRAM ist (siehe Ronning et al. 2005):

a) Es seien h_l die Häufigkeiten der einzelnen Ausprägungen a_l in der Variable x.

b) Gegeben sei eine Wechselwahrscheinlichkeit α ($0<\alpha<1$) für die Ausprägungen.

Die Wahrscheinlichkeit für die Veränderung der Ausprägung a_k in die Ausprägung a_l ist:

$$p_{kl} \equiv P(a_l|a_k) = \alpha h_l$$

Die Wahrscheinlichkeit für den Erhalt der Ausprägung a_l bei der Ausprägung a_l ist:

$$p_{ll} \equiv P(a_l|a_l) = (1-\alpha) + \alpha h_l$$

Damit ergibt sich die Markov-Matrix als:

$$P = (1-\alpha)E + \alpha h \underline{1}^t$$

mit:

E – Einheitsmatrix.

h – Spaltenvektor der h_i als Häufigkeiten der einzelnen Ausprägungen a_i.

$\underline{1}$ – Spaltenvektor aus 1-Elementen.

In der Arbeit von de Wolf u. a. (1998) werden auch noch zwei andere Varianten für invariante Postrandomisierung vorgeschlagen. Für die direkte Berechnung einer invarianten Markov-Matrix schlagen sie folgendes Vorgehen vor:

Zuerst werden die Kategorien k absteigend nach ihren Häufigkeiten $T_o(k)$ sortiert. Die Ausprägung mit der kleinsten existierenden absoluten Häufigkeit sei K. Damit gilt $T_o(k) \geq T_o(K) > 0$ für $k=1,2,...,K$. Dann ergibt sich eine invariante Markov-Matrix für alle $k,l \leq K$ als:

$$p_{kl} = \begin{cases} 1 - \alpha T_o(K)/T_o(k) & ;wenn \quad k = l \\ \alpha T_o(K)/(K-1)T_o(k) & ;wenn \quad k \neq l \end{cases}$$

Hierbei kann α wieder als Parameter zwischen 0 und 1 gewählt werden. Die erhaltenen Werte p_{kl} bezeichnen dabei die Übergangswahrscheinlichkeit, dass eine Ausprägung k im Verfahren in die Ausprägung l geändert wird. Der Unterschied zwischen beiden Verfahren liegt darin, dass während bei dem ersten Verfahren die Nichtdiagonalelemente der Markov-Matrix innerhalb der Zeile als gleich unterstellt werden, gewährleistet das Verfahren von de Wolf, dass die Nichtdiagonalelemente einer Spalte gleich sind.

Der zweite Vorschlag von de Wolf (de Wolf u. a. 1998) stellt erst einmal keine Voraussetzung an P. Dabei wird die Invarianz einfach dadurch erzeugt, dass zweistufig einmal mit P und danach mit $P^{\leftarrow}$ die Postrandomisierung durchgeführt wird. Die zweite Markov-Matrix $P^{\leftarrow}$, die die Verzerrungen durch P wieder ausgleicht, wird dabei folgendermaßen berechnet:

$$p_{lk}^{\leftarrow} = \frac{p_{kl} T_o(k)}{\sum_j p_{lj} T_o(j)}$$

Wird ein Datenbestand, der mit einer Markov-Matrix P anonymisiert wurde, noch einem zweiten PRAM mit Markov-Matrix $P^{\leftarrow}$ unterzogen, so sind die Verzerrungen wieder korrigiert. Das Produkt der beiden Matrizen $R = PP^{\leftarrow}$ stellt somit eine invariante Matrix dar.

Für alle drei Varianten von invariantem PRAM gilt bei der Anwendung folgendes zu beachten:

- Datenintegrität innerhalb von Sätzen

  Bei vielen kategorialen Merkmalen bestehen Abhängigkeiten innerhalb von Sätzen. Beim Anwenden von PRAM sollte verhindert werden, dass durch die Merkmalsveränderung unsinnige Ausprägungskombinationen entstehen. Um Kombinationen wie „Steinkohlebergbau in Berlin“, „Hochseefischerei in Bayern“ oder ähnliches zu verhindern, bestehen zwei Möglichkeiten. Die kategorialen Merkmale (wie z. B. Region und Wirtschaftszweig) können einerseits zusammen, d. h. als Ausprägungskombinationen, betrachtet werden. Damit können auch bei der Anonymisierung nur bereits

vorher existierende Kombinationen erzeugt werden, da beide Ausprägungen gleichzeitig verändert werden. Der Nachteil besteht darin, dass die Häufigkeiten der Ausprägungskombinationen oft sehr klein sein können und damit nur sehr kleine Übergangswahrscheinlichkeiten existieren. Ein solches Vorgehen kann deshalb zu einer sehr hohen Varianz der mit den anonymisierten Daten geschätzten Häufigkeiten führen. Die zweite Variante besteht darin, das Verfahren immer nur auf Blöcke des Datenbestandes anzuwenden. Werden für den PRAM-Austausch Blöcke von Merkmalsausprägungen bestimmt, für die ähnliche Eigenschaften gelten, sollte auch der Tausch von Ausprägungen innerhalb der Blöcke die Datenintegrität erhalten. Sortiert man die Ausprägungen nach diesen Gruppen, so ergibt sich die Markov-Matrix aus einzelnen kleinen Blöcken entlang der Hauptdiagonale, während die Blöcke außerhalb nur 0-Elemente enthalten.

Neben der Verletzung der Datenintegrität innerhalb von Sätzen können auch Zusammenhänge zwischen den Sätzen verletzt werden. Bei Betriebserhebungen sollten alle Betriebe eines Unternehmens in den unternehmenspezifischen Merkmalen (WZ-Klassifikation, Rechtsform u. Ä.) gleiche Ausprägungen haben. Das kann jedoch nur erhalten werden, wenn diese Zusammenhänge von vornherein berücksichtigt werden.

- Sicherheitsprobleme

  Neben der Datenintegrität bestehen auch Sicherheitsprobleme bei der Anwendung von PRAM. Wenn ein Datenbestand Hochrechnungsfaktoren enthält, so weisen identische Hochrechnungsfaktoren darauf hin, dass vor der PRAM-Anonymisierung einheitliche Ausprägungen in den Merkmalen bestanden, die für die Schichtung herangezogen wurden. Die jetzt im Datenbestand dominierenden Ausprägungen sind dabei mit hoher Wahrscheinlichkeit die richtigen.

  Ein anderes Problem besteht in der Bereitstellung von mehreren gleichartig anonymisierten Lösungen. Diese können einerseits durch den Wunsch von Wissenschaftlern nach mehreren projektbezogenen anonymisierten Daten (jeweils einschließlich des PRAM-anonymisierten Merkmals) für verschiedene Projekte entstehen. Ein anderes Problem ist die Verwendung von PRAM für die Anonymisierung kategorialer Merkmaler bei multipler Imputation (siehe Abschnitt 2.2.5), um Wissenschaftlern zu ermöglichen, die Varianzschätzungen selbst vorzunehmen. Wenn an Stelle der Transformationsmatrix verschiedene anonyme Lösungen bereitgestellt werden, besteht ein erhöhtes Schutzrisiko dann, wenn die Tauschwahrscheinlichkeit im Verhältnis zur Bleibewahrscheinlichkeit relativ klein gewählt wird. Wird versucht, bei der Postrandomisierung den Tausch der Ausprägungen nur sehr sparsam einzusetzen, so besitzen die Merkmalsträger eine geringere Wahrscheinlichkeit, dass ihre Ausprägung getauscht wurde, als dass sie unverändert bleibt. Werden jetzt mehrere gleichartige Lösungen bereitgestellt, kann die am häufigsten auftretende Merkmalsausprägung für einen Merkmalsträger mit ziemlicher Sicherheit als die Originalausprägung unterstellt werden (siehe auch de Wolf u. a. 1998). Eine gleichwertige Sicherheit gegenüber der Herausgabe der Transformationsmatrix kann bei der Bereitstellung mehrerer anonymer PRAM-Lösungen nur dann gewährleistet werden, wenn die Wechselwahrscheinlichkeiten in der Transformationsmatrix stark erhöht werden. Diese Vorgehensweise würde aber dem eigentlichen Ziel wiedersprechen, die Veränderungen durch die Anonymisierung möglichst klein zu halten.

2.2.3 Zufallsvertauschungen

Grundprinzip der Zufallsvertauschungen ist die Verschiebung von existierenden Merkmalsausprägungen zwischen verschiedenen Merkmalsträgern. Diese Verschiebungen werden durch stochastische Faktoren beeinflusst (zufällige Wahl des Tauschpartners), damit sie nicht reproduziert und somit rückgängig gemacht werden können. Sowohl das Auffinden der gesuchten Merkmalskombinationen als auch dann erhaltene Zusatzinformationen sind nur mit einer gewissen Unsicherheit richtig. Zufallsvertauschungen sind je nach Verfahrensart sowohl für kategoriale als auch metrische Merkmalsträger möglich.

– **Einfaches Data-Swapping**

  Beim einfachen Data-Swapping (Boyd und Vickers 1999) werden die Merkmalsträger inhaltlich, d. h. anhand ausgewählter kategorialer Merkmale, gruppiert. Die übrigen Merkmalswerte werden dann innerhalb der Gruppen für jedes Merkmal getrennt zufällig getauscht.

– **Rank-Swapping**

  Nach einem absteigenden Sortieren der Datensätze nach einem Merkmal j werden die Merkmalswerte in der Datenspalte X_j in einem festgelegtem Nachbarschaftsbereich $P(a)$ zufällig getauscht:

  $$|i\text{-}l| < P(a)*n/100$$

  mit:

  $P(a)$ – Anteil des Datenbestandes in dem getauscht wird.

  n – Größe des Datenbestandes.

  i,l – i-tes oder l-tes Element nach absteigender Sortierung.

  Der Tausch der Werte erfolgt damit durch:

  $$x^a_{i,j} = x^o_{l,j} \quad ;i = 1,2,...,n$$
  $$x^a_{l,j} = x^o_{i,j} \quad ;i < l$$

  Das i-te Element ist dabei die aktuell bearbeitete Position im Datenbestand und l wird durch Zufallsauswahl im Nachbarschaftsbereich $P(a)$ gewählt. Wurde das i-te Element bereits als Tauschpartner verwendet, so wird $i+1$ zur aktuellen Position. Wurde das l-te Element bereits getauscht, so wird das erste noch nicht getauschte Folgeelement zu l verwendet.

  Damit werden möglichst ähnliche Merkmalswerte getauscht und so die Rangstatistiken gut erhalten. Die Bearbeitung erfolgt für jedes Merkmal getrennt (siehe Dalenius und Reiss 1982 oder Moore 1996).

– **Swapping nach Carlson und Salabasis**

  Als neueres Swappingverfahren ist das Data-Swapping nach Carlson und Salabasis (2002) bekannt. Der Datenbestand wird in einem ersten Schritt in zwei Merkmalsgruppen geteilt. Die unkritischen Merkmale (nicht zu bearbeiten) seien die ersten

k Merkmale des Datenbestandes ($k<m$). Dann werden die übrigen Merkmale durch folgende Prozedur geschützt, die für jedes zu anonymisierende Merkmal j $(j=k+1,k+2,...,m)$ durchzuführen ist.

1) Der Datenbestand wird zufällig in zwei Teilbestände X^{o1} und X^{o2} aufgeteilt.
2) Beide Teilbestände werden nach dem Merkmal j absteigend sortiert. Da beide Teilbestände unabhängige Stichproben des Gesamtbestandes darstellen, sind sie auch näherungsweise gleich verteilt.
3) Die anonymen Werte für die Spalte j ergeben sich als:

$$x_{i,j}^{a1} = x_{i,j}^{o2} \quad ;i = 1,2,...,n/2$$

$$x_{i,j}^{a2} = x_{i,j}^{o1} \quad ;i = 1,2,...,n/2$$

4) Die beiden Teildatenbestände werden wieder zusammenkopiert und gemischt.

Carlson und Salabasis liefern auch Simulationsergebnisse, die das asymptotische Verhalten und die gute Qualität der Ergebnisse zeigen.

Die näherungsweise gleiche Verteilung führt vor allem im Bereich der kleinen Werte dazu, dass $x_{i,j}^{o1} \approx x_{i,j}^{o2}$ gilt. Wenn die Werte jedoch sehr ähnlich sind, so erzeugt der Austausch der Werte keine ausreichende Schutzwirkung mehr, da auch die getauschten Werte gute brauchbare Schätzungen für die Originalwerte darstellen.

Es besteht auch die Möglichkeit, Zufallsvertauschungen und Zufallsüberlagerungen zusammen anzuwenden. Ein bekanntes Verfahren ist das Verfahren von Winkler (siehe Kim und Winkler 1995). Merkmalsträger, die durch Zufallsüberlagerung nicht genügend anonymisiert werden können, werden dabei nachträglich noch einem Data-Swapping unterzogen. Hier werden somit die Vorteile der beiden Verfahren gekoppelt, indem die höhere Datensicherheit, die durch Data-Swapping erreicht wird, mit der höheren Datenqualität der Zufallsüberlagerung verbunden wird, da nur bei den Problemfällen Data-Swapping eingesetzt wird.

Während die oben aufgeführten Entwickler des Verfahrens mit den Ergebnissen des Rank-Swapping sehr zufrieden waren, ergaben verfahrensvergleichende Untersuchungen (siehe Ronning et al. 2005) völlig unzureichende Ergebnisse mit dem Rank-Swapping-Verfahren. Dieser scheinbare Widerspruch kann mit Hilfe der Arbeit von Moore (siehe Moore 1996) aufgeklärt werden. Moore leitet einen Beweis über den Einfluss des Rank-Swapping auf die Datenqualität in Form der Korrelation zwischen den Merkmalen her.

$$E[R(x._j^a, x._k^a)] = R(x._j^o, x._j^a)\, R(x._k^o, x._k^a)\, R(x._j^o, x._k^o)$$

$x._j^o, x._k^o$ – Spaltenvektoren der beiden originalen Variablen j und k.

$x._j^a, x._k^a$ – Spaltenvektoren der beiden anonymen Variablen j und k.

$R(..,..)$ – Korrelation zwischen den angegebenen Variablen.

Dieser Zusammenhang, der von Moore in Moore (1996, S. 8 ff. – siehe auch den Anhang zu diesem Band, S. 134 ff.) bewiesen wird, bedeutet, dass die Verzerrung der Korrelation zwischen zwei anonymisierten Variablen entscheidend durch den Erhalt des korrelativen Zusammenhangs zwischen den Variablen im Original und in der anonymisierten

Form bestimmt wird. Außerdem wird eine Formel zur Abschätzung der Korrelationsverlustes in $R(x._j^o, x._j^a)$ durch das Rank-Swapping hergeleitet, die auch zur Bestimmung von möglichen Swappingbereichen $P(a)$ bei Vorgabe der gewünschten Korrelation $R_0^{1/2}=R(x._j^o, x._j^a)$ oder der relativen Veränderung der Einzelwerte K_0 benutzt werden kann.

Für einen vorgegebenen Korrelationsverlust ($1\text{-}R_0$) (zu erhaltende Korrelation $R_0^{1/2}=R(x._j^o, x._j^a)$) gilt:

$$P(a) = 100 * \frac{\sqrt{2\,Var(x._j^o)\,(1-R_0)}}{(x._{j_{topc}}^o - x._{j_{botc}}^o)} = 100 * \sqrt{2(1-R_0)}\,\frac{\sigma(x._j^o)}{(x._{j_{topc}}^o - x._{j_{botc}}^o)}$$

mit:

$x._{j_{topc}}^o$ *und* $x._{j_{botc}}^o$ – Maximal- und Minimalwert in der Variablenspalte *j*
(In der Arbeit von Moore wurden sie durch vorheriges Topcoding und Bottomcoding gesetzt.)

$Var(x._j^o)$, $\sigma(x._j^o)$ – Varianz bzw. Standardabweichung im Vektor $x._j^o$

$1\text{-}R_0$ – Gesamtkorrelationsverlust durch die Rank-Swapping-Anonymisierung der Variablen $x._j^o$ und $x._k^o$. Es wird deshalb unterstellt, dass $R_0^{1/2}=R(x._j^o, x._j^a)$ gilt.

Für ein vorgegebenes K_0 (relative Mindestveränderung der Einzelwerte) gilt:

$$P(a) = 100\,\sqrt{\frac{8}{3}}\,\frac{K_0\;\overline{x._j^o}}{(x._{j_{topc}}^o - x._{j_{botc}}^o)}$$

mit:

K_0 – Vorgabe für die minimale relative Veränderung der Werte
$K_0 = |\,x._j^a - x._j^o\,| / x._j^o$

$\overline{x._j^o}$ – Mittelwert von $x._j^o$

Damit hängt der Swappingbereich $P(a)$ neben den Vorgabeparametern vom Verhältnis aus Spannweite zur Standardabweichung bzw. Spannweite zum Mittelwert der jeweiligen Variablen ab. Diese Verhältnisse sind jedoch stark vom konkret untersuchten Datenbestand abhängig. Während bei dem von Moore untersuchten Public-Use-File des „Annual Housing Survey" (1993) die Spannweite das ca. 5- bis 11-fache der Standardabweichung betrug, war es bei der von Ronning et al. (2005) untersuchten Kostenstrukturerhebung das 50- bis 100-fache der Standardabweichung. Im Verhältnis zum Mittelwert betrug die Spannweite beim „Annual Housing Survey" ca. 3- bis 16-fache bei der Kostenstrukturerhebung wieder das 300- bis 2 700-fache. Dabei wurde für die Kostenstrukturerhebung das Merkmal der Bestandsveränderungen bereits vernachlässigt, da es durch starke Schwankungen um Null gekennzeichnet war, was sich bei der Berechnung von Verhältnisgrößen zum Durchschnitt negativ auswirkte.

Beim Versuch der Berechnung des Swappingbereiches und der Veränderung der Einzelwerte ergab sich für vorgegebene Korrelationsverluste von 10 % bzw. 1 %:

Korrelationsverlust ($1- R_0$)	10 %	1 %
R_0	0,9	0,99
Datenveränderung: K_0	3,286	1,039
Swappingbereich: $P(A)$	0,545 %	0,172 %
in Sätzen	92	29

Damit dürften nur jeweils 0,545 % bzw. 0,172 % des Datenbestandes als Tauschpartner für die Rank-Swapping-Prozedur verwendet werden, wenn der vorgegebene Korrelationsverlust nicht überschritten werden soll. Bei den Berechnungen in Ronning, et al. (2005) lagen die Werte bedeutend höher. Da diese kleinen relativen Anteile auch nur 92 bzw. 29 Sätze des Gesamtdatenbestandes bedeuten, wird klar, dass eine solche Anwendung des Rank-Swapping nicht praktikabel ist. Sie würde zwar den großen Werten des Datenbestandes einen ausreichenden Schutz liefern. Für kleine Merkmalsträger ist die Veränderung der Daten aber völlig unzureichend, da hier sehr viele Sätze so ähnlich sind, dass ein Swappingbereich von maximal 29 sortierten Sätzen diese Merkmalsträger nicht bzw. nicht ausreichend verändert. Dass die Formeln von Moore jedoch eine mittlere Datenveränderung um das 3-fache (bei 10 % Korrelationsverlust) und immer noch um das 1-fache (bei 1 % Korrelationsverlust) erwarten, liegt darin begründet, dass die vereinfachende Hypothese einer Gleichverteilung der Werte in den Swappingbereichen die Grundlage der Herleitungen war. Diese weicht bei der Kostenstrukturerhebung jedoch sehr stark von den realen Gegebenheiten ab. Die Aussage von Moore, dass sich beide Formeln trotz der vereinfachenden Hypothese gut für die Prognose der Datenveränderungen und die Abschätzung der Intervalle eignen, wurde mit den KSE-Daten nicht bestätigt.

Im Ergebnis kann man feststellen, dass sich Datenbestände nicht mit Rank-Swapping bearbeiten lassen, wenn sie sehr schief verteilt sind. Ein Ausweg würde ein vorheriges Anwenden von Top-/Bottomcoding-Regeln bilden. Dann müsste man allerdings bei der Beurteilung des Verfahrens auch die gemeinsame Anwendung beider Verfahren betrachten. Hier wäre nämlich primär die einseitige Verzerrung durch das Topcoding zu erwähnen. Auf Grund der typischerweise sehr schiefen Verteilung von Wirtschaftsdaten ist es üblicherweise so, dass das Topcoding immer im stärkeren Umfang als das Bottomcoding für die Anonymisierung angewendet werden muss. Für die meisten Variablen gilt in der Regel eine Nichtnegativität der Werte bereits per Definition, d. h. nur wenige Variablen (wie z. B. die Veränderung an Lagerbeständen) haben auch Ausreißer am unteren Rand des Wertebereiches. Die meisten wirtschaftsstatistischen Variablen besitzen die Ausreißer nur am oberen Rand des Wertebereiches, da die Nichtnegativität per Definition bedeutet, dass an der unteren Grenze des Wertebereiches keine vereinzelten Ausreißer sondern eine Häufung vieler kleiner Einheiten mit Werten nahe Null existiert. Damit ist die Anwendung von Bottomcoding meist nicht erforderlich. Die ausschließliche Anwendung von Topcoding bedingt aber auch immer eine einseitige Verzerrung der Mittelwerte und der Standardabweichungen. Sie werden systematisch verkleinert. Diese einseitige Verzerrung muss beim gemeinsamen Anwenden von Rank-Swapping mit Top-/Bottomcoding auch stets mit bewertet werden, weil auf diese Maßnahmen in der Regel nicht verzichtet werden kann. Deshalb führt Rank-Swapping trotz der theoretisch herleitbaren guten Eigenschaften selten zu brauchbaren Ergebnissen bei wirtschaftsstatistischen Daten,

wenn man die Anonymisierung insgesamt bewertet. Für „Spezialfälle“ wie z. B. Mietdaten oder Einkommensdaten (wie im Falle von Moore) kann Rank-Swapping aber nutzbar sein, wenn gewährleistet ist, dass die Daten nicht zu schief verteilt sind.

2.2.4 Simulationsverfahren

Völlig synthetische Merkmalsträger erzeugen Simulationsverfahren. Dazu wird meist ein stochastisches Verfahren verwendet. Bei Simulationsverfahren muss die Anzahl der synthetischen Merkmalsträger nicht mit der Anzahl im originalen Datenbestand übereinstimmen. Es lassen sich durch Simulation auch viel kleinere oder größere Testdatenbestände erzeugen. Die Testdaten lassen sich dann nicht mehr auf die Originaldaten zurückführen.

– Resampling

  Das Resamplingverfahren (Fienberg 1997) basiert auf der Idee, die mehrdimensionale Kerndichte des gesamten Datenbestandes zu schätzen. Mit Hilfe dieser Dichte wird dann die gewünschte Anzahl der synthetischen Datensätze erzeugt. Das Verfahren ist bei stetigen Variablen nur sehr schwer einsetzbar. Es gibt auch noch keine konkreten Erfahrungen mit dieser Methode.

– Latin Hypercube Sampling (LHS)

  Beim Latin Hypercube Sampling (Dandekar, Cohen und Kirkendall 2002) erfolgt zuerst ausgehend von der Anzahl an gewünschten synthetischen Datensätzen eine Simulation der eindimensionalen Merkmalswerte. Diese werden mit Hilfe der geglätteten empirischen Verteilungsfunktion oder einer theoretischen Verteilungsfunktion für die einzelnen Variablen aus gleichverteilten Zufallswerten erzeugt. Diese synthetischen Merkmalswerte werden in einem zweiten Schritt durch ein Swapping-Verfahren so umgeordnet, dass die Rangkorrelationen approximiert werden.

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

Simulationsverfahren scheinen bezüglich der Anonymität die sichersten Verfahren zu sein. Gleichzeitig ermöglichen sie, durch die Beeinflussung der bei der Simulation mit heranzuziehenden Eigenschaften der Originaldaten, „bedarfsgerechte“ Eigenschaften zu erzeugen. Da die Simulationen jedoch relativ unabhängig von den Originalwerten erfolgen, sind die nicht berücksichtigten Eigenschaften auch nur sehr „zufällig“ erhalten bzw. nicht erhalten. Bei den Tests von Simulationsverfahren mit Originaldaten schnitten sie deshalb im Vergleich zu anderen Verfahren relativ schlecht ab, während die Darstellung bei den Veröffentlichungen der Entwickler besser abschließt. Ein typisches Beispiel ist z. B. das LHS-Verfahren von Dandekar: Grundlage der Simulation ist hier die Erzeugung von unabhängigen Wertespalten, die den eindimensionalen Verteilungen der originalen Variablen genügen. Da diese Werte im zweiten Schritt des Verfahrens so miteinander kombiniert werden, dass die Rangkorrelationen möglichst gut reproduziert werden, ist im Ergebnis die eindimensionale Verteilung gut und Rangkorrelation zufriedenstellend erhalten. Die Bravais-Pearson-Korrelation, aber auch die Verteilung innerhalb von Teilmassen sind völlig unzureichend reproduziert (siehe z. B. Ronning et al. 2005).

2.2.5 Imputationsverfahren

Imputationsverfahren erzeugen synthetische Merkmalswerte. Die Grundlagen von Imputationsverfahren sind durch die Methoden gelegt, die in der Nonresponseforschung für Antwortausfälle verwendet werden. Im Gegensatz zur Imputation fehlender Werte werden bei der Anonymisierung besonders schutzwürdige Informationen erst aus dem Datenbestand entfernt und dann wie fehlende Werte wieder eingeschätzt. Diese Idee wurde zuerst von Rubin (1993) vorgeschlagen. Die Imputationen können auf der Basis von Regressionsmodellen erfolgen. Neben der einfachen Imputation (einmalige Schätzung auf der Grundlage eines Modells) ist auch die multiple Imputation möglich. Dabei werden die Schätzungen mit mehreren Bootstrap-Stichproben und ggf. auch mehreren Modellen durchgeführt, so dass dem Datennutzer mehrere anonymisierte Datensätze zur Verfügung stehen (Raghunatan et al. 2003). Neuere Versuche mit multipler Imputation gehen von einem stufenweisen Entfernen aller Originalwerte aus, so dass im Ergebnis nur noch synthetische Merkmalswerte vorhanden sind. Man erhält dadurch völlig synthetische Merkmalsträger wie bei den Simulationsverfahren.

Imputationsverfahren haben zwei grundsätzliche Probleme: Da die für die Imputation unterstellten Modelle häufig den bei der Analyse später verwendeten Modellen ähneln, führen die Imputationen zu einer systematischen Überschätzung des Bestimmtheitsmaßes. Die Stärke der Überschätzung hängt dabei davon ab, in welchem Umfang Imputationen vorgenommen wurden. Handelt es sich bei den verwendeten Modellen um enge Zusammenhänge in Originaldaten, so ist außerdem die Schutzwirkung nicht sehr stark, weil die erhaltenen Schätzungen dem Original sehr ähnlich sind.

Multiple Imputationen ermöglichen es dem Datennutzer, auch einen Eindruck über die durch die Imputation erzeugte Streuung in den Daten zu erhalten. Hier besteht zwar ein Vorteil für die Nutzer der Daten, aber auch ein erhöhtes Sicherheitsrisiko. Ergeben die multiplen Imputationen bei bestimmten Einzelangaben nur eine sehr kleine Streuung, kann sich der Datenangreifer auch relativ sicher sein, dass die metrischen Angaben dem Original sehr ähnlich bei kategorialen Merkmalen sogar identisch zum Originalwert sind. Zu viele Imputationen erhöhen also wieder das Re-Identifikationsrisiko.

2.2.6 Mikroaggregation

Grundprinzip von Mikroaggregationen ist die Gruppierung von möglichst ähnlichen Merkmalsträgern und deren Vereinheitlichung durch das Ersetzen der gruppierten Merkmalswerte durch ihren Durchschnitt.

Fast alle Verfahren dieser Gruppe (bis auf SAFE) lassen die kategorialen Merkmale unbehandelt. Sollen diese mit behandelt werden, so stehen aber die am Anfang des Kapitels beschriebenen Möglichkeiten (siehe Abschnitt 2.2.1) zur Verfügung.

Im Folgenden werden allgemeine Eigenschaften von Mikroaggregationsverfahren dargestellt. Die dabei beschriebenen numerischen Probleme zur Bestimmung einer exakten Lösung führen zu zwei grundlegenden Verfahrensgruppen, die anschließend ausführlicher dargestellt werden. Bei den eindimensionalen Mikroaggregationen werden die Probleme durch eine eindimensionale Transformation der Merkmalsträger umgangen. Die mehrdimensionalen Mikroaggregationen sind Näherungsverfahren zur Bestimmung einer Lösung.

Alle Gruppierungsverfahren gehen von Gruppengrößen von mindestens drei Werten ($k \geq 3$) aus, da bei nur zwei Merkmalsträgern das Risiko der Reidentifikation eines Merkmalsträgers bei Kenntnis der Werte des anderen Merkmalsträgers noch vorhanden ist. Die Möglichkeit, dass ein Merkmalsträger selber Datenangreifer ist, bzw. jemanden mit dem Datenangriff beauftragt, führt automatisch zu der Annahme, dass dem Datenangreifer das Wissen über die Werte mindestens eines Merkmalsträgers (seine eigenen) unterstellt werden muss (Fallzahlproblem).[7]

Die Vereinheitlichung in den Gruppen (durch Durchschnittsbildung) erfolgt theoretisch über alle Merkmalswerte gleichzeitig. Dann könnten Datenangreifer keine eindeutige Zuordnung mehr vornehmen, da für jede beliebige Merkmalskombination beim Angriffswissen mindestens 3 anonymisierte Einheiten gleich gut entsprechen. Es ist aber auch immer möglich, die Verfahren für Teilmengen der Merkmale separat durchzuführen (Blockung der Merkmale).

Mikroaggregationsverfahren reduzieren die Möglichkeit der eindeutigen Zuordnung der Merkmalsträger, weil durch die Vereinheitlichung innerhalb der Gruppen mehrere Merkmalsträger gleiche Merkmalswerte erhalten. Gleichzeitig erzeugt die Durchschnittsbildung eine Unsicherheit in den Daten, die den Wert der Information für den Datenangreifer reduziert. Beide Schutzwirkungseffekte treten im Datenbestand unterschiedlich auf. Wenn keine Durchschnittsbildung über alle Merkmale gleichzeitig erfolgt, bewirkt die Durchschnittsbildung im Bereich der vielen kleinen und sehr ähnlichen Einheiten kaum eine Veränderung der Daten. Die Möglichkeit der richtigen, eindeutigen Zuordnung wird aber noch weiter erschwert. Im Bereich der großen Einheiten kann die richtige Zuordnung weiterhin kaum verhindert werden. Hier bewirkt die Durchschnittsbildung wegen der großen Unterschiede zwischen den Einheiten jedoch, dass der Nutzen der anonymisierten Daten wegen ihrer starken Veränderung für den Datenangreifer nicht mehr besteht.

Die Verfahren der Gruppenbildung der Merkmalsträger haben dabei eine enge Verwandtschaft zu den Clusterverfahren und sind zum Teil als Abwandlung bekannter Clusterverfahren entstanden (z. B. Ward-Algorithmus; siehe Domingo-Ferrer und Mateo-Sanz 2002). Auch die Clusterverfahren gruppieren die Merkmalsträger nach Kriterien der größten Ähnlichkeit. Bei beiden Verfahrensgruppen wird für die Bildung der Cluster/Merkmalsträgergruppen die Ähnlichkeit/Homogenität an der internen Varianz innerhalb der Gruppen gemessen. Der Unterschied besteht im Zweck der Gruppierung, die auch entscheidende Unterschiede bei der Verfahrensdurchführung bewirken. Ziel der Clusterverfahren ist eine Klassierung des Datenbestandes. Damit soll eine begrenzte Anzahl von Gruppen gebildet werden, die in ihren Eigenschaften möglichst homogen ist und so inhaltlich interpretiert werden kann. Bei Mikroaggregationsverfahren besteht das Ziel in der Erzeugung einer

7 Es gibt auch Fälle, in denen diese Minimalgröße von Mikroaggregationsgruppen ($k=3$) nicht ausreichend ist. Sind z. B. in einer Statistik die erhobenen Einheiten Betriebe, müssen auch mehrere zu einem Unternehmen zusammengehörende Betriebe mit ihren Angaben als verfügbares Angriffswissen unterstellt werden. In speziellen Branchen oder Regionen kann es dann durchaus vorkommen, dass diese Betriebe in der gleichen Mikroaggregationsgruppe zusammengefasst werden. Ist dem Datenangreifer dann eine nach Branchen/Regionen geblockte Anwendung des Verfahrens bekannt, kann er die Angabe des einen unbekannten Unternehmens in der Mikroaggregationsgruppe zurückrechnen. Hier zeigt sich wieder, dass in speziellen Konstellationen der Umfang der bereitgestellten Information über die Anwendung der Anonymisierungsverfahren das Sicherheitsrisiko erhöhen kann.

Mehrdeutigkeit der einzelnen Einheiten des Datenbestandes, die für die Einheiten Anonymität bewirkt, weil die Einheiten nicht mehr eindeutig zuzuordnen sind. Außerdem bewirkt die Veränderung der Merkmalswerte, dass die Werte in der Regel nicht mehr mit dem Original identisch sind und somit der Nutzen für Datenangreifer eingeschränkt ist. Anonymität sichernde Mehrdeutigkeit kann aber bereits mit 3 Einheiten pro Mikroaggregationsgruppe erreicht werden. Um die statistischen Eigenschaften des Datenbestandes möglichst gut zu erhalten, sollen gleichzeitig möglichst geringe Veränderungen am Datenbestand vorgenommen werden. Damit besteht eine optimale Lösung für Mikroaggregationen aus möglichst vielen Mikroaggregationsgruppen mit jeweils mindestens 3 Einheiten. Die Lösungen von Clusterverfahren sind möglichst wenige Gruppen, die im Extremfall auch weniger als 3 Einheiten enthalten könnten, wenn in den Daten „exotische" Merkmalsträger (Ausreißer) enthalten sind.

Mikroaggregation stellt sich allgemein als Lösung folgender Aufgabe dar (siehe Oganian und Domingo-Ferrer 2001):

Es sei $X^0=\{x_1^0,x_2^0,x_3^0,...,x_n^0\}^T$ die Matrix der Originaldaten, die aus n Zeilen (Anzahl der Objekte) und m Spalten (Anzahl der Merkmale) besteht. Gegeben sei weiterhin ein Sicherheitsparameter k, der die Anzahl der in einer Mikroaggregationsgruppe enthaltenen Objekte angibt (üblicherweise ist $k\geq 3$). Eine k-Teilung $T=\{G_1,G_2,G_3,...,G_{g(T)}\}$ von X ist eine Aufteilung der einzelnen Objekte x_i^o der Originaldaten auf die Gruppen G_i bei der die Größe der einzelnen Gruppen $|G_i|$ (mit $1\leq i\leq g(T)$) mindestens k beträgt. Es sei $\boldsymbol{T}_k$ die Menge aller möglichen k-Teilungen der Originaldaten X^0. Die optimale Mikroaggregation besteht im Finden derjenigen k-Teilung der Originaldaten, die die Summe der quadrierten euklidischen Abstände der x_i^0 vom Zentroiden der jeweiligen Mikroaggregationsgruppe minimiert.

Als Formel lautet das Problem somit:

$$\min_{T\in T_k}\left(\sum_{i=1}^{g(T)}\sum_{x_j^o\in G_i}\left\| x_j^o-\overline{x_{G_i}}\right\|^2\right)$$

mit dem Zentroiden $\overline{x_{G_i}}$: **(2.2.6 – 1)**

$$\overline{x_{G_i}}=\frac{\sum_{x_j^o\in G_i}x_j^o}{|G_i|}$$

$|G_i|$ – Größe der Gruppe i (Anzahl Objekte in der Gruppe)

$\|\cdot\|$ – Euklidische Norm

Oganian und Domingo-Ferrer beweisen, dass das Problem so komplex ist, dass sich der Aufwand zum Finden der exakten Lösung polynomial zum Produkt aus der Anzahl der Objekte und der Anzahl der möglichen Gruppen verhält. Deshalb lässt sich das Problem für reale Datenbestände nur über heuristische Verfahren lösen, die zu Näherungslösungen führen. Aus diesem Grunde haben sich im Laufe der Zeit verschiedene Mikroaggregationsverfahren entwickelt.

Ein Ansatz zur Lösung der numerischen Probleme ist die Bestimmung der Ähnlichkeit mit Hilfe einer Transformation in ein eindimensionales Maß und anschließender Gruppenbildung anhand dieses Maßes. Diese Varianten der Mikroaggregation werden deshalb auch als eindimensionale Mikroaggregation bezeichnet. Für diesen Ansatz besteht dann die Möglichkeit sehr schnell eine optimale Lösung zu finden (siehe Hansen und Mukherjee 2003 oder den Anhang 1 in diesem Band, S. 131 ff.).

Domingo-Ferrer und Mateo-Sanz ist noch ein weiterer grundlegender Beweis für die Klasse der Mikroaggregationsverfahren gelungen (in Domingo-Ferrer und Mateo-Sanz 2002). Hier zeigen sie, dass für eine optimale Lösung die Größe der einzelnen Gruppen G_i (mit $1 \le i \le g(T)$) zwischen $k \le |G_i| \le 2k-1$ liegt. Hat man eine Näherungslösung erreicht, bei der für mindesten eine Gruppe $2k \le |G_i|$ gilt, so existiert eine Teilungsmöglichkeit dieser Gruppe in 2 neue Gruppen mit mindestens jeweils k Elementen, die das Optimierungsziel nicht verschlechtert. Damit wurde eine enorme Reduzierung der zu untersuchenden Gruppierungsmöglichkeiten bei der Suche nach der Näherungslösung möglich.

Anders als im zitierten Beweis lässt sich aber auch zeigen, dass für jede beliebige Teilung dieser Gruppe mit mindestens $2k$ Elementen gilt, dass sie das Optimierungsziel nicht verschlechtert.

Bei der Bearbeitung der Mikroaggregation existiere eine Gruppe i, für deren Größe $|G_i|$ gilt, dass $2k \le |G_i|$ ist. Dann lässt sich eine Größe für die erste Teilgruppe G_{iA} so wählen, dass sowohl $k \le |G_{iA}|$ als auch $k \le |G_{iB}| = (|G_i| - |G_{iA}|)$ gilt. Werden anschließend beliebige Elemente der Gruppe i der Teilgruppe A (G_{iA}) und die anderen der Teilgruppe B (G_{iB}) zugeordnet, so bleiben alle anderen bereits gebildeten Gruppen unbeeinflusst. Damit wird in 2.2.6 – 1 die Summe aller quadrierten Abstände vom Zentroiden der jeweiligen Gruppe nur für die Gruppe i verändert, die jetzt aus den beiden Teilgruppen A und B besteht. Für den Beweis, dass eine beliebige Gruppenteilung das Ergebnis der Mikroaggregation nicht verschlechtert, ist zu zeigen, dass:

$$\sum_{x_j^o \in G_i} \left\| x_j^o - \overline{x_{G_i}} \right\|^2 \ge \sum_{x_j^o \in G_{iA}} \left\| x_j^o - \overline{x_{G_{iA}}} \right\|^2 + \sum_{x_j^o \in G_{iB}} \left\| x_j^o - \overline{x_{G_{iB}}} \right\|^2 \qquad \textbf{(2.2.6 – 2)}$$

bzw.

$$0 \le \sum_{x_j^o \in G_i} \left\| x_j^o - \overline{x_{G_i}} \right\|^2 - \sum_{x_j^o \in G_{iA}} \left\| x_j^o - \overline{x_{G_{iA}}} \right\|^2 - \sum_{x_j^o \in G_{iB}} \left\| x_j^o - \overline{x_{G_{iB}}} \right\|^2$$

gilt.

Für das Quadrat der euklidischen Norm gilt:

$$\sum_{x_j^o \in G_i} \left\| x_j^o - \overline{x_{G_i}} \right\|^2 = \sum_{x_j^o \in G_i} \sum_{l=1}^{m} \left(x_{j,l}^o - \overline{x_{G_i,l}} \right)^2 \qquad \textbf{(2.2.6 – 3)}$$

Somit gelte: **(2.2.6 – 4)**

$$0 \le \sum_{x_j^o \in G_i} \sum_{l=1}^{m} \left(x_{j,l}^o - \overline{x_{G_i,l}} \right)^2 - \left(\sum_{x_j^o \in G_{iA}} \sum_{l=1}^{m} \left(x_{j,l}^o - \overline{x_{G_{iA},l}} \right)^2 + \sum_{x_j^o \in G_{iB}} \sum_{l=1}^{m} \left(x_{j,l}^o - \overline{x_{G_{iB},l}} \right)^2 \right)$$

$$= \sum_{l=1}^{m} \left(\begin{array}{l} \sum\limits_{x_j^o \in G_i} \left(\left(x_{j,l}^o\right)^2 - 2x_{j,l}^o \overline{x_{G_i,l}} + \overline{x_{G_i,l}}^2 \right) - \sum\limits_{x_j^o \in G_{iA}} \left(\left(x_{j,l}^o\right)^2 - 2x_{j,l}^o \overline{x_{G_{iA},l}} + \overline{x_{G_{iA},l}}^2 \right) \\ - \sum\limits_{x_j^o \in G_{iB}} \left(\left(x_{j,l}^o\right)^2 - 2x_{j,l}^o \overline{x_{G_{iB},l}} + \overline{x_{G_{iB},l}}^2 \right) \end{array} \right)$$

$$= \sum_{l=1}^{m} \left(\begin{array}{l} \sum\limits_{x_j^o \in G_i} \left(\overline{x_{G_i,l}}^2 \right) - \sum\limits_{x_j^o \in G_{iA}} \left(\overline{x_{G_{iA},l}}^2 \right) - \sum\limits_{x_j^o \in G_{iB}} \left(\overline{x_{G_{iB},l}}^2 \right) \\ - \sum\limits_{x_j^o \in G_i} \left(2x_{j,l}^o \overline{x_{G_i,l}} \right) + \sum\limits_{x_j^o \in G_{iA}} \left(2x_{j,l}^o \overline{x_{G_{iA},l}} \right) + \sum\limits_{x_j^o \in G_{iB}} \left(2x_{j,l}^o \overline{x_{G_{iB},l}} \right) \end{array} \right)$$

$$= \sum_{l=1}^{m} \left(\begin{array}{l} |G_i| \overline{x_{G_i,l}}^2 - |G_{iA}| \overline{x_{G_{iA},l}}^2 - |G_{iB}| \overline{x_{G_{iB},l}}^2 \\ + \left(\overline{x_{G_{iA},l}} - \overline{x_{G_i,l}} \right) \sum\limits_{x_j^o \in G_{iA}} \left(2x_{j,l}^o \right) + \left(\overline{x_{G_{iB},l}} - \overline{x_{G_i,l}} \right) \sum\limits_{x_j^o \in G_{iB}} \left(2x_{j,l}^o \right) \end{array} \right)$$

$$= \sum_{l=1}^{m} \left(\begin{array}{l} |G_i| \overline{x_{G_i,l}}^2 - |G_{iA}| \overline{x_{G_{iA},l}}^2 - |G_{iB}| \overline{x_{G_{iB},l}}^2 \\ + 2|G_{iA}| \left(\overline{x_{G_{iA},l}}^2 - \overline{x_{G_i,l}}\, \overline{x_{G_{iA},l}} \right) + 2|G_{iB}| \left(\overline{x_{G_{iB},l}}^2 - \overline{x_{G_i,l}}\, \overline{x_{G_{iB},l}} \right) \end{array} \right)$$

$$= \sum_{l=1}^{m} \left(\begin{array}{l} \left(|G_{iA}| + |G_{iB}| \right) \overline{x_{G_i,l}}^2 + |G_{iA}| \overline{x_{G_{iA},l}}^2 + |G_{iB}| \overline{x_{G_{iB},l}}^2 \\ - 2|G_{iA}| \left(\overline{x_{G_i,l}}\, \overline{x_{G_{iA},l}} \right) - 2|G_{iB}| \left(\overline{x_{G_i,l}}\, \overline{x_{G_{iB},l}} \right) \end{array} \right)$$

$$= \sum_{l=1}^{m} \left(|G_{iA}| \left(\overline{x_{G_i,l}}^2 - 2\left(\overline{x_{G_i,l}}\, \overline{x_{G_{iA},l}} \right) + \overline{x_{G_{iA},l}}^2 \right) + |G_{iB}| \left(\overline{x_{G_i,l}}^2 - 2\left(\overline{x_{G_i,l}}\, \overline{x_{G_{iB},l}} \right) + \overline{x_{G_{iB},l}}^2 \right) \right)$$

$$= \sum_{l=1}^{m} \left(|G_{iA}| \left(\overline{x_{G_i,l}} - \overline{x_{G_{iA},l}} \right)^2 + |G_{iB}| \left(\overline{x_{G_i,l}} - \overline{x_{G_{iB},l}} \right)^2 \right) \geq 0 \qquad ; \, qed \qquad \textbf{(2.2.6 – 5)}$$

Damit ist jede beliebige Teilung von großen Gruppen aus mindestens 2k Elementen ebenfalls eine effektive Lösung von (2.2.1 – 1), die nicht schlechter als die Ausgangssituation ist.

Generell lassen sich Mikroaggregationen als eine Transformation des Datenbestandes in der Form

$$X^a = AX^o$$

mit **(2.2.6 – 6)**

$$A = \left(\begin{array}{ccc|c} A_1 & 0 & 0 & 0 \\ 0 & A_2 & 0 & 0 \\ 0 & 0 & A_3 & 0 \\ \hline 0 & 0 & 0 & A_g \end{array} \right)$$

mit den g Blockmatrizen A_i und den Nullmatrizen 0 zusammenfassen.

$$A_i = \begin{pmatrix} 1/|G_i| & 1/|G_i| & | & 1/|G_i| \\ 1/|G_i| & 1/|G_i| & | & 1/|G_i| \\ \hline 1/|G_i| & 1/|G_i| & | & 1/|G_i| \end{pmatrix} \qquad |G_i| = k, k+1, ..., 2k-1$$

Welche Merkmalsträger durch diese Transformationsvorschrift jeweils in eine Gruppe zusammengefasst werden, wird dabei durch die Sortierung der Merkmalsträger bestimmt. Diese ist so zu wählen, dass die in eine Gruppe zusammenzufassenden Merkmalsträger in der Sortierreihenfolge jeweils zusammen stehen. Für die Bestimmung der Sortierreihenfolge gibt es verschiedene Varianten, die die jeweilige Art der Mikroaggregation bestimmen.

Die im Laufe der Zeit entwickelten verschiedenen Mikroaggregationsverfahren sollen im Folgenden näher erläutert werden.

1) Eindimensionale Mikroaggregation

Bei eindimensionaler Mikroaggregation findet die Bestimmung der Ähnlichkeit durch eine Transformation in ein eindimensionales Maß statt. Nach diesem Maß werden die Merkmalsträger sortiert und dann k aufeinanderfolgende gruppiert. Dabei gibt es folgende Varianten (siehe auch Mateo-Sanz und Domingo-Ferrer 1998a):

– Mikroaggregation nach einer Variable

  Es wird eine dominierende Variable herausgesucht und der Datenbestand danach sortiert. Anschließend werden absteigend immer drei benachbarte Merkmalsträger in einer Gruppe zusammengefasst und alle ihre stetigen Merkmalswerte durch den Durchschnitt der Werte ersetzt. Die dominierende Variable sollte dabei mit möglichst vielen weiteren Merkmalen stark korreliert sein.

– Mikroaggregation nach mehreren Variablen

  Die Sortierung erfolgt an Hand von Hilfsvariablen. Die Hilfsvariablen sind dabei z. B. die Hauptkomponente (als eine durch Transformation gebildete neue Variable mit möglichst hoher Korrelation zu den anderen Variablen) oder die Z-Scores (als die Summe der standardisierten Originalvariablen).

– Unabhängige eindimensionale Mikroaggregation

  Die einzelnen stetigen Merkmale werden unabhängig voneinander bearbeitet. Sie werden sortiert und die k ($k = 3,4,5$) benachbarten Werte durch ihren Durchschnitt ersetzt. Danach werden sie an die Originalposition zurücksortiert und das nächste Merkmal bearbeitet. Im Gegensatz zu den beiden anderen Verfahrensvarianten wird hier die Matrix A (siehe 2.2.6 – 6) nicht für die gesamte Matrix X verwendet, sondern für jede Datenspalte x_j neu bestimmt.

  Die unabhängige eindimensionale Mikroaggregation hat gegenüber den anderen Mikroaggregationsverfahren ein zusätzliches Sicherheitsrisiko, weil für jeden Merkmalswert eine Ober- und Untergrenze der originalen Werte durch die Durchschnitte der benachbarten Gruppen verfügbar ist. Innerhalb dieser beiden Grenzen befinden sich die originalen Werte mit absoluter Sicherheit. Diese Werte können jedoch den Be-

reich des originalen Wertes sehr stark einschränken. Dieses Risiko besteht sonst nur bei der Mikroaggregation nach einer Variable für die ausgewählte dominierende Variable, wenn dem Datenangreifer diese Variable bekannt ist.

Neben der Verwendung einer festen Gruppengröße besteht bei der eindimensionalen Mikroaggregationen auch die Möglichkeit, einfach eine exakte Lösung für variable Gruppengrößen zu finden (siehe Domingo-Ferrer und Mateo-Sanz 2002 sowie Hansen und Mukherjee 2003).

2) Mehrdimensionale Mikroaggregation

Bei den mehrdimensionalen Mikroaggregationen findet für die Bestimmung der Ähnlichkeit eine mehrdimensionale Bestimmung des Abstandes zwischen den Merkmalsträgern statt. Dabei gibt es folgende Varianten:

– Mehrdimensionale Mikroaggregation mit fester Gruppengröße

  Es werden die beiden Merkmalsträger herausgesucht, die den größten Abstand untereinander haben (euklidischer Abstand der normierten Werte). Danach werden diesen beiden jeweils die zwei dichtesten Merkmalsträger dazugruppiert. Die verbleibenden, noch nicht gruppierten Merkmalsträger werden wieder analog behandelt. (Für eine detailliertere Beschreibung vgl. Mateo-Sanz und Domingo-Ferrer 1998b.)

– Mehrdimensionale Mikroaggregation mit variabler Gruppengröße

  Ziel der Mikroaggregationen mit variabler Gruppengröße ist eine noch stärkere datenorientierte Gruppenbildung durch die Möglichkeit von Gruppen größer als k (z. B. $k=3$). Dabei werden wieder Gruppen bis $2k-1$ (bis 5) angestrebt. Gruppen mit einer Größe vom doppelten der minimalen Gruppengröße lassen sich ja ohne Qualitätsverlust weiter teilen (siehe oben).

  Bei diesen Verfahren werden die einzelnen Objekte nach dem Kriterium der größten Ähnlichkeit ggf. auch Gruppen zugeordnet, die bereits 3 oder mehr Elemente haben. Erreichen Gruppen eine Größe von $2k$ oder mehr Elementen, werden sie durch hierarchische Anwendung des Verfahrens wieder geteilt (Mateo-Sanz und Domingo-Ferrer 1998b).

– Stochastische Mikroaggregationen

  Pohlmeier schlägt vor, bei der Bildung von Gruppen auf die Ähnlichkeit der Merkmalsträger als Kriterium zu verzichten. Dafür sollen die Gruppen durch eine stochastische Auswahl getroffen werden. Erfolgt die Zufallsauwahl ohne Zurücklegen, handelt es sich um die einfache stochastische Mikroaggregation. Erfolgt die Auswahl mit Zurücklegen, d. h. die Merkmalsträger können mehrfach Gruppen zugeordnet werden, handelt es sich um die Bootstrap-Mikroaggregation (siehe Lechner und Pohlmeier 2003).

3 Erweiterungen der Mikroaggregationsverfahren

Die Mikroaggregationsverfahren besitzen einige Verfahrensprobleme, die durch folgende Erweiterungen gelöst wurden:

3.1 Nummerische Probleme der mehrdimensionalen Mikroaggregation

Für die Anwendung von Mikroaggregationsverfahren kann auf die von Eurostat im Rahmen des CASC-Projektes[8] bereitgestellte Software μ-Argus zurückgegriffen werden. Die dort enthaltenen Programme wurden von Domingo-Ferrer und Mateo-Sanz bereitgestellt. Sie entsprechen damit den von ihnen veröffentlichten Verfahrensbeschreibungen in Mateo-Sanz und Domingo-Ferrer (1998b). Bei der Durchführung von Testrechnungen mit den Daten der Kostenstrukturerhebung und der Umsatzsteuerstatistik im Rahmen des Projektes „Faktische Anonymisierung von wirtschaftsstatischen Einzeldaten" wurde jedoch festgestellt, dass die Programme sehr lange liefen, ohne eine Lösung zu erreichen. Kleinere Datenbestände konnten problemlos bearbeitet werden.

Bei der mehrdimensionalen Mikroaggregation wurde folgende Vorgehensweise von Domingo-Ferrer und Mateo-Sanz umgesetzt.

- Bestimmung der beiden Merkmalsträger, die den größten Abstand untereinander besitzen.
- Bestimmung von je zwei ähnlichsten Merkmalsträgern zu den beiden gefundenen Merkmalsträgern und Bildung von zwei Satzgruppen.
- Herausnehmen der beiden Gruppen aus dem Datenbestand und Bearbeitung der verbleibenden Merkmalsträger mit dem gleichen Verfahren.
- Für die so gefundenen Satzgruppen wird an Stelle der originalen stetigen Werte jeweils der Mittelwert der Gruppe eingesetzt.

Damit ergibt sich für die Bestimmung der ersten beiden Gruppen folgende Anzahl von Satzvergleichen im Datenbestand (n-Anzahl der Merkmalsträger).

Im 1. Schritt: $n_1 (n_1\text{-}1)$ Satzvergleiche
(jeder mit jedem anderen Merkmalsträger)

Im 2. Schritt: $2 (n_1\text{-}2)$ Satzvergleiche
(die beiden gefundenen Merkmalsträger werden jeweils mit allen übrigen verglichen)

Für beide Schritte: $(n_1^2+n_1\text{-}4)$ Satzvergleiche.

8 Das Projekt Computational Aspects of Statistical Confidentilaty (CASC) befasste sich im Schwerpunkt mit der Analyse vorhandener Methoden und der Bereitstellung von Software für die Lösung von Fragen der Tabellen- und Mikrodatengeheimhaltung. Im Ergebnis wurden die zwei Softwarepakete τ-Argus und μ-Argus entwickelt, in denen die Programme zur Tabellenanonymisierung (τ-Argus) und zur Mikrodatenanonymisierung (μ-Argus) zusammengefasst sind. Nähere Informationen zum CASC-Projekt sind im Internet unter www.neon.vb.cbs.nl/casc/ erhältlich. Dort können die Programmpakete auch kostenlos heruntergeladen werden.

Nach dem Herausnehmen der gefunden Sätze ist der Algorithmus für den Rest des Datenbestandes zu wiederholen ($n_2=n_1-6$).

Für die Reihe mit n = 6,12,18, . . . (n - Anzahl der Sätze der Mikrodatendatei) usw. ergäbe sich deshalb die Funktion

$$f(n) = \frac{1}{18}n^3 + \frac{7}{12}n^2 + \frac{5}{6}n \qquad ;n = 6, 12, 18, ... \qquad \textbf{(3.1 – 1)}$$

die die Anzahl der durchzuführenden Satzvergleiche bei der Mikroaggregation in Abhängigkeit von der Größe des Datenbestandes bestimmt.[9] Da die einzelnen Satzvergleiche einen zeitlich gleichen Aufwand benötigen, bewirkt dieser Zusammenhang, dass sich der Rechenaufwand kubisch zur Größe der Mikrodatendatei verhält.

Sätze der Datendatei	Erforderliche Satzvergleiche
6	38
60	14 150
120	104 500
300	1 552 750
600	12 210 500
900	40 973 250
1 800	325 891 500
3 000	1 505 252 500
6 000	12 021 005 000
60 000	12 002 100 050 000

Damit entstehen sehr schnell Rechenzeiten, die es unmöglich machen, das Verfahren auf größere Datenbestände geschlossen anzuwenden. Ein Ausweg bildet zwar das Zerlegen des Datenbestandes in mehrere Teile, das erfordert aber eine vorherige intensive inhaltliche Untersuchung des Datenbestandes.

Ursache für die langen Rechenzeiten im Verfahrens von Domingo-Ferrer ist die Auswahl der Merkmalsträger vom äußeren Rand des Datenbestandes. Diese Auswahl ist gewollt, weil hier die Entscheidung zum Gruppieren mit den beiden ähnlichsten Sätzen am einfachsten zu treffen ist. Für Sätze innerhalb des Gesamtbestandes ist es schwerer, die optimale Gruppe zu bilden, weil einerseits die Ähnlichkeit zwar für die Gruppierung herangezogen werden kann, es aber nicht passieren darf, dass ungruppierte einzelne Sätze zwischen den gebildeten Gruppen übrig bleiben. Soll eine Gruppierung nicht später noch ein-

9 Mikrodatendateien mit nicht durch 6 teilbarer Anzahl an Merkmalsträgern würden in der Anzahl der benötigten Satzvergleiche jeweils zwischen den benachbarten durch 6 teilbaren Zahlen liegen. Sie hätten die Besonderheit, dass die letzten beiden Gruppen aus mehr als 3 Sätzen bestehen. Die Anzahl der Gruppierungsschritte ist identisch mit der Anzahl der Gruppierungsschritte für die nächst kleinere durch 6 teilbare Zahl. Die Anzahl der Satzvergleiche ist jedoch etwas größer, weil bei jedem Schritt immer 1 – 5 Sätze mehr für den Vergleich vorhanden sind.

mal revidiert werden, muss sie also so getroffen werden, dass es für diese Sätze einerseits keine bessere Lösung geben kann und andererseits die verbleibenden Sätze nicht als einzeln verstreute „Reste“ zwischen den bereits gebildeten Gruppen übrig bleiben. Deshalb sollen die Gruppierungen immer am äußersten Rand des verbleibenden Datenbestandes vorgenommen werden. Bei einer so gebildeten Gruppe kann es nicht vorkommen, dass sie später zwischen noch nicht mikroaggregierten Sätzen liegt. Damit wird weitestgehend verhindert, dass die Gruppierung von restlichen Sätzen des Datenbestandes zu besseren Lösungen führt, wenn diese Gruppe noch einmal mit einbezogen und die Aufteilung neu entschieden wird. Die Bestimmung der Sätze erfolgte über die Suche der beiden am weitesten auseinanderliegenden Merkmalsträger. Da dieses Verfahren zu sehr aufwändigen Vergleichsoperationen führt, war eine Alternative erforderlich, um mehrdimensionale Mikroaggregation auch für größere Datenbestände nutzen zu können. Hier wurde der Abstand zum Zentroiden der Menge (Datenpunkt gebildet aus den Durchschnittswerten der jeweiligen Variablen) als Maß für die Bestimmung des am meisten außerhalb liegenden Merkmalsträgers verwendet. Für jeden Gruppierungsschritt ist der Zentroid des Datenbestandes eine feste Größe. Damit existiert für jeden Merkmalsträger auch nur ein Abstand zum Zentroiden, so dass mit n Vergleichen die Bestimmung des am meisten außerhalb liegenden Merkmalsträgers möglich ist. Zu diesem Merkmalsträger werden die zwei ähnlichsten gesucht, die dann wieder eine Mikroaggregationsgruppe bilden (n-1 Vergleiche). Danach wird der Rest des Datenbestandes analog bearbeitet. Von Vorteil ist dabei, das der Zentroid der verbleibenden Menge sich leicht über den Zentroiden der Gesamtmenge und die Merkmalswerte der drei im vorigen Schritt bestimmten Merkmalsträger bestimmen lässt. Es ist damit kein erneuter Durchlauf durch den Datenbestand zur Zentroidenberechnung Z^n erforderlich.

$$Z_j^{n-3} = \frac{nZ_j^n - \sum_{i=n-2}^{n} x_{ij}}{n-3}$$

mit:

Z^n – Zentroid bei n Merkmalsträgern mit Z_j^n als Mittelwert für Merkmal j bei n Merkmalsträgern

x_{ij} – Merkmalswert des Merkmals j beim Merkmalsträger i

Die Merkmalsträger i=n-2,n-1,n sind die im Gruppierungsschritt mikroaggregierten Merkmalsträger (ggf. Umsortierung erforderlich).

Für einen Gesamtlauf des Verfahrens ergibt sich deshalb für die Anzahl der Satzvergleiche:

– n Vergleiche mit dem Zentroiden zu Bestimmung des äußersten Merkmalsträgers.

– n-1 Vergleiche des ermittelten äußersten Merkmalsträgers mit den übrigen Merkmalsträgern zur Bestimmung der zwei ähnlichsten Merkmalsträger.

Danach sind die Vergleiche für die verbleibenden *n-3* Merkmalsträger zu wiederholen. In jedem Gruppierungsschritt sind somit *2n-1* Vergleichsoperationen nötig. Für die Reihe mit 3, 6, 9, 12 usw. ergäbe sich deshalb die Funktion:

$$f(n) = \frac{1}{3}n^2 + \frac{2}{3}n - 5 \qquad ;n = 3,6,9,12,... \qquad \textbf{(3.1 – 2)}$$

Sie bestimmt die Anzahl der durchzuführenden Satzvergleiche bei dieser Vorgehensweise der Mikroaggregation in Abhängigkeit von der Anzahl der Merkmalsträger. Da die einzelnen Satzvergleiche wieder einen zeitlich gleichen Aufwand benötigen (die Anzahl der Merkmale je Satz ist ja konstant) bewirkt dieser Zusammenhang, dass sich der Rechenaufwand quadratisch zur Größe der Mikrodatendatei verhält.

Sätze der Datendatei	Satzvergleiche
6	9
60	1 199
120	4 801
300	30 003
600	120 005
900	270 007
1 800	1 080 009
3 000	3 000 011
6 000	12 000 013
60 000	1 200 000 015

Gegenüber dem ursprünglichen Algorithmus der mehrdimensionalen Mikroaggregation hat sich die Rechenzeit bei 6 000 Sätzen der Mikrodatendatei auf 1/1 000 und bei 60 000 Sätzen der Mikrodatendatei auf 1/10 000 reduziert. Mit einem solchen Rechenzeitverlauf lässt sich das Verfahren auch bei weitaus größeren Datenbeständen noch gut anwenden.

3.2 Sicherheitsprobleme der unabhängigen eindimensionalen Mikroaggregation

Bei der eindimensionalen Mikroaggregation tritt bei unabhängiger Anwendung für jede Merkmalsspalte, das Problem auf, dass nach einem Sortieren der Datei die benachbarten Werte in den anonymisierten Merkmalsspalten eine sichere obere und untere Schranke für die originalen Werte darstellen.

Beweis:

Zur eindimensionalen Mikroaggregation werden die Merkmalsträger nach ihrer Größe absteigend sortiert. Damit gilt für die originalen Merkmalswerte

$$x^o_i \geq x^o_{i+1} \geq x^o_{i+2} \geq x^o_{i+3} \geq x^o_{i+4} \geq x^o_{i+5} \geq x^o_{i+6} \geq x^o_{i+7} \geq x^o_{i+8} \geq x^o_{i+9} \geq .... \qquad \textbf{(3.2 – 1)}$$

Es werden anschließend immer drei benachbarte Werte durch ihren Durchschnitt ersetzt. Damit gilt für die anonymen Werte:

$$x^{a}_{i,i+1,i+2} = \frac{x^{o}_{i} + x^{o}_{i+1} + x^{o}_{i+2}}{3}$$

$$\textit{und somit} \qquad \textbf{(3.2 – 2)}$$

$$x^{o}_{i} \geq x^{a}_{i,i+1,i+2} \geq x^{o}_{i+2}$$

Aus den beiden obigen Ungleichungen folgt deshalb:

$$x^{a}_{i,i+1,i+2} \geq x^{o}_{i+3} \geq x^{o}_{i+4} \geq x^{o}_{i+5} \geq x^{a}_{i+6,i+7,i+8} \qquad \textbf{(3.2 – 3)}$$

Die veröffentlichten anonymen Merkmalswerte der benachbarten größeren und der benachbarten kleineren Gruppe stellen bei der eindimensionalen Mikroaggregation sichere obere und untere Schranken für die originalen Merkmalswerte dar. Diese können bei sehr dicht belegten Merkmalswerten ein sehr enges Intervall um den originalen Merkmalswert bilden.

Dieser offensichtliche Nachteil der eindimensionalen unabhängigen Mikroaggregation hat aber auch den Vorteil, dass die Änderungen an den Merkmalswerten sehr klein sind. Dieses Verfahren zeichnete sich bei Testrechnungen durch einen sehr geringen Einfluss auf die Einzeldaten aus (geringe Veränderung der Werte). Obwohl die Art der Mikroaggregation zu systematischen Fehlern in ökonometrischen Modellen führt (siehe Lechner und Pohlmeier 2003), sind auf Grund der geringen Datenveränderung die Ergebnisse äußerst stabil (Ronning et al. 2005).

Deshalb stellt sich die Aufgabe, die Eigenschaften der geringen Datenveränderung (durch Mikroaggregation von nach dem Sortieren benachbarter Werte) und des sicheren Bestimmungsintervalls für die Einzeldaten voneinander zu trennen, um so ein Verfahren zu entwickeln, welches die gute Qualität ohne das Sicherheitsrisiko besitzt.

Das Problem kann durch eine bedarfsabhängige Vergrößerung der Datengruppen gelöst werden. Sollten die Ober- und Unterschranken für die Einzelwerte eine zu genaue Vorhersage der Werte ermöglichen, werden die Merkmalsgruppen vergrößert. Da der Effekt nur dann auftritt, wenn die benachbarten Gruppen sehr ähnlich sind, sollte sich der Einfluss auf die Originalwerte und somit den Qualitätsverlust ebenfalls in Grenzen halten. Die variable Gruppengröße wird jetzt aber nicht mehr primär durch die Minimierung der gruppeninternen Varianz, sondern durch die Gewährung eines ausreichenden Schutzes der Daten bestimmt.

Für die Entwicklung der Entscheidungsregel bei der eindimensionalen Mikroaggregation ist aber vorher zu bestimmen, welche Aussagen über die Einzelwerte getroffen werden können. Neben den beiden Schranken durch die Durchschnittswerte der benachbarten Gruppen, können auch zwei sehr dicht benachbarte Werte (unabhängig von den übrigen Werten) ein Sicherheitsrisiko darstellen.

Für das Verhältnis der anonymen Werte zweier benachbarter Mikroaggregationsgruppen (p) gilt:

$$p = \frac{x^a_{i,i+1,i+2}}{x^a_{i+3,i+4,i+5}} \geq 1$$

damit folgt aus :

$$x^a_{i,i+1,i+2} = \frac{x^o_i + x^o_{i+1} + x^o_{i+2}}{3}$$

dass

$$x^o_i = 3p x^a_{i+3,i+4,i+5} - x^o_{i+1} - x^o_{i+2} \quad ; mit \ p \geq 1$$

und aus **(3.2 – 4)**

$$x^a_{i+3,i+4,i+5} \leq x^o_{i+2} \leq x^o_{i+1} \leq x^o_i$$

folgt :

$$x^o_i \leq 3p x^a_{i+3,i+4,i+5} - 2x^a_{i+3,i+4,i+5} \quad ; mit \ p \geq 1$$

Daraus ergibt sich :

$$x^a_{i+3,i+4,i+5} \leq x^o_{i+2} \leq x^o_{i+1} \leq x^o_i \leq (3p-2)x^a_{i+3,i+4,i+5} \quad ; mit \ p \geq 1$$

Für Werte von p nahe 1 ergibt auch $(3p\text{-}2)$ einen Wert nahe 1 und somit ein sehr kleines Intervall für den originalen Wert x^o_i. Soll z. B. für einen originalen Wert ein berechenbares Intervall mindestens 5 % des originalen Wertes betragen, damit der Wert noch als nicht deanonymisiert betrachtet wird, so dürfen die beiden anonymen Werte nicht weniger als 1,67 % Unterschied haben ($1{,}05 \leq (3p - 2)$ bedeutet $p \geq 1{,}0167$). Damit sind die originalen Werte der Gruppe $x^a_{i,i+1,i+2}$ nicht mehr sicher, wenn der Durchschnitt der kleineren Gruppe mehr als 98,36 % des Durchschnitts der nächst größeren Gruppe beträgt. Da x^o_i der größte Wert der drei Originalwerte der Gruppe ist, gelten diese Schranken natürlich auch für x^o_{i+1} und $^o_{i+2}$.

Analog können auch die Einzelwerte der unteren Merkmalsträgergruppe unsicher sein.

Wenn für zwei Mikroaggregationsgruppen gilt

$$p = \frac{x^a_{i,i+1,i+2}}{x^a_{i+3,i+4,i+5}} \geq 1$$

folgt aus:

$$x^a_{i+3,i+4,i+5} = \frac{x^o_{i+3} + x^o_{i+4} + x^o_{i+5}}{3}$$

dass

$$x^o_{i+5} = \frac{3}{p} x^a_{i,i+1,i+2} - x^o_{i+3} - x^o_{i+4} \quad ; mit \ p \geq 1$$

und aus

$$x^o_{i+5} \leq x^o_{i+4} \leq x^o_{i+3} \leq x^a_{i,i+1,i+2}$$

folgt:

$$x^{o}_{i+5} \geq \frac{3}{p} x^{a}_{i,i+1,i+2} - 2\, x^{a}_{i,i+1,i+2} \quad ; mit\ p \geq 1 \qquad \textbf{(3.2 – 5)}$$

Damit ergibt sich:

$$\left(\frac{3}{p} - 2 \right) x^{a}_{i,i+1,i+2} \leq x^{o}_{i+5} \leq x^{o}_{i+4} \leq x^{o}_{i+3} \leq x^{a}_{i,i+1,i+2} \quad ; mit\ p \geq 1$$

Für Werte von p nahe 1 ergibt auch *(3/p-2)* einen Wert nahe 1 und somit ein sehr kleines Intervall für den originalen Wert x^{o}_{i+5}. Soll z. B. für einen originalen Wert ein berechenbares Intervall mindestens 5 % betragen, damit der Wert noch als nicht deanonymisiert betrachtet wird, dürfen die beiden anonymen Werte nicht weniger als 1,7 % Unterschied haben (*0,95≥(3/p - 2)* bedeutet p≥1,0169). Damit sind die originalen Werte der Gruppe $x^{a}_{i+3,i+4,i+5}$ nicht mehr sicher, wenn der Durchschnitt der kleineren Gruppe mehr als 98,3 % des Durchschnitts der nächst größeren Gruppe beträgt. Da x^{o}_{i+5} der kleinste Wert der drei Originalwerte der Gruppe ist, gelten diese Schranken natürlich auch für x^{o}_{i+3} und x^{o}_{i+4}.

Beispiel:

Lfd. Nr.	Anonymisierte Werte	Schranken aus der Regel der benachbarten Gruppen (3.2 – 3)		Schranken aus beiden Regeln des Gruppenabstandes (3.2 – 3 bis 3.2 – 5)	
		oben	unten	oben	unten
1	100 000	keine	99 000	102 000	99 000
2	100 000	keine	99 000	102 000	99 000
3	100 000	keine	99 000	102 000	99 000
4	99 000	100 000	90 000	100 000	97 000
5	99 000	100 000	90 000	100 000	97 000
6	99 000	100 000	90 000	100 000	97 000
7	90 000	99 000	...	99 000	...
8	90 000	99 000	...	99 000	...
9	90 000	99 000	...	99 000	...
...	...	...	...	...	...

Sowohl die obere Schranke von 102 000 für die Einheiten 1 bis 3 als auch die untere Schranke von 97 000 für die Einheiten 4 bis 6 stellen einen erheblichen Informationsgewinn und somit eine starke Erhöhung des Sicherheitsrisikos für die Originalwerte dar. Dieses Risiko kann auch mitten im Datenbestand auftreten, da es nur durch den Abstand zwischen den veröffentlichten Durchschnittswerten bestimmt wird.

Für einen Entscheidungsalgorithmus zur Gewährleistung der Sicherheit sei S ein Sicherheitsintervall (als relativer Anteil am Originalwert), welches angibt, wie groß das durch die

sichere obere und untere Schranke bestimmbare Intervall für die originalen Einzelwerte sein darf, ohne dass die Merkmalsträger als deanonymisiert gelten. Dann bedeutet z. B. *S=0,05* – ein Einzelwert kann durch seine obere und untere Schranke nur so weit bestimmt werden, dass mindestens ein 5 % Unsicherheitsintervall verbleibt. Dann lassen sich aus 3.2 – 4 und 3.2. – 5 Mindestabstände für die benachbarten anonymisierten Werte ermitteln. Aus 3.2 – 4 folgt:

$1 \leq (3p_1 - 2)\,(1-S)$; d. h. $p_1 \geq (3-2S)/(3-3S)$.

Und aus 3.2 – 5 folgt:

$(3/p_2 - 2) \leq (1-S)$; d. h. $p_2 \geq 3/(3-S)$.

Da beide Parameter p_1 und p_2 größer 1 sind und S zwischen 0 und 1 liegt, sind diese Umstellungen der Gleichungen möglich. Außerdem ist die aus 3.2. – 5 folgende Ungleichung restriktiver, so dass für einen Gesamtbestand ein möglicher Algorithmus für die unabhängige eindimensionale Mikroaggregation mit variabler Gruppengröße (abhängig vom Sicherheitsbedarf) herleitbar ist:

1. Lege ein Sicherheitsintervall *S* (als Anteil am Originalwert) fest, welches angibt, wie genau das sichere Intervall für die originalen Einzelwerte bestimmbar sein darf, ohne dass die Merkmalsträger als deanonymisiert gelten (z. B. *S=0,05* bedeutet, der Einzelwert ist nur mit 5 % Unsicherheitsintervall eingrenzbar).

   Wähle für die folgenden Schritte einen Parameter *p* mit $p \geq 3/(3-S)$ aus.

2. Bearbeite jede Merkmalsspalte *j* der Datei nach der Schrittfolge (3. bis 9.):

3. Sortiere die Datei absteigend nach Merkmal *j*.

4. Es seien G_k die Menge der Merkmalsträger in der Mikroaggregationsgruppe *k* und $|G_k|$ die Anzahl der Merkmalsträger in der Menge G_k

5. G_1 sei $G_1 = \{ x^o_{1,}\, x^o_{2,}\, x^o_{3} \}$ und somit
   $x^a_{1,2,3,j} = (x^o_{1,j} + x^o_{2,j} + x^o_{3,j})/3$.

6. Setze $k=2$ und $G_2 = \{ x^o_{4,}\, x^o_{5,}\, x^o_{6} \}$ und $l=6$
   (l bezeichne einen Zähler der absteigend sortierten Merkmalsträger).

7. Prüfe:
$$\overline{x_{k,j}} = \frac{\sum_{x_l \in G_k} x_{l,j}}{|G_k|} \leq \frac{\overline{x_{k-1,j}}}{p}$$

Wenn die Ungleichung nicht erfüllt ist, erweitere die Gruppe G_k um das Element *l+1* und prüfe erneut (erhöhe Zähler *l* auf *l+1* und $|G_k| = |G_k|+1$).

Ist die Ungleichung erfüllt, so gehe zu Schritt 8.

8. Sind noch 3 weitere Merkmalsträger *l+1, l+2, l+3* verfügbar, bilde eine neue Gruppe G_{k+1} mit $G_{k+1} = \{ x^o_{l+1,}\, x^o_{l+2,}\, x^o_{l+3} \}$ und erhöhe die Zähler *k* auf *k+1* und *l* auf *l+3*. Wiederhole Schritt 7.

Waren keine 3 weiteren Merkmalsträger verfügbar, so erweitere die Menge G_k um die restlichen Merkmalsträger *l+1* und ggf. *l+2* und gehe zu Schritt *9*.

9. Für alle Gruppen G_k und alle Elemente x^o_{ij} in den Gruppen bilde die anonymen Werte x^a_{ij} der Merkmalsspalte *j* als:

$$x^a_{ij} = \overline{x_{G_k,j}} = \frac{\sum_{x_l \in G_k} x^o_{lj}}{|G_k|} \quad ; \forall\ x^o_{ij} \in G_k$$

10. Wenn eine weitere Merkmalsspalte vorhanden ist, gehe zu Schritt 3 für Spalte *j=j+1*. Wenn keine weitere Merkmalsspalte vorhanden ist, beende das Verfahren.

Dieser Algorithmus der Mikroaggregation wurde im Rahmen des Projektes „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ für die SAFE2A-Verfahren verwendet (siehe Zwischenbericht des Projektes in Gnoss et al. 2003, S. 89 f.). Bei diesem Verfahren wurde die Behandlung der metrischen Merkmale nach obigem Algorithmus und die Behandlung der kategorialen Merkmale nach dem SAFE-Verfahren vorgenommen (siehe Abschnitt 3.4). Bei den Kriterien, die nur die Qualität der metrischen Merkmale beurteilen, konnte das Verfahren mit der unabhängigen eindimensionalen Mikroaggregation vergleichbare Ergebnisse erzielen, während die Bestimmung von deanonymisierenden sicheren Intervallen für die Daten verhindert wurde. Trotz der Gewährleistung hinreichend großer sicherer Intervalle für die Daten hatte das Verfahren bei einzelnen Einheiten eine unzureichende Schutzwirkung, weil das Verfahren keine Mindestveränderung der Daten sichert. Die Veränderungen der kategorialen Merkmale durch den SAFE-Ansatz konnte ein Restrisiko bei Großunternehmen nicht verhindern (siehe Zwischenbericht des Projektes in Gnoss et al. 2003, S. 92). Hier wären weitere Maßnahmen für die kategorialen Merkmale erforderlich.

3.3 Mikroaggregation mit Varianzerhalt

Bei den Mikroaggregationsverfahren werden die originalen Werte der einzelnen Merkmalsträger durch den Durchschnitt der jeweiligen Gruppe von Merkmalsträgern ersetzt. Damit haben die Lösungen folgende Eigenschaften:

1. Es entsteht innerhalb der Merkmalsträgergruppen eine Mehrdeutigkeit, da alle Merkmalsträger den gleichen Durchschnittswert erhalten. Aus dieser Mehrdeutigkeit resultiert die Anonymität des mit Mikroaggregation bearbeiteten Datenbestandes, da eindeutige Zuordnungen verhindert bzw. erschwert werden. Wird die Mikroaggregation nur spaltenweise oder auf Variablenblöcke unabhängig voneinander angewendet, so besteht noch ein gewisses Re-Identifikationsrisiko, wenn ein Angriffswissen, das aus Merkmalen der verschiedenen Variablenblöcke besteht, nicht ausgeschlossen werden kann. Die Mehrdeutigkeit ist in diesem Fall für Merkmalskombinationen im Angriffswissen aus verschiedenen Variablenblöcken nicht automatisch gegeben. Zuordnungen werden aber trotzdem erschwert, weil die Werte nicht mehr den Originaldaten entsprechen, sondern nur noch eine Ähnlichkeit besitzen.

2. Die Durchschnitte und Summen in den einzelnen Merkmalsträgergruppen werden erhalten. Deshalb werden auch die Summen und Durchschnitte in der Gesamtheit und in bei der Mikroaggregation berücksichtigten Teilgruppierungen erhalten.

3. Die Varianz innerhalb der einzelnen Merkmalsträgergruppen ist Null. Daraus resultiert, dass die Varianzen des Gesamtbestandes und aller Teilgruppierungen systematisch unterschätzt werden, weil nur noch die Varianz zwischen den Merkmalsträgergruppen erhalten bleibt.

   Diese Eigenschaft von systematisch unterschätzten Varianzen wirkt sich negativ bei der Durchführung ökonometrischer Tests aus.

Die Korrelationen werden bei mit Mikroaggregationsverfahren anonymisierten Daten unter bestimmten Voraussetzungen erwartungstreu reproduziert. Dazu kann entweder die verwendete Aggregationsmatrix unabhängig von den betrachteten Merkmalen gebildet werden, z. B. bei stochastischer Mikroaggregation (siehe Pohlmeier und Lechner 2003). Es besteht aber auch die Möglichkeit, die Mikroaggregation für die einzelnen Variablen getrennt und somit unabhängig voneinander durchzuführen (siehe Abschnitt 3.2). Die Korrelationen werden bei diesen Varianten asymptotisch (für $n \to \infty$) erhalten (siehe Schmid 2007).

Systematische Verzerrungen treten auf, wenn die Mikroaggregationsmatrix direkt von einzelnen Merkmalen des Datenbestandes abhängig ist, wie z. B. bei der Variante der Mikroaggregation nach einer Variable. Dann führt der ungleichmäßige Varianzverlust in den Daten dazu, dass ökonometrische Schätzungen systematisch verzerrt sind, wenn die für die Mikroaggregation ausgewählte Variable als abhängige Variable im Modell verwendet wird. Die zur Bestimmung der Mikroaggregationsmatrix verwendete Variable hat dabei den kleinsten Varianzverlust, da der Varianzverlust dieser Variable bei der Bestimmung der Matrix minimiert wurde, während die anderen Variablen nicht berücksichtigt wurden (siehe Schmid 2007).

Um die Varianzen erhalten zu können, wären zwei Varianten möglich: die eindimensionale Varianzkorrektur nach Kim wie bei der Zufallsüberlagerung (siehe z. B. Brand 2000) oder die Bildung von Gruppenaggregaten mit Erhalt der Varianz in den Teilgruppen.

3.3.1 Eindimensionale Varianzkorrektur nach Kim

Die eindimensionale Varianzkorrektur nach Kim (siehe Kim 1986) erfolgt für die Merkmalsspalte j über:

$$x_{i,j}^{a_2} = \frac{\sigma\left(x_j^o\right)}{\sigma\left(x_j^{a_1}\right)}\left(x_{i,j}^{a_1} - \overline{x_j^{a_1}}\right) + \overline{x_j^o} \quad ; i = 1,2,3,\ldots,n$$

$x_{i,j}^{a_1}$ – durch Anonymisierungsschritt 1 (Zufallsüberlagerung oder Mikroaggregation) anonymisierte Werte für Merkmalsträger i beim Merkmal j

$x_{i,j}^{a_2}$ – anonymisierte Werte mit Varianzkorrektur (Anonymisierungsschritt 2)

$\overline{x_j^o}, \overline{x_j^{a_1}}$ – Durchschnitt der Werte für Merkmal j im Original bzw. nach Anonymisierungsschritt 1

$\sigma\left(x_j^o\right),$ – Standardabweichung der Werte für Merkmal j im Original bzw. nach

$\sigma\left(x_j^{a_1}\right)$ – Anonymisierungsschritt 1.

Diese Darstellung der Korrektur hat gegenüber dem Original bei Kim den Vorteil, dass sie die Parameter der Transformation aus den realen Ergebnissen des ersten Anonymisierungsschrittes ableitet. Es verwendet keine Anonymisierungsparameter der stochastischen Überlagerung (für die Kim ursprünglich die Transformation herleitete) und ist somit erstens unabhängig vom eigentlichen Anonymisierungsverfahren im Anonymisierungsschritt 1 und korrigiert zweitens zusätzlich eventuelle kleine stochastische Fehler in den Daten, die gerade bei kleineren oder sehr schiefen Datenbeständen dazu führen, dass im ersten Anonymisierungsschritt nicht unbedingt Mittelwert und Standardabweichung entsprechend der theoretischen Herleitung erzeugt werden können.

Die eindimensionale Varianzkorrektur nach Kim hat den Vorteil, dass sie geeignet ist, die Mittelwerte und Varianzen der anonymisierten Merkmale mit den originalen Merkmalswerten in Übereinstimmung zu bringen. Gleichzeitig werden die linearen Abhängigkeiten zwischen den Merkmalen nicht verändert. Damit bleibt die Korrelationsmatrix so erhalten, wie sie nach dem ersten Anonymisierungsschritt war. Lieferte das Anonymisierungsverfahren im ersten Schritt erwartungstreue Korrelationsmatrizen, so ist ein derart anonymisierter Datenbestand eine optimale Grundlage für Untersuchungen mit linearen Modellen.

Trotzdem hat die Varianzkorrektur nach Kim auch Nachteile. Da die Korrektur der Varianz als lineare Transformation zum Mittelwert des Gesamtbestandes erfolgt, ist die Veränderung der Daten gerade bei den am oberen und unteren Rand des Datenbestandes liegenden Werten am größten. Erfolgt die Korrektur im Ergebnis einer stochastischen Überlagerung, so gilt es die im ersten Schritt erhaltene Varianzvergrößerung zu verkleinern (d. h. $\sigma(x_j^o)/\sigma(x_j^a) < 1$). Die Werte werden in Richtung Mittelwert verschoben. Bei einer Korrektur nach einer Mikroaggregation gilt es die im ersten Schritt erhaltene Varianzverkleinerung wieder auszugleichen (d. h. $\sigma(x_j^o)/\sigma(x_j^a) > 1$). Die Werte werden vom Mittelwert weg verschoben.

Die Varianzkorrektur nach Kim bewirkt gerade bei sehr schief verteilten Daten, wie sie für wirtschaftsstatische Einzeldaten oft üblich sind, eine starke systematische Verzerrung der kleinen Einheiten. Denn bei wirtschaftsstatischen Daten befinden sich in der Regel viele kleine Einheiten am unteren Rand des Datenbestandes, während nur wenige große Einheiten am oberen Rand existieren. Gerade bei einer Korrektur von Varianzverlust (bei Mikroaggregation) besteht auch noch zusätzlich das Problem, dass die korrigierten Werte inhaltlichen Bedingungen (wie z. B. Nichtnegativität oder Abschneidegrenzen der Erhebung) widersprechen.

Die Abbildungen 6 und 7 zeigen die Auswirkung einer Varianzkorrektur am Beispiel von zufallsüberlagerten Werten und bei mikroaggregierten Werten.

Gerade bei der Auswertbarkeit von Teilgesamtheiten wirkt sich das Korrekturverfahren negativ aus. Wenn Teilgesamtheiten des Datenbestandes per Definition größenabhängig sind, bewirkt die Korrektur starke systematische Verzerrungen. Ein Beispiel wäre hier die Untersuchung von kleinen Unternehmen. Die Varianzkorrektur nach Kim wurde deshalb in Ronning (2005, S. 190 ff.) als unbrauchbar eingestuft.

Ein Ausweg könnte die Anwendung des Korrekturverfahrens getrennt auf kleinere „homogenere“ Einheiten sein. Auch dieser Ansatz wurde in Ronning (2005) getestet. Es wurde die Anwendung auf einen anonymisierten, absteigend sortierten Datenbestand blockweise auf jeweils 100 Sätze betrachtet. Die systematische Verzerrung in den Daten wird dabei stark verkleinert (siehe Abbildung 8).

Problematisch sind dabei folgende zwei Punkte: Erstens schließt auch eine solche Anwendung Vorzeichenwechsel in den Daten beim Ausgleich von Varianzverlust nicht grundsätzlich aus. Hier müssten andere Regeln zusätzlich implementiert werden. Zweitens wird durch die Korrektur in kleineren Gruppen gerade bei den kleinen Einheiten die Schutzwirkung des ersten Anonymisierungsschrittes reduziert. Es ist deshalb unbedingt notwendig, dass die Daten nach der Größe der anonymisierten Werte absteigend sortiert werden. Bei den originalen Werten kommt es nämlich gelegentlich vor, dass viele kleine Unternehmen (im Bereich der unteren Erhebungsgrenze) existieren, die die gleichen Werte besitzen. Vor allem bei Variablen wie Umsatz und/oder Beschäftigte, die für die Festlegung der Abschneidegrenzen der Erhebung verwendet wurden, kann es vorkommen, dass die Varianz im Bereich der Abschneidegrenze für einen ganzen Block aus Originaldaten fasst Null ist. Dann reproduziert die Varianzkorrektur die Originalwerte, so dass die anonymisierende Wirkung des ersten Schrittes rückgängig gemacht wird. In diesem Fall ist dann intensiv zu testen, ob die Mehrdeutigkeit in den Daten dieser Merkmale genügend Schutz auch für die anderen Merkmale liefert. Die durch Zufallsüberlagerung erzeugte Datenveränderung und somit „Unsicherheit“ bezüglich der Brauchbarkeit der Daten ist hier nicht mehr gegeben.

Abbildung 6
Kim-Korrektur bei multiplikativer Zufallsüberlagerung

Die Abbildung 6 enthält die aufsteigend sortierten Originalwerte (log_org). Diese wurden im ersten Anonymisierungsschritt einer multiplikativen Zufallsüberlagerung (log_mult) und anschließend einer Kim-Korrektur (log_kim) unterzogen. Hier ist vor allem die systematische Verzerrung der kleinsten 4 000 Einheiten durch die Kim-Korrektur zu erkennen.

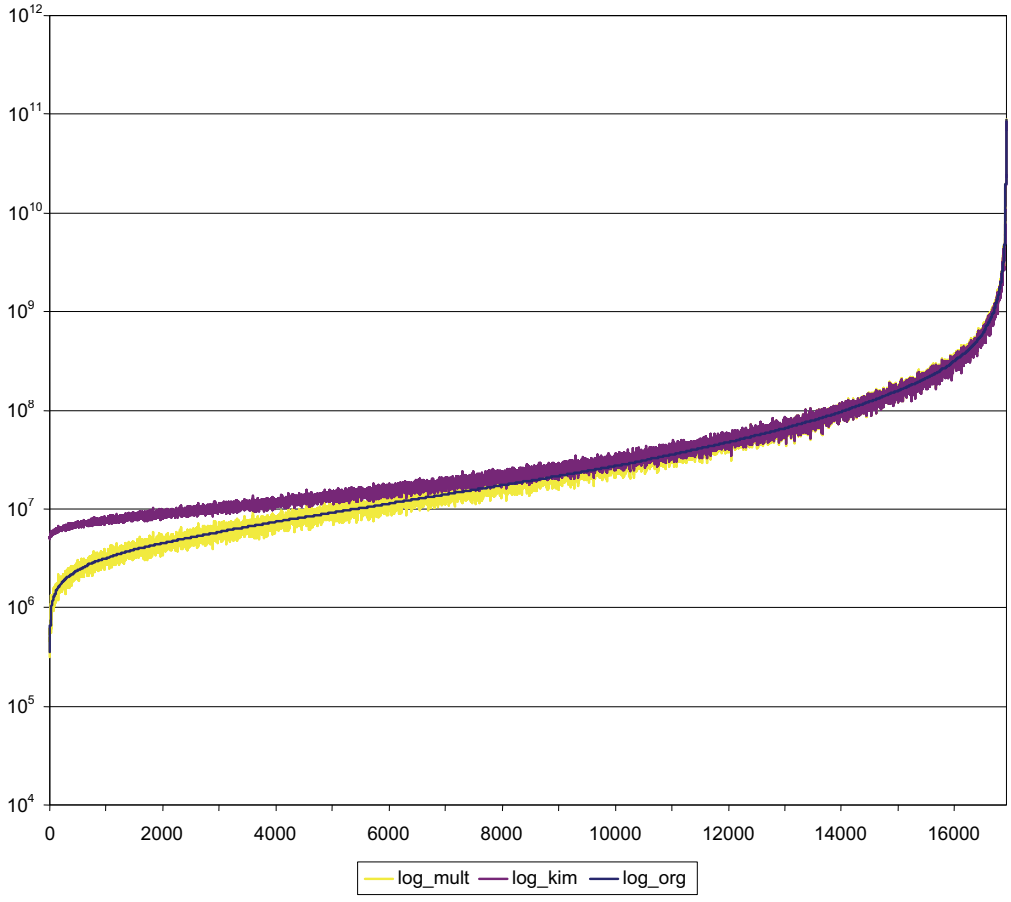


Abbildung 7
Kim-Korrektur bei Mikroaggregation

Die Abbildung 7 enthält die aufsteigend sortierten Originalwerte (log_org). Diese wurden im ersten Anonymisierungsschritt einer eindimensionalen Mikroaggregation und anschließende einer Kim-Korrektur (log_kim) unterzogen. Hier ist vor allem die systematische Verzerrung der kleinsten 2 000 Einheiten durch die Kim-Korrektur mit einigen Hundert negativen Werten zu erkennen. Die Kurven log_org und log_MA sind dagegen fast deckungsgleich.

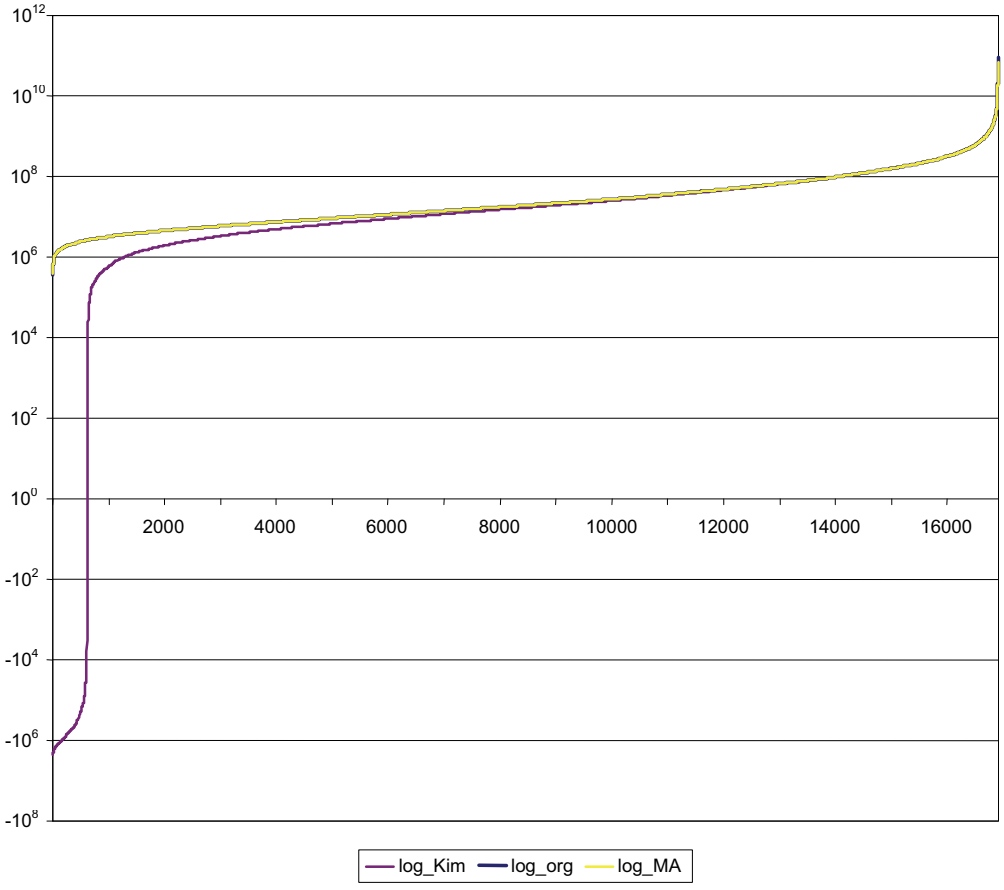
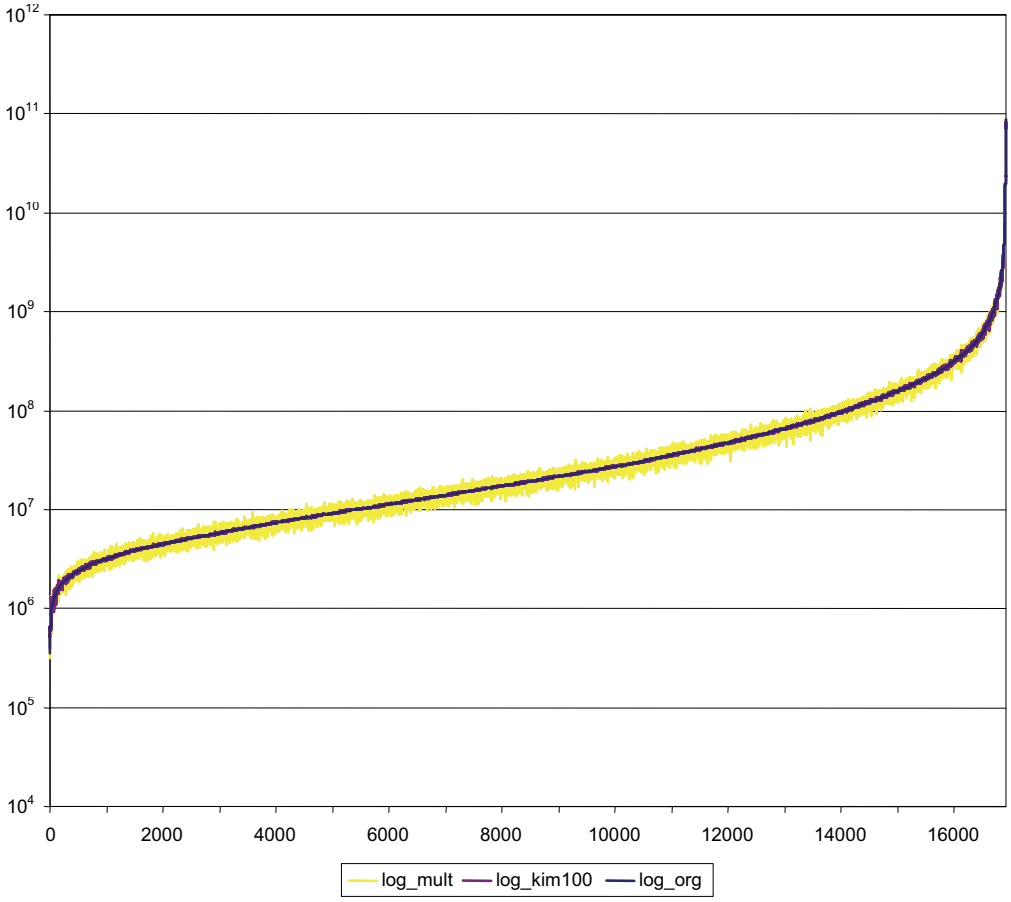


Abbildung 8
Blockweise Kim-Korrektur bei Zufallsüberlagerungen

Die Abbildung 8 enthält die aufsteigend sortierten Originalwerte (log_org). Diese wurden im ersten Anonymisierungsschritt einer multiplikativen Zufallsüberlagerung (log_mult) und anschließend einer blockweisen Kim-Korrektur (log_kim100) unterzogen. Die Kurven log_org und log_kim100 sind fast deckungsgleich.



3.3.2 Mikroaggregation mit Varianzerhalt in den Teilgruppen

Das folgende Verfahren basiert auf einem klassischen Gruppierungsverfahren. Die methodische Weiterentwicklung besteht in der Berechnung der anonymen Werte für die Mikroaggregationsgruppe. Es wurde in Höhne (2004b) erstmals vorgestellt.

Durch Gruppierungsverfahren (siehe z. B. Mateo-Sanz und Domingo-Ferrer 1998b), oder obige Abschnitte) wurden die Gruppen von 4 ähnlichen Merkmalsträgern bestimmt und dann der Datenbestand danach aufsteigend sortiert (k -Teilung mit $k=4$ siehe 2.2.6 – 1). Die Festlegung auf eine feste Gruppengröße $k=4$ wird hier zur vereinfachten einführenden Darstellung vorgenommen. Im späteren Teil werden die Formeln auch für die verallgemeinerte Form (variable Gruppengrößen) dargestellt.

Damit gehören die Sätze x_i, x_{i+1}, x_{i+2} und x_{i+3} in eine Mikroaggregationsgruppe G_h (mit $i=1,5,9,...$ und $h=(i+3)/4$). Die Werte innerhalb der Mikroaggregationsgruppe G_h werden jedoch nicht durch ihren Durchschnitt ersetzt, sondern es ergibt sich für die Merkmalsträger:

$$\overline{x_{G_h,j}} = \frac{x^o_{4h-3,j} + x^o_{4h-2,j} + x^o_{4h-1,j} + x^o_{4h,j}}{4} \quad ; \begin{matrix} h = 1,2,3,... \\ j = 1,2,3,...,m \end{matrix}$$

$$\sigma(x_{G_h,j}) = \sqrt{\frac{\sum_{i=0}^{3}\left(x^o_{4h-i,j} - \overline{x_{G_h,j}}\right)^2}{4}} \qquad \textbf{(3.3 – 1)}$$

$$x^{a-}_{G_h,j} = x^a_{4h-3,j} = x^a_{4h-2,j} = \overline{x^o_{G_h,j}} - \sigma(x^o_{G_h,j})$$

$$x^{a+}_{G_h,j} = x^a_{4h-1,j} = x^a_{4h,j} = \overline{x^o_{G_h,j}} + \sigma(x^o_{G_h,j})$$

Damit werden für jede Gruppe von 4 Merkmalsträgern zwei anonyme Werte $x_{l,j}^{a-}$ und $x_{l,j}^{a+}$ bestimmt. Beide weichen um genau eine Standardabweichung vom Gruppendurchschnitt ab. Bei je zwei Merkmalsträgern werden dann die originalen Werte durch die anonymen Werte $x_{l,j}^{a+}$ bzw. $x_{l,j}^{a-}$ ersetzt. Wird das Verfahren spaltenweise unabhängig angewandt, wird $x_{l,j}^{a+}$ den zwei größeren und $x_{l,j}^{a-}$ den zwei kleineren Originalwerten zugewiesen. Erfolgt die Anwendung mehrdimensional (siehe Abschnitt 2.2.6), sollte auch beachtet werden, dass nach der Zuweisung die Mehrdeutigkeit über die gleichzeitig bearbeiten Merkmale erhalten bleibt. Deshalb müsste die Gruppenbildung vorher stattfinden. Die Mikroaggregationsgruppe ist nochmals so zu teilen, dass zwei gleich große möglichst homogene Teilgruppen entstehen (z. B. durch nochmalige Anwendung der Mikroaggregationsregeln aber mit der Gruppengröße *k/2*). Die anonymen Werte $x_{l,j}^{a+}$ und $x_{l,j}^{a-}$ werden dann jeweils allen Merkmalswerten $x_{l+i,j}^{a}$ $(i=0,1,...)$ einer Gruppe zugewiesen, was am besten auf der Grundlage des Durchschnitts der Originalwerte $x_{l+i,j}^{o}(i=0,1,...)$ entschieden wird. Der größere anonyme Wert $x_{l,j}^{a+}$ wird der Gruppe mit dem größeren Durchschnitt der Originalwerte zugewiesen und $x_{l,j}^{a-}$ analog der Gruppe mit dem kleineren Durchschnitt. Diese Zuweisung nach dem Kriterium des Durchschnitts ist erforderlich, um systematische Verzerrungen in der Kovarianz und der Korrelation zu vermeiden. Wird die Mikroaggregation mit Varianzerhalt mehrdimensional angewendet und einer Teilgruppe immer systematisch die größeren Werte $x_{l,j}^{a+}$ und der anderen die kleineren Werte $x_{l,j}^{a-}$ für alle Merkmalsspalten zugewiesen, bedeutet dies, dass innerhalb des Mikroaggregates die Korrelation zwischen den Variablen 1 beträgt. Da die Varianz aber weiterhin mit dem Original identisch ist, wird offensichtlich die Kovarianz in den Mikroaggregaten systematisch überschätzt. Diese systematische Überschätzung der internen Kovarianzen in den Mikroaggregaten führt dann allerdings auch zu einer systematischen Verzerrung (Überschätzung) der Kovarianz und somit der Korrelationen im Gesamtbestand. Deshalb darf diese stark vereinfachte Zuweisung der Werte $x_{l,j}^{a+}$ und $x_{l,j}^{a-}$ nicht vorgenommen werden. Bei den Regeln für die Zuordnung der Werte $x_{l,j}^{a+}$ und $x_{l,j}^{a-}$ bestehen aber noch Entwicklungsmöglichkeiten für das Verfahren. Es wurde bisher in erster Linie eine unabhängige Anwendung auf die einzelnen Variablenspalten *j* untersucht.

Die Bildung von Gruppenaggregaten mit Varianzerhalt widerspricht eigentlich der Idee, die Werte der Merkmalsgruppe durch identische Werte (den Durchschnitt) zu ersetzen. Es muss mehr als ein Wert für die Repräsentanz der Gruppe verwendet werden (mindestens zwei Repräsentanten). Nur so ist es möglich, dass innerhalb des Mikroaggregates eine Varianz erhalten bleibt. Wenn der Effekt der Mehrdeutigkeit als primäre Quelle der Anonymität erhalten bleiben soll, bedeutet das ebenfalls, dass mindestens 4 Elemente in einem Mikroaggregat enthalten sein müssen (je zwei gleiche).

Alle bisherigen Mikroaggregationsverfahren vereinheitlichten mindestens 3 Werte innerhalb einer Aggregationsgruppe. Beim oben beschriebenen Ansatz werden jedoch nur mindestens 2 Werte vereinheitlicht. Ursache für die 3 Werte ist ein Restrisiko in den Aggregationsgruppen, das besteht, wenn ein Datenangreifer über die Originalinformation eines Merkmalsträgers der Gruppe bereits verfügt. Wenn ein Datenangreifer versucht, Wissen z. B. über einen Wirtschaftskonkurrenten zu beschaffen, dürfte es ihm leicht möglich sein, Informationen über den Auftraggeber als Hilfsinformationen zu verwenden. Bei den anderen Mikroaggregationsverfahren erfolgte die Anonymisierung durch das Ersetzen der Originalwerte mit dem Gruppendurchschnitt. Damit werden die Summen der Angaben in der Gruppe wieder fehlerfrei reproduziert und können somit als bekannt unterstellt werden. Bei einer Gruppengröße von 2 Objekten hätte damit leicht die Möglichkeit bestanden, durch Differenzbildung die Werte des Konkurrenten zu ermitteln, wenn man über die Kenntnis seiner eigenen Werte verfügt. Die zwei identischen Werte im Datenbestand beschreiben bei Mikroaggregation mit Varianzerhalt aber nicht mehr die gesamte Aggregationsgruppe. Bei den beiden identischen Merkmalsträgern gilt der Erhalt der Durchschnittswerte nicht. In der Regel gilt:

$$x^a_{l,j} + x^a_{l+1,j} \neq x^o_{l,j} + x^o_{l+1,j} .$$

Für die gesamte Aggregationsgruppe bleiben die Durchschnittswerte zwar erhalten, diese Gruppe besteht jedoch erstens aus mindestens 4 Merkmalsträgern. Zweitens sind die beiden Gruppen nicht mehr als zum gleichen Mikroaggregat zusammengehörig erkennbar, was selbst Dominanzprobleme innerhalb von Gruppen nicht mehr erkennen lässt. Deshalb sind auch Mikroaggregatgruppen aus 4 Sätzen ausreichend, obwohl im Verfahren nur jeweils zwei identische Sätze entstehen.

Da die Datenbestände nicht immer aus einer durch 4 teilbaren Anzahl bestehen, ergibt sich das Problem, dass die Gruppen nicht immer nur aus 4 Elementen bestehen können. Das Problem verstärkt sich, wenn man auch Gruppen des Datenbestandes einzeln behandeln will, um z. B. das Vorzeichen der Werte oder auch strukturelle Nullen zu erhalten. Es können auch Gruppen mit mehr als 4 Elementen notwendig sein. Größere Mengen von Mikroaggregationsgruppen mit einer Gruppengröße oberhalb der vorgegebenen Minimalgrenze k (hier $k=4$) entstehen auch, wenn man Verfahren der Mikroaggregation mit variabler Gruppengröße (siehe Abschnitt 2.2.6) verwendet. Dabei sind noch die Gruppengrößen $g = k+1, k+2, \ldots, 2k-1$ möglich. Größere Gruppengrößen sind wieder weiter in zwei Gruppen aus jeweils mindestens k Elementen teilbar, was auch ohne Qualitätsverlust möglich ist (siehe S. 48). Für beliebige Gruppen-Größen $g = k+1, k+2, \ldots$ ist als erstes eine Zahl g_1 vorzugeben, die die Anzahl der Elemente in der ersten Teilgruppe angibt (analog ist

$g_2 = g - g_1$ die Anzahl der Elemente in der zweiten Teilgruppe). Dabei ist g_1 so zu wählen, dass $g_1 \geq 2$ und $g_2 = (g - g_1) \geq 2$ gilt. Für den Durchschnitt der Gruppe gilt:

$$\overline{x^o_{G_h,j}} = \frac{\sum_{i \in G_h} x^o_{i,j}}{|G_h|} \quad ; \begin{matrix} h = 1,2,... \\ j = 1,2,...,m \end{matrix}$$

Die anonymen Werte werden durch addieren bzw. subtrahieren der Korrekturwerte a^+ bzw. a^- gebildet.

$$x^{a-}_{G_h,j} = x^a_{l(G_h)+i,j} = \overline{x^o_{G_h,j}} - a^-_{G_h,j} \quad ; i = 0,...,g_1 - 1$$
$$x^{a+}_{G_h,j} = x^a_{l(G_h)+i,j} = \overline{x^o_{G_h,j}} + a^+_{G_h,j} \quad ; i = g_1,...,g - 1$$

mit:

$l(G_h)$ – Position des ersten (kleinsten) Elements der Gruppe G_h

Um die Eigenschaft des Durchschnitts zu erhalten, muss folgende Beziehung gelten:

$$\overline{x^a_{G_h,j}} = \frac{\sum_{i=0}^{g-1} x^a_{l(G_h)+i,j}}{g} = \frac{\sum_{i=0}^{g_1-1}\left(\overline{x^o_{G_h,j}} - a^-_{G_h,j}\right) + \sum_{i=g_1}^{g-1}\left(\overline{x^o_{G_h,j}} + a^+_{G_h,j}\right)}{g} \quad ; \begin{matrix} h = 1,2,... \\ j = 1,2,...,m \end{matrix}$$
$$= \frac{g\,\overline{x^o_{G_h,j}} - \sum_{i=0}^{g_1-1}\left(a^-_{G_h,j}\right) + \sum_{i=g_1}^{g-1}\left(a^+_{G_h,j}\right)}{g} = \overline{x^o_{G_h,j}} + \frac{(g-g_1)\,a^+_{G_h,j} - g_1 a^-_{G_h,j}}{g}$$

daraus folgt:

$$\overline{x^a_{G_h,j}} = \overline{x^o_{G_h,j}} \quad \Leftrightarrow \quad 0 = (g-g_1)a^+_{G_h,j} - g_1 a^-_{G_h,j}$$
$$a^-_{G_h,j} = \frac{(g-g_1)}{g_1} a^+_{G_h,j}$$

Um die Eigenschaft der Varianz/Standardabweichung zu erhalten, muss folgende Beziehung gelten:

$$\sigma\left(x^{a}_{G_h,j}\right)=\sqrt{\frac{\sum_{i=0}^{g}\left(x^{a}_{l(G_h)+i,j}-\overline{x^{a}_{G_h,j}}\right)^2}{g}} \quad ;\begin{matrix} h=1,2,... \\ j=1,2,...,m \end{matrix}$$

aus $\overline{x^{a}_{G_h,j}}=\overline{x^{o}_{G_h,j}}$ folgt:

$$\sigma\left(x^{a}_{G_h,j}\right)=\sqrt{\frac{(g-g_1)\left(a^{+}_{G_h,j}\right)^2+g_1\left(a^{-}_{G_h,j}\right)^2}{g}}$$

bzw.

$$\sigma(x^{a}_{G_h,j})=\sqrt{\frac{(g-g_1)\left(a^{+}_{G_h,j}\right)^2+g_1\left(\frac{g-g_1}{g_1}a^{+}_{G_h,j}\right)^2}{g}}=\sqrt{\left(a^{+}_{G_h,j}\right)^2\left(\frac{(g-g_1)}{g}+\frac{(g-g_1)^2}{g_1 g}\right)}$$

$$=\left(a^{+}_{G_h,j}\right)\sqrt{\frac{g_1(g-g_1)+\left(g^2-2gg_1+g_1^2\right)}{gg_1}}=\left(a^{+}_{G_h,j}\right)\sqrt{\frac{g-g_1}{g_1}}$$

Für Varianzerhalt muss damit gelten:

$$a^{+}_{G_h,j}=\sqrt{\frac{g_1}{g-g_1}}\,\sigma(x^{o}_{G_h,j})$$

$$a^{-}_{G_h,j}=\frac{g-g_1}{g_1}a^{+}_{G_h,j}=\sqrt{\frac{g-g_1}{g_1}}\,\sigma(x^{o}_{G_h,j})$$

Mit diesen beiden Formeln besteht somit die Möglichkeit, für jede beliebige Gruppengröße $g \geq 4$ nach Festlegung der Größe der ersten Teilgruppe (g_1) die Korrekturwerte a^+ und a^- so zu bestimmen, dass der Durchschnitt und die Varianz in der Aggregationsgruppe erhalten bleibt. Für symmetrische Gruppen ($g_1 = g - g_1 = g_2$) sind die Korrekturwerte identisch mit der Standardabweichung.

3.3.3 Auswirkungen varianzerhaltender Mikroaggregation auf die Ergebnisse von OLS-Schätzungen

Im Folgenden soll untersucht werden, wie sich die unabhängige varianzerhaltende Mikroaggregation auf die Ergebnisse des linearen Korrelationskoeffizenten von Bravais-Pearson auswirkt.

Vereinfachend sei angenommen, dass die Gruppengröße über den gesamten Datenbestand 4 beträgt (mit je 2 Objekten in den identischen Teilgruppen) und die Mikroaggre-

gation unabhängig für jedes Merkmal (eindimensional) durchgeführt wird. Damit ist es zufällig, ob sich innerhalb einer Zeile die beiden Variablen als Mittelwert + Standardabweichung oder Mittelwert – Standardabweichung ergeben. Außerdem ist es eher unwahrscheinlich, dass sich die Mittelwerte und Standardabweichungen auf die gleiche Gruppe von Merkmalsträgern beziehen. Trotzdem gilt durch die varianzerhaltende Mikroaggregation der Summenerhalt ($\Sigma(x^a)=\Sigma(x^o)$) und der Varianzerhalt ($\sigma(x^a)=\sigma(x^o)$) innerhalb der einzelnen bearbeiteten Gruppen und somit auch jeweils für die gesamte Datenspalte.

Für den Korrelationskoeffizienten r zwischen den Merkmalen j und k gilt:

$$r = \frac{1/n\sum_{i=1}^{n}(x_{ij} - \overline{x_{.j}})(x_{ik} - \overline{x_{.k}})}{\sigma(x_{.j})\sigma(x_{.k})};$$

mit $\sigma(x_{.j})$ und $\sigma(x_{.k})$ als Standardabweichungen der Datenspalten $x_{.j}$ und $x_{.k}$.

Die Auswirkungen dieser Form der Mikroaggregation auf den Korrelationskoeffizienten lässt sich am besten an der Differenz $r^d=r^a-r^o$ untersuchen.

Für die Differenz gilt:

$$r_{jk}^d = r_{jk}^a - r_{jk}^o = \frac{1/n\sum_{i=1}^{n}(x_{ij}^a - \overline{x_{.j}^a})(x_{ik}^a - \overline{x_{.k}^a})}{\sigma(x_{.j}^a)\sigma(x_{.k}^a)} - \frac{1/n\sum_{i=1}^{n}(x_{ij}^o - \overline{x_{.j}^o})(x_{ik}^o - \overline{x_{.k}^o})}{\sigma(x_{.j}^o)\sigma(x_{.k}^o)}$$

Da die Standardabweichungen und Mittelwerte durch die Art der Anonymisierung nicht verändert wurden gilt:

$$r_{jk}^d = r_{jk}^a - r_{jk}^o = \frac{1/n\sum_{i=1}^{n}(x_{ij}^a - \overline{x_{.j}^o})(x_{ik}^a - \overline{x_{.k}^o})}{\sigma(x_{.j}^o)\sigma(x_{.k}^o)} - \frac{1/n\sum_{i=1}^{n}(x_{ij}^o - \overline{x_{.j}^o})(x_{ik}^o - \overline{x_{.k}^o})}{\sigma(x_{.j}^o)\sigma(x_{.k}^o)}$$

bzw.

$$r_{jk}^d = r_{jk}^a - r_{jk}^o = \frac{\sum_{i=1}^{n}(x_{ij}^a - \overline{x_{.j}^o})(x_{ik}^a - \overline{x_{.k}^o}) - \sum_{i=1}^{n}(x_{ij}^o - \overline{x_{.j}^o})(x_{ik}^o - \overline{x_{.k}^o})}{n\sigma(x_{.j}^o)\sigma(x_{.k}^o)}$$

Es sei $x_{ij}^d = x_{ij}^a - x_{ij}^o$ die Veränderung, die bei der Anonymisierung vorgenommen wird. Dann gilt innerhalb der Mikroaggregationsgruppen $\Sigma(x_{ij}^d)=0$, und somit:

$$r_{jk}^{d} = r_{jk}^{a} - r_{jk}^{o} = \frac{\sum_{i=1}^{n}(x_{ij}^{d} + x_{ij}^{o} - \overline{x_{.j}^{o}})(x_{ik}^{d} + x_{ik}^{o} - \overline{x_{.k}^{o}}) - \sum_{i=1}^{n}(x_{ij}^{o} - \overline{x_{.j}^{o}})(x_{ik}^{o} - \overline{x_{.k}^{o}})}{n\sigma(x_{.j}^{o})\sigma(x_{.k}^{o})}$$

$$= \frac{\sum_{i=1}^{n}\left(\begin{pmatrix} x_{ij}^{d}x_{ik}^{d} + x_{ij}^{d}x_{ik}^{o} - x_{ij}^{d}\overline{x_{.k}^{o}} + x_{ij}^{o}x_{ik}^{d} + x_{ij}^{o}x_{ik}^{o} \\ - x_{ij}^{o}\overline{x_{.k}^{o}} - \overline{x_{.j}^{o}}x_{ik}^{d} - \overline{x_{.j}^{o}}x_{ik}^{o} + \overline{x_{.j}^{o}}\,\overline{x_{.k}^{o}} \end{pmatrix} - \begin{pmatrix} x_{ij}^{o}x_{ik}^{o} - x_{ij}^{o}\overline{x_{.k}^{o}} \\ - \overline{x_{.j}^{o}}x_{ik}^{o} + \overline{x_{.j}^{o}}\,\overline{x_{.k}^{o}} \end{pmatrix}\right)}{n\sigma(x_{.j}^{o})\sigma(x_{.k}^{o})}$$

$$= \frac{\sum_{i=1}^{n}\left(x_{ij}^{d}x_{ik}^{d} + x_{ij}^{d}x_{ik}^{o} - x_{ij}^{d}\overline{x_{.k}^{o}} + x_{ij}^{o}x_{ik}^{d} - \overline{x_{.j}^{o}}x_{ik}^{d}\right)}{n\sigma(x_{.j}^{o})\sigma(x_{.k}^{o})}$$

Für die einzelnen Spalten j und k gilt bei varianzerhaltender Mikroaggregation in jeder Mikroaggregationsgruppe aus 4 Einheiten $\Sigma_i(x_{ij}^{\ d})=0$ und $\Sigma_i(x_{ik}^{\ d})=0$. Damit existieren immer Abweichungen größer und kleiner 0 in jeder Gruppe. Anders als beim Rank-Swapping wird die Datenveränderung nicht mit Transformationspartnern zufällig aus dem Datenbestand, sondern mit den nach einspaltiger Sortierung unmittelbar benachbarten durchgeführt. Damit tritt für die $x_{ij}^{\ d}$ einerseits eine Minimierung ein, andererseits werden sie unabhängig, weil die einspaltige Sortierung für zwei verschiedene Datenspalten in der Regel nicht identische Reihenfolgen ergibt. Ob die einzelnen Werte jedoch einer Veränderung nach oben oder unten unterzogen werden hängt von der konkreten Position in der Sortierreihenfolge und den Größenunterschieden zwischen den $x_{ij}^{\ o}$ in der Mikroaggregationsgruppe ab (bestimmt Gruppenmittelwert und Varianz). Gerade in der Kombination zu einer anderen Merkmalsspalte können diese Veränderungen für die Einheit *i* deshalb nur als zufällig betrachtet werden.[10]

Die Abweichungen können deshalb als unabhängige, zufällige Ereignisse betrachtet werden. Wegen der im Verfahren streng gesicherten Eigenschaft $\Sigma_i x_{ij}^{\ d}=0$ für jede Teilgruppe, gilt die Eigenschaft auch für den Gesamtbestand $\Sigma_i x_{ij}^{\ d}=0$, so dass für den Fehler im Korrelationskoeffizienten durch die Mikroaggregation gilt:

10 Um das zu illustrieren, soll die Transformation an einem vereinfachten Beispiel dargestellt werden. Die Anonymisierung sei exemplarisch nach folgendem Algorithmus erfolgt. Der Datenbestand wird nach Merkmal *j* absteigend sortiert und dann abwechselnd um den Betrag *D* erhöht beziehungsweise verringert. Damit wäre

$x_{ij}^{\ a} = x_{ij}^{\ o} + D$ *und* $x_{i+1,j}^{\ a} = x_{i+1,j}^{\ o} - D$; $i=1,3,5,\ldots$

Für die Datenspalte *m* erfolgt die Transformation analog. Insgesamt besteht aber keine Abhängigkeit zwischen $x_{ij}^{\ d}$ und $x_{im}^{\ d}$, da einerseits für beide Differenzen nur die Möglichkeit von (*-D* und *+D*) besteht, die Wertekombination $x_{ij}^{\ d}$ und $x_{im}^{\ d}$ in der Zeile i aber zufällig ist. Für die Wertepaare $(x_{ij}^{\ d}, x_{im}^{\ d})$ treten die Kombinationen (*-D, -D*), (*-D, +D*), (*+D, +D*), (*+D, -D*) gleichwahrscheinlich auf.

$$r_{jk}^{d} = \frac{\left(\sum_{i=1}^{n}\left(x_{ij}^{d}x_{ik}^{d}\right)+\sum_{i=1}^{n}\left(x_{ij}^{d}x_{ik}^{o}\right)-\overline{x_{.k}^{o}}\sum_{i=1}^{n}\left(x_{ij}^{d}\right)+\sum_{i=1}^{n}\left(x_{ij}^{o}x_{ik}^{d}\right)-\overline{x_{.j}^{o}}\sum_{i=1}^{n}\left(x_{ik}^{d}\right)\right)}{n\sigma(x_{.j}^{o})\sigma(x_{.k}^{o})}$$

$$= \frac{\left(\sum_{i=1}^{n}\left(x_{ij}^{d}x_{ik}^{d}\right)+\sum_{i=1}^{n}\left(x_{ij}^{d}x_{ik}^{o}\right)+\sum_{i=1}^{n}\left(x_{ij}^{o}x_{ik}^{d}\right)\right)}{n\sigma(x_{.j}^{o})\sigma(x_{.k}^{o})}$$

Fehler in der Korrelation zwischen den Variablen $r_{jk}{}^{d}$ durch die Anonymisierung sind somit die Summe von Produkten aus Einzelveränderunge $x_{ij}{}^{d}$. Da die Veränderungen der Einzelwerte $x_{ij}{}^{d}$ jedoch gleichwahrscheinlich ihre Vorzeichen wechseln, kommt es zu starken Kompensationseffekten so dass $r_{jk}{}^{d}$ sehr klein wird. Dieser Effekt verstärkt sich, je größer die Anzahl an Objekten im Datenbestand wird (siehe z. B. Schmid 2007).

Anhand einer Simulation soll diese Aussage näher untersucht werden. Dazu wurden jeweils 250 Testdatenbestände mit jeweils 12, 120 und 1 200 Einheiten und zwei Variablen sowie einer theoretischen Korrelation von 0,7 zwischen den Variablen generiert. Anschließend wurde die Korrelation für die Testdatenbestände vor und nach der Anonymisierung ermittelt. Die Ergebnisse sind in folgenden Abbildungen dargestellt. Bei sehr kleinen Testdatenbeständen ist sowohl die theoretische Korrelation schlecht erhalten als auch ein relativ großer Fehler durch die Anonymisierung erzeugt. Mit zunehmender Stichprobengröße nehmen beide Fehlereffekte stark ab. In allen Fällen bleibt der Fehler durch die Stichprobenauswahl (horizontale Abweichung der Punkte von der theoretischen Korrelation von 0,7) jedoch größer als der Fehler durch die Anonymisierung (Differenz der Testdatenkorrelation anonym zu original als Abweichung der Punkte von der eingezeichneten Diagonale).

Abbildung 9
Mikroaggregation mit Varianzerhalt bei 12 Einheiten (n=12)

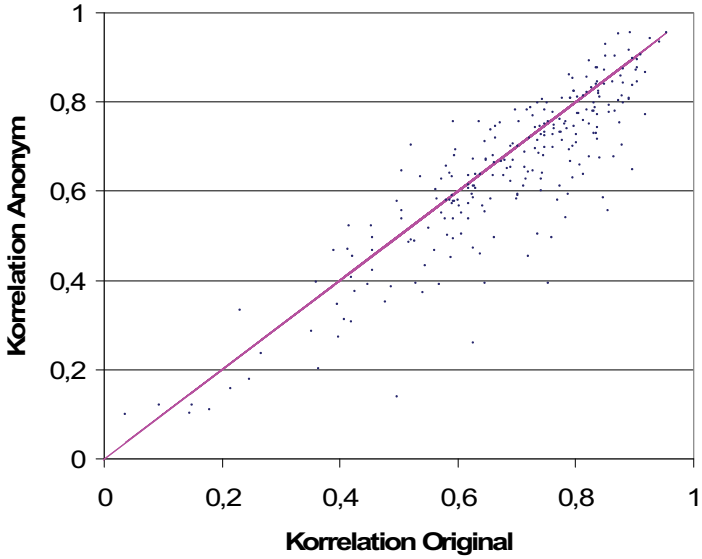


Abbildung 10
Mikroaggregation mit Varianzerhalt bei 120 Einheiten (n=120)

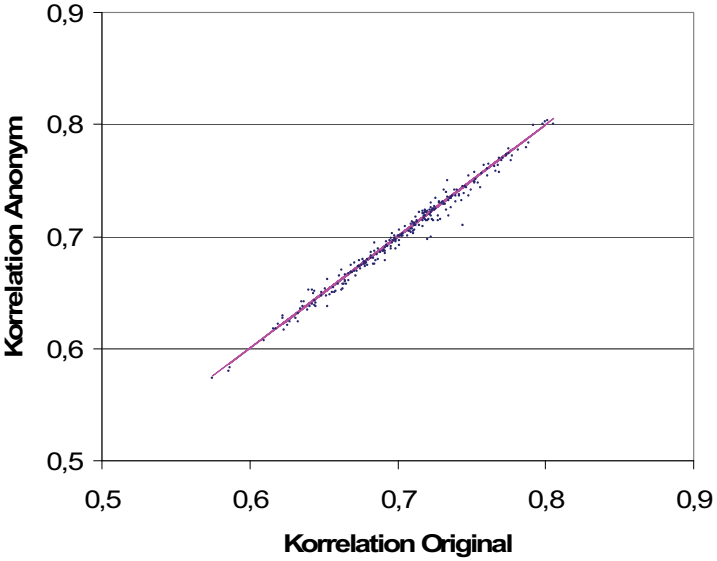
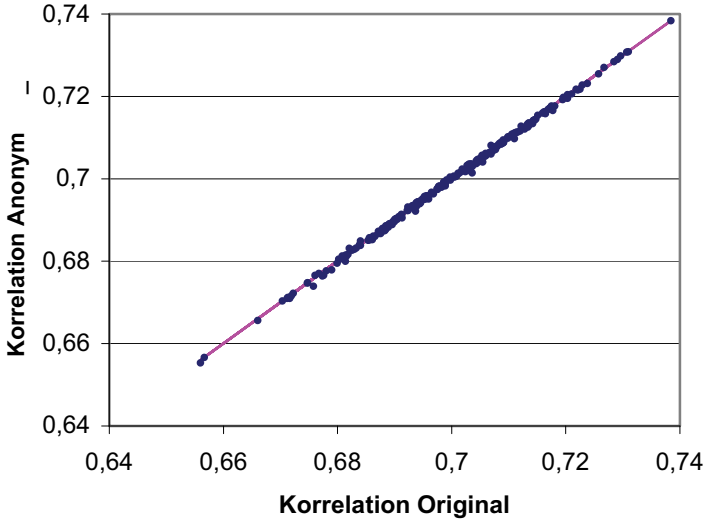


Abbildung 11
Mikroaggregation mit Varianzerhalt bei 1 200 Einheiten (n=1200)



Ähnlich gering ist der Einfluss auf die Rangkorrelationen nach Spearman. Da die Mikroaggregation spaltenweise unabhängig erfolgt, werden die Merkmalsträger für jedes Merkmal unabhängig absteigend also nach ihrer Rangfolge sortiert. Die zwei größeren Merkmalsträger einer Gruppe erhalten als neue Ausprägung Mittelwert + Standardabweichung und die zwei kleineren als Mittelwert – Standardabweichung. Je zwei Sätze werden damit

identisch, was somit auch für ihre Ränge gilt. Die Rangfolge zwischen den großen und den kleinen Merkmalsträgern einer Gruppe bleibt aber erhalten. Ebenso wird die Rangfolge zwischen den Gruppen in der Regel nicht verändert.

Bei der Rangfolge zwischen den Gruppen gibt es jedoch auch Ausnahmen, wenn die Verteilung innerhalb der Mikroaggregationsgruppe sehr schief ist.

Tabelle 1: Datenbeispiel für Mikroaggregation mit Varianzerhalt

Lfd. Nr.	Sortierte originale Werte	Mittelwert	Standard-abweichung	Anonyme Werte
1	1 200	1 215	11,2	1 203,8
2	1 210	1 215	11,2	1 203,8
3	1 220	1 215	11,2	1 226,2
4	1 230	1 215	11,2	1 226,2
5	1 240	1 255	11,2	1 243,8
6	1 250	1 255	11,2	1 243,8
7	1 260	1 255	11,2	1 266,2
8	1 270	1 255	11,2	1 266,2
9	1 280	1 340	86,9	1 253,1
10	1 290	1 340	86,9	1 253,1
11	1 300	1 340	86,9	1 426,9
12	1 490	1 340	86,9	1 426,9
Mittelwert	1 270			1 270,0
Standard-abweichung	72,9			72,9

Für die sehr gleichmäßig verteilten originalen Werte bildet der 12. Wert eine Ausnahme. Innerhalb der Mikroaggregationsgruppe (Werte 9 bis 12) liegt der Mittelwert nicht mehr zentral sondern zur unteren Schranke der Werte verzerrt. Dadurch führt die Anonymisierung der kleineren Werte mit Mittelwert – Standardabweichung zu kleineren Werten als in der Gruppe selbst. Damit kann es passieren, dass die Reihenfolge der Originalwerte nach der Anonymisierung nicht mehr erhalten bleibt. Die Veränderung der Reihenfolge ist für die Anonymität der Daten positiv, weil so im Gegensatz zu sonstigen eindimensionalen Mikroaggregationen keine obere und untere Schranken mehr bestimmbar sind (siehe Abschnitt 3.2). Die Rangkorrelationen werden aber durch derartige Datenkonstellationen nicht grundlegend verzerrt.

Im Extremfall kann die Standardabweichung aber auch größer als der Mittelwert sein, was zu negativen anonymen Werten führen würde. Aus diesem Grund wurde bei Testrechnungen eine Mindestgröße der Gruppen von $k>4$ (z. B. $k=6$) verwendet. Dann besteht die Möglichkeit, bei derartigen Datenkonstellationen die Teilgruppe der größeren Werte zu verkleinern, so dass die Abweichung nach unten auf mehr Werte aufgeteilt werden kann. Mit der Wahl von $g_2=2$ konnten dann negative Werte fasst immer verhindert werden.

Führt man mit derartig anonymisierten Daten einfache lineare Regressionsschätzungen (OLS) durch, so gilt:

Für das lineare Modell

$$y = X\beta^a + u$$

ergibt sich der OLS-Schätzer durch:

$$\hat{\beta}^a = (X'X)^{-1} X'y$$

Um die Auswirkungen der unabhängigen varianzerhaltenden Mikroaggregation abschätzen zu können muss deshalb der Einfluss auf $(X'X)^{-1}$ und $X'y$ näher untersucht werden. Es seien X^o und y^o die originalen Werte von X und y, sowie X^a und y^a die anonymisierten Werte. Die Veränderung durch die Anonymisierung sei

$$X^d = X^a - X^o$$

und

$$y^d = y^a - y^o.$$

Für die Matrix $X'X$ gilt dann:

$$\left(X^{a\prime}X^a\right) = \left(\left(X^d + X^o\right)'\left(X^d + X^o\right)\right)$$

Für ihr Element k,l gilt:

$$\left(X^{a\prime}X^a\right)_{kl} = \sum_{i=1}^{n}\left(\left(x_{ik}^d + x_{ik}^o\right)\left(x_{il}^d + x_{il}^o\right)\right) = \sum_{i=1}^{n}\left(x_{ik}^d x_{il}^d + x_{ik}^o x_{il}^d + x_{ik}^d x_{il}^o + x_{ik}^o x_{il}^o\right)$$

Und wegen der Unabhängigkeit der Anonymisierung der Spalten und der deshalb auftretenden Kompensationseffekte bei den ständig ihr Vorzeichen wechselnden Summanden in den ersten drei Teilsummen wird:

$$\left(X^{a\prime}X^a\right)_{kl} = \sum_{i=1}^{n}\left(x_{ik}^d x_{il}^d\right) + \sum_{i=1}^{n}\left(x_{ik}^o x_{il}^d\right) + \sum_{i=1}^{n}\left(x_{ik}^d x_{il}^o\right) + \sum_{i=1}^{n}\left(x_{ik}^o x_{il}^o\right) \approx \sum_{i=1}^{n}\left(x_{ik}^o x_{il}^o\right)$$

$$X^{a\prime}X^a \approx X^{o\prime}X^o$$

Analog gilt für $X'y$:

$$\left(X^{a\prime}y^a\right) = \left(\left(X^d + X^o\right)'\left(y^d + y^o\right)\right)$$

$$\left(X^{a\prime}y^a\right)_k = \sum_{i=1}^{n}\left(\left(x_{ik}^d + x_{ik}^o\right)\left(y_i^d + y_i^o\right)\right) = \sum_{i=1}^{n}\left(x_{ik}^d y_i^d + x_{ik}^o y_i^d + x_{ik}^d y_i^o + x_{ik}^o y_i^o\right)$$

und wegen der Unabhängigkeit der Anonymisierung der Spalten gilt für Erwartungswert:

$$\left(X^{a\prime}y^a\right) = \left(\left(X^d + X^o\right)'\left(y^d + y^o\right)\right)$$

$$\left(X^{a\prime}y^a\right)_k = \sum_{i=1}^{n}\left(x_{ik}^d y_i^d\right) + \sum_{i=1}^{n}\left(x_{ik}^o y_i^d\right) + \sum_{i=1}^{n}\left(x_{ik}^d y_i^o\right) + \sum_{i=1}^{n}\left(x_{ik}^o y_i^o\right) \approx \sum_{i=1}^{n}\left(x_{ik}^o y_i^o\right)$$

$$\left(X^{a\prime}y^a\right) \approx X^{o\prime}y^o$$

Für den OLS-Schätzer ergibt sich deshalb:

$$\hat{\beta}^a = \left(X^{a\prime}X^a\right)^{-1} X^{a\prime}y^a \approx \left(X^{o\prime}X^o\right)^{-1} X^{o\prime}y^o \approx \hat{\beta}^o$$

OLS-Schätzungen reproduzieren bei unabhängiger, spaltenweiser varianzerhaltender Mikroaggregation zu den Originalschätzungen ähnliche Werte, wobei die Genaugikeit der Ergebnisse mit der Größe des Datenumfangs zunimmt.

Für Anonymisierungsverfahren, die einerseits den Mittelwert und die Varianz erhalten und andererseits unabhängig auf die einzelnen Variablenspalten angewendet werden, leitet *Moore (1996)* folgenden Zusammenhang her:

$$E[R(x._j^a, x._k^a)] = R(x._j^o, x._j^a) * R(x._k^o, x._k^a) * R(x._j^o, x._k^o)$$

Dieser Zusammenhang wurde von Moore zwar für das Rankswappingverfahren hergeleitet, die unterstellten Voraussetzungen sind bei unabhängiger spaltenweiser Mikroaggregation jedoch analog gegeben. Im Anhang 2 (siehe S. 134 ff. in diesem Band) ist der Beweis deshalb in der allgemeineren Form dargestellt.

3.4 SAFE-Ansatz

Das Verfahren SAFE wurde im Statistischen Landesamt Berlin entwickelt. Die Entwicklung dauerte über mehrere Jahre, in denen verschiedene Ansätze getestet wurden. Im Folgenden soll deshalb eine stark verallgemeinerte Beschreibung des Ansatzes vorgenommen werden und anschließend zwei Varianten vorgestellt werden, die auch bei der statistischen Produktion eingesetzt wurden.

3.4.1 Grundidee des SAFE-Verfahrens

Mit dem SAFE-Verfahren soll eine Anonymisierung von Mikrodaten so vorgenommen werden, dass sie auch für die Kriterien der Tabellengeheimhaltung ausreichend ist.

Grundidee des SAFE-Verfahrens ist die Vereinheitlichung von Merkmalsträgern im Mikrodatenbestand derart, dass mindestens 3 Merkmalsträger völlig identisch sind. Wenn die Merkmalsträger sowohl in ihren kategorialen Merkmalen als auch in den metrischen Merkmalen identische Ausprägungen haben, wird erreicht, dass aus diesen Mikrodaten erzeugte Auswertungstabellen keine Fallzahlprobleme als Geheimhaltungsfälle besitzen können, da immer mindestens drei Objekte in den veröffentlichten Tabellenfeldern zusammengefasst wurden. Eventuelle Dominanzprobleme der Tabellengeheimhaltung werden im Laufe des Verfahrens ebenfalls gelöst. Haben Wertefelder in den Auswertungstabellen ein Dominanzproblem, darf die zu schützende Originalinformation nicht veröffentlicht werden. Im Gegensatz zur klassischen Tabellengeheimhaltung, wo dieser Schutz der Information durch Feldersperrung („Auspunkten" von Tabellenfeldern) erfolgt, besteht bei SAFE auch die Möglichkeit, Informationen zu erzeugen, die von der Originalangabe ausreichend abweichen (keine Brauchbarkeit mehr für den Angreifer besitzen). In diesem Punkt gleicht dieser Ansatz den anderen Verfahren der Einzeldatenanonymisierung (und nicht der Tabellenanonymisierung), geht aber auch über diese hinaus, weil diese den Anspruch nur für die Reidentifikationsrisiken bei eventuell erfolgreichen Einzeldatenzuordnungen besitzen. Anonymität bei allen Tabellierungen ist dort nicht Gegenstand der Untersuchungen.

Bei den metrischen Merkmalen wird die Vereinheitlichung von mindestens 3 Merkmalsträgern durch Durchschnittsbildung erreicht. Damit ist SAFE ein Verfahren aus der Klasse der Mikroaggregationen. Der Unterschied besteht jedoch darin, dass kategoriale Merkmale in einem separaten Anonymisierungsschritt behandelt und vereinheitlicht werden.

Bei den anderen Mikroaggregationsverfahren konnten kategoriale Merkmale nur durch eine Transformation in metrische Variablen mit anonymisiert werden (siehe Abschnitt 2.2.1). Während andere Verfahren der Mikroaggregation die Auswahl für die Gruppenbildung ausschließlich anhand von Abstandsmaßen zwischen den Merkmalsträgern treffen, ist bei SAFE vor allem der Einfluss auf die Tabellierungseigenschaften wichtig. Veränderungen an den kategorialen Merkmalen sind deshalb so vorzunehmen, dass sie sich im Rahmen des Gesamtbestandes möglichst kompensieren.

Somit kann das Verfahren wegen der Grundidee der Durchschnittsbildung als ein Mikroaggregationsverfahren betrachtet werden. Die einzelnen anonymen Merkmalsträger sind nicht unbedingt das Ergebnis einer eindeutigen Transformation aus den originalen Werten der gruppierten Merkmalsträger, sondern die anonymen Daten sind mit dem primären Ziel erzeugt worden, qualitative Eigenschaften des Gesamtdatenbestandes möglichst gut zu erhalten. Gerade die nicht bestehenden direkten funktionalen Abhängigkeiten zwischen den originalen und den anonymisierten Merkmalsträgern sowie die gleichzeitig bestehende Mehrdeutigkeit der existierenden Kombinationen von Merkmalsausprägungen sind die Grundlage der hohen Schutzwirkung der anonymisierten Lösung. Andererseits hat der fehlende direkte funktionale Zusammenhang auch entsprechend stärkere Einbußen in der Analysequalität zur Folge. Die Lösungen sollte man deshalb in ihrer Qualität eher mit der von Simulations- oder Imputationsverfahren vergleichen.

Die Grundidee des SAFE-Ansatzes geht auf Appel zurück (z. B. Appel et al. 1993). Erste Ansätze zur Bestimmung einer solchen Lösung wurden im Statistischen Landesamt Berlin bereits gegen Ende der 1980er Jahre unternommen. Da die Verfahrensansätze im Laufe der Zeit für verschiedene Datenbestände und von verschiedenen Personen weiterentwickelt wurden, haben sich verschiedene Lösungsansätze herausgebildet.

Im Folgenden wird versucht, das Problem der SAFE-Geheimhaltung in seiner allgemeinsten Form darzustellen. Es wird dabei klar, dass es in dieser Form und bei den Größenordnungen realer Datenbestände nicht eindeutig lösbar ist. Es bestehen dabei die gleichen Probleme, wie bei der mehrdimensionalen Mikroaggregation. Dort wurde bereits von Oganian und Domingo-Ferrer gezeigt (siehe Oganian und Domingo-Ferrer 2001), dass die Anzahl der Möglichkeiten so groß ist, dass das Problem für reale Datenbestände nicht eindeutig lösbar ist. Daraus erklärt sich auch das Entstehen von verschiedenen Ansätzen zur Bestimmung von Näherungslösungen.

3.4.2 Mathematische Formulierung der Grundidee

Das Ziel der SAFE-Geheimhaltung lässt sich folgendermaßen darstellen:

Der Datenbestand der metrischen Werte der originalen Daten sei die Matrix X^o, bei der jede statistische Einheit eine Zeile $i=1,2,...,n$ darstellt. X_{ij}^o enthält deshalb die Wertangabe des Unternehmens i im Merkmal j $(j=1,2,...,m)$.

Die k kategorialen Merkmale des Datenbestandes seien durch Zuordnungsmatrizen Z_j^o $(j=1,...,k)$ abgebildet. Diese bestehen aus s_j Spalten (für jede auftretende Merkmalsausprägung des kategorialen Merkmals k) und wieder n Zeilen (für jede statistische Einheit i). Diese Zuordnungsmatrizen bestehen aus Zeilen mit lauter Nullen und nur einer 1 je Zeile an der Stelle, die der Merkmalsausprägung der statistischen Einheit i entspricht. Für

jedes kategoriale Merkmal lässt sich eine solche Zuordnungsmatrix angeben. Für Kombinationen von kategorialen Merkmalen lassen sich ebenfalls die dazugehörigen Zuordnungsmatrizen (Z_{jl}) bestimmen. Für diese Merkmalskombinationen kann die dazugehörige Zuordnungsmatrix Z_{jl} theoretisch ermittelt werden, indem man die i Zeilen der Matrix über

$$Z_{i,jl} = Z_{i,j} \otimes Z_{i,l};\ i = 1,2,...,n \text{ und } j \neq l$$

mit

$\otimes$ – Kroneckerprodukt der beiden Vektoren

berechnet.

Damit besteht auch jede Zuordnungsmatrix für Merkmalskombinationen aus n Zeilen, die Anzahl der Spalten beträgt aber $s_j * s_l$. Weiterhin besteht jede Zeile aus lauter Nullen und nur einer 1 an der Stelle, an der $Z_{i,j} = Z_{i,l} = 1$ gilt.

Bei mehr als zwei Merkmalen erfolgt die Erstellung analog durch wiederholte Anwendung der Berechnungsvorschrift. (z. B. $Z_{i,jlh} = Z_{i,jl} \otimes Z_{i,h}$).

Gerade wenn mehrere stark untergliederte kategoriale Merkmale kombiniert werden, kann es leicht dazu führen, dass die Anzahl der Spalten der Zuordnungsmatrix (theoretische Anzahl an Kombinationen der Merkmalsausprägungen) größer als die existierende Anzahl an Kombinationen ist, so dass viele Nullvektoren als Spalten der Matrix auftreten. Das kann einerseits daran liegen, dass bestimmte Kombinationen von Merkmalsausprägungen sich inhaltlich ausschließen[11] oder aber der Datenbestand nicht groß genug ist, um für jede theoretische Kombination einen Repräsentanten zu enthalten. Im Extremfall ist die Anzahl der theoretischen Ausprägungskombinationen größer als die Anzahl der Einheiten im Datenbestand. Die Anzahl der im Bestand existierenden Ausprägungskombinationen und somit der notwendigen Spalten der Zuordnungsmatrizen ist aber immer durch die Anzahl der Einheiten im Datenbestand beschränkt. Die Zuordnungsmatrizen sollten deshalb aus Gründen der Größe des mathematischen Modells und somit der Effektivität bei der Berechnung für Merkmalskombinationen aus den bestehenden Ausprägungskombinationen und nicht aus den theoretisch möglichen hergeleitet werden.

Für Auswertungen werden die Tabellierungen dann einfach über die Multiplikation $T_j^o = (Z_j^o)' X^o$ vorgenommen. Die so erhaltene Tabelle T_j^o enthält dann die Summen aller metrischen Werte (in der gleichen Spaltenreihenfolge wie die Originaldatei) für die Ausprägungen, die durch die Spalten der Zuordnungsmatrix Z_j^o definiert sind. Für Häufigkeitstabellierungen wird die Anzahl der Einheiten (wie z. B. Unternehmen oder Betriebe) in den Merkmalsausprägungen durch $A_j^o = (Z_j^o)' Z_j^o$ bestimmt. Das Ergebnis ist die Diagonalmatrix A_j^o mit der Anzahl der Einheiten auf der Hauptdiagonale. Die zugehörigen Ausprägungen ergeben sich durch die Spalten-/Zeilenposition des Diagonalelements und die entsprechende Definition der Spalte von Z_j^o. Durchschnittsangaben können ggf. durch Division mit diesen Häufigkeiten ermittelt werden. Aus den Werten der Matrix lassen sich aber auch alle gewünschten Verhältnisangaben oder Zeitraumvergleiche ermitteln.

11 So können sich bestimmte Kombinationen aus Wirtschaftszweigen und Regionen wie z. B. Landwirtschaft oder Bergbau in innerstädtischen Bereichen oder aber Kombinationen aus Alters- und Berufsgruppen (z. B. Berufe bei Kindern) inhaltlich ausschließen. In diesem Fall hätte eine theoretisch hergeleitete Zuordnungsmatrix für diese Ausprägungskombinationen Spalten die reine Nullvektoren wären.

Eine anonyme Lösung, die dem SAFE-Ansatz gerecht wird, bildet eine Matrix ($X^a \neq X^o$) mit einem Satz von Zuordnungsmatrizen Z_j^a die analog zu Z_j^o definiert sind. X^a und Z_j^a bestehen aus nur noch maximal *n/3* Zeilen. Zusätzlich zur Matrix X^a und den Matrizen Z_j^a gehört zur Lösung eine Häufigkeitsmatrix H^a. H^a ist eine Diagonalmatrix (in der Dimension identisch mit der Anzahl der Zeilen von X^a), für deren Diagonalelemente $h_{ii}^a \geq 3$ und Ganzzahligkeit gilt. Die Diagonalelemente von H^a geben an, wie viele identische Einheiten in der anonymen Lösung existieren, die durch die zugehörige Zeile in X^a beschrieben werden. Wenn $\Sigma h_{ii}^a = n$ gilt, bleibt die Anzahl der statistischen Einheiten erhalten. Wird für jede Tabellenauswertung $T_j^a=(Z_j^a)' H^a X^a$ berechnet, so ist gewährleistet, dass keine Einzel- und Zweierfälle in Auswertungstabellen ausgewiesen werden können. Die Matrix X^a enthält somit in den Zeilen die metrischen Angaben für die Einheiten, die die Lösung repräsentieren. Die Matrizen Z_j^a ($j=1,\ldots,k$) beschreiben wieder ihre kategorialen Merkmalseigenschaften. Die Häufigkeit der Repräsentanten in der anonymen Lösung enthält die Matrix H^a.[12] Häufigkeitsauszählungen, d. h. die Anzahl der Einheiten (wie z. B. Unternehmen oder Betriebe) in den Merkmalsausprägungen lässt sich durch $A_j^a=(Z_j^a)' H^a Z_j^a$ bestimmen. Das Ergebnis ist die Diagonalmatrix A_j^a mit der Anzahl der Einheiten auf der Hauptdiagonale.

Sollen anonyme Mikrodaten für weitere Untersuchungen bereitgestellt werden, so können für jede Zeile der Matrix X^a die metrischen Werte und die kategorialen Ausprägungen aus den gleichen Zeilen der Matrizen Z_j^a einfach in h_{ii}^a-facher Anzahl als identische Zeilen übernommen werden. Damit sind dann immer mindestens 3 Sätze des anonymen Datenbestandes völlig identisch.

Für die SAFE-Tabellengeheimhaltung besteht die Aufgabe nun darin, X^a, Z_j^a und H^a so zu bestimmen, dass alle für Veröffentlichungen vorgesehenen Häufigkeitstabellen $A_j^a=(Z_j^a)' H^a Z_j^a$ und die Wertetabellen $T_j^a = (Z_j^a)' H^a X^a$ möglichst ähnlich zu den Originalen $A_j^o= (Z_j^o)' Z_j^o$ und $T_j^o=(Z_j^o)' X^o$ sind. Für die Bestimmung der Ähnlichkeit ist allerdings zu berücksichtigen, dass es ggf. Felder in den Auswertungstabellen gibt, die auf Grund von Tabellengeheimhaltungsregeln der Dominanzgeheimhaltung bzw. der Fallzahlgeheimhaltung[13] nicht im Original veröffentlicht werden dürfen. Hier müssen die Elemente von T^o und T^a einen Mindestabstand besitzen, der die Anonymität sicherstellt.

SAFE-Geheimhaltungsansatz

Es sei ein originaler Datenbestand mit seinen metrischen Werten durch X^o und mit seinen kategorialen Ausprägungen durch die Zuordnungsmatrizen Z_j^o ($j=1,2,\ldots,k$; k - Anzahl kategorialer Merkmale und s_j - Spalten der Matrix Z_j^o) gegeben (zur Erläuterung von X^o und Z_j^o siehe vorigen Abschnitt). Zusätzlich existiere eine Menge von t geplanten Auswertungs-

12 Auch für die zuvor beschriebenen Originaldaten kann eine Häufigkeitsmatrix H^o unterstellt werden. Die Formeln zur Berechnung der Auswertungstabellen T_j^o und der Häufigkeitstabellen A_j^o wären dann für den originalen und anonymen Datenbestand identisch. H^o wäre aber eine n-dimensionale Einheitsmatrix, so dass sie bei den Berechnungen vernachlässigt werden kann.

13 Trotz der Sicherstellung einer Mindestanzahl von mindestens 3 identischen Einheiten kann noch eine besondere Form der Fallzahlgeheimhaltung auftreten, wenn als bekannt unterstellt werden muss, dass nur eine Einheit zu einem metrischen Wert beigetragen hat (z. B. nur eine Firma der Branche hatte Exportumsatz). Dann hat auch eine Veröffentlichung der Summenangaben noch ein Reidentifikationsrisiko (siehe Bemerkungen in Abschnitt 1.1).

tabellen T_j *(j=1,2,...,t;* $t \geq k$*)*. Die Anzahl *t* der Auswertungstabellen wird in der Regel größer als die Anzahl der *k* der kategorialen Merkmale sein, weil auch Kombinationen aus mehreren kategorialen Merkmalen auswertbar sind. Für die Erstellung der *t* Auswertungstabellen sind damit die Zuordnungsmatrizen Z_j^o *(j=1,2,...,t)* teilweise bereits gegeben, bzw. lassen sich bei Kombination mehrerer Merkmale oder bei Anwendung von Aggregationsvorschriften aus diesen ermitteln. Die originalen Auswertungstabellen können bei Häufigkeitstabellen als $A_j^o=(Z_j^o)'Z_j^o$ und originale Wertetabellen als $T_j^o=(Z_j^o)'X^o$ berechnet werden. Alle Wertetabellen T_j^o sind auf Probleme der Dominanz- und Fallzahlgeheimhaltung zu testen und auftretende Geheimhaltungsfälle in den Matrizen G_j^o so zu vermerken, dass für die Elemente der Matrizen G_j^o gilt:

$\{G_j^o\}_{il} = 1$; wenn ein Geheimhaltungsproblem in der Tabelle im Tabellenfeld in Zeile i und Spalte l existiert

und

$\{G_j^o\}_{il} = 0$; wenn kein Geheimhaltungsproblem in der Tabelle im Tabellenfeld in Zeile i und Spalte l existiert.

(die Dimension der Matrizen G_j^o ist identisch zu den Wertetabellen T_j^o)

Gesucht sind die Matrizen X^a, Z_j^a und H^a so dass gilt:

Minimiere F_H und F_T

Mit:

$$F_H = \min\left(\max_{j\in(1,t)}\left(\max_{i\in(1,s_j)} \left(\left| a_{j,i,i}^a - a_{j,i,i}^o \right| \right) \right) \right) \qquad \textbf{(3.4 – 1)}$$

$$F_T = \sum_{j=1}^{t}\sum_{i=1}^{s_j}\sum_{l=1}^{m} \left| t_{j,i,l}^a - t_{j,i,l}^o \right| (1 - g_{j,i,l}^o)$$

Unter den Bedingungen:

$$\begin{aligned} A_j^a &= \left(Z_j^a\right)^T H^a Z_j^a \quad ; \quad j = 1,2,...,t \\ A_j^o &= \left(Z_j^o\right)^T Z_j^o \quad ; \quad j = 1,2,...,t \\ T_j^a &= \left(Z_j^a\right)^T H^a X^a \quad ; \quad j = 1,2,...,t \\ T_j^o &= \left(Z_j^o\right)^T X^o \quad ; \quad j = 1,2,...,t \end{aligned} \qquad \textbf{(3.4 – 2)}$$

$$g_{j,i,l}^o * t_{j,i,l}^o * f_{j,i,l} \leq \left| t_{j,i,l}^a - t_{j,i,l}^o \right| \quad \begin{array}{l} ; \quad j = 1,2,...,t \\ ; \quad i = 1,2,...,s_j \\ ; \quad l = 1,2,...,m \end{array}$$

Dabei sind:

$f_{j,i,l}$ – Schranke für eine minimale relative Abweichung im geheimzuhaltenden Tabellenfeld *i, l* in der Tabelle *j*. Diese Schranke könnte auch für alle Tabellen einheitlich festgelegt werden. Das empfiehlt sich jedoch nicht, da das Risiko in den einzelnen Tabellenfeldern unterschiedlich hoch ist.

$t^a_{j,i,l}$ – Element in Zeile i und Spalte l der Wertetabelle T^a_j

$t^o_{j,i,l}$ – Element in Zeile i und Spalte l der Wertetabelle T^o_j

$a^a_{j,i,i}$ – Diagonalelement in Zeile i und Spalte i der Häufigkeitstabelle A^a_j

$a^o_{j,i,i}$ – Diagonalelement in Zeile i und Spalte i der Häufigkeitstabelle A^o_j

$h^a_{i,i}$ – Diagonalelement in Zeile i und Spalte i von H^a mit $h^a_{i,i} = 3,4,5...$ und $\sum h^a_{i,i} = n$

$Z^a_j = \{z^a_{j,i,l}\}$ – Zuordnungsmatrix mit $z^a_{j,i,l} = 0,1$ und $\sum_l z^a_{j,i,l} = 1$

$X^a = \{x^a_{i,l}\}$ – Matrix der anonymen metrischen Werte.

Für die Spalten der Matrix X^a können in Abhängigkeit von ihrer inhaltlichen Bedeutung noch weitere inhaltlich bedingte Nebenbedingungen gesetzt werden, wie z. B. Nichtnegativität, Ganzzahligkeit o. Ä.

Die beiden Zielfunktionen F_H als Fehler in den Häufigkeitstabellierungen und F_T als Fehler in den Wertetabellen sind hierbei als zwei unabhängige Ziele formuliert worden, da für die meisten Analysen sichere Fehlerschranken in den Häufigkeiten als höherwertiges Ziel formuliert werden sollten. Die Minimierung der maximalen Abweichungen hatte sich bei Testrechnungen gegenüber einer Minimierung der durchschnittlichen Abweichung durchgesetzt, da diese einzelne extreme Ausreißer nicht verhindern konnte. Außerdem hatte die Angabe fester Schranken als Maximalfehler bei Datennutzern eine höhere Akzeptanz.

In der Fehlerfunktion der Auswertungen für metrische Werte F_T werden durch die obige Formulierung nur Tabellenfelder berücksichtigt, bei denen keine Geheimhaltungsfälle vorliegen. Es kann ggf. erforderlich sein (abhängig vom Datenbestand), die Fehlerfunktion der metrischen Werte F_T um Gewichtungsfaktoren zu erweitern, mit denen eine unterschiedliche Gewichtung verschiedenartiger metrischer Merkmale (wie z. B. Umsatz- und Beschäftigtenangaben) vorgenommen werden kann. Denkbar wäre auch, vor der Anonymisierung eine Normierung der metrischen Merkmale vorzunehmen, wie sie z. B. von Domingo-Ferrer vorgeschlagen wurde (siehe Domingo-Ferrer und Mateo-Sanz 2002). Eine Normierung erschwert jedoch ggf. die Formulierung zusätzlich erforderlicher Nebenbedingungen, z. B. für den Erhalt der Nichtnegativität.

Die erforderlichen relativen Abweichungen (Fehler) $f_{j,i,l}$ müssen je nach der anzuwendenden Geheimhaltungsregel bestimmt werden. Solche relativen Abweichungen sind auch für die Tabellierungen metrischer Werte bei 1- und 2-Fallzahlproblemen zu bestimmen. Das Modell sichert zwar die Existenz von mindesten drei identischen Einheiten in den Ausprägungskombinationen, ob die Summe der Wertangaben jedoch mit der Angabe des ursprünglichen Fallzahlproblems identisch ist, wird sonst nicht kontrolliert. Einen sehr guten methodischen Ansatz für die Bestimmung der Intervallgrenzen bietet die p %-Regel (zu Tabellengeheimhaltungsregeln siehe Abschnitt 1.1). Aber auch die 1-k-Dominanzregel lässt die Bildung eines solchen Intervalls zu, wenn man aus der Geheimhaltungspflicht von Tabellenwerten mit mehr als k %-Anteil (z. B. bei 80 %) am Gesamtwert den Umkehrschluss zieht, dass der veröffentlichbare Wert um mehr als 100*(100-k)/k vom größten Einzelwert abweicht (bei 80 % z. B. 100*(100-80)/80=25 % Fehler). Der erforderliche Datenfehler aus der 1-k-Dominanzregel ist jedoch bedeutend höher als bei der p %-Regel, was sich bei der zusätzlichen Anwendung auf alle Fallzahlprobleme bemerkbar macht.

Damit sind für die Aufgabe mit den Originaldaten *m*n* vorgegebene metrische Werte sowie *k*n* kategoriale Zuordnungen als Parameter gegeben. Die Klassifizierung der Wertefelder in den Auswertungstabellen als Geheimhaltungsfälle (die Matrizen G_j^o) sind keine Parameter, da sie sich direkt aus den Originaldaten auf der Grundlage der jeweils anzuwendenden Geheimhaltungsregeln ergeben.

Die unbekannte Lösung besteht aus jeweils *m* metrischen Werten sowie *k* kategorialen Zuordnungen für die maximal *n/3* Repräsentanten in der anonymen Lösung (Matrizen X^a und Z_j^a) sowie die maximal *n/3* Häufigkeiten dieser Repräsentanten (Diagonalelemente von H^a). Damit sind gegenüber den Parametern der Aufgabe ca. 1/3 unbekannte Werte zu bestimmen.

Trotzdem ist diese Aufgabe nur schwer exakt lösbar. Das liegt einerseits daran, dass es ein mehrkriterielles Problem ist (F_H und F_T sind zu minimieren). Außerdem ist die Dimension des Problems bei Echtdaten oft sehr groß, weil sowohl viele Einheiten (n) als auch viele metrische (m) und/oder kategoriale (k) Merkmale vorhanden sind. Weiterhin bedingt die Forderung $h_{i,i}=3,4,5,...$, dass die Aufgabe ganzzahlig und wegen der multiplikativen Verknüpfung der Unbekannten in den Formeln für A_j^a und T_j^a nichtlinear ist. Deshalb wurden im Laufe der Zeit verschiedene Ansätze entwickelt, mit denen nur Näherungslösungen gefunden werden können. Weil kein Verfahren die exakte Lösung bestimmen konnte, haben alle Verfahren entsprechende Nach- aber auch Vorteile. Zwei dieser Ansätze sollen in den folgenden Abschnitten näher vorgestellt werden, da sie in statistischen Ämtern noch angewendet werden.

Einen vergleichbaren Ansatz für eine ausschließliche Tabellengeheimhaltung verfolgt Dandekar in der Arbeit „Synthetic Tabular Data – An Alternative to Complementary Cell Suppression“ (Dandekar und Cox 2002). Dort ist die am feinsten gegliederte Auswertungstabelle, d. h. die Auswertungstabelle, die alle kategorialen Merkmale berücksichtigt, Ausgangspunkt der Untersuchungen.

Für die automatisierte Tabellenanonymisierung existiert ein großer Bedarf vor allem für Registerdatenbestände. Registerdatenbestände sind z. B. das Berliner Einwohnerregister (Datenbestand der Einwohnermeldeämter), das Kraftfahrzeugregister (der Zulassungsstellen), das Unternehmensregister u. a. Diese Datenbestände enthalten gegenüber anderen Mikrodaten eine bedeutend größere Anzahl an Objekten. Außerdem verfügen sie über sehr viele kategoriale Merkmale mit entsprechend vielen verschiedenen Ausprägungskombinationen. Die Anzahl der stetigen Merkmale ist dagegen sehr begrenzt, teilweise sind keine vorhanden (z. B. beim Einwohnerregister). Einerseits können bei der Erzeugung von Auswertungstabellen aus Registerdaten umfassende Geheimhaltungsprobleme durch tabellenübergreifende Geheimhaltungsfälle auftreten. Andererseits besteht bei Registerdaten immer ein reges Interesse an Ad hoc- Auswertungen und damit zusätzlichen Tabellen. Viele Nachfragen nach Registerdaten lassen sich nicht durch ein Standardveröffentlichungsprogramm befriedigen und lösen so immer wieder neue Geheimhaltungsprüfungen aus. Ein Schwerpunkt der Methodenentwicklung war deshalb die automatische Lösung dieses Geheimhaltungsproblems.

Ein weiterer Bereich mit großem Bedarf an automatisierter statistischer Geheimhaltung sind Konjunkturstatistiken (z. B. Monatsbericht im Bergbau und verarbeitenden Gewerbe).

Hier entsteht die Notwendigkeit vor allem aus dem regelmäßigen und kurzfristigen Erscheinen der Statistiken (monatlich), die ständige neue Prüfungen der statistischen Geheimhaltung erforderlich macht.

3.4.3 Ein erstes Verfahren

Zu Beginn der Untersuchungen wurde der Datenbestand des „Monatsbericht im Bergbau und Verarbeitenden Gewerbe“ für die Versuche herangezogen. Dieser Datenbestand zeichnet sich in Berlin durch eine relativ geringe Anzahl an Merkmalsträgern aus (ca. 1 000 Firmen). Gleichzeitig bestanden durch die vielen existierenden Schlüsselausprägungen bei den kategorialen Merkmalen (Wirtschaftsklassifikationen WZ, Regionalangaben und Beschäftigtengrößenklassen) sehr viele Geheimhaltungsfälle auf Grund der niedrigen Anzahl an Firmen in einzelnen Ausprägungskombinationen. Das Problem verschärfte sich durch große Firmen, die in eigentlich unkritischen Ausprägungskombinationen Dominanzprobleme auslösten.

Diese Randbedingungen bewirkten, dass der Datenbestand des Monatsberichtes nur mit viel manuellem Aufwand für Veröffentlichungen anonymisiert werden konnte.

Die durch die erste Version des SAFE-Verfahrens bereitgestellte Unterstützung, war eine Automatisierung von Gruppierungen. In Bezug auf die allgemeine Formulierung des SAFE-Geheimhaltungsansatz bestand der Weg zur Lösungsbestimmung ausschließlich darin, dass die Zuordnungsmatrizen Z_j^o manipuliert wurden. Es wurden durch das Verfahren, das in Form von Excel-Makros automatisiert wurde, die kritischen Merkmalsträger bestimmt und die Veränderung seiner kategorialen Merkmale (i.d.R. der Wirtschaftszweig) unterstützt. Der Ansatz besteht somit in einer Gruppierung von kategorialen Merkmalen. Das Verfahren hat den Vorteil, dass die Zusammenfassungen durch diese Gruppierungen dokumentiert werden konnten. Dadurch wurde auch ermöglicht, die Veränderungen in den Folgezeiträumen automatisiert in der gleichen Form vorzunehmen. Diese Zusammenfassungen werden mit den veröffentlichten Tabellen herausgegeben. Außerdem werden sie über möglichst lange Zeiträume konstant durchgeführt, um die Vergleichbarkeit der Werte im Zeitverlauf zu gewährleisten. Geheimhaltungsprobleme auf der Grundlage von Dominanzen in Tabellenwerten wurden ebenfalls durch Gruppierungen gelöst. Waren alle erforderlichen Gruppierungen zum Lösen der Tabellengeheimhaltung vorgenommen, konnten die stetigen Werte bei Nachfrage nach den Mikrodaten über Durchschnittsbildung von mindestens 3 benachbarten Merkmalsträgern (siehe eindimensionale Mikroaggregation Abschnitt 2.2.6) vereinheitlicht werden, ohne die Ergebnisse der Tabellengeheimhaltung zu verändern oder zu gefährden. Von Januar 1994 bis Dezember 2005 wurde der statistische Bericht zum „Monatsbericht für Betriebe des Verarbeitenden Gewerbes“ für Berlin auf dieser Grundlage erstellt. Die Vorgehensweise ist jedoch nicht mit dem traditionellen Verfahren der Gruppierung (siehe Abschnitt 2.1.3) identisch, weil die Zusammenfassung von Kategorien nicht über den gesamten Datenbestand, sondern nur bei kleinen Gruppen von Einheiten erfolgt.

Abbildung 12
Auszug aus den Umbuchungen der Wirtschaftsklassen

Zusammenfassungen und Umbuchungen von Wirtschaftsklassen aufgrund der statistischen Geheimhaltung – Betriebe des Verarbeitenden Gewerbes in Berlin im Januar 2003

WZ 2003	Betriebe vor der Geheimhaltung	Betriebe nach der Geheimhaltung	Umbuchung zur WZ 03	Umbuchung einschließlich WZ 2003			
	1	2	3	4			
11.10	1	–	23.30				
14.21	1	–	23.30				
15.13	26	26					
15.20	1	–	15.51				
15.33	5	5					
15.51	1	4		15.20	15.61		
15.61	2	–	15.51				
15.81	71	71					
15.82	3	–	15.88				
15.84	11	11					
15.86	5	5					
15.87	2	–	15.88				
15.88	2	11		15.82	15.87	15.89	15.98
15.89	1	–	15.88				
15.91	5	5					
15.96	3	3					
15.98	3	–	15.88				

WZ 2003	Betriebe vor der Geheimhaltung	Betriebe nach der Geheimhaltung	Umbuchung zur WZ 03	Umbuchung einschließlich WZ 2003		
	1	2	3	4		
23.30	1	4		11.10	14.21	19.30
24.11	1	–	24.14			
24.14	3	5		24.11	24.20	
24.20	1	–	24.14			
24.30	5	5				
24.41	3	–	24.42			
24.42	20	23		24.41		
24.51	1	–	24.52			
24.52	3	4		24.51		
24.63	1	–	24.64			
24.64	3	7		24.63	24.66	24.70
24.66	2	–	24.64			
24.70	1	–	24.64			
25.12	1	–	25.13			
25.13	6	7		25.12		
25.21	7	7				
25.22	4	4				

Quelle: „Verarbeitendes Gewerbe in Berlin“ Januar 2003, Statistischer Bericht, S. 9

In den statistischen Tabellen des Berichtes werden nur die verbleibenden Wirtschaftsklassen dargestellt. Bei Zusammenfassungen sind entsprechende Kennzeichnungen mit einer Fußnote vorgenommen (siehe folgende Abbildung 13). Als Vergleich ist in Abbildung 14 die Ergebnisdarstellung mit dem klassischen Anonymisierungsverfahren der Tabellenfeldsperrung (Punkte in der Tabelle) dargestellt.

Abbildung 13
Auszug aus der Ergebnisdarstellung nach Wirtschaftsklassen 2003

1.2 Betriebe des Verarbeitenden Gewerbes (sowie Bergbau und Gewinnung von Steinen und Erden) in Berlin im Januar 2003 nach Wirtschaftsklassen

WZ 2003	Unterabschnitt, Abteilung, Klasse	Betriebe	Beschäftigte insgesamt	Beschäftigte darunter Arbeiter	Geleistete Arbeitsstunden	Bruttolohnsumme	Bruttogehaltssumme	Umsatz
					1 000	1 000 EUR		
		1	2	3	4	5	6	7
DA	**Ernährungsgewerbe und Tabakverarbeitung**	**145**	**13 507**	**9 182**	**1 862**	**18 816**	**13 347**	**921 658**
15	Ernährungsgewerbe	141	11 475	7 706	1 601	14 150	11 230	289 168
15.13	Fleischverarbeitung	26	1 456	1 144	213	1 611	681	18 494
15.33	Obst- und Gemüseverarbeitung a.n.g.	5	255	185	32	255	307	7 855 [2]
15.51 [1]	Milchverarbeitung	4	188	136	27	256	178	6 284 [2]
15.81	H.v. Backwaren	71	3 621	2 343	503	3 232	2 317	28 059
15.84	H.v. Süßwaren	11	1 804	1 298	254	2 525	1 855	60 398
15.86	Verarbeitung v. Kaffee und Tee, H.v. Kaffee-Ersatz	5	575	415	82	1 146	541	104 028
15.88 [1]	H.v. homogenisierten und diätetischen Nahrungsmitteln	11	2 739	1 760	366	3 625	3 853	39 015
15.91	H.v. Spirituosen	5	165	95	25	141	219	10 944
15.96	H.v. Bier	3	672	330	99	1 358	1 278	14 091

Quelle: „Verarbeitendes Gewerbe in Berlin“ Januar 2003, Statistischer Bericht, S. 15

Dieses erste Verfahren implizierte allerdings noch viele manuelle Entscheidungen (z. B. Auswahl der Partner-WZ für Zusammenfassungen) beim Neuauftreten von Geheimhaltungsfällen. Deshalb ist es nur bei kleineren Datenbeständen mit einem festen Kreis an Berichtspflichtigen praktikabel. Sowohl regelmäßig wechselnde Berichtspflichtige im Rahmen von Stichprobenerhebungen als auch eine sehr große Anzahl von Berichtspflichtigen machen diesen Ansatz der Geheimhaltung unhandlich und nicht beherrschbar. Für die Analysequalität der Daten ist dieses Verfahren sehr gut, weil alle Veränderungen dokumentiert sind und somit in der Analyse berücksichtigt werden können. Die Auswertung von Teilgesamtheiten, die durch das Verfahren mit anderen gruppiert wurden, ist dabei nicht mehr möglich. Dies stellt aber kein Problem dar, da in diesem Falle das berechtigte Schutzbedürfnis des Einzelnen sowieso die Herausgabe der Daten verhindert hätte. In der aggregierten Darstellung muss sich der Datennutzer die für die Zusammenfassung aus geheimzuhaltenden Gruppen noch verfügbare Information nicht selber durch Differenzbildung herleiten.

Abbildung 14
Auszug aus der Ergebnisdarstellung nach Wirtschaftsklassen 2006

1.2 Betriebe des Verarbeitenden Gewerbes (sowie Bergbau und Gewinnung von Steinen und Erden) in Berlin im Januar 2006 nach Wirtschaftsklassen

WZ 2003	Wirtschaftszweig a = Januar 2006 b= Veränderung zum gleichen Vorjahresmonat in %		Betriebe	Beschäftigte	Geleistete Arbeitsstunden	Bruttoentgelte	Umsatz insgesamt	Umsatz darunter Ausland
			Anzahl		1 000	1 000 EUR		
C	Bergbau u.Gew.v.Steinen u.Erden	a	2	•	•	•	•	•
		b	-33,3	•	•	•	•	•
11.1	Gew.v.Erdöl u.Erdgas	a	1	•	•	•	•	•
		b	–	•	•	•	•	•
14.11	Gew.v.Naturwerksteinen u.Natursteinen ang	a	1	•	•	•	•	•
		b	–	•	•	•	•	•
D	Verarbeitendes Gewerbe	a	795	•	•	•	•	•
		b	-3,1	•	•	•	•	•
DA	Ernährungsgewerbe u. Tabakverarbeitung	a	119	11 846	1 671	29 380	849 639	63 553
		b	-0,8	-3,7	-1,4	-2,7	+7,3	+17,3
15	Ernährungsgewerbe	a	115	9 986	1 413	22 836	•	•
		b	-0,9	-4,4	-2,9	-4,1	•	•
15.13	Fleischverarbeitung	a	19	1 171	166	1 823	16 192	•
		b	-5,0	-7,0	-1,7	-8,0	+17,5	•
15.2	Fischverarbeitung	a	1	•	•	•	•	•
		b	–	•	•	•	•	•
15.33	Obst-u.Gemüseverarbeitung ang	a	3	97	14	238	•	•
		b	–	-20,5	-24,5	-4,1	•	•
15.51	Milchverarbeitung	a	2	•	•	•	•	•
		b	–	•	•	•	•	•
15.61	Mahl-u.Schälmühlen	a	2	•	•	•	•	•
		b	–	•	•	•	•	•
15.81	H.v.Backwaren	a	56	3 300	466	5 185	29 411	•
		b	+3,7	+0,9	+3,9	-0,8	+8,9	•
15.82	H.v.Dauerbackwaren	a	2	•	•	•	•	•
		b	–	•	•	•	•	•
15.84	H.v.Süßwaren	a	11	1 869	267	5 021	62 454	9 742
		b	–	-0,8	-1,5	-6,5	+6,0	+13,9
15.86	Verarb.v.Kaffee u.Tee, H.v.Kaffee-Ersatz	a	5	603	90	1 942	73 421	•
		b	–	–	+9,0	-0,4	-12,5	•
15.87	H.v.Würzmitteln u.Saucen	a	2	•	•	•	•	•
		b	–	•	•	•	•	•
15.88	H.v.homogenisierten u. diät.Nahrungsmitteln	a	1	•	•	•	•	•
		b	-50,0	•	•	•	•	•
15.89	H.v.sonst.Nahrungsmitteln	a	1	•	•	•	•	•
		b	–	•	•	•	•	•
15.91	H.v.Spirituosen	a	5	134	21	305	10 828	•
		b	–	-5,0	-4,7	+8,2	+7,4	•

Quelle: „Verarbeitendes Gewerbe in Berlin“ Januar 2003, Statistischer Bericht, S. 15

3.4.4 Ein automatisiertes Lösungsverfahren

Die hier beschriebene Methode ist das Ergebnis längerer Testreihen und wurde an den Daten des „Monatsberichts für Betriebe“ und der „Jahreserhebung für Betriebe des Verarbeitenden Gewerbes“ für Berlin einer ersten Evaluation unterzogen. Der Teil der kategorialen Geheimhaltung wird außerdem am Einwohnerregister Berlins regelmäßig angewendet. Das Einwohnerregister stellt durch seine Größenordnung (ca. 3,5 Mill. Sätze) eine Herausforderung für die numerischen Verfahren dar. Das Programm ist Teil des DUVA-Programmpaketes, dass vom KOSIS-Verbund für die Städtestatistiker bereitgestellt wird.

Der Algorithmus orientiert sich ständig an einem vorher festgelegten Satz von Häufigkeitstabellen A_j^o und Auswertungstabellen T_j^o, der mit den anonymen Einzeldaten erzeugt werden soll. An diesem Satz von Tabellen werden alle Geheimhaltungsfälle G_j^o markiert und für diese Zellen Unzulässigkeitsbereiche (Schranke *f*) bestimmt. Das komplexe Problem wird dabei in zwei separate Probleme zerlegt, indem zuerst die Funktion F_H und anschließend F_T minimiert wird. Die Anonymisierung erfolgt in folgenden Schritten:

1. Lösung der qualitativen Geheimhaltung
   (nur kategoriale Merkmale werden bearbeitet)
   Bestimmung von Matrizen Z_j^a und H^a zur Minimierung von F_H.
2. Zuordnung der Lösung
   Zuordnung der Zeilen der Matrix X^o zur erhaltenen Lösung der Z_j^a mit dem Ziel des kleinsten Abstandes zwischen den Zeilen von Z_j^a und Z_j^o.
3. Lösung der quantitativen Geheimhaltung
   (nur metrische Merkmale werden bearbeitet)
   Veränderung der Matrix X^o so, dass die Nebenbedingungen zur Geheimhaltung von Tabellenfeldern beachtet werden (zulässige Lösung bestimmen).
4. Optimierung der Lösung
   Veränderung der Matrix X^o zur Minimierung von F_T. Hierbei ist zu beachten, dass die Nebenbedingungen zur Geheimhaltung von Tabellenfeldern weiterhin eingehalten werden.
5. Gruppierung und Durchschnittsbildung

Die numerischen Algorithmen sind in Höhne (2003b) näher beschrieben.

3.4.5 Eigenschaften von SAFE-Lösungen

Das SAFE-Verfahren ist ein Verfahren der Mikroaggregation. Einzelne, sich unterscheidende Datensätze einer Basisdatei werden durch gezielte Auswahl und Gruppenbildung so vereinheitlicht, dass jeder Datensatz in der Basisdatei mit mindestens zwei weiteren Sätzen in der Datei identisch ist.

Der eigentliche Vorteil des SAFE-Verfahrens soll an folgendem Beispiel veranschaulicht werden. In einem Amt aus drei Gemeinden existiere in jeder Gemeinde ein Bäcker, ein Fleischer und Friseur. Jeder Handwerksbetrieb ist allein durch sein Handwerk und durch die Gemeinde eindeutig identifizierbar. Die Veröffentlichung von Angaben wäre somit nicht möglich. Eine SAFE-Lösung besteht nun darin, dass in jeder Gemeinde genau drei Handwerksbetriebe eines Handwerks existieren, deren Wertangaben ebenfalls identisch

sind. Die Aufteilung wird dabei so gewählt, dass die Angaben sowohl bei einer Auswertung nach Branchen als auch nach der Region möglichst dem Original entsprechen. Damit wird einerseits die Information auf der kleinsten Gliederungsebene im Einzeldatenbestand erhalten. Sowohl bei eindimensionalen Auswertungen auf der feinsten Gliederungsebene als auch bei mehrdimensionalen Auswertungen auf aggregierten Ebenen sollten die Auswertungen den Originalauswertungen möglichst ähnlich sein. Eine mehrdimensionale Auswertung auf der feinsten Gliederungsebene, die aus Datenschutzgründen nicht mit dem Original identisch sein darf, liefert dann die sehr unwahrscheinliche und falsche Aufteilung der Angaben auf völlig identische Einheiten. Der Datennutzer wird aber durch das Vergröbern oder Entfernen von kategorialen Merkmalen die Fehler in den Angaben reduzieren können. Dabei ist es für die Nutzung egal, ob entweder auf das Merkmal der Region oder des Wirtschaftszweiges verzichtet wird oder nur eine Vergröberung der Merkmale (z. B. von Gemeinde auf Amt) stattfindet. Eine Vorauswahl der verbleibenden Auswertungsmöglichkeiten, wie sie bei den entsprechenden traditionellen Anonymisierungsverfahren „Variablenunterdrückung“ oder „Gruppierung“ (siehe Abschnitte 2.1.1 und 2.1.3) vorgenommen wird, findet hier nicht statt.

Der entscheidende Vorteil einer solchen anonymen Einzeldatendatei besteht jedoch darin, dass gegenüber den sonst verwendeten Tabellenanonymisierungen die Auswertungen immer zu untereinander konsistenten Ergebnissen führen. Tabellenübergreifende Datenangriffe können höchstens zur Offenlegung der anonymen Einzeldatendatei führen.

Damit ergeben sich für die einzelnen Deanonymisierungsrisiken folgende Sicherheiten:

Fallzahlprobleme können in den Auswertungstabellen nicht mehr auftreten, da mindestens 3 Sätze zu jedem Tabellenwert beitragen. Das bedeutet, dass entweder die in der Realität auftretenden Fallzahlprobleme in den Tabellen entfernt wurden, oder durch die Aggregation die Häufigkeit der Ausprägungskombination auf mindestens 3 erhöht wurde.

Zuordnungsversuche mit Matching-Algorithmen können somit nur zu mehrdeutigen Zuordnungen führen oder fehlschlagen. Wenn ein Satz mehrere Entsprechungen in der anonymisierten Basisdatei hat, kann nicht daraus geschlossen werden, dass die zusätzlich gewonnenen Eigenschaften für das Original gelten, da diese wiederum durch Durchschnittsbildung und ggf. durch Zusammenfassung mit nicht zur Gruppe der kategorialen Ausprägungskombination gehörenden Objekten entstanden sind.

Randsummenprobleme der Tabellengeheimhaltung können zwar theoretisch bei Auswertungstabellen generiert werden, aber auch hier ist es durchaus möglich, dass diese Probleme nur das Ergebnis der Anonymisierungstechnik sind. Die Probleme können entstehen, weil durch das Gruppieren Objekte mit qualitativ verschiedenen Eigenschaften aber sehr geringen Häufigkeiten im Datenbestand zusammengefasst werden. Somit ist kein Rückschluss bei der Auswertung der Basisdatei möglich, mit dem aus Tabellenauswertungen Eigenschaften den Einzelobjekten mit Sicherheit zugeordnet werden können.

Dominanzprobleme der Tabellengeheimhaltung sind die einzigen Geheimhaltungsprobleme, die nicht direkt durch das Verfahrensprinzip der Mikroaggregation gelöst werden können. Deshalb war es erforderlich, bei der Bestimmung der Lösung Unzulässigkeitsbereiche zu definieren (in den Nebenbedingungen (3.4 – 2) durch die Parameter $f_{j,i,l}$ festgelegt), die durch anonymisierte Einzeldaten nicht belegt werden dürfen.

Damit treten in den Auswertungstabellen keine Geheimhaltungsprobleme auf. Selbst wenn einem Datenangreifer die Existenz von Geheimhaltungsfällen der Tabellengeheimhaltung (Fallzahl- oder Dominanzprobleme) bekannt wäre, kann er nicht auf die entsprechenden Tabellenfelder zugreifen, weil sie für einen Datenangriff wegen ihres gesicherten Abstandes zum Originalwert unbrauchbar sind.

Der entscheidende Vorteil aus der SAFE-Anonymisierung der Einzeldaten besteht allerdings darin, dass keine sekundäre Geheimhaltung mehr erforderlich ist. Andere Verfahren der Tabellengeheimhaltung gewährleisten die Sicherheit der Einzelangaben in Auswertungstabellen durch Feldersperrung, d. h. durch Entfernen der den Datenschutz gefährdenden Information. Neben den Primärsperrungen (Tabellenfelder, die direkte Rückschlüsse auf Einzelangaben ermöglichen), sind aber auch Sekundärsperrungen (zusätzlich geheimzuhaltende Tabellenfelder) erforderlich, weil die meisten Tabellen aus untereinander abhängigen Tabellenfeldern bestehen. Zwischensummen und Randsummen in Tabellen bedeuten lineare Abhängigkeiten in Tabellen, so dass auch das Rückrechnen von gesperrten Tabellenfeldern mit Hilfe dieser Abhängigkeiten verhindert werden muss. Die Notwendigkeit von Sekundärsperrungen kann sich auch aus Abhängigkeiten zwischen verschiedenen Auswertungstabellen ergeben. Deshalb kann eine sichere Tabellengeheimhaltung nur in Kenntnis aller bereits veröffentlichten Auswertungen zum Datenbestand erfolgen. Zusätzliche Auswertungstabellen (wie es im Rahmen flexibler Auswertungssysteme gewünscht wäre) erschweren jedoch mit jeder neuen Tabelle das zu lösende Geheimhaltungsproblem. Eine Lösung dieses Problems liegt in anonymisierten Einzeldaten, die für die Tabellengeheimhaltung anonymisiert wurden (siehe Abschnitt „Einleitung und Begriffsbestimmung, S. 12 ff.).

Tabellenauswertungen aus anonymen Einzeldaten ermöglichen immer nur einen Rückschluss auf diese veränderten Einzeldaten. Primärsperrungen und somit auch Sekundärsperrungen sind somit nicht mehr notwendig. Außerdem basieren alle Auswertungen auf dem gleichen Basismaterial (veränderte Einzeldaten). Damit sind die Auswertungen untereinander konsistent, was ein erheblicher Vorteil gegenüber Ansätzen der unabhängigen Tabellenanonymisierung darstellt, die eine Datenveränderung an den einzelnen zu schützenden Auswertungstabellen vorschlagen. Wichtig ist jedoch, dass den Datennutzern die Qualität der Daten bekannt sein muss. Zu erwartende Abweichungen sollten deshalb in allgemeiner Form dokumentiert werden.

Bei mit dem SAFE-Verfahren anonymisierten Daten bedeuten z. B. kleine Häufigkeiten ein hohes Risiko, dass die Daten auch durch Austausch der kategorialen Merkmale verändert wurden. In solchen Fällen kann mit der Bildung größerer Datengruppen (durch Entfernen oder Vergröbern einzelner kategorialer Merkmale) eine stärkere Zusammenfassung erreicht werden, wodurch die Auswertungsqualität steigt.

Bei den verfahrensvergleichenden Untersuchungen in Gnoss et al. (2003, S. 48 ff.) waren einzelne Teilmassenauswertungen bei mit SAFE generierten Lösungen (dort Verfahrensvariante SAFE1A [14]) stark verzerrt. Dieses Problem liegt darin, dass bei stark durch Ein-

14 Die in Gnoss, R. et al. (2003) beschriebene andere SAFE-Variante SAFE2A ist eine Kombination aus der Behandlung der kategorialen Merkmale nach dem obigen SAFE-Algorithmus und einer unabhängigen eindimensionalen Mikroaggregation mit ungefährlichen sicheren Intervallen für die metrischen Einzeldaten (siehe Abschnitt 3.2).

zelunternehmen dominierten Branchen wie Post, Bahn, Bundesbank auch entsprechend starke Veränderungen erforderlich sind, um die Anonymität zu gewährleisten. Bei diesen Umbuchungen kommt es zu dem Effekt, dass die Bereiche, denen diese Unternehmen zugruppiert wurden, ggf. bis zur Unbrauchbarkeit der Angaben verändert werden. Da diese Einheiten vorher oft allein eine WZ-Klasse bildeten, wurden sie anderen WZ-Klassen zugruppiert. Das wäre nur vertretbar, wenn diese Zusammenfassungen auch entsprechend dokumentiert sind, womit die Auswirkungen für den Datennutzer ggf. abschätzbar wären. Das ist jedoch bei den großen Datenbeständen (z. B. Umsatzsteuerdaten) nicht mehr möglich. In den übrigen Auswertungen (z. B. OLS-Schätzungen; siehe Gnoss et al. (2003, S. 64) waren die Ergebnisse mit anderen Mikroaggregationsverfahren durchaus vergleichbar. Bei Datenbeständen, die nicht nur kleine und mittlere Unternehmen enthalten, müsste das automatische SAFE-Verfahren deshalb mit vorherigen manuellen, dokumentierten Gruppierungsentscheidungen für die größten Einheiten starten, was jedoch dem Ziel einer automatischen Anonymisierung widerspräche.

Mit dem SAFE-Verfahren steht ein Verfahren bereit, das es ermöglicht, die grundlegenden Geheimhaltungsansprüche sowohl für aus der Datei erstellten Tabellen als auch für Einzeldaten zu sichern. Gleichzeitig bleibt die flexible Auswertbarkeit der Einzeldaten gewährleistet.

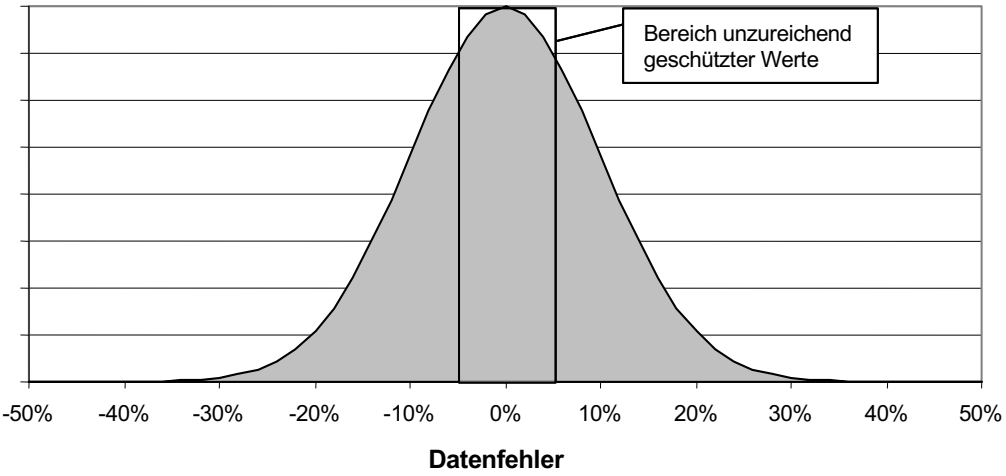
Gerade in Bezug auf die Erstellung von Scientific-Use-Files erweisen sich die Mikroaggregationsverfahren als brauchbar. Hier wären jedoch bei SAFE noch Anpassungen der Zielfunktionen/Teilschritte erforderlich, weil das Verfahren an der Qualität der erzeugbaren Auswertungstabellen optimiert wurde, was teilweise für die Qualitätsansprüche an den Einzeldaten kontraproduktiv ist.

4 Erweiterungen von Verfahren der Zufallsüberlagerung

4.1 Additive Überlagerung mit Mischungsverteilungen (Adaption des Verfahrens von Roque)

Die Idee von Mischungsverteilungen besteht darin, durch die geschickte Kombination von mehreren Verteilungen eine Zufallsüberlagerung zu erreichen, die den Nachteil einer hohen Wahrscheinlichkeitsdichte in der Nähe des benötigten Erwartungswertes der Zufallszahlen (z. B. Null bei additiver Überlagerung) nicht mehr besitzt, die Vorteile einer additiven Zufallsüberlagerung bezüglich der Kontrollierbarkeit der Varianz-Kovarianz-Matrix aber behält. Eine Überlagerung mit einer einfachen Normalverteilung bewirkt eine hohe Wahrscheinlichkeitsdichte im Bereich des Originalwertes, in dem die Werte noch unzureichend geschützt sind (siehe Abbildung 15).

Abbildung 15
Zufallsüberlagerung mit einfacher Normalverteilung



Eine höhere Schutzwirkung könnte durch Erhöhung der Standardabweichung der Überlagerung erreicht werden (siehe folgende Abbildung 16). Es tritt dabei jedoch der Effekt ein, dass nicht nur die unzureichend geschützten Werte verringert werden, sondern sich auch der Anteil der sehr stark veränderten Werte stark erhöht.

Eine Mischung aus mehreren Normalverteilungen kann bei günstiger Parameterkonstellation die geforderten Eigenschaften der Zufallsüberlagerung gewährleisten und gleichzeitig den hohen Anteil unzureichend geschützter Werte reduzieren, ohne den Anteil sehr stark veränderter Werte zu erhöhen (siehe folgende Abbildung 17).

Abbildung 16
Zufallsüberlagerung mit erhöhter Standardabweichung

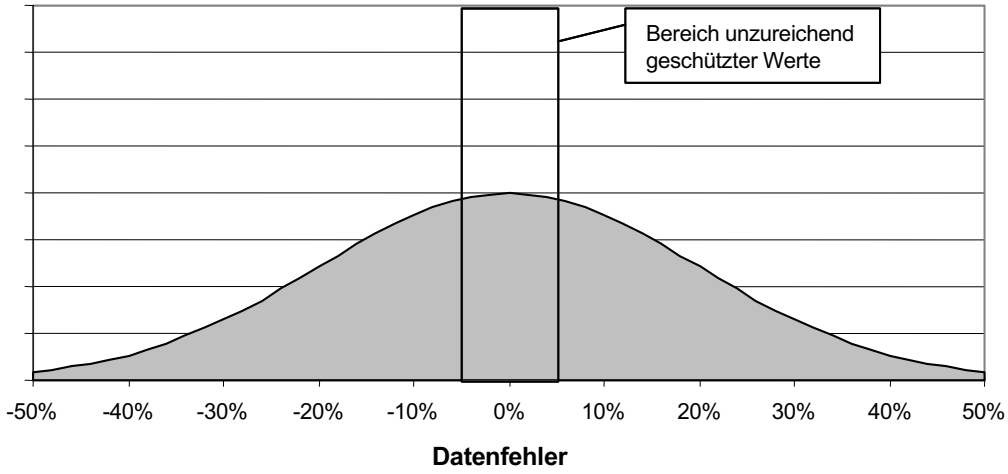
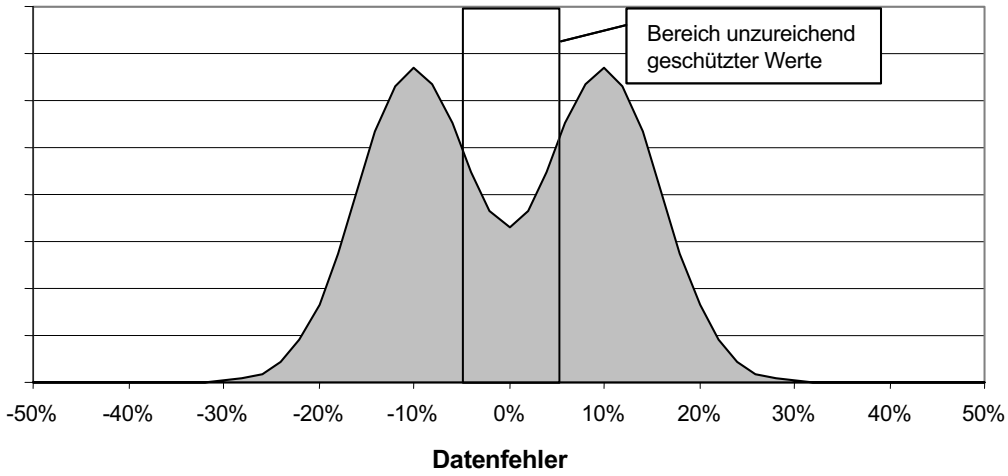


Abbildung 17
Zufallsüberlagerung mit einer Mischungsverteilung



Im Folgenden soll die Verfahrensempfehlung von Roque (Roque 2000), die in den Arbeiten von Yancey und Ronning weiterentwickelt wurde, zusammengefasst dargestellt werden (siehe auch Yancey 2002 und Ronning 2004), um darauf aufbauend die Erweiterungen zu erläutern.

Ziel ist eine additive Zufallsüberlagerung, d. h. die anonyme Wertematrix X^a wird durch

$$X^a = X^o + W$$

erzeugt. Dabei ist *W* eine Matrix aus Zufallszahlen, die die gleiche Dimension wie die Matrix der Originalwerte X^o besitzt. Für die Matrix der Zufallszahlen gilt generell bei additiver Überlagerung:

$E(W) = \underline{0}$; $\underline{0}$ – Nullmatrix der Dimension $n*m$

$\Sigma(W) = d\,\Sigma(X^o)$; d – Parameter zur Regelung der Überlagerungsstärke

Dann gilt auch

$$\Sigma(X^a) = (1+d)\,\Sigma(X^o)$$

mit:

$\Sigma(X^o), \Sigma(W), \Sigma(X^a),$ – Kovarianzmatrizen der Originalwerte, der Überlagerungen und der anonymen Werte

Auf Vorschlag von Roque wird die Matrix der Zufallsüberlagerungen nicht mehr aus einer einfachen Normalverteilung mit $E(W) = 0$ und $\Sigma(W) = d\,\Sigma(X^o)$ (im Folgenden nur als Σ bezeichnet) generiert, sondern aus mehreren einzelnen Normalverteilungen zusammengesetzt, deren Vorteil darin liegt, dass ihre Mittelwerte ungleich Null sind. Damit soll die hohe Wahrscheinlichkeitsdichte der Zufallszahlen in unmittelbarer Nähe von Null verhindert werden. Die Mischung aus den einzelnen Verteilungen ergibt sich in Form der folgenden Dichtefunktion der Zufallsvariablen:

$$f(x;0,\Sigma) = \sum_{j=1}^{k} \omega_j f_j(x;\mu_j,\Sigma_j)$$

mit:

- k – Anzahl der einzelnen Verteilungen in der Mischung
- f_j – Dichtefunktion der Verteilung j
- Σ_j – Varianz- Kovarianzmatrix der Verteilung j
- μ_j – Mittelwertvektor der Verteilung j
- ω_j – Gewicht der Verteilung j in der Mischung
- Σ – Varianz- Kovarianzmatrix der gewünschten Überlagerung mit $\Sigma = d\,\Sigma(X^o)$

Da alle f_i und f Dichtefunktionen mit gegebenen Parametern sind, gelten folgende Abhängigkeiten:

$$\begin{aligned} 1 &= \int f dx && \text{und} && 1 = \int f_j dx \\ \mu &= \int x f dx && \text{und} && \mu_j = \int x f_j dx \\ \Sigma &= \int (x-\mu)(x-\mu)^T f dx && \text{und} && \Sigma_j = \int (x-\mu_j)(x-\mu_j)^T f_j dx \end{aligned} \qquad \textbf{(4.1 – 1)}$$

Aus diesen Eigenschaften für Dichtefunktionen und dem benötigten Erwartungswert von Null und der Varianz $\Sigma = d\Sigma(X^o)$ für die Dichtefunktion f lassen sich folgende Zusammenhänge aus 4.1 – 1 herleiten:

Für die Gewichte der einzelnen Verteilungen gilt:

$$1 = \int f dx = \int \left(\sum_{j=1}^{k} \omega_j f_j \right) dx = \sum_{j=1}^{k} \omega_j \int f_j dx = \sum_{j=1}^{k} \omega_j 1$$
$$\sum_{j=1}^{k} \omega_j = 1 \quad ; mit\ \omega_j \geq 0 \qquad \textbf{(4.1 – 2)}$$

Für den Erwartungswert μ insgesamt gilt:

$$\mu = 0 = \int x f dx = \int \left(\sum_{j=1}^{k} \omega_j x f_j \right) dx = \sum_{j=1}^{k} \omega_j \int x f_j dx = \sum_{j=1}^{k} \omega_j \mu_j$$
$$\sum_{j=1}^{k} \omega_j \mu_j = 0 \quad ; mit\ \omega_j \geq 0 \qquad \textbf{(4.1 – 3)}$$

Für die Kovarianzmatrizen Σ_j der einzelnen Verteilungen gilt:

$$\begin{aligned} \Sigma_j &= \int (x - \mu_j)(x - \mu_j)^T f_j dx \\ &= \int x x^T f_j dx - \left(\int x f_j dx \right) \mu_j{}^T - \mu_j \left(\int x^T f_j dx \right) + \mu_j \mu_j{}^T \\ &= \int x x^T f_j dx - \mu_j \mu_j{}^T \end{aligned} \qquad \textbf{(4.1 – 4)}$$

und somit

$$\int x x^T f_j dx = \Sigma_j + \mu_j \mu_j{}^T$$

Für die Kovarianzmatrix der Mischungsverteilung gilt analog:

$$\begin{aligned} \Sigma &= d\Sigma(X^o) \\ &= \int (x - \mu)(x - \mu)^T f dx \\ &= \int x x^T f dx - \mu\mu^T \end{aligned}$$

und wegen:

$$\begin{aligned} \int x x^T f dx &= \sum_{j=1}^{k} \omega_j \int x x^T f_j dx \\ &= \sum_{j=1}^{k} \omega_j \left(\Sigma_j + \mu_j \mu_j{}^T \right) \end{aligned}$$

gilt auch:

$$\Sigma = \sum_{j=1}^{k} \omega_j \left(\Sigma_j + \mu_j \mu_j{}^T \right) - \mu\mu^T \qquad \textbf{(4.1 – 5)}$$

Aus diesen Beziehungen folgt somit, dass eine nutzbare Mischungsverteilung folgendem System von Gleichungen genügen muss:

$$1 = \sum_{j=1}^{k} \omega_j \quad ; \omega_j \geq 0$$

$$0 = \sum_{j=1}^{k} \omega_j \mu_j \qquad \textbf{(4.1 – 6)}$$

$$\Sigma = \sum_{j=1}^{k} \omega_j \left(\Sigma_j + \mu_j \mu_j{}^T \right)$$

Innerhalb dieses Gleichungssystems ist die gewünschte Überlagerung Σ durch die Varianz-Kovarianzmatrix der Originaldaten $\Sigma(X^o)$ und den Überlagerungsparameter d (d - Stärke der Überlagerung) mit $\Sigma = d\Sigma(X^o)$ bestimmt. Die Anzahl k der verwendeten Verteilungen in der Mischungsverteilung sowie deren Eigenschaften - ω_j (Gewicht in der Mischung), μ_j (Vektor Erwartungswerte) und Σ_j (Varianz- Kovarianzmatrix) sind unter den „Nebenbedingungen“ 4.1 – 6 frei wählbar. Das Gleichungssystem besteht somit aus $1+m+m^2$ Gleichungen, während $(k-1)*(m+1)+k*m^2$ Parameter gesucht sind. Für jede Mischung aus mindestens 2 Komponenten ist das Gleichungssystem völlig unterbestimmt. Deshalb werden in den vorgenannten Arbeiten von Roque, Yancey und Ronning vereinfachende Annahmen unterstellt.

Die gravierendste Annahme ist die Unterstellung:

$$\Sigma_j = g_j \Sigma \qquad \textbf{(4.1 – 7)}$$

Die Matrizen der Varianz-Kovarianz seien in allen Mischungskomponenten proportional zur benötigten Varianz-Kovarianz der gesamten Mischungsverteilung Σ und somit auch direkt zum Datenbestand $\Sigma(X^o)$. Daraus ergibt sich aber direkt:

$$\Sigma = \sum_{j=1}^{k} \omega_j \left(g_j \Sigma + \mu_j \mu_j{}^T \right)$$

$$\left(1 - \sum_{j=1}^{k} \omega_j g_j \right) \Sigma = \sum_{j=1}^{k} \omega_j \left(\mu_j \mu_j{}^T \right) \qquad \textbf{(4.1 – 8)}$$

Mit Hilfe einer Analyse des Ranges der beiden Seiten der Gleichung leitet Roque (Roque 2000) dabei die Notwendigkeit her, dass die Anzahl k der einzelnen Verteilungen in der Mischung mindestens der Anzahl m der Merkmale im Datenbestand entsprechen muss.

Es muss gelten ($m \geq k$). [15]

Weitere Vereinfachungen bei Roque sind die gleiche Größe für alle Gewichte ω_j ($\omega_j = 1/k$), sowie die Identität der g_j für alle j. Damit sind nur noch die Vektoren der Erwartungswerte der Mischungsverteilungen μ_j frei wählbar.

$$k(1-g_1)\Sigma = \sum_{j=1}^{k}\left(\mu_j \mu_j^{T}\right) \qquad \textbf{(4.1 – 9)}$$

Damit hat sich das Problem in ein nichtlineares Gleichungssystem verwandelt. Die Lösung dieses Gleichungssystems ist eine komplexere Aufgabe (Roque 2000, S. 48 ff.), die Roque durch die Anpassung nichtlinearer Optimierungssoftware löste.

Yancey schlägt deshalb zur Lösung des Problems eine andere Vorgehensweise vor (Yancey 2002). Er empfiehlt eine Mischungsverteilung im ersten Schritt als „white noise" Zufallsverteilung W_1 zu erzeugen. Damit gelte:

$E(W_1) = \underline{0}$ $\quad$ $\underline{0}$ – Nullmatrix der Dimension n,m

$\Sigma(W_1) = \underline{I}$ $\quad$ $\underline{I}$ – Einheitsmatrix der Dimension m

Durch die Transformation:

$$W_2 = \sqrt{d} W_1 \left(\Sigma^{1/2}\right)^{T} \qquad \textbf{(4.1 – 10)}$$

(mit $\Sigma^{1/2}$ ist die Choleskimatrix zur Kovarianzmatrix Originaldaten $\Sigma(X^o)$)

gilt:

$$E(W_2) = 0$$
$$\Sigma(W_2) = d\Sigma(X^o)$$

15 Grundlage der Herleitung sind folgende Beziehungen:

Der Rang einer Varianz-Kovarianzmatrix der Originaldaten X^o ist üblicherweise mit der Anzahl der Variablen m identisch $Rang(\Sigma(X^o))=m$. Die Ausnahmefälle von linear abhängigen Variablen oder dass die Anzahl Variablen größer ist als die Anzahl an Objekten seien hier nicht betrachtet. Weiterhin ist der $Rang(\mu_j \mu_j^T)=1$, weil μ_j Vektoren sind und somit den Rang 1 besitzen. Für die Summe von zwei Matrizen gilt allgemein

$$Rang(A+B) \leq Rang(A) + Rang(B).$$

Somit gilt für die beiden Seiten in Formel 4.1 – 9

$$Rang\left[\left(1-\sum_{j=1}^{k}\omega_j g_j\right)\Sigma\right] = Rang[\Sigma] = m$$

$$Rang\left[\sum_{j=1}^{k}\omega_j\left(\mu_j \mu_j^{T}\right)\right] = \min(k,m)$$

Damit ergibt sich direkt $k \geq m$ als notwendige Lösbarkeitsbedingung, weshalb die Anzahl der Mischungskomponenten (k) mindestens der Anzahl an Merkmalen entsprechen muss (m).

Damit ist die Kovarianzmatrix von W_2 wieder proportional zur Kovarianzmatrix der Originaldaten $\Sigma(X^o)$ und der Erwartungswert der Überlagerungen ist Null. Für die „white noise" Zufallsverteilung W_1 können laut Yancey auch Mischungen aus nur 2 Mischungskomponenten verwendet werden. Hier scheint ein Widerspruch zu der Forderung von Roque zu existieren, dass $k \geq m$ gelten muss. Dieser Widerspruch löst sich jedoch darin auf, dass die Choleskimatrix eine Dreiecksmatrix der Dimension m ist. Damit wird durch die Transformation 4.1. – 10 die Zufallsmatrix W_2 als gewichtete Linearkombination der einzelnen Spalten aus W_1 erzeugt. Die in W_2 zu Grunde liegende Verteilung ist damit keine Mischungsverteilung aus zwei Mischungskomponenten, sondern ist abhängig von der betroffenen Spalte eine Linearkombination aus bis zu m Mischungsverteilungen. Das Ergebnis ist somit nicht unbedingt selbst eine Mischungsverteilung, sondern eine Linearkombination aus Zufallszahlen.

Beide Ansätze haben den großen Nachteil, dass die Anzahl der Mischungskomponenten einerseits sehr groß ist ($k \geq m$) und andererseits keine direkte Einflussmöglichkeit auf die Erwartungswerte der Mischungskomponenten besteht. Die Stärke der Nutzung von Mischungsverteilungen besteht aber gerade darin, dass man Mischungskomponenten verwendet, die nicht in der Nähe des Erwartungswertes der Gesamtverteilung von Null liegen. Um dieses Problem zu lösen, wurde getestet, welche Annahmen von Roque und Yancey wirklich zu Vereinfachungen führen. Die Annahme einer zur gesamten Kovarianzmatrix proportionalen Kovarianzmatrix für alle Mischungskomponenten führte zur Gleichung 4.1 – 9:

$$k(1-g_1)\Sigma = \sum_{j=1}^{k}\left(\mu_j \mu_j{}^T\right)$$

Diese Gleichung bedeutet aber direkt, dass es möglich sein muss, aus den Erwartungswerten der Mischungskomponenten (einzige Parameter der rechten Seite) die Kovarianzstruktur der Originaldaten zu erzeugen. Deshalb wurde in Höhne (Höhne 2004a) auf diese Annahme verzichtet und nur die Gleichheit der Kovarianzmatrizen für alle Mischungskomponenten unterstellt. Damit ergibt die Gleichung 4.1 – 8:

$$\Sigma = \sum_{j=1}^{k} \omega_j \left(g_j \Sigma_M + \mu_j \mu_j{}^T\right)$$

$$\sum_{j=1}^{k} \omega_j g_j \Sigma_M = \Sigma - \sum_{j=1}^{k} \omega_j \left(\mu_j \mu_j{}^T\right) \qquad \textbf{(4.1 – 11)}$$

(mit Σ_M als gemeinsame Kovarianzmatrix aller Mischungskomponenten und $\Sigma \neq \Sigma_M$).

Die Anzahl der Mischungskomponenten k ($k \geq 2$) sowie deren Erwartungswerte ist dadurch frei wählbar. Die gemeinsame Kovarianzmatrix der Mischungskomponenten wird dann so bestimmt, dass sie die „Reststreuung" der Zufallswerte erzeugt, die nicht durch die Verschiebung der Erwartungswerte von Null erzeugt wird. Jetzt sind auch vereinfachende Annahmen wählbar, ohne einen Widerspruch bezüglich der Lösbarkeit zu erzeugen. Als vereinfachende Annahmen wurde k=2 und die gleiche Größe für alle Gewichte ω_j ($\omega_j = 1/k = 0{,}5$), sowie die Identität der g_j für alle j gewählt. Aus $k=2$ folgt direkt $\mu_2 = -\mu_1$ (siehe 4.1 – 6). Damit gilt auch $\mu_2 \mu_2{}^T = (-\mu_1)(-\mu_1)^T = \mu_1 \mu_1{}^T$. Somit vereinfacht sich 4.1 – 11 zu:

$$\Sigma_M = \Sigma - \left(\mu_1 \mu_1^T\right) = d\Sigma(X^o) - \left(\mu_1 \mu_1^T\right) \qquad \textbf{(4.1 – 12)}$$

Σ_M – mit der Identität der Proportionalitätsfaktoren g_j für alle j ist es möglich, die Matrix Σ_M gleich um diesen Faktor zu korrigieren, so dass der Faktor g_j in 4.1 – 12 entfallen kann.

Die einzige Restriktion bei der Wahl von μ_1 besteht darin, dass die Verschiebung nur so groß gewählt wird, dass die positive Definitheit von Σ_M nicht gefährdet ist. Nur dann kann eine Zufallsmatrix mit der Varianz-Kovarianzmatrix Σ_M erzeugt werden. Soll eine größere Verschiebung der Mischungskomponenten (μ_1) genutzt werden, muss auch eine größere Kovarianz der Überlagerung Σ insgesamt akzeptiert werden (größerer Parameter d erforderlich).

Algorithmus für die Überlagerung

1. Wahl eines Überlagerungparameters d, der die Stärke der Überlagerung als Anteil an der Varianz-Kovarianz der Originaldaten bestimmt. $\Sigma = d\,\Sigma(X^o)$.

2. Bestimmung des Vektors μ_1. der Erwartungswerte der ersten Mischungsverteilung. Auch für diesen Vektor empfiehlt sich eine vorherige Analyse des Datenbestandes. Die Elemente des Vektors dürfen nicht größer sein, als das d-fache der Standardabweichung der zum Element gehörigen Variable in den Originaldaten, da dann die durch die Verschiebung erzeugte Varianz der Variable bereits größer ist als die zu erzeugende Varianz und somit die zu erzeugende Varianz nicht mehr durch eine zusätzliche Streuung in den Mischungskomponenten generiert werden kann. Es bietet sich deshalb z. B. an, den Vektor μ_1 direkt über einen weiteren Parameter p $(0<p<1)$ aus der Standardabweichung zu bestimmen. $\mu_1 = pdS(X^o)$.

3. Die Wahl der Parameter ist erfolgreich, wenn Σ_M mit

   $$\Sigma_M = d\Sigma(X^o) - \left(\mu_1 \mu_1^T\right)$$

   positiv definit ist. Diese Überprüfung kann auch sehr leicht im Rahmen der Bestimmung der Choleski-Matrix $(\Sigma_M^{1/2})$ mit $(\Sigma_M^{1/2})(\Sigma_M^{1/2})^T = \Sigma_M$) erfolgen.

   Ist die positive Definitheit von Σ_M nicht gegeben, können andere Werte für p oder aber auch ein direkter Eingriff in den Vektor μ_1 erfolgen. [16]

4. Erzeugung von 2 Sätzen von „white noise“ Zufallszahlen mit m-Spalten und $n/2$-Zeilen.

   $E(W_1) = \underline{0}$ $\quad$ $\underline{0}$ – Nullmatrix der Dimension $n/2,m$

   $\Sigma(W_1) = \underline{I}$ $\quad$ $\underline{I}$ – Einheitsmatrix der Dimension m

16 Eigene Berechnungen haben gezeigt, dass es oft dann schwierig ist, passende Parameter zu finden, wenn fast lineare Abhängigkeit zwischen einzelnen Merkmalsspalten auftritt. Hier erweist es sich als günstig, die Streuung der Variablen mit unterschiedlicher Stärke im Vektor μ_1 zu berücksichtigen.

Für die Erzeugung von „white noise“ Zufallszahlen sei hier auf Idee von Brand (2002) verwiesen, mit der das Problem der Qualität vor allem bei kleinem Datenumfang n gelöst werden kann. [17]

5. Transformation der beiden „white noise“ Zufallsmatrizen mit:

$$W_2 = W_1\left(\Sigma_M^{1/2}\right)^T \pm \underline{\iota}\mu_1$$

$\underline{\iota}$ – Spaltenvektor aus lauter Einsen der Dimension n

$\pm\ \underline{\iota}\ \mu_1$ – bedeutet, dass bei der einen Zufallsmatrix W_2 das Produkt $\underline{\iota}\ \mu_1$ addiert und bei der anderen Zufallsmatrix subtrahiert wird.

Damit genügen die Zufallszahlen der beiden Matrizen den Verteilungen

$f_1(x;\ \mu_1,\ \Sigma_M)$ bzw. $f_2(x;\ -\mu_1,\ \Sigma_M)$.

Die Mischungsverteilung wird jetzt dadurch erzeugt, dass die Zeilen der beiden Zufallsmatrizen zufällig gemischt werden.

$$f(x;0,d\Sigma(X^o))=0{,}5\ f_1(x;\ \mu_1,\ \Sigma_M)+0{,}5\ f_2(x;\ -\mu_1,\ \Sigma_M)$$

Die so generierte Mischungsverteilung ist für alle Variablenspalten eine echte zweigipfelige Mischungsverteilung und gewährleistet gleichzeitig, dass sie im Erwartungswert der Überlagerungen Null ist und in der Varianz-Kovarianzmatrix der geforderten Struktur Σ entspricht.

4.2 Multiplikative Überlagerung mit Mischungsverteilungen

Additive Zufallsüberlagerungen haben einen Nachteil, dass die Größe der Überlagerung unabhängig von der Größe der Merkmalswerte bestimmt wird. Die Verteilung der Zufallszahl hängt in ihrer Varianz von der Originalvarianz des Merkmals ab. Die Merkmale selbst sind jedoch in der Regel sehr schief verteilt. [18] Damit werden die sehr vielen kleinen Merk-

17 Da die erzeugten Zufallszahlen W nur Realisationen der Zufallsverteilung sind, kommt es vor, dass die erhaltenen Mittelwerte und Varianz-Kovarianzmatrizen unterschiedlich stark von der theoretischen Größe 0 (beim Vektor der Mittelwerte) bzw. I (Varianz-Kovarianz-Matrix) abweichen. Dann wird von Brand, R. (2002) und *Yancey, W. E. (2002)* folgende Korrektur empfohlen:

$$W' = \left(W - \overline{W}\right)\left(\left(\Sigma_W^{1/2}\right)^{-1}\right)^T$$

mit :

W – Matrix der Zufallszahlen

$\left(W - \overline{W}\right)$ – bedeutet, dass in jeder Zeile von W der Mittelwertvektor $\overline{W}$ subtrahiert wird

$\Sigma_W^{1/2}$ – Ist die Choleskimatrix der Varianz – Kovarianzmatrix von W

Die Zufallszahlen gewährleisten dann exakt, dass $\overline{W'} = 0$ und $\Sigma_{W'} = I$ gilt.

18 Dieser Effekt tritt vor allem bei wirtschaftsstatistischen Daten auf, wenn die Angaben von sehr großen Unternehmen mit denen kleiner Unternehmen in einer Statistik zusammen erhoben werden. Bei gesamtwirtschaftlichen Analysen müssen die Daten jedoch zusammen anonymisiert werden, da eine Beschränkung z. B. auf die kleinen Einheiten die Ergebnisse stark verzerren würde.

malswerte im Datenbestand mit der gleichen Varianz der Zufallszahlen überlagert, wie die wenigen großen Merkmalswerte. Sollen auch die großen Merkmalswerte einen ausreichenden Schutz erhalten, ist die Varianz sehr groß zu wählen. Dadurch müssen die kleinen Merkmalswerte jedoch sehr stark verändert werden und es besteht das Risiko, dass sie auch negativ werden können. Negative Werte sind einerseits für viele Merkmale inhaltlich sinnlos als auch für bestimmte ökonometrische Modelle störend, weil z. B. keine Logarithmen berechnet werden können.

Einen Ausweg bieten deshalb multiplikative Überlagerungen.

4.2.1 Verfahrensbeschreibung

Bei Multiplikation eines Originalwertes mit einer Zufallszahl ist die erzeugte Veränderung in den Daten sowohl von der Zufallszahl als auch von der Größe der Originalwerte abhängig. Wird für jeden Originalwert eine eigene Zufallszahl generiert, berechnen sich die anonymen Werte als:

$$X^a = X^o \odot W \qquad \textbf{(4.2 – 1)}$$

$\odot$ – Hadarmardprodukt für die elementweise Multiplikation der Matrizen X^o und W

Dabei werden in W positive Zufallszahlen mit einem Erwartungswert von 1 verwendet. Positive Zufallswerte sichern den Erhalt der Vorzeichen der Merkmalswerte. Gleichzeitig bleiben Nullwerte erhalten, unabhängig davon, ob sie strukturell bedingt sind oder nur bei der speziellen Datenkonstellation auftreten. Während der Erhalt struktureller Nullwerte die Datenqualität erhöht, führt der sichere Erhalt sonstiger Nullwerte dazu, dass das Deanonymisierungsrisiko steigt. Ist z. B. bekannt, dass ein gesuchtes Unternehmen die Besonderheit besitzt, als einziges Unternehmen in der Region und/oder Branche nicht zu exportieren, dann schränkt der Erhalt der Nullen bei den Außenhandelswerten (wie z. B. Außenhandelsumsatz) die möglichen Zuordnungen stark ein. Diese erhöhten Zuordnungsrisiken müssen bei der Verwendung der multiplikativen Überlagerung mit berücksichtigt werden.

Der Erhalt struktureller Nullwerte, d. h. Nullwerte, die bei bestimmten Merkmalskombinationen inhaltlich bedingt auftreten müssen (z. B. Umsätze aus eigener Handelstätigkeit bei Unternehmen im Produzierenden Gewerbe ohne eigene Handelstätigkeit), erhöhen das Deanonymisierungsrisiko nicht, da sie bei allen Unternehmen mit diesen Merkmalskombinationen analog bedingt sind. Ist diese Merkmalskombination (Unternehmen im Produzierenden Gewerbe ohne eigene Handelstätigkeit) selbst kein Sicherheitsrisiko, weil sie häufig genug auftritt, erzeugt auch der erhaltene Nullwert kein höheres Risiko. Für die Plausibilität anonymisierter Daten ist der Erhalt dieser Nullwerte aber sehr hilfreich.

4.2.2 Bestimmung der Überlagerungsparameter

Generell sollte man die Zufallswerte für die multiplikatvie Überlagerung so wählen, dass gilt:

$E(W) = \underline{1}$ $\qquad$ $\underline{1}$ – Matrix aus lauter Einsen der Dimension n,m

Damit gilt wegen der Unabhängigkeit der Zufallswerte von den Originalwerten auch:

$$E(X^a) = E(X^o \odot W) = E(X^o) \odot E(W) = E(X^o) \qquad \textbf{(4.2 – 2)}$$

Die Erwartungswerte der Originaldaten werden somit erhalten.

Vom Aspekt der Steuerung der Schutzwirkung bietet die multiplikative Überlagerung mehr Möglichkeiten als die additive Überlagerung. Wie in 4.2 – 1 bereits festgestellt wurde, hat der Erhalt aller Nullwerte im Datenbestand für die Datensicherheit nicht nur Vorteile. Anders verhält es sich mit der Brauchbarkeit der übrigen Werte. Mit den Eigenschaften der Zufallswerte kann der relative Fehler in den Daten direkt kontrolliert werden. Damit bietet multiplikative Überlagerung auch die Möglichkeit der Einflussnahme auf die Brauchbarkeit der Daten. Wird die Brauchbarkeit der Daten für Datenangreifer dadurch definiert, wie viele Daten eindeutig zugeordnet werden können und innerhalb einer bestimmten Fehlerschranke liegen (siehe Lenz et al. 2004), so kann auch die Zufallsverteilung optimal darauf abgestimmt werden. In dem Vorschlag wurde unterstellt, dass ein risikoaverser Datenangreifer dann von unbrauchbaren Mikrodaten für seine Datenangriffe ausgeht, wenn es ihm nicht gelingt, bei seinen Zuordnungen mit einer hinreichenden Wahrscheinlichkeit p brauchbare Daten zuzuordnen. Die Brauchbarkeit seiner zugeordneten Daten wird dadurch bestimmt, dass der Fehler in den Daten unterhalb einer Nutzbarkeitsgrenze f liegt.

Beispiel

Es werde eine Zuordnungswahrscheinlichkeit p von 90 % und eine Nutzbarkeitsgrenze f von 5 % unterstellt. Damit ist für einen Datenangreifer ein Massenfischzug nur dann lukrativ, wenn es ihm gelingt, mindestens 90 % der Einzelwerte mit einem Fehler von maximal 5% den Unternehmen zuzuordnen.

(Einzelangriffe bedürfen einer gesonderten Betrachtung, weil hier geprüft werden muss, inwieweit erstens die richtige Zuordnung und zweitens der Fehler in den Daten noch stochastischen Einflüssen unterliegt. Das hängt von den einzelnen Anonymisierungsmaßnahmen vor allem bei den kategorialen Merkmalen ab.)

Für die Zufallsverteilung der multiplikativen Überlagerung stellen die Parameter p und f somit Vorgaben dar, die direkt bei der Wahl der Verteilung berücksichtigt werden können.

Soll die Überlagerung z. B. mit einer einfachen Normalverteilung der Form W~N(1,s) erfolgen und unabhängig von den Erfolgen einer Zuordnung der kategorialen Merkmale allein durch die fehlende Brauchbarkeit der anonymisierten Werte einen Schutz für die Daten bieten, so gilt für den Parameter s:

$$F(x) = \frac{1}{\sqrt{2\pi}s} \int_{-\infty}^{1-f} e^{-\frac{(t-1)^2}{2s^2}} dt > \frac{1-p}{2}$$

(Da eine Normalverteilung symmetrisch ist, ist der Parameter s so zu wählen, dass ein Anteil von mehr als *(1-p)/2* der Werte um mehr als *f* nach unten abweichen. Wird diese Ungleichung nach *s* aufgelöst, lässt sich für beliebige Konstellationen der Parameter *p* und *f* die notwendige Standardabweichung bestimmen. Für ausgewählte Parameterkonstellationen ist die notwendige Standardabweichung *s* in der folgenden Tabelle 2 dargestellt.)

Tabelle 2: Minimale Standardabweichung der multiplikativen Überlagerung (normalverteilt) zu Erreichung von Anonymität

Erwarteter Anteil nutzbarer Werte (p)	Wertangaben brauchbar mit Fehlern unter . . . (f)		
	2 %	5 %	10 %
50%	0,021	0,052	0,104
80%	0,012	0,030	0,060
90%	0,010	0,025	0,049

Eine weitere einfache Variante besteht darin, gestutzte Verteilungen zu verwenden. Dabei werden für die gewählte Verteilung zusätzlich Ausschlussbereiche definiert und für den Fall, dass eine Zufallszahl aus dem Ausschlussbereich gezogen wird, wird das Ziehen der Zufallszahl wiederholt. Diesen Ansatz findet man z. B. bei Kim und Winkler (2003) und Gottschalk (2004). Während Kim und Winkler eine normalverteilte Zufallszahl ($w \sim N(1;0,15)$) verwenden und sie auf den Bereich $0,01 \leq |w_j - 1| \leq 0,6$ beschränken, verwendet Gottschalk eine Gleichverteilung im Bereich $0,5 \leq w_j \leq 1,5$. Mit beiden Beschränkungen wird sowohl eine geringe Häufigkeit in unmittelbarer Nähe der Originalwerte als auch gleichzeitig die Nichtnegativität zum Erhalt der Vorzeichen gewährleistet. Die Wahl von 0,01 bei Kim/Winkler erscheint für die oben beschriebenen Vorstellungen vom Sicherheitsparameter f ($f=0,05$) relativ klein, da so noch sehr viele brauchbare Zufallszahlen generiert werden. Der Wert von 0,01 sichert nur eine Mindestveränderung der Daten von 1 %. Für einen Datenangreifer nützliche Werte (z. B. bei Datenfehler kleiner als $f=0,05$) sind aber immer noch 30,3 % des Datenbestandes (bei $s=0,1$). Dafür wird aber versucht, die Zufallszahlen im Schwerpunkt möglichst nahe an 1 zu erzeugen. Die Parameter bei Gottschalk erscheinen dagegen sehr groß, was jedoch darin begründet liegt, dass im Ergebnis auch nur simulierte Daten erzeugt werden sollen.

Beide Varianten der multiplikativen Zufallsüberlagerung könnten durch die Wahl der Überlagerungsparameter in ihrer Wirkung so optimiert werden, dass die generierten Zufallszahlen nur in kleinem Umfang stark veränderte Zufallszahlen (für die Analyse unbrauchbare Werte) bzw. kaum veränderte Zufallszahlen (unsichere Werte im Datenbestand) erzeugen. Trotzdem soll im Folgenden ein weiteres Modell vorgeschlagen werden, weil beide Modelle den Nachteil besitzen, dass sie von einer unabhängigen Generierung der Zufallszahlen für jeden Datenwert ausgehen. Besser wäre es jedoch, wenn man die Richtung der multiplikativen Überlagerung für die einzelnen Einheiten konstant halten könnte. Dadurch wird der Erhalt der Abhängigkeiten zwischen den Merkmalen besser gewährleistet. Die Idee, die Abhängigkeiten zwischen den Merkmalen zu erhalten, kann durch folgendes Modell verfolgt werden (siehe z. B. Rosemann 2006):

$$X^a = WX^o \qquad \textbf{(4.2 – 3)}$$

W – Diagonalmatrix mit normalverteilten Zufallszahlen mit $E(W)=I$ (I-Einheitsmatrix)

Dieses Modell hat jedoch einen entscheidenden Nachteil. Da jeder Merkmalswert einer Einheit mit dem gleichen Zufallswert verändert wird, gilt:

$$\begin{aligned} x_{ij}^{a} &= w_{ii} x_{ij}^{o} \quad ; i = 1,...,n \wedge j = 1,...,m \\ x_{ij}^{o} &= x_{ij}^{a} (x_{il}^{o} / x_{il}^{a}) \quad ; i = 1,...,n \wedge j = 1,...,m \wedge j \neq l \end{aligned} \qquad \textbf{(4.2 – 4)}$$

Für jede einmal eindeutig zugeordnete Einheit i ist es bei Kenntnis von nur einem originalen Wert (Merkmal l) möglich, alle anderen m-1 numerischen Merkmale exakt mit der Gleichung 4.2 – 4 wieder herzuleiten. Damit wird die Erfolgsquote für einen Datenangreifer stark vergrößert, da nur noch eine erfolgreiche Zuordnung und ein Originalwert (z. B. eine Angabe aus einer Selbstveröffentlichung des Unternehmens) für die vollständige Offenlegung eines Unternehmens erforderlich ist. Deshalb soll dieser Ansatz hier nicht weiter betrachtet werden. Das Modell hätte aber auch den entscheidenden Vorteil:

$$\frac{x_{ij}^{a}}{x_{il}^{a}} = \frac{x_{ij}^{o}}{x_{il}^{o}} \quad ; i = 1,...,n \wedge j,l = 1,...,m$$

Die Verhältnisse zwischen den numerischen Merkmalswerten werden originalgetreu erhalten. Modelle, die ausschließlich Abhängigkeiten zwischen Verhältniswerten untersuchen, sind somit fehlerfrei analysierbar.

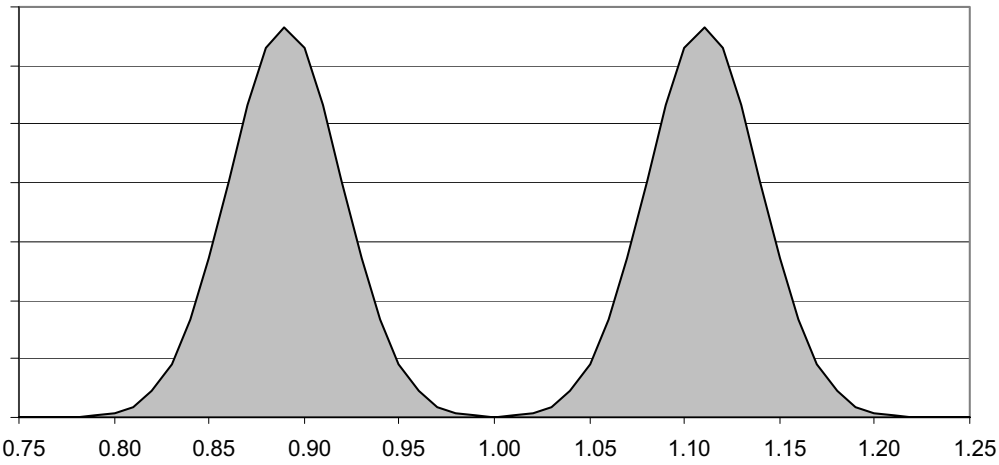
Um beide Eigenschaften (sichere Anonymisierungswirkung und ungefährer, aber nicht exakter Erhalt der Verhältnisse zwischen den Merkmalen) kombinieren zu können, bietet sich folgendes Modell an:

$$\begin{aligned} x_{ij}^{a} &= w_{ij} x_{ij}^{o} = (1 + f w_{i}^{\#} + w_{ij}^{*}) x_{ij}^{o} \quad ; i = 1,...,n \wedge j = 1,...,m \\ w_{i}^{\#} &\in (1;-1) \quad ; P(w_{i}^{\#} = 1) = P(w_{i}^{\#} = -1) = 0{,}5 \\ w_{ij}^{*} &\sim N(0, s) \end{aligned} \qquad \textbf{(4.2 – 5)}$$

Damit wird zwar für jeden Datenwert ein zufälliger Überlagerungsfaktor erzeugt, die Überlagerungsfaktoren eines Unternehmens erzeugen aber alle die gleiche Richtung der Datenveränderung (Vergrößern bzw. Verkleinern die Einzelangaben). Die Anonymisierungswirkung wird bei dem Modell 4.2 – 5 durch die beiden Parameter f und s bestimmt. Der Parameter f kennzeichnet dabei die Niveauverschiebung der Werte eines Unternehmens. Die Standardabweichung s der Normalverteilung beschreibt die Stärke einer zwischen den Werten unabhängigen zusätzlichen Zufallsüberlagerung.

Analysiert man die Verteilung der Überlagerungsfaktoren w_{ij}, so zeigt sich, dass diese wieder einer Mischungsverteilung entsprechen. Die Verteilung der Überlagerungsfaktoren entspricht einer Mischung aus zwei Normalverteilungen (siehe folgende Abbildung 18).

Abbildung 18
Mischungsverteilung des Überlagerungsmodells 4.2 – 5 mit f=0,11 und s=0,03



Der Vorteil dieser Art der Ziehung der Zufallszahlen besteht allerdings darin, dass die Ziehung für die Werte der einzelnen Einheiten nicht völlig unabhängig aus der Gesamtverteilung erfolgt. Die Abhängigkeiten der Merkmale untereinander werden dadurch besser erhalten, als bei einer völlig unabhängigen Überlagerung der Einzelwerte. Werden bei der Analyse die Verhältnisse zwischen zwei Merkmalen näher untersucht (x_{ij}/x_{ik} ; $j \neq k$), so gilt:

$$\frac{x_{ij}^a}{x_{ik}^a} = \frac{w_{ij} x_{ij}^o}{w_{ik} x_{ik}^o} = \frac{(1 + f w_i^{\#} + w_{ij}^*) x_{ij}^o}{(1 + f w_i^{\#} + w_{ik}^*) x_{ik}^o} = \frac{x_{ij}^o}{x_{ik}^o} + \frac{(w_{ij}^* - w_{ik}^*)}{(1 + f w_i^{\#} + w_{ik}^*)} \frac{x_{ij}^o}{x_{ik}^o}$$

$$; i = 1,\ldots,n \wedge j = 1,\ldots,m \wedge j \neq k$$

wegen :

$w_{ij}^* \sim N(0, s)$

gilt :

$E(w_{ij}^*) = E(w_{ik}^*) = 0$

$E(w_{ij}^* - w_{ik}^*) = 0$

$E(1 + f w_i^{\#} + w_{ik}^*) = 1$

$$E\left(\frac{x_{ij}^a}{x_{ik}^a}\right) = E\left(\frac{x_{ij}^o}{x_{ik}^o}\right)\left(1 + E\left(\frac{w_{ij}^*}{1 + f w_i^{\#} + w_{ik}^*}\right) - E\left(\frac{w_{ik}^*}{1 + f w_i^{\#} + w_{ik}^*}\right)\right)$$

Wegen der Unabhängigkeit der Zufallszahlen w_{ij}^{*} *und* $w_i^{\#}$ gilt:

$$E\left(\frac{w_{ij}^{*}}{1+fw_i^{\#}+w_{ik}^{*}}\right)=E\left(w_{ij}^{*}\right)E\left(\frac{1}{1+fw_i^{\#}+w_{ik}^{*}}\right)=0$$

Wegen w_{ik}^{*} sind Zähler und Nenner im zweiten Ausdruck nicht unabhängig. Für diesen Fall ist folgende Approximation möglich:

$$E\left(\frac{X}{Y}\right)\approx \underbrace{\frac{E(X)}{E(Y)}}_{\text{wahrer Anteil}}+\underbrace{\frac{E(X)}{(E(Y))^3}V(Y)-\frac{1}{(E(Y))^2}\operatorname{cov}(X,Y)}_{\text{bias}} \qquad \textbf{(4.2 – 6)}$$

(siehe z. B. Ronning 2008).

Damit ergibt sich

$$E\left(\frac{w_{ik}^{*}}{1+fw_i^{\#}+w_{ik}^{*}}\right)\approx\frac{E(w_{ik}^{*})}{E(1+fw_i^{\#}+w_{ik}^{*})}+\frac{E(w_{ik}^{*})\,V(1+fw_i^{\#}+w_{ik}^{*})}{\left(E(1+fw_i^{\#}+w_{ik}^{*})\right)^3}-\frac{\operatorname{cov}(w_{ik}^{*},(1+fw_i^{\#}+w_{ik}^{*}))}{\left(E(1+fw_i^{\#}+w_{ik}^{*})\right)^2}$$

Für die beiden Möglichkeiten der zweiwertigen Zufallszahl $w_i^{\#} \in (1;\text{-}1)$ ergeben sich deshalb die beiden Approximationen:

$$E\left(\frac{w_{ik}^{*}}{1+fw_i^{\#}+w_{ik}^{*}}\middle| w_i^{\#}=-1\right)\approx-\frac{s^2}{(1-f)^2}$$

und

$$E\left(\frac{w_{ik}^{*}}{1+fw_i^{\#}+w_{ik}^{*}}\middle| w_i^{\#}=1\right)\approx-\frac{s^2}{(1+f)^2}$$

Da beide Realisationen der Zufallszahl $w_i^{\#}$ mit der Wahrscheinlichkeit 0.5 auftreten, gilt für den Erwartungswert des Quotienten die Näherung:

$$E\left(\frac{x_{ij}^{a}}{x_{ik}^{a}}\right)=E\left(\frac{x_{ij}^{o}}{x_{ik}^{o}}\right)\left(1-E\left(\frac{w_{ik}^{*}}{1+fw_i^{\#}+w_{ik}^{*}}\right)\right)$$

$$\approx E\left(\frac{x_{ij}^{o}}{x_{ik}^{o}}\right)\left(1+0.5\left(\frac{s^2}{(1+f)^2}+\frac{s^2}{(1-f)^2}\right)\right)$$

Damit werden die Quotienten zweier Variablen systematisch überschätzt. Die Größe der Verzerrung hängt aber von den beiden Parametern der Mischungsverteilung ab. Für eine realistische Kombination von $f=0.11$ und $s=0.03$ beträgt sie z. B.:

$$E\left(\frac{x_{ij}^{a}}{x_{ik}^{a}}\right)\approx E\left(\frac{x_{ij}^{o}}{x_{ik}^{o}}\right)\left(1+0.5\left(\frac{0.03^2}{(1+0.11)^2}+\frac{0.03^2}{(1-0.11)^2}\right)\right)=1{,}00093\ *\ E\left(\frac{x_{ij}^{o}}{x_{ik}^{o}}\right)$$

Der Erwartungswert der Quotienten überschätzt das Ergebnis somit um weniger als ein Promille und sollte sich somit in einer noch vertretbaren Größenordung befinden.

4.2.3 Varianzprobleme

Bei der Anwendung der multiplikativen Zufallsüberlagerung auf Wirtschaftsstatistiken (Kostenstrukturerhebung) zeigte sich eine Besonderheit, die diese Überlagerungsart gegenüber anderen Verfahren benachteiligte. Bei Testrechnungen fielen die besonders hohen Abweichungen in den Varianzen auf, die mit den anonymisierten Daten reproduziert wurden. Dass multiplikative Zufallsüberlagerung zu systematischen Fehlern in den Varianzen führen, wurde bereits in Ronning (2005) gezeigt. Leider brachten auch die dort hergeleiteten Korrekturformeln keine grundlegende Verbesserung in den Varianzschätzungen, obwohl mit relativ großen Datenbeständen gearbeitet wurde. Deshalb soll die Ursache hier näher untersucht werden.

Es wurde außerdem festgestellt, dass die Abweichungen sehr empfindlich auf die Initialisierung der Zufallszahlgeneratoren reagierten. Zufallszahlen sollen ja einerseits das Ergebnis von stochastischen Ereignissen sein. Zufallszahlgeneratoren sind allerdings Funktionen, mit denen man aus einer beliebigen Zahl eine neue Zahl eindeutig ermitteln kann. Die so erhaltene Reihe von Zahlen entspricht zwar der gewünschten Verteilung von Zufallszahlen, ist jedoch nicht „richtig“ zufällig entstanden. Damit hat man die Möglichkeit, den Startwert eines Zufallszahlengenerators zu beeinflussen. Man kann den Startwert aus externen Angaben (z. B. interne Rechnerzeit) „quasi zufällig“ automatisch genieren lassen oder direkt vorgeben. Das direkte Vorgeben ist vor allem dann erforderlich, wenn man bestimmte Einzelergebnisse nachverfolgen und deshalb genau reproduzieren will.

Da die Auswirkungen bei allen Merkmalen auftraten, bei großen Merkmalswerten jedoch intensiver nachweisbar waren, konzentrieren sich die Untersuchungen auf ein großes Merkmal des Datenbestandes. Ausgewählt wurde das Merkmal Umsatz aus der Kostenstrukturerhebung. Grundlage der Untersuchungen ist die Anonymisierung der Daten über folgendes Modell:

> Es erfolgt eine multiplikative Überlagerung mit einer Mischungsverteilung M. Die Mischungsverteilung setzt sich aus zwei Normalverteilungen (N_1 und N_2) zusammen, die jeweils um 11 % von 1 abweichen und eine Standardabweichung von 3 % besitzen.
>
> $$M \sim 0{,}5*N_1(1{,}11;0{,}03) + 0{,}5*N_2(0{,}89;0{,}03)$$
>
> Das Überlagerungsmodell entspricht (4.2 – 5) bei eindimensionaler Anwendung mit den speziellen Parametern f=0,11 und s=0,03 (siehe auch Abbildung 18).

Die Anwendung dieser Mischungsverteilung entspricht den üblichen Parametern, mit denen Anonymität dadurch erreicht wird, dass der größte Teil der Angaben um mehr als 10 % vom Original abweicht.

Für die Mischungsverteilung gilt dann (siehe Yancey 2002):

$$E(M) = \sum_{i=1}^{2} 0,5 * E(N_i) = 0,5 * 0,89 + 0,5 * 1,11 = 1$$

$$V(M) = \sum_{i=1}^{2} \left(0,5 * \left(V(N_i) + E(N_i)^2\right)\right) - E(M)^2$$

$$= 0,5 * \left(0,03^2 + 0,89^2\right) + 0,5 * \left(0,03^2 + 1,11^2\right) - 1^2$$

$$= 0,013$$

$$S(M) = \sqrt{V(M)} = 0,114$$

Die Standardabweichung der Mischungsverteilung beträgt somit 0,114 (11,4 %).

Hypothese

Als Ursache für diese Effekte wird vermutet, dass wirtschaftsstatische Datenbestände in der Regel sehr schief verteilt sind, wenn sie die Gesamtwirtschaft umfassen und keine Einschränkung, z. B. auf kleine und mittlere Unternehmen, existiert. Bei Stichprobenerhebungen sind Großunternehmen dazu noch überrepräsentiert (oft in Totalschichten), während kleinere und mittlere Unternehmen nur mit bedeutend kleineren Anteilen erhoben werden. Damit sind im Datenbestand wenige extrem große Einheiten vorhanden, die natürlich die Werte der einzelnen Momente der Verteilung der Daten (Mittelwert, Varianz, Schiefe und Exzess) sehr stark beeinflussen. Auf Grund der geringen Anzahl dieser Einheiten ist es jedoch möglich, dass die zufällige Veränderung dieser Werte auch für mehrere große Einheiten in die gleiche Richtung gerichtet ist. Dadurch wird der Effekt der Verzerrung der Momente nicht kompensiert, sondern bleibt nach der Anonymisierung der großen Einheiten erhalten und kann auch durch eine große Anzahl an ebenfalls zufällig überlagerten kleinen Einheiten nicht mehr kompensiert werden.

Test

Um diese Hypothese zu testen wurden Monte-Carlo-Simulationen durchgeführt. Dazu wurden 10 000 verschiedene Startwerte für Zufallszahlenfolgen gleichverteilt aus der Menge aller möglichen Startwerte (ca. 32 000) generiert. Damit wurde die Anonymisierung mit 10 000 verschiedenen Zufallszahlenfolgen durchgeführt. Die Variation der Ergebnisse ist in der folgenden Tabelle 3 dargestellt. Die relativen Abweichungen sind immer als Abweichungen zu den bei den Simulationen nicht veränderten Originalwerten berechnet. Anschließend sind daraus wieder Mittelwert, Minimum, Maximum und Standardabweichung berechnet worden.

Tabelle 3: Ergebnisse der ersten 4 Momente der Verteilung nach 10 000 Simulationen

	Mittelwert		Standardabweichung	
	absolut	relativ zum Original	absolut	relativ zum Original
Original	120 808 541		1 289 556 816	
Ergebnisse bei 10 000 Simulationen im:				
Mittelwert	120 813 450	*0,00* %	1 296 814 275	*0,56* %
Minimum	117 217 343	*– 2,97* %	1 146 141 178	*– 11,12* %
Maximum	124 616 816	*3,15* %	1 442 649 066	*11,87* %
Standard-abweichung	1 129 880	*0,94* %	58 572 913	*4,54* %

	Schiefe		Exzess	
	absolut	relativ zum Original	absolut	relativ zum Original
Original	47,1		2 695	
Ergebnisse bei 10 000 Simulationen im:				
Mittelwert	47,7	*1,07* %	2 774	*2,93* %
Minimum	41,1	*– 12,77* %	2 050	*– 23,91* %
Maximum	53,6	*13,73* %	3 508	*30,17* %
Standard-abweichung	2,4	*5,00* %	280	*10,39* %

Erkennbar ist hier, dass allein durch die verschiedenen Zufallszahlenfolgen Abweichungen bei den Eigenschaften auftreten, die teilweise sehr erheblich sind. Während beim Mittelwert der anonymisierten Daten die Maximalfehler mit 3,15 % noch relativ klein sind und ein relativer mittlerer Fehler im Mittelwert (Standardabweichung) von 0,94 % vertretbar erscheint, sind die Fehler, die bei der Standardabweichung und den höheren Momenten auftreten können, schon erheblich.

Erstaunlich ist, dass im Durchschnitt über alle 10 000 Simulationen alle 4 Momente trotzdem sehr gut reproduziert werden.

Dass Fehler bei der Varianz/Standardabweichung durch multiplikative Überlagerung auftreten, ist bereits theoretisch herleitbar. Für die Varianz zweier unabhängiger multiplikativ verknüpfter Zufallszahlen gilt:

$$V(X*W) = V(X)*V(W) + V(X)*E(W)^2 + V(W)*E(X)^2$$

mit :

$$E(W) = 1$$

$$E(X) = E(X*W)$$

gilt :

$$V(X*W) = (1+V(W))*V(X) + V(W)*E(X*W)^2$$

$$V(X) = \frac{V(X*W) - V(W)*E(X*W)^2}{1+V(W)}$$

Damit lässt sich aus dem beobachteten Mittelwert und beobachteter Varianz die Varianz korrigieren. Für das obige Modell 4.2 – 5 mit den Parametern f=0,11 und s=0,03 gilt dann:

$$V(X) = \frac{V(X*W) - 0{,}013 * E(X*W)^2}{1{,}013}$$

Für den relativen Fehler in der Varianz durch die Anonymisierung müsste gelten:

$$\frac{V(X*W)}{V(X)} - 1 = V(W)\left(1 + \frac{E(X)^2}{V(X)}\right)$$

Damit müssten mit obigen Modellparametern zufallsüberlagerte Umsatzwerte die Varianz um das 0,01311-fache (ca. 1,3 %) überzeichnen. Weil die Standardabweichung im Gesamtbestand größer ist als das 10-fache des Mittelwertes, ist der Ausdruck $E(X)^2/V(X)$ für die Korrektur fast unbedeutend (‹1/100). Kritisch ist jedoch, dass diese entwickelten Korrekturformeln nicht geeignet sind, um diese Fehler brauchbar zu korrigieren. Als Beispiele seien hier erst mal nur 3 zufällige Überlagerungsergebnisse aufgeführt.

Tabelle 4: Wirksamkeit der Formeln zur Varianzkorrektur

Mittelwert	Standardabweichung		Fehler in Standardabweichung	
	unkorrigiert	korrigiert	unkorrigiert	korrigiert
122 246 804	1 389 153 778	1 380 141 896	*7,72* %	*7,02* %
119 619 313	1 202 424 614	1 194 607 387	*– 6,76* %	*– 7,36* %
119 812 490	1 232 190 321	1 224 183 098	*– 4,45* %	*– 5,07* %

Die Korrektur führt hier zu meist unerheblichen Veränderungen des Fehlers gegenüber dem Original. Dabei sind selbst Verschlechterungen der Ergebnisse nicht ausgeschlossen.

Bei näheren Untersuchungen der drei obigen Simulationen bestätigte sich der Verdacht, dass die ungleichmäßige Anonymisierung der großen Einheiten als Ursache für diese Effekte auftritt. Während die verstärkte Veränderung der großen Einheiten nach oben im ersten Beispiel den Mittelwert nur um 1,19 % nach oben verschob, wurde die Standardabweichung bereits um 7,72 % verändert. Bei den beiden anderen Beispielen war zwar die Richtung entgegengesetzt, die Auswirkungen sind jedoch analog. Den relativ kleinen Fehlern von – 0,98 % bzw. – 0,82 % beim Mittelwert folgen die starken Fehler von – 6,76 % bzw. – 4,45% bei der Standardabweichung.

Um die Abhängigkeit der Variation in der Standardabweichung von der Anzahl sehr großer Einheiten und somit indirekt von der Schiefe des Datenbestandes näher zu untersuchen, wurde der Datenbestand aufsteigend sortiert und in mehreren Stufen nur teilweise verarbeitet. Damit wurden bei den folgenden Simulationen immer die größten Einheiten ausgeschlossen und nur der verbleibende Datenbestand aus n Einheiten (Spalte „Auswahl Sätze“) untersucht. Es wurden wieder jeweils 10 000 Simulationen durchgeführt.

Tabelle 5: Die ersten 4 Momente im reduzierten Datenbestand (Mittelwerte aus 10 000 Simulationen)

Auswahl Sätze		Eigenschaften der 1. – 4. Momente im Original			
absolut	in %	Mittelwert	σ	Schiefe	Exzess
16 824	*100*	120 808 541	1 289 556 816	47,1	2 695
16 818	*99,96*	100 741 670	589 648 974	28,1	1 031
16 812	*99,93*	92 650 440	400 188 934	20,5	655
16 806	*99,89*	88 001 587	308 817 559	11,9	198
16 800	*99,86*	85 478 111	278 384 670	10,4	150
16 750	*99,56*	74 431 660	183 485 937	6,3	52
16 700	*99,26*	68 724 665	150 486 176	5,1	33
16 656	*99,00*	65 088 046	132 909 643	4,5	25
15 983	*95,00*	42 704 752	60 520 762	2,4	6
15 142	*90,00*	31 560 889	37 318 131	1,9	3

Mit dem Entfernen der größten Einheiten ist automatisch eine systematische Verkleinerung der Momente im verbleibenden Originalbestand verbunden.

Tabelle 6: Relative Fehler im Mittelwert bei 10 000 Simulationen

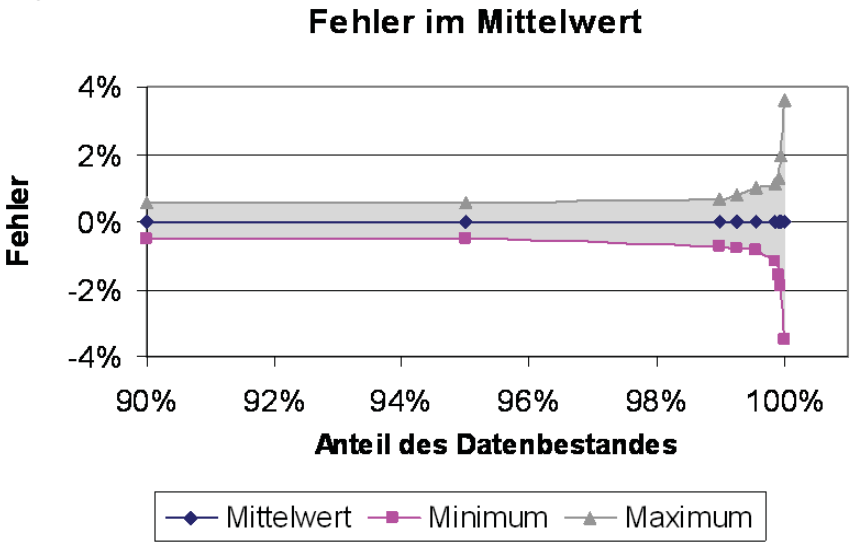
Auswahl Sätze		Relativer Fehler im Mittelwert			
absolut	in %	mittlerer	minimaler	maximaler	σ
		in %			
16 824	*100*	*0,00*	*– 2,97*	*3,15*	*0,94*
16 818	*99,96*	*0,00*	*– 1,90*	*1,96*	*0,52*
16 812	*99,93*	*0,00*	*– 1,51*	*1,57*	*0,39*
16 806	*99,89*	*0,00*	*– 1,15*	*1,37*	*0,32*
16 800	*99,86*	*0,00*	*– 1,19*	*1,13*	*0,30*
16 750	*99,56*	*0,00*	*– 0,85*	*1,00*	*0,24*
16 700	*99,26*	*0,00*	*– 0,77*	*0,81*	*0,21*
16 656	*99,00*	*0,00*	*– 0,72*	*0,69*	*0,20*
15 983	*95,00*	*0,00*	*– 0,53*	*0,55*	*0,16*
15 142	*90,00*	*0,00*	*– 0,53*	*0,55*	*0,14*

Tabelle 7: Relative Fehler in der Standardabweichung bei 10 000 Simulationen

Auswahl Sätze		Relativer Fehler in der Standardabweichung			
absolut	in %	mittlerer	minimaler	maximaler	σ
		in %			
16 824	*100*	*0,56*	*– 11,12*	*11,87*	*4,54*
16 818	*99,96*	*0,64*	*– 8,39*	*10,03*	*2,83*
16 812	*99,93*	*0,66*	*– 6,78*	*7,38*	*2,27*
16 806	*99,89*	*0,69*	*– 3,61*	*5,25*	*1,26*
16 800	*99,86*	*0,70*	*– 3,86*	*4,93*	*1,10*
16 750	*99,56*	*0,75*	*– 1,89*	*3,20*	*0,68*
16 700	*99,26*	*0,79*	*– 1,32*	*2,84*	*0,56*
16 656	*99,00*	*0,80*	*– 1,20*	*2,77*	*0,50*
15 983	*95,00*	*0,97*	*– 0,27*	*2,20*	*0,32*
15 142	*90,00*	*1,11*	*– 0,04*	*2,28*	*0,29*

Wie zu erwarten war, treten beim Mittelwert nur sehr kleine Abweichungen auf. Selbst die Maximalfehler sind unterhalb von 1 %, wenn auf weniger als 0,5 % des Datenbestandes verzichtet wird (siehe auch Abbildung 19). Im Mittel über alle Simulationen zeigt sich, dass die theoretische Eigenschaft der Erwartungstreue gegeben ist.

Abbildung 19
Streuung der Mittelwerte bei 10 000 Simulationen



Bei der Standardabweichung erhält man ein widersprüchliches Ergebnis. Die maximalen Abweichungen und die Standardabweichung über alle Fehler zum Original gehen zwar

zurück, es wird jedoch die theoretisch nachgewiesene systematische Verzerrung der Ergebnisse mit dem Verzicht auf die großen Einheiten immer stärker sichtbar (siehe auch Abbildung 20). Das liegt daran, dass der Ausdruck $E(X)^2/V(X)$ immer größer wird und somit an Einfluss gewinnt. Bei einer Beschränkung des Datenbestandes auf 90 % hat sich die systematische Verzerrung ungefähr verdoppelt. Sie liegt aber weiterhin unter der theoretischen Varianzverzerrung von + 2,23 %.

Abbildung 20
Streuung der Standardabweichung bei jeweils 10 000 Simulationen

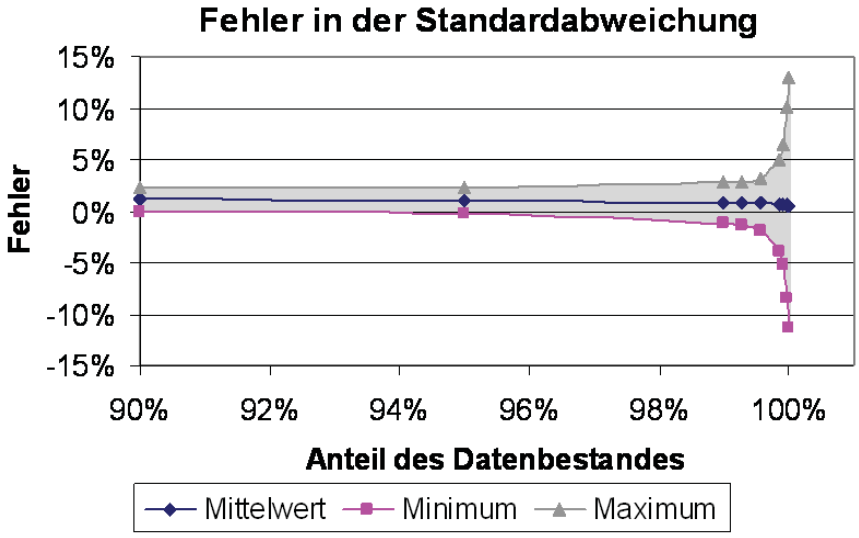


Tabelle 8: Relative Fehler in der Schiefe bei 10 000 Simulationen

Auswahl Sätze		Relativer Fehler in der Schiefe			
absolut	in %	mittlerer	minimaler	maximaler	σ
		in %			
16 824	*100*	*1,07*	*– 12,77*	*13,73*	*5,00*
16 818	*99,96*	*1,42*	*– 12,46*	*12,46*	*4,27*
16 812	*99,93*	*1,42*	*– 14,83*	*17,96*	*6,61*
16 806	*99,89*	*1,82*	*– 78,34*	*10,57*	*2,71*
16 800	*99,86*	*1,92*	*– 6,73*	*9,62*	*2,88*
16 750	*99,56*	*2,06*	*– 2,65*	*6,75*	*1,37*
16 700	*99,26*	*1,96*	*– 1,96*	*5,88*	*1,96*
16 656	*99,00*	*2,26*	*– 2,38*	*5,74*	*0,95*
15 983	*95,00*	*3,12*	*– 1,23*	*5,23*	*0,55*
15 142	*90,00*	*3,85*	*– 1,96*	*5,60*	*0,51*

Tabelle 9: Relative Fehler im Exzess bei 10 000 Simulationen

Auswahl Sätze		Relativer Fehler im Exzess			
absolut	in %	mittlerer	minimaler	maximaler	σ
		in %			
16 824	*100*	*2,93*	*– 23,91*	*30,17*	*10,39*
16 818	*99,96*	*4,04*	*– 23,63*	*31,89*	*9,66*
16 812	*99,93*	*3,73*	*– 29,53*	*42,63*	*14,98*
16 806	*99,89*	*4,87*	*– 17,99*	*24,94*	*6,47*
16 800	*99,86*	*5,01*	*– 15,10*	*23,11*	*5,95*
16 750	*99,56*	*5,60*	*– 7,12*	*16,66*	*3,48*
16 700	*99,26*	*6,12*	*– 3,36*	*15,29*	*2,45*
16 656	*99,00*	*6,31*	*– 5,58*	*14,73*	*2,46*
15 983	*95,00*	*10,47*	*– 4,62*	*16,68*	*1,68*
15 142	*90,00*	*15,41*	*– 8,94*	*21,82*	*1,81*

Auch bei den höheren Momenten (Schiefe und Exzess) sind die Effekte nachweisbar, die bereits bei der Standardabweichung auftraten (siehe auch die folgenden Abbildungen 21 und 22). Mit zunehmendem Verzicht auf die großen Einheiten verringert sich die Streuung der Simulationsergebnisse. Dabei sind jedoch die beiden Fehlerquellen „systematische Verzerrung durch das Verfahren“ und „Fehler durch ungünstige Konstellation der Zufallszahlen“ in ihren Auswirkungen erheblich. Es dürfte nur schwer zu entscheiden sein, ob eine systematische Verzerrung von 3,85 % (bei der Schiefe) bzw. 15,41 % (beim Exzess) besser ist, als ein zufälliger Fehler von ca. 5 % (bei der Schiefe) bzw. 10,39 % (beim Exzess). Die systematischen Verzerrungen sind nur dann „die bessere Wahl“, wenn der Wissenschaftler bei der Analyse in der Lage und bereit ist, sie im Modell mit zu berücksichtigen und Korrekturen vorzunehmen. Der große Einfluss der Ausreißer im Datenbestand auf die Verteilungseigenschaften (Standardabweichung, Schiefe und Exzess) bewirkt in den Testdaten, dass durch die Erwartungstreue der Datenveränderung im Ergebnis vieler Simulationen die Verteilungseigenschaften im Mittel gut reproduziert werden, obwohl eigentlich systematische Verzerrungen vorliegen müssten. Die systematische Verzerrung wird erst mit dem Verzicht auf die Ausreißer im Datenbestand sichtbar.

Abbildung 21
Streuung der Schiefe bei jeweils 10 000 Simulationen

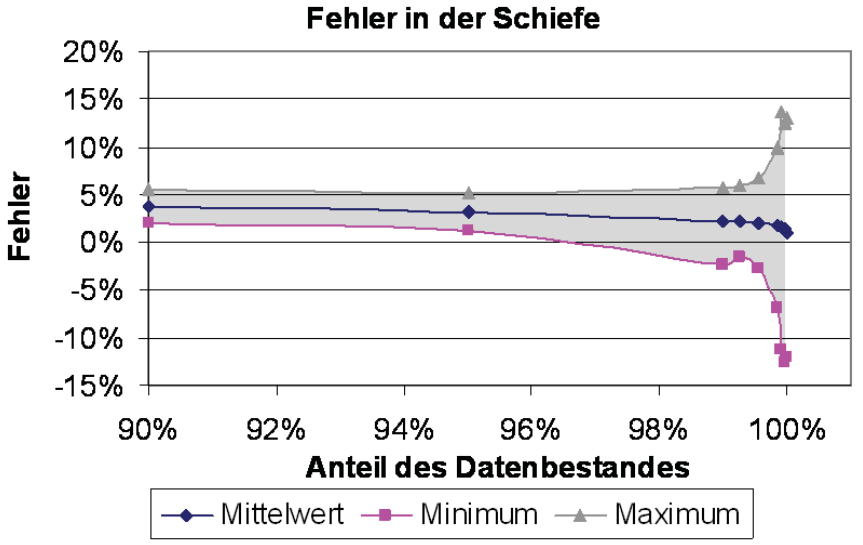
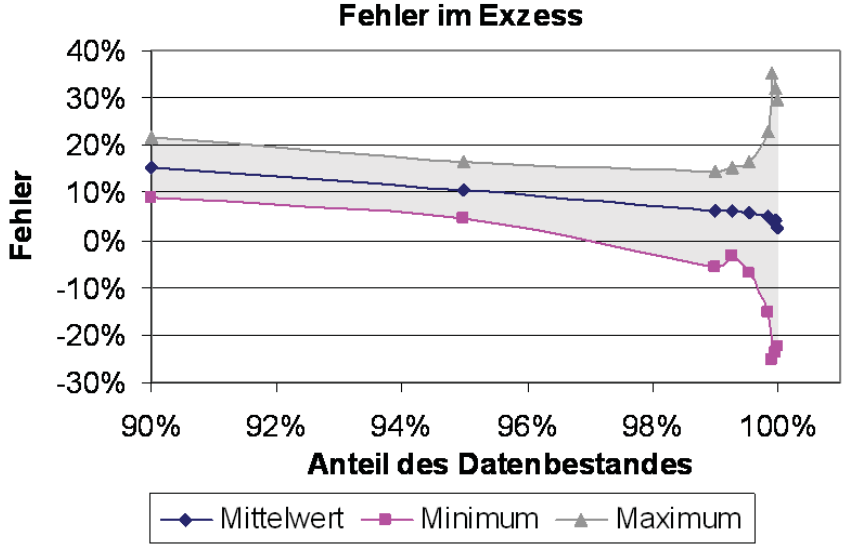


Abbildung 22
Streuung des Exzess bei jeweils 10 000 Simulationen

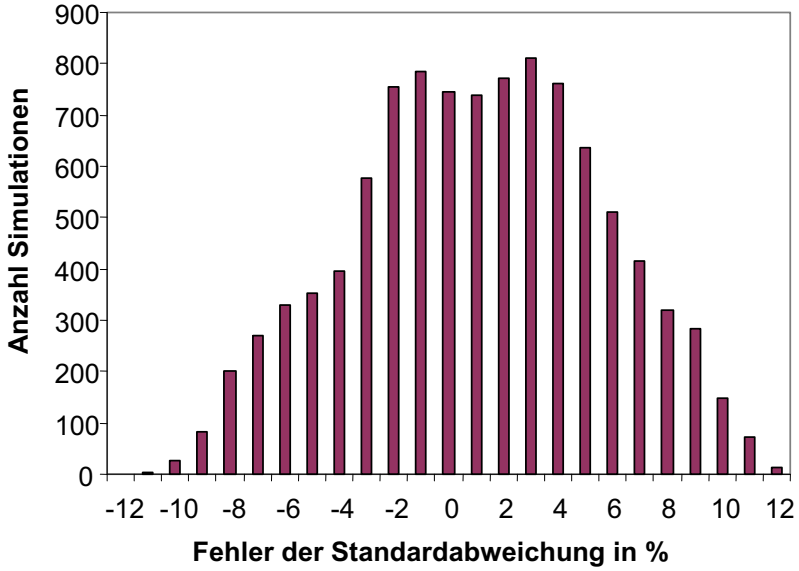


Im Folgenden wurde die Verteilung der Fehler bei der Reproduktion der Standardabweichung etwas näher untersucht. Dazu wurden die relativen Fehler der einzelnen Simulationen in den Standardabweichungen auf Ganzzahligkeit gerundet und dann nach diesem Fehler gruppiert. Die Häufigkeiten des Auftretens der Fehler ist in der folgenden Abbildung 23 dargestellt. Als Datenbestand wurde der Gesamtbestand verwendet. Erkennbar

ist, dass sich die Fehler um 0 % konzentrieren (im Bereich von – 2 % bis 4 % am meisten auftreten). Hieraus erklärt sich auch, dass der mittlere Fehler in der Standardabweichung über alle Simulationen nur bei 0,56 % lag. Dabei sind in der Verteilung sowohl zwei Spitzen als auch eine leichte Verschiebung zu 0 % erkennbar, so dass man nicht von einer völlig symmetrischen Verteilung ausgehen kann. Während einerseits ca. 23 % der simulierten Zufallszahlenfolgen die Standardabweichung der Gesamtbestandes sehr gut reproduzieren (Fehler ≤1 %), sind andererseits ca. 37 % der simulierten Zufallszahlenfolgen nicht geeignet, eine Analyse der Standardabweichung für den Gesamtbestand zu ermöglichen (Fehler ≥5 %). Jede dritte multiplikative Überlagerung liefert damit unbrauchbare Ergebnisse, wenn mit den Daten höhere Momente analysiert werden sollen.

Der mittlere Fehler in der Standardabweichung von + 0,56 % entspricht nicht der theoretischen systematischen Verzerrung, für die ja oben bereits + 1,31 % hergeleitet wurde. Dieser Effekt, dass der mittlere Fehler bei ca. 50 % des theoretischen Fehlers liegt, tritt selbst dann noch auf, wenn der Datenbestand durch Entfernen der größten Werte auf 90 % reduziert wurde.

Abbildung 23
Verteilung der relativen Fehler in der Standardabweichung bei 10 000 Simulationen mit dem Gesamtbestand



Im Ergebnis der Untersuchungen stellen sich die folgenden Fragen:

1. Kann man multiplikative Zufallsüberlagerungen bei diesen Ergebnissen noch als brauchbares Verfahren einstufen?
2. Gibt es Möglichkeiten, bessere Ergebnisse zu erreichen und die Fehler in den höheren Momenten zu kontrollieren?

Für die Beurteilung des Verfahrens muss nochmals hervorgehoben werden, dass die schlechten Ergebnisse einerseits das Ergebnis der erzwungenen relativen Veränderungen sind. Diese Veränderungen sind aus dem Schutzbedarf heraus entstanden, welcher die Möglichkeit kleiner Veränderungen der Daten nur mit sehr geringer Wahrscheinlichkeit zulässt. Andererseits ist dieses Ergebnis abhängig von der Schiefe der Originaldaten. Während der Schutzbedarf einerseits ja das eigentliche Ziel der Anonymisierungsverfahren darstellt und somit nur insofern hinterfragt werden kann, ob die Anonymisierungsparameter wirklich in dieser Größe erforderlich sind, ist die zweite Ursache der schlechten Ergebnisse von dem speziellen Datenbestand abhängig. Diese Konstellation muss so nicht immer auftreten, so dass das Verfahren bei nicht so schief verteilten Daten völlig unproblematisch bezüglich der Reproduktion der höheren Momente sein kann. Im Vergleich zu anderen Anonymisierungsverfahren hat somit die multiplikative Zufallsüberlagerung mit Mischungsverteilungen genau die gleichen Probleme. Der Erhalt der Schutzwirkung geht bei schief verteilten Daten ggf. zu Lasten der Analysefähigkeit. Da aber gerade die Erzeugung einer ausreichenden Schutzwirkung bei schiefen Originaldaten die Stärke der multiplikativen Zufallsüberlagerung ist, soll die zweite Frage „Gibt es Möglichkeiten, bessere Ergebnisse zu erreichen und die Fehler in den höheren Momenten zu kontrollieren?“ im Folgenden näher untersucht werden.

Als Triviallösung kann folgende Vorgehensweise betrachtet werden: Nach der Berechnung und Analyse aller möglichen Zufallszahlenfolgen wird diejenige Zufallszahlenfolge ausgewählt, die die Ergebnisse am wenigsten beeinflusst. Der mit der ausgewählten Zufallszahlenfolge erzeugte anonyme Datenbestand wird für die Weitergabe an die Wissenschaft verwendet. Diese Herangehensweise ist einerseits robust und ermöglicht andererseits beliebige Qualitätskriterien in die Analyse mit aufzunehmen. Sie hat aber einen entscheidenden Nachteil: Diese Herangehensweise setzt voraus, dass die Anonymisierung mit möglichst vielen, ggf. allen Zufallszahlenfolgen, berechnet und analysiert wird. Was bei dem oben beschriebenen Beispiel des Umsatzmerkmals aus der Kostenstrukturerhebung (ca. 16 000 Datensätzen) und den Qualitätsmerkmalen 1. – 4. Momente noch im Bereich ca. 30 Minuten möglich war, würde bei Einbeziehung aller Merkmale und der Anwendung auf große Datenbestände (Steuerstatistiken mit mehreren Millionen Datensätzen) große Rechenzeitprobleme auslösen. Deshalb ist ein Verfahren gesucht, mit dem ohne viele Simulationsläufe eine gute Lösung gefunden werden kann.

Es soll deshalb eine Kontrolle der Überlagerung, vor allem bei den großen Einheiten durch eine **kontrollierte** Vorgehensweise, bei der Überlagerung vorgestellt werden.

4.2.4 Kontrollierte Überlagerungen

Das Hauptproblem der multiplikativen Überlagerung ist offensichtlich die unabhängige Überlagerung der einzelnen großen Einheiten. Diese kann zu dem Effekt führen, dass die großen Einheiten schwerpunktmäßig in eine Richtung verändert werden, was durch die analoge Überlagerung der vielen kleinen Einheiten im Datenbestand nicht mehr kompensiert werden kann. Deshalb ergibt sich auch für die Gesamtsummen im Datenbestand eine starke Verzerrung. Die im Folgenden beschriebenen Verfahren versuchen dieses Problem zu lösen. Die stärkste Veränderung wird im Modell 4.2 – 5 durch den Überlagerungsfaktor $w_i^{\#}$ erzeugt. Da diese Zufallszahl (Realisationen 1 und – 1) nur einmal für alle

Werte einer Einheit bestimmt wird, ist es schwierig, die Auswahl so zu treffen, dass sie für alle Variablen die optimale Richtung bestimmt. Theoretisch gibt es drei verschiedene Lösungsmöglichkeiten:

a) Bestimmung im Zuge eines Optimierungsproblems.
b Bestimmung im Ergebnis eines eindimensionalen Rankingverfahrens.
c) Bestimmung im Ergebnis eines mehrdimensionalen Rankingverfahrens.

Während die Variante a) die globale Lösung finden würde, dienen die beiden Varianten b) und c) zur Bestimmung von Näherungslösungen durch ein heuristisches Verfahren, d. h., dass zusätzliche Restriktionen/Berechnungsvorschriften eingeführt werden, die einerseits die Lösbarkeit stark erleichtern, aber andererseits dazu führen, dass es sich nur noch um die optimale Lösung aus einer vorher eingeschränkten Menge von möglichen Lösungen handelt. Wenn diese Einschränkungen jedoch plausibel und nicht zu restriktiv sind, sollten auch diese Ergebnisse nahe bei einem globalen Optimum liegen. Der Vorteil von Näherungslösungen besteht jedoch darin, dass ihre Bestimmung mit bedeutend kleinerem Rechenaufwand möglich ist. Damit sind solche Verfahren bei kleinen Datenbeständen sehr effizient und bei sehr großen Datenbeständen oft die einzigen mit begrenzten Ressourcen noch realisierbaren Verfahren. Eine analoge Herangehensweise findet z. B. bei den mehrdimensionalen Mikroaggregationsverfahren statt.

Hauptproblem aller Verfahren ist, dass den Modellen Zielfunktionen (Optimalitätskriterien) zu Grunde gelegt werden müssen, mit denen es möglich ist, zwei beliebige Lösungen zu vergleichen und die bessere auszuwählen. Da der Schwachpunkt der rein stochastischen Überlagerung in der Qualität der erhaltenen Momente lag, sollten auch möglichst viele Eigenschaften in die Zielfunktion mit einfließen. Für ökonometrische Analysen sind dabei der Erhalt von Mittelwerten und Varianzen/Standardabweichungen besonders wichtig. Hier stellt sich leider folgendes Problem: Die Mittelwerte und Standardabweichungen von verschiedenen Variablen/Merkmalen haben alleine auf Grund ihres ökonomischen Inhalts und der sich daraus ergebenden Dimension der Daten ganz verschiedene Größenordnungen. Diese Werte lassen sich nur bedingt zu einem gemeinsamen Zielkriterium zusammenfügen. Der Trick, die Daten vorher zu normieren und somit in allen Daten den gleichen Mittelwert (Null) und die gleiche Standardabweichung (Eins) zu erzeugen (siehe mehrdimensionale Mikroaggregation im Abschnitt 2.2.6, S. 51), ist für die Verfahrensgruppe der multiplikativen Überlagerungen völlig ungeeignet, weil dann auch die Vorzeichen der Daten in großem Umfang geändert werden müssten und somit nicht mehr durch das Verfahren automatisch erhalten werden können. Die multiplikative Überlagerung würde dann einen großen methodischen Vorteil gegenüber der additiven Überlagerung verlieren. Deshalb sollten die Fehler in Mittelwert und Varianz/Standardabweichung erst innerhalb der Zielfunktion normiert werden. Bei den beiden verschiedenen Eigenschaften der Variablen (Mittelwert und Streuung) müsste dann mit Hilfe von Gewichten ein Ausgleich zwischen den beiden Teilzielen geschaffen werden.

Eine mögliche Zielfunktion wäre deshalb:

$$ZF = \alpha \sum_{j=1}^{m} \left(\frac{\left| \overline{x_{.j}^{a}} - \overline{x_{.j}^{o}} \right|}{m \overline{x_{.j}^{o}}} \right) + (1-\alpha) \sum_{j=1}^{m} \frac{\left| \sigma(x_{.j}^{a}) - \sigma(x_{.j}^{o}) \right|}{m \sigma(x_{.j}^{o})}$$

$$x_{ij}^{a} - x_{ij}^{o} = (f w_i^{\#} + w_{ij}^{*}) x_{ij}^{o} \qquad \textbf{(4.2 – 6)}$$

$$w_i^{\#} \in (1;-1) \quad ; P(w_i^{\#} = 1) = P(w_i^{\#} = -1) = 0{,}5$$

$$w_{ij}^{*} \sim N(0, s)$$

m – Anzahl der stetigen Merkmale im Datenbestand

f – kennzeichnet dabei die Niveauverschiebung der Werte eines Unternehmens

s – Standardabweichung der Normalverteilung, beschreibt die Stärke der zwischen den Werten unabhängigen zusätzlichen Zufallsüberlagerung.

$\overline{x^{o}_{.j}}$, $\overline{x^{a}_{.j}}$ – Mittelwerte in der Datenspalte j im originalen bzw. anonymen Datenbestand

$\sigma(x^{o}_{.j})$, $\sigma(x^{a}_{.j})$ – Standardabweichungen in der Datenspalte j im originalen bzw. anonymen Datenbestand

$w_i^{\#}, w_{ij}^{*}$ – Zufallszahlen mit i=1,...,n und j=1,...,m

Bei dieser Zielfunktion werden die mittleren relativen Fehler für die Mittelwerte und die Standardabweichungen bestimmt. Beide werden dann mit einem Gewichtungsfaktor α zu einem Gesamtmaß der relativen Fehler zusammengefasst.

a) Lösung als Optimierungsproblem

Innerhalb der zu lösenden Aufgabe 4.2 – 6 beschreiben f und s die Stärke der Anonymisierung. Die Mittelwerte und die Standardabweichungen der Originalwerte sind extern durch die Originaldaten fest gegebene Größen. Somit bleiben nur die Zufallszahlen $w_i^{\#}$ und w_{ij}* als mögliche Optimierungsvariablen. Das w_{ij}* normalverteilt sein sollen lässt sich schwer durch Nebenbedingungen sicherstellen. Diese Zufallszahlen müssten deshalb unabhängig generiert und als gegeben vorausgesetzt werden.

Die bimodale Verteilung der $w_i^{\#}$ lässt sich jedoch durch folgende Nebenbedinungen leicht ins Modell mit aufnehmen:

$$w_i^{\#} \in (1;-1)$$

und

$$\left| \sum_{i=1}^{n} w_i^{\#} \right| \leq 1$$

Die Bedingungen sind zwar härter formuliert als bei echten Zufallszahlen, sie haben aber den Vorteil, dass sie sich leichter kontrollieren lassen. Die Summe der $w_i^{\#}$ sei gleich Null,

für den Fall, dass die Anzahl der Einheiten gerade ist, ansonsten weiche die Anzahl der positiven Zufallszahlen von den negativen Zufallszahlen um maximal 1 ab. Leider ist dieses Optimierungsproblem nichtlinear und ganzzahlig, so dass eine Lösung nicht so einfach zu ermitteln ist. Wegen der Ganzzahligkeit der $w_i^{\#}$ mit den beiden möglichen Realisationen $w_i^{\#} \in (-1, 1)$ ist die Anzahl der möglichen Lösungen endlich, so dass ggf. auch ein Durchtesten der Lösungen möglich ist. Die Anzahl der möglichen Kombinationen von $w_i^{\#}$ ist jedoch abhängig von der Anzahl der Einheiten im Datenbestand. Die zu testenden Möglichkeiten nehmen sehr stark zu, je größer der Datenbestand ist. Für kleine Datenbestände wäre es aber durchaus möglich, die optimale Lösung zu bestimmen.

b) Lösung als eindimensionales Rankingverfahren

Da die Anzahl der möglichen Kombinationen von $w_i^{\#}$ sehr groß werden kann, stellt sich die Frage, ob es nicht möglich ist, weitere vereinfachende Annahmen vorzunehmen. Die Forderung, dass $E(w_i^{\#})=0$ gilt, bedingt für optimale Lösungen, dass Veränderungen von großen Merkmalswerten durch entgegengesetzte Veränderungen anderer großer Einheiten kompensiert werden. Wäre das nicht der Fall, müssten bedeutend mehr kleine Einheiten eine zu den großen Einheiten entgegengesetzte Datenveränderung erfahren, um die Veränderung weniger großer Einheiten zu kompensieren. Die Forderung, $E(w_i^{\#})=0$, wäre dann nicht zu realisieren. Dieser Effekt soll beim folgenden Lösungsansatz ausgenutzt werden, um das Bestimmungsverfahren in seinem Berechnungsaufwand stark zu reduzieren.

Zu lösen sei wieder die Aufgabe 4.2 – 6. Für eine schrittweise Bearbeitung des Datenbestandes, der eine Kompensation der Veränderung der Daten bei möglichst gleichwertigen Einheiten vorsieht, ist als erstes eine Sortierung der Daten vorzunehmen, da so gewährleistet ist, dass gleichwertige Daten zusammen bearbeitet werden. Vorgeschlagen wird ein eindimensionales Ranking, bei dem nach Auswahl eines dominierenden Merkmals (siehe eindimensionale Mikroaggregationsverfahren im Abschnitt 2.2.6, S. 50 f.) der Datenbestand nach diesem Merkmal absteigend sortiert wird. Die absteigende Sortierung ist erforderlich, damit die großen absoluten Veränderungen, die bei der Anonymisierung der großen Einheiten erfolgen, am Anfang der Anonymisierung durchgeführt werden und so noch im Laufe des Verfahrens korrigiert werden können und nicht erst bei den letzten zu bearbeitenden Einheiten erzeugt werden.

Der Datenbestand wird dann nach folgendem Algorithmus bearbeitet:

1. Vorgabe der Sicherheitsparameter f und s.
2. Auswahl eines dominierenden Merkmals k (Merkmal mit möglichst hoher Korrelation zu allen anderen zu bearbeitenden Merkmalen).
3. Absteigende Sortierung des Datenbestandes nach dem Merkmal k.
4. Generierung einer Zufallszahlenmatrix w_{ij}* mit:
   $$w_{ij}^* \sim N(0,s)\ ; \quad i = 1,...,n \wedge j = 1,...,m\ .$$
5. Anonymisierung der größten Einheit mit $w_1^{\#} = -1$. und somit
   $$x_{1j}^a = (1 - f + w_{1j}^*) x_{1j}^o\ .$$

6. Schrittweise Anonymisierung der weiteren Einheiten i=2,...,n mit

$$x_{ij}^{a} = (1 - f w_i^{\#} + w_{ij}^{*}) x_{ij}^{o}$$

wobei:

$w_i^{\#} = -1$, wenn $\sum_{i=1}^{l-1} \left(x_{ik}^{a} - x_{ik}^{o} \right) > 0$

, sonst $w_i^{\#} = 1.$

(Es wird hier nur der Fehler in der Summe der Werte des dominierenden Merkmals k kontrolliert und $w_i^{\#}$ so gewählt, dass die Datenveränderung diesem Fehler entgegengerichtet ist.)

Dieses Verfahren wurde mit mehreren Datenbeständen im Rahmen des Projektes faktische Anonymisierung wirtschaftsstatistischer Einzeldaten getestet. In Ronning et al. (2005) wurde es im Vergleich zu anderen Verfahren der Zufallsüberlagerung mit verschiedenen Datenbeständen getestet. Es wird dort als das Verfahren von Höhne bezeichnet (siehe Ronning et al. 2005, S. 282 ff.).

c) Lösung als mehrdimensionales Rankingverfahren

Das eindimensionale Ranking hat leider einige Nachteile:

Mit der Auswahl eines dominierenden Merkmals k ist bereits ein besonderer Qualitätserhalt für dieses Merkmal festgelegt, der bei den anderen Merkmalen nur dann erreicht werden kann, wenn diese mit dem Merkmal k sehr stark korrelliert sind.

Innerhalb des Merkmals k dürfen keine Missingwerte auftreten, da dann keine richtige Einsortierung der Einheiten erfolgen kann. Da üblicherweise diese Werte analog Null-Werten ans Ende der absteigenden Sortierung sortiert werden, könnten damit selbst sehr große Unternehmen erst am Ende des Verfahrensdurchlaufs behandelt werden. Dieses Problem tritt z. B. bei Paneldaten verstärkt auf, da dann die Existenz der Unternehmen nicht für alle Perioden im Datenbestand gewährleistet sein muss. Hier können z. B. Unternehmensausgliederungen, -fusionen oder ähnliche Veränderungen dazu führen, dass selbst sehr große Unternehmen im Datenbestand neu auftreten, bzw. verschwinden, so dass bei der Auswahl eines Merkmals k Missingwerte für dieses Merkmal nicht verhindert werden können.

Sind bei dem dominierenden Merkmal auch negative Werte möglich (z. B. Bestandsveränderungen, Gewinnangaben, o. Ä.), führt die eindimensionale Betrachtung dazu, dass große Veränderungen auch bei der Anonymisierung der letzten Einheiten erzeugt werden. Diese können im weiteren Durchlauf des Verfahrens nicht mehr korrigiert werden. Deshalb müsste immer nach den Absolutwerten sortiert werden.

Aus diesen Gründen wurde eine Weiterentwicklung des Verfahrens getestet, die mehrere Merkmale gleichzeitig bei der Anonymisierung berücksichtigt. Die Notwendigkeit ergab sich vor allem bei dem Versuch, die kontrollierte Überlagerung auf Paneldaten zu übertragen. Gerade bei Paneldaten treten wegen des Ausfalls und der Neuaufnahme von Einheiten im Panel verstärkt Missingfälle auf, so dass es nicht möglich ist, ausschließlich mit einer Variablenspalte ein Ranking der Einheiten vorzunehmen.

Der Datenbestand wird dann nach folgendem Algorithmus bearbeitet:

1. Vorgabe der Sicherheitsparameter f und s.
2. Bestimmung eines Scores S über alle für die Einheit i existierenden Merkmale, mit dem die „Größe“ der Einheiten beschrieben wird. Der Score soll dabei die mögliche Veränderungswirkung von multiplikativen Überlagerungen über alle Merkmale bewerten.

$$S_i = \sum_{j=1}^{m} \left(\frac{\left| \overline{x_{ij}^o} \right|}{m \overline{x_{.j}^o}} \right).$$

   Treten bei einer Einheit Missingwerte auf, wird die Variable im Score nicht berücksichtigt (m wird reduziert).
3. Absteigende Sortierung des Datenbestandes nach dem Score S.
4. Generierung einer Zufallszahlenmatrix w_{ij}* mit:

$$w_{ij}^* \sim N(0,s)\ ; \quad i = 1,...,n \wedge j = 1,...,m$$

5. der größten Einheit mit $w_1^{\#} = -1$. und somit

$$x_{1j}^a = (1 - f + w_{1j}^*) x_{1j}^o$$

6. Schrittweise Anonymisierung der weiteren Einheiten i=2,...,n mit

$$x_{ij}^a = (1 - f w_i^{\#} + w_{ij}^*) x_{ij}^o$$

   wobei die Zielfunktion ZF für die beiden Möglichkeiten $ZF_1 = ZF(w_i^{\#} = -1)$ und $ZF_2 = ZF(w_i^{\#} = 1)$ mit ZF als:

$$ZF = \sum_{j=1}^{m} \left(\frac{\left| \overline{x_{.j}^a} - \overline{x_{.j}^o} \right|}{m \overline{x_{.j}^o}} \right)$$

$$x_{kj}^a = x_{kj}^o \quad ; \forall\ k > i$$

   bestimmt wird. Auswahl der kleineren Variante ZF_1 oder ZF_2 und Verwendung des zugehörigen $w_i^{\#} = -1$ bzw. $w_i^{\#} = 1$ bei der Anonymisierung der Einheit i.
   (Es wird hier der normierte Fehler im Mittelwert / in der Summe der Werte aller Merkmale kontrolliert, und $w_i^{\#}$ so gewählt, dass die Datenveränderung den kleineren Fehler erzeugt. Die Werte der Zielfunktion ZF_1 bzw. ZF_2 bei der Anonymisierung der Einheit i können auch kleiner sein als bei der Einheit $i-1$, wenn der Anonymisierungsschritt in größerem Umfang alte Veränderungen kompensiert.)

   Die im Anonymisierungsschritt 6 verwendete Zielfunktion entspricht der Zielfunktion in 4.2 – 6 für den Wert $\alpha = 1$. Das wurde vorgenommen, weil die Berechnung der einzelnen Zielfunktionswerte ZF_1 bzw. ZF_2 für die Einheit i iterativ durch Korrekturformeln aus den Werten für die Einheit $i-1$ vorgenommen werden sollte. Hier wären aber Erweiterungen denkbar, mit denen auch die Standardabweichungen mit kontrolliert werden könnten.

5 Zusammenfassung

Für die statistische Geheimhaltung von Einzeldaten wurden eine Vielzahl von Verfahren entwickelt. Das Entstehen dieser Vielfalt von Verfahren resultierte sowohl aus der Verschiedenartigkeit der zu anonymisierenden Datenbestände als auch der Unterschiedlichkeit der vorgesehenen Auswertungen. So zeichnen sich z. B. Personendaten durch ein große Anzahl an kategorialen Merkmalen mit ggf. wenigen nicht sehr schief verteilten metrischen Merkmalen aus. Bei wirtschaftsstatistischen Daten überwiegen in der Regel die metrischen Merkmale. Diese sind je nach Statistik mehr oder weniger schief verteilt (z. B. sehr schief bei Unternehmensstatistiken im Vergleich zu Handwerksstatistiken). In Abhängigkeit von den vorgesehenen Auswertungen erweisen sich deshalb die unterschiedlichsten Anonymisierungsverfahren als vorteilhaft. Im Rahmen dieser Arbeit wurde ein Überblick über Verfahren zur statistischen Geheimhaltung von Einzeldaten gegeben. Neben der Kurzbeschreibung der Verfahren, wurden analytische Eigenschaften von mit diesen Verfahren anonymisierten Daten als auch ggf. bestehende Restrisiken dargestellt. Nach Möglichkeit wurden Hinweise für die Verhinderung dieser Risiken gegeben. Dabei wurden sowohl Zuordnungsversuche (als Massenfischzüge oder Einzeldatenangriffe) als auch Datenangriffe über Tabellenauswertungen (Tabellengeheimhaltung) betrachtet.

Für die Anonymisierung wirtschaftsstatistischer Daten erweist sich der Einsatz datenverändernder Verfahren als erforderlich, weil sie meistens sehr schief verteilte metrische Merkmale enthalten. Von den im Projekt „faktische Anonymisierung wirtschaftsstatistischer Einzeldaten“ untersuchten Verfahrensgruppen erwiesen sich die Verfahren der Mikroaggregation und der Zufallsüberlagerung als überlegen, wenn die Anonymisierung mit dem Ziel der Erstellung eines Scientic-Use-Files, d. h. ohne vorherige Festlegung der Auswertungsziele durch den Wissenschaftler, durchgeführt wurde (siehe Ronning et al. 2005). Werden projektbezogen Daten bereitgestellt, so erweisen sich z. B. auch die informationsreduzierenden Verfahren als vorteilhaft, da man bereits durch das Entfernen aller für die Analysen nicht benötigten Merkmale das Risiko einer Offenlegung von Informationen stark senken kann. Sind außerdem die vorgesehenen Modelle bekannt, so kann man das Verfahren so auswählen, dass gerade die benötigten Eigenschaften am besten erhalten bleiben.

In dieser Arbeit werden Verfahrenserweiterungen dargestellt, die durch den Autor für die Verfahrensgruppen der Mikroaggregation und der Zufallsüberlagerung entwickelt wurden. Die eindimensionale unabhängige Mikroaggregation mit Varianzerhalt stellt dabei ein Verfahren dar, dass sich durch einen hochwertigen Erhalt der Analysemöglichkeiten auszeichnet. Der für Mikroaggregationsverfahren typische Effekt des Varianzverlustes wird durch das Verfahren verhindert. Unabhängig durchgeführte eindimensionale Mikraggregationen können die Sicherheit dabei wegen der relativ geringen Veränderung der Originalwerte jedoch nur gewährleisten, wenn dieses Verfahren zusammen mit anderen Verfahren der Informationsreduktion eingesetzt wird. Zufallsüberlagerungen bieten dagegen durch die Möglichkeit der Vorgabe von abgestimmten Überlagerungsparametern die Möglichkeit, das Verfahren so anzuwenden, dass die Sicherheit der Daten allein durch die Veränderung der Originalwerte gesichert wird. Setzt man die Parameter jedoch so, dass die Originaldaten nur im notwendigen Umfang verändert werden, so liefert das entwickelte Verfahren der kontrollierten multiplikativen Überlagerung mit Mischungsverteilungen gute

Ergebnisse. Die beiden Verfahren wurden deshalb für die Erstellung von Scientific-Use-Files in den Projekten zur „faktischen Anonymisierung von wirtschaftsstatistischen Einzeldaten“ und „faktische Anonymisierung wirtschaftsstatistischer Paneldaten“ eingesetzt (siehe z. B. Ronning et al. 2005).

Literaturverzeichnis

Appel, G., Kinzel, S., Nölte, D. (1993): SAFE – A Generally Usable Program System for the Anonymization of Individual Data in Official Statistics, in: Proceedings of the International Seminar on Statistical Confidentiality, Dublin, Ireland, 8 – 10 September 1992, S. 201 – 228.

Borchsenius, L. (2005): New Developments in the Danish System for Access to Microdata, in: Joint UNECE/Eurostat Work Wession on Statistical Data Confidentiality, Geneva, 9 – 11 November 2005.

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

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

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

Bundesdatenschutzgesetz (BDSG) in der Fassung der Bekanntmachung vom 14. Januar 2003; Stand: 31. August 2006.

Bundesstatistikgesetz (BStatG) in der Fassung vom 22. Januar 1987; Stand: 21. Juni 2005.

Carlson, M., und Salabasis, M. (2002): A Data Swapping Technique for Generating Synthetic Samples: A Method for Disclosure Control, in: Research in Official Statistics, 6, S. 35 – 64.

Corsini, V., Franconi, L., Pagliuca, D. und Seri, G. (1998): An Application of Microaggregation Methods to Italian Business Surveys, in: Statistical Data Protection Proceedings of the Conference, Eurostat 1999.

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

Dandekar, R. A., Cox, L. H. (2002): „Synthetic Tabular Data – An Alternative to Complementary Cell Suppression“ (unpublished paper).

Dandekar, R. A., Cohen, M. und Kirkendall, N (2001): Applicability of Latin Hypercube Sampling to Create Multi Variate Synthetic Micro Data. Download available:
epp.eurostat.ec.europa.eu/portal/research_methodology/documents/83.pdf

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

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

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

Domingo-Ferrer, J. und Mateo-Sanz J. M. (2001): An Empirical Comparison of SDC Methods for Continuous Microdata in Terms of Information Loss and Disclosure Risk, Working Paper at the Joint ECE/Eurostat Work Session on Statistical Data Confidentiality 14 – 16 March.

Domingo-Ferrer, J. und Mateo-Sanz, J. M. (2002): Practical Data-Oriented Microaggregation for Statistical Disclosure Control, in: IEEE Transactions on Knowledge and Data Engineering, Vol. 14, No. 1.

Dück, W., Körth, H., Runge, W., Wunderlich, L. (1984): Mathematik für Ökonomen, Verlag Die Wirtschaft, Berlin.

Evers, K. und Höhne, J. (1999): SAFE – Ein Verfahren zur Anonymisierung und statistischen Geheimhaltung wirtschaftsstatistischer Einzeldaten, in: Schriftenreihe „Spektrum Bundesstatistik", Bd. 14, S. 136 –147, hrsg. vom Statistischen Bundesamt, Wiesbaden.

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

Gießing, S. (1999): Statistische Geheimhaltung in Tabellen, in: Methoden zur Sicherung der statistischen Geheimhaltung, Schriftenreihe „Forum der Bundesstatistik", Bd. 31, S. 6 – 26, hrsg. vom Statistischen Bundesamt, Wiesbaden.

Gießing, S. (2004): „Kurs zur Tabellengeheimhaltung", Unterlagen für die interne Fortbildung zur Tabellengeheimhaltung im Gemeinsamen Fortbildungsprogramm der Statistischen Ämter des Bundes und der Länder, November 2004.

Gnoss, R., Ronning, G., Arndt, C., Höhne, J., Lenz, R., Rosemann, M., Sturm, R., Vorgrimler, D., Wiegert, R. (2003): Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten, Zwischenbericht zum Forschungsprojekt an das Bundesministerium für Bildung und Forschung (BMBF).

Gnoss, R., Ronning, G., Arndt, C., Höhne, J., Lenz, R., Rosemann, M., Sturm, R., Vorgrimler, D., Wiegert, R. (2003a): Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten, Anlage zum Zwischenbericht zum Forschungsprojekt an das Bundesministerium für Bildung und Forschung (BMBF).

Göhler, W (1989): Höhere Mathematik – Formeln und Hinweise, Leipzig, 10. Auflage.

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

Heitzig, J. (2004): Protection of Confidential Data when Publishing Correlation Matrices in COMPSTAT'2004 Symposium, Physica-Verlag/Springer.

Heitzig, J. (2005): The "Jackknife" Method: Confidentiality Protection for Complex Statistical Analyses, in: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, 9 – 11 November.

Höhne, J. (2003a): Methoden zur Anonymisierung wirtschaftsstatistischer Einzeldaten, in: Gnoss, R. und Ronning, G. (Hrsg.):, Anonymisierung wirtschaftsstatistischer Einzeldaten, Schriftenreihe „Forum der Bundesstatistik", Bd. 42, S. 69 – 94, hrsg. vom Statistischen Bundesamt, Wiesbaden.

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

Höhne, J. (2004a): Varianten von Zufallsüberlagerungen. Arbeitspapier des Projektes „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten"

Höhne, J. (2004b): Weiterentwicklung von Mikroaggregationsverfahren. Arbeitspapier des Projektes „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten".

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

Hundepool, A. und de Wolf, P.-P. (2005): OnSite@Home: Remote Access at Statistics Netherlands, in: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, 9 – 11 November.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E. S., Seri, G. und de Wolf, P.-P. (2009): Handbook on Statistical Disclosure Control. ESSNET SDC.

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

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

Kim, J. J. and Winkler, W. E. (2003): Multiplikative Noise for Masking Continous Data, Research Report Series (Statistics #2003-01) Statistical Research Division U.S. Bureau of the Census, Washington, DC 20233.

Köhler, S. (1999): Anonymisierung von Mikrodaten in der Bundesstatistik und ihre Nutzung – Ein Überblick, in: Methoden zur Sicherung der statistischen Geheimhaltung, Schriftenreihe „Forum der Bundesstatistik", Bd. 31, S. 133 – 149, hrsg. vom Statistischen Bundesamt, Wiesbaden.

Kommision zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik (Hrsg., 2001): Wege zu einer besseren informationellen Infrastruktur. Gutachten der vom Bundesministerium für Bildung und Forschung eingesetzten Kommission zur Verbesserung der informationellen Infrastruktur zwischen Wissenschaft und Statistik, Baden-Baden.

Lambert, D. (1993): Measures of Disclosure Risk and Harm, in: Statistics Sweden, Journal of Official Statistics, Vol. 9, No. 2, S. 313 – 331.

Lechner, S., Pohlmeier, W. (2003): Schätzung ökonometrischer Modelle auf der Grundlage anonymisierter Daten, in: Gnoss, R. und Ronning, G. (Hrsg.):, Anonymisierung wirtschaftsstatistischer Einzeldaten, Schriftenreihe „Forum der Bundesstatistik", Bd. 42, S. 115 – 137, hrsg. vom Statistischen Bundesamt, Wiesbaden.

Lenz, R. (2003): Disclosure of confidential information by means of multi objective optimizations (CD-ROM Publikation der "Comparative Analysis of (micro) Enterprise Data Conference" CAED 2003), 15. – 16. September 2003, London.
www.statistics.gov.uk/events/caed/abstracts/lenz.asp

Lenz, R., Sturm,R., Vorgrimler, D. (2004): Maße für die faktische Anonymität von Mikrodaten, in: Wirtschaft und Statistik, 6/2004, S. 621 – 638.

Mateo-Sanz, J. M. und Domingo-Ferrer, J. (1998a): A Comparative Study of Microaggregation Methods. *www.etse.urv.es/~jdomingo/*

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

Moore, R. A. Jr. (1996): Controlled data-swapping techniques for masking public use microdata sets, U.S. Bureau of the Census, Washington, DC (manuscript published only by internet).

Müller, W., Blien, U., Knoche, P., Wirth, H., u. a. (1991): Die faktische Anonymität von Mikrodaten, in: Schriftenreihe „Forum der Bundesstatistik", Bd. 19, hrsg. vom Statistischen Bundesamt, Wiesbaden.

Muralidhar, K. und Sarathy, R. (1999): Security of Random Data Pertubation Methods in ACM Transactions on Database Systems, Vol. 24, No. 4, December 1999.

Muralidhar, K. und Sarathy, R. (2003): A Theoretical Basis for Pertubation Methods in Statistics and Computing, Vol. 13, Kluwer Academic Publishers.

Oganian, A. und Domingo-Ferrer, J. (2001): On the complexity of optimal microaggregation for Statistical Disclosure Control, Universitat Rovira I Virgili, Dept. of Computer Engineering and Mathematics, Tarragona, Catalonia, Spain. (Arbeitspapier im Rahmen des CASC-Projektes präsentiert auf der „Joint ECE/Eurostat Work Session on Statistical Data Confidentiality" in Skopje 14. – 16. März 2001.)

Raghunatan, T., Reiter, J., Rubin, D. (2003): Multiple Imputation for Statistical Disclosure Limitation, in: Journal of Statistics, Bd. 19, S. 1 – 16.

Repsilber, D. (1999): Das Quaderverfahren, in: Methoden zur Sicherung der statistischen Geheimhaltung, Schriftenreihe „Forum der Bundesstatistik", Bd. 31, S. 150 – 157, hrsg. vom Statistischen Bundesamt, Wiesbaden.

Ronning, G. (2004): Mischung von Verteilungen und Anonymisierung, Wirtschaftswissenschaftliche Fakultät der Eberhard-Karls-Universität Tübingen (unveröffentlichtes Arbeitspapier zum Projekt Faktische Anonymisierung von wirtschaftsstatistischen Einzeldaten).

Ronning, G. (2005): Additive und multiplikative stochastische Überlagerungen, Wirtschaftswissenschaftliche Fakultät der Eberhard-Karls-Universität Tübingen (unveröffentlichtes Arbeitspapier zum Projekt Faktische Anonymisierung von wirtschaftsstatistischen Einzeldaten).

Ronning, G. (2009): Stochastische Überlagerungen mit Hilfe der Mischungsverteilung. Schätzung linearer (Panel-)Modelle auf Basis anonymisierter Daten. IAW Discussion Paper 48, March 2009.
www.iaw.edu/iaw/EN:Publications:IAW-Series:IAW-Discussion-Papers

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

Ronning, G., Sturm, R., Höhne, J., Lenz, R., Rosemann, M., Scheffler, M., Vorgrimler, D.: (2005): Handbuch zur Anonymisierung wirtschaftsstatistischer Mikrodaten, in: Schriftenreihe „Statistik und Wissenschaft", Bd. 4, hrsg. vom Statistischen Bundesamt, Wiesbaden.

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

Rosemann, M. (2006): Auswirkungen datenverändernder Anonymisierungsverfahren auf die Analyse von Mikrodaten, Dissertation, Institut für Angewandte Wirtschaftsforschung, Tübingen.

Rubin, D. (1993): Discussion. Statistical Disclosure Limitation, in: Journal of Official Statistics, Bd. 9(2), S. 461 – 468.

Rubin, D., Schenker, N. (1991): Multiple Imputation in Health-Care Databases: An Overview and Some Applications, in: Statistics in Medicine, Bd. 10, S. 585 – 598.

Schneeweiß, H. (2005): Ökonometrie, 4. überarbeitete Auflage, Heidelberg.

Schmid, M. (2007): Estimation of a Linear Regression with Microaggregated Data. Dissertation, Universität München, Verlag Dr. Hut, München, ISBN: 978-3-89963-507-2.

Söderberg, L.-J. (2005): MONA – Microdata on-line Access at Statistics Sweden, in: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Genf, 9 – 11 November 2005.

Statistische Ämter des Bundes und der Länder und Institut für angewandte Wirtschaftsforschung – IAW (2003): Forschungsprojekt: „Faktische Anonymisierung wirtschaftsstatistischer Einzeldaten" – Zwischenbericht an das Bundesministerium für Bildung und Forschung (BMBF), Statistisches Bundesamt, Wiesbaden.

Steel, P. und Reznek, A. (2005): Issues in Designing a Confidentiality Preserving Model Server, in: Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Geneva, 9 – 11 November 2005.

Sullivan, G. (1989): The Use of Added Error to Avoid Disclosure in Microdata Releases. Unpublished PhD Thesis, Iowa State University.

Torra, V. (2004): Microaggregation for Categorical Variables: A Median Based Approach, in: Privacy in Statistical Databases. Lecture Notes in Computer Science 3050, S. 162 – 174, Springer.

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

Wirth, H. (2003): Szenarien für Angriffe auf wirtschaftsstatistische Einzeldaten – Ein Überblick, in: Gnoss, R. und Ronning, G. (Hrsg.): Anonymisierung wirtschaftsstatistischer Einzeldaten, Schriftenreihe „Forum der Bundesstatistik", Bd. 42, S. 11 – 24, hrsg. vom Statistischen Bundesamt, Wiesbaden.

de Wolf P.-P., Gouweleeuw, J. M., Kooimann P. und Willenborg L.C.R.J. (1998): Reflections on PRAM, in: Proceedings of SDP'98, Amsterdam, IOS Press.

Yancey, W. E. (2002): Working Papers for Mixture Model Additive Noise for Microdata Masking, Research Report Series (Statistics #2002-03), Statistical Research Division U.S. Bureau of the Census, Washington, DC, 20233.

Anhang

1 Eindimensionale Mikroaggregationsverfahren mit variabler Gruppengröße nach Hansen und Mukherjee

Dieser Abschnitt enthält eine ausführlichere Darstellung des Verfahrens zur Bestimmung der optimalen Teilung des Datenbestandes in Mikroaggregationsgruppen mit variabler Größe von Hansen und Mukherjee (siehe Hansen und Mukherjee 2003).

Ziel ist die Bestimmung einer optimalen Teilung des Datenbestandes bei eindimensionaler Sortierung und variabler Gruppengröße.

Es sei $X^0=\{x_1^0,x_2^0,x_3^0,...,x_n^0\}^T$ die Matrix der sortierten Originaldaten, die aus n Zeilen (Anzahl der Objekte) und m Spalten (Anzahl der Merkmale) besteht. Welches Kriterium bei der Sortierung zugrunde gelegt wurde (siehe S. 50 f.) ist für die weitere Betrachtung nicht entscheidend. Gegeben sei weiterhin ein Sicherheitsparameter k, der die Anzahl der in einer Mikroaggregationsgruppe enthaltenen Objekte angibt (üblicherweise ist $k\geq 3$). Eine k-Teilung $T=\{G_1,G_2,G_3,...,G_{g(T)}\}$ von X ist eine Aufteilung der einzelnen Objekte x_i^0 der Originaldaten auf die Gruppen G_i bei der die Größe der einzelnen Gruppen $|G_i|$ (mit $1\leq i\leq g(T)$) mindestens k beträgt. Innerhalb der Gruppen G_i sollen sich jedoch nur in der Sortierung benachbarte Objekte befinden.

Es sei $\boldsymbol{T}_k(X^0)$ die Menge aller möglichen k-Teilungen der Originaldaten X^0. Die optimale eindimensionale Mikroaggregation besteht im Finden derjenigen k-Teilung der Originaldaten $\boldsymbol{T}_k^{opt}(X^0)$, die die Summe der quadrierten euklidischen Abstände der x_i^0 vom Zentroiden der jeweiligen Mikroaggregationsgruppe minimiert.

Als Formel lautet das Problem somit:

$$T_k^{opt}(X^o)=\min_{T\in T_k}\sum_{i=1}^{g(T)}\sum_{x_j^o\in G_i}\left\|x_j^o-\overline{x_{G_i}}\right\|^2$$

mit dem Zentroiden $\overline{x_{G_i}}$: **(1.1 – 1)**

$$\overline{x_{G_i}}=\frac{\sum_{x_j^o\in G_i}x_j^o}{|G_i|}$$

und

- $|G_i|$ – Größe der Gruppe i (Anzahl Objekte in der Gruppe)
- $\|.\|$ – Euklidische Norm

Während Oganian, A. und Domingo-Ferrer, J. beweisen, dass das Problem für unsortierte Datenbestände so komplex ist, dass sich der Aufwand zum Finden der exakten Lösung polynomial zum Produkt aus der Anzahl der Objekte und der Anzahl für möglichen Gruppen verhält, lässt sich das Problem für sortierte Datenbestände exakt lösen.

Dadurch dass bei eindimensionaler Mikroaggregation nur in der Sortierung unmittelbar benachbarte Objekte zu einer Gruppe zusammengefasst werden dürfen, besteht für jede Mikroaggregationsgruppe die Möglichkeit den Datenbestand zu teilen in einen Teil X^A bestehend aus den Objekten der Gruppe G_t und allen in der Sortierreihenfolge kleineren Objekte des Datenbestandes und einen Teil X^B der größeren (restlichen) Objekte.

Theorem 1

Wenn die Gruppe G_t eine gebildete Gruppe der optimalen Teilung des Gesamtbestandes X ist, dann sind auch die Teilungen $T^A=\{G_1,G_2,G_3,...,G_t\}$ und $T^B=\{G_{t+1},G_{t+2},G_{t+3},...,G_{g(T)}$ die optimalen Teilungen von X^A und X^B.

Beweis

Es bezeichne F(T) die Summe der quadrierten euklidischen Normen in den Mikroaggregationsgruppen bei der Teilung T.

$$F(T)=\sum_{i=1}^{g(T)}\sum_{x_j^o\in G_i}\left\|x_j^o-\overline{x_{G_i}}\right\|^2$$

$\boldsymbol{T}_k^*(X^A)$ sei beliebige aber nicht optimale k Teilung von X^A . Dann gilt: (siehe 1.1 – 1)

$$F(\boldsymbol{T}_k^*(X^A))> F(\boldsymbol{T}_k^{opt}(X^A))$$

Jede k-Teilung $\boldsymbol{T}_k(X)$ des Gesamtbestandes X, die für den Teilbestand X^A in diese Teilung $\boldsymbol{T}_k^*(X^A)$ verwendet, ist dann ebenfalls keine optimale Teilung von X. Denn mit:

$$F(\boldsymbol{T}_k(X))=F(\boldsymbol{T}_k^*(X^A))+F(\boldsymbol{T}_k(X^B))> F(\boldsymbol{T}_k^{opt}(X^A)) +F(\boldsymbol{T}_k(X^B))$$

kann $F(\boldsymbol{T}_k(X))$ bei Verwendung der Teilung $\boldsymbol{T}_k^*(X^A)$ somit nicht das Minimum der Aufgabe 1.1 – 1 sein.

Der Beweis für die Teilung T^B ist analog.

Auf Grund von Theorem 1 und der Erkenntnis, dass eine optimale Gruppengröße für eine k-Teilung von k bis maximal 2k-1 ist (siehe S. 48), lässt sich für jeden Teilbestand der sortierten Daten aus den ersten h Objekten des Datenbestandes $X^0{}_h=\{x_1{}^0,x_2{}^0,x_3{}^0,...,x_h{}^0\}^T$ die optimale Teilung $\boldsymbol{T}_k^{opt}(X^0{}_h)$ bestimmen als:

$$\mathrm{T}_k^{opt}(X_h^o)=\min_{i=k,...,2k-1}\left(\sum_{j=0}^{i-1}\left\|x_{h-j}^o-\frac{1}{i}\sum_{j=0}^{i-1}x_{h-j}^o\right\|^2+F(\mathrm{T}_k^{opt}(X_{h-i}^o))\right) \qquad \textbf{(1.1 – 2)}$$

Damit müssen nur die k-1 möglichen Gruppierungen untersucht werden, mit denen $x_h{}^0$ als letztes Objekt im Datenbestand zusammengefasst werden könnte. Die quadrierte euklidische Norm für diese Gruppe und für die optimale Teilung für die übrigen Objekte bis h reichen dann aus, um die optimale Teilung der ersten h Objekte zu bestimmen.

Hansen und Mukherjee schlagen deshalb vor, die optimale Teilung rekursiv zu bestimmen. Dazu werden zwei Vektoren der Größe n für das Verfahren definiert. Der erste Vektor L enthalte in den Elementen l_h die Größe der letzten Gruppe (Anzahl Objekte) bei

optimaler Teilung des Datenbestandes bis zum Objekt h. Der Vektor F enthalte in den Elementen f_h die Summe der euklidischen Normen für diese optimale Teilung. Dann berechnen sich die Elemente l_h und f_h wie folgt:

1. $l_h = h$;h=k,...,2k-1

$$f_h = \sum_{j=1}^{h} \left\| x_j^o - \frac{1}{h} \sum_{j=1}^{i} x_j^o \right\|^2 \quad ;h=k,...,2k-1$$

2. Für die weiteren h=2k,2k+1,...,3k-1 berechnen sich die Elemente l_h und f_h:

$$f_h = \min_i \left(\sum_{j=0}^{i-1} \left\| x_{h-j}^o - \frac{1}{i} \sum_{j=0}^{i-1} x_{h-j}^o \right\|^2 + f_{h-i})) \right) \quad ; \forall\, i \mid i \geq k \wedge (h-i) \geq k$$

(Bei kleinen Teilbeständen können nur solche Teilungen untersucht werden, die gewährleisten, dass beide Gruppen mindestens k Elemente besitzen.)

3. Für die weiteren h=3k,3k+1,...,n berechnen sich die Elemente l_h und f_h:

$$f_h = \min_{i=k,...,2k-1} \left(\sum_{j=0}^{i-1} \left\| x_{h-j}^o - \frac{1}{i} \sum_{j=0}^{i-1} x_{h-j}^o \right\|^2 + f_{h-i})) \right)$$

4. Abschließend lassen sich die Mikroaggregationsgruppen der optimalen Teilung aus dem Vektor L über folgende Regel ablesen.
   Setze k=n und i=1. Dann umfasst die Gruppe G_i die Objekte:

   $$x_k^o, x_{k-1}^o, ..., x_{k-l_k+1}^o$$

   Setze $k=k-l_k$ und i=i+1. und ermittle die Objekte der nächsten Gruppe G_i mit der gleichen Zuordnungsregel.

Die so ermittelte Teilung des Datenbestandes stellt die optimale eindimensionale Teilung des Datenbestandes dar.

2 Fehler auf den Korrelationskoeffizienten, der durch spaltenweise unabhängige Anonymisierungsverfahren mit Erhalt der Mittelwerte und Varianzen erzeugt wird

Die Bedingungen des Erhalts der Mittelwerte und der Varianzen sowie der unabhängigen Anonymisierung der Merkmale sind bei Rankswappingverfahren (siehe Abschnitt 2.2.3) aber auch bei spaltenweiser varianzerhaltender Mikroaggregation (siehe Abschnitt 3.3) gegeben. Die Beweisführung wurde in Moore (1996) hergeleitet.

Zusammenfassung

Die Werte $x_{ij}{}^a$ seien die anonymisierten Werte von $x_{ij}{}^o$ und analog $x_{ik}{}^a$ seien die anonymisierten Werte von $x_{ik}{}^o$. Das Anonymisierungsverfahren erhalte die Mittelwerte ($\Sigma_i x_{ij}{}^a = \Sigma_i x_{ij}$) und die Varianzen des Datenbestandes ($\sigma(x_{.j}{}^a) = \sigma(x_{.j}{}^o)$) und werde unabhängig auf die Spalten j und k angewendet. Weiterhin sei $R(x_{.j}{}^o, x_{.j}{}^a)$ der Korrelationskoeffizient zwischen den Wertepaaren $x_{ij}{}^o$ und $x_{ij}{}^a$ und $R(x_{.k}{}^o, x_{.k}{}^a)$ der Korrelationskoeffizient zwischen den Wertepaaren $x_{ik}{}^o$ und $x_{ik}{}^a$. Wenn $R(x_{.j}{}^o, x_{.j}{}^a)$ und $R(x_{.k}{}^o, x_{.k}{}^a)$ näherungsweise 1 sind, gilt:

$$E[COV(x_{.j}{}^a, x_{.k}{}^a)] = R(x_{.j}{}^o, x_{.j}{}^a) * R(x_{.k}{}^o, x_{.k}{}^a) * COV(x_{.j}{}^o, x_{.k}{}^o)$$

Lemma A1

Es seien für $x_{ij} = m * x_{ik} + c + e_i$, m und c so zu wählen, dass sie

$$\Sigma_i e_i^2$$

minimieren, dann gilt:

$$\sum_{i=0}^{n} e_i \left(x_{ik} - \overline{x_{.k}} \right) = 0$$

mit

$$\overline{x_{.k}} = \frac{1}{n} \sum_{i=0}^{n} \left(x_{ik} \right)$$ – Mittelwert der Werte in der Datenspalte k

und

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

Beweis

Es sei

$$z = \sum_{i=0}^{n} e_i^2 = \sum_{i=0}^{n} (x_{ij} - m * x_{ik} - c)^2 \quad .$$

Man bestimme die partielle Ableitung von z nach m und die partielle Ableitung von z nach c. Setzt man beide gleich 0 und löst sie gemeinsam als Gleichungssystem, erhält man die beiden obigen Gleichungen.

Theorem 2

Es seien x_{ij}^{a} die nach den obigen Bedingungen (Mittelwert- und Varianzerhalt) anonymisierten Werte von x_{ij}^{o}. Dann sind für

$$\left(x_{ij}^{a}-\overline{x_{.j}^{a}}\right)=m*\left(x_{ij}^{o}-\overline{x_{.j}^{o}}\right)+c+e_i$$

1) $m = R(x_{.j}^{o},x_{.j}^{a})$ der Korrelationskoeffizient zwischen $x_{.j}^{o}$ und $x_{.j}^{a}$, und

2) $c = 0$.

Beweis

Es sei

$$z=\sum_{i=0}^{n}e_i^2=\sum_{i=0}^{n}\left[\left(x_{ij}^{a}-\overline{x_{.j}^{a}}\right)-m*\left(x_{ij}^{o}-\overline{x_{.j}^{o}}\right)-c\right]^2 \quad .$$

Die partielle Ableitung nach c und Division durch -2 ergibt:

$$\frac{z_c}{-2}=\sum_{i=0}^{n}\left(x_{ij}^{a}-\overline{x_{.j}^{a}}\right)-m*\sum_{i=0}^{n}\left(x_{ij}^{o}-\overline{x_{.j}^{o}}\right)-nc \quad .$$

Die ersten beiden Summanden auf der rechten Seite sind 0. Deshalb bedingt $z_c=0$, dass auch $c=0$.

Die partielle Ableitung nach m und Division durch 2n ergibt:

$$\frac{z_m}{2n}=\frac{1}{n}\sum_{i=0}^{n}\left(x_{ij}^{a}-\overline{x_{.j}^{a}}\right)\left(x_{ij}^{o}-\overline{x_{.j}^{o}}\right)-m\frac{1}{n}\sum_{i=0}^{n}\left(x_{ij}^{o}-\overline{x_{.j}^{o}}\right)^2+\frac{c}{n}\sum_{i=0}^{n}\left(x_{ij}^{o}-\overline{x_{.j}^{o}}\right) \quad .$$

Der letzte Term auf der rechten Seite ist 0. Wenn $z_m=0$ und $c=0$, müssen der erste und der zweite Term gleich sein. Dass impliziert, dass die $COV(x_{.j}^{o},x_{.j}^{a})=mVar(x_{.j}^{o})$. Nach Division beider Seiten durch $Var(x_{.j}^{o})$, erhält man das Ergebnis $R(x_{.j}^{o},x_{.j}^{a})=m$.

Deshalb ist die am besten angepasste Gerade:

$$\left(x_{ij}^{a}-\overline{x_{.j}^{a}}\right)=R(x_{.j}^{o},x_{.j}^{a})*\left(x_{ij}^{o}-\overline{x_{.j}^{o}}\right)+e_i$$

Theorem 3

Angenommen, es seien die Werte x_{ij}^{a} die nach den obigen Bedingungen anonymisierten Werte von x_{ij}^{o} und die Werte x_{ik}^{o} nicht anonymisiert. Dann gilt:

$$COV(x_{.j}^{a},x_{.k}^{o})=R(x_{.j}^{o},x_{.j}^{a})*COV(x_{.j}^{o},x_{.k}^{o})+\frac{1}{n}\sum_{i=1}^{n}e_i\left(x_{ik}^{o}-\overline{x_{.k}^{o}}\right)$$

Beweis

$$COV(x_{.j}^{a},x_{.k}^{o})=\frac{1}{n}\sum_{i=1}^{n}\left(x_{ij}^{a}-\overline{x_{.j}^{a}}\right)\left(x_{ik}^{o}-\overline{x_{.k}^{o}}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[R(x^o_{.j}, x^a_{.j}) * \left(x^o_{ij} - \overline{x^o_{.j}}\right) + e_i\right]\left(x^o_{ik} - \overline{x^o_{.k}}\right)$$

$$= R(x^o_{.j}, x^a_{.j})COV(x^o_{.j}, x^o_{.k}) + \frac{1}{n}\sum_{i=1}^{n} e_i \left(x^o_{ik} - \overline{x^o_{.k}}\right)$$

Wird neben der Merkmalsspalte *j* auch die Anonymisierung der Merkmalsspalte *k* betrachtet, so ergibt sich analog zu Theorem 2 und 3:

Theorem 4

Angenommen, es seien die Werte $x_{ij}{}^a$ die nach den obigen Bedingungen anonymisierten Werte von $x_{ij}{}^o$ und die Werte $x_{ik}{}^a$ die anonymisierten Werte von $x_{ik}{}^o$. Weiterhin sei:

$$\left(x^a_{ij} - \overline{x^a_{.j}}\right) = R(x^o_{.j}, x^a_{.j}) * \left(x^o_{ij} - \overline{x^o_{.j}}\right) + e_i$$

$$\left(x^a_{ik} - \overline{x^a_{.k}}\right) = R(x^o_{.k}, x^a_{.k}) * \left(x^o_{ik} - \overline{x^o_{.k}}\right) + f_i$$

Dann gilt:

$$COV(x^a_{.j}, x^a_{.k}) = R(x^o_{.j}, x^a_{.j})R(x^o_{.k}, x^a_{.k})COV(x^o_{.j}, x^o_{.k}) + \frac{1}{n}\sum_{i=1}^{n} e_i \left(x^a_{ik} - \overline{x^a_{.k}}\right) + \frac{1}{n}\sum_{i=1}^{n} f_i \left(x^a_{ij} - \overline{x^a_{.j}}\right) - \frac{1}{n}\sum_{i=1}^{n} e_i f_i$$

Beweis

Aus Theorem 3 folgt:

$$COV(x^a_{.j}, x^a_{.k}) = R(x^o_{.j}, x^a_{.j})COV(x^o_{.j}, x^a_{.k}) + \frac{1}{n}\sum_{i=1}^{n} e_i \left(x^a_{ik} - \overline{x^a_{.k}}\right)$$

$$= R(x^o_{.j}, x^a_{.j})R(x^o_{.k}, x^a_{.k})COV(x^o_{.j}, x^o_{.k}) + R(x^o_{.j}, x^a_{.j})\frac{1}{n}\sum_{i=1}^{n} f_i \left(x^o_{ij} - \overline{x^o_{.j}}\right) + \frac{1}{n}\sum_{i=1}^{n} e_i \left(x^o_{ik} - \overline{x^o_{.k}}\right)$$

Setzt man jetzt

$$R(x^o_{.j}, x^a_{.j}) * \left(x^o_{ij} - \overline{x^o_{.j}}\right) = \left(x^a_{ij} - \overline{x^a_{.j}}\right) - e_i$$

ein, ergibt sich:

$$COV(x^a_{.j}, x^a_{.k}) = R(x^o_{.j}, x^a_{.j})R(x^o_{.k}, x^a_{.k})COV(x^o_{.j}, x^o_{.k}) + \frac{1}{n}\sum_{i=1}^{n} f_i \left(\left(x^a_{ij} - \overline{x^a_{.j}}\right) - e_i\right) + \frac{1}{n}\sum_{i=1}^{n} e_i \left(x^a_{ik} - \overline{x^a_{.k}}\right)$$

und somit

$$COV(x^a_{.j}, x^a_{.k}) = R(x^o_{.j}, x^a_{.j})R(x^o_{.k}, x^a_{.k})COV(x^o_{.j}, x^o_{.k}) +$$
$$\frac{1}{n}\sum_{i=1}^{n} f_i \left(x^a_{ij} - \overline{x^a_{.j}} \right) + \frac{1}{n}\sum_{i=1}^{n} e_i \left(x^a_{ik} - \overline{x^a_{.k}} \right) - \frac{1}{n}\sum_{i=1}^{n} f_i \; e_i$$

Da die Anonymisierung für beide Spalten unabhängig voneinander vorgenommen wurde, gilt $E(e_i)=0$, unabhängig von den Werten $x_{ik}{}^a$; und $E(f_i)=0$, unabhängig von den Werten $x_{ij}{}^a$. Die letzten 3 Terme im Theorem 4 haben deshalb den Erwartungswert 0.

Damit ergibt sich:

Theorem 5

Wenn die Werte der Spalten j und k unabhängig mit einem Mittelwert und Varianz erhaltendem Verfahren anonymisiert wurden, dann gilt:

$$E\left(COV(x^a_{.j}, x^a_{.k})\right) = R(x^o_{.j}, x^a_{.j})R(x^o_{.k}, x^a_{.k})COV(x^o_{.j}, x^o_{.k})$$

und wegen des Varianzerhalts ($Var(x_{.j}{}^o) = Var(x_{.j}{}^a)$ und $Var(x_{.k}{}^o) = Var(x_{.k}{}^a)$) gilt nach Division durch

$$\sqrt{VAR(x^o_{.j})VAR(x^o_{.k})}$$

auch

$$E\left(R(x^a_{.j}, x^a_{.k})\right) = R(x^o_{.j}, x^a_{.j})R(x^o_{.k}, x^a_{.k})R(x^o_{.j}, x^o_{.k}) .$$