Walter Müller, Uwe Blien, Peter Knoche, Heike Wirth
unter Mitarbeit von
Petra Beckmann, Stefan Bender, Thomas Helmcke
und Michael Müller

Die faktische Anonymität von Mikrodaten

Band 19 der Schriftenreihe
Forum der Bundesstatistik
herausgegeben vom
Statistischen Bundesamt

METZLER
POESCHEL

CIP-Titelaufnahme der Deutschen Bibliothek:
Die faktische Anonymität von Mikrodaten/
Hrsg.: Statistisches Bundesamt, Wiesbaden.
Walter Müller ... unter Mitarbeit von
Petra Beckmann ... – Stuttgart:
Metzler-Poeschel, 1991
(Schriftenreihe Forum der Bundesstatistik; Bd. 19)
ISBN 3-8246-0231-8

NE: Müller, Walter; Deutschland /
Statistisches Bundesamt; GT

Erschienen im Dezember 1991
Preis: DM 23,20
Bestellnummer 1030419-91900
ISBN 3-8246-0231-8
Verlagsauslieferung:
Hermann Leins GmbH & Co. KG
Holzwiesenstr. 2
7408 Kusterdingen

Copyright: Statistisches Bundesamt, Wiesbaden 1991
Alle Rechte vorbehalten. Es ist insbesondere nicht gestattet, ohne ausdrückliche Genehmigung des Statistischen Bundesamtes diese Veröffentlichung oder Teile daraus zu übersetzen, zu vervielfältigen, auf Mikro-film/-fiche zu verfilmen oder in elektronische Systeme einzuspeichern.

Geleitwort

Für die moderne amtliche Statistik war von Anbeginn an die Öffentlichkeit ihrer Ergebnisse selbstverständlich. Im vorigen Jahrhundert konvergierten dabei die Technologien der Erhebung und der Veröffentlichung: die "Tabelle" war Inbegriff dieser Konvergenz. Seit sich die Statistik - wie auch weite Bereiche der Gesellschaft einschließlich der Wissenschaft - der heutigen Informationsverarbeitung bedient, ist der Wunsch nach Öffentlichkeit statistischer Daten nicht mehr auf druckbare "Tabellen" beschränkt.

Für die allseits im Grundsatz begrüßte Freigabe anonymisierter Individualdaten, insbesondere für die wissenschaftliche Forschung, erwies sich jedoch ein Hindernis als fast unüberwindlich: eine ausreichend zuverlässige Anonymität der Daten und der Ausschluß einer Reidentifikation - mit anderen Worten, die Sicherung des Statistik-Geheimnisses, welches so alt ist wie die amtliche Statistik selbst (und neuerdings durch den "Datenschutz"-Gedanken Verstärkung erfährt).

In diesem Dilemma hatten sich die meisten Beteiligten viel von einem neuen Begriff versprochen, der eine realistische Lösung der gedanklichen Konflikte in Aussicht stellte: "faktische Anonymität", welche ausreichen sollte, wenn über die Weitergabe anonymisierter Individualdaten zu entscheiden war. Dieser Begriff "faktische Anonymität" teilte indes das Schicksal vieler neuer Begriffe: seine Bedeutung war, zumindest für die Anwendung, nicht klar genug.

So wird mit diesem Buch ein in mehrfacher Hinsicht ungewöhnliches Forschungsprojekt in seinen Ergebnissen dargestellt. Sein Ziel war die empirisch gestützte Präzisierung des Begriffs "faktische Anonymität", und es hat dieses Ziel erreicht.

Das Projektteam hat intensiv das untersucht, was ausgeschlossen werden soll: die Reidentifizierung. Dabei sind Sozialwissenschaftler an der Identifizierung einzelner Individuen in der Regel gar nicht interessiert, denn sie wollen Aussagen über Kollektive, nicht über einzelne Individuen machen; aber es ist für viele Fragen von entscheidender Bedeutung, ob und mit welchem Aufwand Reidentifizierung aus Datensätzen der amtlichen Statistik möglich ist. Zu diesem Thema überwog bisher die Spekulation und die Errechnung von Modellen mit unrealistischen Annahmen; der Irrglaube, Daten, die in einem Computer gespeichert seien, seien im allgemeinen auch so fehlerfrei, daß das

Wiederauffinden des "Datenschattens" eines Individuums mittels EDV problemlos sei, war weit verbreitet. Aber gesichertes Wissen zu diesem Thema gab es kaum.

So entstand aus einem mehrjährigen Diskussionszusammenhang von amtlicher Statistik und Sozialforschern die Idee, der Frage der Reidentifizierung nun endlich empirisch auf den Grund zu gehen. Selbstverständlich ist das Problem des Datenschutzes berührt, weswegen wir dankbar dafür sind, daß kundige Vertreter der Datenschutzbeauftragten in dem Beirat mitgewirkt haben, der das Projekt begleitet hat.

Niemandes Phantasie reicht aus, sich alles vorzustellen, was potentiellen Reidentifizierern, potentiellen Straftätern also, alles in den Sinn kommen könnte; kein einzelnes Forschungsprojekt kann alles, was nur denkbar ist, erschöpfend überprüfen. Der projektbegleitende Beirat hat deswegen sehr eingehend die Szenarien erörtert, welche realistischerweise in Betracht kommen und die zu untersuchen sind. Während die Ergebnisse des Forschungsvorhabens selbst von den Autoren zu vertreten sind, hat der Beirat darüber gewacht, daß die wesentlichen Annahmen, welche Voraussetzung der Untersuchung waren, als realistisch gelten können.

Mochte das Projekt auf den ersten Blick paradox erscheinen, spätestens auf den zweiten Blick schien das Vorhaben sogar zunächst an gesetzlichen Bestimmungen zu scheitern, welche die Paradoxie noch steigerten. Die Möglichkeit der Reidentifikation muß für eine Übermittlung von Daten der amtlichen Statistik ja ausgeschlossen sein, und die Kenntnis der Identitäten verbietet sich von selbst. So umfangreich auch die Datensätze waren, mit denen die Projektgruppe arbeitete: sie kannte Namen und Adressen der in den Datensätzen abgebildeten Personen nicht, und durfte sie nicht erfahren. Eine ingeniöse Lösung solcher Probleme ist bekannt: die Einschaltung eines "Treuhänders". Nur so kann man sicher sein, daß die vermeintlich "Identifizierten" auch wirklich identisch sind. Nur so kann man "Treffer" von "irrtümlichen Zuordnungen" unterscheiden, und beides zählen. Daß die Tätigkeit des "Treuhänders" in diesem Sinn äußerst wichtig, aber gänzlich frei von intrinsischer Gratifikation ist, versteht sich. Ohne diese Tätigkeit würde man wohl auch von "Identifikation" reden, aber es wäre bloß eine Annahme; wirklich wissen würde man es nicht. Ihm, ebenso wie der Projektgruppe, möchte ich an dieser Stelle den Dank des Beirats ausdrücken.

Die Ergebnisse des Projekts haben für eine keineswegs esoterische Frage, die für amtliche Statistik, Sozialforschung und Datenschutz von großer Bedeutung ist, eine empirisch gesicherte Grundlage geschaffen. Seine zum Teil überraschenden Ergebnisse geben auch künftiger Forschung eine neue Richtung.

Frankfurt am Main, im November 1991

Prof. Dr. Klaus Allerbeck
Vorsitzender des projekt-
begleitenden Beirats

Vorwort

Mikrodaten der amtlichen Statistik sind in den letzten Jahren eine zunehmend wichtigere Datenquelle in den Sozial- und Wirtschaftswissenschaften geworden. Voraussetzung ihrer Weitergabe an Hochschulen und sonstige Einrichtungen mit der Aufgabe unabhängiger Forschung ist nach dem Bundesstatistikgesetz von 1987 die faktische Anonymität der Daten. Unter welchen Bedingungen ist faktische Anonymität gegeben und wie kann sie in einer Weise hergestellt werden, die sowohl im Sinne des Datenschutzes wirksam ist als auch den Interessen der amtlichen Statistik und der Forschung gerecht wird? Dies zu untersuchen, war das Ziel eines Forschungsprojektes, dessen Ergebnisse mit diesem Buch vorgelegt werden.

Viele Institutionen und Personen haben an diesem Projekt mitgewirkt. Im Bericht ist dieses nicht mehr sichtbar. Deshalb soll an dieser Stelle den vielen Beteiligten gedankt werden, ohne deren Beiträge die erzielten Ergebnisse nicht zustande gekommen wären.

Das Anonymisierungsprojekt konnte durchgeführt werden, weil das Bundesministerium für Forschung und Technologie (BMFT) dafür großzügig Personal- und Sachmittel zur Verfügung gestellt hat. Eine erste Förderungsphase dauerte vom 1. April 1988 bis zum 30. September 1989. Auf der Basis eines Zwischenberichtes wurde das Projekt bis Ende Juni 1990 weitergefördert. Wenn als Ergebnis des Projektes der Zugang der Wissenschaften zu Mikrodaten der amtlichen Statistik wesentlich erleichtert und erweitert wird und sich dadurch die Forschungsmöglichkeiten wesentlich verbessern, dann hat die Unterstützung, die das BMFT diesem Vorhaben gewährt hat, entscheidend dazu beigetragen. Der Dank hierfür an das BMFT richtet sich insbesondere an Harald Uhl für sein Interesse und die wohlwollende Geduld dem Projekt gegenüber sowie an Peter Kanzler für die Mitwirkung im projektbegleitenden Beirat.

Das Projekt selbst wurde unter der Leitung von Walter Müller in Kooperation des Statistischen Bundesamtes, des Zentrums für Umfragen, Methoden und Analysen ZUMA (e.V.), Mannheim, und des Lehrstuhls für Methoden der empirischen Sozialforschung und angewandte Soziologie der Universität Mannheim durchgeführt. Diese Institutionen haben einen wesentlichen Teil der Projektarbeiten durch Eigenmittel mitgefördert und insbesondere die abschließenden Arbeiten und die vorliegende Publikation ermöglicht. Sie haben über die gesamte Projektlaufzeit sehr kooperativ und aufgeschlossen zusammenge-

arbeitet. Eine außerordentlich hoch einzuschätzende Nebenfolge dieser Zusammenarbeit ist das im Verlaufe der Arbeiten und des intensiven Austausches zunehmend gewachsene Verständnis für die unterschiedlichen Wahrnehmungen, Bewertungen, Ziele und Interessenlagen der beteiligten Partner. Dadurch erst konnte ein abschließender Bericht zustandekommen, an dem Mitarbeiter aus jeder der beteiligten Institutionen mitgearbeitet haben und den die Autoren gemeinsam verantworten.

Die Mitarbeit von Walter Müller an dem Projekt erfolgte zum Teil während eines Aufenthalts als Fellow am Center for Advanced Study in the Behavioral Sciences, Stanford, CA., im akademischen Jahr 1989/90. Dafür ist sowohl dem Center für die vorzügliche Arbeitsumgebung wie auch der Stiftung Volkswagenwerk für ein Akademiestipendium der National Science Foundation, grant BNS87-00864, und der Spencer Foundation zu danken.

Die Arbeiten des Projektes wurden von einem Beirat begleitet. Ihm gehörten Vertreter der statistischen Ämter des Bundes und der Länder, der Datenschutzbeauftragten des Bundes und der Länder, der Wissenschaft und des BMFT an. Der Beirat hat eine Vielzahl entscheidender Anregungen gegeben und den Verlauf des Projektes wesentlich mitgeprägt. Sehr hilfreich und für den Erfolg des Projektes wichtig waren insbesondere seine Bemühungen, für die im Projekt durchgeführten diffizilen empirischen Untersuchungen datenschutzgerechte Lösungen zu finden. Der Grundhaltung sehr konstruktiver Unterstützung und der Bereitschaft der Beiratsmitglieder, ihre aus dem jeweils eigenen Erfahrungs- und Verantwortungsbereich stammenden Kenntnisse zum Nutzen der Ziele des Projektes einzubringen, ist es auch zu verdanken, daß Empfehlungen erarbeitet werden konnten, die schließlich von allen von der Weitergabe von Mikrodaten betroffenen Institutionen - den statistischen Ämtern, dem Datenschutz und der Wissenschaft - mitgetragen werden. Der große Anteil an den Arbeiten des Projektes kommt in den insgesamt sieben ganztägigen Sitzungen zum Ausdruck, an denen der Beirat auf der Basis umfangreicher Materialen die jeweils erzielten Arbeitsfortschritte diskutiert und die nächsten Schritte des Projektfortganges mitbeeinflußt hat.

Für das außerordentlich engagierte Mitwirken und die vielseitigen Beiträge, die sie zum Gelingen des Projektes geleistet haben, sei allen Mitgliedern des Beirates herzlich gedankt. Im einzelnen waren es, in Vertretung:

- des *Bundesministeriums für Forschung und Technologie* :

  Peter Kanzler (Bundesministerium für Forschung und Technologie, Bonn)

- *der statischen Ämter* :

  Rolf Deininger (Statistisches Landesamt Baden-Württemberg, stellvertretender Vorsitzender des Beirats)

  Werner Nowak (Statistisches Bundesamt, Wiesbaden)

  Wilhelm Vorschulte (Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen)

- *der Datenschutzbeauftragten des Bundes und der Länder* :

  Henning Kaysers (Amt des Bundesbeauftragten für den Datenschutz, Bonn; im Jahre 1988)

  Bertram Raum (Amt des Bundesbeauftragten für den Datenschutz, Bonn; in den Jahren 1989 und 1990)

  Peter Schneider (Amt des Landesbeauftragten für den Datenschutz, Baden-Württemberg)

  Jürgen Werner (Amt des Bundesbeauftragten für den Datenschutz, Bonn, von 1988-1989)

- *der Wissenschaft* :

  Klaus Allerbeck (Universität Frankfurt, Vorsitzender des Beirats)

  Richard Hauser (Universität Frankfurt)

  Ekkehard Mochmann (Zentralarchiv für empirische Sozialforschung, Köln)

  Peter Ph. Mohler (Zentrum für Umfragen, Methoden und Analysen, Mannheim)

  Brigitte Preißl (Deutsches Institut für Wirtschaftsforschung, Berlin)

Neben den Mitgliedern des Beirates ist für sachkundige Beratung, vor allem bei der Entwicklung von Anonymisierungsregeln für die EVS und das Regionalfile zu danken: Erika Schulz und Martin Gornich (Deutsches Institut für Wirtschaftsforschung, Berlin), Richard Hauser und Heinrich Schlomann (Universität Frankfurt), Jürgen Hertel (Statistisches Bundesamt, Berlin), Ferdinand Böltken (Bundesforschungsanstalt für Landeskunde und Raumordnung, Bonn - Bad Godesberg) und Bruno Schmaus (Amt für Statistik der Stadt Heidelberg). Wertvollen Rat in methodischen Fragen hat uns Willibald Nagel (Universität Konstanz) gegeben. Mit juristischen Expertisen haben Hans-Wolfgang Arndt und Joachim Schrey die Arbeit des Projektes weitergebracht. Von großer Be-

deutung war die Unterstützung, die Christian Alt, Walter Bien, Hans Bertram und Bernhard von Rosenbladt dem Projekt zukommen ließen. Für Beratungen zu Problemen der Neuroinformatik sei Bernhard Mechler und Hans-Günter Lindner gedankt.

Um die Zuordnungsexperimente durchführen zu können, die den innovativen Kern der im Projekt geleisteten empirischen Arbeit bilden, mußten Institutionen gefunden werden, die bereit waren, für diesen nicht alltäglichen Forschungszweck Daten zur Verfügung zu stellen. Für die Bereitschaft dieses Experiment zu ermöglichen und aufgeschlossen zu unterstützen, sind wir dem Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen und zwei sozialwissenschaftlichen Forschungsinstituten sehr zu Dank verpflichtet. Einen ebenso verbindlichen Dank schulden wir in diesem Zusammenhang Robert Steffen vom Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen für die sorgfältigen und sehr aufwendigen Arbeiten, die er für das Projekt als Treuhänder geleistet hat.

Zu besonderem Dank verpflichtet sind wir Gerhard Paaß und Udo Wauschkuhn (Gesellschaft für Mathematik und Datenverarbeitung, St. Augustin bei Bonn) für die Überlassung des von ihnen entwickelten Programmsystems zum Matching von Datensätzen auf der Basis eines diskriminanzanalytischen Zuordnungsverfahrens. Gerhard Paaß hat das Projekt außerdem in außergewöhnlichem Maße bei der Implementation des Verfahrens auf der dem Projekt zur Verfügung stehenden Rechenanlage und bei allen Problemen, die im Verlaufe der Nutzung dieses Programms auftraten, unterstützt.

Das Projekt hat in vielfältiger Weise profitiert von den Diskussionen, die wir im Mannheimer Kontext insbesondere mit verschiedenen Kolleginnen und Kollegen bei ZUMA führen konnten. Auch die kritischen Hinweise zu Entwürfen einzelner Teile dieses Berichts waren sehr hilfreich. Wir danken dafür besonders Siegfried Gabler, Peter Hartmann, Wolfgang Karle, Paul Lüttinger, Günther Rothe, Bernhard Schimpl-Neimanns, Achim Wackerow, Michael Wiedenbeck und Georgios Papastefanou, der maßgeblich auch bei der Entwicklung der Projektidee und der Vorbereitung der Projektanträge mitgewirkt hat.

Stefan Bender und Michael Müller haben einen wesentlichen Anteil daran, daß es gelungen ist, das Programmpaket von Paaß/Wauschkuhn auf der benutzten Rechenanlage zu implementieren, in seine Tiefen und Untiefen einzudringen, für die besonderen Bedürfnisse des Projektes zu adaptieren und es für die aufwendigen Zuordnungsexperimente zu nutzen. Wir können nur hoffen, daß sie das Gefühl des Erfolges als Lohn ihrer Mühen sehen können.

Schließlich danken wir den vielen Mitarbeiterinnen und Mitarbeitern, die uns bei den vielfältigen Alltagsarbeiten während der fast vierjährigen Laufzeit des Projektes unermüdlich geholfen haben: bei der Beschaffung von Literatur, der Aufbereitung von Daten, den umfangreichen Rechenarbeiten, bei der Erstellung und Korrektur der Manuskripte, bei der Administration und den Abrechnungen für das Projekt. In unterschiedlicher Funktion waren dabei beteiligt: Dagmar Groh, Dirk Hovemann, Helga Koné, Ingo Koßmann, Heike Lieser-Schwendler, Renate Ludwig, Holger Mahr, Irene Mettner, Aloys Müller, Thilo Ritz, Hans Joachim Schwamb, Christa Sutter und Cornelia Vogt. Nicht zuletzt danken wir den mit der Projektabwicklung befaßten Mitarbeiterinnen und Mitarbeitern der Verwaltung der Universität Mannheim und der Rechenzentren der Universitäten Mannheim und Heidelberg, die in zuvorkommender Weise unsere Sonderwünsche bei der Zuteilung von Rechenzeiten und EDV-Infrastruktur, wann immer möglich, erfüllt haben.

Die Verantwortung für die Aussagen dieses Buches und seine Unzulänglichkeiten tragen allein seine Autoren. Als Folge intensiver Diskussionen und gegenseitiger Überarbeitung von Entwürfen sind Ideen und Formulierungen nicht in jedem Fall einzelnen Autoren eindeutig zuzuordnen. Bei jedem Kapitel sind jedoch als Fußnote zum Titel die Namen der Personen vermerkt, denen ein Kapitel oder Teile davon hauptsächlich zuzurechnen sind.

November 1991

Walter Müller, Mannheim
Uwe Blien, Nürnberg
Peter Knoche, Wiesbaden
Heike Wirth, Mannheim

Inhalt

Geleitwort

Vorwort

Kurzfassung IX

1 Ziele und Schwerpunkte einer Untersuchung der faktischen Anonymität von Mikrodaten 1

1.1 Die faktische Anonymität von Mikrodaten und ihr Beitrag zum Datenschutz 2

1.2 Der Bedarf der Wissenschaft an Mikrodaten der amtlichen Statistik und die gesetzlichen Voraussetzungen ihrer Nutzung 6

1.3 Schwerpunkte der Untersuchung 9

2 Zum Stand der Forschung und zur gegenwärtigen Praxis der Datenweitergabe 15

2.1 Forschungsstand 15

2.1.1 Forschungsschwerpunkte im Inland 16
2.1.2 Forschungsschwerpunkte im Ausland 23
2.1.2.1 Allgemeine Beiträge zur Frage der Anonymität und Anonymisierung 23
2.1.2.2 Forschung über einzelne Anonymisierungsmaßnahmen . . 25
2.1.2.3 Verfahren zum einfachen Abgleich von Datenfiles 27
2.1.2.4 Forschungsarbeiten zum Anonymisierungsproblem in den nationalen statistischen Ämtern 28
2.1.2.5 Fazit 31

2.2 Die gegenwärtige Praxis der Datenweitergabe 31

2.2.1 Inland 31
2.2.2 Ausland 35

3 Methodische Grundlagen von Reidentifikationsversuchen 41

3.1 Grundprinzipien einer Reidentifikation 41

3.2 Technik des Abgleichs von Merkmalsausprägungen 50

3.2.1 Dateninkompatibilitäten und Doppelgängerproblematik 50

3.2.2 Zuordnungsmechanismen bei Techniken des Abgleichs von Merkmalsausprägungen 52
3.2.3 Zusammenfassung 56

3.3 Zuordnung durch Distanzminimierung 56

3.4 Diskriminanzanalytische Reidentifikationstechnik nach Paaß/Wauschkuhn 59

3.4.1 Konzeption der Methode 60
3.4.1.1 Der Bayessche Schätzansatz 67
3.4.1.2 Annahmen über die Verteilungsform von Datenunkompatibilitäten als Voraussetzung zur Anwendung der linearen Diskriminanzanalyse 70
3.4.1.3 Die Mahalanobisdistanz 73
3.4.1.4 Die Stichprobenproblematik 75
3.4.1.5 Zur Interpretation des Verfahrens 76
3.4.1.6 Zusammenfassung 79
3.4.2 Umsetzung der diskriminanzanalytischen Reidentifikationstechnik in einen Algorithmus 80
3.4.2.1 Das erste Schätzproblem: Die Bestimmung von p(zk) . . . 81
3.4.2.2 Das zweite Schätzproblem: Die Schätzung von p(z) 83
3.4.3 Die diskriminanzanalytische Reidentifikationstechnik und die faktische Anonymität 84

4 Bestimmungsfaktoren des Reidentifikationsrisikos 87

4.1 Ein allgemeines Konzept für das Reidentifikationsrisiko 89

4.1.1 Das Szenario der gezielten Suche 89
4.1.2 Das Reidentifikationsrisiko in Fischzugsszenarien und bei Wissen um die Zugehörigkeit zum Mikrodatenfile 93
4.1.3 Das Reidentifikationsrisiko bei verschiedenen Reidentifikationsstrategien und Informationsgrundlagen 95

4.2 Einzigartigkeit als Schlüssel zur Reidentifikation 99

4.2.1 Bestimmungsfaktoren der Einzigartigkeit von Ausprägungskombinationen 99
4.2.2 Die Wahrscheinlichkeit von Einzelfällen: empirische Befunde 102
4.2.3 Formale Modelle zur Schätzung der Wahrscheinlichkeit von Einzelfällen in der Population 107

4.3 Kompatibilität der Überschneidungsmerkmale 112

4.3.1 Datenfehler 115
4.3.2. Unterschiedlicher sachlicher und zeitlicher Bezug von Identifikationsfile und Mikrodatenfile 121

4.3.3 Zum Ausmaß und zur Bedeutung von Dateninkompatibilitäten für das Reidentifikationsrisiko 123
4.3.3.1 Empirische Befunde für sozialwissenschaftliche Erhebungen 123
4.3.3.2 Empirische Befunde für Erhebungen in der amtlichen Statistik 126
4.3.3.3 Quantitative Schätzung der Bedeutung von Dateninkompatibilitäten für das Reidentifikationsrisiko . . . 128
4.3.4 Folgerungen für die Erfolgsaussichten und den Aufwand von Reidentifikationsversuchen 131

5 Deanonymisierungsmotive 132

5.1 Die professionelle Handlungslogik der Sozialwissenschaften 133

5.1.1 Ziele der Sozialwissenschaften 134
5.1.2 Praxis und Methoden empirischer Untersuchungen in den Sozialwissenschaften 136
5.1.3 Zur Organisation der wissenschaftlichen Forschung und der Sozialisation von Forschern 140
5.1.4 Einige Sonderprobleme 144
5.1.4.1 Verwendung zur Stichprobenziehung 144
5.1.4.2 Synthese von Einzeldaten aus einem Mikrodatenfile und anderen Datenquellen 147
5.1.4.3 Methodenvergleich 148
5.1.5 Fazit 150

5.2 Nicht an der sozialwissenschaftlichen Handlungslogik orientierte Deanonymisierungsmotive 151

5.2.1 Der Stellenwert der außerberuflichen Motive 152
5.2.2 Charakterisierung der außerberuflichen Motive 153
5.2.3 Ausgewählte außerberufliche Motive 155
5.2.4 Der Nachweis der prinzipiellen Deanonymisierbarkeit eines Mikrodatenfile als eigenständiges wissenschaftliches Motiv . . 156

6 Das Zusatzwissen für Deanonymisierungsvorhaben im Bereich der Wissenschaft 158

6.1 Kompatibilität und Umfang von Mikrodatenfile und Identifikationsfile 161

6.1.1 Kompatibilität 161
6.1.2 Die Bedeutung der Kenntnis über die Zusammensetzung eines File 164

6.2 Zusatzwissen aus öffentlichen Registern und privaten Datensammlungen 165

6.2.1 Der Auflösungsgrad der Überschneidungsmerkmale 168

6.2.2 Der Zugang der Wissenschaft zu Daten aus öffentlichen Registern und privaten Datenquellen 174
6.2.3 Kumulation von Informationen aus verschiedenen Registern/privaten Datenquellen 181

6.3 Zusatzwissen aus Umfragen in den Sozialwissenschaften 183

6.3.1 Umfragen, bei denen Zugang zu personenbezogenen Daten besteht 184
6.3.1.1 Pretests 184
6.3.1.2 Eigene Haupterhebungen 185
6.3.2 Umfragen, bei denen kein Zugang zu personenbezogenen Daten besteht 186

6.4 Fazit 188

7 Allgemeine Kosten-Nutzen-Überlegungen zu Deanonymisierungsversuchen 212

7.1 Nutzenkomponenten eines Deanonymisierungsversuchs 213

7.1.1 Eingesparte Ausgaben 214
7.1.2 Nutzenkomponenten bei beruflicher Motivation 215
7.1.3 Nutzenkomponenten bei außerberuflicher Motivation 216
7.1.3.1 Publikationserlöse 216
7.1.3.2 Weitere monetäre Verwertungsmöglichkeiten 217
7.1.3.3 Nichtmonetäre Nutzenelemente 218

7.2 Kostenkomponenten eines Deanonymisierungsversuchs 220

7.2.1 Kosten zur Beschaffung von Zusatzwissen 220
7.2.2 Kosten zur Bearbeitung der Files 221
7.2.2.1 Übernahme auf maschinenlesbare Datenträger 221
7.2.2.2 Fehlerabschätzungen 222
7.2.2.3 Anpassung der Überschneidungsmerkmale 223
7.2.3 Programmentwicklung und -ablauf 224
7.2.4 Kosten zur Verschleierung von Deanonymisierungsversuchen 225
7.2.5 Rechtliche Sanktionen 225
7.2.6 Entgangener Nutzen 226

7.3 Alternative Datenbeschaffungsmethoden 227

7.3.1 Eigene Recherchen versus Identifikation 228
7.3.2 Eigene Recherchen versus Fischzug 230

8 Entwicklung konkreter Angriffsszenarien für die Operationalisierung der faktischen Anonymität und Datenbasis für die empirische Überprüfung einzelner Szenarien 233

8.1 Deanonymisierungsmotive und Zusatzwissen als Grundlage für die Entwicklung von Angriffsszenarien im Wissenschaftsbetrieb 235

8.1.1 Angriffsszenario 1: Erweiterung der Informationsbasis einer wissenschaftlichen Untersuchung 238
8.1.2 Angriffsszenario 2: Erweiterung der Information über einen Bekannten zur Befriedigung der persönlichen Neugier 239
8.1.3 Angriffsszenario 3: Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe 240
8.1.4 Angriffsszenario 4: Gewinnung von Informationen über eine Persönlichkeit des öffentlichen Lebens 240
8.1.5 Angriffsszenario 5: Gewinnung ökonomisch verwertbarer Informationen 241
8.1.6 Zusammenfassung 243

8.2 Datenbasis der empirischen Zuordnungsexperimente 243

8.2.1 Doppelt-Blind-Verfahren 243
8.2.2 Datenbasis von Angriffsszenario 1: Erweiterung der Informationsbasis einer wissenschaftlichen Untersuchung 246
8.2.2.1 Beschreibung der Datenbasis 246
8.2.2.2 Datenaufbereitung 249
8.2.3 Datenbasis von Angriffsszenario 2: Erweiterung der Informationen über einen Bekannten zur Befriedigung der persönlichen Neugier . . 255
8.2.3.1 Beschreibung der Datenbasis 255
8.2.3.2 Datenaufbereitung 259
8.2.4 Zusammenfassung 264

9 Empirische Überprüfung des Reidentifikationsrisikos 265

9.1 Zuordnungen zwischen Mikrozensus und "Kürschners Deutschem Gelehrtenkalender" mit einfacher Abgleichtechnik 266

9.1.1 Praktische Umsetzung der einfachen Reidentifikationstechnik 267
9.1.2 Berücksichtigung der alternativen Verkodung einzelner Merkmale im Gelehrtenkalender 270
9.1.3 Ergebnisse des Zuordnungsexperiments 273
9.1.4 Überprüfung der Ergebnisse durch den Treuhänder 277
9.1.5 Mögliche Ursachen für das Versagen der einfachen Abgleichtechnik 279
9.1.6 Folgerungen für die faktische Anonymität 280

9.2 Empirischer Test der Anonymität des Mikrozensus. Verwendung der diskriminanzanalytischen Reidentifikationstechnik von Paaß/Wauschkuhn und "Kürschners Deutschem Gelehrtenkalender" als Identifikationsfile . . . 283

9.2.1 Vorbereitung der Zuordnungsexperimente 284
9.2.1.1 Berücksichtigung der Alternativvariablen 285
9.2.1.2 Vorbereitung eines Massenfischzugs 287
9.2.1.3 Eingrenzung der Population 288
9.2.2 Generierung der Fehlerverteilungen 289
9.2.2.1 Nutzung vorhandener Informationsquellen 289
9.2.2.2 Alternative Annahmen über den Fehlerprozeß 290
9.2.2.3 Generierung der Fehlerfiles 292
9.2.3 Durchführung und Ergebnisse der Rechnungen 295
9.2.3.1 Zuordnung durch Distanzminimierung 295
9.2.3.2 Darstellung der Ergebnisse der Zuordnungsexperimente . . 296
9.2.3.3. Die Überprüfung der Ergebnisse durch den Treuhänder . . 300
9.2.3.4 Erklärung der erhaltenen Ergebnisse (I) 302
9.2.3.5 Vergleich mit den Ergebnissen einer Testrechnung bei Verwendung eines synthetischen Identifikationsfile 306
9.2.3.6 Erklärung der erhaltenen Ergebnisse (II) 308
9.2.4 Folgerungen für die faktische Anonymität von Daten 310
9.2.5 Ausblick 311

9.3 Zuordnung zwischen Mikrozensus und einer sozialwissenschaftlichen Stichprobe mit einer einfachen Abgleichtechnik 323

9.3.1 Untersuchungsdesign 325
9.3.2 Ergebnisse des Zuordnungsexperiments 329
9.3.3 Überprüfung der Ergebnisse durch den Treuhänder 332
9.3.4 Dateninkompatibilitäten als natürlicher Schutz vor Reidentifikationsversuchen 334
9.3.4.1 Mögliche Ursachen für die aufgetretenen Dateninkompatibilitäten 338
9.3.5 Weitergehende Überlegungen zu dem untersuchten Szenario 342
9.3.5.1 Regionalinformationen 343
9.3.5.2 Erfolgschancen bei Wissen um Teilnahme 344
9.3.6 Folgerungen für die faktische Anonymität von Mikrodaten 348

10 Argumentative Analyse einzelner Szenarien 351

10.1 Angriffsszenario 3: Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe 352

10.1.1 Beschreibung des Szenarios 352
10.1.2 Identifikations- und Mikrodatenfile 353
10.1.3 Abgrenzungs- und Fehlerproblematik 354
10.1.4 Praktische Durchführung 355
10.1.5 Kosten-Nutzen-Analyse 355
10.1.5.1 Deanonymisierungskosten 356

10.1.5.2 Alternative Datenbeschaffung 362
10.1.5.3 Deanonymisierungsnutzen 362
10.1.5.4 Fazit 363

10.2 Angriffsszenario 4: Gewinnung von Informationen über eine Persönlichkeit des öffentlichen Lebens mit dem Ziel der anonymen Publikation 364

10.2.1 Beschreibung des Szenarios 364
10.2.2 Identifikations- und Mikrodatenfile 365
10.2.3 Abgrenzungs- und Fehlerproblematik 366
10.2.4 Praktische Durchführung 366
10.2.5 Kosten-Nutzen-Analyse 366
10.2.5.1 Mögliche Vorgehensweise des Angreifers 366
10.2.5.2 Deanonymisierungskosten 368
10.2.5.3 Alternative Datenbeschaffung und Fazit 372

10.3 Angriffsszenario 5: Gewinnung ökonomisch verwertbarer Informationen aus dem Mikrozensus zum Verkauf an einen Adressenhändler 373

10.3.1 Beschreibung des Szenarios 373
10.3.2 Identifikations- und Mikrodatenfile 374
10.3.3 Abgrenzungs- und Fehlerproblematik 375
10.3.4 Praktische Durchführung 375
10.3.5 Kosten-Nutzen-Analyse 376
10.3.5.1 Deanonymisierungskosten 376
10.3.5.2 Alternative Datenbeschaffungsmethoden 378
10.3.5.3 Deanonymisierungsnutzen 381
10.3.5.4 Fazit 384

11 Anonymisierungsmaßnahmen 386

11.1 Kriterien für die Auswahl von Anonymisierungsmaßnahmen 388

11.2 Darstellung ausgewählter Anonymisierungsmaßnahmen 391

11.2.1 Maßnahmen, die das Einbringen von falschen Angaben in das Datenmaterial vorsehen 391
11.2.2 Maßnahmen mit Datenmodifikation ohne Einbringen von falschen Angaben in das Datenmaterial: Vergröberung von Merkmalen 394
11.2.2.1 Vergröberung von Regionalangaben 395
11.2.2.2 Vergröberung von schwach besetzten oder extremen Merkmalswerten 396
11.2.3 Maßnahmen, die keine direkte Datenmodifikation vorsehen 398
11.2.3.1 Entfernen einzelner Merkmale 398
11.2.3.2 Die Weitergabe veralteter Daten 398
11.2.3.3 Systemfreie Sortierung 399
11.2.3.4 Ziehung von zufälligen Substichproben 400

11.3 Empirische Überprüfung ausgewählter Anonymisierungsmaßnahmen 402

11.3.1 Vergröberung von Merkmalsausprägungen 403
11.3.2 Ziehung zufälliger Substichproben 406

11.4 Zusammenfassung 408

12 Technisch-organisatorische Sicherungsmaßnahmen gegen unbefugte Datenzugriffe bei faktisch anonymen Daten 409

12.1 Faktische Anonymität und Datensicherung 409

12.2 Datensicherungsmaßnahmen, die überwiegend der Abschottung gegenüber externen Personen dienen 413

12.3 Sicherungsmaßnahmen gegenüber berechtigten Personen 422

12.4 Eine 'Standardliste' an Sicherungsmaßnahmen 428

13 Zusammenfassende Bewertung der untersuchten Szenarien und Folgerungen für die Übermittlung faktisch anonymer Daten 431

13.1 Zusammenfassung und Bewertung der Untersuchungsergebnisse 431

13.2 Folgerungen für die Übermittlung faktisch anonymer Daten 436

14 Empfehlungen für die Weitergabe von Einzelangaben aus dem Mikrozensus und der EVS an die Wissenschaft 440

14.1 Allgemeine Schutzvorkehrungen 441

14.2 Datenfile-spezifische Anonymisierungsmaßnahmen 443

14.3 Begründung 445

14.3.1 Reidentifikation durch Massenfischzug 446
14.3.2 Gezielte Suche 449

Literaturverzeichnis 454

Ergebnisse des Anonymisierungsprojekts und Empfehlungen für die Weitergabe faktisch anonymer Mikrodaten an die Wissenschaft*

Kurzfassung

Nach §16 Abs.6 BStatG 1987 dürfen "für die Durchführung wissenschaftlicher Vorhaben (...) vom Statistischen Bundesamt und den statistischen Ämtern der Länder Einzelangaben an Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung übermittelt werden, wenn die Einzelangaben nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können und die Empfänger Amtsträger, für den öffentlichen Dienst besonders Verpflichtete oder Verpflichtete nach Absatz 7 sind" (Dorer/Mainusch/Tubies 1988:87f). Diese gesetzliche Definition entspricht dem Begriff der *faktischen* Anonymität. Sie hat die bis dahin geltende gesetzliche Regelung (§11 Abs.5 BStatG 1980) ergänzt, die eine Übermittlung von Einzelangaben an die Wissenschaft nur bei *absoluter* Anonymität vorsah und durch die der Nutzungsbedarf der Wissenschaft nicht befriedigt und das wissenschaftliche Potential der Daten auch nicht annäherend ausgeschöpft werden konnte.

Dieser Erkenntnis wird in §16 Abs.6 BStatG durch die Einführung des "Unverhältnismäßigkeitskriteriums" insofern entsprochen, als bei übermittelten Daten eine Deanonymisierung nicht mehr mit Sicherheit ausgeschlossen sein muß, sondern implizit ein Rationalkalkül eines Datenangreifers zwischen dem Wert, bzw. dem durch einen Reidentifikationsversuch erzielten Nutzen, und den hierfür anfallenden Kosten zugrundegelegt wird. Allerdings ist im Gesetzestext nicht näher konkretisiert, wie die faktische Anonymität eines konkreten Datenbestandes sichergestellt werden kann.

In Anknüpfung an eine von der amtlichen Statistik schon 1986 aufgegriffene Initiative hat ein in Kooperation zwischen dem Statistischen Bundesamt, der Universität Mannheim und dem Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim durchgeführtes Forschungsprojekt die Aufgabe übernommen, die gesetzliche Regelung der faktischen Anonymität mit dem Ziel zu operationalisieren, eine unter den Beteiligten konsensfähige Lösung für die

* Autoren: Walter Müller, Uwe Blien, Peter Knoche, Heike Wirth.

Übermittlung von Einzelangaben an die Wissenschaft zu finden.[2] Dazu hat wesentlich ein projektbegleitender Beirat beigetragen, in dem Vertreter der Statistischen Ämter des Bundes und der Länder, der Datenschutzbeauftragten von Bund und Ländern, der Wissenschaft und des Bundesministeriums für Forschung und Technologie zusammengearbeitet haben.

Durch Untersuchungen zahlreicher einschlägiger Teilaspekte wurde versucht, der Komplexität der Reidentifikationsproblematik gerecht zu werden und eine solide Basis für die praktische Umsetzung des §16 Abs.6 BStatG zu schaffen. Die zentralen Arbeiten des Anonymisierungsprojektes bezogen sich auf die Analyse des vorhandenen Risikos der Reidentifikation von an die Wissenschaft übermittelten anonymisierten Einzelangaben und auf die dabei anfallenden Kosten. Hierfür wurden zunächst umfassend die Randbedingungen des Wissenschaftsbereiches untersucht, die für einen hypothetischen Reidentifikationsversuch von Bedeutung sein könnten. Auf dieser Grundlage wurden fünf Szenarien entwickelt, für die das Reidentifikationsrisiko und der damit verbundene Aufwand eingehend überprüft wurden. Die Arbeiten haben ein weit gefaßtes Spektrum denkbarer Möglichkeiten von Deanonymisierungsrisiken berücksichtigt und ermöglichen konkrete Empfehlungen für die Weitergabe amtlicher Einzeldaten nach Maßgabe der faktischen Anonymität.

Das Grundprinzip einer Reidentifikation beruht darauf, daß Einzeldatensätze einer Datei mit anonymen Mikrodaten den Einzeldatensätzen einer anderen, personenbezogenen Datei (dem sogenannten Identifikationsfile, das das Zusatzwissen enthält und das nicht maschinenlesbar sein muß) in einer eins-zu-eins Entsprechung zugeordnet werden. Dabei setzen Deanonymisierungsversuche bei den Merkmalen an, die Zusatzwissen und Mikrodatenfile gemeinsam sind (Überschneidungsmerkmale). Durch einen Abgleich dieser Überschneidungsmerkmale auf Identität oder große Ähnlichkeit wird angestrebt, jene Datensätze in Zusatzwissen und Mikrodatenfile herauszufinden, die von ein- und derselben Person stammen .

Bei Deanonymisierungsversuchen können im wesentlichen zwei Strategien angewandt werden: Im Falle einer *gezielten Suche* sollen bestimmte, dem Angreifer bekannte Personen aus einem Mikrodatenfile deanonymisiert werden. Bei einem *Fischzug* dagegen werden nicht vorgebene, sondern beliebige Per-

---

[2] Vom 3. - 5. März 1986 fand im Statistischen Bundesamt das wissenschaftliche Kolloquium "Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik - Bedingungen und Möglichkeiten" (vgl. Statistisches Bundesamt 1987) statt. Die Finanzierung der Hauptkosten des Anonymisierungsprojektes (1988-1990) erfolgte durch das Bundesministerium für Forschung und Technologie.

sonen in einem Mikrodatenfile gesucht (Paaß/Wauschkuhn 1985). Beide Varianten wurden im einzelnen untersucht.

Bei den Reidentifikationstechniken wurden als wichtigste Verfahren einbezogen:

a) *Abgleichverfahren*: hierbei wird durch Sortier- oder Selektionsroutinen ein Abgleich der Überschneidungsmerkmale auf Identität durchgeführt.

b) *Diskriminanzanalytische Reidentifikationstechnik nach Paaß*: bei dieser Methode die auf der Diskriminanzanalyse beruht, aber sie mit zahlreichen anderen mathematisch-statistischen Verfahren zu einem neuen Algorithmus kombiniert, wird über Wahrscheinlichkeitsberechnungen die Doppelgängerproblematik und das Problem der Fehlerhaftigkeit von Daten mitberücksichtigt.

Bei der Übermittlung von Einzelangaben an die *Wissenschaft* ist für die Operationalisierung des Unverhältnismäßigkeitskriteriums maßgeblich von den im Bereich der *Wissenschaft* gegebenen nutzen- und kostenrelevanten Randbedingungen für Deanonymisierungsvorhaben auszugehen.

1) Um den potentiellen Nutzen von deanonymisierten Daten im *Wissenschafts*kontext zu ermitteln, wurde geprüft, welche Logik der *wissenschaftlichen* Datennutzung zugrundeliegt und welche Motive sich hieraus für einen Deanonymisierungsversuch ergeben könnten. Diese Analyse führte zu dem Ergebnis, welches auch schon das Bundesverfassungsgericht ausdrücklich anerkannt hat: "(...) der Wissenschaftler ist regelmäßig nicht an der einzelnen Person interessiert, sondern an dem Individuum als Träger bestimmter Merkmale" (BVerfG NJW 1984:428). Plausible Motive für Deanonymisierungsversuche sind aufgrund der beruflichen Interessenlage des Wissenschaftlers kaum überzeugend rekonstruierbar. Es erschien allenfalls als vorstellbar, daß aus wissenschaftlichem Interesse ein Mikrodatensatz deanonymisiert werden könnte, um eine informative Auswahlbasis für eine eigene Erhebung zu gewinnen oder um eine eigene Datenbasis um Informationen zu ergänzen, die in einem Mikrodatenfile enthalten sind. Ergänzend wurde jedoch auch der Nutzen eines Deanonymisierungsversuchs im Hinblick auf wissenschaftsfremde Motive geprüft. Hierbei sind von persönlicher Neugier über ökonomisch motivierte Deanonymisierungsversuche (beispielsweise der Verkauf von Adressen über bestimmte Personengruppen) bis hin zu eindeutig krimineller Motivation eine Vielzahl von Motiven hypothetisch vorstellbar.

2) Die Reidentifikation einer Person ist nur dann möglich, wenn ein Datenangreifer über entsprechendes Zusatzwissen verfügt. Deshalb wurde untersucht, welche Arten und Qualitäten von relevantem Zusatzwissen in den Sozialwissenschaften vorhanden oder typischerweise zugänglich sind. Im Detail wurden sowohl die als einschlägig erscheinenden öffentlich oder für einen beschränkten Personenkreis zugänglichen Register und private Datenquellen als auch die in den Sozialwissenschaften aus eigenen Erhebungen zur Verfügung stehenden Daten im Hinblick auf ihre Verwendbarkeit als Zusatzwissen analysiert.

3) Als eine weitere wichtige Komponente wurden die Bedingungen der Datenverarbeitung innerhalb des Wissenschaftsbereiches untersucht. Darauf aufbauend wurden Empfehlungen über erforderliche technisch-organisatorische Schutzmaßnahmen bei der Verarbeitung anonymisierter Einzelangaben erarbeitet, die sicherstellen sollen, daß die gelieferten Daten den Bereich der Wissenschaft nicht verlassen bzw. nur das datenempfangende Institut Zugriff auf die Mikrodaten hat. Diese Maßnahmen sind im übrigen auch insofern für die faktische Anonymität relevant, als sie den Aufwand eines Angreifers erhöhen können.

Im Hinblick auf die Gefährdungssituationen, die sich aus der Analyse des potentiellen Nutzens von deanonymisierten Daten und des im Wissenschaftsbereich potentiell vorhandenen oder erwerbbaren Zusatzwissens ergeben, wurden fünf Szenarien erarbeitet. Sie können als repräsentativ gelten für verschiedene - sowohl berufliche wie außerberufliche - Angriffsmotive und Deanonymisierungsstrategien. Drei dieser Szenarien wurden unter Zugrundelegung von plausiblen Randbedingungen einer argumentativen Aufwandsanalyse unterzogen. Für zwei Szenarien wurden das Reidentifikationsrisiko und der dabei entstehende Aufwand empirisch überprüft, indem die Vorgehensweise eines hypothetischen Datenangreifers nachvollzogen wurde. Diese Experimente weichen insofern von anderen Untersuchungen ab, als hier so realistisch wie möglich vorgegangen wurde. Sowohl Mikrodatenfile wie Identifikationsfile bestanden aus empirischen Daten, und ein Treuhänder hat überprüft, ob Datensätze aus dem Mikrodatenfile tatsächlich richtig den Personen zugeordnet wurden, von denen sie stammten.

Bei den empirischen Untersuchungen wurde als zu deanonymisierendes Mikrodatenfile der Mikrozensus von Nordrhein-Westfalen zugrunde gelegt. Die Datensätze des Mikrozensus waren im wesentlichen nur insoweit anonymisiert, als sie keine Namen und Anschriften enthielten. Sie enthielten vergleichsweise differenzierte Regionalinformationen (Bundesraumordnungsregion und

Gemeindegrößenklasse). Als Zusatzwissen wurde für ein erstes Angriffsszenario Kürschners Deutscher Gelehrtenkalender 1987 herangezogen. Damit wurde eine der denkbar riskantesten öffentlich zugänglichen Informationsquellen ausgewählt.[3] Nach bisherigen Annahmen mußte mit einem hohen Reidentifikationsrisiko, insbesondere auch wegen der großen Zahl einzigartig vertretener Ausprägungskombinationen, gerechnet werden.

Als zweite wichtige Quelle von Zusatzwissen wurde exemplarisch für die sozialwissenschaftlichen Datenbestände eine der umfangreichsten, gegenwärtig in den Sozialwissenschaften der Bundesrepublik durchgeführten, repräsentativen empirischen Erhebungen herangezogen. Auch diese Datenbasis mußte, insbesondere wegen der hohen Zahl der berücksichtigten Überschneidungsmerkmale zum Mikrozensus - sie enthielt insgesamt 35 vorwiegend erwerbs-, haushalts- und familienbezogene Überschneidungsmerkmale -, als riskant betrachtet werden.

Entgegen diesen Erwartungen fanden sich bei den rund 8 000 berücksichtigten Fällen des Gelehrtenkalenders mit einer einfachen Abgleichtechnik und bei Unkenntnis über die Teilnahme an der Mikrozensuserhebung lediglich 14 Fälle mit einzigartigen Ausprägungskombinationen, die im Mikrozensus ebenfalls eine gleiche einzigartige Ausprägungskombination aufwiesen. Die Überprüfung dieser Zuordnungen durch einen Treuhänder ergab, daß es sich hierbei nur in vier Fällen auch um eine identische Person handelte. Bezogen auf das Identifikationsfile ergab sich eine Reidentifikationswahrscheinlichkeit an korrekten Zuordnungen von 0,0005. Bei den sehr wenigen Fällen, bei welchen eine eindeutige Zuordnung vorgenommen werden konnte, war die Wahrscheinlichkeit einer Falschzuordnung mit 0,71 wesentlich höher als die Wahrscheinlichkeit einer korrekten Zuordnung. Von den 53 Fällen, die nach den Nachprüfungen des Treuhänders tatsächlich in beiden untersuchten Datensätzen (Mikrodatenfile und Identifikationsfile) enthalten waren, wurden nur vier korrekt als eindeutig identisch identifiziert.

---

3 Der Gelehrtenkalender enthält Informationen zu zehn teilweise sehr differenzierten Überschneidungsmerkmalen zum Mikrozensus, u.a. Geburtsjahr, Wirtschaftszweig, Beruf und Region. Diese Informationen dürften relativ fehlerfrei sein, da die in das Handbuch aufgenommenen Personen die Möglichkeit haben, die Angaben selbst zu überprüfen. Die Personen, über die im Gelehrtenkalender Zusatzwissen vorhanden ist, bilden eine durch ihre Erwerbstätigkeitsmerkmale deutlich von der Durchschnittsbevölkerung abgegrenzte Subpopulation, die zudem in sich stark differenziert ist.

Diese Quoten kamen erst dadurch zustande, daß die knapp 8 000 Fälle des Gelehrtenkalenders in 16 unterschiedlichen Verkodungsvarianten auf identische Ausprägungskombinationen geprüft wurden. Außerdem ist zu betonen, daß die vier korrekten Zuordnungen nur dann zu finden waren, wenn neben der Entfernung von Name und Anschrift keine Anonymisierungsmaßnahmen getroffen wurden. Wurden beispielsweise die Regionalinformationen auf die Ebene des Bundeslandes vergröbert, konnte trotz Einbezug der sehr differenzierten Merkmale Geburtsjahr, Beruf und Wirtschaftszweig kein Fall korrekt zugeordnet werden. Auch durch Vergröberungen bei anderen Merkmalen oder Substichprobenziehung wurde das Reidentifikationsrisiko deutlich kleiner und das Risiko einer Falschzuordnung nahm zu.

Bei der Verwendung der sehr viel aufwendigeren diskriminanzanalytischen Reidentifikationstechnik von Paaß/Wauschkuhn, die das Vorhandensein von Datenfehlern und Doppelgängern in der Grundgesamtheit berücksichtigt, war entgegen allen Erwartungen die Quote der Falschzuordnungen höher als bei Abgleichverfahren.

Beim Szenario mit der sozialwissenschaftlichen Erhebung enthielt das Identifikationsfile 2 685 Fälle. Davon wurden bei Berücksichtigung aller 35 Überschneidungsmerkmale zwar 35 eindeutige Zuordnungen zu einem Fall im Mikrodatenfile des Mikrozensus gefunden. Bei der Überprüfung durch den Treuhänder stellte sich jedoch heraus, daß es sich dabei in keinem einzigen Fall um eine in Wirklichkeit identische Person handelte. Von den nach den Feststellungen des Treuhänders sowohl im Identifikationsfile wie im Mikrodatenfile gleichzeitig enthaltenen zehn Personen konnte in den Experimenten keine einzige gefunden werden.

Nach all diesen Szenarien überstieg der Aufwand, der für eine korrekte Deanonymisierung betrieben werden müßte, bei weitem den Betrag, der für eine alternative Informationsbeschaffung aufzuwenden wäre, und das Kriterium der faktischen Anonymität war mit Sicherheit erfüllt.

Das Anonymisierungsprojekt hat damit eine Reihe wichtiger Ergebnisse erbracht, die in ihrer Bedeutung für die faktische Anonymität von Einzelangaben zusammenfassend wie folgt bewertet werden können.

Ganz allgemein hat sich gezeigt, daß bei der Anwendung auf empirische Daten das Reidentifikationsrisiko weitaus niedriger ist, als dies bei bisher durchgeführten Experimenten mit ganz oder teilweise synthetisch produzierten Daten der Fall war.

Beim Massenfischzug auf der Basis des Gelehrtenkalenders hat sich gezeigt, daß es unmöglich ist, mit leicht verfügbarem Zusatzwissen ohne großen Aufwand eine größere Zahl von Personen zu reidentifizieren, wenn Einzelangaben aus einer Untersuchung übermittelt werden, die auf einer Stichprobenerhebung basieren, deren Auswahlsatz so klein wie beispielsweise beim Mikrozensus ist.

Die wesentliche Kostenkomponente bei einem Massenfischzug mit einfachen Zuordnungstechniken ist nicht die EDV-Software oder die Rechenzeit, sondern die Verfügung über umfangreiche Adreßdateien mit differenzierten Überschneidungsmerkmalen, die mit Informationen im Mikrodatenfile im zeitlichen und sachlichen Bezug exakt und fehlerfrei übereinstimmen. Das Risiko der Erfolgschancen eines Massenfischzugs und die damit verbundenen Kosten werden deshalb nicht von den zukünftigen Entwicklungen in der EDV-Technologie oder den EDV-Kosten bestimmt, sondern allenfalls davon, ob innerhalb der Wissenschaft sehr umfangreiche (mehrere hunderttausend Fälle umfassende) personenbezogene Dateien leicht zugänglich werden, die in differenzierten Überschneidungsmerkmalen vollständig mit amtlichen Mikrodaten kompatibel und als EDV-lesbare Dateien erfaßt sind. Wenn ein solcher Fall absehbar wird, müßte man neue Aufwandsuntersuchungen durchführen und eventuell neue Anonymisierungsmaßnahmen entwickeln.

Bei Strategien der gezielten Suche haben die Experimente das Ergebnis erbracht, daß die untersuchten Einzelangaben selbst bei Berücksichtigung einer Vielzahl familien-, haushalts- und erwerbsbezogener Merkmale faktisch anonym waren. Die Datensätze einer repräsentativ ausgewählten Stichprobe der Bevölkerung erwiesen sich als faktisch anonym auch unter der Bedingung, daß unter den vielfältigen haushalts- und familientypisierenden Merkmalen keinerlei Anonymisierungsmaßnahmen vorgenommen wurden.

Bei den Experimenten mit der diskriminanzanalytischen Reidentifikationstechnik von Paaß/Wauschkuhn hat sich erwiesen, daß das diskriminanzanalytische Verfahren bei empirischen Daten nicht 'leistungsfähiger' ist als Abgleichverfahren. Gleichzeitig haben diese Experimente deutlich gemacht, daß die Anwendung der diskriminanzanalytischen Reidentifikationstechnik im Hinblick auf den potentiellen Deanonymisierungsertrag einen prohibitiv hohen Aufwand bedeutet. Es kann auf absehbare Zeit ausgeschlossen werden, daß für Deanonymisierungszwecke im Kontext der Sozialwissenschaften dieses oder ein ähnlich komplexes Verfahren eingesetzt wird.

Eine wesentliche Erkenntnis der durchgeführten Experimente besteht darin, daß das Vorhandensein einmaliger Ausprägungskombinationen keineswegs

schon eine hinreichende Bedingung für eine Reidentifikation bedeutet. Bei den Experimenten wiesen sowohl im Mikrodatenfile wie im Identifikationsfile die überwiegende Zahl der relevanten Fälle einmalige Ausprägungskombinationen auf. Obwohl beim Gelehrtenkalender das Identifikationsfile einer Vollerhebung nahekam, war die Zahl der mit Sicherheit richtig vorgenommenen Reidentifikationen nahezu gleich Null. Der wichtigste Grund dafür ist die praktische Unvermeidbarkeit von Inkompatibilitäten zwischen den Datenfiles.

Derartige Inkompatibilitäten, die auf Erhebungs-, Übertragungs- und Kodierfehler, unterschiedliche Abgrenzung der Kategorien, differierende Erhebungskontexte, Erhebungsziele oder Erhebungszeitpunkte etc. zurückzuführen sind, wurden bislang in ihrer Wirkung auf das Reidentifikationsrisiko nicht angemessen eingeschätzt. Im Ergebnis bewirken solche Inkompatibilitäten zwischen Mikrodatenfile und Identifikationsfile bei Reidentifikationsversuchen eine Vielzahl von Falsch- und Nichtzuordnungen. Aus diesem Grunde führt auch eine Abschätzung des Reidentifikationsrisikos, die lediglich auf einer wahrscheinlichkeitstheoretischen Berechnung von einzigartigen Fällen in einem Mikrodatenfile (Uniqueness-Konzept) beruht, zwangsläufig zu einer erheblichen Überschätzung des real bestehenden Risikopotentials.

Eine Dateninkompatibilität wird um so wahrscheinlicher, je mehr Variablen benötigt werden, um eine einmalige Ausprägungskombination herzustellen. Da jedes Merkmal mit einer gewissen Wahrscheinlichkeit inkompatibel ist, wächst die Wahrscheinlichkeit, daß zumindest bei einem Merkmal eine Dateninkompatibilität vorliegt, mit jedem zusätzlich einbezogenen Merkmal, sofern die Unstimmigkeiten nicht vollständig korreliert sind. Allein aufgrund 'natürlicher' Dateninkompatibilitäten sind Datensätze also umso mehr gegen das Risiko der Reidentifikation geschützt sind, je mehr Variablen benötigt werden, um zu einzigartigen Ausprägungskombinationen zu gelangen. Schon bei fünf bis acht benötigten Merkmalen sind Datensätze allein wegen impliziter Dateninkompatibilitäten gegen Deanonymisierung mit Hilfe einfacher Sortier- und Selektionsroutinen in hohem Maße geschützt. Als Maßnahme zum Schutz gegen Deanonymisierung ist deshalb nicht vorrangig auf die Verhinderung einmaliger Ausprägungskombinationen abzustellen, sondern es ist eher darauf zu achten, daß keine Merkmalsausprägungen ausgewiesen werden, die so selten sind, daß durch sie allein einzelne Personen leicht identifiziert werden könnten (vgl. hierzu auch Brennecke 1980:163).

Die bei den Experimenten genutzten Datensätze enthielten nur diejenigen Merkmale, bei denen ein Deanonymisierungsversuch ansetzen müßte, d.h. die typischen Überschneidungsmerkmale mit dem möglicherweise vorhandenen

oder beschaffbaren Zusatzwissen (Hamacher 1980:223). Die Übermittlung zusätzlicher Merkmale beeinflußt das Reidentifikationsrisiko nicht in entscheidender Weise: Zum einen ist in aller Regel das Zusatzwissen, das für die Nutzung dieser Merkmale zu Deanonymisierungszwecken erforderlich wäre, nicht vorhanden. Zum andern sinkt mit jedem zusätzlich genutzten Merkmal die Wahrscheinlichkeit, daß eine eindeutige Zuordnung auch eine richtige Zuordnung ist, und wirkt damit dem höheren Informationspotential des Zusatzwissens entgegen. Für die Anonymisierung von Datenfiles mit Einzelangaben sind deshalb maßgeblich nur die Merkmale zu berücksichtigen, für die typischerweise Zusatzwissen vorhanden ist.

Bei den vielfältigen untersuchten Datenkonstellationen und Deanonymisierungsstrategien zeigte sich, daß nur bei einer gezielten Suche oder Einzelfischzügen in besonders gelagerten, sehr seltenen Einzelfällen unter Umständen mit verhältnismäßig geringem Aufwand die Möglichkeit einer Reidentifikation besteht. Dieser Fall kann wie folgt charakterisiert werden:

1) Eine im Mikrodatenfile gesuchte Person gehört einer sehr kleinen, durch ein spezielles Merkmal identifizierbaren Subpopulation an;

2) Der Mikrodatenfile enthält differenzierte Regionalinformationen, so daß in den Regionaleinheiten nur wenige Personen der spezifischen Subpopulation leben;

3) der Datenangreifer weiß, daß die gesuchte Person im Mikrodatenfile enthalten ist;

4) die Merkmale der Person sind genau in der Weise im Mikrodatenfile erfaßt, wie es der Forscher vermutet.

Damit es zu einem Datenangriff kommt, muß außerdem vorausgesetzt werden, daß der betreffende Forscher ein subjektives Interesse daran hat, das die denkbaren Kosten der Konsequenzen des Angriffs (Reputationsverlust, Vertragsstrafen, gesetzliche Strafen) übersteigt. Beim gleichzeitigen Zusammentreffen aller dieser speziellen Bedingungen erscheint die Möglichkeit der Reidentifikation eines Einzelfalles ohne großen Aufwand als gegeben. Bereits wenn eine der Bedingungen nicht gegeben ist, kann eine sichere Reidentifikation ohne den Aufwand unverhältnismäßig hoher Kosten nach den durchgeführten Experimenten als äußerst gering betrachtet werden. Das gleichzeitige

Zusammentreffen aller Bedingungen kann bei Stichprobenerhebungen als außergewöhnlich seltenes Ereignis betrachtet werden. Dennoch sollten bei der Datenübermittlung Vorkehrungen getroffen werden, damit auch eine solche (unwahrscheinliche) Risikokonstellation ausgeschlossen ist.

Zur Sicherstellung der faktischen Anonymität von Einzelangaben werden sowohl allgemeine Schutzvorkehrungen wie datenfile-spezifische Maßnahmen empfohlen, die sowohl im Sinne des Datenschutzes wirksam sind, als auch die vielfältigen Interessen der Forschung berücksichtigen. Die Vorschläge beziehen sich nur auf den Mikrozensus und die Einkommens- und Verbrauchsstichprobe, weil die Untersuchungen des Anonymisierungsprojektes auf diese Datenfiles ausgerichtet waren. Die Übertragbarkeit der Ergebnisse auf andere Datenbestände müßte gesondert untersucht werden. Im einzelnen sind die Empfehlungen im vorletzten Kapitel dieses Buches dargestellt und in der folgenden Übersicht zusammengefaßt. Die allgemeinen Schutzvorkehrungen sind weitgehend bereits durch Gesetze vorgegeben und können durch vertragliche Bindung des Empfängers faktisch anonymer Daten konkretisiert und ergänzt werden. Des weiteren wird eine möglichst strikte Geheimhaltung der lokalen Umsetzung der Stichprobenpläne durch die amtliche Statistik empfohlen, sowie eine systemfreie Anordnung der weitergegebenen Daten.

Die datenfile-spezifischen Anonymisierungsmaßnahmen sollten so gestaltet sein, daß durch sie das wissenschaftliche Analysepotential der Daten möglichst wenig beeinträchtigt wird. Als adäquateste Lösung bietet sich an, bei der Weitergabe von Daten, bei denen eine tiefgegliederte regionale Differenzierung sinnvoll ist, für den Regelfall zwei unterschiedliche Datenfiles vorzusehen: ein sogenanntes Grundfile und ein sogenanntes Regionalfile. Dies trifft für den Mikrozensus zu. Bei der Einkommens- und Verbrauchsstichprobe ist aus Stichprobengründen eine über die Ebene der Bundesländer hinausgehende Regionalisierung nicht zu empfehlen.

Beim Grundfile sind die Regionalinformationen nur in undifferenzierter Form (Bundesländer und grobe Charakterisierung des Ortstyps, z.B. durch vergröberte Gemeindegrößenklasse), die übrigen Informationen jedoch in möglichst großer Differenzierung enthalten. Das Regionalfile enthält dagegen stärker differenzierte Regionalinformationen, grenzt dafür die Differenzierungstiefe bei den übrigen Variablen ein.

Ergänzend wird die Weitergabe einer Einprozent-Substichprobe aus dem Mikrozensus empfohlen. Sie enthält mit Ausnahme der Regionalangaben sämtliche Merkmale des Mikrozenus ohne weitere Anonymisierungsmaßnahmen.

Durch den Ausfall der Regionalinformationen ist die Substichprobenziehung eine hinreichende Schutzmaßnahme. Dieses File ist für Analysen gedacht, in welchen eine Massendatenbasis nicht erforderlich ist oder für Wissenschaftler, die eine Massendatenanalyse nicht selbst vornehmen, jedoch an einem kleineren Datenfile geplante Analysen testen wollen.

Übersicht 1: Empfehlungen zu Anonymisierungsmaßnahmen für die Übermittlung faktisch anonymer Mikrodaten aus Mikrozensus und EVS

Allgemeine Schutzvorkehrungen

1) *vertragliche Bindung* des Empfängers faktisch anonymer Daten;
- geeignete *technisch-organisatorische* Maßnahmen zur Kontrolle der Datennutzung;
- *Vertragsstrafe* bei Reidentifikationsversuch;
- Nutzungsbegrenzung auf das angegebene *wissenschaftliche Vorhaben*;
- *keine Datenweitergabe* an Dritte;
- *Löschung bzw. Rückgabe* der Daten nach Abschluß des wissenschaftlichen Vorhabens;
- Behandlung von Datenauszügen oder -duplikaten wie von Originalen;
- *keine Nachforschungen* hinsichtlich der lokalen Umsetzung der Stichprobenpläne.

2) *Geheimhaltung der lokalen Umsetzung der Stichprobenpläne* durch die amtliche Statistik.

3) *Systemfreie Anordnung* der Daten.

Anonymisierungsmaßnahmen für Mikrozensus - *Grundfile*:

1) *Regionalangaben* nur stark vergröbert (nur Bundesland und Siedlungsstrukturtyp bzw. vergröberte Gemeindegrößenklasse);
- keine einzelne Gemeinde mit weniger als 500 000 Einwohnern darf identifizierbar sein;
- ein Gemeindetyp, dem mehrere Gemeinden zugehören, darf in keinem Bundesland weniger als 400 000 Einwohner umfassen.

2) Keine *Nationalität* oder Gruppe von Nationalitäten mit weniger als 50 000 Einwohnern in der Bundesrepublik darf identifizierbar sein.

3) *Merkmalsvergröberungen* bei allen übrigen Variablen - falls erforderlich - so, daß in den univariaten Randverteilungen jede ausgewiesene Merkmalsausprägung für die Bundesrepublik mindestens 5000 Fälle umfaßt.

4) Nur eine *Substichprobe* von 70 Prozent wird weitergegeben.

Anonymisierungsmaßnahmen für Mikrozensus - *Regionalfile*:

1) Durch die Kombination von Regionalklassifikationen soll keine Regionaleinheit ermittelbar sein, die eine Einwohnerzahl von weniger als 100 000 Personen aufweist.

2) Vergröberung von Beruf, Wirtschaftszweig, Nationalität und Alter dergestalt, daß keine Merkmale ausgewiesen werden, denen
- in der Grundgesamtheit der Bundesrepublik Deutschland nicht wenigstens 50 000 Einwohnern zukommen,
- pro übermittelter Regionaleinheit (ohne Substichprobenziehung) nicht mindestens drei Fällen im Mikrodatenfile zukommen.

3) Alle übrigen Variablen werden - falls erforderlich - so aggregiert, daß jede ausgewiesene Merkmalsausprägung für die Grundgesamtheit der Bundesrepublik mindestens 5000 Fälle umfaßt.

4) Nur eine *Substichprobe* von (mindestens) 85 Prozent wird weitergegeben.

Anonymisierungsmaßnahmen für EVS:

1) *Regionalangaben* wie Mikrozensus-Grundfile

2) *Nationalität* wie Mikrozensus-Grundfile

3) *Merkmalsvergröberung*
- *Bei "sichtbaren" oder über die Zeit stabilen Merkmalen (z.B. Geburtsjahr, Stellung im Beruf oder Besitz auffälliger Konsumgüter):*
  Merkmalsvergröberung wie Mikrozensus-Grundfile (Punkt 3).
- *Bei öffentlich wenig bekannten oder wenig stabilen, jedoch differenziert erfaßten Merkmalen (insb. Einkommens-, Vermögens- und Ausgabenbeträge):*
  Die fünf niedrigsten und fünf höchsten Ausprägungen eines Merkmals werden nur als Mittelwert dieser Ausprägung ausgewiesen. Die übrigen Ausprägungen im untersten und obersten Dezil der Verteilung eines solchen Merkmals werden mit einem Zufallsfehler von bis zu plus oder minus ein Prozent des jeweiligen Merkmalswertes überlagert.

4) *Substichprobe* in Abhängigkeit von der Anzahl der weitergegeben Erhebungsteile:
   98 Prozent: Haushalts- und Personenmerkmale + 1 Erhebungsteil
   90 Prozent: Haushalts- und Personenmerkmale + 2 Erhebungsteile
   80 Prozent: Haushalts- und Personenmerkmale + 3 Erhebungsteile

1 Ziele und Schwerpunkte einer Untersuchung der faktischen Anonymität von Mikrodaten*

Empirische Forschung in den Human-, Sozial- und Wirtschaftswissenschaften beruht unvermeidlich auf der Sammlung von Informationen über einzelne Personen, Haushalte oder andere Einheiten, die Gegenstand einer Untersuchung sind. Ob die Informationen durch systematische Beobachtung, durch Befragung, durch die Aufzeichnungen bei der Geschäftstätigkeit von Behörden oder privaten Einrichtungen oder durch andere Methoden gewonnen werden, ist dabei nicht entscheidend. Auch wenn in den veröffentlichten Statistiken und Forschungsergebnissen diese Einheiten nicht mehr erkenntlich sind, weil die Daten der Elementareinheiten durch Aggregation oder andere Verfahren der statistischen Informationsverarbeitung zusammengefaßt sind, setzt Forschung letztlich immer bei der gezielten Datensammlung über solche Elementareinheiten an. Statistiken und die Arbeit der statistischen Ämter sind ein Teil solcher Forschung.

Die Informationen über die Elementareinheiten bezeichnen wir als *Mikrodaten*. Konkrete Beispiele sind etwa die Angaben, die jede einzelne Person in der Volkszählung oder in einer sozialwissenschaftlichen Erhebung über sich macht oder die Aufzeichnungen, die amtliche Stellen oder private Einrichtungen über einzelne Personen halten: über ihr Alter, ihr Geschlecht, ihren Besitz, ihr Einkommen, ihre gekauften Konsumgüter, ihre Krankheiten, über ihre Einstellungen zu bestimmten Sachverhalten oder ihre politischen Präferenzen. Da sich Mikrodaten auf die Informationen beziehen, die über jede einzelne Einheit in einer Datensammlung enthalten sind, werden sie oft auch als Einzelangaben bezeichnet. Bei diesen Einheiten muß es sich nicht um Personen oder Individuen handeln. Einheiten können auch Grundstücke, Betriebe, Verbände, Institutionen oder andere Objekte sein, die Gegenstand einer Datensammlung sind. Da diese Einheiten jedoch vielfach Individuen sind, ist auch der Begriff Individualdaten gebräuchlich.[2]

---

* Autor: Walter Müller

2 Mikrodaten unterscheiden sich von Aggregatdaten oder Makrodaten. Aggregatdaten oder Makrodaten sind Daten, die in der Regel durch statistische Operationen für bestimmte Kollektive aus den Mikrodaten der Einheiten gewonnen werden, die diesen Kollektiven angehören, beispielsweise die Zahl der Männer oder Frauen, die in einer bestimmten Gemeinde leben oder das Durchschnittseinkommen der Angehörigen einer Berufsgruppe. Wenn ein Aggregatdatum oder ein Makrodatum aus

Informationen über identifizierbare Personen gehören nach den Wertvorstellungen unserer Gesellschaft zu dem zu schützenden Bereich der Privatsphäre dieser Personen; Informationen über Betriebe oder andere Institutionen - je nach Art der Institutionen und Informationen - zu dem zu schützenden Bereich der Interessenssphäre der betroffenen Institutionen. Die Einsichtnahme in solche Informationen durch Unbefugte oder ihre öffentliche Bekanntmachung gilt als Eingriff in die Integrität der Person oder die Interessenslage einer Institution. Insbesondere im Hinblick auf den Schutz der Privatsphäre hat das Bundesverfassungsgericht dem Schutz von Informationen eine sehr hohe Bedeutung beigemessen. In seinem Volkszählungsurteil von 1983 hat es aus dem Grundsatz des Schutzes der Würde der Person das Recht auf informationelle Selbstbestimmung abgeleitet und dieses Recht in den Kanon der Grundrechte aufgenommen. Die Aufgabe des Datenschutzes ist es, diesen Wertvorstellungen und entsprechenden Rechtsgrundsätzen und Gesetzen Geltung zu verschaffen.

Für weite Bereiche der Statistik und der Forschung in den Human-, Sozial- und Wirtschaftswissenschaften ergibt sich daraus ein grundsätzliches Dilemma. Auf der einen Seite führt der lange Weg des Zustandekommens von Forschungsergebnissen unweigerlich auf eine Phase der Sammlung und Verarbeitung von Daten über einzelne Individuen, Institutionen oder andere Elementareinheiten zurück. Andererseits ist es eine grundlegende Norm der Wissenschaft, "Forschung zu wissenschaftlichen Zwecken so durchzuführen, daß den Untersuchungsobjekten kein Schaden entsteht" (Mangold 1980:243), und das heißt auch dafür zu sorgen, daß die Integrität der Privatsphäre nicht verletzt wird. Auch die Statistik ist diesem Prinzip verpflichtet und ist wie die Wissenschaft darauf angewiesen, daß sich niemand, der Informationen bereitstellt, dadurch in seiner persönlichen Sphäre oder seinen Interessen beeinträchtigt sieht (Allerbeck 1987).

1.1 Die faktische Anonymität von Mikrodaten und ihr Beitrag zum Datenschutz

Ein höchst bedeutsamer Weg, dieses grundlegende Dilemma für einen großen Teil der Forschung zu lösen, besteht in der Anonymisierung der Mikrodaten, das heißt, in Vorkehrungen, die es verhindern, daß die Informationen mit den

---

einer Vielzahl von Mikrodaten gebildet ist, können in der Regel die Mikrodaten aus den Makrodaten nicht mehr rekonstruiert werden.

Personen oder Institutionen, auf die sie sich beziehen, in Verbindung gebracht werden (können). Daten werden zwar als Informationen über einzelne Personen oder Institutionen genutzt, jedoch nur anonym, ohne Bezug zu den konkreten, bestimmten oder bestimmbaren Personen oder Institutionen, von denen sie stammen. An Stelle von identifizierbaren Individuen stehen nur noch abstrakte, unbekannte Einheiten und Merkmalsträger.

Diese Lösung ist allerdings nicht für alle Forschungzwecke gangbar. Eine Grenzziehung zwischen den Fällen, in denen die Anonymisierung ohne Einschränkung der Forschung möglich ist, und den Fällen, für die dieses nicht gilt, ist in 31 Thesen zu Datenzugang und Datenschutz in hervorragender, auch heute noch gültiger Weise von einer Gruppe von Wissenschaftlern unterschiedlicher Disziplinen und einem führenden Experten des Datenschutzes als Ergebnis gründlicher Untersuchungen versucht worden und muß deshalb hier nicht weiter erörtert werden (Kaase/Krupp/Pflanz/Scheuch/Simitis 1980). Ist Anonymität aufgrund spezieller Aufgaben der Forschung oder in bestimmten Forschungsphasen nicht möglich, müssen an Stelle der Anonymität andere Schutzmechanismen treten, z.B. die Garantie der vertraulichen Behandlung der Informationen durch einen sehr kleinen Personenkreis und die funktionale Abschottung derjenigen, die Zugang zu personenbezogenen Daten haben.

In den Fällen, in denen Anonymität möglich ist, kommt dieser in mehrfacher Hinsicht große Bedeutung zu:

- Der gesamte Komplex des Datenschutzes ist vom Gesetzgeber in einer Weise konzipiert, daß die Regeln und Normen des Datenschutzes nur Daten betreffen, die entweder unmittelbar personenbezogen oder personenbeziehbar sind. Anonymisierte Daten, die keinen Bezug auf eine bestimmte oder bestimmbare Person aufweisen, unterliegen nicht den gesetzlichen Regeln des Datenschutzes. Der Gesetzgeber hat damit also anerkannt, daß durch die Anonymisierung Daten eine Qualität bekommen, die keine besonderen, gesetzlich zu regelnden Maßnahmen zu Zwecken des Persönlichkeits- oder Interessenschutzes erfordern.

- Die Zusicherung der Anonymität und der anonymen Behandlung von Daten ist in der Regel auch für die Personen, die Informationen geben, ein entscheidender Faktor ihrer Mitteilungsbereitschaft. Unter dem Siegel der Anonymität werden Außenstehenden Einsichten in persönliche Verhältnisse, Probleme, Verhaltensweisen und Einstellungen gewährt, die dieselben Personen ansonsten entweder überhaupt niemandem oder allenfalls engsten Vertrauten offenlegen. Anonymität scheint in diesem Sinne in

der modernen Gesellschaft auch im öffentlichen Leben als Schutz der Persönlichkeit und in der Folge als Äquivalent von Datenschutz wahrgenommen zu werden. Sei es in der Straßenbahn, im Restaurant, auf öffentlichen Plätzen oder auch in Selbsthilfegruppen zur Bewältigung persönlicher Probleme (Anonyme Alkoholiker, Anonyme Spieler, Anonyme Overeaters, ...): Immer wieder ist beobachtbar, daß Menschen unter der Annahme der Anonymität Informationen preisgeben, mit denen sie unter der Erwartung, sie könnten jemals von Dritten mit ihrer Person in Verbindung gebracht werden, gewiß oft sehr viel zurückhaltender umgehen würden. Wie sich ein solches Verständnis von Privatsphäre und ihres Schutzes in der Öffentlichkeit durch Anonymität herausgebildet hat, wäre selbst eine interessante soziologische und sozialhistorische Analyse wert.

Durch Anonymisierung kann damit zwar dem Grundsatz nach für einen wesentlichen Teil der Forschung in den Human-, Sozial- und Wirtschaftswissenschaften das Dilemma zwischen Bedarf an Mikrodaten und Schutz der Privatsphäre oder der Interessenlage des Datengebers gelöst werden. Die Frage ist jedoch, unter welchen Bedingungen die Anonymität von Mikrodaten gesichert ist und wie sichergestellt werden kann, daß eine Deanonymisierung oder Reidentifikation anonymisierter Mikrodaten, d.h. eine nachträgliche Wiederherstellung des Personenbezuges, ausgeschlossen ist. Unter den Experten besteht Übereinstimmung darin, daß eine *absolute* Sicherheit gegenüber Reidentifikation in der Regel nur mit Maßnahmen erreichbar ist, die ein so hohes Maß an Verfremdung der Mikrodaten implizieren, daß sie für viele Forschungszwecke unbrauchbar werden oder ihre Nutzbarkeit in nicht zu vertretender Weise eingeschränkt ist.

In der Diskussion hat sich deshalb das Konzept der *faktischen* Anonymität als Lösung durchgesetzt. Damit ist gemeint, daß die Möglichkeit einer Deanonymisierung nicht mit *absoluter* Sicherheit ausgeschlossen werden muß. Vielmehr ist im Sinne einer Güterabwägung zwischen dem Bedarf der Forschung an der Daten*nutzung* und den berechtigten Interessen der Datenspender an Daten*schutz* nur ein hinreichend hohes Maß an Sicherheit vor Deanonymisierung zu gewährleisten und ein begrenztes Restrisiko hinzunehmen. Wie eine Straßenverkehrsordnung, die absolute Sicherheit zu gewährleisten hätte, von niemandem gefordert wird, weil sie praktisch jeden Verkehr ausschließen müßte, so ist es in analoger Weise auch nicht sinnvoll, auf *absolut* sicheren Maßnahmen bei der Anonymisierung von Daten zu bestehen. Das Prinzip der faktischen Anonymität beherrscht die Alltagspraxis in der einschlägigen wissenschaftlichen Forschung in den meisten Ländern, in denen systematische sozialwissen-

schaftliche Forschung stattfindet und in denen ein Verständnis für die schutzwürdigen Interessen der von der Forschung betroffenen Personen und Institutionen gegeben ist. Das Prinzip ist inzwischen auch in Datenschutzgesetzen explizit etabliert, so im Bundesdatenschutzgesetz von 1990 (Dörr/Schmidt 1991) und in einzelnen Landesdatenschutzgesetzen.

Selbst wenn das Grundprinzip der faktischen Anonymität damit weitgehend anerkannt ist, ergeben sich jedoch eine Reihe von offenen Fragen:

- Wie läßt sich faktische Anonymität operational bestimmen und welches Restrisiko der Reidentifikation kann hingenommen werden?

- Von welchen Faktoren hängt das Reidentifikationsrisiko ab und wie groß ist es unter bestimmten angebbaren Bedingungen?

- Welche Anonymisierungsmaßnahmen bieten einen wirksamen Schutz und wie wirken sich diese Maßnahmen auf das wissenschaftliche Analysepotential der Daten aus?

Diese allgemeinen Fragen stellen sich für jede Art von Mikrodaten. Ihre Beantwortung kann jedoch nur in Grenzen allgemein und für jede Art von Daten in gleicher Weise zutreffend erfolgen. Das verbleibende Risiko, das hinnehmbar ist, wird beispielsweise von der Art der Informationen abhängen müssen. So wird ein verbleibendes Risiko bei sensiblen Informationen, bei deren Offenlegung Betroffene einen großen Schaden erleiden würden, niedriger anzusetzen sein als bei Informationen, die ohnehin weithin bekannt sind oder bei Informationen, bei denen nicht zu erwarten ist, daß Betroffene durch Offenlegung beeinträchtigt werden könnten. Wie sich zeigen wird, variieren die Bestimmungsgrößen des Reidentifikationsrisikos erheblich nach der Art der in den Mikrodaten enthaltenen Informationen und nach dem beim Datenempfänger vorhandenen oder erwerbbaren Wissen. Je nach Forschungszweck beeinträchtigen unterschiedliche Anonymisierungsmaßnahmen das Analysepotential der Daten in unterschiedlicher Weise.

Damit ist es vorteilhaft und erforderlich, eine diesen Problemen gewidmete Untersuchung auf präzise, bestimmte Bedingungen zuzuschneiden. Dieses geschieht in dieser Arbeit durch die Konzentration auf Mikrodaten der amtlichen Statistik und ihre Nutzung innerhalb von Einrichtungen der wissenschaftlichen Forschung. Selbst bei Daten der amtlichen Statistik stellt sich das Problem unterschiedlich für verschiedene Typen solcher Daten und deshalb wird eine weitere Einschränkung der Untersuchung auf Mikrodaten der amtlichen Statistik

erfolgen, die auf dem Wege der Befragung bei einzelnen Personen und Haushalten der Bevölkerung zustande kommen.

In diesem Buch wird eine umfassende Klärung dieser Problematik versucht, angefangen bei der bisherigen Weitergabepaxis der statistischen Ämter auf nationaler und internationaler Ebene, der Klärung des Begriffs der faktischen Anonymität, der Untersuchung der Randbedingungen der Verarbeitung anonymisierter Mikrodaten in wissenschaftlichen Instituten, der Klärung der Frage möglicher Deanonymisierungsmotive, der empirischen Überprüfung von Reidentifikationsrisiken bis hin zu der darauf basierenden Entwicklung von Empfehlungen von Maßnahmen, die eine hinreichende faktische Anonymisierung sicherstellen.

In diesem Kapitel wird zunächst der Hintergrund der Untersuchung dargestellt. Dann werden die einzelnen Forschungsfragen ausführlicher entfaltet und damit verbunden ein Wegweiser durch diesen Forschungsbericht gegeben.

1.2 Der Bedarf der Wissenschaft an Mikrodaten der amtlichen Statistik und die gesetzlichen Voraussetzungen ihrer Nutzung

Neben den erwähnten grundsätzlichen Fragen war der konkrete Anlaß für diese Untersuchung die Novellierung des Bundesstatistikgesetzes (BStatG) von 1987, das in §16 Abs. 6 erstmals eine Regelung vorsieht, nach der auf der Grundlage der faktischen Anonymität amtliche Mikrodaten an Hochschulen und andere Einrichtungen der unabhängigen Forschung übermittelt werden können. In der Folge dieser gesetzlichen Neuregelung lag ein Interesse dieser Untersuchung also auch darin, die Bedingungen zu klären, unter denen diese Datenübermittlung erfolgen kann. Die Studie sollte einen Beitrag dafür leisten, die Kooperation zwischen der amtlichen Statistik und der Wissenschaft zu verstärken und eine intensivere Nutzung amtlicher Mikrodaten in der Wissenschaft zu ermöglichen.

Daten der amtlichen Statistik stellen für viele Bereiche und viele Zwecke der wissenschaftlichen Forschung seit langem eine außerordentlich wichtige und reichhaltige Datenressource dar. Allein die Tatsache, daß diese Daten in vielfältiger Weise in der wissenschaftlichen Forschung tatsächlich genutzt werden, ist hinreichender Beleg dafür. Es ist hier nicht der Ort, die konkreten Verwendungszwecke im einzelnen zu erörtern. Für einzelne Datenbasen ist dies

andernorts geschehen (vgl. z.B. für den Mikrozensus: Esser u.a. 1989, Statistisches Bundesamt 1990).

In den letzten Jahrzehnten hat sich jedoch die Verwendungsart dieser Daten nachhaltig verändert. Zumindest seit Beginn der siebziger Jahre besteht ein von seiten der wissenschaftlichen Forschung immer wieder mit Nachdruck artikulierter Bedarf, diese Daten als anonymisierte Mikrodaten, d.h. in der ursprünglichen Form von Einzelangaben für Analysezwecke zur Verfügung zu haben. In anderen Ländern, besonders in den USA, werden Mikrodaten aus der amtlichen Statistik in der wissenschaftlichen Forschung seit langem mit großer Selbstverständlichkeit und Gewinn hauptsächlich in den Sozial- und Wirtschaftswissenschaften in dieser Weise genutzt. Der Datenbedarf resultiert aus vielen inhaltlichen Fragestellungen der gegenwärtigen Forschung. Der Anspruch, die Daten als Mikrodaten nutzen zu können, ergibt sich jedoch vor allem aus der Entwicklung der methodischen Verfahren und verbesserten Möglichkeiten der Datenanalyse. Verschiedene Verfahren der multivariaten Analyse und der Mikrosimulation setzen die Möglichkeit des Rückgriffs auf Mikrodaten unabdingbar voraus. Nur um den Preis des Zurückbleibens hinter der wissenschaftlichen Entwicklung oder der Datenduplizierung durch die Sammlung entsprechender Daten durch die Wissenschaft selbst, könnte auf die Übermittlung von Mikrodaten aus der amtlichen Statistik an Institutionen der wissenschaftlichen Forschung verzichtet werden. Dieser Sachverhalt und der Bedarf sind in zahlreichen Beiträgen ausführlich erörtert und begründet worden (Vgl. Krupp 1980, Kaase u.a. 1980, Krupp 1982, Müller 1982, Müller/Hauser 1987, Statistisches Bundesamt 1987, Blien/Papastefanou 1988, Grohmann 1989, Krupp/Preißl 1989, Rainwater 1990). Es kann deshalb hier darauf verzichtet werden, diese Begründungen zu wiederholen. Die Verarbeitung entsprechender Mikrodaten gehört schlichtweg zum international etablierten Forschungsstandard. Verbesserte Möglichkeiten des Zugangs zu Mikrodaten werden auch von internationalen Forschungsorganisationen (z.B. European Science Foundation) seit langem angemahnt. Auch in höchstrichterlichen Urteilen wird der Bedarf grundsätzlich anerkannt und seine Befriedigung in datenschutzgerechter Weise als realisierbar angesehen.

Mit dem Bundesstatistikgesetz von 1987 hat der Gesetzgeber durch die Einführung einer spezifischen Wissenschaftsklausel die rechtlichen Voraussetzungen für Datenübermittlungen geschaffen, die den Bedürfnissen der Wissenschaft entgegenkommen.

Nach §16 Abs.6 BStatG dürfen "für die Durchführung wissenschaftlicher Vorhaben (...) vom Statistischen Bundesamt und den statistischen Ämtern der Län-

der Einzelangaben an Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger Forschung übermittelt werden, wenn die Einzelangaben nur mit einem *unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft* (Hervorhebung durch d.V.) zugeordnet werden können und die Empfänger Amtsträger, für den öffentlichen Dienst besonders Verpflichtete oder Verpflichtete nach Absatz 7 sind" (Dorer/Mainusch/Tubies 1988:87f). Diese gesetzliche Regelung folgt dem Verständnis der faktischen Anonymität, wie er bereits vom Bundesverfassungsgericht verwendet wurde und auch in einer Stellungnahme der European Science Foundation empfohlen wurde. Sie soll eine den Bedürfnissen der Wissenschaft gerecht werdende Nutzung von mit öffentlichen Ressourcen gesammelten Daten der amtlichen Statistik ermöglichen.

Mit der Einführung dieser spezifischen Wissenschaftsklausel wurde die bis dahin geltende gesetzliche Regelung (§11 Abs.5 BStatG 1980) ergänzt, die eine Übermittlung von Einzelangaben an die Wissenschaft nur bei absoluter Anonymität vorsah. Das hierbei ursprünglich angestrebte Ziel, "der Wissenschaft und anderen Stellen im gewissen Umfang Daten zur eigenen Aufbereitung unter Wahrung des Datenschutzes zur Verfügung zu stellen" (Deutscher Bundestag 1986:21), hatte sich nur sehr eingeschränkt verwirklichen lassen. Die Datenweitergabe mußte so restriktiv gehandhabt werden, oder konnte nur mit solch weitreichenden datenmodifizierenden Anonymisierungsmaßnahmen erfolgen, daß die Nutzungsbedarfe der Wissenschaft nicht befriedigt und das wissenschaftliche Potential der Daten auch nicht annähernd ausgeschöpft werden konnte.

Dieser Erkenntnis wird in §16 Abs.6 BStatG durch die Einführung des "*Unverhältnismäßigkeitskriteriums*" insofern entsprochen, als bei übermittelten Daten eine Deanonymisierung nicht mehr mit Sicherheit ausgeschlossen sein muß, sondern implizit ein Rationalkalkül eines Datenangreifers zugrunde gelegt wird. Allerdings ist im Gesetzestext über die angegebene Formulierung hinaus nicht näher konkretisiert, wie die faktische Anonymität in der Praxis umzusetzen ist. Das Abwägungskriterium der Unverhältnismäßigkeit ist im Gesetzestext nicht explizit formuliert. Es kann ihm nur implizit entnommen werden. Im wesentlichen handelt es sich um das Verhältnis zwischen dem Wert bzw. dem durch einen Reidentifikationsversuch erzielten Nutzen und den hierfür anfallenden Kosten. In der vorliegenden Untersuchung wird davon ausgegangen werden, daß sich der Wert bzw. der Nutzen objektiv am ehesten im Vergleich zu den Kosten einer alternativen Informationsbeschaffung ermessen läßt. An späterer Stelle wird ausführlich erörtert werden, wie sich die Unverhältnis-

mäßigkeit in der Regel durch einen solchen Vergleich mit alternativen Informationsbeschaffungskosten bestimmen läßt.

An dieser Stelle ist nur darauf zu verweisen, daß der Gesetzgeber ein spezifisches Verständnis des Unverhältnismäßigkeitskriteriums vorgegeben hat. Die Unverhältnismäßigkeit wird am Kriterium des Aufwandes der Deanonymisierung festgemacht. Es gäbe auch andere Möglichkeiten, dieses Kriterium zu definieren. Dies würde unter Umständen zu anderen Maßstäben dafür führen, wann Daten als faktisch anonym zu betrachten sind. Beispielsweise könnte die Unverhältnismäßigkeit am Schaden gemessen werden, der der Gesellschaft dadurch entsteht, daß der Wissenschaft Daten vorenthalten werden, die sie benötigt, um unabhängige Erkenntnisse zu gewinnen, die für die gesellschaftliche Planung und für politische Entscheidungen oder für die Kontrolle der Effizienz und Wirkung politischer Maßnahmen wichtig sind.

1.3 Schwerpunkte der Untersuchung

Wenn man die faktische Anonymität im Sinne des durch den Gesetzestext implizierten Unverhältnismäßigkeitskriteriums versteht, erweist sich die konkrete Umsetzung dieser Vorschrift in der Praxis als keineswegs trivial. Es fehlen Kenntnisse zu den zentralen Problemen der faktischen Anonymität, vor allem ein auf realistischen Voraussetzungen folgendes Wissen über das Reidentifikationsrisiko von anonymisierten Mikrodaten, die an die Wissenschaft übermittelt werden, über den mit einer Reidentifikation verbundenen Aufwand und den Wert möglicherweise reidentifizierter Angaben innerhalb des Wissenschaftsbereiches.

Eine sorgfältige Untersuchung möglichst aller für die faktische Anonymität relevanten Aspekte ist gerade im Zusammenhang mit Mikrodaten der amtlichen Statistik von besonderer Bedeutung:

- Die Beteiligung der Bevölkerung an Erhebungen der amtlichen Statistik erfolgt zum Teil auf der Grundlage gesetzlicher *Verpflichtung* (z.B. beim Mikrozensus). Bei der Verwendung solcher durch Pflichtbeteiligung der Bürger zustande gekommenen Daten ist deshalb ein besonders hoher Grad an Datenschutz und eine besonders gründliche Prüfung der Wirksamkeit einzelner Schutzmaßnahmen gefordert.

- Die Erhebungen der amtlichen Statistik sind in aller Regel sehr umfangreich. Beim Mikrozensus handelt es sich um eine Stichprobe von einem Prozent der Gesamtbevölkerung und bei der Einkommens- und Verbrauchsstichprobe um eine Stichprobe von ca. 0,2 Prozent der Gesamtbevölkerung. Ohne entsprechende Schutzmaßnahmen wären bei derart umfangreichen Datenbasen viele Menschen in ihren Datenschutzinteressen potentiell gefährdet.

- Die amtliche Statistik ist, gerade weil ihre Erhebungen so umfangreich sind, in hohem Maße auf das Vertrauen der Bürger in einen lückenlosen Datenschutz angewiesen. Sie kann es sich nicht leisten, durch einen "Datenunfall" dieses Vertrauen aufs Spiel zu setzen und muß deshalb bei der Weitergabe von Mikrodaten dem Datenschutzaspekt ein hohes Gewicht beimessen.

- Weil die Erhebungen der amtlichen Statistik so umfangreich sind, eignen sie sich in besonderer Weise zur Untersuchung des Deanonymisierungsrisikos. Bei dem großen Datenumfang der amtlichen Statistik können selbst kleinste Risiken, sofern sie bestehen, empirisch untersucht werden.

Im Zusammenhang mit Mikrodaten der amtlichen Statistik ist es deshalb nicht nur von der Sache her besonders dringend, die faktische Anonymität in allen ihren Voraussetzungen möglichst umfassend abzuklären. Es bestehen auch besonders gute Voraussetzungen für eine solche Untersuchung. In diesem Buch werden in den folgenden Kapiteln die wesentlichen Teilaspekte aufgegriffen, aufeinander bezogen und in jedem Aspekt soweit untersucht, daß zum Ende konkrete Empfehlungen für die Operationalisierung des Kriteriums der faktischen Anonymität bei zwei zentralen Erhebungen der amtlichen Statistik, dem Mikrozensus und der Einkommens- und Verbrauchsstichprobe, abgeleitet werden können. Weil die zum Schluß dargebotenen Empfehlungen sich auf diese beiden Erhebungen beziehen, sind auch die verschiedenen Untersuchungen hauptsächlich auf die spezifischen Gegebenheiten dieser Erhebungen ausgerichtet.

Wegen der Komplexität des Untersuchungsgegenstandes müssen sich die einzelnen Kapitel mit zunächst sehr heterogen erscheinenden Fragen befassen. Die Logik der Sache erfordert eine unterschiedliche Vorgehensweise. Der folgende Ausblick soll dem Leser helfen, die einzelnen Aspekte besser einzuordnen und aufeinander beziehen zu können.

Faktische Anonymität von Mikrodaten im Sinne des §16 Abs.6 BStatG bedeutet, daß die Identität der konkreten Person, auf die sich Einzelangaben beziehen, soweit unkenntlich gemacht ist, daß der Bezug der Einzelangaben auf diese Person, d.h. die Reidentifikation oder Deanonymisierung der Einzelangaben, nur mit einem unverhältnismäßig großen Aufwand möglich ist. Es müssen deshalb vor allem die Wahrscheinlichkeiten der Reidentifikation von an die Wissenschaft gelieferten anonymisierten Mikrodaten und die bei einem Reidentifikationsversuch entstehenden Kosten ermittelt werden.

Das Grundprinzip einer Reidentifikation beruht darauf, daß Einzeldatensätze einer Datei mit anonymen Mikrodaten Einzeldatensätzen einer anderen Datei, die Namen und sonstige Identifikationsmerkmale enthält, in eine eins-zu-eins-Zuordnung gebracht werden. Die für eine Reidentifikation unabdingbare personenbezogene Datei wird auch als Identifikationsfile oder als sogenanntes Zusatzwissen bezeichnet und muß sich - zumindest in Teilen - auf die gleichen Personen wie das anonymisierte Mikrodatenfile beziehen und neben den Informationen zur Identifikation der Person, Informationen zu Merkmalen enthalten, die auch im Mikrodatenfile vorhanden sind (Überschneidungsmerkmale). Durch einen Abgleich dieser Überschneidungsmerkmale auf Identität oder sehr große Ähnlichkeit wird angestrebt, jene Datensätze im Zusatzwissen und Mikrodatenfile herauszufinden, die von ein und derselben Person stammen.

In Kapitel 2 wird einführend der Forschungsstand zu Deanonymisierungsrisiken sowie zu Anonymisierungsmaßnahmen skizziert. Außerdem enthält dieses Kapitel einen Überblick über die bisherige nationale sowie internationale Weitergabepraxis der statistischen Ämter.

In Kapitel 3 werden die verschiedenen Strategien, methodischen Grundprinzipien und Verfahren der Reidentifikation näher erörtert, und es werden einige der mit diesen Verfahren bislang durchgeführten Versuche und ihre Ergebnisse dargestellt sowie in ihrer Aussagekraft gewürdigt.

In Kapitel 4 wird die Frage erörtert, von welchen Faktoren das Reidentifikationsrisiko abhängt. Einschlägige Befunde bisheriger Forschung werden diskutiert. Insbesondere wird untersucht, inwieweit Inkompatibilitäten zwischen den Informationen des Zusatzwissens und den Informationen des Mikrodatenfile zu erwarten sind und wie solche Inkompatibilitäten das Reidentifikationsrisiko beeinflussen.

Ob, abgesehen von der Frage der technischen Machbarkeit, überhaupt Deanonymisierungsversuche zu erwarten sind, hängt ab vom potentiellen

Nutzen der deanonymisierten Daten. Hierbei sind im wesentlichen zwei grundlegend verschiedene Nutzen- und Motivkategorien zu unterscheiden: Da faktisch anonyme Mikrodaten nur an unabhängige Institutionen der Wissenschaft zum Zwecke der Forschung weitergegeben werden dürfen, ist zunächst zu klären, welche Logik der wissenschaftlichen Datennutzung zugrunde liegt und welche Motive sich hieraus für einen Deanonymisierungsversuch ergeben könnten. Selbst wenn Daten nur für die genannten Forschungszwecke weitergegeben werden dürfen, sind jedoch - um das gesamte Gefährdungspotential abzuprüfen - auch denkbare wissenschafts*fremde* Motive für Deanonymisierungsversuche zu untersuchen. Der Versuch einer realistischen Einschätzung der Verwendungsmöglichkeiten deanonymisierter Mikrodaten in wissenschaftlichen und nicht-wissenschaftlichen Verwendungskontexten ist Gegenstand von Kapitel 5.

Die Reidentifikation einer Person in einem anonymisierten Mikrodatenfile ist nur dann möglich, wenn ein Datenangreifer über entsprechendes Zusatzwissen verfügt. Als entscheidende Voraussetzung einer Deanonymisierung ist also zu ermitteln, über welches Zusatzwissen ein Datenangreifer verfügt. Allgemein läßt sich das erlangbare Zusatzwissen kaum abschließend abgrenzen (Burkert 1980). Die in §16 Abs.6 BStatG enthaltene Beschränkung des Empfängerkreises von anonymisierten Mikrodaten auf den wissenschaftlichen Kontext bedeutet jedoch, daß die Analyse der faktischen Anonymität nicht vor dem Hintergrund eines beliebig zur Verfügung stehenden Zusatzwissens erfolgen muß, sondern spezifisch auf den Datenempfängerkreis ausgerichtet sein kann. Deshalb wird in Kapitel 6 untersucht, welche Arten und Qualitäten von relevantem Zusatzwissen in den wissenschaftlichen Institutionen, an welche die Datenlieferung erfolgen kann, vorhanden oder typischerweise zugänglich ist.

Wie bereits mehrfach betont, bemißt das BStatG die faktische Anonymität an dem Unverhältnismäßigkeitskriterium des Aufwandes an Zeit, Kosten und Arbeitskraft. Da diesem Kriterium eine entscheidende Bedeutung zukommt, wird es in Kapitel 7 ausführlich erörtert. Dabei werden auch die Elemente konkretisiert, die in einer entsprechenden Kosten-Nutzen-Analyse zu berücksichtigen sind.

Der zentrale Kern der Untersuchung liegt in den Kapiteln 8 bis 11. In Kapitel 8 werden aus der Analyse der denkbaren Deanonymisierungsmotive sowie aus den Untersuchungen des Zusatzwissens fünf Gefährdungssituationen abgeleitet. Für sie soll dann näher überprüft werden, welches Deanonymisierungsrisiko und welche Kosten-Nutzen-Bilanz faktisch besteht. Bisherige Untersuchungen basieren im wesentlichen entweder auf Simulationen mit synthetisch erzeugten

Daten oder auf wahrscheinlichkeitstheoretischen Berechnungen. Im Vergleich dazu wird in Kapitel 9 insofern ein neuer Weg beschritten, als in empirischen Deanonymisierungsexperimenten möglichst realistisch diejenigen Arbeitsschritte nachgestellt werden, die ein Datenangreifer ebenfalls vollziehen müßte. Die Realitätsnähe der durchgeführten Experimente besteht vor allem darin, daß mit verschiedenen Deanonymisierungstechniken versucht wird, aus einem aus tatsächlich erhobenen Daten bestehenden Mikrodatenfile Zuordnungen zu einem ebenfalls reale Daten enthaltenden Identifikationsfile vorzunehmen. In datenschutzgerechter Weise wird von einem Treuhänder überprüft, ob die Zuordnungen richtig oder falsch sind. Die Untersuchungen sind dabei so angelegt, daß Bedingungen untersucht werden, die nach dem bisherigen Wissen als besonders riskant erscheinen, und daß die aus der Literatur ableitbaren Hypothesen überprüft werden können. Auf der Grundlage der experimentellen Ergebnisse werden dann in Kapitel 10 für weitere Szenarien Kosten-Nutzen-Berechnungen argumentativ durchgeführt. Schließlich wird in Kapitel 11 empirisch die Schutzwirkung einzelner Anonymisierungsmaßnahmen analysiert. Damit werden in diesen Experimenten erstmals in größerem Umfang Reidentifikationsrisiken und Anonymisierungsregeln auf der Basis von realistischen Szenarien und Realdaten überprüft und bewertet. Ihre Ergebnisse können an die Stelle der hypothetischen Annahmen und entsprechend oftmals unrealistischen Dramatisierungen treten, die die bisherige Diskussion vielfach beherrscht haben.

Eine wesentliche Festlegung des §16 Abs.6 BStatG besteht darin, daß faktisch anonyme Mikrodaten der amtlichen Statistik nur an "Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger Forschung" übermittelt werden dürfen. Aus diesem Grunde kann die Untersuchung Randbedingungen unterstellen, die in solchen Institutionen gegeben sind. Sie muß weder ein potentiell unbeschränktes Zusatzwissen noch unkalkulierbare Motive eines hypothetischen Angreifers oder unbeschränkt hohe Mittel zugrundelegen. Solche Annahmen unterstellen aber, daß gesichert ist, daß gelieferte Daten den Bereich der Wissenschaft nicht verlassen bzw. nur das datenempfangende Institut Zugriff auf Mikrodaten hat (§16 Abs. 8 BStatG). Kapitel 12 befaßt sich deshalb mit den Bedingungen der Datenverarbeitung innerhalb des Wissenschaftsbetriebes und prüft technisch-organisatorische Sicherheitsmaßnahmen, die verhindern oder es zumindest erschweren können, daß Daten den Wissenschaftsbereich verlassen. Solche Maßnahmen sind auch insofern für die faktische Anonymität relevant, als sie den Aufwand eines potentiellen Datenangreifers erhöhen oder einen Angriffsversuch eher erkennbar werden lassen.

Auf der Grundlage einer zusammenfassenden Würdigung der Ergebnisse der verschiedenen Untersuchungen in Kapitel 13 werden schließlich in Kapitel 14 Empfehlungen für Anonymisierungsmaßnahmen abgeleitet, die sowohl im Sinne des Datenschutzes wirksam sind als auch die vielfältigen Interessen der Forschung berücksichtigen. Diese Empfehlungen beziehen sich auf den Mikrozensus und die Einkommens- und Verbrauchsstichprobe.

Wenn viele Einzelbefunde dieser Untersuchungen durchaus von allgemeiner Gültigkeit sind, so muß dennoch nachdrücklich darauf hingewiesen werden, daß eine eins-zu-eins-Übertragung auf andere Daten- und Bedingungskonstellationen in der Regel nicht möglich sein wird. Die Untersuchungen können Anhaltspunkte geben. Bei ihrer Übertragung auf andere Situationen müssen vor allem die jeweils spezifischen Bedingungen sowohl der in den Mikrodaten enthaltenen Informationen als auch des verfügbaren Zusatzwissens, das sich mit diesen Informationen überschneidet, sorgfältig untersucht werden. Auch werden sich je nach Forschungszweck unterschiedliche Anonymisierungsmaßnahmen als günstig und das Analysepotential der Daten möglichst wenig einschränkend erweisen.

2 Zum Stand der Forschung und zur gegenwärtigen Praxis der Datenweitergabe*

2.1 Forschungsstand

Die Durchsicht der Literatur zum Forschungsgebiet "Anonymisierung" statistischer Daten kann in einer groben Zuordnung in drei Felder eingeteilt werden. Einmal wird diese Thematik in ihrer grundsätzlichen Bedeutung aufgegriffen. Als Stichworte seien nur genannt: Geheimhaltung, Schutz der Privatsphäre, Akzeptanz bei statistischen Erhebungen, Recht auf informationelle Selbstbestimmung aber auch auf freien Zugang zu Informationen usw. Hierzu können auch Fragen der Definition einer Reidentifikation (Disclosure) gezählt werden. Das zweite Feld betrifft die Anonymisierungsmaßnahmen. Dabei wird meistens eine Zweiteilung zwischen "Tabellen" einerseits und "Mikrodatensätzen" andererseits vorgenommen. Die wissenschaftlichen Ausarbeitungen dazu befassen sich überwiegend mit den Fragen:

- Wie kann Anonymität erreicht werden (Anonymisierungsverfahren)?
- Welche Reidentifikationsrisiken verbleiben bei den einzelnen Verfahren?
- Welche Auswirkungen haben Anonymisierungsverfahren (z.B. Informationsverluste, Auswirkungen auf das Analysepotential)?

Zur Frage der "Deanonymisierung" (Strategien, Methoden, Verfahren usw.), dem dritten Feld, ist die Literatur dagegen äußerst spärlich, und die Recherchen haben ergeben, daß speziell zu diesem Thema gegenwärtig wahrscheinlich nur wenig Forschung betrieben wird. Überwiegend handelt es sich bei den wissenschaftlichen Publikationen um theoretische Abhandlungen über bestehende Reidentifikationsrisiken bei unterschiedlichen Annahmen und Hypothesen.

---

* Autor: Peter Knoche

2.1.1 Forschungsschwerpunkte im Inland

Nach einer Reihe früherer Arbeiten[2] wurde in der Bundesrepublik Deutschland die neuere Diskussion über potentielle Reidentifikationsrisiken im wesentlichen durch zwei äußerst unterschiedliche Studien bestimmt. Hierbei handelt es sich zum einen um die im Vorfeld der Volkszählung 1987 an der Universität Hamburg im Fachbereich Informatik durchgeführte Untersuchung "Zur Anonymität und Reidentifizierbarkeit statistischer Daten" (Fischer-Hübner 1986, Brunnstein 1987); zum anderen um das von der Gesellschaft für Mathematik und Datenverarbeitung (GMD) durchgeführte AIMIPH-Projekt "Konstruktion und Erprobung eines anonymen integrierten Mikrodatenfile der bundesdeutschen Privathaushalte" unter Federführung von Paaß/Wauschkuhn (1985).

Grundlage der Hamburger Untersuchung war ein künstlich generiertes Mikrodatenfile mit 100 000 Einzeldatensätzen und 21 Merkmalen.[3] Diese Datenbank wurde spezifischen D-BaseIII-Selektionsroutinen mit dem Ziel unterworfen, Datensätze mit einzigartigen Ausprägungskombinationen zu bestimmen. Personen wurden dann als reidentifiziert angesehen (Fischer-Hübner 1986:57), "wenn sie innerhalb des Datenbestands die einzigen mit einer bestimmten Kombination von direkt über sie und eventuell über Mitbewohner gespeicherten Merkmalswerten waren."

Ausgehend von den Ergebnissen der Selektionsroutinen wurden folgende allgemeine Aussagen getroffen (Fischer-Hübner 1986):

- 5 Prozent der Erwerbspersonen sind anhand des Geburtsjahres/Geburtshalbjahres, Geschlecht und Berufsgruppe reidentifizierbar.

- Etwa 50 Prozent sind mit wenigen weiteren Merkmalen, wie Schul- und Ausbildungsabschlüsse sowie Wirtschaftszweig reidentifizierbar.

- Nur ca. 10 Prozent der Erwerbspersonen können nicht anhand der direkt über sie gespeicherten Merkmale reidentifiziert werden, sondern nur wenn zusätzlich haushaltsbezogene Informationen hinzukommen.

---

2 Vgl. u.a. Brennecke 1980, Schlörer 1977, 1978a, 1978b, Kaase et al 1980.

3 Identifikationsnummer, Geburtsjahr, -monat, -tag, Geschlecht, Staatsangehörigkeit, Familienstand, Religion, Schulabschluß, Schulbesuch, Ausbildungsabschluß, Studium, Beschäftigung, Beruf, Wirtschaftszweig, Hausnummer, Gebäudenummer, Ehepartner, Mutter, Vater, Einzelhaushalt (Fischer-Hübner 1986:25).

Derartige Folgerungen sind allerdings insofern problematisch, als sie nur durch den Befund von einzigartigen Datensätzen in einem *künstlich generierten Datenfile* gestützt werden, d.h. nur einer von vielen Bestimmungsfaktoren für die Bewertung eines Reidentifikationsrisikos herangezogen wird. Es wird weder die Zugänglichkeit von Zusatzwissen (Identifikationsfile) als notwendige Voraussetzung für eine Reidentifikation, noch die sich daran direkt anschließende Frage der Kompatibilität von Mikrodatenfile und Identifikationsfile problematisiert. Die Folgerung, daß "eine Reidentifikation in den allermeisten Fällen durch einen einfachen Vergleich auf einzelne Merkmale möglich ist" (a.a.O:58), muß auf dem Hintergrund dieser die Realität sehr simplifizierenden Annahmen doch als reichlich überzogen gewertet werden, jedenfalls kann unter Rückgriff auf die zweite wichtige, zeitlich noch vor der Hamburger Untersuchung durchgeführten Studie, dem AIMIPH-Projekt, die Vermutung ausgesprochen werden, daß ein einfacher Abgleich von Merkmalsausprägungen keinesfalls Reidentifikationen in größerem Ausmaß ermöglicht. Diese von der GMD durchgeführte Untersuchung hat auf dem Gebiet der Anonymität bzw. der Verfahren zur Reidentifikation wesentliche Pionierarbeit geleistet und fand nicht nur im nationalen sondern auch im internationalen Bereich große Beachtung. Das in diesem Projekt entwickelte Verfahren ist bisher das in sich geschlossenste und methodisch anspruchvollste auf diesem Forschungsgebiet.

Aufgabe des AIMIPH-Projekts war es, praktikable Kriterien für die *absolute* Anonymität von Daten zu erarbeiten und einen Beitrag zu leisten für die Schaffung von Regelungen und Verfahren, die der Wissenschaft und Planung den Zugang zu Einzelangaben eröffnen sollten. Ausgangspunkt war, analog zu der vorliegenden Untersuchung, daß ein Datenangreifer Datensätze eines Mikrodatenfile reidentifizieren will. Dazu benötigt er aus anderen Quellen Zusatzwissen in Form eines Identifikationsfile, welches gemeinsame Merkmale mit dem Mikrodatenfile enthält, sowie direkte Identifikationsmerkmale, so daß die Identität der Datensätze im Identifikationsfile dem Angreifer bekannt ist. Der Angreifer kann nun die Überschneidungsmerkmale der Datensätze des Identifikationsfile mit denen des Mikrodatenfile vergleichen. Stimmen diese für zwei Datensätze überein, so kann zumindest auf den ersten Blick davon ausgegangen werden, daß beide Datensätze von der gleichen Person bzw. dem gleichen Haushalt oder einer anderen Erhebungseinheit stammen. Damit wäre ein Datensatz des Mikrodatenfile deanonymisiert. Der Angreifer kann die Informationen in den restlichen Merkmalen, die nicht Überschneidungsmerkmale sind, für seine Zwecke nutzen. Doch die Realität läßt sich nicht so einfach abbilden, denn durch Datenfehler bzw. Dateninkompatibilitäten und dem Auftreten von Doppelgängern (bei Stichproben) ist eine Identität der Überschneidungsmerk-

male im Mikrodatenfile und Identifikationsfile nicht immer gleichbedeutend mit einer gelungenen eindeutigen und richtigen Zuordnung.

Paaß/Wauschkuhn (1985) haben hierzu ein Verfahren entwickelt, das sowohl Lösungen für die Dateninkompatibilitätsproblematik wie für die Doppelgängerproblematik verspricht. Die von Paaß/Wauschkuhn (1985) entwickelten Methoden werden, da sie auch für das vorliegende Projekt von großer Bedeutung sind, in Abschnitt 3.4 im einzelnen dargestellt und gewürdigt. Ihre Leistungsfähigkeit wird später auch empirisch überprüft. An dieser Stelle soll nur ein knapper Überblick über die Anlage, Ergebnisse und Implikationen der Studie von Paaß/Wauschkuhn gegeben werden.

Für die Ausrichtung der hier vorliegenden Studie ist die von Paaß/Wauschkuhn angewandte Szenarien-Technik von großer Relevanz. Paaß/Wauschkuhn orientieren ihre Arbeit an sechs unterschiedlichen Szenarien, die sie im wesentlichen aus drei Faktoren konstruieren, die für das Reidentifikationsrisiko von zentraler Bedeutung sind. Diese ergeben das Motiv des Datenangreifers, den Umfang und Informationsgehalt der Überschneidungsmerkmale, die ihm im Rahmen des ihm zugänglichen Zusatzwissens zur Verfügung stehen, sowie schließlich - stark von den bereits genannten Faktoren abhängig - die beim Datenangriff verfolgte Deanonymisierungsstrategie.

Im einzelnen haben Paaß/Wauschkuhn (1985) die folgenden sechs Szenarien entwickelt und überprüft:

1)	Staatsanwalts-Szenario:	Gezielte Suche mit 68 Überscheidungsmerkmalen,
2)	Steuerfahndungs-Szenario:	Gezielte Suche mit 45 Überschneidungsmerkmalen,
3)	Kripo-Szenario:	Gezielte Suche mit 15 Überschneidungsmerkmalen,
4)	Journalisten-Szenario:	Einzel-Fischzug mit 10 Überschneidungsmerkmalen,
5)	Adreßverlag-Szenario:	Massen-Identifikation mit 7 Überschneidungsmerkmalen,
6)	Konzern-Szenario:	Massen-Identifikation mit 11 Überschneidungsmerkmalen.

Aus dieser Übersicht geht hervor, daß neben der verfolgten Deanonymisierungsstrategie die Szenarien sich vor allem in der Zahl der berücksichtigten Überschneidungsmerkmale - in Anlehnung an das bei einem bestimmten Angreifer unterstellte Zusatzwissen - voneinander unterscheiden. Dabei ist nicht

nur die Zahl der Überschneidungsmerkmale bedeutsam, sondern im einzelnen auch ihr Differenzierungsgrad (siehe dazu im einzelnen: Paaß/Wauschkuhn 1985:79,82). Aber nicht nur Art, Zahl der Merkmale und Strategie wurden variiert; darüber hinaus wurden verschiedene Sicherheitsschwellen für korrekte Zuordnung eingebaut. Ferner wurden die Höhe der Datenfehler und das Wissen des Angreifers um diese Fehler variiert. Außerdem wurden die beiden Alternativen, daß der Angreifer von der Teilnahme des Gesuchten an einer Erhebung weiß (damit entfällt der Schutz der Stichprobeneigenschaft), bzw. daß er davon nichts weiß, in die Rechnungen aufgenommen.

Als Datenmaterial standen Einzelangaben aus der Einkommens- und Verbrauchsstichprobe 1978 und für das Konzern-Szenario Einzelangaben des Mikrozensus 1978 aus drei Bundesländern zur Verfügung. Aus diesem Material wurde sowohl das Mikrodatenfile als auch das Identifikationsfile gebildet, letzteres also synthetisch als Surrogat für Zusatzwissen aus Verwaltungsakten, polizeilichen Ermittlungen, öffentlichen Registern usw.

Tabelle 2.1: Anteil der erfolgreichen Reidentifikationsversuche bei normalen Datenfehlern in Prozent

Szenario (Anzahl der ÜM)	Fehlerverteilung des Angreifers	Variante des Szenarios: Suche vorgegebener Haushalte MDF-Teiln. bekannt ? nein[2]	Massenfischzug[1] ja[3]	
Staatsanwalt (68 MM)	exakt	0.11	43.0	- [5]
Steuerfahndung (45 MM)	exakt	0.11	41.0	17.0
	vierfach	0.11	42.0	16.0
	zufällig	0.12	43.0	17.0
Kripo (15 MM)	exakt	0.01	22.0	2.5
	vierfach	0.02	21.0	4.0
	zufällig	0.02	23.0	4.3
Adreßverlag (7 MM)	exakt	0.00	0.0	- [5]
	zufällig	0.00	0.0	- [5]
Konzern (11 MM des Mikrozensus)	exakt[4]	0.00	0.2	- [5]

Quelle: Paaß/Wauschkuhn, a.a.O.:83

1) umfaßt 25 Prozent der Grundgesamtheit; Sicherheitsschwelle 99,0 Prozent
2) Sicherheitsschwelle 99,9 Prozent
3) Es wurde nur eine Teilstichprobe freigegeben; EVS: 50 Prozent Mikrozensus: 20 Prozent; Sicherheitsschwelle 90 Prozent
4) Es lagen keine Datenfehler vor
5) Nicht ausgewertet

Das Journalisten-Szenario ist nicht in die Tabelle 2.1 aufgenommen worden, da es nicht sinnvoll erscheint, die Anzahl der reidentifizierten Haushalte auf die Population hochzurechnen, wie es bei den anderen fünf Szenarien der Fall ist, da keine Zufallsstichprobe von Haushalten untersucht wurde, sondern die be-

sonders gefährdeten Haushalte ausgesucht wurden. Bei unterstelltem normal hohem Fehlerniveau und einer exakten Kenntnis des Anteils über den tatsächlichen Datenfehler wurden drei der 25 Haushalte mit "extremen" Merkmalsausprägungen identifiziert. Die Tabelle zeigt, daß der Informationsgehalt der Überschneidungsmerkmale die wesentliche Determinante für die Deanonymisierungsquote ist. Bei den Szenarien mit weniger als 15 Überschneidungsmerkmalen war der Informationsgehalt so gering, daß keine Einzeldatensätze zugeordnet werden konnten. Bei einem höheren Informationsgehalt, etwa 15 bis 20 Merkmale, konnte dagegen ein teilweise beträchtlicher Anteil der Einzeldatensätze deanonymisiert werden; deshalb führt bei umfangreichem Zusatzwissen eine Freigabe mit sämtlichen oder sehr vielen Merkmalen einer Erhebung selbst bei erheblicher Verfälschung - so die Autoren - zu einem hohen potentiellen Deanonymisierungsrisiko. Außerdem zeigt sich, daß bei unterstellter Kenntnis von der Teilnahme an der das Mikrodatenfile bildenden Erhebung trotz Ziehung einer Teilstichprobe, um die Schutzwirkung der Stichprobeneigenschaft wieder herzustellen, ein wesentlich höherer Prozentsatz der Haushalte reidentifiziert werden konnte, als bei der Variante, bei der die Teilnahme des Haushaltes nicht bekannt ist. Das Ziehen einer Stichprobe allein scheint daher bei diesen Fällen nicht ausreichend zu sein, um das Deanonymisierungsrisiko genügend zu vermindern.

Bei der Beurteilung der Ergebnisse sind Aspekte zu beachten, die tendenziell zu einer Unterschätzung und solche, die zu einer Überschätzung der Erfolgsquote gelungener Reidentifikationen führen, und zwar:

- das grundlegende Datenmaterial war "klassisch anonymisiert", d.h. es wurde folgenden Anonymisierungsmaßnahmen unterzogen:

  * keine direkten Identifikatoren,
  * keine Regionalangaben (bei EVS vier Gemeindegrößenklassen),
  * EVS-Material war eine 90 Prozent-Stichprobe,
  * teilweise Vergröberung der Merkmalsausprägungen (bei Mikrozensus z.B. nur zwei Staatsangehörigkeiten, bei EVS bei 16 diskreten Merkmalen, um eine zweifache Besetzung jeder Merkmalsausprägung zu erreichen);

- das Identifikationsfile wurde synthetisch erzeugt, d.h. es wurde nicht mit realistischem Datenmaterial gearbeitet; insofern liegt keine durchgängig

empirische Untersuchung vor, sondern eher eine empirisch orientierte Modellrechnung,

- einige der getroffenen Annahmen sind nicht sehr realitätsnah, wie z.B. das exakte Wissen des Angreifers über die vorliegenden Datenfehler,

- es ist zu vermuten, daß die Erhebungsfehler und andere Dateninkompatibilitäten unterschätzt wurden,

- speziell bezogen auf die Wissenschaft ist zunächst erstaunlich, daß von den untersuchten Szenarien keines auf die spezifische Situation in der Wissenschaft ausgerichtet ist, obwohl es eine der Zielsetzungen der Untersuchung war, speziell für die Wissenschaft die Bedingungen beim Zugang zu Einzelangaben zu verbessern. Man kann jedoch feststellen, daß die für eine erfolgreiche Reidentifikation benötigte Zahl von Überschneidungsmerkmalen, d.h. der Umfang des benötigten Zusatzwissens, in der Regel im Wissenschaftsbereich nicht vorhanden ist. Bei dem noch am ehesten mit der Situation von Wissenschaft und Forschung vergleichbaren Adreßverlags-Szenario ergaben sich keine korrekten Reidentifikationen. Selbst bei der äußerst unplausiblen Annahme, daß die Teilnahme an der EVS für eine Vielzahl von Haushalten (Massenidentifikation) dem Adreßverlag bekannt war, ergab sich bei einer Sicherheitsschwelle von 90 Prozent zwar eine Reihe von Falschzuordnungen, aber bei keiner der untersuchten Sicherheitsschwellen von 90, 99 und 99,9 Prozent konnte ein Haushalt richtig zugeordnet werden.

Paaß und Wauschkuhn haben im Anschluß an ihre Analysen auch einzelne Anonymisierungsmaßnahmen vorgestellt und ihre Vor- und Nachteile diskutiert, insbesondere die Bildung von Merkmalsscheiben, die Erzeugung von Aggregat-Haushalten (Mikroaggregation) und die Bildung synthetischer Haushalte. Den letzten Ansatz griff Paaß (1989) wieder auf und entwickelte einen stochastischen Algorithmus zur Entwicklung eines synthetischen Datensatzes. Die Generierung synthetischer Datensätze ist auch Gegenstand eines Artikels von Radtke (1988). Das darin vorgestellte Verfahren beruht zum einen auf der Zusammenfassung von Erhebungseinheiten des zu anonymisierenden Materials in verwandten Gruppen über einen geeigneten Nachbarschaftsbegriff und der daran sich anschließenden zyklischen Vertauschung eines Teils der Merkmale von mehreren benachbarten Erhebungseinheiten innerhalb verwandter Gruppen. Das Verfahren wurde anhand von Mikrozensusmaterial getestet. Der erste Verfahrensschritt greift dabei auf Forschungen zurück, die im Statistischen

Bundesamt zur Bildung von Tripel-Aggregaten (Mikroaggregation) durchgeführt werden und Grundlage für die Freigabe sog. Kurzbandsätze - das sind Merkmalsscheiben mit ca. 10 bis 15 Variablen - wurden (Kühn 1987, Kühn/Radtke 1987).

2.1.2 Forschungsschwerpunkte im Ausland

Die folgenden Ausführungen können und sollen kein erschöpfender Bericht über die Forschung im Ausland zum Problem der Übermittlung von Einzeldatensätzen sein, sondern beschränken sich bewußt auf die Darstellung der wichtigsten erkennbaren Forschungslinien und -tendenzen.

Die Literaturrecherchen beschränkten sich auf die Quellen, die sich aus langjähriger Erfahrungspraxis mit der Anonymisierung statistischer Daten als einschlägig herausgebildet haben. Hervorgehoben werden sollen insbesondere die Berichte zum jährlichen Treffen der Amerikanischen Statistischen Gesellschaft und einschlägige Publikationen im Journal of the American Statistical Association/JASA sowie die Tagungsberichte über die seit 1985 jährlich stattfindende Forschungskonferenz des US Bureau of the Census, bei der - mit Ausnahme von 1986 - stets auch Fragen zu Verfahren zur Anonymisierung bzw. deren Auswirkungen auf die Datennutzung diskutiert werden. Eine Umfrage bei ausgewählten nationalen statistischen Ämtern zusammen mit der Frage nach der Weitergabepraxis von Einzeldatenmaterial war leider nicht sehr ergiebig.

2.1.2.1 Allgemeine Beiträge zur Frage der Anonymität und Anonymisierung

Wichtiger Ausgangspunkt für Forschungsaktivitäten zur Anonymisierung war der Aufsatz von Fellegi (1972) zur Frage der statistischen Vertraulichkeit. Am Ende seiner Überlegungen - die im Kern Probleme der Geheimhaltung bei aggregierten Daten betreffen - spricht sich der Verfasser tendenziell ablehnend zu einer Übermittlung von Einzeldatensätzen aus. Bei Einzeldatensätzen werde es sich - so der Verfasser - wegen der Vielzahl der Variablen stets um "Unikate" handeln und die Reidentifikation sei daher relativ leicht möglich und nur Kosten-Nutzen-Überlegungen könnten einen potentiellen Angreifer von einem derartigen Unterfangen abhalten. Fellegi stellt dabei auf das erste bekanntgewordene

public use file ab, welches das US Bureau of the Census 1963 aus dem Volkszählungsmaterial von 1960 herausgab.

Eine wegweisende Publikation über die Problematik und die angewandte Praxis einzelner Verfahren der Anonymisierung und der Datenweitergabe veröffentlichte das US Bureau of the Census (US Department of Commerce, 1978).[4] Dabei zeigt sich, daß die Anonymisierungsmaßnahmen zunächst hauptsächlich auf der Grundlage von Plausibilitätsüberlegungen sowie praktischer Erfahrungen mit statistischem Material entstanden und nicht aus stringenten methodischen Modellen abgeleitet wurden. Schon zu diesem Zeitpunkt wurde jedoch deutlich, daß das "Fellegi Kriterium" (no unique records in the file) ein unrealistischer Standard ist, da damit eine Weitergabe nicht möglich ist und es zur Sicherung der Geheimhaltung auch nicht nötig ist. Publikationen zur methodischen Weiterentwicklung von Anonymisierungsmaßnahmen waren dann recht zahlreich und für die statistischen Ämter auch sehr hilfreich. So wurde es z.B. von Cox (1986) aus der Sicht der statistischen Ämter ausdrücklich begrüßt, daß Duncan und Lambert (1986), zunächst mit Priorität für Tabellen, dann erweitert auf die Freigabe von Mikrodaten (1987), ein entscheidungstheoretisches Rahmenwerk (Disclosure Limitation Framework) entwickelt haben, das zwar den Ansprüchen an eine allgemeingültige methodologische Grundlage nicht voll entsprechen kann, aber trotzdem für die Entwicklung und Evaluierung von Anonymisierungsmaßnahmen durch die statistischen Ämter hilfreich sein kann - hilfreich nicht nur für die Abschätzung verbleibender Reidentifikationsrisiken, sondern auch zur Vermeidung zu weitgehender Anonymisierungen. Duncan und Lamberts (1986) Konzeption beruht allerdings auf einem sehr weitgefaßten "Disclosure"-Verständnis. Der im angelsächsischen Sprachgebrauch übliche Begriff "Disclosure" (Enthüllung) ist grundsätzlich weitergehend als der in der vorliegenden Arbeit verwandte Zuordnungsbegriff, der sich an der dem Bundesstatistikgesetz zugrunde liegenden Vorstellung der eindeutigen Zuordnung eines Datensatzes zu einer bestimmten Person, orientiert (vgl. dazu ausführlich Abschnitt 3.1).

Ein anderes relativ breit diskutiertes Problem in wissenschaftlichen Publikationen zur Anonymisierung ist die Frage, ob Anonymisierungsmaßnahmen am Datenmaterial vorgenommen oder ob nicht vertragliche Regelungen präferiert werden sollen, z.B. Verbot der Weitergabe an Dritte und der Zusammenführung mit anderen Datenfiles (matching procedures). Dabei wird auch auf die einschlägigen Bestimmungen des Bundesstatistikgesetzes (§§ 21, 22) verwiesen

4 Zu nennen wäre allerdings auch Dalenius (z.B. 1977, 1984, 1988) mit einer Reihe von Veröffentlichungen umfassender Art aber auch zu Einzelproblemen.

(Gates 1988:251 ff.), in dem die allgemeinen - vertraglich festzulegenden - Sicherungsvorkehrungen in groben Zügen bereits vorgegeben sind. Bei der gegenwärtigen Rechtslage müssen sowohl solche vertraglichen Sicherungsvorkehrungen, als auch Anonymisierungsmaßnahmen am Datenmaterial vorgegeben werden.

2.1.2.2 Forschung über einzelne Anonymisierungsmaßnahmen

Bekanntlich gibt es eine Vielzahl unterschiedlicher Anonymisierungsmaßnahmen, die auch überwiegend in irgendeiner Form bei den statistischen Ämtern zur Anwendung kommen. Von den verschiedenen Konzepten, die Einzelmaßnahmen zu ordnen, sei hier die von Skinner et al. (1990:13) bzw. Skinner (1990:8) vorgenommene Differenzierung aufgegriffen, in der unterschieden wird zwischen

- "Kontaminierungsmethoden"[5] (contamination methods), erzeugen Daten, die nicht mehr als "repräsentativ" eingestuft werden können. Darunter wird die Bildung synthetischer Datensätze, Mikroaggregation und die Einführung von Fehlern (addition of noise, random noise) verstanden, und

- "Censoring-Methoden", welche die "Integrität" der Daten erhalten - darunter werden (Sub-)Stichprobenziehung, Streichen bzw. Vergröberung bestimmter Variablen subsumiert.

Vor allem die mit merklichen Manipulationen an den originären Daten verbundenen Anonymisierungsmaßnahmen spielen nach unserem Eindruck eine dominierende Rolle in der wissenschaftlichen Auseinandersetzung über Anonymisierungsmaßnahmen. Die Bildung synthetischer Datensätze ist eine Methode, die immer zu deutlichen Einbußen hinsichtlich der Verwendungsmöglichkeit der Daten für wissenschaftliche Zwecke führt. Die Forschung der letzten Jahre geht daher nach und nach in die Richtung, auf den Einsatz dieser Methoden bei Datensätzen, die für sozialwissenschaftliche Forschungen bestimmt sind, zu verzichten und stattdessen nach Wegen zu suchen, wie derartige Maßnahmen auf Wirtschaftsdaten angewendet werden können (Spruill

---

5 Entspricht "perturbation".

1983, Paaß 1989).[6] Bei der Bildung synthetischer Datensätze werden zwei unterschiedliche Ansätze präferiert, zum einen die Bildung von *Algorithmen zur stochastischen Modifikation* von Datensätzen (Paaß 1989) und zum anderen die *Mikroaggregation* mit Hilfe der Bildung von Clustern (Spruill 1983). Vereinfacht ausgedrückt bilden die (gewogenen) Durchschnitte in den Datensätzen eines Clusters die Werte des neuen synthetischen Datensatzes, der dann weitergegeben wird (Cox et al. 1985:209, Greenberg 1989:457). Weitere, erst in neuester Zeit entwickelte Ansätze zur "Maskierung" von Variablen oder Datensätzen sind "*Resampling-Methoden*", insbesondere Bootstrap und Jackknife. Einen einführenden Überblick über die statistischen Aussagen bei komplexen Erhebungsdaten geben Rao und Wu (1988). Anwendung auf das "Disclosure Avoidance"-Problem findet das Bootstrap- und Jackknifeverfahren bei Jones und Adam (1989). Auch bei diesen Ansätzen werden letztlich die originären Daten durch Verfahren zur Gewinnung synthetischer Verteilungen und Werte ersetzt, die dann an Dritte weitergegeben werden. Eine abschließende Beurteilung dieses Verfahrens ist noch nicht möglich. Allerdings erscheint es durchaus überlegenswert, bei nur gering besetzten Wertfeldern, z.B. Einkommens- oder Ausgabengrößen, diese durch einen mit Jackknife gewonnenen Wert zu ersetzen, statt sie zu unterdrücken oder mit anderen zu einer größeren Kategorie zusammenzufassen, wie das derzeit gängige Praxis ist.

Schon lange praktiziert und weit verbreitet ist das *Einschleusen von (zusätzlichen) Fehlern*, z.B. durch eine zufallsgesteuerte Addition oder Subtraktion kleinerer Beträge bei Wertgrößen (random noise). Diese Verfahren werden u.a. eingesetzt bei gering besetzten Tabellenfeldern oder Merkmalsausprägungen bei Mikrodaten (z.B. am oberen bzw. unteren Rand einer Verteilung), aber auch bei Variablen, über die "leicht zugängliches" Zusatzwissen vorhanden ist, beispielsweise bei Merkmalen, die in öffentlichen Registern geführt werden und spärlich besetzt sind. Durch intensive Forschung wurde dieser Ansatz wesentlich verfeinert und außer für kontinuierliche auch für diskrete Variablen weiterentwickelt. Initiator der Forschung war unter anderem das US Bureau of the Census (Greenberg 1990:156). Neben Spruill (1983) hat vor allem Kim (1990) diesen Ansatz weiterentwickelt und seine Auswirkungen für die Nutzung der Daten untersucht. Auch wenn dadurch Fortschritte hinsichtlich der Verbesserung der Qualität derartiger Daten für statistische Analysen erzielt werden - z.B. können die Varianz- und Korrelationsstrukturen der Originaldaten sowie die univariaten Verteilungen erhalten werden (Greenberg 1990:161, Skinner et al.

---

6 Derartige Daten sind bekanntlich besonders schwer zu anonymisieren, mit dem Ergebnis, daß sie nach dem vorliegenden Erkenntnisstand bisher von den nationalen statistischen Ämtern nicht freigegeben werden (vgl. Abschnitt 2.2.1.2).

1990:140) - können derart anonymisierte Daten noch immer nur für sehr spezifische Zwecke Verwendung finden. Als generelle Verfahren bei der Bildung von public use files sind sie jedoch nicht geeignet und sollten daher nur sehr eingeschränkt Verwendung finden.

Verfahren, die nicht an der Verfälschung der Originaldaten ansetzen und in der Praxis verbreitet sind, sind - neben der Ziehung einer Stichprobe - Streichen oder Zusammenfassen bzw. Vergröbern der Merkmale bzw. ihrer Ausprägungen. Die dahinterstehenden Überlegungen gehen davon aus, daß einzigartige oder seltene (exotische) Ausprägungen oder Ausprägungskombinationen problemlos zur Identifizierung der zugehörigen Betroffenen dienen können. Eine derartige Zuordnung auf bestimmte Personen ist sicher dann leicht möglich, wenn die "Einzigartigkeit" so offensichtlich ist, wie bei dem oft genannten Beispiel der Berufsangabe "Bundeskanzler".[7] Die Vermeidung der Weitergabe einzigartiger oder seltener Ausprägungen bzw. Ausprägungskombinationen ist gängige Praxis bei den statistischen Ämtern. Dies ist darauf zurückzuführen, daß durch entsprechende gesetzliche Bestimmungen eine Zuordnung der für statistische Zwecke erhobenen Daten auf die Betroffenen bei Publikation bzw. Weitergabe der Ergebnisse nicht mehr möglich sein darf. Ein anderer Grund ist das Eigeninteresse der statistischen Ämter, die aus Akzeptanzgründen den Befragten zusichern, daß ihre Daten nicht weitergegeben werden bzw. zumindest nicht in einer Form, die einen Rückschluß auf einzelne Individuen oder Wirtschaftseinheiten zuläßt (Keller/Bethlehem 1987, 1990:2 f.).

2.1.2.3 Verfahren zum einfachen Abgleich von Datenfiles

Von Ausnahmen abgesehen sind die Forschungsschwerpunkte auf dem Gebiet der Anonymisierung nicht präzise auf die Probleme abgestellt, die beim einfachen Abgleich von großen Datenmengen mit Hilfe der automatischen Datenverbeitung auftreten, obwohl bei vielen der aufgezeigten Forschungen unterstellt wird, daß potentielle Deanonymisierungsversuche über den Abgleich von Datensätzen aus zwei oder mehreren Quellen erfolgen. Es liegt daher nahe zu eruieren, ob zu dieser Problematik unabhängig von Anonymisierungsproblemen Forschungsergebnisse vorliegen. Der Ausgangspunkt kann dabei durchaus konträr zum bisher betrachteten liegen, indem Probleme untersucht werden , die beim erlaubten Zusammenführen unterschiedlicher Datenbasen entstehen,

---

7 Zur Erläuterung: Die Kategorie "Bundeskanzler" existiert in der amtlichen Statistik nicht.

z.B. im Rahmen polizeilicher Fahndungsmaßnahmen. Die Recherche war jedoch nicht ergiebig, da bei den Untersuchungen ein Bezug zu statistischen Fragestellungen und Methoden bestehen sollte. Hingewiesen wird deshalb auf eine Untersuchung neueren Datums von Copas und Hilton 1990. Dabei wurden anhand empirischer Daten, die bei der Einreise nach Großbritannien (landing card) und Ausreise (embarkation card) anfielen, die Probleme bei der Zusammenführung untersucht und Lösungsvorschläge mit Hilfe von Modellen aus der Wahrscheinlichkeitstheorie diskutiert. Im Mittelpunkt steht die Fehlerhaftigkeit des Materials und die Frage, wie mit Hilfe der erwähnten Modelle ein Ausgleich geschaffen werden kann. Erhoben werden neun Variablen (Nationalität, Geschlecht, Geburtsjahr, -monat und -tag, Familienname, 1., 2. und 3. Vorname). Dabei ist zu berücksichtigen, daß die Namensangaben nach einem phonetischen Code-System verschlüsselt und damit auch fehleranfällig werden. Ohne die Prozeduren hier näher beschreiben zu wollen, ist es interessant, festzustellen, daß bei 8 601 Datensätzen, die eigentlich hätten identisch sein müssen, 3 999, also knapp die Hälfte, Diskrepanzen aufweisen, und das bei erhebungstechnisch so einfachen Variablen. Einen Fehler wiesen rund 23 Prozent der Datensätze auf und bei rund 24 Prozent wurden sogar mehrere festgestellt; die Spitze lag bei acht Fehlern bei insgesamt neun Variablen, allerdings nur bei einem Datensatz (Copas/Hilton 1990:291). Die anschließend erörterten statistischen Modelle zur Überwindung dieses "Fehlerhandicaps" stehen natürlich auch einem Angreifer zur Verfügung; doch welche theoretischen und praktischen Probleme dabei auftreten und welcher Aufwand an Zeit, Kosten und Arbeitskraft sich damit grundsätzlich verbindet, ist unter anderem Gegenstand der vorliegenden Studie.

2.1.2.4 Forschungsarbeiten zum Anonymisierungsproblem in den nationalen statistischen Ämtern

Für alle statistischen Ämter ist die Frage des Datenschutzes bei der Datenweitergabe ein zentrales Problem. Über Forschungsaktivitäten auf dem Gebiet der Anonymisierung in den statistischen Ämtern ist jedoch nur wenig bekannt. Die folgenden Ausführungen decken allerdings mit Sicherheit nicht das gesamte Spektrum gegenwärtiger Forschung ab. Die Durchsicht der deutschsprachigen und angelsächsischen Literatur legt jedoch die Vermutung nahe, daß in den meisten statistischen Ämtern der Bereich "Anonymisierung" kein Forschungsschwerpunkt ist. Dennoch sind in einigen Ämtern in den letzten Jahren vermehrte Forschungsaktivitäten zu beobachten. Dabei kommt der

"Uniqueness-Problematik" - in der deutschen Sprachweise häufger als "Doppelgänger-Problem" bezeichnet - eine besondere Bedeutung zu.

Das Niederländische Zentralamt für Statistik konzentriert sich schon seit vielen Jahren auf das "Uniqueness-Konzept". In mehreren wissenschaftlichen Publikationen haben Mitarbeiter des Amtes das Konzept dargestellt und weiterentwickelt sowie die Anwendungsbreite vergrößert (z.B. Bethlehem/Keller/Pannekoek 1988 und 1990, Willenborg/Mokken/Pannekoek 1990, Koolmann/Willenborg 1990). Diese Autoren haben ein fortgeschrittenes formales Model entwickelt, in dem mehrere für das Reidentifikationsrisiko wichtige Elemente verbunden werden; u.a.

- die Wahrscheinlichkeit, daß eine Person in den für eine Reidentifikation verwendbaren Schlüsselvariablen eine populations-einmalige Ausprägungskombination aufweist,
- die Wahrscheinlichkeit, daß ein Angreifer diese Person kennt und weiß, daß ihre Ausprägungskombination einmalig ist,
- die Wahrscheinlichkeit, daß die Person Teil des dem Angreifer zugänglichen Mikrodatenfile ist.

Zum Teil auf weitreichenden Annahmen basierend werden diese Wahrscheinlichkeiten geschätzt und daraus das Reidentifikationsrisiko bei der Weitergabe von Einzelangaben mit unterschiedlichen Differenzierungsgraden der Schlüsselvariablen berechnet. Die Autoren versuchen aus den Modellrechnungen insbesondere abzuleiten, mit welcher regionalen Gliederungstiefe solche Einzelangaben ohne Reidentifikationsgefahr an die Öffentlichkeit weitergegeben werden können.

Skinner et al. (1990) und Skinner (1990) haben dieses Modell modifiziert, u.a. durch die Berücksichtigung von Informationsfehlern realistischer gemacht und durch den Versuch der empirischen Bestimmung einzelner Parameter, für die Anwendung bei der Weitergabe von Einzelangaben aus dem Britischen Census, konkretisiert. Als Ergebnis wird herausgestellt, daß ausgehend von der Zielsetzung "to provide a flexible data resource for unknown and unanticipated user demand" (Skinner et al. 1990:140), Methoden zur Datenverfälschung (contamination methods) nicht möglich und auch als Ergebnis der durchgeführten Evaluierung des Reidentifikationsrisikos nicht nötig sind. Zusam-

menfassungen bzw. Vergröberungen von Variablen, z.B. von regionalen Identifikatoren sind ausreichend für den Schutz der Geheimhaltung bei der Freigabe von Mikrodaten aus der Volkszählung 1991 in Großbritannien. In der Version von Skinner et al. (1990) wird dieses Modell in Kapitel 4 aufgegriffen werden. Es wird sich dort als sehr hilfreich erweisen, um die wesentlichen Bestimmungsfaktoren des Reidentifikationsrisikos bei unterschiedlichen Daten- und Informationsvoraussetzungen und unterschiedlichen Deanonymisierungsstrategien miteinander in Beziehung zu setzen und in ihrer Bedeutung zu gewichten.

Das "Uniqueness-Konzept" der Niederländischen Autorengruppe wird gegenwärtig auch in einem Forschungsprojekt des Nationalen Statistischen Amt Italiens auf seine Brauchbarkeit für Volkszählungsdaten getestet. Anhand der Forschungsergebnisse soll dann entschieden werden, ob in Italien Mikrodaten freigegeben werden.

Im Australischen Bureau of Statistics wird ein umfangreiches Forschungsprogramm durchgeführt, das sich unter anderem auch mit Anonymisierungstechniken bei der Freigabe von Einzelangaben in Form sogenannter "unit record tapes" befaßt. Auch hier ist die Eignung des "Uniqueness"-Ansatzes Forschungsschwerpunkt. Ferner wurden Vergleichsforschungen verschiedener Techniken durchgeführt, insbesondere zwischen der vom Amt benutzten Cluster-Methode (Mikroaggregation) und der auf Kim zurückgehende Methode des "random noise".

Ein breites Forschungsspektrum wird vom US Bureau of the Census abgedeckt. Einen Einblick vermitteln Gates (1988) und Greenberg (1990). Da in den USA (und Kanada) für viele sozialwissenschaftlich relevante Erhebungen schon seit langem "public use files" freigegeben werden, richtet sich die Forschung auf die Verbesserung der eingesetzten Anonymisierungsinstrumente, die Suche nach geeigneten Verfahren zur Anonymisierung von Mikrodaten aus dem Wirtschaftsbereich und auf optimale Ersatzmöglichkeiten, falls die Herausgabe eines public use files aus bestimmten Gründen nicht möglich ist. Einer der Schwerpunkte bei der Weiterentwicklung der Anonymisierungsmaßnahmen sind Forschungen zum Problem, wie sich bei unterschiedlicher regionaler Gliederungstiefe die Wahrscheinlichkeit von Einzelfällen verändert (Greenberg 1990:158 f, Greenberg/Voshell 1990).

Das US Bureau of the Census arbeitet eng mit externen Forschern zusammen, z.B. mit der Harvard-Universität in Form eines "Joint Statistical Agreement". Seit 1977 gibt es außerdem ein "Census Research Program", das vom Census

Bureau und der National Science Foundation finanziert und gemeinsam von Bureau und der American Statistical Association organisiert wird. In seinem Rahmen werden auch Probleme, die mit der Freigabe von Mikrodaten verbunden sind, analysiert. Ganz speziell zur Verstärkung der wissenschaftlichen Forschungsansätze und ihrer Ergebnisse bei der Evaluierung der Risiken einer Freigabe von Mikrodaten hat das Census Bureau einen permanenten Beauftragten eingesetzt. Er ist u.a. verantwortlich für die Kooperation mit externen Forschern. Schon allein daraus wird die Bedeutung sichtbar, die diesen Fragen beigemessen wird.

2.1.2.5 Fazit

Die Darstellung der Forschungsaktivitäten auf dem Gebiet der Freigabe anonymisierter Mikrodatensätze, der dabei bestehenden Reidentifikationsrisiken und der Auswirkungen auf die Nutzungsmöglichkeiten der freigegebenen Daten hat gezeigt, daß quantitativ und qualitativ ein beachtlicher Stand erreicht ist, insbesondere durch Forschungen in den achtziger Jahren. Dabei hat die Arbeit von Paaß/Wauschkuhn entscheidende Impulse gegeben. Die Schwerpunkte liegen neben den allgemeinen Fragen zur Anonymität auf der theoretisch-methodologischen Durchdringung spezieller Konzeptionen und Verfahren, ihren Anwendungsmöglichkeiten und der Evaluierung ihrer Restrisiken bzw. ihrer Auswirkungen auf die Datenqualität. Der Untersuchungsgegenstand der vorliegenden Studie, insbesondere die Berechnung und Bewertung des Deanonymisierungsrisikos und des damit verbundenen Aufwands anhand spezieller empirischer Experimente unter realistischen Ausgangsbedingungen, findet jedoch keine Parallelität. Es ist deshalb zu hoffen, daß mit ihren Ergebnissen eine wesentliche Lücke geschlossen werden kann, die auch über die Grenzen der Bundesrepublik hinaus von Bedeutung sind.

2.2 Die gegenwärtige Praxis der Datenweitergabe

2.2.1 Inland

Die vorliegende Untersuchung erfolgt auf dem Hintergrund einer etablierten Praxis der Anonymisierung und Weitergabe statistischer Daten durch die statistischen Ämter. Es ist deshalb sinnvoll, neben dem Stand der Forschung auch einen Überblick über die geübte Praxis und die damit gemachten Erfahrungen

im In- wie im Ausland zu geben. Für die Bundesrepublik wird von der Praxis des Statistischen Bundesamtes ausgegangen. Die Statistischen Landesämter praktizieren jedoch gleiche oder ähnliche Verfahren; zum einen, weil auch für sie bei der Durchführung von Bundesstatistiken die Geheimhaltungsvorschriften des Bundesstatistikgesetzes Geltung haben und zum anderen, weil die statistischen Ämter des Bundes und der Länder bemüht sind, sowohl bei den generellen Anonymisierungsverfahren und Richtlinien der Datenweitergabe, als auch bei der Lösung spezieller Geheimhaltungsprobleme einzelner Statistiken, Übereinstimmung zu erzielen.

Grundsätzlich wird zwischen folgenden drei Formen der Datenweitergabe unterschieden:

a) Tabellen
b) Kurzdatensätze
c) Langdatensätze

Für alle drei Arten der Datenweitergabe gilt entsprechend §16 Abs.1 Nr.4 BStatG, daß Einzelangaben nur übermittelt werden dürfen, wenn sie den Befragten oder Betroffenen nicht zugeordnet werden können (absolute Anonymität). Ausnahmen von diesem Grundsatz der statistischen Geheimhaltung sind in Absatz 1 Nr. 1 bis 3 geregelt. Weitere besondere Vorschriften zur Weitergabe von Einzeldaten enthalten die Absätze 2 bis 6. Absatz 6 regelt, wie oben ausgeführt, die Übermittlung von Einzeldaten für die Durchführung wissenschaftlicher Vorhaben.

Die Weitergabe *absolut* anonymisierter Einzelangaben erfolgt in Form von:

a) *Tabellen*; die Aggregierung von Einzelangaben in Tabellen ist die übliche und häufigste Form der Weitergabe statistischer Ergebnisse.

b) *Kurzdatensätze*; Mikrodatensätze, die nur eine relativ geringe Zahl von Merkmalen aus einem umfangreichen Datensatz enthalten (Merkmalsscheibe).

c) *Langdatensätze*; Mikrodatensätze, die umfangreiche Merkmale bzw. Merkmalsausprägungen aus dem gesamten vorliegenden Mikrodatensatz einer Erhebung enthalten.

ad (a) Bei der Durchführung von Verfahren zur Geheimhaltung gemäß §16 Abs.1 Nr.4 BStatG wird bei Tabellen üblicherweise davon ausgegangen,

daß bei mit "1" oder "2" besetzten Tabellenfeldern eine Zuordnung zum Befragten/Betroffenen nicht zweifelsfrei ausgeschlossen werden kann. Zellen mit derart schwachen Besetzungszahlen werden meistens unterdrückt, wobei durch weitere Maßnahmen sichergestellt wird, daß die geheimzuhaltenden Angaben nicht durch Differenzbildung wieder errechnet werden können (Sekundärgeheimhaltung).

Im Statistischen Bundesamt wurde für *Fallzahlentabellen* aus dem Mikrozensus ein Geheimhaltungsverfahren durch Zufallsüberlagerung entwickelt (Kühn/Pfrommer/Schrey 1984). Das Verfahren überlagert in der Originaltabelle jedes Tabellenfeld mit einer Zufallszahl mit Erwartungswert 0 und Varianz 3. Dabei bleibt die Gesamtmasse der Tabelle "im Mittel" erhalten. Weiterhin ist gewährleistet, daß die tabellierten Verteilungen in ihrer Struktur unverändert bleiben. Bei Fallzahlen unter 50 Einheiten kann es jedoch zu hohen relativen Abweichungen kommen.

Bei *Wertetabellen* (z.B. Umsätze, Investitionsausgaben) findet die sogenannte "Dominanzregel" Anwendung. Sie besagt, daß bei einem Anteil eines Befragten oder Betroffenen an einer zusammengefaßten Angabe von mehr als 85 Prozent diese Angabe geheimzuhalten ist; bei einem Anteil zwischen 50 und 85 Prozent ist die zusammengefaßte Angabe dann geheimzuhalten, wenn die Gefahr besteht, daß dadurch die Verhältnisse eines einzelnen Befragten/Betroffenen offengelegt werden.

Neben diesen generellen Verfahren bestehen noch weitere Anonymisierungsmaßnahmen bei bestimmten Statistiken oder bestimmten Veröffentlichungstabellen einzelner Statistiken, z.B. bei der Todesursachenstatistik.

Die dargestellte Geheimhaltungspraxis bezieht sich auf die Bestimmungen des §16 Abs. 1 Nr.4 BStatG, nicht jedoch auf die Weitergabe von Tabellen für wissenschaftliche Zwecke. Verfahren für eine *faktische Anonymisierung* von Tabellen liegen derzeit noch nicht vor. Überlegungen dazu waren nicht Gegenstand dieser Untersuchung.

ad (b) Auch bei der Weitergabe von Kurzdatensätzen gilt die Grundregel, daß jeder einzelne Kurzdatensatz als Ganzes in mindestens dreifacher Besetzung vorhanden sein muß. Einzeldatensätze werden also so verändert, daß in dem entstehenden anonymisierten Material kein Satz einzig bleibt. Dabei bleibt natürlich das Ziel bestehen, daß der statistische Informationsgehalt möglichst wenig verfälscht wird. Das Vorgehen entspricht dem der Bildung sogenannter "Tripelaggregate" (Kühn 1987, Kühn/Radtke 1987). Das Verfahren kann bezüglich der international üblichen Systematisierung der Geheimhaltungsverfahren als eine Art "Mikroaggregation" eingestuft werden. In einem ersten Schritt wird das Material zu Gruppen mit ähnlich strukturierten Einzeldatensätzen zusammengefaßt. Im zweiten Schritt wird dann für die jeweilige Gruppe ein - eventuell künstlicher - Datensatz ausgewählt, der die gesamte Masse der betreffenden Gruppe enthält. Die Zuordnung zu einer bestimmten Gruppe, also die Entscheidung über die Frage nach der Ähnlichkeit der Einzeldatensätze erfolgt über eine Sortierung der Datensätze nach bestimm-

ten, für die Auswertungszwecke wesentlichen Merkmale. Oberster Sortierbegriff ist demzufolge das Merkmal, dessen Verteilung durch die Sortierung am wenigsten beeinträchtigt werden darf. Diese Konstruktion von Gruppen "verwandter" Einzeldatensätze ist sehr aufwendig und führt, wie Tests mit dem Verfahren gezeigt haben, doch vielfach zu erheblichen Verzerrungen bei den gängigen statistischen Maßen. Das Anonymisierungsverfahren dürfte daher für eine Datenübermittlung für Zwecke der Wissenschaft kaum in Betracht kommen.

ad (c) Bei Langdatensätzen kommt zur Anonymisierung nach §16 Abs. 1 Nr. 4 BStatG ein Katalog von Maßnahmen zur Anwendung, die in einem "Leitfaden zur Anonymisierung statistischer Einzelangaben" aufgelistet sind. Im Prinzip kommen alle genannten Maßnahmen zur Anwendung. Da es sich aber um *Empfehlungen* zur Anonymisierung handelt, besteht ein gewisser Spielraum, der sich darin äußert, daß je nach Materialanforderung die eine oder andere Maßnahme variiert werden kann (Südfeld 1987). Allerdings besteht auch die Verpflichtung, bei der jeweiligen Prüfung des konkreten Einzelfalls zusätzliche, d.h. über den Katalog hinausgehende Anonymisierungsmaßnahmen vorzusehen, die am spezifischen Informationsgehalt der jeweiligen Statistik anknüpfen. Die Kernpunkte des Katalogs sind:

- keine direkten Identifikatoren,
- systemfreie Anordnung der Datensätze,
- Substichprobenziehung,
- Mindestalter der übermittelten Daten,
- nur typisierte Regionalangaben,
- fünffache Besetzung jeder Merkmalsausprägung,
- Klassifizierung der Merkmale in sensiblen Bereichen,
- Klassifizierung von Merkmalen, über die sehr einfach Zusatzinformationen zu erhalten sind, und
- mindestens dreifache Besetzung von Kombinationen sensibler Merkmale mit Merkmalen, über die sehr einfach Zusatzwissen zu erhalten ist.

Auf der Basis dieses Maßnahmenkatalogs wurden im Statistischen Bundesamt bestimmte Techniken entwickelt und implementiert. Die Verfahren führen zweifelsohne zu einer Einschränkung des Analysepotentials, wobei schwach besetzte Merkmale bzw. Merkmalsausprägungen besonders betroffen sind.

Durch die Bestimmung des §16 Abs.6 BStatG und damit die Einführung der *faktischen Anonymität* ist nun für die statistischen Ämter der Weg offen, nach Verfahren zu suchen und diese zu implementieren, die zu keinen merklichen Restriktionen in der Nutzungsmöglichkeit für wissenschaftliche Auswertungen führen. Es ist Ziel dieser Untersuchung, hierfür Empfehlungen zu entwickeln.

2.2.2 Ausland

Die Ausführungen zur Weitergabe von Einzeldatensätzen durch nationale statistische Ämter müssen sich auf ausgewählte Staaten beschränken. Es wird dabei auch nur die Übermittlungspraxis von Daten erörtert, die für die sozialwissenschaftliche Forschung besondere Relevanz haben, also personen- bzw. haushaltsspezifisch sind.

Grundlage sind Informationen, die dem Statistischen Bundesamt nach entsprechender Kontaktaufnahme mit den statistischen Ämtern bestimmter Länder zugegangen sind; u.a. wurden dabei auch Österreich und die Schweiz miteinbezogen, da diese Länder über eine - der deutschen vergleichbaren - Mikrozensuserhebung verfügen. Eine ausgezeichnete Grundlage - auch für weitergehende Informationen zum Problem Datenschutz und -zugang in ausgewählten Ländern, speziell aus Sicht der Wissenschaft - ist eine kürzlich erschienene Publikation, herausgegeben von Guchteneire und Mochmann (1990). Die Frage nach dem Zugang zu statistischen Mikrodaten ist dabei allerdings nur eine von sieben Schwerpunkten, so daß detaillierte Angaben, z.B. zu verwendeten Anonymisierungsmaßnahmen, nicht für alle Ämter zu finden sind. Die genannte Publikation verfügt darüber hinaus über eine umfangreiche internationale Bibliographie zu Datenschutz- und Datenzugangsfragen. Eine weitere interessante aktuelle Veröffentlichung wurde vom Niederländischen Zentralbüro für Statistik (Citteur/Willenborg 1991) vorgelegt. Sie zielt auf die gegenwärtige Praxis nationaler Ämter bei der Herausgabe von "public use files" ab, enthält aber gleichzeitig wichtige Informationen über Alternativen bzw. andere Formen des Datenzuganges in den betreffenden Staaten. Dabei sind die ausgewählten Länder in beiden genannten Publikationen überwiegend deckungsgleich. Neben der Bundesrepublik Deutschland sind dies die USA und Kanada, die skandinavischen Länder Dänemark, Schweden und Norwegen, ferner Großbritannien, die Niederlande, Frankreich und Italien. In der Publikation des Niederländischen Amtes sind außerdem Informationen zu Australien und Neuseeland enthalten; aus beiden Ländern sind auch dem Statistischen Bundesamt Informationen zur Weitergabepraxis (und zu Forschungsaktivitäten) zugegangen. Auf Ungarn, das in der Publikation von Guchteneire und Mochmann außerdem noch vertreten ist, soll hier nicht eingegangen werden.

Die Auswertung der Literatur zeigt, daß in den verschiedenen Staaten die einzelnen Reidentifikationsrisiken (z.B. in bezug auf Regionalmerkmale) unterschiedlich eingeschätzt werden, bei der Anwendung von Anonymisierungsmaßnahmen aber durchaus Parallelen und Kongruenzen erkennbar werden. Außerdem wird deutlich, daß die Praxis der Ämter teilweise nicht nur durch die

Datenschutzgesetzgebung bestimmt wird, sondern daß zumindest in einigen Staaten - insbesondere gilt dies für Schweden und die Niederlande, aber auch für andere - "Eigeninteressen" der statistischen Ämter der Grund für eine eher zurückhaltende Weitergabepraxis von Einzeldatensätzen sind. Unter "Eigeninteressen" ist vor allem die Sorge der Ämter zu verstehen, daß eine Weitergabe die Akzeptanz ihrer Erhebungen bei den Befragten negativ beeinflussen könnte. Interessant ist, daß möglicherweise dieses Akzeptanzproblem in mehreren Ländern - ähnlich wie in der Bundesrepublik Deutschland - dazu geführt hat, daß die Freigabe von Volkszählungsdaten im Vergleich zu sonstigen demographischen Erhebungen restriktiver gehandhabt wird. Ersichtlich wird auch, daß die meisten Staaten keine Unterschiede zwischen einzelnen Empfängergruppen bei der Weitergabe bzw. dem Grad der Anonymisierung kennen, einige aber Wissenschaft und Forschung privilegieren.

In mehreren nationalen Beiträgen in der Publikation von Guchteneire und Mochmann wird deutlich, daß nicht nur restriktive gesetzliche Regelungen eine Weitergabe an die Wissenschaft erschweren, sondern es sich vielfach um eine Kostenfrage handelt, die den Bezug verhindert. Die meisten statistischen Ämter sind verpflichtet, die Kosten der Bereitstellung derartiger Datensätze den Nutzern in Rechnung zu stellen. Dabei werden teilweise Kosten in einer Höhe in Rechnung gestellt, die es Forschern oder sogar Forschungseinrichtungen unmöglich machen, solche Mikrodatenfiles zu beziehen. Selbst bei der Bereitstellung von "public use files", die im Prinzip billiger verkauft werden können als "Einzelanfertigungen" für jede Datenanforderung, können die Kosten von einem Teil der Wissenschaft nicht aufgebracht werden. Das Problem des Preises bzw. der Kosten für faktisch anonymisierte Mikrodatenfiles aus den statistischen Ämtern der Bundesrepublik Deutschland ist ebenfalls ein noch nicht gelöstes Problem und könnte durchaus den eben angesprochenen Effekt bewirken.

Wird die Weitergabe von Einzelangaben in Form von "public use files" angesprochen, werden an erster Stelle meistens *die USA und Kanada* genannt. In beiden Staaten hat die Weitergabe derartiger Files schon eine lange Tradition; das US Bureau of the Census hat mit seinem File aus der Volkszählung 1960 Pionierarbeit geleistet.

Zur Sicherstellung der Geheimhaltung und zur Entwicklung standardisierter Anonymisierungsverfahren wurde 1981 ein Microdata Review Panel eingerichtet. Diese Institution hat auch einen siebenseitigen Leitfaden bzw. Fragebogen entwickelt, anhand dessen das Einzelmaterial vor einer Freigabe anonymisiert werden muß. Die vier großen Bausteine dieser Leitlinien, auf die sich die Anonymisierungsmaßnahmen ausrichten, sind: geographische Informationen,

außergewöhnliche Risiken hinsichtlich einer Deanonymisierung, die Möglichkeiten, das File mit anderen Datenbasen abzugleichen und sonstige für die Anonymisierung wichtige Tatbestände. Als Anonymisierungsmaßnahmen sind insbesondere vorgesehen:

- *data grouping* (mit den verschiedenen Unterformen aggregation, top-coding, recoding or range coding, reducing geographic detail);
- *suppressing* (Unterdrückung bestimmter Ausprägungen oder ganzer Merkmale);
- *data perturbation* (Einführung von Zufallsfehlern in das Datenmaterial);
- *data rounding* (zufälliges Runden der Merkmalsausprägungen) (Cox 1987:32).

"Public use files" werden u.a. wegen der Dezentralisierung der amtlichen Statistik auf verschiedene Ministerien bzw. Behörden auch von anderen Stellen als dem US Bureau of the Census herausgegeben. Ähnliches trifft auch auf *Kanada* zu. Auch dort gibt es eine Einrichtung (Microdata Release Committee), die für die Freigabe verantwortlich ist. Die Anonymisierungsverfahren ähneln denen der USA wobei hier ebenfalls zwischen den einzelnen public use files (Citteur/Willenborg 1991:9) differenziert wird. Neben den Verfahren der Unterdrückung, Vergröberung, Bildung von Kategorien und Einschleusung von Fehlern findet auch die (Sub-)Stichprobenziehung Anwendung. Historische Files werden weniger streng anonymisiert als solche mit aktuellen Daten.

In *Australien* gibt es "public use files" für die Volkszählung und eine Reihe anderer Erhebungen, außerdem einen Microdata Review Panel, der das dortige Statistische Amt bei der Sicherstellung der Geheimhaltung berät. Die Anonymisierungsverfahren entsprechen im Prinzip denen in den USA und Kanada und schließen auch die (Sub-)Stichprobenziehung mit ein. Besonderes Augenmerk wird auch hier - wie in allen Staaten - regionalen Merkmalen gewidmet.

Das Department of Statistics von *Neuseeland* gibt dagegen keine Mikrodaten an die Wissenschaft und andere externe Nutzer weiter (nur eingeschränkt für wissenschaftliche Zwecke an behördliche Einrichtungen).

In *Großbritannien* werden Mikrodaten in Form von "public use files" für verschiedene Erhebungen erstellt, aber zumindest bisher nicht für Volkszählungsdaten. Die Gründe dafür werden ausführlich von Denham (1987) dargelegt. Es ist jedoch beabsichtigt, Einzeldaten aus der jüngsten Volkszählung freizugeben (Marsh et al. 1991). Ein entsprechendes Forschungsprojekt ist bereits abgeschlossen (vgl. Abschnitt 2.2.2). Die Grundlagen für die Freigabe von Einzel-

daten durch das Central Statistical Office von Großbritannien sind in einem "Code of Practice on the Handling of Data obtained from Statistical Inquiries" aus dem Jahr 1984 festgelegt. Dort wird auch entsprechend der Regelung des deutschen Statistikgesetzes definiert, daß die Anonymität gewährleistet ist, wenn zur Reidentifizierung unverhältnißmäßig hoher Aufwand an Zeit, Kosten und Arbeit notwendig wären.

Die Techniken zur Sicherstellung der Geheimhaltung bei der Freigabe von Mikrodaten konzentrieren sich auch in Großbritannien auf die regionalen Merkmale und solche Variablen, die sehr differenziert ausgewiesen werden, wie z.B. die Berufsangabe. Streichungen und Vergröberungen sind in der Regel die in diesen Fällen praktizierten Verfahren.

Keine "public use files" sind in den skandinavischen Ländern *Dänemark, Schweden* und *Norwegen* von den dortigen statistischen Ämtern erhältlich. In allen drei Staaten ist die Weitergabe von Mikrodaten nur auf der Basis eines Vertrages möglich. In Norwegen ist eine vertragliche Weitergabe auf die Wissenschaft und Forschung beschränkt. Norwegen ist damit ein Land, in dem der Forschungsbereich privilegiert wird. Dabei wird bei den Anonymisierungsmaßnahmen darauf geachtet, daß der Forschungszweck erreicht werden kann. Auch in Dänemark scheint eine gewisse Bevorzugung der Wissenschaft zu bestehen; jedoch geben die zur Verfügung stehenden Quellen darüber keinen eindeutigen Aufschluß. Anwendung finden in allen drei Ländern die bekannten Verfahren, wie Streichen und Vergröbern von Merkmalen, Runden, (Sub-)Stichprobenziehung usw. In Norwegen wird z.B. der fünfstellige Regionalcode auf einen einstelligen vergröbert. Im Schwedischen Statistischen Amt existiert ein internes Gremium, das letztlich über die Freigabe entscheidet.

In den *Niederlanden* hat das dortige Zentralbüro für Statistik im Lauf der Jahre seine Freigabemodalitäten geändert. Während früher die Freigabe von Daten über "public use files" relativ großzügig gehandhabt wurde, wurden dann verstärkt Anonymisierungsmaßnahmen ergriffen, die das Reidentifizierungsrisiko de facto auf Null reduzieren sollen. So werden z.B. Beruf, Wirtschaftszweig und Ausbildung nur auf dem Niveau eines einstelligen Schlüssels freigegeben. Der Haushaltszusammenhang darf nicht erkennbar sein, in der univariaten Verteilung muß jede Zahl mindestens 5 Prozent der Gesamterhebung repräsentieren, mit einem absoluten Minimum von 20 Einheiten im File und im bivariaten Fall sind keine Einzelfälle erlaubt. Diese und andere im internationalen Vergleich relativ restriktive Anonymisierungstechniken - verbunden mit hohen Abgabepreisen - haben dazu geführt, daß die Wissenschaft kaum noch Gebrauch von den angebotenen "public use files" machte. Inzwischen ist auf

einzelvertraglicher Grundlage für Forschungsinstitutionen der Zugang zu Einzelangaben erleichtert worden. So werden z.B. regionale Merkmale detaillierter freigegeben als in den gewöhnlichen Files.

Das Institut National de la Statistique et des Etudes Economiques (INSEE) *Frankreichs* gibt für die laufenden personen- und haushaltsspezifischen Erhebungen ebenso "public use files" heraus wie für Volkszählungen. Wissenschaft und Forschung werden nicht bevorzugt. Über Anonymisierungsmaßnahmen liegen keine Informationen vor; es ist bekannt, daß z.B. regionale Merkmale, Haushaltszusammensetzung, Alter, Beruf und Ausbildung weitergegeben werden.

In *Italien* dagegen werden bisher von der amtlichen Statistik im Prinzip keine Einzeldaten weitergegeben, zumindest nicht für Wissenschaft und Forschung. Volkszählungsdaten werden in ihrer ursprünglichen Form an die regionalen Behörden übermittelt.

Das Bundesamt für Statistik der *Schweiz* kann - wie aus dem uns vom Schweizerischen Amt zugeleiteten Material hervorgeht - Einzelangaben entsprechend den Bestimmungen der für die jeweilige Erhebung geltenden Verordnung zur Verfügung stellen. Für den Mikrozensus ist in § 10 der Verordnung die Weitergabe folgendermaßen geregelt:

"Das Bundesamt kann die Daten des Mikrozensus auf Verlangen amtlichen Stellen des Bundes, der Kantone oder der Gemeinden, oder Privatpersonen im Dienste der Forschung, für bestimmte statistische Arbeiten zur Verfügung stellen, wenn:

a) die übermittelten Daten nicht mehr direkt auf die befragten Personen Bezug nehmen;
b) der Empfänger die erhaltenen Daten nicht an Dritte weiterleitet und sie nach Beendigung der Arbeit dem Bundesamt zurückgibt oder vernichtet;
c) die nötigen Sicherheitsmaßnahmen ergriffen und der Datenschutz gewährleistet ist."

Für die Verbrauchserhebung gilt eine ähnliche Regelung. Der Forschungsbereich wird - zumindest verglichen mit anderen privaten Nutzern - bevorzugt. Das Bundesamt hat "Weisungen für die Diffusion von Einzel- und statistischen Daten" entwickelt, die Grundlage für Einzelvereinbarungen sind, die bei Weitergabe mit den Nutzern getroffen werden. Die räumliche Komponente spielt auch

in der Schweiz bei den Anonymisierungsmaßnahmen eine wichtige Rolle; ferner werden für Forschungszwecke nicht benötigte Merkmale gestrichen.

In *Österreich* wird vom dortigen Zentralamt ebenfalls jede Datenanforderung einer Einzelprüfung unterzogen. Einzeldaten aus der Volkszählung und die meisten Merkmale der Häuser- und Wohnungszählungen werden nicht freigegeben. Mikrozensusangaben, z.B. solche aus den Mikrozensus-Sonderprogrammen, werden an einen beschränkten Adressatenkreis weitergegeben. Dabei handelt es sich um renommierte wissenschaftliche Institutionen bzw. Forscher im Hochschulsektor. Damit ist auch in Österreich eine Bevorzugung des Wissenschaftssektors gegeben. Anonymisierungsmaßnahmen konzentrieren sich primär auf die räumlichen/regionalen Merkmale; dagegen werden die Haushalts- und Familienbezüge nicht zerstört.

Die hier kurz skizzierte Weitergabepraxis ausgewählter Staaten hatte das Ziel, das Anliegen dieser Studie im internationalen Zusammenhang darzustellen, sowie eine grobe Einordnung der Gegebenheiten in der Bundesrepublik Deutschland in den internationalen Kontext zu ermöglichen, mit dem speziellen Bezug auf den Wissenschafts- und Forschungsbereich. Weitergehende Informationen über die Weitergabepraxis bei Einzelangaben können den eingangs erwähnten Publikationen entnommen werden.

3 Methodische Grundlagen von Reidentifikationsversuchen*

Nachdem im vorangegangenen Kapitel verschiedene Forschungsschwerpunkte zu Deanonymisierungsrisiken dargestellt wurden, sollen im folgenden die Grundprinzipien einer Reidentifikation sowie die hierbei entstehenden Probleme präzisiert werden. Ein wesentlicher Teil dieses Kapitels dient der Darstellung unterschiedlicher Reidentifikationstechniken und der Einschätzung ihrer Leistungsfähigkeit. Hierbei wird zum einen die Methode des einfachen Abgleichs (Abschnitt 3.2), zum anderen die bereits mehrfach erwähnte diskriminanzanalytische Reidentifikationstechnik von Paaß/Wauschkuhn (Abschnitt 3.4) detailliert dargestellt, da ihre jeweilige Leistungsfähigkeit im weiteren empirisch überprüft werden soll. Darüber hinaus soll mit dem sogenannten Distanzminimierungsverfahren eine weitere mögliche Reidentifikationstechnik (Abschnitt 3.3) erörtert werden, die jedoch als "Zwischenstück" von einfacher Abgleichtechnik und diskriminanzanalytischer Reidentifikationstechnik keiner gesonderten empirischen Überprüfung unterzogen werden soll.

3.1 Grundprinzipien einer Reidentifikation

Das Grundprinzip einer Reidentifikation beruht darauf, daß Einzeldatensätze einer Datei mit anonymen Mikrodaten (Mikrodatenfile) den Einzeldatensätzen einer anderen personenbezogenen Datei in einer eins-zu-eins Entsprechung zugeordnet werden (vgl. u.a. Schlörer 1980, Dittrich/Schlörer 1987). Die für einen Reidentifikationsversuch unabdingbare personenbezogene Datei wird als Identifikationsfile bezeichnet. Sie entspricht dem sogenannten Zusatzwissen (vgl. Kapitel 6) und muß sich - zumindest in Teilen - auf die gleichen Personen wie das anonymisierte Mikrodatenfile beziehen. Im Extremfall kann dieses Zusatzwissen aus einem einzigen Datensatz bestehen, der nicht maschinenlesbar dargestellt sein muß.

Deanonymisierungsversuche setzen bei den Merkmalen an, die Zusatzwissen und Mikrodatenfile gemeinsam sind (Überschneidungsmerkmale). Durch einen Vergleich der Merkmalsausprägungen der Einzeldatensätze beider Files wird

---

* Autoren: Heike Wirth (3.1, 3.2)
Uwe Blien, Heike Wirth (3.3)
Stefan Bender, Uwe Blien (3.4)

hierbei angestrebt, auf der Basis identischer oder sehr ähnlicher Werte jene Datensätze zuzuordnen, die von ein und derselben Person stammen. Damit können die im Mikrodatenfile enthaltenen Nutzmerkmale (beispielsweise Einkommen, Schulden etc.) identifizierbaren Personen zugewiesen werden. Ein Ziel eines solchen Reidentifikationsversuchs könnte darin bestehen, das daraus resultierende personenbezogene Wissen für wissenschaftliche oder andere Zwecke zu nutzen (vgl. Übersicht 3.1).

Übersicht 3.1: Die einem Reidentifikationsversuch zugrundeliegende Datenbasis

	IDENTIFIKATIONS-MERKMALE (z.B. Name, Anschrift)	ÜBERSCHNEIDUNGS-MERKMALE (z.B. Beruf, Alter)	NUTZMERKMALE (z.B. Einkommen)
MIKRODATENFILE	fehlen	vorhanden	vorhanden
IDENTIFIKATIONSFILE (ZUSATZWISSEN)	vorhanden	vorhanden	fehlen

Bei Deanonymisierungsversuchen können im wesentlichen zwei Strategien[2] angewandt werden: Im Falle einer gezielten Suche (*Einzelsuche*) sollen bestimmte, dem Angreifer bekannte Personen aus einem Mikrodatenfile deanonymisiert werden. Bei einem *Fischzug* dagegen werden nicht vorgegebene, sondern beliebige Personen in einem Mikrodatenfile gesucht (Paaß/Wauschkuhn 1985).

Ein Reidentifikationsversuch ist dann relativ problemlos, wenn beide benötigten Datenfiles, also Zusatzwissen und Mikrodatenfile, die in ihnen enthaltenen Einzelfälle in den Datensätzen völlig kompatibel abbilden und mindestens einer der beiden Datenfiles die gesamte Population enthält. In diesem Fall muß lediglich geprüft werden, ob sich für einen Einzelfall des Identifikationsfile ein in den Merkmalsausprägungen identischer Fall im Mikrodatenfile findet und die Beziehung eindeutig ist. Ist eine solche eins-zu-eins Zuordnung (im folgenden als *eindeutige Zuordnung* bezeichnet) möglich, kann der Angreifer zweifelsfrei sicher sein, daß die zugeordneten Datensätze von ein und derselben Person stammen. Da für den Datensatz des Mikrodatenfile damit ein eindeutiger Per-

[2] Die sich aus den unterschiedlichen Strategien für die Reidentifikationswahrscheinlichkeit ergebenden Konsequenzen werden in Kapitel 4 ausführlich erläutert.

sonenbezug hergestellt wäre, würde eine Reidentifikation im Sinne des BStatG vorliegen.

Unter den gleichen Annahmen (keine Inkompatibilitäten, Mikrodatenfile und/oder Identifikationsfile ist eine Vollerhebung) kann ein eindeutiger Personenbezug allerdings schon dann nicht mehr hergestellt werden, wenn eine spezifische Ausprägungskombination im Mikrodatenfile und/oder Identifikationsfile mehrfach besetzt ist, d.h. *statistische Doppelgänger* auftreten. Liegt aufgrund von statistischen Doppelgängern eine Zuordnungsrelation von 1:n, n:1 oder n:n, d.h. eine *mehrdeutige Zuordnung* vor, ist ein Angreifer ohne zusätzliche Informationen nicht in der Lage, für einen Einzeldatensatz einen spezifischen Personenbezug herzustellen. Zusatzwissen steht in aller Regel jedoch nicht unbegrenzt zur Verfügung. Wäre dies der Fall würde sich ein Reidentifikationsversuch erübrigen. Aus diesem Grund werden mehrdeutige Zuordnungen im allgemeinen als weniger riskant und daher als untergeordnetes Problem aus der Diskussion ausgeklammert (Paaß/Wauschkuhn 1985).

Vorwiegend in der englischsprachigen und vereinzelt auch in der deutschsprachigen Literatur wird der Deanonymisierungsbegriff allerdings zum Teil in einem umfassenderen Sinn einer Aufdeckung bzw. Enthüllung (*Disclosure*) von Informationen verwandt, bei welcher nicht notwendigerweise eine Reidentifikation (d.h. die Herstellung eines Personenbezugs für einen spezifischen Datensatz) vorliegen muß (Dalenius 1977, Schlörer 1980, eine zusammenfassende Darstellung findet sich bei Duncan/Lambert 1987). Die oben gegebene Definition der eindeutigen Zuordnung zweier Datensätze aus unterschiedlichen Datenfiles mit dem Ziel einer Reidentifikation einer oder mehrerer spezifischen/spefizischer Person(en) wird in diesen Konzept[3] als sogenannte 'identity disclosure' (vereinzelt auch als 'personal disclosure') bezeichnet (vgl. unter anderem Bethlehem et al. 1990, Paaß 1988a, Skinner et al. 1990, Marsh et al. 1991). Von 'statistical disclosure' wird dann gesprochen, wenn es gelingt einen Einzeldatensatz korrekt mit anderen zuverlässigen Informationen zu verknüpfen, d.h. also eine Datensatzerweiterung vorzunehmen, auch wenn die hinter diesem Einzeldatensatz stehende Person nicht bekannt ist (Cox/Sande 1979). Dalenius (1977) spricht von 'attribute disclosure', wenn es durch die in

[3] Wenn hier von einem Konzept gesprochen wird, dann in dem Sinne, daß unterschiedlichste Formen von Informationsenthüllung unter dem Oberbegriff Disclosure zusammengefaßt werden. Sowohl Schlörer (1980) wie auch Duncan/Lambert (1987) weisen darauf hin, daß unterschiedliche Autoren ein und denselben Begriff für unterschiedliche Sachverhalte verwenden und daher keinesfalls ein in sich geschlossenes Disclosure Konzept zur Verfügung steht.

einem Mikrodatenfile enthaltenen Informationen möglich ist, Merkmalsausprägungen für einen spezifischen Datensatz einer Person genauer zu bestimmen, als dies ohne die in dem Mikrodatenfile enthaltenen Informationen möglich wäre. Auch hier ist die Identität der Person ohne Belang. Palley und Simonoff (1986) schließlich beziehen 'disclosure' nicht mehr auf einzelne Fälle sondern auf spezifische Subpopulationen ('population disclosure'). 'Population disclosure' würde dann vorliegen, wenn es gelingt eine Verbindung zwischen spezifischen Gruppencharakteristika (als Beispiel wird die Verbindung von Lohn- und Angestelltencharakteristika angeführt) herzustellen, und man auf diese Weise zu statistischen Aussagen über bestimmte Personengruppen gelangt.

Wie aus dieser äußerst knappen Darstellung unterschiedlicher disclosure-Typen hervorgeht, muß sich disclosure nicht auf erkennbare Einzelpersonen beziehen. Prinzipiell kann jede Form einer Enthüllung von statistischen Daten sowohl bezogen auf Gruppen mit mehreren tausenden Mitgliedern aber auch auf Einzelpersonen als 'disclosure' bezeichnet werden. Da der Zuordnungsbegriff im Sinne des BStatG die Herstellung eines eindeutigen Personenbezugs voraussetzt, ist daher nur die 'identity' bzw. 'personal disclosure', die der oben gegebenen Definition der eindeutigen Zuordnung entspricht, für die Bestimmung der faktischen Anonymität von Relevanz. In bezug auf eine Einordnung der mehrdeutigen Zuordnungen, muß allerdings auch die sogenannte 'attribute disclosure', d.h. die Enthüllung von Ausprägungen in aller Kürze erörtert werden.

Wie oben ausgeführt, bedeutet eine mehrdeutige Zuordnung, daß eine spezifische Ausprägungskombination in den Überschneidungsmerkmalen $\ddot{U}_1$ bis $\ddot{U}_n$ nicht nur auf eine Person, sondern auf n Personen zutrifft. Um einen spezifischen, eindeutigen Personenbezug im Sinne einer Reidentifikation herzustellen, müßte ein Angreifer über weitergehende Information verfügen. Dies trifft insbesondere dann zu, wenn die mehrdeutig zugeordneten Datensätze zwar für $\ddot{U}_1$ bis $\ddot{U}_n$ identische Ausprägungskombinationen aufweisen, aber sich in den Nutzmerkmalen N_1 bis N_n unterscheiden. Hiervon zu unterscheiden ist allerdings der Sonderfall, bei welchem ein Angreifer aus mehrdeutigen Zuordnungen auch ohne Herstellung eines eindeutigen Personenbezugs zumindest einen Informationsgewinn erhalten könnte.

Bei dieser spezifischen Bedingungskonstellation, die unter der erwähnten 'attribute disclosure' subsumiert werden könnte, ist vorausgesetzt, daß alle in einer mehrdeutigen Zuordnung enthaltenen, statistischen Doppelgänger, nicht

nur in den Überschneidungsmerkmalen $Ü_1$ bis $Ü_n$ sondern mindestens auch in einem der Nutzmerkmale N_1 bis N_n identisch sind (vgl. Übersicht 3.2).

Übersicht 3.2: Enthüllung von Ausprägungen bei mehrdeutigen Zuordnungen

IFz	$Ü_1$	$Ü_2$	...	$Ü_n$	MFy	N_1	N_2	...N_n
z_1	0	1	...	0	y_1	2	.	....
	0	1	...	0	y_2	2	.	....
	0	1	...	0	y_3	2	.	....
.					.			
.					.			
z_6	1	0	...	1	y_6	1	.	....
z_7	1	0	...	1	y_7	1	.	....
z_8	1	0	...	1	y_8	1	.	....

Unter dieser Bedingung könnte ein Angreifer für eine bestimmte Person die betreffende Ausprägung des Nutzmerkmals erschließen, auch wenn es ihm nicht möglich ist, für den spezifischen Datensatz einen Personenbezug herzustellen. Unter der Annahme, daß nicht nur *ein* Nutzmerkmal, sondern *alle* Nutzmerkmale N_1 bis N_n identisch wären, könnte ein Angreifer allerdings auch ohne weiteren Aufwand auf die Nutzmerkmale der ihn interessierenden Person schließen. Allerdings handelt es sich hierbei eher um eine Modellannahme, die für analytische Zwecke konstruiert ist, als um ein in der Realität häufig anzutreffendes Ereignis. Denn mit jedem zusätzlich einbezogenen Merkmal steigt die Wahrscheinlichkeit, daß sich zwei Datensätze bezüglich dieser Information unterscheiden. Daher ist es äußerst unwahrscheinlich, daß mehrdeutige zugeordnete Datensätze auch über alle Nutzmerkmale N_1 bis N_n identische Ausprägungen aufweisen. Eine bloße Enthüllung von Merkmalsausprägungen, die sich nicht - wie bei einer eindeutigen Zuordnung - erkennbar auf eine spezifische Person, sondern auf mehrere Personen beziehen, kann noch nicht als eine Reidentifikation im Sinne des BStatG bezeichnet werden, da diese immer die Herstellung eines eindeutigen Personenbezugs erfordert.

Vor diesem Hintergrund können sich die weitergehenden Erörterungen auf eindeutige Zuordnungen beschränken, wobei die in diesem Fall auftretenden Probleme analog auch bei mehrdeutigen Zuordnungen gegeben sind.

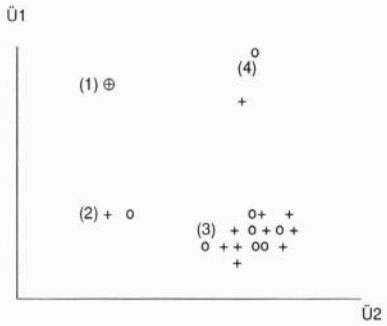
Wie aus den späteren, ausführlicheren Erörterungen zum Zusatzwissen und zu Dateninkompatibilitäten deutlich werden wird, spiegelt der oben dargestellte Reidentifikationsvorgang, bei welchem die eindeutige Zuordnung gleichbedeutend mit einer Reidentifikation war, eine für den Angreifer idealtypische Situation wider. Von einer solchen kann unter realistischen Bedingungen allerdings nicht ausgegangen werden. Sehr viel wahrscheinlicher ist, daß sowohl das Mikrodatenfile als auch das Zusatzwissen lediglich als Stichproben zur Verfügung stehen und die Daten nicht vollständig kompatibel sind. Bereits beim Vorliegen einer dieser Bedingungen kann auch bei einer eindeutigen Zuordnung von Fällen des Identifikationsfile und des Mikrodatenfile nicht mehr mit letzter Sicherheit geschlossen werden, daß diese Zuordnung korrekt ist, d.h. die zugeordneten Datensätze auch in der Realität von ein und derselben Person stammen. Bei Dateninkompatibilitäten kann es vorkommen, daß ein Datensatz des Identifikationsfile einem Datensatz im Mikrodatenfile mit identischer Merkmalskombination zugeordnet wird. Die Datensätze im Identifikationsfile und Mikrodatenfile beziehen sich aber auf unterschiedliche Personen, deren Ausprägungskombinationen aufgrund inkompatibler Abbildungen identisch geworden sind. Sind sowohl Identifikationsfile wie Mikrodatenfile nur Stichproben, ist nicht mehr mit Sicherheit entscheidbar, ob identische Datensätze des Identifikationsfile und Mikrodatenfile sich auch auf eine identische Person beziehen, weil die Stichproben von Identifikationsfile und Mikrodatenfile unterschiedlich zusammengesetzt sein können. Die Identität von Datensätzen könnte daher darauf beruhen, daß statistische Doppelgänger in den jeweiligen Stichproben enthalten sind. Aus diesen Gründen sind also auch eindeutige Zuordnungen nur noch mit einer gewissen Wahrscheinlichkeit korrekt. Der Erfolg eines Datenangreifers ist daher im wesentlichen dadurch bestimmt, inwieweit eine ihm zur Verfügung stehende Reidentifikationstechnik mit den aufgezeigten Unsicherheitsfaktoren umgehen kann. Im folgenden wird die hier nur kurz skizzierte Zuordnungsproblematik präzisiert. Hieran schließt sich eine ausführliche Erörterung ausgewählter Reidentifikationstechniken an.

Die sich bei einem Reidentifikationsversuch ergebenden Unsicherheitsfaktoren können anhand von Abbildung 3.1 verdeutlicht werden. Dargestellt werden vier denkbare Zuordnungssituationen.[4] Die beiden Achsen stellen die Ausprägungen der beiden Überschneidungsmerkmale Ü1 und Ü2 dar. Die Datensätze des Identifikationsfile sind durch Kreise, die Datensätze des Mikrodatenfile durch

[4] Natürlich sind noch eine Vielzahl anderer Zuordnungssituationen denkbar. Die mit einer Zuordnung verbundenen Probleme können jedoch anhand dieser vier grundlegenden Zuordnungskonstellationen aufgezeigt werden.

Kreuze gekennzeichnet. Vereinfachend wird angenommen, daß die Zuordnung nur auf der Basis von Ü1 und Ü2 erfolgen soll.

Abbildung 3.1: Verschiedene Zuordnungssituationen



+: Ein Fall im Mikrodatenfile
o: Ein Fall im Identifikationsfile

Wie aus den obigen Ausführungen deutlich wurde, ist die Zuordnungssituation vergleichsweise einfach, wenn die Datenbasis kompatibel und das Mikrodatenfile oder Identifikationsfile die gesamte Population umfaßt. Ist eine eindeutige Zuordnung möglich, d.h. findet sich im Zusatzwissen ein Fall, der in seinen Ausprägungen identisch mit einem Fall des Mikrodatenfile ist, hat eine Reidentifikation stattgefunden. In Abbildung 3.1 würde dies Situation (1) entsprechen.

Sind die Daten hingegen nicht vollständig kompatibel, können Datensätze, obwohl sie von ein und derselben Person stammen, in verschiedenen Datenbeständen unterschiedlich abgebildet sein. Für Zuordnungen kommen daher nicht nur Datensätze in Betracht, die identische, sondern auch solche, die ähnliche Ausprägungskombinationen aufweisen. In Situation (1) fällt die Entscheidung noch vergleichsweise leicht. Hier unterscheiden sich zwei Fälle aus dem Identifikationsfile und dem Mikrodatenfile nicht. Im Merkmalsraum liegen keine anderen Fälle in der Nähe. Ein Angreifer kann daher annehmen, daß die beiden Datensätze mit hoher Wahrscheinlichkeit von einer Person stammen. In Situa-

tion (2) liegt zwar eine geringfügige Abweichung auf der Dimension Ü2 vor, dennoch sind die betreffenden Datensätze im "Merkmalsraum" einerseits stark von anderen Datensätzen isoliert und liegen andererseits sehr nah zusammen. So könnte man annehmen, daß eine Dateninkompatibilität vorliegt und die Angaben trotz der Abweichung von der gleichen Person stammen.

Anders in Situation (3). Hier liegen zwar auch Fälle aus dem Identifikationsfile in der Nähe von Fällen des Mikrodatenfile. Es tritt jedoch eine solche lokale Häufung von ähnlichen Ausprägungskombinationen auf, daß nicht zu unterscheiden ist, welcher Datensatz des Mikrodatenfile einem Fall des Identifikationsfile zugeordnet werden kann. Der Vergleich der Situationen (2) und (3) zeigt, daß Fälle mit vergleichsweise 'exotischen' Ausprägungskombinationen potentiell leichter identifizierbar sind als der durchschnittliche 'Merkmalsträger', dessen Ausprägungen nur wenig von denen der anderen Befragten variieren.

In Situation (4) schließlich sind zwei Fälle derart unähnlich, daß nicht zu erwarten ist, daß sie von der gleichen Person stammen.

Das durch die Abbildung verdeutlichte Entscheidungsproblem verschärft sich, wenn beide Datenfiles nur Stichproben widerspiegeln, da dann viele statistische Doppelgänger in der dem Angreifer unbekannten Population auftreten können. Jede vorgenommene Zuordnung könnte daher auch auf einer Verwechslung durch statistische Doppelgänger beruhen. So wüßte ein Angreifer in Situation (2) beispielsweise nicht, ob es sich hier um die Datensätze einer Person handelt, die aufgrund von Dateninkompatibilitäten unterschiedlich abgebildet sind oder ob die Datensätze von unterschiedlichen Personen stammen und die jeweiligen Partnerdatensätze nicht in den Stichproben enthalten sind.

Anhand dieser Überlegungen wird deutlich, daß eine Zuordnung nicht mehr mit Sicherheit, sondern nur noch mit einer gewissen Wahrscheinlichkeit erfolgen kann, wenn Identifikationsfile und Mikrodatenfile Stichproben sind und/oder wenn Datenfehler auftreten. Für einen Angreifer stellt sich daher das Problem, wie einerseits das Reidentifikationspotential der Daten auch bei Inkompatibilitäten weitestgehend ausgeschöpft ist und wie andererseits das Risiko einer Falschzuordnung (d.h. die zugeordneten Datensätze stammen von unterschiedlichen Personen) möglichst gering gehalten werden kann.

In der Literatur wird hauptsächlich auf zwei Reidentifikationstechniken hingewiesen, die in bezug auf Leistungsfähigkeit und Aufwand jeweils Extrempunkte einer Skala möglicher Angriffstechniken markieren. Am unteren Ende der Skala steht ein einfaches Abgleichverfahren (Block/Olsson 1976, Dittrich/Schlörer

1987). Bei dieser Technik wird als einziges Zuordnungskriterium die Identität zweier Datensätze in ihren Ausprägungskombinationen berücksichtigt. Diese Methode enthält keine statistischen Kriterien, um die oben aufgezeigten Unsicherheitsfaktoren durch mögliche Dateninkompatibilitäten oder Stichprobeneigenschaften der Daten abzuschätzen. Bei dieser Vorgehensweise können Datensätze, die von einer Person stammen aber unterschiedlich abgebildet sind, nicht zugeordnet werden. Zugleich stehen auch keine statistischen Hilfsmittel zur Verfügung, um die Wahrscheinlichkeit einer korrekten Zuordnung zu ermitteln. Die einfache Abgleichtechnik muß somit nach den bisherigen Ausführungen als eine nur bedingt leistungsfähige und sehr unsichere Reidentifikationsmethode angesehen werden.

Als das - auch international - leistungsfähigste Verfahren gilt eine von der Gesellschaft für Mathematik und Datenverarbeitung (GMD) im Rahmen des AIMIPH-Projektes entwickelte Reidentifikationstechnik (Paaß/Wauschkuhn 1985). Diese Technik versucht, die oben aufgezeigten Probleme durch Abschätzungen von Inkompatibilitäten und des Auftretens von statistischen Doppelgängern in der Grundgesamtheit zu lösen. Der auf der Diskriminanzanalyse und Dichteschätzung beruhende Algorithmus ermittelt die Wahrscheinlichkeiten von bestimmten Ausprägungskombinationen und setzt sie zu der Wahrscheinlichkeit von Datenfehlern - über die bestimmte Annahmen getroffen werden - in Beziehung. Auf diese Weise kann für jeden Zieldatensatz des Identifikationsfile derjenige Datensatz des Mikrodatenfile ermittelt werden, der in seiner Ausprägungskombination die höchste Übereinstimmung aufweist.

Während die einfache Abgleichtechnik in der Literatur in der Regel nur als eine mögliche und vergleichsweise einfach durchzuführende Reidentifikationstechnik erwähnt wird, ohne daß hierfür empirische Befunde vorliegen, wurde die diskriminanzanalytische Reidentifikationstechnik der GMD einer empirischen Überprüfung unterzogen. Hierfür stand als Datenmaterial ein empirisches Mikrodatenfile und ein daraus synthetisch erzeugtes - mit Fehlern überlagertes - Identifikationsfile zur Verfügung. Das Verfahren hat bei diesem Material unter bestimmten Randbedingungen vergleichsweise hohe Reidentifikationsquoten angezeigt und die Einschätzung der Deanonymisierungsrisiken stark beeinflußt. Damit stand erstmals eine Reidentifikationstechnik zur Verfügung, die den Anspruch hatte, eine leistungsfähige Methode für das Problem von Dateninkompatibilitäten und statistischen Doppelgängern darzustellen.

Da sowohl die einfache Abgleichtechnik als auch die diskriminanzanalytische Reidentifikationstechnik für einen Datenangriff herangezogen werden könnten, sollen sie im folgenden im Hinblick auf ihre Funktionsweise näher erörtert und

später ihre konkrete Leistungsfähigkeit unter empirischen Bedingungen untersucht werden.

Ergänzend wird eine weitere Reidentifikationstechnik, das sogenannte Distanzminimierungsverfahren, näher untersucht, da es den Anspruch erhebt, Schwächen der einfachen Abgleichtechniken mit relativ niedrigem Aufwand zu lösen.

3.2 Technik des Abgleichs von Merkmalsausprägungen

Die denkbar einfachste Reidentifikationstechnik, auch unter den Bezeichnungen "Backwards Identification" (Block/Olsson 1976) bzw. "Hintertreppenidentifikation" oder "Sortier- und Selektionsverfahren" (Dittrich/Schlörer 1987, Fischer-Hübner 1986; Brunnstein 1987) bekannt, beruht auf einem Abgleich von Datensätzen aus Mikrodatenfile und Identifikationsfile auf Identität in den Merkmalsausprägungen. Ausgehend von dem Mikrodatenfile wird jeder darin enthaltene Datensatz sukzessive mit jedem in dem Identifikationsfile enthaltenen Datensatz auf Identität der Überschneidungsmerkmale abgeglichen (vgl. Block/Olsson 1976). Findet sich im Mikrodatenfile genau ein Fall, der im Hinblick auf seine Wertekombination exakt einem Fall des Identifikationsfile entspricht, so kann dieser als reidentifiziert betrachtet werden, allerdings nur unter der Annahme, daß die Datenbestände kompatibel und das Mikrodatenfile und/oder das Identifikationsfile eine Vollerhebung darstellt. Statistische Zuordnungskriterien, etwa in Form von Fehlerabschätzungen oder der wahrscheinlichkeitstheoretischen Berechnung des Auftretens von statistischen Doppelgängern in der Grundgesamtheit werden bei dieser Technik nicht berücksichtigt.

3.2.1 Dateninkompatibilitäten und Doppelgängerproblematik

Der Vorteil einer solchen Reidentifikationstechnik liegt insbesondere in der einfachen Realisierbarkeit, die keinen großen technischen Aufwand erfordert und gegebenenfalls auch manuell durchführbar ist. In einer Reihe von Untersuchungen zum Reidentifikationsrisiko amtlicher Datenbestände wird diesen einfachen Abgleichtechniken daher explizit (Fischer-Hübner 1986, Brunnstein 1987) oder implizit (Dalenius 1986, Bethlehem/Keller/Pannekoek 1990) ein sehr hohes Gefährdungspotential zugeschrieben. Diese Argumentation beruht auf

dem sogenannten Uniqueness-Konzept, bei welchem das Reidentifikationsrisiko für anonymisierte Datensätze ausschließlich darüber bestimmt wird, ob der jeweilige Datensatz in seinen Ausprägungskombinationen einzigartig ist oder nicht. In einigen Untersuchungen wird daher - etwas vereinfacht - der Anteil der einzigartigen Datensätze in einem Mikrodatenfile einem potentiellen Reidentifikationsrisiko gleichgesetzt. In der Untersuchung von Fischer-Hübner 1986 und Brunnstein 1987 ist dies auf der einen Seite insofern gerechtfertigt, als die Argumente dieser Autoren sich auf die Volkszählung, also eine Totalerhebung, beziehen. Auf der anderen Seite sind die vorgenommenen Abschätzungen des Reidentifikationsrisikos insofern unrealistisch, als von vollständig kompatiblen Informationen zwischen Mikrodatenfile und Identifikationsfile ausgegangen wird. Komplexere Analysen hingegen versuchen den Anteil der einzigartigen Ausprägungskombinationen in der Grundgesamtheit zu bestimmen und darüber den Anteil der "echten" Einzelfälle in einer Stichprobe zu schätzen, indem sie univariate Randverteilungen sowie die statistische Abhängigkeit der Merkmale berücksichtigen (Bethlehem/Keller/Pannekoek 1990). Unabhängig von diesen konzeptionellen Unterschieden stimmen die Vertreter des Uniqueness-Konzepts darin überein, daß einzigartig vertretene Datensätze - sofern ein Angreifer über das geeignete Zusatzwissen verfügt - durch einen einfachen Abgleich der Merkmalsausprägungen auf Identität, d.h. entsprechend Situation (1) in Abbildung 3.1, vergleichsweise problemlos reidentifizierbar sind (vgl. Fischer-Hübner 1986, Brunnstein 1987, Dalenius 1987, Bethlehem et al. 1990).

Eine Reidentifikationstechnik, die als einziges Zuordnungskriterium identische Ausprägungskombinationen zwischen zwei Datensätzen voraussetzt, hat aber entscheidende Defizite: Einerseits können Datensätze, die von einer Person stammen, aber in Mikrodatenfile und Identifikationsfile unterschiedlich abgebildet sind, nicht zugeordnet werden. Treten Dateninkompatibilitäten auf und ist ein gesuchter Fall hiervon betroffen, wird der Datensatz im Verlauf des Zuordnungsprozesses als nicht passend aussortiert. Andererseits kann es auch dazu kommen, daß Datensätze, die bei korrekter Abbildung nur sehr ähnliche Merkmalskombinationen aufweisen würden, aufgrund von Fehlern zu Datensätzen mit identischer Ausprägungskombination werden. In diesem Fall würden Dateninkompatibilitäten also zu Falschzuordnungen führen. Abweichungen von den "wahren" Werten wären bei einfachem Abgleich nur dann problemlos für Deanonymisierungsversuche, wenn ein betroffener Fall im Identifikationsfile und Mikrodatenfile exakt die gleichen Abweichungen aufweisen würde und damit die Abbildung trotz Fehler kompatibel wäre. Schließlich kann es, wenn sowohl das Mikrodatenfile als auch das Zusatzwissen nur eine Stichprobe darstellen, auf-

grund der Doppelgängerproblematik zu Falschzuordnungen kommen. Das Ausmaß von Nicht- und Falschzuordnungen kann durch die einfache Abgleichtechnik jedoch nicht kontrolliert werden.

Einfache Abgleichtechniken werden nicht nur durch die Stichprobeneigenschaft der Daten und/oder mögliche Dateninkompatibilitäten nachhaltig gestört, sondern auch der Zuordnungsmechanismus erweist sich bei näherer Betrachtung als ein wesentlicher Unsicherheitsfaktor bei dem Versuch, ausschließlich auf der Basis von Ausprägungsidentität einen Personenbezug bei anonymisierten Daten herzustellen. Die Gründe hierfür werden im folgenden aufgezeigt.

3.2.2 Zuordnungsmechanismen bei Techniken des Abgleichs von Merkmalsausprägungen

Der Grad der Gefährdung, d.h. die Quote der potentiell reidentifizierbaren Einzeldatensätze eines Mikrodatenfile hängt unter anderem vom Informationsgehalt eines hypothetisch unterstellten Zusatzwissens ab. Je höher der Informationsgehalt des Zusatzwissens ist, desto schärfer unterscheiden sich einzelne Datensätze in ihren Ausprägungskombinationen und desto wahrscheinlicher ist es, daß einzelne Personen im Merkmalsraum eindeutig voneinander unterscheidbare Positionen einnehmen. Deshalb sinkt bei einem hohen Informationsgehalt der Überschneidungsmerkmale die Wahrscheinlichkeit, daß zwei Datensätze (aus Mikrodatenfile und Identifikationsfile) aufgrund einer zufälligen Übereinstimmung in ihren Ausprägungen zugeordnet werden (Paaß/Wauschkuhn 1985, Bethlehem et al. 1990, siehe im einzelnen dazu Abschnitt 4.2). Bei niedrigem Informationsgehalt der Überschneidungsmerkmale hingegen ist aufgrund der geringeren Diskriminierungsfähigkeit in erhöhtem Maße von zufälligen Übereinstimmungen zwischen Ausprägungskombinationen auszugehen, die zu Falschzuordnungen führen können. In komplexen Reidentifikationstechniken wird dieser Sachverhalt berücksichtigt, indem die Wahrscheinlichkeit einer Zuordnung unter anderem durch die Berücksichtigung des Informationsgehalts der Überschneidungsmerkmale ermittelt wird. Fällt insbesondere die Anzahl der Überschneidungsmerkmale unter ein bestimmtes Minimum, so "verweigern" komplexe Reidentifikationstechniken unter Umständen eine Zuordnung (Paaß/Wauschkuhn 1985:12).

Bei einfachen Reidentifikationstechniken fehlen derartige statistische Kontrollmechanismen. Das Potential der Zuordnungen wird ausschließlich durch die Regeln der Kombinatorik bestimmt. Dabei besteht ein negativer Zusammen-

hang zwischen der Höhe des Informationsgehalts der Überschneidungsmerkmale und der absoluten Anzahl identischer Ausprägungskombinationen in Mikrodatenfile und Identifikationsfile. Je geringer der Informationsgehalt der Überschneidungsmerkmale und damit die Variation in den Ausprägungskombinationen ist, desto mehr ein- und mehrdeutige Zuordnungen sind potentiell möglich. Je höher der Informationsgehalt ist, d.h. je breiter die Ausprägungskombinationen über einen Merkmalsraum streuen, desto geringer wird die Wahrscheinlichkeit, daß sich identische Ausprägungskombinationen zwischen Mikrodatenfile und Zusatzwissen finden. Die Zahl der potentiell möglichen Zuordnungen sinkt daher.

Die Logik einfacher Reidentifikationstechniken und der damit einhergehende Unsicherheitsfaktor kann vereinfacht an einem Beispiel verdeutlicht werden. Übersicht 3.3 spiegelt die Ergebnisse aus unterschiedlichen Zuordnungssituationen (Z1 bis Z4) mit einer einfachen Abgleichtechnik wider. Für alle vier Situationen wurde die in Übersicht 3.4 dargestellte Datenbasis zugrunde gelegt. Die Zuordnungssituationen Z1 bis Z4 unterscheiden sich lediglich bezüglich der jeweils berücksichtigten Überschneidungsmerkmale. Datensätze im Mikrodatenfile und Identifikationsfile, die hypothetisch von einer Person stammen, sind durch die gleichen Buchstaben gekennzeichnet. Im konkreten Beispiel sind dies die Fälle H und I.

Übersicht 3.3: Variabilität in den Zuordnungen bei einer einfachen Abgleichtechnik in Abhängigkeit der zur Verfügung stehenden Überschneidungsmerkmale

Z1	Z2	Z3	Z4
Ü1 — Ü6 MF IF	Ü1 — Ü5 MF IF	Ü2 — Ü6 MF IF	Ü2 — Ü5 MF IF
H ⟶ H I ⟶ I	B ⟶ Q C/H ⟶ H A/I ⟶ I E ⟶ N G ⟶ K	F ⟶ Q H ⟶ H I ⟶ I G ⟶ K/O	B/F ⟶ Q C/H ⟶ H A/I ⟶ I E ⟶ N D ⟶ P

Übersicht 3.4: Hypothetische Datenbasis (bezogen auf Überschneidungsmerkmale) eines Reidentifikationsversuchs

Fall	MIKRODATENFILE					
	V1	V2	V3	V4	V5	V6
A)	5	1	3	0	16	7
B)	4	2	7	2	15	6
C)	1	2	4	0	16	6
D)	4	1	3	3	14	4
E)	3	2	5	4	13	1
F)	5	2	7	2	15	7
G)	2	2	7	2	11	3
H)	1	2	4	0	16	2
I)	5	1	3	0	16	6

Fall	IDENTIFIKATIONSFILE					
	V1	V2	V3	V4	V5	V6
H)	1	2	4	0	16	2
I)	5	1	3	0	16	6
K)1	2	7	2	11	3	
L)	3	1	3	1	16	4
M)	4	1	3	1	15	7
N)	3	2	5	4	13	7
O)	5	2	7	2	11	1
P)	2	1	3	3	14	6
Q)	4	2	7	2	15	7

Wie Übersicht 3.3 entnommen werden kann, stellt die Zuordnungssituation Z1 unter der Bedingung, daß die Daten kompatibel sind und Ü1 bis Ü6 (Übersicht 3.4) das maximal erreichbare Zusatzwissen widerspiegelt, die höchste Erfolgschance eines Angreifers dar. Für die Zuordnungssituationen Z2 bis Z4 wird hingegen unterstellt, daß das Zusatzwissen des Angreifers jeweils die Überschneidungsmerkmale Ü1 bis Ü5, Ü2 bis Ü6 bzw. Ü2 bis Ü5 umfaßt.

Durch den Vergleich der Zuordnungssituationen Z1-Z4 werden verschiedene Besonderheiten dieser einfachen Abgleichtechniken deutlich:

- Die Logik der einfachen Reidentifikationstechnik schließt aus, daß sich die Zahl der - aufgrund von Ausprägungsidentität - möglichen Zuordnungen durch die Einbeziehung zusätzlicher Überschneidungsmerkmale erhöht. Datensätze, die in mindestens einem Merkmal abweichen, können, wie der Vergleich von Z4 und Z1 zeigt, auch dann nicht identisch werden, wenn zusätzliche Überschneidungsmerkmale berücksichtigt werden. Hat ein Angreifer bei der Berücksichtigung von sieben Überschneidungsmerkmalen 20 Zuordnungen vornehmen können, so wird sich diese Zahl auch dann nicht erhöhen, wenn weitere 15 oder 40 Überschneidungsmerkmale berücksichtigt werden.

- Während zusätzliche Informationen die Anzahl der absolut möglichen Zuordnungen nicht zu erhöhen vermögen, kann eine reduzierte Zahl von

berücksichtigten Merkmalen zu einem quantitativen "Mehr" an Zuordnungen führen: Datensätze, die bei der Betrachtung des gesamten Merkmalspektrums unterschiedlich abgebildet sind, können in bezug auf einen spezifischen Spektrumsausschnitt vollkommen identische Ausprägungen aufweisen (vgl. Z1 bis Z4).

- Zuordnungen, die auf der Basis von identischen Ausprägungskombinationen bei einer Vielzahl von Überschneidungsmerkmalen korrekt vorgenommen werden können, sind bei einem reduzierten Zusatzwissen weiterhin möglich. Sie können dann allerdings, wie aus dem Beispiel hervorgeht, von einer Vielzahl weiterer falscher Zuordnungen begleitet sein, die eine unmittelbare Selektion der korrekten Zuordnungen nahezu unmöglich macht.

- Sofern das Zusatzwissen des Angreifers nur einen begrenzten Ausschnitt der Merkmale eines Mikrodatenfile enthält (und hiervon ist auszugehen), kann es aufgrund dieser unvollständigen Informationen zu relativ willkürlichen Zuordnungen kommen (vgl. Z2 bis Z4). Wie aus den obigen Übersichten hervorgeht, können Datensätze, die von unterschiedlichen Personen stammen und bei der Berücksichtigung der gesamten Informationsmenge auch unterschiedlich abgebildet sind, bei der Betrachtung eines Merkmalsausschnitts durchaus identische Wertekombinationen aufweisen. In Abhängigkeit der zur Verfügung stehenden Überschneidungsmerkmale kann deshalb nicht nur die Zahl der Zuordnungen variieren, sondern ein und derselbe Datensatz des Mikrodatenfile kann auch unterschiedlichen Datensätzen des Identifikationsfile vice versa zugeordnet werden. So wird der Fall Q aus dem Identifikationsfile einmal dem Fall B (Z2) und einmal dem Fall F (Z3) des Mikrodatenfile zugeordnet. Betrachtet man sich hierzu die in Übersicht 3.2 dargestellte Datenbasis, zeigt sich, daß beide Zuordnungen falsch wären. Diese Fälle werden nur deshalb zugeordnet, weil das Wissen des Angreifers unvollständig ist. Für einen Angreifer, dessen Ziel jedoch gerade darin besteht, aus einem Mikrodatenfile Informationen zu gewinnen, bleibt dieser Sachverhalt verborgen, da sein Zusatzwissen ja annahmegemäß nicht alle Merkmale des Mikrodatenfile aufweisen wird.

Die hier aufgezeigten Effekte werden mit einer zunehmenden Anzahl von Überschneidungsmerkmalen sicherlich an Bedeutung verlieren. Solange das Zusatzwissen des Angreifers jedoch nicht den vollständigen Merkmalsbereich des Mikrodatenfile umfaßt, geht mit diesen einfachen Zuordnungsmechanismen ein Unsicherheitsfaktor einher, der durch das Auftreten von echten statistischen

Doppelgängern bzw. das Auftreten von Dateninkompatibilitäten zusätzlich verstärkt wird.

3.2.3 Zusammenfassung

Die hier nur skizzierten Defizite einfacher Reidentifikationstechniken machen deutlich, daß mit Hilfe dieser Techniken zwar vergleichsweise einfach Fälle mit identischen Ausprägungskombinationen im Mikrodatenfile und Identifikationsfile gefunden werden können. Aber selbst unter der Bedingung, daß eine spezifische Ausprägungskombination im Mikrodatenfile und Identifikationsfile jeweils nur einfach besetzt ist, kann daraus keinesfalls gefolgert werden, daß bereits eine Reidentifikation vorliegt, d.h. daß die zugeordneten Datensätze in der Realität von ein und derselben Person stammen. Eine solche Zuordnung kann zwar auf einer realen Eindeutigkeit beruhen, sie kann aber ebenso auf Verwechslungen mit statistischen Doppelgängern, auf Datenfehler oder schlicht auf unvollständige Informationen des Angreifers zurückzuführen sein. Eine zweite - neben möglichen Falschzuordnungen - entscheidende Schwäche dieser Technik besteht darin, daß inkompatibel abgebildete Datensätze nicht zugeordnet werden können, ein Reidentifikationsversuch daher unter Umständen völlig ins Leere läuft.

Die hier angestellten Überlegungen sind bislang allerdings nur theoretischer Art und unseres Wissens liegt bislang keine Untersuchung vor, in welcher ein Datenangriff mit solchen Techniken realitätsnah überprüft wurde. Zugleich ist anzunehmen, daß diese einfachen Techniken - da technisch problemlos realisierbar - bei einem Datenangriff sehr wahrscheinlich an erster Stelle stehen. Deshalb wird ihre konkrete Leistungsfähigkeit später empirisch untersucht werden.

3.3 Zuordnung durch Distanzminimierung

Dem Problem von Nichtzuordnungen aufgrund von Dateninkompatibilitäten könnte mit einer anderen Zuordnungstechnik begegnet werden, die mit der Berechnung von Distanzmaßen arbeitet. In diesem Fall würde nicht mehr nur die Ausprägungsidentität, sondern auch die Ausprägungsähnlichkeit über die Zuordnung von Datensätzen entscheiden. Der Datenangreifer könnte annehmen, daß Datensätze zusammengehören, deren Distanz einen bestimmten

Schwellenwert unterschreiten, d.h. von ein und derselben Person stammen. Mit Zuordnungstechniken, die auf Distanzschätzungen zwischen Datensätzen beruhen, könnte also dem Problem begegnet werden, daß aufgrund möglicher inkompatibler Abbildungen der Merkmale einer Person im Mikrodatenfile und Identifikationsfile durch einfache Techniken an sich korrekte Zuordnungen nicht vorgenommen werden können.

Aber auch Verfahren der Zuordnung aufgrund minimaler Distanz sind problembehaftet. Dies wird bei der folgenden Überlegung unmittelbar deutlich. Bei einem Zuordnungskriterium aufgrund minimaler Distanz müßten als erstes Datensätze als von der gleichen Person stammend zugeordnet werden, die identische Ausprägungskombinationen und damit die Distanz Null aufweisen. Identität in den Ausprägungskombinationen heißt, wie in den vorangegangenen Abschnitten ausgeführt, jedoch nicht ohne weiteres, daß die Datensätze von ein und derselben Person stammen. Identität in den Ausprägungskombinationen kann durch Datenfehler zustandekommen oder aber auf statistische Doppelgänger zurückzuführen sein. Ein Distanzminimierungsverfahren würde deshalb in bezug auf Zuordnungen mit der Distanz Null die gleiche Fehlerwahrscheinlichkeit wie einfache Abgleichtechniken aufweisen. Wenn eine korrekte Zuordnung jedoch schon bei der Null-Distanz nicht mit absoluter Sicherheit gegeben ist, stellt sich die Frage, welches Distanzkriterium dann bei der Zuordnung von nur "ähnlichen" Datensätzen für eine Reidentifikation vorausgesetzt werden soll.

Die Festlegung eines Kriteriums für eine Maximaldistanz, bei welcher von einer erfolgreichen Zuordnung ausgegangen werden könnte, wäre auch bei einer technischen Umsetzung dieser Methode mit großen Problemen verbunden. So ist es möglich, daß zunächst der Datensatz A des Mikrodatenfile dem Datensatz B des Identifikationsfile zugeordnet wird, obwohl in einer späteren Phase des Suchprozesses ein Datensatz C des Mikrodatenfile gefunden wird, der eine geringere Distanz zu B aufweist als A zu B. Der Versuch, dieses Problem durch die Berechnung aller Distanzen zwischen den Datensätzen des Identifikationsfile gegenüber denen des Mikrodatenfile zu lösen, würde bereits ab einer relativ geringen Anzahl von Fällen an zu hohen Rechenzeiten scheitern. Nach den Regeln der Kombinatorik ist leicht ersichtlich, daß diese prohibitiv hoch werden.

Sind einfache Selektionsverfahren in gewisser Weise zu wenig sensitiv für die in den Daten enthaltene Information, so wäre die Methode "Zuordnung durch Distanzminimierung" unter vielen Umständen zu reagibel. Eine geringe Distanz zwischen zwei Datensätzen kann dadurch zustande kommen, daß

a) zwei Datensätze zwar von der gleichen Person stammen, in den Datenfiles jedoch leicht unterschiedlich abgebildet werden;

b) zwei Datensätze von zwei verschiedenen Personen stammen, die sich in der Realität geringfügig voneinander unterscheiden und in den Datenfiles korrekt abgebildet sind;

c) zwei Datensätze von zwei sich in der Realität geringfügig unterscheidenden Personen stammen und in den Datenfiles zusätzlich mit geringfügigen Fehlern abgebildet sind.

Das Verfahren könnte zwischen diesen Fällen nicht unterscheiden und würde auf diese Weise das Potential von zuordnungsfähigen Fällen in kaum kontrollierbarer Weise erhöhen. Diese Reidentifikationstechnik könnte daher nur unter sehr speziellen Umständen anwendbar sein. Solche wären in erster Linie dann gegeben, wenn sehr viele Überschneidungsmerkmale vorliegen würden und der Angreifer wüßte, daß die gesuchte Person im Mikrodatenfile enthalten ist. Denn wenn der Datenangreifer von einer bestimmten Person in seinem Identifikationsfile wüßte, daß sie im Mikrodatenfile enthalten ist, wäre das Problem unter anderem rechentechnisch einfacher zu handhaben, weil dann nur von einer Person ausgehend die Distanzen zu dem im Mikrodatenfile enthaltenen Fällen berechnet werden müßten.

Das von einem Distanzminimierungsverfahren möglicherweise ausgehende Gefährdungspotential braucht hier allerdings nicht mehr weiter präzisiert zu werden, da mit der diskriminanzanalytischen Reidentifikationstechnik von Paaß/Wauschkuhn eine Methode vorliegt, die den Anspruch erhebt, die aufgezeigten Probleme zu lösen, wobei die Distanzminimierung eine der von dem Verfahren berücksichtigten Komponenten ist.

3.4 Diskriminanzanalytische Reidentifikationstechnik nach Paaß/Wauschkuhn

Von der Gesellschaft für Mathematik und Datenverarbeitung (GMD) in St. Augustin wurde im Rahmen des AIMIPH-Projekts eine statistische Reidentifikationstechnik für die Beurteilung von Reidentifikationsrisiken entwickelt, die den Anspruch erhebt, die Defizite der bisher dargestellten Reidentifikationstechniken zu vermeiden. Dieser auf der Diskriminanzanalyse beruhende Algorithmus gilt als die "ultima ratio", die einem Angreifer zur Verfügung stehen könnte: "Im Rahmen des Projekts AIMIPH sollte ... die maximale Höhe des Reidentifikationsrisikos bestimmt werden. Das Risiko ist offensichtlich maximal, wenn der Angreifer das beste Verfahren benutzt, welches die verfügbaren Informationen in optimaler Weise ausnutzt. Da das ... (diskriminanzanalytische Verfahren - die Verf.) das beste uns bekannte Verfahren war, wurde es für die Reidentifikationsexperimente verwendet" (Paaß/Wauschkuhn 1985:74).

Das Verfahren der GMD basiert zwar auf der linearen Diskriminanzanalyse, kombiniert diese jedoch mit zahlreichen anderen mathematischen und statistischen Methoden zu einem neuen Algorithmus. Es stellt somit eine Eigenentwicklung der GMD dar und wird in der vorliegenden Arbeit als "diskriminanzanalytische Reidentifikationstechnik" bezeichnet. Mit diesem Verfahren der GMD wird eine Zuordnung in zwei Schritten vorgenommen. Im ersten Schritt wird für jeden gegebenen Datensatz des Identifikationsfile derjenige Datensatz des Mikrodatenfile ermittelt, der in den Ausprägungen der Überschneidungsmerkmale die beste Übereinstimmung aufweist. Das Verfahren erlaubt eine Zuordnung auch dann, wenn die beiden Datensätze keine identischen Ausprägungen aufweisen (vgl. Situation 2 in Abbildung 3.1). Es sucht sich den "ähnlichsten" Fall des Mikrodatensatzes heraus und beinhaltet daher ein Distanzminimierungsverfahren.

Im zweiten Schritt wird die Wahrscheinlichkeit berechnet, daß die Zuordnung korrekt ist. Diese ist um so höher, je weniger weitere ähnliche Datensätze im Mikrodatenfile vorhanden sind (man vergleiche die Situation 2 und 3 in Abbildung 3.1). Der betreffende Wert der ermittelten Wahrscheinlichkeit wird mit einer vorgegebenen Sicherheitsschranke konfrontiert. Wird diese Schranke überschritten, so kann man davon ausgehen, daß beide Datensätze von derselben Person stammen und damit eine Reidentifikation stattgefunden hat (Paaß/Wauschkuhn 1985, Paaß 1985a, 1985b, 1987, 1988a, Bender 1990).

Das Verfahren wurde bisher allerdings nur an Datenmaterial überprüft, das aus einem empirischen Mikrodatenfile und einem hieraus synthetisch gewonnenen

Identifikationsfile bestand. Letzteres war aus dem Mikrodatenfile durch Überlagerung mit zufälligen Abweichungen generiert worden. Auf diese Weise wurden Datenfehler und andere Inkompatibilitäten simuliert. Bei diesem Material ergaben sich unter bestimmten Randbedingungen vergleichsweise hohe Reidentifikationsrisiken.

3.4.1 Konzeption der Methode

Grundansatz und Ablauf der diskriminanzanalytischen Reidentifikationstechnik werden im folgenden rekonstruiert. Ihr entscheidender Vorzug gegenüber einfacheren Verfahren ist die explizite Berechnung der Wahrscheinlichkeit, daß Datensätze im Mikrodatenfile und Identifikationsfile von der gleichen Person stammen. Der Angreifer hat damit die Möglichkeit eine Sicherheitsschwelle (im Sinne der statistischen Testtheorie) zu wählen, die er für ausreichend hält. Wird bei einer Anwendung z.B. ermittelt, daß Fall A aus dem Identifikationsfile und Fall B aus dem Mikrodatenfile mit 90%iger Wahrscheinlichkeit zusammengehören, kann er das je nach Neigung oder Ziel als Zuordnung akzeptieren oder nicht.

Für die Berechnung der Wahrscheinlichkeit einer korrekten Zuordnung wird auf die Diskriminanzanalyse zurückgegriffen. Bei diesem Verfahren wird davon ausgegangen, daß die Grundgesamtheit in mehrere disjunkte Teilgesamtheiten (im folgenden "Klassen" genannt) zerlegbar ist. Die Anzahl der Klassen und die Verteilung der Häufigkeiten in den Klassen müssen bekannt sein. Jeder Fall soll dabei genau einer Klasse zugehörig sein.

Die Aufgabenstellung, die mittels der Diskriminanzanalyse gelöst werden soll, kann man allgemein folgendermaßen beschreiben: Ein Fall, von dem nicht bekannt ist welcher Klasse er angehört, soll anhand einer Reihe von bekannten Merkmalen, die zur Klassenzugehörigkeit in einer Abhängigkeitsbeziehung stehen, einer der möglichen Klassen zugeordnet werden. Die Zuordnung dieser Fälle erfolgt mit bestimmten Wahrscheinlichkeiten, ist also mit einer angebbaren Unsicherheit belastet.

Der Grundansatz der Diskriminanzanalyse[5] soll an einer typischen Fragestellung aus dem Bereich der Wahlforschung, nämlich der Weigerung von Wählern ihre Parteipräferenz in Umfragen anzugeben, erläutert werden:

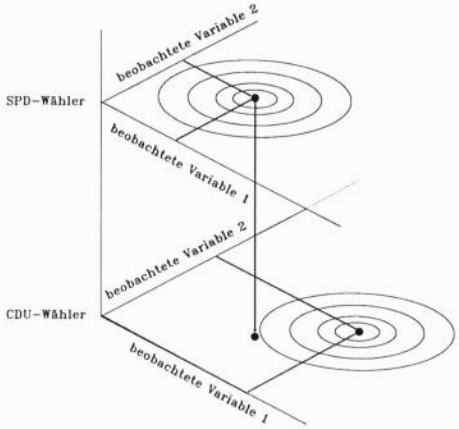
Hierbei wird die Klassenzugehörigkeit durch die verschiedenen Parteipräferenzen von Personen (CDU, SPD) definiert. Unter der Annahme, die Parteipräferenz einer Anzahl von Personen sei nicht bekannt, kann mittels der Diskriminanzanalyse eine Zuordnung zu den Klassen der SPD- bzw. CDU-Wähler vorgenommen werden. Dies ist dann möglich, wenn für die verschiedenen Personen zusätzliche (z.B. demographische) Merkmale vorliegen und sich der "typische" CDU-Wähler hinsichtlich seiner Merkmalsausprägungen von einem "typischen" SPD-Wähler unterscheidet. Die Wahrscheinlichkeit einer korrekten Zuordnung steigt, je stärker diese Unterschiede sind und je mehr sich die Wähler mit gleicher Parteipräferenz hinsichtlich ihrer Merkmalsausprägungen ähneln.

Normalerweise wird zur Anwendung der Diskriminanzanalyse zunächst eine sogenannte "Lernstichprobe" (vgl. Fahrmeir/Häußler/Tutz 1984:310) benötigt. Für die Personen in dieser Stichprobe sind sowohl die Klassenzugehörigkeit (im Beispiel: Parteipräferenzen) als auch die Ausprägungen der weiteren Merkmale bekannt, so daß der Zusammenhang zwischen den beobachteten Variablen und der Klassenzugehörigkeit geschätzt werden kann. Auf dieser Grundlage kann im nächsten Schritt die unbekannte Klassenzugehörigkeit von Personen geschätzt werden.

---

5 Allgemein dazu vgl. Fahrmeir/Häußler/Tutz (1984), eine Einführung bieten Backhaus et al. (1989).

Abbildung 3.2: Hypothetische Verteilung von SPD- und CDU-Wählern über zwei beliebige Merkmale



Im allgemeinen Fall spannen die Merkmale einen Raum auf, der so viele Dimensionen hat wie Variablen für die Zuordnung verwendet werden. In Abbildung 3.2 wird, zur Verdeutlichung, von lediglich zwei Merkmalen ausgegangen. Ihre Wertekombinationen liegen in einer Ebene. In der dritten Dimension wurden die einzelnen Klassen (hier: Parteipräferenz) abgetragen. Bestimmte Werte für die Wahrscheinlichkeit, daß die Zugehörigkeit zu der jeweiligen Klasse mit einer bestimmten Ausprägungskombination der Merkmalen verknüpft ist, werden hier durch konzentrisch angeordnete "Höhenlinien" dargestellt. Weiter innen liegende Höhenlinien symbolisieren höhere Werte der Wahrscheinlichkeitsdichte. Punkte auf den inneren Linien liegen näher bei der typischen Ausprägungskombination der Wähler einer bestimmten Partei.

Die Abbildung zeigt, daß sich die beiden Klassen (hier: SPD- und CDU-Präferenz) hinsichtlich der Wahrscheinlichkeiten für die beiden dargestellten Merkmale unterscheiden. Die Höhenlinien der Dichten liegen nicht genau übereinander, sondern versetzt in der ihnen zugehörigen Ebene. Der "typische" SPD-Anhänger hat in beiden beobachteten Variablen eine niedrigere Aus-

prägung als der "typische" CDU-Anhänger. Die betreffenden Wahrscheinlichkeitsdichten werden auf der Basis der Lernstichprobe geschätzt.

Um die Diskriminanzanalyse für das spezifische Problem der Zuordnung von Personen eines Identifikationsfile zu Personen eines Mikrodatenfile anwenden zu können, muß die Person k, die Träger des Merkmalsvektors y_k (k=1,...,N) des Mikrodatenfile ist, als jeweils eigene Klasse {k} aufgefasst werden. y_k symbolisiert also den mehrdimensionalen Vektor für die k-te Person aus der Gesamtheit der N Fälle des Mikrodatenfile.[6] Die Datensätze des Identifikationsfile werden nun diesen Klassen zugeordnet. Die Klasse, der der Fall des Identifikationsfile zugeordnet wird, soll k* genannt werden.

Bei der praktischen Durchführung der diskriminanzanalytischen Reidentifikationstechnik wird für einen Fall des Identifikationsfile (z) die Wahrscheinlichkeit berechnet, daß dieser zu einer Klasse des Mikrodatenfile gehört. Ist diese Wahrscheinlichkeit hoch, wird angenommen, daß der betreffenden Fall von der gleichen Person abstammt, wie der k-te Fall des Mikrodatenfiles.

Die Zuordnung erfolgt anhand der Überschneidungsmerkmale von Identifikationsfile und Mikrodatenfile. Dabei ist die Art der Zuordnung eine andere als dies im einfachen Zuordnungsverfahren durch Datenabgleich der Fall ist. Im einfachen Zuordnungsverfahren wird in beiden Files nach identischen Ausprägungen gesucht. Sobald Identität vorliegt, erfolgt eine Zuordnung.

Im Fall der diskriminanzanalytischen Reidentifikationstechnik wird dagegen die Ähnlichkeit der Fälle in den beiden Files als Entscheidungskriterium herangezogen, d.h. eine Entscheidung für eine Zuordnung wird wahrscheinlicher, je näher die Ausprägungskombination eines Falles des Identifikationsfile an der Ausprägungskombination eines Falles des Mikrodatenfile liegt und je unähnlicher die restlichen Fälle des Mikrodatenfile und des Identifikationsfile zu dieser spezifischen Ausprägungskombination sind (vgl. Abbildung 3.1 in Abschnitt 3.1.1).

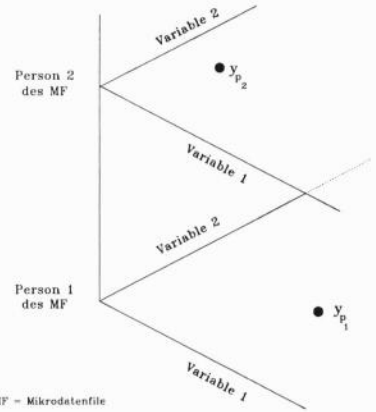
Abweichend von der bisherigen Darstellung besteht im Reidentifikationsprozeß jede Klasse jeweils nur aus einer Person, d.h. es wird nicht wie im obigen Beispiel eine Dichte im Raum aufgespannt. Die jeweilige Klasse ist vielmehr durch

---

6 Im vorliegenden Fall soll zunächst angenommen werden, daß eine gleiche Anzahl von Klassen und Personen im Mikrodatenfile vorliegt, d.h. es wird unterstellt, daß jeder Fall des Mikrodatenfile durch eine einzigartige Ausprägung von Merkmalen gekennzeichnet ist. Von dieser Annahme kann auch abgewichen werden (vgl. Abschnitt 3.4.2.2).

einen über den Ausprägungen der Überschneidungsmerkmale definierten Punkt y_p im Merkmalsraum repräsentiert.

Abbildung 3.3: Hypothetische Verteilung von zwei Überschneidungsmerkmalen für zwei Personen aus dem Mikrodatenfile

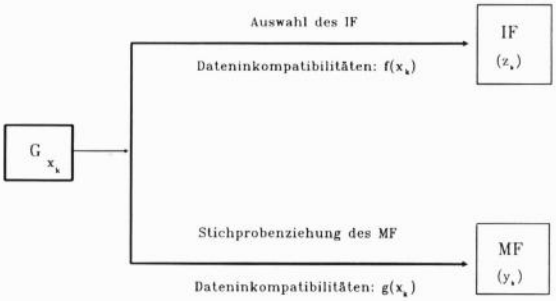


Die Diskriminanzanalyse ist dennoch anwendbar: Es wurde bereits dargelegt, daß unter empirischen Bedingungen die Daten von Mikrodatenfile und Identifikationsfile in der Regel nicht vollständig kompatibel sein werden. Gelingt es, diese Dateninkompatibilitäten durch spezifische Dichten zu beschreiben, kann aus ihnen eine Dichte um jene Punkte y_p (d. h. um die Fälle im Mikrodatenfile) geschätzt werden (vgl. Abbildung 3.3).

Die Dichteschätzungen sind Teil der von Paaß/Wauschkuhn gewählten Strategie, mit der die Anwendung einer Lernstichprobe (wie oben allgemein dargestellt) ersetzt werden soll. Eine solche Lernstichprobe steht nämlich nicht zur Verfügung

Bei der Modellierung von Dateninkompatibilitäten wird angenommen, - wie in Abbildung 3.4 - daß sowohl das Mikrodatenfile als auch das Identifikationsfile Abweichungen von den "wahren" Werten der Grundgesamtheit aufweisen. Dabei wird unterstellt, daß solche "wahren" Werte existieren. Die Abweichungen im Mikrodatenfile und im Identifikationsfile werden demgemäß als "Fehler" aufgefaßt (vgl. Paaß/Wauschkuhn 1985:64ff.), hier werden sie generell als Inkompatibilitäten bezeichnet. In diesen Begriff sind sowohl einfache Erhebungs- und Verkodungsfehler eingeschlossen, als auch unterschiedliche zeitliche und sachliche Bezüge der Datenbestände, die dem Identifikationsfile und dem Mikrodatenfile zugrundeliegen (für eine systematische Behandlung vgl. Abschnitt 4.3).

Abbildung 3.4: Zusammenhang zwischen Mikrodatenfile und Identifikationsfile bei dem diskriminanzanalytischen Verfahren von Paaß/Wauschkuhn (vgl. Bender 1990:18)



G = Grundgesamtheit IF = Identifkationsfile MF = Mikrodatenfile

Unter der Annahme, daß Dateninkompatibilitäten auftreten können, müssen für die Beschreibung einer Person k, deren Daten im Mikrodatenfile und im Identifi-

kationsfile enthalten sind, drei verschiedene Vektoren der Überschneidungsmerkmale herangezogen werden:

x_k: Vektor mit den wahren Merkmalsausprägungen der k-ten Person.

y_k: Merkmalsvektor, welcher aus den von den wahren Werten abweichenden Merkmalsausprägungen der k-ten Person im Mikrodatenfile hervorgeht.

z_k: Merkmalsvektor mit den von den wahren Werten abweichenden Merkmalsausprägungen der k-ten Person im Identifikationsfile.

Man sollte sich hierbei allerdings vergegenwärtigen, daß die k-te Person nicht an der k-ten Stelle in beiden Files stehen muß, d.h. die Files können eine unterschiedliche Ordnung besitzen. Ein personenbezogenes Merkmal ist zumindest für den Mikrodatenfile nicht gegeben, da für den Angreifer die Identität der einzelnen Personen nicht bekannt ist.

Wird die durch Inkompatibilitäten bei der Sammlung und Verkodung der Daten auftretende Verzerrung von wahren Merkmalsausprägungen als zufällig aufgefaßt, so gilt:

$$z_k = f(x_k) \qquad \text{und} \qquad y_k = g(x_k) \qquad \text{für } k=1,...,N$$

Die Veränderung der x_k über f und g werden in Abbildung 3.4 jeweils durch Pfeile symbolisiert. Es soll angenommen werden, daß ein Angreifer sich einen beliebigen Datensatz y_i aus dem Mikrodatenfile und einen beliebigen Datensatz z_j aus dem Identifikationsfile herausnimmt. Es ist dann davon auszugehen, daß er nicht exakt weiß, welche Arten von Abweichungen bei der Abbildung der realen Fälle x_k in den Datensätzen aufgetreten sind, da die durch f und g[7] gekennzeichneten Prozesse zufällige Veränderungen beinhalten.[8]

Um den Abweichungsprozeß nachzuvollziehen, müssen *externe Informationen* - d.h. solche, die nicht im Mikrodatenfile und Identifikationsfile enthalten sind -

---

7 Im Grenzfall mögen allerdings die Inkompatibilitäten bei den zum Identifikationsfile gehörenden Datensätzen gegen 0 gehen, nämlich dann, wenn der Angreifer die zugehörigen Personen genau kennt. Jedoch ist auch das Alltagswissen über bekannte Personen potentiell fehlerbehaftet. Es ist möglich, sich z.B. im Alter eines Arbeitskollegen zu irren. Außerdem ist die Realität vieldeutig und so wird es in vielen Fällen schwierig sein, selbst die genaue Berufsbezeichnung dieses Kollegen sicher anzugeben.

8 Wiederum könnte es unter bestimmten Umständen möglich sein, im Fall des Identifikationsfile genauere Aussagen zu treffen.

für die *Abschätzung der Inkompatibilitäten* herangezogen werden. Für die Anwendung der diskriminanzanalytischen Reidentifikationstechnik wird hierbei nicht nur vorausgesetzt, daß der Angreifer die Verteilungstypen von f und g (vgl. Abbildung 3.4) kennt, sondern auch die Parameter der Verteilungen.

Die Notwendigkeit, einen Prozeß zu spezifizieren, der Dateninkompatibilitäten nachzeichnen kann, setzt Kenntnisse in erheblichem Umfang auf seiten des Angreifers voraus. Er muß nicht nur "einfache" Erhebungs- und Verkodungsfehler berücksichtigen, sondern auch andere Dateninkompatibilitäten.[9] Er muß z.B. eine quantifizierbare Vorstellung von der Häufigkeit von Veränderungen einzelner Merkmale im Zeitablauf haben, wenn sein Identifikationsfile und sein Mikrodatenfile nicht am gleichen Tag erhoben wurden. Er muß schließlich die Wahrscheinlichkeit von Unterschieden im Antwortverhalten von ein und derselben Person bei der Erhebung seiner beiden Datenquellen abschätzen.

Diese Anforderungen überschreiten, wie in Abschnitt 4.3 aufgezeigt werden wird, zum Teil die in den Sozialwissenschaften überhaupt vorhandenen Kenntnisse und begründen damit Zweifel an der Einsatzfähigkeit der diskriminanzanalytischen Reidentifikationstechnik unter realistischen Bedingungen, d.h. bei der Verwendung von empirischen Daten.

Diese Bedenken können jedoch zunächst zurückgestellt werden, da sich im AIMIPH-Projekt zeigte, daß das Verfahren sehr robust gegen eine Fehlspezifikation der Dateninkompatibilitäten durch den Angreifer ist (vgl. Paaß, Wauschkuhn 1985:190). Danach wären bereits sehr ungenaue Angaben ausreichend. Auch die generelle Höhe der Dateninkompatibilitäten muß, gemäß den vorab vorliegenden Ergebnissen, nicht sehr genau getroffen sein. Unter diesen Vorausetzungen müßte es daher auch bei Verwendung von empirischen Daten möglich sein, einen sinnvoll verwendbaren Fehlerprozeß zu spezifizieren.

3.4.1.1 Der Bayessche Schätzansatz[10]

Paaß/Wauschkuhn bedienen sich bei der Berechnung des Reidentifikationsrisikos der Bayesschen Entscheidungsregel (vgl. Iversen 1984). Dieser Ansatz stellt ein generelles Schätzprinzip dar. Danach läßt sich die *Wahrscheinlichkeit einer korrekten Einordnung eines bestimmten Falls z aus dem Identifikationsfile*

---

9 Eine ausführliche Erörterung über Ursachen, Ausmaß und Konsequenzen von Dateninkompatibilitäten erfolgt in Abschnitt 4.3.

10 Für die genauere Herleitung vgl. Bender (1990).

in eine gegebene "Klasse" k des Mikrodatenfile wie folgt bestimmen (vgl. Paaß/Wauschkuhn 1985:108; Fahrmeir et al. 1984:303f.):

$$(1) \quad p(k|z) = \frac{p(z|k)\, p(k)}{\sum_{k=1}^{N} p(z|m)\, p(m)}$$

Zu $p(k|z)$:

Die Dichte[11] $p(k|z)$ ist dabei die a posteriori Wahrscheinlichkeit, daß ein Fall mit beobachteten Merkmalsvekor z im Identifikationsfile der Klasse k angehört (vgl. Fahrmeir/Häußler/Tutz 1984:303). Diese wird als *Reidentifikationsrisiko* bezeichnet und gibt an, mit welcher Wahrscheinlichkeit ein gegebener Fall z einem bestimmten Fall y_k aus dem Mikrodatenfile zugeordnet werden kann (vgl. Paaß/Wauschkuhn 1985:108).

Zu $p(z|k)$:

Die Informationen, die einem Angreifer über Dateninkompatibilitäten zur Verfügung stehen, dienen - neben Identifikationsfile und Mikrodatenfile - als *zusätzlicher Input* für eine Schätzung der bedingten Dichten $p(z|k)$. Für $k=1,2,...,N$ ist $p(z|k)$ die gesuchte Klassendichte. An ihr ist abzulesen, mit welcher Wahrscheinlichkeit eine bestimmte Wertekombination z unter Berücksichtigung von Dateninkompatibilitäten in Klasse k einzuordnen ist. Gemäß der Bayesschen statistischen Theorie werden diese Dichten auch Klassenverteilungen genannt. Diese Dichten sind die Grundlage für die Wahrscheinlichkeit, daß ein gegebener Datensatz z des Identifikationsfile einem Datensatz y_k des Mikrodatenfile zugehörig ist, d.h. von der gleichen Person stammt (vgl. Paaß/Wauschkuhn 1985:66).

Zu $p(k)$:

Für $k=1,...,N$ ist $p(k)$ die "a priori" Wahrscheinlichkeit für Klasse k. Die a priori Wahrscheinlichkeit für Klasse k beträgt 1/N, da jede Klasse einelementig ist. Diese Implikation geht sowohl in den Zähler ($p(k)$) als auch in den Nenner ($p(m)$) ein.

---

[11] Der Einfachheit halber wird in diesem Abschnitt immer von Dichten gesprochen, obwohl bei diskreten Zufallsvariablen Massefunktion der exaktere Begriff wäre.

Wie bereits erwähnt, kann mit dem Bayesschen Ansatz die Wahrscheinlichkeit einer korrekten Zuordnung berechnet werden. Da $p(k|z)$ Werte zwischen O (Zuordnung ist mit Sicherheit falsch) und 1 (Zuordnung ist mit Sicherheit richtig) annehmen kann, hat ein Angreifer die Wahl: Er kann sich gemäß seiner *Risikoaversion* für ein bestimmtes *Sicherheitsniveau* entscheiden und muß dazu eine bestimmte Schranke (z. B. 90 oder 99 Prozent) festlegen, oberhalb derer er eine Zuordnung als richtig betrachten will.

Die Implikationen von Gleichung (1) werden deutlicher, wenn sie umgeformt und vereinfacht werden. Als Konstante kann p(m) als Multiplikator vor das Summenzeichen gesetzt werden. Gleichung (1) vereinfacht sich durch Kürzen von p(k) im Zähler und Nenner zu:

$$(1a) \quad p(k|z) = \frac{p(z|k)}{\sum_{k}^{N} p(z|m)}$$

Unter Verwendung des Satzes der totalen Wahrscheinlichkeit kann der Nenner in Gleichung (1) jedoch auch durch folgende Beziehung dargestellt werden:

$$\sum_{k}^{N} p(z|m)p(m) = p(z)$$

Gleichung (1) vereinfacht sich durch diese Transformation zu:

$$(2) \quad p(k|z) = \frac{p(z|k)\,p(k)}{p(z)} \quad \text{, bzw da } p(k)=1/N \text{ zu}$$

$$(2a) \quad p(k|z) = \frac{p(z|k)}{N\,p(z)}$$

Zu p(z):

Die Dichte p(z) ist die Wahrscheinlichkeit für die Ausprägungskombination z im Identifikationsfile. Sie wird als Dichte der Identifikationsfile-Verteilung bezeichnet. Den Wert 0 besitzt sie nur, wenn z nicht im Identifikationsfile ist (Paaß/Wauschkuhn 1975:67).

Da bei gegebenen z der Ausdruck im Nenner auf der rechten Seite von Gleichung (2a) konstant ist, variiert $p(k|z)$ nur in Abhängigkeit vom Ausdruck im Zähler, d.h. konkret von der Wahl der Klasse k. Nach Gleichung (2a) erreicht $p(k|z)$ deshalb, bei vorgegebenen z, genau dann sein Maximum, wenn $p(z|k)$ maximal ist. D.h. ein Angreifer wird z derjenigen Klasse k^* zuordnen, die die höchste Klassendichte besitzt. Diese Dichte für die oben definierte Klasse k^* soll mit $p(z|k^*)$ bezeichnet werden. Für diese bestimmte Klasse k^* ergibt sich folgende Beziehung:

$$(3) \quad p(k^*|z) = \frac{p(z|k^*)}{N\,p(z)}$$

3.4.1.2 Annahmen über die Verteilungsform von Dateninkompatibilitäten als Voraussetzung zur Anwendung der linearen Diskriminanzanalyse

Würden keine Dateninkompatibilitäten vorliegen und beide Files jeweils die Population umfassen, dann hätte $p(z|k^*)$ bei einzigartigen Ausprägungskombinationen (d.h. bei Einzelfällen) folgendes Aussehen (Paaß/Wauschkuhn 1975:119):

$$p(z|k^*) = \begin{cases} 1 & \text{falls } z=y_{k^*} \\ 0 & \text{falls } z \neq y_{k^*} \end{cases}$$

Hätte man einen Fall y_{k^*}, der identisch mit z ist, dann könnte man diese beiden Fälle mit absoluter Sicherheit zuordnen. Da in der Realität allerdings Dateninkompatibilitäten vorliegen werden, müssen bestimmte Annahmen über die Verteilungen von $p(z|k^*)$ gemacht werden. Selbst wenn $z=y_{k^*}$ wäre, würde $p(z|k^*)$ voraussichtlich einen Wert kleiner 1 erreichen.

Von Paaß/Wauschkuhn wurde zur Lösung des Zuordnungsproblems der spezifische Ansatz der linearen Diskriminanzanalyse gewählt.[12] Zur Begründung für die Wahl dieser Reidentifikationstechnik verweisen Paaß/Wauschkuhn (1985:128) auf empirische Untersuchungen von Titterington et al. (1981), die verschiedene Algorithmen verglichen, wobei die lineare Diskriminanzanalyse besonders gute Ergebnisse erzielte.[13]

Die Anwendung der linearen Diskriminanzanalyse auf das Problem der Reidentifikation ist jedoch mit bestimmten Restriktionen verbunden. Mehrere dieser Restriktionen betreffen die Dichte $p(z|k^*)$. Es ist aufgrund der unten dargestellten Annahmen möglich, eine Dichte aufzuspannen (vgl. Abbildung 3.2), hierbei sind jedoch folgende Restriktionen zu beachten:

---

12 Das Problem der Zuordnung von Fällen zu vorgegebenen Klassen, das für die Diskriminanzanalyse kennzeichnend ist, kann mit verschiedenen Algorithmen gelöst werden. Eine Übersicht findet sich in Fahrmeir et al. (1984), Jain (1987) oder auch in Dillon/Goldstein (1984).
In jüngerer Zeit sind Ansätze bekannt geworden, die auf den ersten Blick für das beschriebene Problem adäquater erscheinen als die lineare Diskriminanzanalyse. Ausgehend von der Theorie der "Fuzzy Sets" wurde eine Diskriminanzanalyse entwickelt, bei der unscharfe Definitionen der Klassen von vornherein berücksichtigt sind. Dieser Ansatz scheint die Fehlerproblematik bereits zu integrieren. Er erscheint um so erfolgversprechender als die Theorie der "Fuzzy Sets" (die auf Zadeh 1965 zurückgeht, für eine allgemeine Darstellung vergleiche die Beiträge in Zimmermann, Zadeh, Gaines 1984, insbesondere als Einführung Carlsson 1984) in jüngster Zeit vielfältige erfolgreiche technische Anwendungen erfahren hat und die betreffende Form der Diskriminanzanalyse auch für die Sozialwissenschaften empfohlen wird (vgl. Smithson 1987, S. 109ff.). Allerdings zeigt eine Analyse der Behandlung des Diskriminationsproblems, die auf einer "Fuzzified Bayes Formula" beruht (vgl. Okunda et al. 1984), daß dabei unterstellt wird, die Klassen wären nicht scharf definiert, sondern die "wahre" Zugehörigkeiten von Objekten ist bereits nur mit einer bestimmten Wahrscheinlichkeit angebbar. Die Identität von Personen im Reidentifikationsproblem ist jedoch immer exakt definiert. Der Lösungsversuch mittels Fuzzy Sets wäre im vorliegenden Fall also unangemessen.

13 Von Titterington et al. (1981) wurden sieben diskriminanzanalytische Schätzverfahren (Unabhängigkeitsmodell, Lancaster mit Interaktionen 1.Ordnung, Latent class models, Dichteschätzer mit Kern-Schätzer, Lineares logistisches Modell, Lineare und Quadratische Diskriminanzfunktion) u.a. hinsichtlich der Zahl von Falschzuordnungen und Zuverlässigkeit (Vergleich von tatsächlichem mit dem erwarteten Dichtewert) untersucht. Die lineare Diskriminanzfunktion war invariant gegen verschiedene Verteilungen der Klassen, und erbrachte in allen Simulationsrechnungen gute Schätzungen.

a) "Quasi"-Normalverteilung in den Klassen.

Die lineare Diskriminanzanalyse unterstellt zunächst die Normalverteilung innerhalb der Klassen. Gegen eine Verletzung dieser Annahme ist das Verfahren zwar relativ robust, die Abweichung darf allerdings nicht zu stark sein. Damit sind aber die Annahmen über die Dateninkompatibilitäten nicht mehr beliebig variierbar.
Im Sinne eines Kompromisses unterstellten Paaß/Wauschkuhn nicht die Normalverteilung, sondern eine Verteilungsform, die eine starke Ähnlichkeit mit dieser hat: Da man keine Kenntnis über den wahren Wert der betreffenden Person des Mikrodatenfile hat, wird angenommen, daß der im Mikrodatenfile enthaltene Wert der wahrscheinlichste ist. Die Dichte hat demnach in jeder Klasse k genau im Punkt y_p ihr Maximum. Dieser Punkt y_p wird durch den Merkmalsvektor y_k festgelegt, d.h. $y_p=y_k$.
Weiterhin wird angenommen, daß Ausprägungskombinationen gleicher Wahrscheinlichkeit auf Ellipsen um diesen Hochpunkt liegen und daß die Dichtewerte von diesem Hochpunkt ausgehend monoton abnehmen. Die Dichten der Dateninkompatibilitäten sind demnach eingipflig und glatt. Paaß/Wauschkuhn konzedieren, daß unter diesen Umständen die Behandlung von ordinalen und nominalen Variablen Schwierigkeiten machen könnte. Damit solche Variablen mit mehr als zwei Ausprägungen überhaupt in die Analyse einbezogen werden können, müssen sie dichotomisiert werden.

b) Die Varianz-Kovarianzmatrizen der Dateninkompatibilitäten sind für alle Klassen k identisch.

Mit dieser Annahme wird unterstellt, daß sich die Dichten in den einzelnen "fehlerbehafteten" Klassen nicht in ihrer Form, sondern nur in ihrer Lage in den Klassenebenen unterscheiden (vgl. Abbildung 3.2). Dies bedeutet, daß die angenommenen Dateninkompatibilitäten im Mikrodatenfile für alle Personen die gleiche Form besitzen. Es erscheint zweckmäßig, diese Homogenitätsvoraussetzung zugrunde zu legen, da andernfalls weitreichende Kenntnisse des Angreifers vorausgesetzt werden müßten. Demgemäß haben Paaß/Wauschkuhn in ihren praktischen Experimenten identische Kovarianzmatrizen in den Klassen vorausgesetzt und nicht angenommen, daß der Angreifer über exakte Kenntnisse der Dateninkompatibilitäten verschiedener Personengruppen verfügt.
Die Annahme klassenweise identischer Kovarianzmatrizen hat Konsequenzen für das zu verwendende statistische Modell. Erst durch die Festlegung von bekannten "Quasi"-Normalverteilungen in den Klassen kann man vom allgemeinen Modell der Diskriminanzanalyse zur linearen Diskriminanzanalyse übergehen. Durch die weitere Annahme der klassenweisen identischen Kovarianzmatrizen wird ein bestimmtes Abstands-

maß, die Mahalanobisdistanz, festgelegt. Dieses Abstandsmaß hat hohe Werte für zwei Fälle, die sich unähnlich sind und niedrige Werte für zwei Fälle, die sich gleichen.

3.4.1.3 Die Mahalanobisdistanz[14]

Mit Hilfe der Mahalanobisdistanzen wird eine Transformation von Wahrscheinlichkeiten, wie sie in Abschnitt 3.4.1.1 für die Definition des Reidentifikationsrisikos Verwendung fanden, auf Distanzen vorgenommen. Unter der Annahme, daß eine bestimmte Verteilung vorliegt, kann die Wahrscheinlichkeit, daß ein gegebener Fall z einer Klasse k zugehörig ist, über eine Metrik durch Distanzen geschätzt werden.

Die Mahalanobisdistanz ist ein Abstandsmaß zwischen z und y_k[15], die folgendermaßen definiert ist:

$$(4) \quad ||y_k\text{-}z||^2 = (z - y_k)'\Sigma^{-1}(z - y_k)$$

Da die Wahrscheinlichkeiten monoton fallend sind, kann der Fall z, der zu einem Klassenmittelpunkt y_k die kleinste Distanz besitzt, als der wahrscheinlichste angenommen werden. Da weiterhin angenommen wurde, daß die a priori Wahrscheinlichkeiten der Klassen k gleich sind, wird z der Klasse zugeordnet, deren quadrierte Mahalanobis-Distanz am kleinsten ist.[16]

D.h. es muß folgendes gelten:

$$(5) \quad ||y_{k^*}\text{-}z||^2 = \min_k(||y_k\text{-}z||^2)$$

Eine Zuordnung erfolgt demnach über eine verallgemeinerte Minimum-Distanz-Regel: z wird vorläufig in diejenige Klasse k^* des Mikrodatenfile mit der geringsten Mahalanobis-Distanz einsortiert.

---

14 Die Betrachtung der Mahalanobisdistanz beschränkt sich im folgenden auf die konkrete Anwendung auf einen Reidentifikationsversuchs. Für eine allgemeine Herleitung der Mahalanobisdistanz vergleiche Fahrmeir/Hamerle 1984:317ff.

15 Die Mahalanobisdistanz könnte damit bei dem in Abschnitt 3.3 erörterten Distanzminimierungsverfahren angewendet werden.

16 Eine Herleitung findet sich bei Fahrmeir/Hamerle 1984:318f.

Im Rahmen dieses Konzeptes ist es möglich, einzelne Überschneidungsmerkmale unterschiedlich zu gewichten, je nachdem welche Wahrscheinlichkeiten der Dateninkompatibilitäten für einzelne Variablen spezifiziert werden. Damit können die Kovariationen von Dateninkompatibilitäten berücksichtigt werden.

Dazu ist es wichtig, sich die Funktion von Σ^{-1} in Gleichung (4) zu vergegenwärtigen. Σ^{-1} ist die Inverse der Varianz-Kovarianz-Matrix. Es werden Differenzen zwischen einem wirklichen Mikrodatenfile und einem künstlich erzeugten fehlerbehafteten Mikrodatenfile gebildet.[17] Aus diesen Differenzen wird eine Varianz-Kovarianz-Matrix der Dateninkompatibilitäten erzeugt, die in Gleichung (4) als Σ^{-1} eingeht.

Für die spätere Betrachtungen (vgl. Abschnitt 9.2) ist es wichtig, sich die Wirkung der Dateninkompatibilitäten auf die Klassendichten zu vergegenwärtigen. Die Klassendichten liegen, wegen der Annahme der Quasi-Normalverteilungen als Ellipsoide vor.

Die durch die Dichte gegebenen Ellipsoide enthalten alle y_k im Mikrodatenfile:

$$(6) \quad M_c^2(z) = \{ y_k: \ ||y_k - z||^2 \leq c^2 \}$$

$M_c^2(z)$ ist das Quadrat der Mahalanobisdistanz.

Würden alle Überschneidungsmerkmalen das gleiche Ausmaß an Dateninkompatibilitäten haben und wären die Dateninkompatibilitäten unkorreliert, so würde Σ^{-1} proportional zur Einheitsmatrix vorliegen. Die Form der Dichte, die durch Gleichung (6) festgelegt wird, wäre eine Kugel. Wenn die Dateninkompatibilitäten in unterschiedlichen Größenordnungen vorliegen, verformt sich die Kugel zu einem Ellipsoid. Je größer die Dateninkompatibilitäten in einem Überschneidungsmerkmal sind, desto größer ist die Streckung der Kugel.

Die Zuordnungsregel für die lineare Diskriminanzanlyse läßt sich in bezug auf die Mahalanobisdistanz auch folgendermaßen formulieren: Ordne den Fall z der Klasse k mit dem geringsten Wert für $M_c^2(z)$ zu. Mit der vorläufigen Zuord-

[17] Eine ausführlichere Darstellung des Programmalgorithmus erfolgt in Abschnitt 3.4.2

nung aufgrund der Mahalanobisdistanz kann vollzogen werden, was einleitend "erster Schritt" bei einem Reidentifikationsversuch mit der diskriminanzanalytischen Reidentifikationstechnik genannt wurde. Im zweiten Schritt wird dann anhand des Bayes Theorem eine Wahrscheinlichkeit berechnet. Übertrifft diese Wahrscheinlichkeit eine vorgegebene Schranke, so wird sich ein Angreifer dafür entscheiden, die Zuordnung für eine korrekte Identifikation zu halten.

3.4.1.4 Die Stichprobenproblematik

In der obigen Herleitung der diskriminanzanalytischen Reidentifikationstechnik wurde bisher davon ausgegangen, daß der Mikrodatenfile als Vollerhebung der Grundgesamtheit vorliegt. Ist das Mikrodatenfile allerdings nur eine Stichprobe, so zerfällt die Population in einen für den Angreifer bekannten und einen unbekannten Teil. Der Angreifer ist daher gezwungen den Nenner in Gleichung (1) zu schätzen.

Es ist die tatsächliche Dichte der Identifikationsfile-Verteilung p(z) bei Vorlage einer Stichprobe unbekannt. Diese wird im Verfahren von Paaß/Wauschkuhn mittels einer Vielfalt statistischer Methoden aus dem Identifikationsfile (oder wahlweise Mikrodatenfile) geschätzt (vgl. nachfolgenden Abschnitt). So schätzen Paaß/Wauschkuhn (1985:112ff.) anstelle von Gleichung (2a):

$$(7) \quad \hat{p}(k^*|z) = \frac{\hat{p}(z|k^*)}{\hat{p}(z|k^*) + (N-1)\,\hat{p}(z)}$$

wobei $\hat{p}(k^*|z)$ mit wachsender Stichprobengröße gegen den theoretischen Wert $p(k^*|z)$ konvergiert.

N bezeichnet dabei den Umfang der Grundgesamtheit (Mikrodatenfile oder Identifikationsfile), der für die Schätzung benutzt wird. $\hat{p}(z)$ ist eine Schätzung für die Dichte p(z).

3.4.1.5 Zur Interpretation des Verfahrens

Der Ausdruck (3) läßt eine Interpretation zu, die ein anschaulicheres Verständnis der Reidentifikationstechnik eröffnet. In (3) werden die bedingte Dichte $p(z|k^*)$, die sich aus den Angaben zu den Dateninkompatibilitäten ergibt, und die unbedingte Dichte $p(z)$, die die Wahrscheinlichkeit von Doppelgängern (oder von "ähnlichen" Personen) angibt, zueinander in Beziehung gesetzt. Je höher der Wert für den Gesamtausdruck, umso wahrscheinlicher ist, daß die beiden Fälle y_{k^*} und z die gleiche Person repräsentieren. Dies hat folgende Konsequenzen:

a) Die zwei fraglichen Fälle werden einander um so eher zugeordnet, je größer der Wert für die bedingte Dichte $p(z|k^*)$ ist. Dies ist plausibel, denn unter den getroffenen Annahmen über die Dateninkompatibilitäten wird $p(z|k^*)$ um so kleiner sein, je "unähnlicher" sich die zwei Fälle sind, d. h. je weiter sie im Merkmalsraum auseinanderliegen. Beim Zuordnungalgorithmus findet gleichzeitig auch eine Bewertung der Dimensionen dieses Merkmalsraumes statt. Variablen, die besonders inkompatibel sind, spielen für die Zuordnung eine geringere Rolle. Dies wird daraus ersichtlich, daß hohen Werten von $p(z|k^*)$ niedrige Werte der Mahalanobisdistanz entsprechen, wobei bei der Berechnung der Mahalanobisdistanzen die Distanzen für die einzelnen Merkmale gewichtet nach dem Grad ihrer Inkompatibilitäten in die Berechnung eingehen.

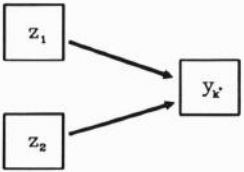
   Die Zuordnungsart, bei welcher ein Fall des Identifikationsfile genau einem Fall des Mikrodatenfile zugewiesen wird, soll, wie bei einfachen Abgleichverfahren, als eindeutige (eins zu eins) Zuordnung bezeichnet werden.

   Eine eindeutige Zuordnung kann hierbei - wie bei einfachen Abgleichverfahren - bedeuten, daß die zugeordneten Datensätze identische Ausprägungskombinationen aufweisen. Datensätze können auch dann in einer eins-zu-eins Form zugeordnet werden, wenn sie nur ähnliche Ausprägungskombinationen aufweisen und im Merkmalsraum keine anderen Fälle in der Nähe liegen.

Abbildung 3.5: Eindeutige Zuordnung

b) Wie Gleichung (2) noch einmal deutlich macht, beruht die diskriminanzanalytische Reidentifikationstechnik auf dem Ansatz der Einzelsuche, wobei jeweils nur ein Fall des Identifikationsfile mit allen Fällen des Mikrodatenfiles verglichen wird. D.h. eine Massenidentifikation wird in eine Reihe von Einzelsuchen aufgelöst, wobei die vorangegangenen Zuordnungswahrscheinlichkeiten der Fälle des Identifikationsfile unberücksichtigt bleiben. Es ist daher grundsätzlich möglich, daß zwei oder mehrere Fälle aus dem Identifikationsfile einem gleichen Fall des Mikrodatenfile (vgl. Abbildung 3.6) zugewiesen werden (mehrdeutige Zuordnung erster Art). Diese Art der mehrdeutigen Zuordnung kann die vom Angreifer gesetzte Wahrscheinlichkeitsschranke für eine korrekte Zuordnung überschreiten.

Abbildung 3.6: Mehrdeutige Zuordnung ("erster Art")

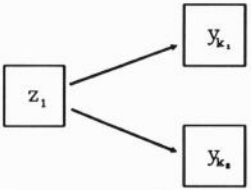


Es können mehr als zwei Fälle (z_1, z_2) gegeben sein.

c) Hingegen sind mehrdeutige Zuordnungen, bei welchen mehrere Fälle des Mikrodatenfile einem Fall des Identifikationsfile zugewiesen werden (vgl.

Abbildung 3.7), mit der Reidentifikationstechnik von Paaß/Wauschkuhn sehr unwahrscheinlich. Hohe Werte von p(z) sind in Gleichung (2) mit niedrigen Zuordnungswahrscheinlichkeiten verknüpft, da p(z) im Nenner steht, und somit den Wert für den Gesamtausdruck senkt. Dies ist eine Lösung für ein Doppelgängerproblem, das im folgenden mehrdeutige Zuordnung zweiter Art genannt werden soll.

Abbildung 3.7: Mehrdeutige Zuordnung ("zweiter Art")



Es können mehr als zwei Fälle (y_{k_1}, x_{k_2}) gegeben sein.

In Abbildung 3.1 (Abschnitt 3.1) ist die mehrdeutige Zuordnung zweiter Art die Situation 3, bei der sich viele Fälle aus dem Identifikationsfile auf relativ engem Raum befinden. Hier ist die Möglichkeit der Verwechslung groß. Wenn richtig geschätzt wird, werden hohe Werte für p(z) ermittelt und niedrige Zuordnungswahrscheinlichkeiten folgen. Umgekehrt heißt dies, daß die diskriminanzanalytische Reidentifikationstechnik Zuordnungen auch bei relativ großen Dateninkompatibilitäten ermöglicht, wenn die Ausprägungskombination z sehr wenig Ähnlichkeit mit anderen Kombinationen hat und damit p(z) sehr klein ist.

In bezug auf mehrdeutige Zuordnungen stellt die diskriminanzanalytische Reidentifikationstechnik daher nicht das Optimum eines Verfahrens dar, da der Angreifer sich ein Verfahren wünschen würde, das überhaupt keine mehrdeutigen Zuordnungen zuläßt. Nur dann kann er Verwechslungen vermeiden.

Der praktischen Umsetzung eines solchen (hypothetischen) Verfahrens, das auch keine mehrdeutigen Zuordnungen erster Art zulassen würde, stehen jedoch entscheidende Schwierigkeiten entgegen. Paaß/Wauschkuhn (1985:105) argumentieren zu Recht, daß eine andere Vorgehensweise als die gewählte schon bei kleinen Fallzahlen zu prohibitiv hohen Rechenzeiten führen

würde, da die Explosion der Kombinationsmöglichkeiten auch den schnellsten Rechner überfordern müßte.

Es bietet sich jedoch ein Ausweg an. Wird die Sicherheitsschwelle für eine korrekte Zuordnung hoch genug gesetzt (und funktioniert das Verfahren in der vorgesehenen Weise), werden Falschzuordnungen und damit mehrdeutige Zuordnungen entsprechend unwahrscheinlich, da es wahrscheinlicher wird, daß nur ein Fall des Identifikationsfile über der Sicherheitsschranke liegt.

3.4.1.6 Zusammenfassung

Zum besseren Überblick sei die Grundidee der diskriminanzanalytische Reidentifikationstechnik kurz zusammengefaßt.

- Jeder Datensatz des Mikrodatenfile wird als eine Klasse im Sinne der Diskriminanzanalyse aufgefaßt.

- Es werden Wahrscheinlichkeitsdichten innerhalb der Klassen geschätzt. Dies geschieht auf der Grundlage von Informationen über Dateninkompatibilitäten (d. h. Verkodungsfehler, Reliabilitätsprobleme, unterschiedliche Erhebungszeitpunkte etc.). Diese Informationen müssen zusätzlich zu den im Mikrodatenfile und im Identifikationsfile enthaltenen Daten als Input für die Schätzungen verwendet werden.

- Die Wahrscheinlichkeit, daß ein Datensatz des Identifikationsfile und einer des Mikrodatenfile die gleiche Person repräsentieren, wird mittels des Bayesschen Theorems berechnet. So werden bei Kontrolle der Gefahr von Falschzuordnungen trotz Dateninkompatibilitäten Zuordnungen möglich.

- Der Angreifer muß die Wahrscheinlichkeit einer korrekten Zuordnung mit einem Sicherheitsniveau vergleichen, das er nach Maßgabe seiner eigenen Risikoaversion zu wählen hat. Ist die berechnete Wahrscheinlichkeit höher, wird er die Zuordnung eines Falls aus dem Identifikationsfile zu einem Fall aus dem Mikrodatenfile als korrekt betrachten.

3.4.2 Umsetzung der diskriminanzanalytische Reidentifikationstechnik in einen Algorithmus

Die Grundidee der diskriminanzanalytische Reidentifikationstechnik ist, wie im vorangegangenen Abschnitt dargestellt, noch relativ einfach. Die Realisierung in Form eines ausführbaren Algorithmus, der dann in ein Computerprogramm umgesetzt werden kann, ist jedoch ein Problem ganz anderer Art. Die in der Bayesschen Gleichung enthaltenen bedingten und unbedingten Wahrscheinlichkeiten müssen geschätzt werden. Dies ist im vorliegenden Fall ein sehr schwieriges und aufwendiges Unterfangen, das nur in einem mehrstufigen Schätzprozeß gelöst werden kann, bei dem verschiedene statistisch-mathematische Methoden aufeinander aufbauen. Im Unterschied zu vielen statistischen Schätzproblemen geht es nicht um globale Parameter von Verteilungen, vielmehr sind ihre lokalen Eigenschaften wichtig, die aber nur schwer zu ermitteln sind. Eine kurze Zusammenstellung bei Paaß/Wauschkuhn (1985:145ff.) zeigt, daß in dem Algorithmus verschiedenste Verfahren (u.a. Jakobitransformationen, loglineare Modelle, multiple Regressionen, Faktorenanalysen, Clusteranalysen) kombiniert werden.

In den Entwurf des Algorithmus der diskriminanzanalytischen Reidentifikationstechnik floß das gesammelte Know-how der GMD ein und dokumentierte sich eine hohe Gewandtheit in der Verknüpfung verschiedener fortgeschrittener Verfahren der statistischen Analyse. Eine umfangreiche Dokumentation der Reidentifikationstechnik wurde bewußt nicht vorgenommen. Der Projektbericht der GMD (Paaß/Wauschkuhn 1985:125 - 156) enthält nur relativ allgemeine Hinweise, die dem kundigen Statistiker lediglich ein allgemeines Verständnis des Problems ermöglichen. Dies ist darauf zurückzuführen, daß das Verfahren der GMD nur für einen begrenzten Projektzusammenhang konzipiert worden war. Eine Weitergabe des im Rahmen des AIMIPH-Projekts erarbeiteten Computerprogramms war ursprünglich nicht vorgesehen. Außerdem wollte man potentiellen Datenangreifern nicht die Arbeit erleichtern.

Die Rekonstruktion des Algorithmus im Rahmen der eigenen Untersuchung mußte daher überwiegend anhand des Quellprogrammtextes erfolgen, in dem der Algorithmus für die diskriminanzanalytische Reidentifikationstechnik vom AIMIPH-Projekt der GMD umgesetzt worden war. Im folgenden wird ein kurzer Überblick über die Struktur des Algorithmus gegeben, der die grundlegenden Schwierigkeiten und die Ansätze zu ihrer Lösung darlegen soll.

Das gestellte Problem der Reidentifikation läßt sich in zwei Teilprobleme gliedern, nämlich die Schätzung der Reidentifikationswahrscheinlichkeit $p(z|k^*)$ und der Wahrscheinlichkeit $p(z)$, die eine Abschätzung ermöglicht, ob eine Person

einzigartig im Identifikationsfile und in der Grundgesamtheit vorhanden ist. Dabei wird gleich auf das Programm Bezug genommen, das unter Verwendung von SAS und FORTRAN von der GMD erstellt worden war.[18]

3.4.2.1 Das erste Schätzproblem: Die Bestimmung von $p(z|k^*)$

Zu Beginn des Programms erfolgt die datentechnische Zusammenstellung von modifizierten Datenfiles, die in der weiteren Analyse verwendet werden. Das Ausgangsmaterial bilden dabei jeweils ein Mikrodatenfile und ein Identifikationsfile. Diese müssen zuvor hinsichtlich der Überschneidungsmerkmale einen gleichen Aufbau haben und die gleiche Reihenfolge der Überschneidungsmerkmale besitzen.[19]

Zu Beginn eines Reidentifikationsversuches besitzt ein Angreifer eine Datenmatrix, die in der Zeilenzahl durch die Anzahl der Fälle und in der Spaltenzahl durch die Anzahl der Überschneidungsmerkmale begrenzt ist.

Für den Angreifer stellt sich - vor dem eigentlichen Reidentifikationsversuch - zunächst das Problem, daß er die Verteilung von Dateninkompatibilitäten simulieren muß. Im Forschungsbericht der GMD wird in erster Linie auf "echte" (z.B. Verkodungs-) Fehler abgestellt. Im folgenden wird die dabei gewählte Begrifflichkeit beibehalten, indem von "Fehlerprozeß", "Fehlerfile" etc. die Rede ist. Diese Terminologie wird jedoch nur zum Zweck der Abkürzung verwendet. Andere Inkompatibilitäten sind eingeschlossen.

Eine spezifische Subroutine des Programms generiert aus Angaben über die Wahrscheinlichkeit von Dateninkompatibilitäten mittels eines Zufallsprozesses ein mit Inkompatibilitäten behaftetes synthetisches Mikrodatenfile, dessen

---

18 SAS ist ein Statistikprogrammpacket, das in der verwendeten Version 5.x (die in der Zwischenzeit ersetzt worden ist) eine spezielle Matrixmanipulationssprache enthielt, nämlich SAS-Matrix. Diese Sprache ist ein sehr mächtiges Instrument, das die schnelle Programmierung komplexer Methoden erlaubt. SAS-Matrix wurde demgemäß im AIMIPH-Programm extensiv verwendet. Rechenzeitkritische Teile des Programms wurden in FORTRAN programmiert.

19 Während in der ersten Spalte die Identifikationsnummer des betreffenden Datensatzes eingetragen ist, folgen danach die Ausprägungen der einzelnen Überschneidungsmerkmale. Im Rahmen der Datenaufbereitung wird ein weiterer Parameter den einzelnen Datensätzen angefügt. Diese Größe fungiert als Hochrechnungsgewicht und repräsentiert das Verhältnis zwischen der Stichprobengröße und der Grundgesamtheit.

Datensätze gemäß der angenommenen Inkompatibilitätsverteilung veränderte Werte enthalten. Nimmt ein Angreifer bspw. an, daß ein Drittel der Einkommensangaben mit Dateninkompatibilitäten behaftet ist, so macht er entsprechende Eingaben und erzeugt eine Datei, in der ein Drittel der Einkommensangaben verändert wurde. Programmintern werden Differenzen zwischen dem so erzeugten Datensatz, der Dateninkompatibilitäten enthält, und dem originalen Datensatz gebildet. Aus diesen Differenzen werden die Wahrscheinlichkeiten der durch Dateninkompatibilitäten hervorgerufenen Abweichungen geschätzt. Im Ergebnis erhält man also eine Abschätzung, mit welcher Wahrscheinlichkeit Abweichungen zwischen zwei Datensätzen (wie z. B. in Abbildung 3.1/Situation (2)) durch Dateninkompatibilitäten hervorgerufen werden. Das Ergebnis dieses Fehlerprozesses wird durch die Erstellung einer Varianz-Kovarianz-Matrix der Differenzen von Mikrodatenfile und Fehlerfile komprimiert. Paaß/Wauschkuhn sehen es als plausibel an, daß die Kenntnis des Datenangreifers von der heuristischen Verteilung der Dateninkompatibilitäten relativ vollständig ist.

Wie im Abschnitt 3.4.1.2 dargestellt wurde, geht ein Angreifer von quasinormalverteilten Klassen und klassenweise identischen Kovarianzmatrizen aus und berechnet daher für jeden einzelnen Datensatz des Mikrodatenfile die Mahalanobis-Distanz zum gegebenen Fall z. Die errechnete Varianz-Kovarianz-Matrix fließt hierzu als "Gewichtung" der einzelnen Überschneidungsmerkmale ein.

Mittels einer Hauptachsentransformation werden die Klassendichten, die eine ellipsoide Form haben und durch die Mahalanobis-Distanz dargestellt werden, in Kugeln transformiert. Im folgenden hat man es mit euklidischen Abständen zu tun.

Im folgenden wird nicht mehr mit den Originaldaten, sondern mit der Matrix der euklidischen Abstände der y_k zu z gerechnet. Diese einzelnen Abstände der y_k zu z werden dann nach der Größe sortiert, kategorisiert und verdichtet. Die derartig ermittelte Verdichtung (Perzentile) wird dann in Abhängigkeit der letzten Stelle kumuliert.[20] Durch die Aggregation der Perzentile erhält man die Verteilungsfunktion von $p(z|k^*)$. Um Zufallsschwankungen zu eliminieren, werden die Schätzwerte durch ein Polynom 2. Ordnung geglättet.

---

20 Es wird nur ein Wert - der nicht gleich dem vorangegangenen Wert ist - zu der Gesamtsumme addiert. So wird eine Verteilungsfunktion von $p(z|k)$ erzeugt, die nie die Steigung 0 besitzt.

3.4.2.2 Das zweite Schätzproblem: Die Schätzung von p(z)

Da die Schätzung von p(z) um einiges aufwendiger ist als die Schätzung der Dichte $p(z|k^*)$, wird für diese Berechnung ein Großteil des Programmes verwendet. Zur Berechnung von p(z) werden u.a. auch Zwischenergebnisse aus der Schätzung von $p(z|k^*)$ benötigt: Für die Bestimmung von p(z) werden zunächst das Mikrodatenfile und das Identifikationsfile sowie die Varianz-Kovarianz Matrix der Dateninkompatibilitäten zugrunde gelegt, um daraus die euklidischen Abstände der Überschneidungsmerkmale zu berechnen.

Aus der Gesamtheit des Mikrodatenfile wird eine Teilmenge (im Programm 500 Fälle) der y_k mit minimaler euklidischer Distanz ausgewählt.[21] Diese in der Anzahl der Fälle gekürzte Auswahl wird als Referenzmenge für die vorzunehmende Dichteschätzung verwendet. Ähnlich wie in Abschnitt 3.4.2.1 wird ein Fall z des Mikrodatenfile zunächst vorläufig einem Fall y_{k^*} des Identifikationsfile zugeordnet.

Aufgrund der Annahmen über die Verteilung der Dateninkompatibilitäten gilt: Je geringer der euklidische Abstand zwischen y_{k^*} und z ist, desto größer wird die Dichte $p(z|k^*)$. Allerdings ist anzunehmen, daß nur sehr wenige Fälle des Identifikationsfile in der Nähe von z liegen. Es wird daher i.d.R. nicht möglich sein, die Dichte p(z) unmittelbar zu schätzen. Als Lösung wird eine künstliche Verdichtung um den Punkt z mittels Distanzkomponenten vorgenommen.[22]

Diese Distanzkomponenten werden in einem nächsten Schritt ebenfalls aggregiert und in zwei oder drei "grobe" Kategorien eingeteilt.[23] Diese Kategorien sind dann Grundlage für Kreuztabellen. Für alle Paare von Distanzkomponenten werden zweidimensionale Kreuztabellen erstellt und die empiri-

---

21 Die restlichen Datensätze (Anzahl der Fälle im Identifikationsfile - 500) liegen im allgemeinen weit außerhalb der interessierenden Umgebung von y_{k^*} und es wird angenommen, daß diese keine Informationen mehr über die lokalen Eigenschaften von p(z) enthalten.

22 Zu diesen Distanzkomponenten werden diejenigen Distanzen zusammengefaßt, die eine hohe Abhängigkeit in der Umgebung von z aufweisen. So entstehen ca. 10 vorläufige Distanzkomponenten, welche mit Hilfe einer Clusteranalyse bestimmt werden. Hochkorrelierte Merkmale werden auf diese Weise der gleichen Distanzkomponente zugeordnet.

23 Das Kriterium für diese Zusammenlegung ist dabei, ob die Mittelwerte der Distanzkomponenten in der Umgebung von y_{k^*} in etwa gleich groß sind. Um eine möglichst genaue Dichteschätzung für z zu erzielen, umfaßt eine nach oben offene Kategorie alle diejenigen Werte, die wesentlich größer sind.

schen Häufigkeiten des Identifikationsfile bestimmt.[24] Anschließend wird unter Zugrundelegung eines loglinearen Modells die gemeinsame Verteilung der Distanzkomponenten geschätzt. Mittels des Iterative Proportional Fitting-Algorithmus[25] (vgl. Fienberg 1980:32ff.) wird für die grobe Zelleinteilung der Distanzkomponenten die Wahrscheinlichkeit der einzelnen Zellen berechnet.

Für jede Zelle der groben Kategorien wird unterstellt, daß die Distanzkomponenten innerhalb der Zellen statistisch unabhängig sind, somit können die Verteilungen der einzelnen Distanzkomponenten ebenfalls unabhängig voneinander geschätzt werden. Hierzu werden die Distanzkomponenten in etwa 200 "feine" Kategorien gleicher Breite eingeteilt, um anschließend die beobachteten Werte des Identifikationsfile diesen Kategorien zuzuordnen. Damit wird die bedingte Wahrscheinlichkeit der "feineren" Zellen innerhalb der "groben" Zellen unter der Annahme der Unabhängigkeit bestimmt.

Gewichtet nach der geschätzten Wahrscheinlichkeit der einzelnen groben Zellen werden in einem nächsten Schritt die Verteilungsfunktionen von z innerhalb der Zellen zu einer globalen Verteilungsfunktion zusammengefaßt. Hieraus kann schließlich die Dichte p(z) bestimmt werden. Um auch hier Zufallsschwankungen zu glätten, werden die Schätzwerte durch das gleiche Polynom wie bei der Schätzung von $p(z|k^*)$ geglättet, und aus einer Ableitung dieses Polynoms ergibt sich letztlich die Dichte p(z).

Nach Abschluß dieses Prozesses sind alle Variablen bekannt, die für die Verwendung des Bayesschen Ansatzes erforderlich sind. Die Wahrscheinlichkeit der korrekten Zuordnung kann nun mittels Gleichung (7) berechnet und damit das Reidentifikationsrisiko ermittelt werden.

3.4.3 Die diskriminanzanalytische Reidentifikationstechnik und die faktische Anonymität

Die Darstellung des Designs der Reidentifikationstechnik und der Überblick über den Algorithmus ermöglichen im Hinblick auf den Aufwand, der beim Einsatz dieses Verfahrens erforderlich ist, bereits einige Folgerungen für das Problem der faktischen Anonymität. Das AIMIPH-Programm, mit dem die diskriminanzanalytische Reidentifikationstechnik umgesetzt wurde, ist eine Spezial-

---

24 Diese Gruppenbildung erfolgt in einer Weise, daß bei jeder der 2-dimensionalen marginalen Randverteilungen jede Zelle mindestens 5 Elemente enthält.

25 Unter der Annahme, daß die Interaktionen 2. und höherer Ordnung gleich 0 sind.

entwicklung, die nicht allgemein zur Verfügung steht. Ein potentieller Datenangreifer müßte den Prozeß der Entwicklung weitgehend selbst vollziehen. Die hier und in den Veröffentlichungen von Paaß, Paaß/Wauschkuhn angegebenen Schritte ergeben lediglich eine allgemeine Struktur, die nur wenig Entwicklungsaufwand spart.

Die von der GMD verwendete Kombination statistischer Methoden ist äußerst rechenintensiv. Die Experimente der GMD zeigten jedoch, daß auch ein derart aufwendiges Verfahren durchführbar ist. Nur große Erfahrung und modernster Wissensstand auf dem Gebiet der numerischen Methoden erlaubten eine so rationelle Gestaltung des Algorithmus, daß er für praktische Zwecke überhaupt einsetzbar ist.

Das generelle Konzept für das Verfahren der GMD beruht nicht auf einer einzelnen "Idee", mit der das gestellte Problem in besonders eleganter Weise gelöst werden könnte. Ein solcher "Königsweg" für den Datenangreifer existiert nach heutigem Wissen nicht. Stattdessen wird das Problem durch den konzentrierten Einsatz zahlreicher verschiedener Verfahren angegangen, die in intelligenter Weise kombiniert und modifiziert werden.

Die für das Verständnis und die praktische Umsetzung der diskriminanzanalytischen Reidentifikationstechnik notwendigen Fähigkeiten engen den Kreis der potentiellen Nutzer[26] stark ein. Der gewählte Ansatz folgt einem abstrakten Konzept. Er weicht von den in den Sozialwissenschaften üblicherweise gewählten Verfahren grundlegend ab. Zu seinem Verständnis ist eine sehr fortgeschrittene mathematisch-statistische Ausbildung erforderlich. Gründliche Kenntnisse der Wahrscheinlichkeitstheorie, der Schätz-, der Stichprobentheorie, der Diskriminanzanalyse und anderer multivariater Methoden werden benötigt. Das Verständnis der Dichteschätzungen für diskriminanzanalytische Zwecke setzt grundlegende (z. B. partielle Integration, Matrixalgebra) und fortgeschrittene (z. B. Jacobi-Transformation, Transformation in multiplen Integralen) mathematische Kenntnisse voraus. Für theoretische Abschätzungen wurde von Paaß/Wauschkuhn Shannons mathematische Informationstheorie (1976) herangezogen. Darüber hinaus sind sozialwissenschaftliche Kenntnisse für die Abschätzung der Inkompatibilitäten und Kenntnisse aus dem Bereich der Infor-

---

26 Damit wird auch die Verfügbarkeit der diskriminanzanalytische Reidentifikationstechnik als Reidentifikationsmethode eingeschränkt.

matik erforderlich, um den Algorithmus in ein Computerprogramm umsetzen zu können.

Selbst wenn jemand über eine Ausbildung in allen genannten Feldern verfügen würde, erforderte allein die Aneignung der theoretischen Grundlagen der Reidentifikationstechnik mehrere Monate. Der eigentliche Entwicklungsaufwand kann an der Länge des Quellprogrammtextes abgeschätzt werden. Obwohl SAS-Matrix eine außerordentlich mächtige Sprache ist, umfaßt das AIMIPH-Programm 130 Druckseiten mit komplizierten Operationen und Verfahren.

Da für die vorliegende Untersuchung das Programm sowie die Hilfe von Herrn Paaß zur Verfügung stand, brauchte der Entwicklungsaufwand nicht selbst geleistet werden. Um das Programm jedoch auf dem vorhandenen Rechner implementieren zu können, waren Änderungen erforderlich, die ein volles Verständnis des Algorithmus verlangten. Für die Aneignung der statistischen Theorie und die Rekonstruktion des Algorithmus inklusive einer Erstellung von Datenflußplänen waren ca. 7 Monate Arbeitszeit wissenschaftlich ausgebildeter Kräfte erforderlich.[27] Auch daran wird die Komplexität der Reidentifikationstechnik deutlich.

Der zu ihrer Entwicklung zu leistende Aufwand ist so hoch, daß dafür ein mehrjähriges Projekt einkalkuliert werden müßte. Aus diesem Grund kann davon ausgegangen werden, daß die diskriminanzanalytische Reidentifikationstechnik im sozialwissenschaftlichen Szenario bzgl. faktischer Anonymität darstellt. Die diskriminanzanalytische Reidentifikationstechnik ist schwierig zu realisieren. Es entstehen so hohe Kosten, daß sich ihre Verwendung zur Informationsbeschaffung nicht lohnt.

Die Durchführung von Experimenten mit der diskriminanzanalytischen Reidentifikationstechnik im Rahmen dieser Untersuchung hat allein den Sinn, die effektivste verfügbare Reidentifikationstechnik zu testen, um eine obere Abschätzung des Gefahrenpotentials zu erhalten, und nicht, weil eine reale Gefahr mit der diskriminanzanalytischen Reidentifikationstechnik verbunden wäre. Die Anwendung des Verfahrens auf Daten aus dem Mikrozensus und "Kürschers Deutschen Gelehrtenkalender" und die dabei erzielten Ergebnisse sind in Abschnitt 9.2 dargestellt.

---

27 Von den Mitarbeitern dieser Untersuchung wurde zusätzlich die Beratung von Experten für die Klärung von Spezialproblemen gesucht.

4 Bestimmungsfaktoren des Reidentifikationsrisikos*

Wie Reidentifikationsversuche angelegt sein können und welche Voraussetzungen sie haben, sollte nach der Erörterung der verschiedenen methodischen Verfahren deutlich geworden sein. Damit sind die Grundlagen geschaffen, um in diesem Kapitel den Versuch zu unternehmen, in einer möglichst umfassenden Systematik und im Hinblick auf unterschiedliche denkbare Angriffsszenarien die Faktoren zu untersuchen, von denen die Chancen (bzw. das Risiko) einer erfolgreichen Reidentifikation abhängen. Dabei soll das potentielle Gewicht der einzelnen Risikofaktoren sowohl durch theoretisch-statistische Überlegungen und Modellierungen bestimmt werden, wie durch die Darstellung empirischer Befunde, die bislang jedoch nur in Teilaspekten vorliegen. Eine Aufgabe dieses Kapitels ist es auch, die Grundlagen zu legen, auf deren Basis dann die Anlage der eigenen empirischen Untersuchungen abgeleitet und ihre Ergebnisse interpretiert werden können.

In der Konzeptualisierung des Problems greifen wir Vorschläge einer Gruppe von Wissenschaftlern auf, die im Zusammenhang der Abklärungen der Reidentifikationsrisiken für Mikrodaten aus der Volkszählung 1991 in England den bislang überzeugendsten Ansatz zur Systematisierung der das Reidentifikationsrisiko bestimmenden Größen vorgelegt haben (Skinner et al. 1990, Marsh et al. 1991). Zu einzelnen Aspekten hat diese Wissenschaftlergruppe Vorstellungen entwickelt (z.B. zum Problem der Kompatibilität von Informationen des Mikrodatenfile und des Zusatzwissens), die unabhängig von diesen Arbeiten bei den empirischen Untersuchungen des Anonymisierungsprojektes überprüft wurden. Mit einigen weitergehenden Überlegungen kann dieses Konzept die wichtigsten denkbaren Angriffsszenarien und Datenkonstellationen in einer einheitlichen Terminologie behandeln und auf vergleichsweise einfache Weise aufzeigen, wie sich die für die Risikoeinschätzung bedeutsamen Faktoren in unterschiedlichen Angriffsszenarien verändern.

Als Ausgangspunkt bietet sich eine Datenangriffssituation an, der auch in der öffentlichen Diskussion der Gefährdungssituationen eine besondere Rolle zukommt. Andere Angriffsszenarien können dann im Vergleich dazu charakterisiert werden. Das Grundszenario ist das Modell der *gezielten Suche*. Es existiert ein Mikrodatenfile, in dem für alle N Fälle einer Population oder für eine

---

* Autoren: Walter Müller (4.1, 4.2)
Uwe Blien, Walter Müller (4.3)

Stichprobe von n Fällen dieser Population je ein Datensatz mit den Messungen (Ausprägungen) für eine bestimmte Anzahl von Variablen (Merkmalen) enthalten ist. Ein Datenangreifer besitzt über einen oder mehrere Fälle der Population durch Zusatzwissen Informationen zu einzelnen Merkmalen, die als Überschneidungsmerkmale auch in den Datensätzen des Mikrodatenfile vorhanden sind (Identifikationsfile). Er möchte diese Fälle im Mikrodatenfile auffinden, um sich Kenntnis über die weiteren dort enthaltenen Informationen zu verschaffen. Der Datenangreifer weiß aber nicht ob die Datensätze für die ihn interessierenden Fälle überhaupt im Mikrodatenfile enthalten sind oder nicht. Die Deanonymisierungstechnik ist ein einfacher Abgleich der Überschneidungsmerkmale auf Identität der Merkmalsausprägungen. Im Vergleich zu dieser Situation werden dann später die Unterschiede in den Bedingungen von Einzel- oder Massenfischzugsszenarien dargestellt. Auch für den Sonderfall, bei dem der Datenangreifer weiß, daß sich der Datensatz einer gesuchten Person im Mikrodatenfile befindet (*response knowledge*), kann gezeigt werden, wie sich das Reidentifikationsrisiko im Vergleich zum gewählten Grundszenario verändert. Im Vergleich zum Grundszenario kann ebenfalls verdeutlicht werden, zu welchen Problemen komplexere Reidentifikationstechniken (z.B. die Methode von Paaß/Wauschkuhn) beanspruchen, eine Lösung gefunden zu haben.

Für den Fall der gezielten Suche kann man mit Skinner et al. (1990) und Marsh et al. (1991) das Reidentifikationsrisiko systematisch bestimmen, wenn man von den folgenden vier Bedingungen ausgeht, die erfüllt sein müssen, damit die Reidentifikation eines in einem Mikrodatenfile gesuchten Datensatzes als erfolgreich vorgenommen gelten kann:

A. Der Datensatz des zu identifizierenden Falles muß im Mikrodatenfile enthalten sein (Repräsentations- oder Selektivitätsproblem).

B. Die Überschneidungsmerkmale müssen im Identifikationsfile und im Mikrodatenfile in identischer Weise abgebildet sein (Kompatibilitätsproblem).

C. Der gesuchte zu identifizierende Fall muß im Hinblick auf die Ausprägungen der Überschneidungsmerkmale in der Population einmalig sein, das heißt: Es darf in der Population kein anderer Fall mit einer identischen Ausprägungskombination der Überschneidungsmerkmale existieren (Problem der Populationseinzigartigkeit, Uniqueness).

D. Der Angreifer muß *wissen*, daß die Ausprägungskombination in der Population einmalig ist. Die Tatsache allein, daß der zu identifizierende

Fall in der Population eine einmalige Ausprägungskombination aufweist, reicht nicht aus (Sicherheitsproblem).

Im folgenden wird zunächst kurz gezeigt, daß diese Bedingungen tatsächlich eine überzeugende Konzeptualisierung und Operationalisierung des Reidentifikationsrisikos in verschiedenen Szenarien ermöglichen. Danach werden sie im einzelnen Zug um Zug gründlicher diskutiert und Quantifizierungen ihrer Bedeutung für das Reidentifikationsrisiko versucht.

4.1 Ein allgemeines Konzept für das Reidentifikationsrisiko

4.1.1 Das Szenario der gezielten Suche

Daß die angeführten Komponenten A bis D eine sehr allgemeine, überzeugende Konzeptualisierung des Reidentifikationsrisikos ermöglichen, ergibt sich aus den folgenden Überlegungen.

A. Das Risiko, daß über einen gesuchten Fall Informationen aus einem Mikrodatenfile gewonnen werden können, besteht in der Tat nur dann, wenn der Datensatz dieses Falls im Mikrodatenfile vorhanden ist. Die erste Komponente des Reidentifikationsrisikos ergibt sich deshalb aus der Wahrscheinlichkeit, daß ein gesuchter Fall im Mikrodatenfile enthalten ist. Diese Wahrscheinlichkeit ist durch den Auswahlprozeß bestimmt, nach dem Fälle aus einer Grundgesamtheit in das Mikrodatenfile aufgenommen werden. Wenn es sich beim Mikrodatenfile um eine Totalerhebung handelt, ist diese Wahrscheinlichkeit gleich 1. Einfach läßt sich die Wahrscheinlichkeit ebenfalls bestimmen, wenn die Auswahl durch Verfahren der zufälligen Stichprobenziehung erfolgt. Dann entspricht sie dem Auswahlsatz der Stichprobe bzw. dem Repräsentationsverhältnis n/N der Anzahl Fälle n des Mikrodatenfile zur Anzahl Fälle N der Grundgesamtheit. Bei anderen Verfahren der Auswahl kann es zu - in der Regel empirisch nur schwer bestimmbaren - mehr oder weniger großen Abweichungen von diesem Repräsentationsverhältnis kommen. Eine besondere Rolle spielt dabei, ob die Wahrscheinlichkeit, in die Auswahl aufgenommen zu werden, positiv oder negativ mit der Ausprägung eines Merkmals des gesuchten Falles korreliert. In der Regel dürfte jedoch das Verhältnis n/N eine brauchbare Annäherung für die Abschätzung der ersten Risikokomponente darstellen.

Konkret bedeutet dies, daß im Beispiel des Mikrozensus in aller Regel nur einer von hundert Fällen, über den Zusatzwissen besteht, auch im Mikrodatenfile des Mikrozensus enthalten sein wird und dort durch einen Angreifer gefunden werden kann. Bei der Einkommens- und Verbrauchsstichprobe (EVS) mit einem Auswahlsatz von ca. 0,2 Prozent gilt dieses für einen von 500 Fällen. Die Stichprobeneigenschaft einer Erhebung stellt damit bereits eine erhebliche Eingrenzung des Risikos dar, daß über eine beliebige Person einer Grundgesamtheit Informationen in einem Mikrodatenfile gefunden werden können. Je kleiner der Auswahlsatz der Stichprobe ist, umso kleiner ist das entsprechende Risiko (Hamacher 1980:224).

B. Wie die Diskussion der einfachen Abgleichsverfahren gezeigt hat, wird das Risiko, daß ein Fall, über den Zusatzwissen besteht und der einem Fall im Mikrodatenfile zugeordnet werden soll, entscheidend davon bestimmt, ob dieser Fall in vollständig kompatibler Weise im Identifikationsfile und im Mikrodatenfile abgebildet ist. Ist diese Bedingung verletzt, kann durch diese Verfahren keine korrekte Zuordnung vorgenommen werden. Die damit verbundene Komponente des Reidentifikationsrisikos entspricht demnach der Wahrscheinlichkeit, mit der Fälle im Mikrodatenfile und im Identifikationsfile vollständig kompatibel abgebildet sind.

C. Im Hinblick auf den Faktor der Einzigartigkeit von Merkmalsausprägungen geht die bisherige Diskussion richtigerweise davon aus, daß eine Voraussetzung für die Reidentifizierbarkeit von Datensätzen die Einzigartigkeit im *Mikrodatenfile* ist. Nur in einem speziellen Fall, auf den wir später eingehen, ist dieses jedoch eine hinreichende Bedingung. Grundlegender ist die Einmaligkeit der Ausprägungskombination in der *Population*. Wann immer Populationseinmaligkeit besteht, kann - kompatible Abbildung vorausgesetzt - eine gegebene Ausprägungskombination auch in der Population entweder nur einmalig oder im Falle einer Stichprobe unter Umständen überhaupt nicht besetzt sein. Wann immer in der Grundgesamtheit aber *mehrere* Fälle eine identische Kombination in den Ausprägungen der Überschneidungsmerkmale aufweisen, ist nicht entscheidbar, von welchem dieser Fälle ein im Mikrodatenfile enthaltener Datensatz mit dieser Ausprägungskombination stammt, selbst wenn diese Ausprägungkombination im Mikrodatenfile eindeutig ist. Eine Reidentifikation ist mit den gegebenen Informationen nicht eindeutig möglich. Voraussetzung für die Möglichkeit einer eindeutigen Reidentifikation ist also die Populations-

einzigartigkeit und nicht die Einzigartigkeit einer Ausprägungskombination im Mikrodatenfile.

Entscheidend für die Bestimmung des Reidentifikationsrisikos ist es deshalb zu klären, welche Anteile einmaliger Kombinationen in der Grundgesamtheit bei solchen Merkmalen und Merkmalsausprägungen bestehen, die man im Mikrodatenfile und im Identifikationsfile als vorhanden und als möglichst kompatibel abgebildet unterstellen kann.

D. Wann immer eine Ausprägungskombination von Merkmalen in der Grundgesamtheit einmalig ist, wird sie auch im Mikrodatenfile einmalig sein, vorausgesetzt, Bedingung A (Repräsentation) und Bedingung B (Kompatibilität) sind erfüllt. Der Umkehrschluß ist aber keineswegs allgemein gültig. Wenn ein Datenangreifer im Mikrodatenfile eine einzigartige Ausprägungskombination findet, kann er im Szenario einer gezielten Suche noch nicht per se von einer gelungenen Reidentifikation ausgehen. Selbst wenn faktisch Populationseinzigartigkeit bestehen sollte, verfügt ein Datenangreifer in der Regel nicht über die Kenntnisse, die ihn dieses mit Sicherheit wissen läßt. Entweder benötigt er vollständiges Wissen über die Grundgesamtheit oder er kann nur mit bestimmten Wahrscheinlichkeiten auf Einzigartigkeit in der Population schließen. Grundlage für entsprechende Wahrscheinlichkeitsschätzungen kann sein, ob die interessierende Ausprägungskombination im Mikrodatenfile einzigartig ist und wie groß der Anteil einzigartiger Ausprägungskombinationen im Mikrodatenfile insgesamt ist. Als weitere Komponente des Reidentifikationsrisikos kommt deshalb die Wahrscheinlichkeit hinzu, mit der richtig geschlossen werden kann, daß die interessierende Ausprägungskombination in der Grundgesamtheit einmalig ist.

Die Abschätzung dieser Gewißheit kann theoretisch erfolgen. Inzwischen gibt es jedoch ebenfalls erste empirische Untersuchungen, die Anhaltspunkte für diese Risikokomponente geben. Auf die Einzelheiten wird später eingegangen. Allgemein kann das Vorhandensein statistischer Doppelgänger umso eher ausgeschlossen und eine Reidentifikation mit umso größerer Sicherheit vorgenommen werden, *je vollständiger entweder das Mikrodatenfile oder das Identifikationsfile die in der Grundgesamtheit existierenden Fälle enthält.* Das Sicherheitsproblem ist gelöst, wenn eines dieser beiden Files alle Fälle der Grundgesamtheit enthält, also eine Totalerhebung darstellt. Unabhängig davon, ob das Mikrodatenfile oder das Identifikationsfile alle Fälle der Grundgesamtheit enthält, kann nämlich aus den Daten des File mit der Gesamterhebung festgestellt werden, ob

es für eine bestimmte Ausprägungskombination in der Grundgesamtheit statistische Doppelgänger gibt oder nicht.

Das Risiko, daß ein gesuchter Datensatz reidentifiziert werden kann, ergibt sich nun - wie Skinner et al. (1990) nachvollziehbar zeigen - aus dem Produkt der Wahrscheinlichkeiten, die für jeden einzelnen der Faktoren A, B, C und D ermittelt werden können, wobei zu berücksichtigen ist, inwiefern die Wahrscheinlichkeiten der einzelnen Faktoren voneinander abhängig sind. Da die Wahrscheinlichkeit des Faktors A - Auswahl eines Datensatzes für die Aufnahme im Mikrodatenfile von keinem der anderen Faktoren abhängig sein dürfte, ergibt sich die Wahrscheinlichkeit einer Reidentifikation W_R damit als Produkt der folgenden Reihe konditionierter Einzelwahrscheinlichkeiten:

$$(1)\ W_R = W_A \cdot W_{B|A} \cdot W_{C|A,B} \cdot W_{D|A,B,C}$$

Es ist plausibel anzunehmen, daß die Wahrscheinlichkeit des Faktors B nicht von A und auch die Wahrscheinlichkeit des Faktors C nicht von A abhängen.[2] Die Sicherheit (Wahrscheinlichkeit) mit der ein Datenangreifer Populationseinzigartigkeit annehmen kann (W_D), ist jedoch statistisch abhängig von den Gegebenheiten bei den Faktoren A, B und C. Man muß wohl auch davon ausgehen, daß es eine Beziehung geben kann zwischen der Kompatibilität der Abbildung eines Datensatzes und der Wahrscheinlichkeit, daß dieser Datensatz eine einzigartige Merkmalskonstellation aufweist.[3] Unter diesen Abhängigkeitsannahmen verändert sich dann die Gleichung (1) zu

$$(2a)\ W_R = W_A \cdot W_B \cdot W_{C|B} \cdot W_{D|A,B,C}$$

bzw. zu

$$(2b)\ W_R = W(R_{MDF}) \cdot W(K) \cdot W(U_P) \cdot S(U_P),$$

---

2 Der Grad der Kompatibilität, mit dem die Überschneidungsmerkmale im Mikrodatenfile und im Identifikationsfile abgebildet sind, ist unabhängig davon, ob ein Fall im Mikrodatenfile enthalten ist. Ob der Fall in der Grundgesamtheit eine einmalige Ausprägungskombination aufweist, ist in der Regel auch unabhängig davon, ob ein Fall im Mikrodatenfile enthalten ist, denn zumindest bei Zufallsstichproben erfolgt die Stichprobenziehung unabhängig von bestimmten Merkmalen eines Falles.

3 Es gibt beispielsweise Hinweise aus der Reliabilitätsforschung, daß bei Merkmalen, die ein sehr differenziertes Kategorienschema aufweisen (und damit eher zu Einzelfällen führen), tendenziell mit einer verringerten Meßreliabilität gerechnet werden muß.

wobei

W_A = $W(R_{MDF})$ = Wahrscheinlichkeit, daß der gesuchte Fall im Mikrodatenfile enthalten (repräsentiert) ist,

W_B = $W(K)$ = Wahrscheinlichkeit, daß alle Überschneidungsmerkmale im Mikrodatenfile und im Identifikationsfile vollständig kompatibel abgebildet sind,

$W_{C|B}$ = $W(U_P)$ = Wahrscheinlichkeit, daß die Ausprägungskombination in der Population einzigartig ist (Uniqueness in Population), bei gegebener kompatibler Abbildung

$W_{D|A,B,C}$ = $S(U_P)$ = Sicherheit (Wahrscheinlichkeit), mit der der Datenangreifer Populationseinzigartigkeit annehmen kann, gegeben daß die Bedingungen A, B und C erfüllt sind.

Haben wir somit - ausgehend vom Szenario der gezielten Suche - das allgemeine Modell für die Bestimmung des Reidentifikationsrisikos formuliert, können wir nun erörtern, wie sich andere Szenarien im Vergleich zu diesem Modell bestimmen lassen.

4.1.2 Das Reidentifikationsrisiko in Fischzugsszenarien und bei Wissen um die Zugehörigkeit zum Mikrodatenfile

Fischzugsszenarien unterscheiden sich vom Szenario der gezielten Suche dadurch, daß eine Reidentifikation nicht für einen bestimmten Fall angestrebt wird, sondern der Angreifer sich damit zufrieden gibt, einen beliebigen Fall (Einzelfischzug) oder beliebige Fälle (Massenfischzug) zu reidentifizieren.[4] In dieser Situation ist zu erwarten, daß *der Angreifer von einzigartigen Fällen des Mikrodatenfile ausgeht und sie durch Abgleich mit dem Identifikationsfile zu*

---

[4] Als Motivation kann beispielsweise unterstellt werden, daß ein Angreifer daran interessiert ist,
- prinzipiell die Machbarkeit einer Deanonymisierung nachzuweisen;
- ein statistisches Amt durch den "Nachweis" mangelnden Datenschutzes zu diskreditieren;
- durch Deanonymisierung eines Datensatzes für eine prominente Persönlichkeit öffentliche Aufmerksamkeit zu provozieren;
- Adressen für Zielgruppen mit bestimmten im Mikrodatenfile erfaßten Eigenschaften zu gewinnen.

identifizieren sucht. In welcher Weise sind bei einem solchen Angriff die Bedingungen gegenüber dem dargestellten allgemeinen Modell verändert?

Die einzige Besonderheit für die Risikoberechnung besteht darin, daß sich im Hinblick auf die Komponente A das Repräsentationsproblem umkehrt: Der aus dem Mikrodatenfile mit dem Ziel der Identifikation ausgewählte Fall muß im Identifikationsfile enthalten sein. An die Stelle der Wahrscheinlichkeit, im Mikrodatenfile enthalten zu sein, tritt die Wahrscheinlichkeit, im Identifikationsfile enthalten zu sein. Alle anderen Komponenten bleiben unverändert. Eine Reidentifikation ist auch bei einem Fischzugsszenario nur möglich, wenn Datensätze im Mikrodatenfile und im Identifikationsfile vollständig kompatibel abgebildet sind und wenn die gesuchte Ausprägungskombination in der Grundgesamtheit einmalig ist. Eine Reidentifikation ist außerdem abhängig vom Grad der Gewißheit, mit dem Populationseinzigartigkeit angenommen werden kann.

Bei der Komponente A bestehen im Fischzugsszenario insofern Unterschiede zum Szenario der gezielten Suche, als das Identifikationsfile wohl nur in den seltensten Fällen eine Zufallsauswahl aus der Grundgesamtheit darstellt. Die Wahrscheinlichkeit, daß ein Datensatz im Identifikationsfile enthalten ist, wird sehr oft nicht genau bestimmt werden können. Skinner et al. (1990) weisen jedoch mit Recht darauf hin, daß ein Angreifer in diesem Szenario u.U. die Möglichkeit hat, ein Identifikationsfile zu benutzen, bei dem eine hohe Wahrscheinlichkeit besteht, daß die aus dem Mikrodatenfile zur Reidentifikation ausgewählten Fälle im Identifikationsfile enthalten sind. Als Identifikationsfile könnte beispielsweise ein Register genutzt werden, das für die Ausprägung eines Merkmals mehr oder weniger vollständig alle Fälle enthält, die diese Ausprägung aufweisen. Wenn der Datensatz, dessen Identifikation angestrebt wird, die Ausprägung "Arzt" aufweist, dann würde ein Ärzteverzeichnis ein Identifikationsfile bilden, in dem der zu identifizierende Fall mit großer Wahrscheinlichkeit enthalten ist; wenn es sich um einen Automobilbesitzer handelt, dann das Register der Fahrzeughalter, etc.. Da in solchen Situationen die Komponente A mit einem Faktor nahe eins angesetzt werden kann, sind ceteris paribus Fischzugsszenarien mit einem größeren Reidentifikationsrisiko verbunden als eine gezielte Suche.

Auch das Szenario, in dem neben den Annahmen der gezielten Suche unterstellt wird, daß ein Datenangreifer Kenntnis davon hat, daß der gesuchte Fall im Mikrodatenfile vorhanden ist, beispielsweise aufgrund der Kenntnis der Teilnahme an einer Erhebung (*response knowledge*), kann als Spezialfall des allgemeinen Modells formuliert werden. Bei diesem Szenario kann man unterstellen, daß Bedingung A erfüllt, d.h. das Repräsentationsproblem gelöst ist.

Der Datenangreifer weiß ja, daß ein gesuchter Fall im Mikrodatenfile enthalten ist. Darüber hinaus kann in diesem Szenario die Wahrscheinlichkeit der Populationseinzigartigkeit (Bedingung C) ersetzt werden durch die Wahrscheinlichkeit einer einzigartigen Ausprägungskombination im Mikrodatenfile. Wird eine einzigartige Ausprägungskombination im Mikrodatenfile gefunden, die mit der Kombination des gesuchten Falls übereinstimmt, dann kann - vollständige kompatible Abbildung (Bedingung B) vorausgesetzt - ein Angreifer davon ausgehen, daß es sich um den von ihm gesuchten Fall handelt, unabhängig davon, ob in der Grundgesamtheit noch weitere Fälle mit gleichen Ausprägungskombinationen existieren oder nicht. Darüber hinaus wird in dieser Konstellation Bedingung D obsolet: Wenn man weiß, daß ein Fall mit einer bestimmten Ausprägungskombination im Mikrodatenfile enthalten ist, und diese Kombination dort einmalig ist, dann muß es sich um den gesuchten Fall handeln. Durch Kenntnis der Zugehörigkeit zum Mikrodatenfile werden also gleichzeitig die Bedingungen A und D erfüllt, wenn die Ausprägungskombination im Mikrodatenfile einmalig ist.

Als weitere Besonderheit der Situation mit response knowledge kommt hinzu, daß bei einem gegebenen Differenzierungsgrad der Überschneidungsmerkmale das Reidentifikationsrisiko umso größer ist, je kleiner der Umfang des Mikrodatenfile ist. Je kleiner der Umfang des Mikrodatenfile nämlich ist, umso eher ist bei gegebener Anzahl und Art der Ausprägungen der Überschneidungsmerkmale damit zu rechnen, daß ihre Kombination im Mikrodatenfile nur einmalig besetzt ist und sich damit eine Reidentifikation vornehmen läßt.

Auch ein Vergleich der Leistungen der einfachen Abgleichtechniken mit der diskriminanzanalytischen Reidentifikationstechnik von Paaß/Wauschkuhn läßt sich im Rahmen des allgemeinen Modells recht einfach deutlich machen. Der Ansatz von Paaß/Wauschkuhn trägt zumindest seinem Anspruch nach entscheidend dazu bei, das Kompatibilitätsproblem zu entschärfen und er trägt zu einer rationalen Lösung der Komponente D bei.

4.1.3 Das Reidentifikationsrisiko bei verschiedenen Reidentifikationsstrategien und Informationsgrundlagen

Um die ausgeführten Sachverhalte nochmals zusammenfassend zu verdeutlichen, sind in Übersicht 4.1 die in den verschiedenen Szenarien wirksamen Komponenten des Reidentifikationsrisikos zusammengestellt. Die in der Übersicht eingefügten Pfeile verdeutlichen, in welcher Richtung die Suche beim

Identifikationsvorgang verläuft. In dieser Übersicht wird zusätzlich die Bedeutung des Umfangs berücksichtigt, zu dem eine Population entweder im Mikrodatenfile oder im Identifikationsfile repräsentiert ist. Der wesentliche Unterschied besteht darin, ob es sich um eine Totalerfassung oder nur um eine Teilerfassung handelt.

Übersicht 4.1: Komponenten des Reidentifikationsrisikos

Identifikationsstrategie	Repräsentation der Population		Komponenten des Reidentifikationsrisikos				
	IF	MDF	Repräsentation	Kompatibilität		Uniqueness	Sicherheit
gezielte Suche	Teile →	Teile	$W(R_{MDF})$	* W(K)	*	$W(U_P)$	* $S(U_P)$
gezielte Suche	Teile →	Total	1	* W(K)	*	$W(U_P)$	* 1
gezielte Suche	Total →	Teile	$W(R_{MDF})$	* W(K)	*	$W(U_P)$	* 1
Fischzug	Teile ←	Teile	$W(R_{IF})$	* W(K)	*	$W(U_P)$	* $S(U_P)$
Fischzug	Teile ←	Total	$W(R_{IF})$	* W(K)	*	$W(U_P)$	* 1
Fischzug	Total ←	Teile	1	* W(K)	*	$W(U_P)$	* 1
gezielte Suche mit response knowledge	Teile →	Teile	1	* W(K)	*	$W(U_{MDF})$	* 1

wobei

$W(R_{IF})$ = Wahrscheinlichkeit, daß der gesuchte Fall im Identifikationsfile enthalten ist;

$W(U_{MDF})$ = Wahrscheinlichkeit, daß die Ausprägungskombination des gesuchten Falles im Mikrodatenfile einzigartig ist. Übrige Definitionen wie oben.

Unter der Voraussetzung, daß es sich in beiden Datenfiles um Teilerhebungen handelt, sind alle Komponenten kleiner als eins, wobei es sich bei einzelnen Komponenten um sehr kleine Werte handeln kann, so daß ihre multiplikative Verknüpfung, dann in aller Regel ebenfalls einen sehr kleinen Wert liefert.

Immer dann, wenn ein Angreifer bei einem der Datenfiles auf eine Totalerfassung der Population zurückgreifen kann, erhöht sich das Reidentifikationsrisiko dadurch erheblich, daß ein Angreifer durch Inspektion seiner Daten vollständige Sicherheit darüber gewinnen kann, ob in der Population eine bestimmte Ausprägungskombination einzigartig ist oder nicht. Je nach Such-

strategie kann bei einer Totalerhebung auch mit Sicherheit damit gerechnet werden, daß der gesuchte Fall in dem durchsuchten File enthalten ist, so daß auch die in der Repräsentationsspalte aufgeführte Komponente des Reidentifikationsrisikos als 1 angenommen werden kann. Dies gilt beispielsweise für die in Zeile 2 charakterisierte Situation, in der das Mikrodatenfile als Vollerhebung (Beispiel: Volkszählung) angenommen ist. Ohne besondere Schutzvorkehrungen sind Mikrodaten einer Volkszählung deshalb besonders reidentifikationsgefährdet, weil die in mehrerer Hinsicht relevante Schutzwirkung der üblicherweise vorhandenen Stichprobeneigenschaft des Mikrodatenfile wegfällt.

Aus der Übersicht wird auch deutlich, daß Fischzugsszenarien vor allem dann als mit einem höherem Risiko verbunden einzuschätzen sind, wenn der Datenangreifer auf ein Identifikationsfile zurückgreifen kann, das für ein prägendes Merkmal des gesuchten Falls (annähernd) eine Totalerhebung darstellt.

Schließlich erweist sich der in der letzten Zeile der Tabelle aufgeführte Sonderfall (response knowledge) als besonders riskant, weil neben der Kompatibilitätskomponente nur die Wahrscheinlichkeit einer einzigartigen Ausprägungskombination im Mikrodatenfile ins Kalkül einzugehen hat. Durch das Vorwissen darüber, daß der gesuchte Fall im Mikrodatenfile enthalten ist, werden die Repräsentations- und die Sicherheitskomponente zu eins, und Einzigartigkeit in der Ausprägungskombination muß nur im Mikrodatenfile gegeben sein.

In allen Szenarien ist die Voraussetzung der Kompatibilität der Abbildung der Fälle im Mikrodatenfile und im Identifikationsfile für die Schutzwirkung bedeutsam, auch bei Wissen um Mikrodatenfilezugehörigkeit. Außerdem muß - mit Ausnahme des Sonderfalls 'response knowledge'- immer die Bedingung der Populationseinzigartigkeit erfüllt sein. Die Möglichkeit einer Reidentifikation ist deshalb in aller Regel umso unwahrscheinlicher, je größer die Population ist, auf die bezogen ein Reidentifikationsversuch unternommen wird.

Aus diesen grundsätzlichen Überlegungen läßt sich bereits in groben Zügen ableiten,

- welche Situationen als besonders riskant zu betrachten sind;

- wie diesbezüglich die faktische Anonymität der in dieser Untersuchung im Vordergrund stehenden Mikrodaten des Mikrozensus und der EVS einzuordnen ist;

- an welcher Stelle Maßnahmen zur Verringerung des Reidentifikationsrisikos und damit zur Sicherung der faktischen Anonymität anzusetzen haben.

Die vorgetragenen Überlegungen sind jedoch auch im Hinblick darauf wichtig, welche Komponenten des Reidentifikationsrisikos bei den weiteren Untersuchungen in besonderer Weise zu berücksichtigen sind.

Als besonders riskant erweisen sich demnach Situationen, in denen entweder das Mikrodatenfile oder das Identifikationsfile eine Totalerhebung oder annähernd eine Totalerhebung darstellen. Bei den Mikrodatenfiles von Mikrozensus und EVS ist dies nicht der Fall. Der Mikrozensus ist eine Stichprobe von 1 Prozent der Bevölkerung und die EVS eine Stichprobe von ca. 0,2 Prozent der Bevölkerung. Von dieser Stichprobeneigenschaft gehen deshalb schon ganz erhebliche Schutzwirkungen aus. Genauer zu untersuchen ist jedoch, ob Zusatzwissen existiert und im Bereich der Wissenschaft zugänglich ist, aus dem Identifikationsfiles gebildet werden können, die eine Totalerhebung darstellen oder einer Totalerhebung nahekommen. Diese Frage wird - neben anderen Aspekten - in dem gesonderten Kapitel 6 zum Zusatzwissen untersucht werden.

Eine starke Erhöhung des Reidentifikationsrisikos erbringt auch das Wissen eines Datenangreifers darüber, ob ein Fall im Mikrodatenfile enthalten ist. Dieses Wissen kann bei einem Datenangreifer existieren, aber bei Mikrozensus und EVS trägt wiederum die Stichprobeneigenschaft dazu bei - im Unterschied etwa zu einer Volkszählung -, daß es unwahrscheinlich ist, daß ein Sozialwissenschaftler über solches Wissen verfügt und insbesondere, daß er von vielen Fällen Kenntnis über die Stichprobenzugehörigkeit hat. Dennoch wird diese Situation bei der Entwicklung von Anonymisierungsmaßnahmen besonders zu beachten sein (vgl. Kapitel 11).

Neben den eher außergewöhnlichen Risikofaktoren - der Vollerhebung und des Wissens um Stichprobenzugehörigkeit - dürfen jedoch die in praktisch allen Szenarien wirksamen Bedingungen der Populationseinzigartigkeit und der Kompatibilität der Informationen in Mikrodatenfile und Identifikationsfile nicht übersehen werden. Auf diese Aspekte soll in den beiden folgenden Abschnitten dieses Kapitels näher eingegangen werden. In Abschnitt 4.2 behandeln wir die mit der Voraussetzung der Populationseinzigartigkeit zusammenhängenden Fragen. Dabei untersuchen wir zunächst die Frage, welche Faktoren zu faktischer Einzigartigkeit von Ausprägungskombinationen in der Population beitragen, und dann die Frage, wie von Einzigartigkeit im Mikrodatenfile auf Ein-

zigartigkeit in der Population geschlossen werden kann. Abschnitt 4.3 ist schließlich einer ausführlichen Analyse der Kompatibilitätskomponente gewidmet, die in der bisherigen Diskussion nur unzureichend berücksichtigt wurde.

4.2 Einzigartigkeit als Schlüssel zur Reidentifikation

Der Schlüssel, der einen anonymisierten Mikrodatensatz zur Reidentifikation aufschließen kann, ist die Einzigartigkeit einer Kombination der Ausprägungen der Überschneidungsmerkmale, die sowohl auf einen Datensatz im Mikrodatenfile wie auf einen Datensatz im Identifikationsfile paßt. Mit Ausnahme des Sonderfalls 'response knowledge' muß diese Einzigartigkeit in der Population gegeben sein, auf die sich das Mikrodatenfile bezieht. Die Einzigartigkeit im Mikrodatenfile reicht nicht aus, es sei denn das Mikrodatenfile stellt selbst eine Vollerhebung der Population dar. Für die Bestimmung des Reidentifikationsrisikos ergeben sich daraus, wie oben dargestellt, vor allem die folgenden Fragen: Erstens: Mit welcher Wahrscheinlichkeit sind einzigartige Ausprägungskombinationen in der Population zu erwarten, sowie zweitens: Mit welcher Sicherheit kann ein Datenangreifer davon ausgehen, daß eine *bestimmte* Ausprägungskombination in der Population einzigartig ist?

Dieser Abschnitt dient einer näheren Klärung dieser Fragen. Wir untersuchen zunächst die wichtigsten Faktoren, durch die die Wahrscheinlichkeit einzigartiger Ausprägungskombinationen bestimmt ist. Dann beschreiben wir kurz einige Ergebnisse empirischer Studien über die Häufigkeit einzigartiger Ausprägungskombinationen. Schließlich stellen wir einen formalen Ansatz vor, durch den sich die Wahrscheinlichkeit einzigartiger Ausprägungskombinationen sowie die Sicherheit darüber, daß eine bestimmte Kombination einzigartig ist, ermitteln läßt.

4.2.1 Bestimmungsfaktoren der Einzigartigkeit von Ausprägungskombinationen

Ob ein Fall einer Population einzigartig in den ihn charakterisierenden Merkmalen und ihren Ausprägungen ist, hängt von zwei Faktoren ab: Vom Umfang der Population, d.h. der Anzahl der Einheiten, die der Population zugehören, sowie dem Auflösungsgrad der Merkmale und Merkmalsausprägungen, die als

Schlüssel dienen und durch die die Einheiten der Population klassifiziert werden.

Was den Umfang der Population betrifft, so werden offensichtlich bei gegebenem Auflösungsgrad der Merkmale, einzigartige Fälle umso seltener sein, je größer der Umfang der Population ist. Je mehr Einheiten einer Population zugehören, umso wahrscheinlicher ist es, daß eine Einheit, die durch spezifische Ausprägungen in den zur Charakterisierung einbezogenen Merkmalen bestimmt ist, diese Eigenschaft mit anderen Einheiten teilt. Wie unten gezeigt werden wird, besteht jedoch in der Regel keine rein lineare Beziehung zwischen dem Umfang N einer Population und der Anzahl einzigartiger Fälle.

In dem in dieser Studie relevanten Zusammenhang ist der Umfang der Population bestimmt durch die Abgrenzungskriterien, die bei der Generierung des Mikrodatenfile gelten. Diese sind nicht immer völlig eindeutig. Für viele praktische Zwecke wird man jedoch davon ausgehen können, daß hinreichend klare Abgrenzungen der Einheiten möglich sind, die einer Population zugehören. Dies gilt vor allem dann, wenn die Abgrenzungskriterien rein territorialer Art sind und wenn damit - wie bei manchen Bevölkerungserhebungen - alle Personen eingeschlossen werden, die innerhalb eines bestimmten Gebietes ihren Wohnsitz haben.[5] Insbesondere auf der Basis von Regionalinformationen lassen sich verhältnismäßig einfach und verläßlich auch relevante Subpopulationen in gleicher Weise innerhalb des Mikrodatenfile und der Grundgesamtheit abgrenzen. Die Möglichkeit der Abgrenzung von Subpopulationen ist vor allem deshalb wichtig, weil bei hinreichend enger Eingrenzung solcher Subpopulationen die Bestimmung der Wahrscheinlichkeit der Einzigartigkeit präziser und mit größerer Sicherheit vorgenommen werden kann (Mokken/Pannekoek/Willenborg 1989, Willenborg 1990) und weil die Beschaffung vollständigen Zusatzwissens unter Umständen wesentlich weniger aufwendig ist. Im folgenden gehen wir jedoch auf das Problem der Subpopulationsbildung nicht weiter ein. Die für die Gesamtpopulation geltenden Aussagen sollen möglichst so formuliert werden, daß sie auch auf Subpopulationen übertragen werden können.

---

5 Aber selbst dieser einfachste Fall macht deutlich, daß es sich bei der Abgrenzung der Population nicht um ein triviales Problem handelt. Selbst die Wohnsitzfrage ist nicht immer eindeutig bestimmt, und in aller Regel gibt es eine Reihe weiterer Unschärfen in der Populationsabgrenzung. Welche Schwierigkeiten damit verbunden sein können, die einem Mikrodatenfile entsprechende Population zu bestimmen, geht eindrucksvoll aus dem Beitrag von Schnell (1991) hervor.

Die Bestimmung des Auflösungsgrades der Merkmale und ihrer Ausprägungen, die als Zuordnungsschlüssel dienen, ist komplexer. Der Auflösungsgrad dieses Schlüssels ist umso größer,

- je größer die Zahl der einbezogenen Merkmale ist,
- je größer die Zahl der Ausprägungen der einzelnen Merkmale ist,
- je weniger die einzelnen Merkmale untereinander korreliert sind,
- je geringer die Konzentration der Fälle auf einzelne Merkmalsausprägungen, bzw. die Kombination von Merkmalsausprägungen ist.

Ein geeignetes Maß für den Auflösungsgrad des Zuordnungsschlüssels ist die Entropie der Verteilung aller Fälle auf die Kombinationen der in den Schlüssel einbezogenen Merkmale und ihrer Ausprägungen. Bezeichnen wir mit C die Gesamtzahl unterschiedlicher Ausprägungskombinationen des Schlüssels, mit i = 1, 2, ..., C je eine dieser Kombinationen und mit p_i die relative Häufigkeit der Fälle, die die Ausprägungskombination i aufweisen, dann ist das Maß für die Entropie dieser Verteilung E definiert als

$$E = - \sum_{i=1}^{C} p_i * \ln(p_i)$$

Die maximale Entropie wird erreicht, wenn jede besetzte Ausprägungskombination durch je einen Fall besetzt ist und die N Fälle sich mit gleichem Gewicht $p_i = 1/N$ auf die besetzten Ausprägungskombinationen verteilen. Je größer der Anteil der nur mit 1 besetzten Ausprägungskombinationen ist, umso größer ist ceteris paribus die Entropie.

Wichtig hierbei ist, daß die Entropie und damit der Auflösungsgrad des Schlüssels nicht nur von der Anzahl C der nicht mit Null besetzten Ausprägungskombinationen bestimmt ist. Entscheidend ist zudem, wie die Fälle auf die einzelnen Kombinationen verteilt sind, d.h. mit welcher Wahrscheinlichkeit p_i die Ausprägungskombinationen auftreten. Nur in dem Sonderfall, in dem alle Merkmale voneinander unabhängig und gleichverteilt sind, wäre für jede Kombination die gleiche Auftretenswahrscheinlichkeit zu erwarten. In der Realität treffen die Annahmen von Gleichverteilung und Unabhängigkeit allerdings nahezu nie zu. Vielmehr treten in der Regel Datensätze in bestimmten Aus-

prägungsregionen verstärkt (Klumpenbildung) auf und nur wenige Datensätze liegen am äußeren Rand oder ganz außerhalb dieser Klumpen. Selbst unter der Annahme, daß alle Ausprägungskombinationen nur einfach besetzt sind, ist bei einem zufälligen Ziehen einer beliebigen Ausprägungskombination die Wahrscheinlichkeit, eine Kombination aus der Mitte zu treffen, wesentlich höher als in den Randzonen. Bezogen auf die Entropie und damit den Auflösungsgrad des Schlüssels transportieren die niedrigen Wahrscheinlichkeiten mehr Informationen als hohe Wahrscheinlichkeiten. Aus diesem Grunde ist eine mit eins besetzte Ausprägungskombination eines "Ausreißers" stärker reidentifikationsgefährdet als eine mit eins besetzte durchschnittliche Ausprägungskombination. Datensätze mit seltenen Ausprägungskombinationen sind daher stärker reidentifikationsgefährdet als solche mit durchschnittlichen Ausprägungskombinationen.

4.2.2 Die Wahrscheinlichkeit von Einzelfällen: empirische Befunde

Empirisch variiert damit die Quote der Einzelfälle offensichtlich mit dem Populationsumfang und dem Auflösungsgrad des Schlüssels. Je nach Bedingungen können damit Einzelfallquoten hergestellt werden, die zwischen 0 Prozent und fast 100 Prozent variieren. Während der Populationsumfang für bestimmte Szenarien vergleichsweise eindeutig festgelegt werden kann, gibt es bei den Annahmen über die zur Verfügung stehenden Überschneidungsmerkmale und die dabei nutzbare Ausprägungsdifferenziertheit vielfältige Ermessensaspekte. Es ist deshalb sehr wichtig, Grundlagen für realistische Abschätzungen über das in bestimmten Szenarien verfügbare Zusatzwissen, das über die Differenziertheit des Schlüssels entscheidet, zu schaffen (vgl. dazu insbesondere Kapitel 6).

Wenn wir im folgenden die Ergebnisse einzelner vorliegender empirischer Untersuchungen kurz resümieren, sollen damit die Grenzen für realistische Annahmen über die Höhe von Einzelfallquoten in der Population abgesteckt werden.

Skinner et al. (1990) referieren die Ergebnisse einer Untersuchung mit Daten der italienischen Volkszählung. Die untersuchte Population sind konkret die 3,5 Millionen Einwohner der Toskana. In dieser Untersuchung wurde verfügbares Zusatzwissen für folgende acht Merkmale und Ausprägungskategorien unterstellt: Regionalgliederung (9 Kategorien), Stellung zum Haushaltsvorstand (4

Kategorien), Wohnsituation (Besitz/Miete, 3 Kategorien), Geschlecht (2 Kategorien), Familienstand (5 Kategorien), Berufsgruppe (8 Kategorien), Stellung im Beruf (15 Kategorien) sowie Alter in verschieden differenzierten Gruppierungen. Bei der differenziertesten Aufgliederung des Regionalmerkmals und bei Einjahresaltersgruppen ergeben sich ungefähr 13 Millionen unterschiedliche Ausprägungskombinationen. Die Autoren gehen mit nicht unplausiblen Argumenten davon aus, daß damit Annahmen über einen Schlüssel gemacht sind, der eine Obergrenze des verfügbaren Wissens darstellt, insbesondere dann, wenn angenommen wird, daß diese Informationen für eine gesamte Population verfügbar sein müßten. In Tabelle 4.1 sind die Ergebnisse über die Einzelfallquoten aufgeführt, die mit diesem Schlüssel bei unterschiedlichen Annahmen über die regionale Gliederungstiefe und Differenziertheit der Altersgliederung erzielt wurden. Sie liegen auch in dem Fall, in dem die kleinste Regionaleinheit nicht mehr als 200 000 Einwohner umfaßt, unter 1 Prozent.

Tabelle 4.1: Anteile einzigartiger Kombinationen der Schlüsselvariablen in der Bevölkerung der Toskana

Regionale Differenzierung	kleinste Regional-einheit (in 1000)	Alter in 1-Jahres Gruppen	Alter in 5-Jahres Gruppen	Alter in 10-Jahres Gruppen
Keine	3 500	0.14%	0.01%	0.00%
Florenz vs. Rest	1 200	0.36%	0.06%	0.03%
Provinzen	200	0.98%	0.26%	0.14%

(Quelle: Skinner et al. 1990:6)

Dagegen findet Schlörer (1975) in einer Untersuchung, in der er die Zahl der genutzten Überschneidungsmerkmale variiert, folgende Einzelfallquoten:

Bei 5 Merkmalen mit insg. 9 000 Ausprägungskombinationen: 11 Prozent;
bei 6 Merkmalen mit insg. 58 000 Ausprägungskombinationen: 32 Prozent;
bei 7 Merkmalen mit insg. 360 000 Ausprägungskombinationen: 56 Prozent;
bei 8 Merkmalen mit insg. 2 200 000 Ausprägungskombinationen: 76 Prozent.

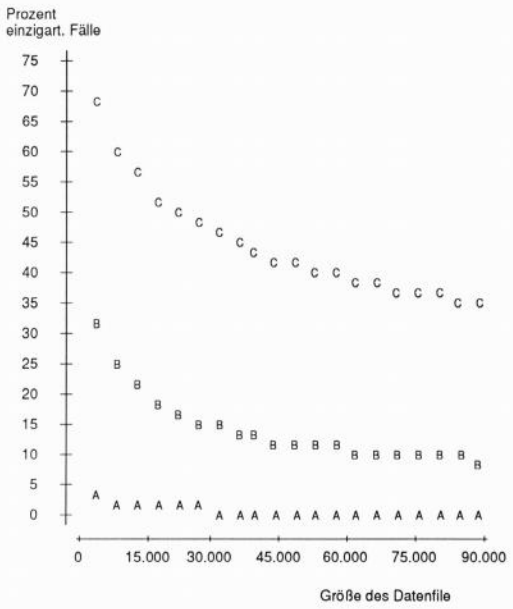
Die massiven Unterschiede in den Ergebnissen resultieren im wesentlichen daraus, daß sich die Untersuchung von Schlörer nur auf eine Population von 31 465 Fällen bezieht und in den Überschneidungsmerkmalen das exakte

Geburtsdatum enthalten ist. Die Population von Schlörer ist also wesentlich kleiner, und mit dem Geburtsdatum wird ein Überschneidungsmerkmal benutzt, mit dem eine außergewöhnlich hohe Auflösung erreicht werden kann, weil es allenfalls sehr schwach mit anderen Merkmalen korreliert.

Von besonderem Interesse ist eine Untersuchung von Greenberg/Voshell (1990), in der Populationsumfang und Entropie des Schlüssels systematisch variiert werden. Hier zeigt sich, daß die Quote der Einzelfälle nicht linear mit dem Populationsumfang N verknüpft ist. Beispielhaft ist dieser Sachverhalt in Abbildung 4.1 dargestellt. In diesem Schaubild wird mit Daten aus dem amerikanischen Zensus gezeigt, wie sich bei zunehmend größer werdendem Populationsumfang der Anteil der Einzelfälle verändert, wenn der Schlüssel zunehmend mehr Variablen umfaßt und damit einen höheren Auflösungsgrad bekommt. Der Schlüssel besteht aus typischen demographischen Variablen sowie Haushalts-, Erwerbstätigkeits- und Einkommensmerkmalen. Bei der Kurve A enthält er 81 200 mögliche Ausprägungskombinationen, bei der Kurve B ca. 1.6 Milliarden mögliche Kombinationen und bei Kurve C ca. 229 Milliarden mögliche Kombinationen. In den Varianten B und C ist also die Zahl der Ausprägungskombinationen um ein Vielfaches größer als der Populationsumfang. Dennoch finden sich nicht einmal bei Variante C und einem sehr kleinen Populationsumfang 100 Prozent Einzelfälle, da die Ausprägungen bei den verschiedenen Merkmalen nicht unabhängig voneinander sind.

Abbildung 4.1: Anteil einzigartiger Haushalte vs. Größe des Datenfile

Die Zeichen stellen die Anzahl der Variablen im Datenfile dar. A: 6 Variablen, B: 10 Variablen, C: 15 Variablen.



(Quelle: Greenberg/Voshell 1990)

Die Kurven illustrieren, daß mit zunehmendem Populationsumfang der Anteil der Einzelfälle nicht linear abnimmt. Je größer der Populationsumfang ist, umso kleiner wird zwar der Anteil der Einzelfälle, aber der Anteil der Einzelfälle vermindert sich mit wachsendem Populationsumfang zunehmend langsamer, und

es entsteht sogar der Eindruck, daß sich dieser Anteil jeweils asymptotisch einer unteren Saturationsschwelle annähert. Ab einer gewissen Größe verringert sich bei einem bestimmten Schlüssel der Anteil der Einzelfälle nicht mehr oder nur noch marginal. Leider ist das Experiment von Greenberg und Voshell auf einen maximalen Populationsumfang von 90 000 Einheiten begrenzt, so daß die Entwicklung der Zahl der Einzelfälle bei größeren Populationen nicht beobachtet werden kann. Dennoch lassen sich aus diesem Realexperiment einige aufschlußreiche Folgerungen im Hinblick auf das Verhältnis des Anteils von Einzelfällen in Stichproben zum Anteil der Einzelfälle in der Grundgesamtheit ziehen:

- Selbst bei dem differenziertesten Schlüssel (mit 229 Milliarden Ausprägungskombinationen!) liegt beim größten berücksichtigten Populationsumfang - der in der Untersuchung mit 90 000 Einwohnern allerdings als vergleichsweise klein angenommen ist - die Einzelfallquote unter 40 Prozent.

- Die in Abbildung 4.1 dargestellten Befunde blieben weitgehend identisch, wenn an Stelle der Populationserweiterung durch Einschluß zusätzlicher benachbarter Regionaleinheiten der Umfang der Datenfiles dadurch vergrößert wurde, daß aus einer Population Zufallsstichproben mit zunehmend größerem Auswahlsatz gezogen wurden. Es ist also damit zu rechnen, daß im Verhältnis einer Stichprobe mit kleineren Fallzahlen zur Population mit größeren Fallzahlen ähnliche Ergebnisse zu erwarten sind.

  Selbst wenn bei einem bestimmten Populationsumfang eine sehr hohe Quote von Einzelfällen gegeben ist, nimmt diese Quote bei wachsendem Populationsumfang sehr schnell ab; die Einzelfallquote nimmt umso stärker ab, je höher sie beim Ausgangspunkt ist. Diese Hypothese würde nur dann nicht gelten, wenn bei einem gegebenen Schlüssel beim Ausgangspunkt die untere Saturationsschwelle schon erreicht wäre. Ist in einer Stichprobe ein bestimmter Anteil von Einzelfällen zu beobachten, so ist zu erwarten, daß in der Population dieser Anteil wesentlich niedriger ist.

  Wenn in einer Stichprobe eine bestimmte Kombination von Merkmalsausprägungen einzigartig ist, kann daraus nicht geschlossen werden, daß die Kombination auch in der Grundgesamtheit einzigartig ist. Dies gilt nicht nur, weil in der Grundgesamtheit eine insgesamt niedrigere Einzelfallquote zu erwarten ist als in der Stichprobe. Dies ist auch deshalb der Fall, weil Ausprägungskombinationen, die in der Stichprobe einzigartig sind, in der Grundgesamtheit mehrfach besetzt sein können, und ebenso,

weil in der Grundgesamtheit Kombinationen einzigartig sein können, die in der Stichprobe überhaupt nicht besetzt sind.

4.2.3 Formale Modelle zur Schätzung der Wahrscheinlichkeit von Einzelfällen in der Population

Schätzungen über Einzelfallquoten in der Grundgesamtheit können durch formale Modelle auch auf der Basis von Stichproben aus dieser Grundgesamtheit gewonnen werden. Dabei können sowohl globale Einzelfallquoten (wie groß ist der Anteil der Einzelfälle in der Population insgesamt) geschätzt werden als auch Einzelfallquoten unter der Bedingung, daß auch in der Stichprobe ein Einzelfall gegeben ist. Solche Modelle haben den Vorzug, daß die interessierenden Informationen über die Grundgesamtheit auf der Basis eines existierenden Mikrodatenfile, das ja in der Regel eine Stichprobe ist, geschätzt werden können. Mit ihnen ist also eine Möglichkeit gegeben, ohne Kenntnis der Grundgesamtheit selbst, die in den Gleichungen zum Reidentifikationsrisiko wichtigen Größen sowohl der Wahrscheinlichkeit der Populationseinzigartigkeit zu bestimmen wie der Wahrscheinlichkeit (Sicherheit), daß eine bestimmte Ausprägungskombination in der Grundgesamtheit ein Einzelfall ist. Bei formalen Modellen muß auf der anderen Seite immer die Einschränkung in Kauf genommen werden, daß vereinfachende Modellannahmen unterstellt werden müssen.

In den letzten Jahren sind solche Modelle in mehreren Arbeiten hauptsächlich von einer Wissenschaftlergruppe im Niederländischen Statistischen Zentralamt (Mokken/Pannekoek/Willenborg 1989, Bethlehem/Keller/Pannekoek 1990, Koolman/Willenborg 1990, Willenborg 1990) erarbeitet worden. Skinner et al. (1990) haben diese Modelle weitergeführt. Das Verdienst der Entwicklung der Modelle kommt zweifellos der holländischen Gruppe zu. Das grundlegende Modell ist das von Mokken/Pannekoek und Willenborg (1989) vorgeschlagene Poisson-Gamma-Modell. Wir resümieren hier nur ganz knapp die wesentlichen Grundlagen und wenden das Modell dann auf den Mikrozensus und die EVS an. Bei der Darstellung folgen wir im wesentlichen den Formulierungen bei Skinner et al. (1990).

Das Modell geht von der Voraussetzung aus, daß die Häufigkeiten, mit der in einer Grundgesamtheit die C unterschiedlichen Ausprägungskombinationen von Merkmalen besetzt sind, eine Poisson-Verteilung aufweisen, bei der die Raten

des Auftretens der Häufigkeiten nach einer Gamma-Verteilung mit den Parametern α und ß variieren. Unter der Restriktion $C\alpha ß=1$ ist nach diesem Modell:

(1) $W(U_P) = (1+Nß)^{-(1+\alpha)}$

(2) $W(U_S) = (1+nß)^{-(1+\alpha)}$

(3) $W(U_P|U_S) = [(1+nß)/(1+Nß)]^{-(1+\alpha)}$, wobei

N = Populationsumfang

n = Stichprobenumfang

C = Anzahl unterschiedlicher Ausprägungskombinationen

$W(U_P)$ = Wahrscheinlichkeit von Einzelfällen in der Population, bzw. Wahrscheinlichkeit, daß eine zufällig aus der Population herausgegriffene Einheit eine einzigartige Ausprägungskombination aufweist,

$W(U_S)$ = Wahrscheinlichkeit von Einzelfällen in der Stichprobe, bzw. Wahrscheinlichkeit, daß eine zufällig aus der Stichprobe herausgegriffene Einheit in der Stichprobe eine einzigartige Ausprägungskombination aufweist,

$W(U_P|U_S)$ = Wahrscheinlichkeit eines Einzelfalls in der Population, unter der Voraussetzung, daß in der Stichprobe ein Einzelfall gegeben ist.

Der entscheidende Parameter dieses Modells ist ß. Er steuert die Entropie in der Verteilung der Besetzungshäufigkeiten der C Ausprägungskombinationen. Je kleiner ß ist, umso größer ist die Einzelfallquote. Bei großem C wird aufgrund der Restriktion $C\alpha ß=1$ der Parameter α in der Regel sehr klein. Er kann unter dieser Bedingung in den Gleichungen (1) und (2) zur Vereinfachung ohne große Folgen vernachlässigt werden. Unter dieser Voraussetzung empfehlen Skinner et al. (1990) einen einfachen Weg zur Schätzung von ß: Setzt man in (2) an Stelle von $W(U_S)$ den tatsächlich beobachteten Anteil p_{us} von Einzelfällen in der Stichprobe, dann kann durch einfache Transformation ß aus der Stichprobe geschätzt werden.[6] Es ergibt sich:

(4) $\hat{ß} = (1/p_{us}-1)/n$,sowie

(5) $\hat{W}(U_P) = [1+(N/n)(1/p_{us}-1)]^{-1}$

---

6 Komplexere Schätzverfahren werden von Mokken/Pannekoek/Willenborg (1989) entwickelt.

Aus (5) ist ersichtlich, daß der Auswahlsatz n/N - bzw. der Hochrechnungsfaktor von der Stichprobe auf die Grundgesamtheit N/n - nach diesem Modell entscheidend die Schätzung des Anteils der Einzelfälle in der Population aus dem Anteil der Einzelfälle in der Stichprobe bestimmt, jedoch nicht ausschließlich.

Beispiele für einige nach dem Poisson-Gamma-Modell geschätzte Einzelfallquoten in der Population sind in Tabelle 4.2 aufgeführt. Als Stichprobenauswahlsatz werden die Auswahlsätze des Mikrozensus und der EVS berücksichtigt. Außerdem wird unterschieden, ob Kenntnis über Einzigartigkeit in der Stichprobe berücksichtigt wird oder nicht. In den ersten beiden Ergebnisspalten sind also die Werte dargestellt, die die nach (5) geschätzten Wahrscheinlichkeiten in (1) ergeben, in der dritten und vierten Spalte dagegen die Wahrscheinlichkeiten (3); die Werte in der dritten und vierten Spalte geben also an, wie wahrscheinlich es ist, daß eine ganz bestimmte Ausprägungskombination, die in der Stichprobe einzigartig ist, auch in der Grundgesamtheit einzigartig ist, während die ersten beiden Spalten die Schätzwerte für die insgesamt erwarteten Einzelfallquoten in der Population enthalten.

Tabelle 4.2: Geschätzte Anteile Einzelfälle (in Prozent) in der Population bei unterschiedlichen Anteilen von Einzelfällen in der Stichprobe, unterschiedlichen Auswahlsätzen und vorhandener oder nicht-vorhandener Kenntnis über Stichprobeneinzigartigkeit

Anteil Einzelfälle in Stichprobe	Kenntnis über Einzigartigkeit in Stichprobe ist			
	nicht gegeben		gegeben	
	bei Auswahlsatz von			
	1%	0.2%	1%	0.2%
5%	0.05%	0.01%	0.00%	0.00%
10%	0.11%	0.02%	0.01%	0.00%
20%	0.25%	0.05%	0.05%	0.01%
30%	0.43%	0.09%	0.13%	0.03%
40%	0.66%	0.13%	0.27%	0.05%
50%	0.99%	0.20%	0.50%	0.10%
60%	1.48%	0.30%	0.89%	0.18%
70%	2.28%	0.47%	1.60%	0.33%
80%	3.85%	0.79%	3.08%	0.64%
90%	8.26%	1.77%	7.43%	1.59%
95%	15.97%	3.66%	15.17%	3.48%

Was nun die für die Population insgesamt geschätzten Einzelfallquoten betrifft, so entsprechen sie weitgehend den mit dem Stichprobenauswahlsatz gewichteten Einzelfallquoten der Stichprobe, wenn die Einzelfallquote in der Stichprobe niedrig ist (20 Prozent oder weniger); bei größeren Einzelfallquoten in der Stichprobe nehmen die Einzelfallquoten in der Population überproportional zu. Dennoch würden sich nach diesem Modell beachtenswerte Einzelfallquoten in der Population erst dann ergeben, wenn - etwa beim Beispiel des Mikrozensus - in der Mikrozensus-Stichprobe Einzelfallquoten von 70 Prozent oder mehr gefunden würden.

Daß eine *bestimmte* in der Stichprobe als einzigartig gefundene Ausprägungskombination auch in der Population einzigartig ist, erweist sich nach der dritten und vierten Spalte als deutlich weniger wahrscheinlich als die Wahrscheinlichkeit von Einzelfällen in der Population insgesamt. Weder beim Mikrozensus noch bei der EVS könnte man überhaupt jemals mit einer hinreichenden Sicherheit schließen, daß eine in diesen Stichproben gefundene einzigartige Ausprägungskombination auch in der Gesamtbevölkerung ein Einzelfall ist.

Falls diese Modell-Ergebnisse auch einer empirischen Überprüfung standhalten, würden sich weitreichende Schlußfolgerungen für einige der in der Übersicht in Abschnitt 4.1.2 angeführten Szenarien ergeben. Die in Tabelle 4.2 angeführten Werte in den ersten beiden Spalten können als Schätzwerte für die Risikokomponente der Populationseinzigartigkeit betrachtet werden. Die in der dritten und vierten Spalten enthaltenen Werte sind dagegen Schätzwerte für die Sicherheitskomponente. Aus den Ergebnissen müßte geschlossen werden, daß im Falle von Mikrozensus und EVS bei Szenarien mit gezielter Suche aus Einzigartigkeit im Mikrodatenfile praktisch nie mit hinreichender Sicherheit auf Einzigartigkeit in der Population geschlossen werden kann.

Nur für Fischzugsszenarien ergäbe sich unter Umständen eine andere Schlußfolgerung. Würde ein Angreifer bei einem Fischzug über ein Identifikationsfile verfügen, das die Grundgesamtheit - oder eine interessierende Teilpopulation der Grundgesamtheit - weitgehend umfaßt, könnte die Wahrscheinlichkeit und Sicherheit der Populationseinzigartigkeit aus dem *Identifikationsfile* geschätzt werden. Nehmen wir als Beispiel ein Identifikationsfile, das 80 Prozent der Fälle der Population enthalte, Zufallsauswahl sei vorausgesetzt. Wenn dieses File 50 Prozent Einzelfälle enthält, dann wäre zu erwarten, daß die Grundgesamtheit 44,4 Prozent Einzelfälle enthält. Von einer einmaligen Merkmalskombination in einem solchen Identifikationsfile wäre mit 22,2 Prozent Sicherheit zu erwarten, daß diese Kombination auch in der Grundgesamtheit einzigartig ist. Enthält ein Identifikationsfile, das 80 Prozent der Fälle der Population abdeckt, 80 Prozent Einzelfälle, dann betrüge die erwartete Einzelfallquote in der Grundgesamtheit 76 Prozent und es könnte mit einer Sicherheit von 61 Prozent erwartet werden, daß eine im Identifikationsfile einmalige Kombination von Merkmalsausprägungen auch in der Population einmalig ist.

Als allgemeine Schlußfolgerung ergibt sich hieraus, daß für Fälle, in denen das Mikrodatenfile eine Stichprobe mit einem Auswahlsatz von höchstens 1 Prozent ist, ein beachtenswertes Reidentifikationsrisiko allenfalls von Szenarien mit 'response knowledge'sowie von Fischzugsszenarien ausgeht, bei denen der Angreifer über ein Identifikationsfile verfügt, das annähernd die gesamte Population enthält und bei dem ein außerordentlich hoch auflösender Schlüssel von Überschneidungsmerkmalen zum Mikrodatenfile gegeben ist.

Neben der Lösung der Einzigartigkeitsproblematik muß jedoch auch das Kompatibilitätsproblem gelöst werden. Den damit zusammenhängenden Fragen wenden wir uns im folgenden Abschnitt zu.

4.3 Kompatibilität der Überschneidungsmerkmale

In den meisten vorliegenden Untersuchungen zum Reidentifikationsrisiko wird das Problem der Kompatibilität der Überschneidungsmerkmale in Mikrodatenfile und Identifikationsfile auf eine der beiden folgenden Weisen behandelt: In einem Teil der Arbeiten wird es überhaupt nicht erwähnt und es wird davon ausgegangen, daß Fälle mit übereinstimmenden einzigartigen Ausprägungskombinationen, vor allem wenn sich diese auf die Grundgesamtheit beziehen, ohne Einschränkungen als reidentifizierbar zu betrachten sind (Fischer-Hübner 1986, Brunnstein 1987, Dalenius 1986). In einem anderen Teil der Literatur wird zwar darauf hingewiesen, daß das Reidentifikationsrisiko durch Datenfehler und Inkompatibilitäten in der Abbildung der Überschneidungsmerkmale in Mikrodatenfile und Identifikationsfile verringert wird. Bei der Schätzung des Reidentifikationsrisikos wird dann jedoch von diesem Problem abgesehen und die Annahme unterstellt, die Daten seien vollständig kompatibel (Schlörer 1977, Brennecke/Schneider 1977, Bethlehem et al. 1990, Spruill 1983). Nur wenige Arbeiten stellen Dateninkompatibilitäten explizit in Rechnung (Paaß/Wauschkuhn 1985, Skinner et al. 1990).

In dieser Studie gehen wir davon aus, daß den Folgen von Dateninkompatibilitäten für das Reidentifikationsrisiko bislang zu wenig Aufmerksamkeit zuteil wurde und deren Mißachtung zu einer erheblichen Überschätzung des Reidentifikationsrisikos geführt hat. Vor allem für Verfahren, die auf dem einfachen Abgleich auf Identität von Merkmalsausprägungen basieren, ist zu erwarten, daß sie durch Inkompatibilitäten in der Abbildung eines Falles im Mikrodatenfile und im Identifikationsfile empfindlich gestört werden. Wann immer ein Fall in einem Merkmal, das bei einem Reidentifikationsversuch zum Abgleich verwendet wird, im Mikrodatenfile und im Identifikationsfile unterschiedlich abgebildet ist, kann keine korrekte Reidentifikation erfolgen. Aus diesem Grund wird das Inkompatibilitätsproblem im folgenden ausführlich behandelt.

Wir verwenden hier bewußt den umfassenden Begriff von Dateninkompatibilität. Von Dateninkompatibilität sprechen wir dann, wenn ein Fall, der gleichzeitig im Mikrodatenfile und im Identifikationsfile enthalten ist, in den entsprechenden Datensätzen dieses File in einem oder mehreren Überschneidungsmerkmalen unterschiedlich abgebildet ist. Dabei ist entscheidend, daß eine Vielzahl von Faktoren zu einem unterschiedlichen Abbild beitragen kann. Wir sprechen auch bewußt von der *Abbildung* eines Falls in einem Datensatz. Damit soll betont werden, daß Datensätze je nach der Zwecksetzung, der Fragestellung und den Bedingungen, unter denen sie entstehen, unterschiedliche Schattierungen der Realität abbilden werden. Datensätze sind darüber hinaus in aller Regel das

Ergebnis eines langen Generierungsprozesses, in den sowohl spezifische Bedingungen bei der Datenerhebung - Stichwort: Datenerhebung als sozialer Prozess (Esser 1984c) - als auch spezifische Bearbeitungsroutinen und Bearbeitungsstandards bei der Datenaufbereitung einfließen können. Jedes in einer Analyse verwandte Datum ist das Ergebnis aller operationalen Schritte, die bei seiner Genese von Bedeutung sind, seien diese Schritte nun explizit geplant und kontrolliert, oder ergeben sie sich naturwüchsig in einem bestimmten Kontext der Datenerhebung und -aufbereitung.

Es dürfte nun die Regel sein, daß Mikrodatenfile und Identifikationsfile aus sehr unterschiedlichen Datengenerierungsprozessen stammen. Die Mikrodatenfiles, um die es in dieser Studie geht, resultieren aus Datenerhebungen der amtlichen Statistik. Die Identifikationsfiles dagegen werden aus gänzlich anderen Datengenerierungskontexten stammen. Es kann sich um Informationen aus Registern handeln, die zu einem bestimmten Zweck erstellt sind und bei denen anzunehmen ist, daß die Informationen in zweckentsprechender Weise aufgenommen werden. Die Informationen des Identifikationsfile können aus alltäglichem Anschauungswissen stammen, das durch Beobachtungsverzerrungen oder ungenaue Kenntnisse beeinträchtigt sein kann. Die Informationen des Identifikationsfile könnten aus einer Befragung stammen, bei der unterschiedliche Fragekontexte aus unterschiedlichen Frage - und Antwortvorgaben oder aus unterschiedlichen Kodierungsregeln Datensätze entstehen lassen, die sich zumindest in Teilen von den Daten unterscheiden, die im Mikrodatenfile enthalten sein werden. Auf jeden Fall kann - vielleicht mit Ausnahme von Merkmalen wie Geschlecht und Lebensalter - nicht damit gerechnet werden, daß eine Kompatibilität der Abbildung problemlos gegeben ist.

Es ist an dieser Stelle zu betonen, daß das Problem der Dateninkompatibilitäten nicht auf ein reines Datenfehlerproblem reduziert werden kann. Datenfehler sind Abweichungen eines bei einem Meßobjekt tatsächlich gemessenen Wertes von dem "wahren Messwert". Nach der klassischen Meßtheorie ist der "wahre Wert" unter spezifischen Voraussetzungen und Annahmen erfaßbar als Mittelwert einer Serie von Messungen mit dem gleichen Instrument am gleichen Objekt unter gleichen Bedingungen. Meßfehler sind die unsystematischen Abweichungen der Ergebnisse der Einzelmessungen vom "wahren Wert", die durch vielfältige spezielle Bedingungen, die die Einzelmessung beeinflussen können, zustande kommen. Radikale Kritiker haben die Vorstellung aufgegeben, daß es, insbesondere bei Messungen am lebenden Objekt, überhaupt einen "wahren Wert" geben könne. Ein Meßvorgang sei letztlich ein sozialer Akt, in dem Personen nicht anders als in anderen Situationen handeln, nämlich als Akteure, die Situationen interpretieren, Bedeutungen stets neu aushandeln

und definieren und dabei ihr Verständnis, ihre Zielsetzungen und Interessen situationsspezifisch aktiv einfließen lassen. Es mache keinen Sinn, überhaupt von stabilen, situationsunabhängig gegebenen, latenten Eigenschaften und Merkmalen von Personen auszugehen. Selbst wenn man diese radikale Sicht nicht teilt, zweifelt heute kein Sozialwissenschaftler ernsthaft daran, daß Daten situations- und kontextbedingt sind und deshalb von Messung zu Messung Variationen aufweisen, die natürlich je nach Meßgegenstand, Untersuchungsmethode, Situationsvariabilität etc. unterschiedlich stark sein können.[7]

Datenfehler im Sinne des klassischen Reliabilitätsverständnisses tragen durchaus auch zu Dateninkompatibilitäten bei. Denn sowohl bei den für das Mikrodatenfile als auch bei den für das Identifikationsfile vorgenommenen Messungen wird, wie bei allen Messungen, mit Datenfehlern zu rechnen sein. Darüber hinaus sind jedoch weitere Inkompatibilitäten zu erwarten, die aus unterschiedlichen Meßintentionen, unterschiedlichen Datenerhebungs- und Datenbearbeitungsverfahren usw. resultieren. Vor allem bei den Arten von Zusatzwissen, die auf Registern basieren, kann als weitere Komponente hinzukommen, daß die im Register enthaltenen Fälle ein Interesse daran haben, in einer bestimmten - für sie vorteilhaften - Weise dargestellt zu werden (siehe dazu Kapitel 6).

Ohne den Anspruch auf Vollständigkeit werden im folgenden zunächst auf Datenfehlern beruhende Inkompatibilitäten näher untersucht. Als Beispiele anderer Elemente wird dann auf Dateninkompatibilitäten eingegangen, die aus unterschiedlichen sachlichen und zeitlichen Bezügen resultieren. Dann werden einige empirische Befunde zur Illustration des Ausmaßes von Dateninkompatibilitäten zusammengetragen und ein Modell zur quantitativen Abschätzung des Effekts von Dateninkompatibilitäten auf das Reidentifikationsrisiko dargestellt.

---

7 Zur Diskussion um das Konzept des "wahren Wertes" vgl. Esser (1986) und dort angeführte Literatur.

4.3.1 Datenfehler [8]

Es ist in diesem Zusammenhang nicht möglich, umfassend die Fehlertheorie und Fehlerlehre zu Erhebungen sozialwissenschaftlicher oder statistischer Daten zu erörtern. Dieses ist ein eigener, weit ausgebauter Bereich der sozialwissenschaftlichen und statistischen Methodenlehre und Methodenforschung. Die wichtigsten allgemeinen theoretischen Ansätze basieren einerseits auf Grundlagen der Kognitions-, Emotions- und Motivationspsychologie (Schwarz et al. 1985, Schwarz et al. 1989, Hippler et al. 1987, Schwarz 1990, Sudman/Bradburn 1974, Bradburn/Sudman 1979), andererseits auf einer allgemeinen Theorie des sozialen Handelns, bei denen das Verhalten in Befragungssituationen als Spezialfall sozialen Handelns angesehen wird (Esser 1984b, Esser 1984d, Esser 1986, Schnell/Hill/Esser 1988).[9]

In aller Kürze soll im folgenden nur auf einige konkrete und in der Forschung gut dokumentierte Faktoren hingewiesen werden, die einen Einfluß darauf haben, wie ein Sachverhalt schließlich als Datum in einem Datensatz abgebildet wird, sowie darauf, wie gleiche Sachverhalte in verschiedenen Datenfiles unterschiedlich abgebildet sein können. Die wichtigsten dieser Fehlerfaktoren sind sogenannte Antwortverzerrungen bei der Datenerhebung (vor allem untersucht für Datenerhebungen durch Befragungen) und Fehler, die bei der Datenregistrierung und -aufbereitung entstehen können.

Zu *Antwortverzerrungen* bei der Datenerhebung tragen vor allem die folgenden Elemente bei:

> *Frageeffekte*: Selbst einfache Fragen werden häufig mißverstanden. Sozialpsychologische Experimente haben gezeigt, daß Befragte die ihnen vorgelegten Fragen oft anders interpretieren als der Fragende es erwartet.[10] Selbst als geringfügig erscheinende Unterschiede in der Frage-

---

8 In allgemeinen Abhandlungen zum Thema "Fehler in Statistiken" werden stets auch Stichprobenfehler thematisiert, die hier jedoch keine Rolle spielen, da es nicht um die Hochrechnung der auf der Basis einer Stichprobe erzielten Ergebnisse geht, sondern um einzelne Fälle.

9 Die Reliabilität von Erhebungsinstrumenten ist eines der wichtigsten Themen in der Methodenlehre der empirischen Sozialforschung (vgl. z.B. die Lehrbücher Friedrichs 1973:100ff., Schnell/Hill/Esser 1988:147ff. und Heidenreich 1987:365ff. sowie die Beiträge Wegener 1983, Lord und Novick 1974, Koch 1986).

10 Schwarz (1988:5) berichtet davon, daß in Befragungen geäußerte Auffassungen darüber, welche Aktivitäten "bezahlte Arbeit" darstellen, mit dem Alter und der sozialen Stellung wechseln. Grenzfälle, wie "Babysitten" und "Rasenmähen",

formulierung können zu deutlich unterschiedlichen Antworten führen. Wenn ein Befragter Probleme mit dem Verständnis von Fragen hat, greift er auf Informationen zurück, die ihm der Fragesteller absichtlich oder unabsichtlich bietet. Werden ihm bestimmte Antwortvorgaben vorgelegt, so beeinflussen diese sein Frageverständnis und die im folgenden abgegebenen Verhaltensberichte und Urteile.

Reihenfolgeeffekte der Fragen und Antwortvorgaben: Jede Frage aktiviert in bestimmter Weise die Aufmerksamkeit und das Gedächtnis des Befragten und ruft bestimmte Assoziationen hervor. Die Reihenfolge der Fragen und die durch sie geschaffenen Aufmerksamkeits- und Assoziationskontexte beeinflussen deshalb die Antworten auf die Fragen. Ähnliche Effekte können bei Antwortvorgaben auftreten. Bekannte Effekte sind hier beispielsweise, daß sich Befragte bei einer Liste von Antwortvorgaben schnell für eine am Anfang stehende Alternative entscheiden und die anderen Alternativen nicht oder nur mit geringerer Aufmerksamkeit betrachten, oder daß sie sich für eine der am Ende der Liste stehenden Alternativen entscheiden, weil ihnen - vor allem beim mündlichen Vorlesen der Alternativen - die am Anfang stehenden Vorgaben nicht mehr präsent sind.[11]

Interviewer-Effekte: Untersuchungsergebnisse zeigen, daß das Antwortverhalten der Befragten von Merkmalen und Verhaltensweisen des Interviewers beeinflußt werden kann (z.B. von Alter, Geschlecht, Schichtzugehörigkeit, Erscheinungsbild, seinen eigenen Einstellungen und signalisierten Werthaltungen und Verhaltenserwartungen an den Befragten; vgl. Allerbeck/Hoag 1985, Atteslander/Kopp 1987:165 ff). Unterschiedliche Eigenschaften des Interviewers können unterschiedliche Antworten als sozial wünschenswert erscheinen lassen. Bestimmte Angaben will man nur gegenüber "vertrauenswürdig" erscheinenden Personen machen.[12]

---

werden von Jugendlichen häufiger als bezahlte Arbeit angesehen als von Erwachsenen.

11 Beinhalten die Antwortvorgaben eine quantitative Einordnung, so ist wichtig, wie stark der vorstellbare obere oder der untere Wertebereich aufgegliedert ist. Befragte tendieren dazu, "mittlere" Kategorien zu wählen. Daraus kann ein Einfluß auf die Verteilung von quantifizierten Angaben (z.B. zum Einkommen oder zur Arbeitszeit) entstehen (vgl. Schwarz/Hippler/Strack 1988:21f. und Schwarz/Hippler/Noelle-Neumann 1989).

12 Auch aus dem Bereich der amtlichen Statistik liegen Ergebnisse vor: Deininger (vgl. 1960:172ff.) führte eine Untersuchung zu Interviewereffekten im Rahmen einer Probeerhebung zum Mikrozensus durch. Er fand unter anderem signifikante Ein-

- *Effekte durch Anwesenheit Dritter oder durch stellvertretende Auskunfterteilung*: Eine wichtige Variable der Befragungssituation ist die Anwesenheit dritter Personen während einer Befragung, durch die das Antwortverhalten beeinflußt werden kann. Angaben können auch je nachdem variieren, ob sie von der betroffenen Person selbst oder stellvertretend durch Dritte gegeben werden (vgl. Statistisches Bundesamt 1978:68).[13]

- *Sponsorship-Effekte*: Das Antwortverhalten von Befragten kann dadurch beeinflußt sein, wer von ihnen als Auftraggeber der Untersuchung und Nutzer der Daten wahrgenommen wird. Erhebungen der Statistischen Ämter, wie z.B. der Mikrozensus und die EVS haben einen amtlichen Charakter und führen zu teilweise anderen Befragtenreaktionen als Erhebungen durch private Datenerhebungsinstitute. Das Befragtenverhalten ist auch dadurch beeinflußt, ob es sich um eine Datenerhebung mit Pflichtteilnahme oder freiwilliger Beteiligung handelt (Esser u.a. 1989, vgl. auch die Diskussion um Institutseffekte in der empirischen Sozialforschung).[14]

- *Effekte sozialer Erwünschtheit*: Hier handelt es sich um Antwortverzerrungen, die dann zustande kommen, wenn Befragte ihre Antworten danach ausrichten, was ihnen aufgrund bestimmter sozialer Normen und Werte allgemein oder in einer gegebenen Befragungssituation als sozial erwünscht erscheint. Hierbei handelt es sich um einen generell wirksamen Mechanismus, der insbesondere auch zur Erklärung der Interviewer-Effekte und der Sponsorship-Effekte beigezogen werden kann.

- *Fehlende Werte*: Als Folge von *Antwortverweigerung*, *Meinungslosigkeit*, oder *Nicht-Wissen* auf seiten der Befragten oder als Folge von Nachlässigkeit des Interviewers kommt es bei Datenerhebungen relativ häufig zu fehlenden Werten bei einzelnen Variablen. Je nach Frage und Situation

---

flüsse auf die Angaben zur Anzahl der Wohnungen in dem betreffenden Haus und auf die Antwortbereitschaft der Befragten. Bei der Personenzahl des Zielhaushalts und beim Geburtsjahr der Zielperson wurden keine signifikanten Interviewereffekte festgestellt.

13 Es sei darauf hingewiesen, daß Ergebnisse von Erhebungen unter Pflichtbeteiligung entgegen vielfach geäußerten Erwartungen keine geringere Qualität aufweisen als Erhebungen unter freiwilliger Beteiligung (Esser et al. 1989, Scheuch/Gräf/Kühnel 1989).

14 Vgl. Edwards/Walker (1961), Phillips/Clancy (1972), Sudman/Bradburn (1974), Schuessler/Hittle/Cardascia (1978), DeMaio (1984).

variiert die Quote fehlender Angaben zwar erheblich, aber fehlende Werte sind eine generelle Erscheinung und selbst bei amtlichen Erhebungen mit Auskunftspflicht festzustellen. Es kann darüber diskutiert werden, ob *fehlende Angaben* sinnvollerweise als Datenfehler zu bezeichnen sind. Im Zusammenhang des Reidentifikationsproblems wirken sich fehlende Angaben jedoch immer dann wie andere Dateninkompatibilitäten aus, wenn im Mikrodatenfile die Angaben zu einem Merkmal fehlen, während sie im Identifikationsfile enthalten sind oder umgekehrt.[15]

Weitere bekannte Momente, durch die Befragungsergebnisse beeinflußt werden oder durch die es zu fehlerhaften Daten kommen kann, sind:

- das Geben von Antworten trotz Nicht-Wissens oder trotz Meinungslosigkeit (Smith 1984);
- die Tendenz der Zustimmung unabhängig vom Inhalt der Fragen;
- das Ausweichen oder Geben bewußt unwahrer Antworten, mit dem insbesondere bei unangenehmen Fragen oder bei Tabuthemen gerechnet werden muß (Esser 1986);
- die Stimmung, in der sich ein Befragter befindet oder der Erfahrungshintergrund, vor dem eine Befragung erfolgt (Schwarz 199?);
- die Methode der Befragung, durch die die Daten erhoben werden (persönliche, schriftliche oder telephonische Befragung).[16]

Alle diese Punkte belegen, daß eine Vielzahl von Bedingungen existieren, die eine Quelle dafür sein können, daß bei unterschiedlichen Datenerhebungen selbst bei gleicher Meßintention das Meßergebnis unterschiedlich ist. Abweichungen der tatsächlich erfaßten Daten von den "wahren Werten" können nach der Phase der Datenerhebung auch in der Phase der *Datenaufbereitung* entstehen:

---

15 Für einen Datenangreifer erscheint es auch praktisch unmöglich, zu wissen, ob im Mikrodatenfile zu einem Merkmal eine Angabe existiert oder nicht.

16 Schwarz/Hippler/Noelle-Neumann berichteten (1989), daß sich die Befragten bei schriftlicher Befragung eher für weiter oben stehende Vorgaben entscheiden. Bei mündlicher Befragung werden hingegen eher zuletzt genannte Vorgaben gewählt. Diese werden bei der Formulierung der Antwort am ehesten erinnert.

- Neben Fehlern, die bei *der Übertragung der erhobenen Daten auf elektronische Speichermedien zustande kommen* können, sind dabei vor allem Fehler bei der *Kodierung der erhobenen Daten* zu erwähnen. Aus Untersuchungen der Kodier-Reliabilität ist bekannt, daß praktisch nie eine hundertprozentige Reliabilität erreicht werden kann. Nach einer im Zusammenhang mit der Volkszählung von 1961 von Szameitat und Deininger (1969) durchgeführten Untersuchung ist das Auftreten solcher Fehler bei manuellem Kodieren stark von formalen und inhaltlichen Eigenschaften des betreffenden Merkmals abhängig. Brauchten nur vorgegebene Kategorien verkodet zu werden, war die Fehlerhäufigkeit niedriger als ein Prozent. Höhere Fehlerraten waren zu verzeichnen, wenn keine fest vorgegebenen Antwortkategorien verwendet wurden. Inhaltlich erlaubt die Doppeldeutigkeit und Ungenauigkeit von Angaben oftmals keine eindeutige Zuordnung zu bestimmten vorgegebenen Kategorien. Szameitat und Deininger stellten bei der Kodierung der Berufskategorie eine Fehlerhäufigkeit von 4,0 Prozent fest.

- Bei *Plausibilitätkontrollen* werden die auf Speichermedien aufgenommenen Daten in der Regel maschinell auf inkonsistente Wertekombinationen geprüft (Beispiel: Personen mit Altersangabe unter 15 Jahren, die verheiratet oder erwerbstätig sind). Solche unplausiblen Wertekombinationen werden bereinigt. Dabei kann aus Gründen des Arbeitsaufwandes nicht in jedem Fall auf die Originalunterlagen zurückgegriffen werden. Auch das Einholen von klärenden Informationen durch Nacherhebungen bei den befragten Personen selbst ist oftmals unmöglich. Bereinigungen können maschinell vorgenommen werden, und die Bereinigung von Inkonsistenzen muß unter Umständen mit neuen Fehlern erkauft werden (Soll beispielsweise im obigen Beispiel die Altersangabe oder die Angabe zum Familienstand verändert werden?).

Die Zusammenstellung aller dieser Punkte mag den desillusionierenden Eindruck erwecken, als wären Datensammlungen wertlose Fehlerhaufen. Dieser Eindruck wäre falsch. Jeder empirische Forscher weiß zwar, daß Fehler unvermeidlich existieren und es völlig unrealistisch wäre, die tatsächliche Existenz von Datenfehlern und unreliablen Messungen in Abrede stellen zu wollen. Aber aus der Zusammenstellung der potentiellen Fehlerquellen kann nicht geschlossen werden, daß jedes Datum auch mit einem entsprechenden Fehler behaftet ist. Viele Einzeldaten sind fehlerfrei und die Fehler erreichen nicht ein Niveau, bei dem die Daten für Zwecke der wissenschaftlichen und statistischen Analyse unbrauchbar würden. Wenn einzelne Daten auch fehlerbehaftet sind, so

gleichen sie sich zumindest teilweise in der statistischen Aggregierung aus (vgl. die Untersuchungen des Statistischen Bundesamtes 1978:68ff). Damit ermöglichen Summenwerte, Durchschnittswerte, Anteilswerte oder andere statistische Kennziffern, die auf vielen Einzelwerten beruhen, durchaus zutreffende Charakterisierungen großer Gesamtheiten. Gleichzeitig hat die statistische Analyse zunehmend Techniken entwickelt, durch die die (mangelnde) Reliabilität von Einzeldaten explizit modelliert und korrigierend in statistischen Analysen kontrolliert werden kann (Faulbaum 1987, Bohmstedt 1983, Jagodzinski/Kühnel 1987, Porst/Schmidt/Zeifang 1987, Jöreskog/Sorböm 1979, Schwartz 1985, Clogg 1984, Duncan 1984, Lessler 1984; Krebs/Schmidt (im Druck), Bohrnstedt/Mohler/Müller 1987). Dennoch ist natürlich die Güte der Ergebnisse empirischer Forschung und statistischer Analyse direkt von möglichst wenig Datenfehlern in der Datenerhebung und Datenaufbereitung abhängig. Je weniger fehlerbehaftet Daten sind, umso höher ist die Qualität der Analyseergebnisse, und es muß eines der obersten Ziele empirischer Forschung sein, ein möglichst hohes Reliabilitätsniveau bei der Sammlung von Daten zu erzielen. Entsprechend werden auch in der amtlichen Statistik vielfältige Anstrengungen unternommen, Datenfehler möglichst zu beheben (Herberger 1985).

Für die Thematik dieser Studie ist aber besonders wichtig, daß in der Phase der Datenanalyse - dann, wenn die Daten bereinigt auf den Speichermedien verfügbar sind - in der Regel nicht mehr feststellbar ist, welches Datum in welchem Ausmaß fehlerbehaftet ist. Mit dieser Situation ist ein Datenangreifer konfrontiert. Er muß nicht nur davon ausgehen, daß das Mikrodatenfile fehlerbehaftet ist. Es ist auch damit zu rechnen, daß das Identifikationsfile fehlerhaft ist. Und da die beiden Files durch unterschiedliche Datengenerierungsprozesse entstanden sind, ist davon auszugehen, daß die Fehler unabhängig voneinander sind. Ein Datum im Mikrodatenfile mag fehlerfrei sein, während das entsprechende Datum im Identifikationsfile fehlerbehaftet ist und umgekehrt. Während sich bei der statistischen Analyse Fehler ausgleichen oder methodisch kontrolliert werden können, trägt beim Versuch einer Reidentifikation jeder Fehler - im Mikrodatenfile wie im Identifikationsfile - zu deren Verhinderung bei.

Auf die Abschätzung des Ausmaßes von Fehlern, soweit dieses aus empirischen Untersuchungen bekannt ist, werden wir in Abschnitt 4.3.4 zurückkommen. Zuvor ist jedoch auf Inkompatibilitäten in der Realitätsabbildung einzugehen, die zusätzlich zu den bislang diskutierten Datenfehlern zwischen Mikrodatenfile und Identifikationsfile zu erwarten sind.

4.3.2 Unterschiedlicher sachlicher und zeitlicher Bezug von Identifikationsfile und Mikrodatenfile

Nur bei wenigen Merkmalen ist die Aufstellung eines Kodierschemas und die Einteilung der Antworten trivial: Geburtsjahr und Geschlecht sind die wenigen Beispiele unproblematischer Kodierschemata. Für die meisten Merkmale bieten sich hingegen verschiedene Arten der Kodierung an. So wird z.B. der Beruf im Mikrozensus gemäß der deutschen amtlichen Berufsklassifikation erfaßt, während in der Sozialforschung häufiger der ISCO-Code verwendet wird. Die unterschiedlichen Kodierschemata sind in der Regel nicht eindeutig ineinander transformierbar; auf der Ebene der dreistelligen Berufskodes ist dies nur bei rund der Hälfte der Dreisteller möglich.[17]

Zwar sind Bestrebungen im Gange, die Vielfalt der Systematiken zu vereinheitlichen. Da die unterschiedlichen Kategorisierungen jedoch - zumindest zum Teil - aus den unterschiedlichen Erkenntnisinteressen resultieren und nicht einfach nur auf historisch verfestigten Gewohnheiten beruhen, ist nicht zu erwarten, daß die Unterschiede in den Systematiken vollständig verschwinden.

Datensammlungen privater oder öffentlicher Art, die nicht auf wissenschaftlicher Basis erfolgen, aber mögliche Quellen von Zusatzwissen sind, verwenden unter Umständen vollständig abweichende Systematiken, die auf die jeweiligen Zwecke der Erhebung ausgerichtet sind.

Der Grad der Entsprechung von Identifikationsfile und Mikrodatenfile in bezug auf die Abgrenzung der Kategorien ist offen. Wichtig ist jedoch, daß der Bereich des verwendbaren Zusatzwissens sich durch nicht zusammenfügbare Kategorien erheblich verkleinert.

Für eine identische Abbildung eines Sachverhaltes im Identifikationsfile und Mikrodatenfile kommt es also nicht allein darauf an, daß die gesammelten Daten fehlerfrei im Sinne der jeweiligen Meßintention sind. Es müssen tatsächlich auch die gleichen Meßintentionen und Konventionen angewandt werden. Faktisch kann man davon ausgehen, daß sich durch solche sachlich nicht vollständig kompatiblen Meß- und Kodierschemata der Umfang des ohne Einschränkungen übereinstimmenden Zusatzwissens erheblich einschränkt. Dies gilt in gleicher Weise für die im folgenden behandelten Unterschiede im zeitlichen Bezug.

---

17 Mündliche Auskunft von Experten bei ZUMA e.V., Mannheim.

Nur in seltenen Fällen beziehen sich Identifikationsfile und Mikrodatenfile auf den exakt gleichen Zeitpunkt. Nur wenige Merkmale sind jedoch in der Zeit unveränderlich (wiederum Geschlecht, Geburtsjahr). Für viele im Mikrozensus enthaltene Variablen, die als Überschneidungsmerkmale bei einem Reidentifikationsversuch in Betracht kommen, sind auf individueller Ebene bedeutende Veränderungen festzustellen: In jedem Jahr treten ungefähr sechs Millionen Personen eine Arbeitsstelle bei einer Firma an, bei der sie zuvor nicht beschäftigt waren (vgl. Rudolph 1986:261ff.). Der innerbetriebliche Arbeitsplatzwechsel tritt ähnlich häufig auf. In beiden Fällen ergeben sich vielfältige Veränderungen bei einer ganzen Reihe von Variablen. Das Einkommen, der ausgeübte Beruf, die Stellung im Beruf, die Stellung im Betrieb, die Branche, in der gearbeitet wird, die Merkmale der jeweiligen Tätigkeit und weitere Variable können betroffen sein. Ortswechsel können ebenfalls damit verbunden sein.

Mehr oder weniger unabhängig von der beruflichen Dimension sind auch für Merkmale, die den Familienstand und die Haushaltssituation beschreiben, erhebliche Veränderungen in der Zeit festzustellen.

Die Abweichungen zwischen Identifikationsfile und Mikrodatenfile werden um so größer sein, je größer der Abstand zwischen den Zeitpunkten ist, auf den sich die Daten in den beiden Files beziehen. Ein Datenangreifer muß darum bestrebt sein, Zusatzwissen zu verwenden, dessen Stichtag in der Nähe des Mikrodatenfile liegt. Das Ausmaß der Veränderungen für den Zeitraum eines Jahres zeigt jedoch, daß selbst Unterschiede von wenigen Monaten zu bedeutenden Dateninkompatibilitäten führen können.

Werden Registerdaten als Identifikationsfile verwendet, so ist der zeitliche Abstand zum Mikrodatenfile in der Regel für die einzelnen Einheiten des Registers variabel. Es gibt keinen festen Bezugspunkt für die gesamte Datensammlung. Die Daten werden vielmehr eingetragen, wenn ein bestimmter Anlaß gegeben ist, wenn z.B. jemand die Wohnung gewechselt, ein Haus erworben, einen Telefonanschluß beantragt hat oder in einen Verein oder Verband eingetreten ist. Register können aus diesem Grund zwar im Hinblick auf ihren primären Zweck korrekt, aber im Hinblick auf andere Information, die sie ebenfalls enthalten, in hohem und unkontrolliertem Grade fehlerhaft bzw. nicht mehr aktuell sein, weil Veränderungen, die sich in der Wirklichkeit vollziehen, keine entsprechende Aktualisierung in den Registereinträgen zur Folge haben.

Unterschiede in der sachlichen Kategorisierung einzelner Variablen sowie Unterschiede im zeitlichen Bezug sind neben Datenfehlern die offensichtlichsten Quellen möglicher Inkompatibilitäten zwischen Mikrodatenfile und

Identifikationsfile. Daneben sind noch weitere Quellen von Inkompatibilitäten, wie beispielsweise eine unterschiedliche Abgrenzung der Population oder unterschiedliche lokale Bezüge denkbar, die ein Angreifer ebenfalls zu berücksichtigen hätte.

Im Unterschied zu den im vorangehenden Abschnitt diskutierten Datenfehlern könnte ein Datenangreifer bei den hier erörterten Quellen von Dateninkompatibilitäten unter Umständen abschätzen, bei welchen Merkmalen mit einer besonders großen Zahl von Unstimmigkeiten zu rechnen ist. Er kann feststellen, bei welchen Variablen (teilweise) unvereinbare Verkodungskonventionen vorliegen. Auch Unterschiede im zeitlichen Bezug wird er in Rechnung stellen können. Allerdings dürfte dieses nicht viel helfen, denn es wird wiederum nicht bekannt sein, bei welchem konkreten Fall mit einer Inkompatibilität zu rechnen ist. Außerdem ist zu erwarten, daß wegen der Systematik der Effekte die Zahl der Fälle, die von derartigen Inkompatibilitäten betroffen ist, größer ist als die Zahl der Fälle, die durch klassische Datenfehler beeinträchtigt sind.

4.3.3 Zum Ausmaß und zur Bedeutung von Dateninkompatibilitäten für das Reidentifikationsrisiko

4.3.3.1 Empirische Befunde für sozialwissenschaftliche Erhebungen

Zur Abschätzung von Fehlern in den Informationsquellen, die möglicherweise als Zusatzwissen dienen könnten, liegen nur wenige empirische Ergebnisse vor. In Abschnitt 10.3.3 wird auf das Beispiel eines Adressenhändlers verwiesen, der einen Datensatz mit Branchenangaben weitergab, die zu mindestens 53 Prozent fehlerhaft waren. Da davon ausgegangen werden kann, daß nach wissenschaftlichen Kriterien durchgeführte Erhebungen weniger fehlerbehaftet sind, wird eine solche im folgenden herangezogen, um eine Abschätzung der Fehlerhäufigkeit nach unten zu erhalten. Damit wird nicht impliziert, daß die Verwendung von Umfragen als Zusatzwissen besonders wahrscheinlich sei.

In der Annahme, daß Erhebungen, die mit wissenschaftlichen Methoden durchgeführt werden, ein vergleichsweise niedriges Fehlerniveau haben, wird im folgenden eine solche Untersuchung herangezogen, um eine Abschätzung der Fehlerhäufigkeit zu erhalten, die an der unteren Grenze liegt.

Die in Tabelle 4.3 dargestellten Ergebnisse zur Häufigkeit von Abweichungen basieren auf einem Vergleich zweier Wellen einer sozialwissenschaftlichen Erhebung mit identischen Fragen. Bei der von ZUMA 1984 im Rahmen des "ALLBUS" (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften) durchgeführten Test-Retest-Studie war einer Teilstichprobe (154 Personen) der gleiche Fragebogen in Monatsabständen insgesamt dreimal vorgelegt worden, um die Frage beantworten zu können: "Wie stabil sind Umfragedaten?" (Porst/Zeifang 1987a, vgl. auch dies. 1987b und Koch 1986).

In Tabelle 4.3 ist der Anteil jener Personen zusammengestellt, die bei der dritten Welle anders antworteten als bei der ersten Welle. Um den Bezug auf die vorliegende Problemstellung zu erleichtern, wurden nur solche Fragen ausgewählt, die auch in der amtlichen Statistik in ähnlicher Form gestellt werden. Soweit dies möglich war, wurden die Daten vor der Berechnung der Anteile unterschiedlicher Antworten in die Anwortkategorien des Mikrozensus umkodiert.

Die festgestellten Abweichungen sind nur zu einem kleineren Teil auf empirische Veränderungen im Status der Befragten zurückzuführen, da der Abstand zwischen den beiden verwendeten Wellen nur rund zwei Monate beträgt. Bei einem Teil der Fragen (Geschlecht, Geburtsjahr) sind ohnehin keine Veränderungen im Zeitablauf möglich. Abweichungen durch unterschiedliche Antwortbereitschaft sind hingegen enthalten, da fehlende Werte nicht ausgeschlossen wurden. Verkodungs- und Übertragungsfehler dürften bei dieser mit besonderer Sorgfalt unter Einschaltung mehrfacher Kontrollen durchgeführten Untersuchung nahe Null sein. Der größte Teil der Abweichungen dürfte somit auf instabiles Antwortverhalten zurückzuführen sein, wie es durch die beschriebenen Einflüsse in der Erhebungssituation zustande kommt.

Tabelle 4.3: Anteil der Befragten der "ALLBUS" Test-Retest-Studie, die bei der *dritten* Welle andere Angaben machten als bei der *ersten* Welle (in Prozent)

Gemeindegröße	2,6
Regierungsbezirk	0,6
Geschlecht	0,0
Geburtsjahr	2,6
Familienstand	1,3
Heiratsjahr	9,1
Schule derzeit	1,3
Erwerbstätig	6,5
Arbeitslosigkeit	5,2
Früher erwerbstätig	8,4
Jahr, in dem zuletzt eine Erwerbstätigkeit ausgeübt wurde	24,7
Branche	11,0
Beruf	7,8
Stellung im Beruf	6,5
Arbeitswochenstunden	18,2
allgemeiner Schulabschluß	7,8
beruflicher Ausbildungsabschluß	16,2
überwiegende Einkünfte	13,0
Einkommen	44,8
Kinderzahl	7,1

Quelle: Eigene Auswertungen der Test-Retest-Studie

Tabelle 4.3 zeigt, daß bestimmte soziodemographische Merkmale wie Geschlecht, Familienstand, Alter und Regionalmerkmale sehr hohe Antwortstabilitäten aufweisen. Andererseits überrascht das Auftreten von immerhin 2,6 bzw. 7,1 Prozent Abweichungen bei den "trivialen" Kategorien Geburtsjahr und Kinderzahl.

Generell höhere Abweichungen sind bei den mit der Erwerbstätigkeit verknüpften Variablen festzustellen. Die Variable Einkommen ist demnach für Reidentifikationsversuche mit einfachen Abgleichsverfahren generell unbrauchbar. Hier spielen das Unwissen der Befragten und die tatsächlichen Einkommensvariationen im Zeitablauf eine Rolle. Die Angaben zum Jahr der letzten Erwerbstätigkeit (bei Nichterwerbstätigen) und zu den Arbeitswochenstunden zeigen ebenfalls relativ niedrige Stabilitäten.

4.3.3.2 Empirische Befunde für Erhebungen in der amtlichen Statistik

Bei den amtlichen Statistiken erfolgen die Entwicklung des Fragebogens und die Durchführung der Erhebung ebenfalls mit besonderer Sorgfalt. Außerdem ist zumindest nicht auszuschließen, daß sich die Befragten ihre Antworten im Fall einer amtlichen Erhebung besonders genau überlegen (Esser u.a. 1989). Dennoch ist auch in diesem Fall keine Reliabilität von 100 Prozent erzielbar.

Vom Statistischen Bundesamt wurden Untersuchungen zur Methode und Verläßlichkeit amtlicher Erhebungen unternommen, die mit der ALLBUS Test-Retest-Studie vergleichbar sind. Dabei wurden die Ergebnisse zweier Mikrozensuszusatzerhebungen, die im April und Juni 1970 durchgeführt wurden, mit denen der Volkszählung von 1970 auf der Basis von Individualdaten verglichen. Die Ergebnisse sind in Tabelle 4.4 eingetragen.

Tabelle 4.4 zeigt sehr niedrige Fehlerwerte für eine Reihe von Variablen. Insbesondere der Mikrozensus ist danach eine sehr zuverlässige Informationsquelle. Allerdings sind für einige Variablen, die bei der ALLBUS Test-Retest-Studie besonders starke Variabilität zeigten, keine Werte ausgewiesen. Da es in der Untersuchung, die die Basisdaten für die Tabelle 4.4 enthielt, um Datenfehler im engeren Sinne ging, wurden keine Werte für Merkmale ausgewiesen, die im Zeitablauf stärker variieren.

Obwohl für die einzelnen Quellen von Abweichungen, wie sie nach den Spalten der Tabelle 4.4 unterschieden werden können, nur sehr niedrige Prozentwerte ausgewiesen sind, erreichen sie in der Summe eine beachtliche Größenordnung (vgl. Spalte 1). Insgesamt werden erhebliche Unterschiede zwischen den Erhebungen deutlich, obwohl es sich in jedem Fall um die Erstellung einer amtlichen Statistik handelt, bei der die Bemühungen um die identische Abbildung von Sachverhalten aufgrund identischer Konzepte besonders groß und die zeitlichen Abstände der Erhebungen relativ gering sind.

Tabelle 4.4: Antwortvariabilität zwischen Volkszählung (VZ) und Mikrozensus (MZ)[18]

	Abweichungen zw. VZ und MZ		Abweichungen in (2) sind zuzurechnen[19]		fehlende Angaben	
	insg.	ohne fehl. Angaben	VZ	MZ	VZ	MZ
	(1)	(2)	(3)	(4)	(5)	(6)
	%	%	%	%	%	%
Geburtsjahr	3,3	3,3	2,7	0,6	0,0	0,0
Familienstand	2,1	1,4	0,6	0,8	0,7	0,0
Stellung innerhalb des Haushalts	8,4	5,0	4,4	0,6	0,7	2,0
Staatsangehörigkeit	1,2	0,5	4,4	0,6	0,7	2,0
Weiterer Wohnraum	8,3	3,5	2,1	1,4	1,9	2,6
Erwerbstätigkeit	7,4	5,2	3,5	1,7	2,2	0,0
Überwiegender Lebensunterhalt	11,3	9,4	5,1	4,3	1,6	0,3
Weitere Tätigkeit	2,4	2,4	1,9	0,5	-	-
Wirtschaftsbereich	10,2	4,1	2,7	1,5	3,6	2,5
Stellung im Beruf	10,0	4,0	2,9	1,1	3,5	2,5

Ähnliche Ergebnisse werden von der Statistikbehörde Großbritanniens berichtet. Im Rahmen einer Nacherhebung zum Zensus von 1981 wurden Anteile unterschiedlicher Klassifikation bei einzelnen Merkmalen im Zensus und in der Nacherhebung ermittelt.[20] Sie betrugen für das Merkmal Haushaltsgröße 2,4 Prozent, für Haus- oder Wohnungseigentum 3,2 Prozent, für die Anzahl von Autos im Besitz des Haushalts 3,5 Prozent, für Erwerbstätigkeit 7,8 Prozent, für

18 Die Tabelle enthält Prozentangaben auf der Basis der Antworten von 58.473 Personen. Quelle: Angaben in StBA (1978:67), sowie eigene Berechnungen auf dieser Grundlage.

19 Die Zurechnung von Abweichungen als "Fehler" des Mikrozensus (Spalte 4) erfolgt immer dann, wenn eine der beiden Mikrozensus-Wellen von der anderen und von der Volkszählung abwich. Als "Fehler" der Volkszählung (Spalte 3) wurde berücksichtigt, wenn die beiden Wellen des Mikrozensus übereinstimmende Werte erbrachten und lediglich die Volkszählung damit nicht übereinstimmte. Zu diesem Fehler haben demnach auch Unterschiede zwischen den Erhebungstypen VZ und MZ beigetragen. Wichen alle drei Erhebungen voneinander ab, wurden anteilig Zuschläge zu den beiden Fehlertypen (Spalte 3 und 4) vorgenommen.

die Zugehörigkeit zu grob definierten Berufsklassen 13 Prozent. Je detaillierter das Kategorienschema war, mit dem der Beruf erfaßt wurde, umso größer war der Anteil unterschiedlicher Einordnungen in beiden Erhebungen. Er betrug 13 Prozent bei einem Schema mit 6 Kategorien, 16 Prozent bei 16 Kategorien, 22 Prozent bei 350 Kategorien. Seltene, wenig besetzte Kategorien wiesen höhere Anteile unterschiedlicher Klassifikationen auf. Nach diesen Befunden würden also differenzierte Klassifikationsschemata, die zwar hilfreich sind, um zu einzigartigen Ausprägungskombinationen zu gelangen, gleichzeitig einen geringeren Grad an Reliabilität aufweisen.

4.3.3.3 Quantitative Schätzung der Bedeutung von Dateninkompatibilitäten für das Reidentifikationsrisiko

Wir können an dieser Stelle nun auf die zu Beginn dieses Kapitels für alle denkbaren Deanonymisierungsszenarien hervorgehobene Bedeutung von Dateninkompatibilitäten für das Reidentifikationsrisiko zurückkommen. Für alle in Übersicht 4.1 (S. 95) zusammengefaßten Deanonymisierungssituationen steht das Reidentifikationsrisiko in einer direkten Beziehung zur Wahrscheinlichkeit, daß die Ausprägungen der Überschneidungsmerkmale im Mikrodatenfile und im Identifikationsfile vollständig kompatibel abgebildet sind. Dies gilt zumindest für die mit wenig Aufwand handhabbare und damit attraktive Reidentifikationstechnik des Abgleichs auf Identität von Merkmalsausprägungen. Diese Technik wird durch Dateninkompatibilitäten in einer doppelten Weise wirksam gestört:

- Damit können zum einen - bei Erfüllung aller anderen in Abschnitt 4.1 genannten Bedingungen - nur Datensätze reidentifiziert werden, bei denen alle Merkmale vollständig kompatibel abgebildet sind.

- Zum anderen kann es bei nicht-kompatibler Abbildung zu Falschzuordnungen kommen. Im Mikrodatenfile und Identifikationsfile kann es in der Folge nicht-kompatibler Abbildungen zwar zu identischen einzigartigen Ausprägungskombinationen kommen, die sich aber auf unterschiedliche Fälle in der Grundgesamtheit beziehen. Die scheinbare Reidentifikation ist damit fehlerhaft.

---

20 Die Ergebnisse sind hier zitiert nach Skinner et al. 1990.

Die für das Reidentifikationsrisiko zentrale Größe der Wahrscheinlichkeit vollständig kompatibler Abbildung kann hier nicht exakt empirisch bestimmt werden. Dafür müßten viel umfassendere empirische Untersuchungen über den Grad der Kompatibilität einzelner Merkmale zwischen spezifizierten Identifikationsfiles und Mikrodatenfiles vorliegen. Unter vereinfachenden Modellannahmen kann diese Wahrscheinlichkeit jedoch auf die folgende Weise theoretisch bestimmt werden: Sei $p(V_i)$ die Wahrscheinlichkeit des Auftretens einer Inkompatibilität bei der Variable (Überschneidungsmerkmal) V_i (i=1,..,I) und I sei die Gesamtzahl aller verwendeten Überschneidungsmerkmale. Es sei angenommen, daß die Wahrscheinlichkeit eines Datenfehlers beim Merkmal V_i unabhängig von jener beim Merkmal V_j ist, d. h.:

$$p(V_i \mid V_j) = p(V_i).$$

Bezeichnet WCK die Wahrscheinlichkeit, daß ein beliebiger Datensatz von keiner Inkompatibilität betroffen ist, so gilt:

$$WCK = (1-p(V_1))(1-p(V_2))(1-p(V_3))(1-p(V_4)) \ldots (1-p(V_I))$$

oder

$$WCK = \prod_{i=1}^{I} (1-p(V_i)) .$$

Wenn $p(V_i) = p(V)$, für alle i=1,...,I, d. h. die Fehlerwahrscheinlichkeit für alle Variablen konstant ist, so gilt:

$$WCK = (1-p(V))^I .$$

Ist die Stichprobe groß, kann WCK als der Anteil der Datensätze interpretiert werden, der von keiner Dateninkompatibilität betroffen ist. Nur solche Datensätze können mittels einfacher Abgleichsverfahren zugeordnet werden.

Dazu können nun einige Beispiele berechnet werden. Es sei von dem einfachen Fall ausgegangen, daß die Wahrscheinlichkeit einer Inkompatibilität für alle Merkmale konstant gleich p(V) = 0,10 ist. Dann hängt WCK lediglich von der Anzahl I der einbezogenen Variablen ab. WCK beträgt:

bei 10 Merkmalen: 0,3487;
bei 20 Merkmalen: 0,1216;
bei 30 Merkmalen: 0,0424.

Benutzt ein Angreifer zehn Merkmale, wird er bei der durchschnittlichen Inkompatibilität von zehn Prozent pro Merkmal nur noch jeden dritten Datensatz korrekt zuordnen können.[21] Bei zehn Merkmalen kann es jedoch sein, daß die Bedingung der Einzigartigkeit der Ausprägungskombination im Datenbestand oder in der Population nicht erfüllt ist. Versucht der Angreifer, dies durch Erhöhung der Variablenzahl zu umgehen und stehen ihm Daten mit entsprechend vielen Überschneidungsmerkmalen zur Verfügung, so fällt der Anteil der kompatiblen Datensätze drastisch. Bei 30 Variablen ist nur noch für jeden 24. Fall zu erwarten, daß alle Merkmale vollständig kompatibel abgebildet sind.

Zur Abschätzung der Wahrscheinlichkeit WCK können die in Tabelle 4.3 aufgelisteten Angaben aus der ALLBUS Test-Retest Studie verwendet werden. Für $p(V_i)$ werden dann die Angaben zu den einzelnen Variablen eingesetzt. Dabei findet einerseits eine Überschätzung der Fehlerhäufigkeit statt: Die verschiedenen Fehler sind nicht unabhängig voneinander, sondern korrelieren positiv.

Andererseits sind in Tabelle 4.3 nur wenige Quellen von Inkompatibilitäten enthalten. Der zeitliche Abstand zwischen den beiden betrachteten Wellen ist denkbar gering. Da zweimal mit der gleichen Untersuchungsanlage die gleichen Fragen gestellt wurden, gibt es keine Schwierigkeiten mit unterschiedlichen Interviewsituationen, Frageformulierungen, Transformierungen von Kategorien ineinander usw. Insofern ist die auf dieser Basis erfolgte Berechnung von WCK lediglich als Anhaltspunkt zu verstehen. Der Wert bei einem realen Deanonymisierungsversuch dürfte vermutlich darunter liegen.

Werden alle 20 Variablen verwendet, beträgt der Anteil unveränderter Datensätze:

$$WCK_{20} = 0{,}11.$$

---

21 Hier wird ein Selektionsverfahren in seiner einfachsten Form zugrundegelegt. Es sind Strategien denkbar, bei denen der Angreifer versucht, mit dem Vorhandensein von Fehlern zurechtzukommen. Ist er sich über die exakte Ausprägung des Wertes einer Variablen nicht im klaren, kann er verschiedene Alternativausprägungen berücksichtigen. Dies wird bei der experimentellen Anwendung einfacher Sortierverfahren versucht werden (vgl. Kapitel 9). Die Anwendung dieser Strategie stößt jedoch an Grenzen, weil eine Unsicherheit über die "richtige" Ausprägung verbleibt und weil bereits die Berücksichtigung nur weniger Alternativen zur Explosion des Rechenaufwandes führt.

Im Durchschnitt ist also nur jeder neunte Datensatz nicht von einer Inkompatibilität betroffen. Wird die Einkommensvariable ausgeklammert, da sie sich als besonders instabil erwies, erhöht sich der Wert für WCK auf:

$$WCK_{19} = 0{,}20.$$

Es sei jedoch noch einmal betont, daß es sich hier nur um eine Modellrechnung handelt. Sie zeigt jedoch an, daß allein aus dem Verzicht auf die Annahme vollständig kompatibler Daten eine erhebliche Reduzierung des Reidentifikationsrisikos resultiert.

4.3.4 Folgerungen für die Erfolgsaussichten und den Aufwand von Reidentifikationsversuchen

Die vorgelegte Analyse zeigt, daß Dateninkompatibilitäten eine wichtige Barriere für Deanonymisierungsversuche sein können. Die 'naive' Anwendung von einfachen Datenabgleichsverfahren wird empfindlich gestört: Eindeutige Zuordnungen setzen die Einbeziehung von möglichst vielen Überschneidungsmerkmalen voraus. Je mehr Merkmale jedoch zur Verfügung stehen, desto höher ist die Wahrscheinlichkeit, daß eine Dateninkompatibilität den Suchprozeß vereitelt, da derartige Verfahren die Identität der Merkmalsausprägungen in Identifikationsfile und Mikrodatenfile voraussetzen. Datenfehler können also Zuordnungen verhindern.

Dateninkompatibilitäten können andererseits auch dazu führen, daß Falschzuordnungen vorgenommen werden. Gerade dann, wenn die Kompatibilität gering ist, dürfte es zu einer großen Zahl zufälliger Zuordnungen kommen, die jedoch falsch sind. Ein Datenangreifer wäre dann mit einer Situation konfrontiert, bei der er unter Umständen eine größere Zahl von Fällen scheinbar identifiziert hat, von denen er jedoch nicht weiß, welche richtig und welche falsch zugeordnet wurden. Dateninkompatibilitäten verringern nicht nur die Möglichkeit von Reidentifikationen schlechthin, sie vermindern auch die Sicherheit des Angreifers über die Korrektheit von Zuordnungen.

Die Analyse zeigt auch, daß mit der Vernachlässigung der Rolle von Dateninkompatibilitäten in der bisherigen Deanonymisierungsforschung wohl einer unrealistischen Einschätzung des Reidentifikationsrisikos Vorschub geleistet wurde. Es ist deshalb wichtig, diesen Aspekt in den eigenen Untersuchungen besonders aufmerksam zu verfolgen.

5 Deanonymisierungsmotive*

Bei der Prüfung der Unverhältnismäßigkeit des Aufwandes (an Zeit, Kosten und Arbeitskraft), den der Gesetzgeber in §16 Abs.6 BStatG von 1987 als Voraussetzung der Weitergabe von Einzeldaten der amtlichen Statistik an die Wissenschaft bestimmt hat, ist offensichtlich als ein wesentlicher Teilaspekt der Nutzen zu berücksichtigen, den ein Wissenschaftler von deanonymisierten Daten haben könnte. Ist dieser Nutzen hoch, sind die Daten eher 'gefährdet', dann lohnt es sich, auch aufwendige Verfahren zur Deanonymisierung einzusetzen.

Um darüber Aufschluß zu gewinnen, werden in diesem Kapitel zunächst Ziele und Maßstäbe des professionellen Handelns der solche Daten nutzenden Sozialwissenschaftler untersucht. Es wird also danach gefragt, welcher "Logik" ihr Handeln folgt, um potentielle Motive für eine Deanonymisierung aufzufinden, die sich aus den typischen Handlungsorientierungen des Personenkreises und der Institutionen ergeben könnten, welche im Gesetz bezüglich einer Weitergabe faktisch anonymer Daten vorgesehen sind. Die Beschränkung auf die professionellen Motivationsstrukturen impliziert, daß dabei rein persönliche Motive nicht analysiert werden.

Wenn sich dabei herausstellen sollte, daß in der Handlungslogik der Wissenschaft selbst keine oder nur geringe Anreize für einen Versuch der Deanonymisierung von überlassenen Mikrodaten festzumachen sind, dann ist allerdings zu beachten, - und darauf beziehen sich entsprechende Befürchtungen in der öffentlichen Diskussion vor allem - daß ein Wissenschaftler oder eine andere Person, die im Wissenschaftsbereich Zugang zu anonymisierten Mikrodaten hat, auch aus nicht berufsbezogenen Motiven ein Interesse an der Deanonymisierung von Daten haben könnte. In einem zweiten Abschnitt sollen deshalb explizit auch solche außerberuflichen Motive und Nutzenelemete untersucht werden.

---

* Autoren: Uwe Blien, Walter Müller (5.1)
Peter Knoche (5.2)

5.1 Die professionelle Handlungslogik der Sozialwissenschaften

Was nun zunächst die Motive betrifft, die sich aus der spezifischen Interessenlage des forschenden Wissenschaftlers als Wissenschaftler ergeben, so ist es sinnvoll, nur solche Bereiche der Forschung zu betrachten, bei denen überhaupt ein Mindestmaß an Interesse für die Art von Daten besteht, die in der amtlichen Statistik typischerweise vorhanden sind. Dies bedeutet in der Konsequenz, daß sich die folgenden Erörterungen an quantitativen Forschungsmethoden orientieren.

Darum ist es möglich, in unserem Zusammenhang den Umstand zu vernachlässigen, daß es bei qualitativen Untersuchungen[2] zu spezifischen Kollisionen des beruflichen Interesses mit Datenschutzgesichtspunkten führen könnte: Ist es die Absicht des Forschers, Fallstudien anzufertigen[3], in denen die besonderen Lebensumstände oder Einstellungsstrukturen einzelner Personen eine wesentliche Rolle spielen, so wird er diese auch im individuellen Detail dokumentieren oder veröffentlichen wollen. Damit besteht eine forschungsimmanente Notwendigkeit der Publikation von Einzelfallinformationen, die mit Datenschutzinteressen durchaus in Konflikt stehen kann. Derartige Probleme sind aber im Zusammenhang der "Massendaten" der amtlichen Statistik und der typischen Art ihrer Verwendung und Veröffentlichung nicht gegeben.

Darüber hinaus wird auch auf eine genauere Untersuchung der Forschung in der kommerziellen Markt- und Meinungsforschung verzichtet, weil im Gesetz die Datenweitergabe mit faktisch anonymen Daten an diese Stellen nicht vorgesehen ist und sich deshalb eine Erörterung erübrigt. Wenn im folgenden von 'Wissenschaft' und 'Forschung' die Rede ist, so wird unter dem institutionellen

---

2 In Analysen mit qualitativer Ausrichtung (für eine Lehrbuchdarstellung vgl. Lamnek 1988, 1989) wird eine viel intensivere und detaillliertere Annäherung an persönliche und individuelle Aspekte der unterstellten Realität angestrebt als mit Massenerhebungen überhaupt möglich ist. Dabei werden Probleme untersucht, von denen der Forscher annimmt, daß sie einer quantifizierenden Sichtweise unzugänglich sind. Als Forschungsmethode spielt vielfach die Interpretation und Analyse von unstrukturierten Texten eine zentrale Rolle (vgl. Fuchs 1984:126ff., Lamnek 1989:35ff.). Für Untersuchungen der qualitativen Art ist die Information der amtlichen Statistiken somit allenfalls als Hintergrundinformation interessant. Dafür reichen im allgemeinen die veröffentlichten Tabellen der offiziellen Statistik aus.

3 Dies wird generell in der qualitativen Sozialforschung angestrebt (vgl. z. B. Lamnek 1989:4) und hat in verschiedenen speziellen Soziologien eine lange Tradition (für die Industriesoziologie vgl. z. B. die 'klassische' Hawthorne-Untersuchung, Roethlisberger/Dickson 1939).

Aspekt demgemäß an "Hochschulen oder sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung" (§16 Abs.6 BStatG) gedacht. Auch die eigene Forschung der Statistischen Ämter wird nicht betrachtet, da diese einen unmittelbaren Datenzugang hat.

Im folgenden wird zwar besonders häufig auf Beispiele aus der Soziologie verwiesen, die vorgetragene Argumentation bezieht sich jedoch - abgesehen von den genannten Einschränkungen - ganz allgemein auf empirisch orientierte Sozialwissenschaften (also auch z. B. auf die politische Wissenschaft, die quantitative Ökonomie, die Sozialgeographie oder die Bevölkerungswissenschaft).

5.1.1 Ziele der Sozialwissenschaften

Die quantitativ orientierte wissenschaftliche Forschung ist darauf ausgerichtet, die Merkmale und den Wandel sozialer Einheiten und sozialer Beziehungen zu beschreiben und zu erklären. Besonderes Charakteristikum des wissenschaftlichen Vorgehens ist dabei das Streben nach Reproduzierbarkeit der Ergebnisse (vgl. z.B. Boudon 1980:182f., Schnell/Hill/Esser 1988), nach Abstraktion vom Einzelfall, nach Generalisierungen, also die Suche nach Gesetzmäßigkeiten und die Etablierung möglichst allgemein gültiger Theorien. Diese können dann für die Erklärung konkreter Gegebenheiten verwendet werden, die unter spezifischen historischen Bedingungen vorfindbar sind und zur Erstellung von Prognosen dienen.

Im sozialwissenschaftlichen Bereich ist hinsichtlich der grundsätzlichen Ausrichtung lediglich umstritten, ob das Ziel sein sollte, Gesetzmäßigkeiten aufzudecken, die unabhängig von einer historischen Situation gelten,[4] oder ob solche Gesetze im sozialen Bereich - etwa im Unterschied zu der Realität der Naturwissenschaften - entweder nicht existieren oder wenigstens beim heutigen Stand der Forschung noch nicht aufgefunden werden können. Sollte dies der Fall sein, muß sich der Wissenschaftler darauf konzentrieren, Zusammenhänge

---

4 Eine solche Haltung wird oft auf Arbeiten des "Kritischen Rationalismus" gestützt, dessen Position gerade von Sozialwissenschaftlern vertreten wird, die Forschung mit quantitativen Methoden betreiben (vgl. z.B. Opp 1976). Der Kritische Rationalismus sieht als Ziel der Wissenschaft, allgemeine Erklärungen aufzufinden und diese, in Form eines "deduktiven Systems", einer Theorie zuzuordnen (vgl. Popper 1972:117). Auch in diesem Zusammenhang wird also Abstraktion von individuellen Einzelfällen gefordert.

in einer bestimmten Gesellschaft, in einer spezifischen Situation nachzuvollziehen.[5]

Allerdings ist anzumerken, daß nicht jede Art wissenschaftlicher Forschung auf Erklärung ausgerichtet ist. Das Ziel kann ebenfalls sein, kulturelle, gesellschaftliche, wirtschaftliche, demographische oder politische Zustände zu beschreiben und damit transparent zu machen. Dies ist oft der Zweck von Analysen, die u.a. für die staatliche Planung wichtig sind, z. B. wenn untersucht wird, wie sich der Anteil der Gymnasiasten an der Bevölkerung einer bestimmten Altersgruppe im Zeitablauf entwickelt (vgl. Blien/Tessaring 1988, Tessaring et al. 1990). Aber auch solche Fragestellungen sind nicht am persönlichen Einzelfall, sondern an verallgemeinernden Aussagen interessiert.

Die Methode der Abstraktion hat demnach eine Implikation, die von zentraler Bedeutung für die vorliegende Problemstellung ist: Da das Ziel des Forschers ist, allgemeine oder historisch-situative Regelmäßigkeiten aufzufinden, richtet sich sein Interesse nicht auf einzelne identifizierbare Personen in ihrer individuellen Besonderheit, sondern in ihrer Eigenschaft als bloße Merkmalsträger[6], von deren persönlichen Spezifika gerade abzusehen ist. Auch dort, wo die Deskription und nicht die Erklärung im Vordergrund steht, interessieren sich Sozialwissenschaftler nicht für einzelne Personen per se, sondern nur insoweit, als aus den Merkmalen einer Mehrzahl von Einzelpersonen Verallgemeinerungen abgeleitet werden können. Aus den Zielen der wissenschaftlichen Tätigkeit ergibt sich somit keine Notwendigkeit einer namentlichen Identifikation der einzelnen Personen, die Gegenstand der Forschung sind.

Die Spezifika der Logik des wissenschaftlichen Forschungsprozesses können im Kontrast zu der Verfahrensweise von staatlichen Verwaltungen verdeutlicht werden (vgl. auch Bayer 1987:93). Generell ist für die tägliche Routine in Ämtern die Prüfung und Abwicklung von Einzelfällen charakteristisch, zu denen auch solcherart differenzierte Informationen vorhanden sein müssen. Darum ist

---

5 Eine dieser Haltung entsprechende wissenschaftstheoretische Position wird von der kritischen Theorie eingenommen. Dort wird die historische Dimension der die sozialen Beziehungen der Menschen betreffenden "Gesetze" oder Regelmäßigkeiten betont (vgl. Adorno 1972a, 1972b). Die verschiedenen wissenschaftstheoretischen Positionen weichen nicht prinzipiell voneinander ab, was die Forderung nach Abstraktion betrifft.

6 Dies wurde vom Bundesverfassungsgericht in dem Urteil zur Volkszählung ausdrücklich anerkannt: "... der Wissenschaftler ist regelmäßig nicht an der einzelnen Person interessiert, sondern an dem Individuum als Träger bestimmter Merkmale" (1984:69). Für ähnliche Aussagen aus dem Bereich der Wissenschaft vgl. Scheuch (1980:253), Müller/Hauser (1987:64) und Kaase (1987:72).

eine exakte Zuordnung der Daten zu den betreffenden Personen im allgemeinen notwendig und gehört die Suche nach Angaben über die jeweiligen 'Klienten' zur natürlichen Aufgabe der Behörde. Die Verwendung der Informationen im praktischen Handeln der betreffenden Behörde kann mit Vorteilen oder Nachteilen für den involvierten Staatsbürger verbunden sein, je nachdem, ob ihm eher Rechte oder eher Pflichten erwachsen oder ob er - als Konsequenz eines Gesetzesverstoßes - sogar mit Sanktionen rechnen muß. Die Daten der Verwaltung stellen somit oft einen Bestand an sogenanntem "Interventionswissen" (Scheuch 1980:253) gegenüber einzelnen Personen dar.

Wie sieht demgegenüber die typische Arbeitsweise von Sozialwissenschaftlern aus? Für eine Projektgruppe, die - etwa im unmittelbaren Auftrag eines Finanzministeriums - die Konsequenzen bestimmter staatlicher Maßnahmen für die Steuereinnahmen oder die betroffenen Bürger erforscht, ist das identifizierbare Individuum ohne Belang.[7] Die Fragestellung verlangt geradezu die Abstraktion von den einzelnen Personen, um den generellen Zusammenhang aufzufinden. Der betreffende Projektleiter käme gar nicht auf die Idee, eine Liste identifizierbarer Personen als Ergebnis seiner Untersuchung vorzulegen.

5.1.2 Praxis und Methoden empirischer Untersuchungen in den Sozialwissenschaften

Die Ziele und Erkenntnisinteressen der Sozialwissenschaft werden in der Forschungspraxis in konkrete Projekte umgesetzt. Auf dieser Ebene werden Daten erhoben und sind demgemäß Konflikte mit den Notwendigkeiten des Datenschutzes möglich. Darum soll hier der typische Ablauf einer sozialwissenschaftlichen Untersuchung kurz dargestellt werden. Dies geschieht auf der Grundlage von Empfehlungen der entsprechenden wissenschaftlichen Lehrbücher[8] und von Erfahrungen, die beim Zentrum für Umfragen, Methoden und

7 Eine typische Fragestellung für sozialwissenschaftliche Untersuchungen des benannten Problemkreises ist z. B. jene nach der Auswirkung von Steuersatzänderungen auf das Einkommen von Familien (vgl. Lietmeyer 1986:149). Es ist leicht zu erkennen, daß dies eine generelle Frage ist, bei deren Beantwortung es nicht auf identifizierbare Personen ankommt. Im übrigen sind viele Bereiche der öffentlichen Verwaltung ebenfalls nicht an Einzelpersonen interessiert. Dies gilt z. B. für den gesamten Bereich der amtlichen Statistik.

8 Vgl. z. B. Friedrichs (1980:51 und 119), Drew (1980:67ff.), Kriz (1981:37), Alemann (1977:81), Bortz (1984), Roth (1984:86f.) in Berufung auf Bunge (1967) und Schnell/Hill/Esser (1988:110ff.).

Analysen in Mannheim gesammelt wurden. Da diese Institution auf Beratungsleistungen für die Forschungspraxis spezialisiert ist, liegen dort vielfältige Erfahrungen hinsichtlich der Planung und Durchführung von empirischen Projekten vor.[9]

1) Jede Untersuchung beginnt mit einer Fragestellung. Diese hat im allgemeinen einen Bezug zu einer sozialwissenschaftlichen Theorie, die ihrerseits mehr oder minder ausgebildet sein kann.

2) Ist auf Basis dessen, was der Wissenschaftler für gesicherte Erkenntnis hält, keine Lösung des Ausgangsproblems möglich, so wird er versuchen, einen empirischen Bezug herzustellen, um eine Prüfung seiner Fragestellung zu ermöglichen. Die Spezifikation von Hypothesen bzw. konkretisierten Forschungsfragestellungen schließt sich an. Das Ausgangsproblem wird soweit eingeengt, daß es einer Bearbeitung fähig wird.

3) Als nächstes wird der Forscher nach Daten suchen, die ihm bei der Beantwortung seiner Frage behilflich sein könnten. Stehen keine geeigneten Daten zur Verfügung, ist an eine Neuerhebung zu denken. Dazu muß zunächst ein genauer Plan, ein Untersuchungsdesign, entworfen werden, in dem auch geeignete Erhebungsinstrumente festgelegt werden. Die in den Sozialwissenschaften meistverwendete Form der Erhebung ist die Befragung.

4) Im nächsten Schritt wird die Datenerhebung selbst durchgeführt. Das dabei gewonnene Datenmaterial wird aufbereitet und oftmals per EDV abgespeichert.

5) In den meisten Fällen wird eine umfangreiche quantitative Auswertung der dabei erhaltenen Informationen erforderlich. Dazu stehen statistische Methoden ganz unterschiedlicher Komplexität zur Verfügung. Dies können einfache Häufigkeitsauszählungen sein wie auch sogenannte multivariate Verfahren, die der Analyse komplexer Abhängigkeitsbeziehungen zwischen mehreren Merkmalen dienen.

   Dabei wird jeweils nach Antworten auf die Fragestellung gesucht, die Anlaß der Untersuchung war. Ist das Ziel der Nachweis eines Zusammenhangs, wird der Forscher bestrebt sein, diesen anhand der Daten zu

---

[9] Für ein Lehrbuch eines ZUMA-Mitarbeiters, das Erfahrungen mit praktischer Umfrageforschung reflektiert, vgl. Porst (1985).

demonstrieren. Umgekehrt kann auch beabsichtigt sein, eine in der wissenschaftlichen Diskussion wichtige These mit empirischen Ergebnissen zu widerlegen. Dann wird der Forscher anhand der Daten zu zeigen versuchen, daß eine vorgeschlagene Hypothese keine adäquate Darstellung der empirischen Verhältnisse ist.

6) Aus den Ergebnissen können Folgerungen für die Theorie oder allgemeiner, den theoretischen Zusammenhang, aus dem heraus die Forschungsfrage gestellt wurde, entwickelt werden. In der Theorie umstrittene Probleme werden besonders häufig empirisch untersucht. Dabei wird analysiert, welche der konkurrierenden Theorien den erhaltenen Befund besser erklären kann.

Die dargelegte Reihenfolge der einzelnen Schritte im Forschungsprozeß wird oft nicht strikt eingehalten. So ist z. B. denkbar, daß die Struktur bereits vorhandener Daten eine theoretische Hypothese nahelegt. Im vorliegenden Kontext ist von zentraler Bedeutung, daß schon an dem typischen Ablauf eines Forschungsvorhabens die Absicht des Forschers erkennbar ist, verallgemeinerbare Erkenntnisse zu gewinnen. Beabsichtigt wird, Aufschluß über die Gültigkeit von Theorien zu gewinnen oder gesellschaftliche Zustände in ihren Bedingungszusammenhängen zu analysieren.

Einzelne Personen spielen dabei nur insofern eine wichtige Rolle, als Daten bei ihnen erhoben werden (vgl. Mohler 1987:129). Während der Erhebungsphase einer Untersuchung ist ein Personenbezug (oft) wichtig und unerläßlich. In vielen Fällen kann die Aussagekraft der Untersuchung nämlich nur gewährleistet werden, wenn von identifizierbaren Personen ausgegangen wird, indem aus einem Register oder einer Kartei über eine Zufallsauswahl eine Stichprobe gezogen wird, um die betreffenden Personen dann aufzusuchen und zu befragen.

Oft wird es notwendig sein, den Personenbezug eine gewisse Zeit beizubehalten, um Rückfragen bei unvollständig oder widersprüchlich ausgefüllten Fragebögen zu ermöglichen (vgl. Mohler/Kaase 1980:69ff.). Dies ist in besonderem Maße bei Panelerhebungen der Fall, bei denen die gleichen Personen im Zeitablauf wiederholt befragt werden und die Daten der einzelnen Befragungen zusammengeführt werden müssen, damit Veränderungen in der Zeit beobachtet werden können.

Beim Problem der Weitergabe von Daten der amtlichen Statistik an die Wissenschaft liegt jedoch eine völlig anders gelagerte Situation vor. Hier wird die Datenerhebung nicht vom Forscher selbst oder in seinem Auftrag durchgeführt. Die Wissenschaft erhält Daten, bei denen die für eine Analyse notwendige Datenaufbereitung bereits abgeschlossen ist. Deshalb ist ein Personenbezug überhaupt nicht erforderlich.

Der auf die Erhebung und Datenaufbereitung folgende Schritt im Rahmen einer sozialwissenschaftlichen Untersuchung, die statistische Analyse, setzt nämlich bereits die Abstraktion von Individuen voraus. Deshalb sind anonyme Daten völlig hinreichend. Deren Auswertung dient geradezu dem Zweck, die enorme Vielfalt der in ihnen enthaltenen Information zu kondensieren, um Kennziffern zu erhalten, die Aussagen über den jeweils interessierenden Zustand oder Zusammenhang ermöglichen.

Wird z. B. die Beziehung zwischen dem Qualifikationsniveau und dem Risiko, arbeitslos zu werden (vgl. z. B. Tessaring 1988), untersucht, so kann die Verteilung in der verwendeten Stichprobe an Hand einer Tabelle - d. h. an Aggregatdaten, nicht an Einzelfällen - im Überblick geprüft werden. Anschließend wird der Forscher in der Regel statistische Kennziffern berechnen, die die Stärke des Zusammenhangs verdeutlichen und zeigen, ob die Beziehung systematisch oder rein zufälliger Natur ist.

Wenn für diese Zwecke der Bezug zu einer identifizierbaren Person auch völlig unerheblich ist, so ist dennoch in aller Regel wichtig, daß für die Analysen auf Einzelangaben der (nicht-identifizierten) Individuen zurückgegriffen werden kann (vgl. dazu Müller/Hauser 1987 und Blien/Papastefanou 1988), da viele multivariate statistische Methoden nur dann angewendet werden können, und die volle Freiheit, Daten in bestimmter Weise abzugrenzen, zu ordnen und zu filtern, nur so besteht.

Als wichtiger Befund dieses Abschnitts kann also festgehalten werden, daß der Forscher zwar Einzelangaben benötigt, daß aber in dem Fall, in dem er diese aus dem Datenbestand der statistischen Ämter erhält, die Daten selbst keinen Personenbezug aufweisen müssen. Für die Nutzung der Daten zum Zweck ihrer Analyse für eigene wissenschaftliche Fragestellungen ist ein Personenbezug weder erforderlich noch hätte der Forscher irgendeinen Nutzen für seine Arbeit, wenn ein Personenbezug vorhanden oder herstellbar wäre.

5.1.3 Zur Organisation der wissenschaftlichen Forschung und der Sozialisation von Forschern

Die dargestellten Prinzipien wissenschaftlicher Forschung finden in der Organisation der Forschungspraxis und der Lehre ihren konkreten Niederschlag. Ein Blick in eines der einschlägigen Lehrbücher belegt, daß die Vermittlung des beschriebenen Umgangs mit Daten und der Bedeutung der Abstraktion von zentraler Bedeutung ist.[10] Die Kenntnis der entsprechenden Methoden und die Fähigkeit zur Abstraktion wird in Examina geprüft und nach dem Studium insbesondere bei jenen Personen vorausgesetzt, die für die wissenschaftliche Weiterqualifikation und die Forschung an den Universitäten und vergleichbaren Forschungsinstitutionen rekrutiert werden.

Im Verlaufe der weiteren wissenschaftlichen Ausbildung und der Einmündung in eine berufliche Laufbahn in der Wissenschaft ist der berufliche Erfolg an die Beachtung der genannten Maßstäbe geknüpft. Veröffentlichungen in den relevanten Fachzeitschriften und die Durchführung von Forschungsprojekten unterliegen im allgemeinen einer Begutachtung durch Wissenschaftler mit bekanntem Namen. Forschungsanträge und Zwischenberichte zum Stand laufender Projekte an Institutionen, die Forschung fördern und finanzieren, werden in entsprechender Weise beurteilt. In der wissenschaftsinternen Fachdiskussion werden ebenfalls die benannten Maßstäbe angelegt.

Der mit dem Versuch einer Deanonymisierung verbundene Aufwand an Kosten und Zeit wäre demzufolge für die Karriere in der Wissenschaft in aller Regel kontraproduktiv, da er im Sinne der üblichen Ziele der Wissenschaft keinen Fortschritt bringen würde.

Die Bedeutung des Datenschutzgesichtspunkts wird den Studierenden der Sozialwissenschaften während der Ausbildung vermittelt. Als Beispiel für ein relevantes Lehrbuch sei Bortz zitiert: "Wenn die Anonymität der persönlichen Angaben nicht gewährleistet werden kann, sollte auf eine empirische Untersuchung verzichtet werden. Jedem Untersuchungsteilnehmer muß versichert werden, daß die persönlichen Daten nur zu wissenschaftlichen Zwecken verwendet und daß die Namen nicht registriert werden. Auskünfte über andere Personen unterliegen dem *Datenschutz*. Vor größeren Erhebungen, in denen auch persönliche Angaben erfragt werden, empfiehlt es sich, die entsprechen-

---

10 Vgl. die in Fußnote 8 genannten Lehrbücher.

den rechtlichen Bestimmungen einzusehen" (1984:22, Hervorhebung im Original).[11]

Mit der Einhaltung der Schutzbestimmungen verbinden sich spezifische Interessen. Für die Arbeit der Sozialwissenschaftler ist nämlich wichtig, daß sich die bei Forschungsvorhaben nach bestimmten Verfahren ausgewählten Personen freiwillig beteiligen und dabei wahrheitsgemäße Angaben zu oft sehr persönlichen Sachverhalten machen. Dies wird nur der Fall sein, wenn die Befragten nicht den Eindruck haben, daß ihre Daten an Dritte weitergegeben oder in anderer Weise zweckentfremdet werden.[12]

Darum sind Ereignisse, die das Verhältnis zwischen der Forschung und ihren Informanten belasten könnten, als schädlich für die Interessen der Forschung einzustufen, da sie dazu beitragen, die Wissenschaft tendenziell von ihrer empirischen Basis abzuschneiden. Ein durch die Aufdeckung eines Deanonymisierungsversuchs entfachter Skandal wäre den Forschungsinteressen abträglich. Der betreffende Wissenschaftler müßte nicht nur den Vertrauensverlust in der Öffentlichkeit einrechnen, der ihm bei einer erneuten Befragung Schwierigkeiten machen würde. Er hätte das zusätzliche Problem, daß Institutionen, die auf ihre Seriosität Wert legen, den Kontakt mit ihm abbrechen und ihm auch keine Daten mehr zur Verfügung stellen.

Unter diesen Voraussetzungen mag die in der Vergangenheit bei einem Teil der Wissenschaftler zu beobachtende, etwas naive Vorstellung (vgl. Eichener 1988:68) von der absoluten Harmlosigkeit des eigenen Tuns erstaunlich sein. Sie war aber hauptsächlich die Konsequenz des hier dargelegten Desinteresses an Einzelpersonen. Weil die anonyme und nicht an der persönlichen Identifikation orientierte Behandlung der Daten selbstverständlich ist, war in der Vergangenheit in vereinzelten Fällen eine gewisse Unbekümmertheit hinsichtlich der konkreten Maßnahmen des Datenschutzes festzustellen. Verschiedentlich fehlte das Bewußtsein, daß dafür besondere Sicherungsmaßnahmen getroffen werden müßten. Die Selbstverständlichkeit, daß im Kontext der Wissenschaft Daten ohnehin vertraulich behandelt werden - im Verein mit technischen Schwierigkeiten, Unbequemlichkeiten und Kosten, die für den Forscher manchmal mit Datenschutzmaßnahmen verbunden sind -, und nicht das anders gelagerte Interesse des Wissenschaftlers oder individueller böser Wille erklärt wohl auch Verstöße, die in der Vergangenheit gegen die Bestimmungen des

---

11 Zu weiteren Darstellungen und "Belehrungen" zur Datenschutzproblematik in der Literatur vgl. Friedrichs (1980:40) und Drew (1980:47f.).

12 Für den Fall der Kriminologie vgl. Heinz (1987:155).

Datenschutzes zu beobachten waren.[13] Zu denkbaren Einschränkungen dieser allgemeinen Aussage wird weiter unten noch Stellung genommen.

Die öffentliche Datenschutzdiskussion der letzten Jahre hat die Sensibilität unter den Wissenschaftlern erhöht. In einer Reihe von wichtigen Fachzeitschriften[14] wurde die Datenschutzproblematik aufgegriffen und das Handeln der Sozialwissenschaftler unter diesem Aspekt reflektiert.

Die gestiegene Sensibilität für den Datenschutz hat die Diskussion um eine einvernehmliche Forschungsethik verstärkt.[15] Von verschiedenen Seiten wurde die explizite Formulierung eines "Code of Ethics" befürwortet[16], in dem Datenschutzaspekte eine zentrale Rolle spielen sollten.

Die Bindung der Wissenschaft an eine ausformulierte Forschungsethik wäre möglich, weil die Wissenschaft die in ihrem Bereich für richtig erachteten Steuerungsprinzipien auch geltend machen kann. Die institutionellen Voraussetzungen in Form einer eigenständigen Entwicklung der Wissenschaft sind gegeben. Diese ist der Verwaltung oder auch der Wirtschaft nicht weisungsgebunden. Der Wissenschaftler ist in ein System eingebunden, das einen erheblichen Grad autonomer Selbststeuerung aufweist, die im Grundgesetz durch das Prinzip der Freiheit von Forschung und Lehre abgesichert ist.

Universitäten verwalten sich weitgehend selbst. Die Inhaber der Lehrstühle sind in einem großen Rahmen frei bei der Definition der Inhalte ihrer Forschung und Lehre. Die unabhängigen Forschungsinstitute bestimmen ebenfalls die Angelegenheiten der Forschung in großen Zügen selbst. Diese Unabhängigkeit wurde durch die Wahl entsprechender rechtlicher Konstruktionen abgesichert.[17]

---

13 Beispiele für Beanstandungen der Datenverarbeitung in den Sozialwissenschaften, deren Anlässe allerdings hier nicht in irgendeiner Weise beurteilt werden sollen, enthalten die Berichte der Datenschutzbeauftragten; vgl. etwa den 7. Tätigkeitsbericht der Landesbeauftragten für den Datenschutz in Baden-Württemberg (1986:51ff.).

14 Vgl. u. a. die Beiträge von Eichener (1988), Bayer (1987), Dorer/Wingen (1987), Mohler (1987) und Eberle (1981).

15 Das Thema Forschungsethik wurde jedoch auch unabhängig davon bereits vorher diskutiert. Vgl. Friedrichs (1980:40), Drew (1980:37ff.), Bower/Gasparis (1978).

16 Vgl. u. a. die Beiträge von Simitis (1980:6 und 1987:50), Scheuch (1980:254ff.), Mangold (1980), Ziegler-Jung (1980), Küchler (1983) und Endruweit (1983).

17 So haben z. B. die Max-Planck-Gesellschaft und die ihr angeschlossenen Institute die Rechtsform eines eingetragenen Vereins. Dies zeigt schon formell, daß sie kein Teil der Staatsverwaltung sind.

So entstehen relativ hohe institutionelle Barrieren zwischen dem Bereich der Forschung und anderen gesellschaftlichen Teilbereichen.

Damit ist eine wichtige Bedingung gegeben, der Wissenschaft eine privilegierte Stellung in Regelungen der Datenhaltung und Datennutzung einzuräumen, denn: "Die wahrscheinlich wichtigste Voraussetzung jeder Sonderregelung zugunsten der wissenschaftlichen Forschung (ist - die Verf.): die strikte funktionale Trennung zwischen den Forschungseinrichtungen einerseits und allen anderen an personenbezogenen Daten interessierten Stellen andererseits" (Simitis 1987:53). Unter Ausnutzung dieser Trennung wird es möglich, beispielsweise die Weitergabe von wissenschaftlichen Daten an die Administration zu verbieten, ohne daß die beteiligten Forscher dies als Restriktion in ihren Handlungen empfinden würden.

Allerdings gibt es behördennahe Forschungsinstitutionen, in denen die Scheidung der Institutionen nicht derart eindeutig vollzogen ist wie an den Universitäten. In diesen Fällen müßte jeweils einzeln geprüft werden, ob eine Gefährdung vorhanden ist.

Die Trennung von Staatsapparat und Wissenschaft ist allerdings in einer Hinsicht unvollständig. In der Datenschutzdiskussion wird zu Recht immer wieder darauf hingewiesen, daß ein Forscher kein Zeugnisverweigerungsrecht hat (vgl. Friedrichs 1980:40). Darum ist möglich, daß er behördlich zur Herausgabe von Daten angewiesen wird, wenn die betreffenden Informationen zur Aufklärung oder Verfolgung einer Straftat erforderlich sind.

So ist z. B. denkbar, daß bei einem Projekt aus dem Bereich der kriminologischen Forschung im Rahmen einer Umfrage den Informanten Datenschutz zugesichert wird, die Daten aber dann polizeilich beschlagnahmt und die verbindlichen Garantien dadurch zunichte gemacht werden. Dieses Problem ist jedoch bei den amtlichen Mikrodaten, um die es hier geht, nicht relevant: Abgesehen von dem Umstand, daß die Stelle, die die Daten anfordert, selbst in der Lage sein müßte, die erhaltenen Informationen zu deanonymisieren, sind die Daten gegen eine derartige Verwendung (durch § 16 BStatG) gesetzlich geschützt. Eine Aufforderung zur Herausgabe der Daten wäre gesetzeswidrig und könnte somit abgewehrt werden.

5.1.4 Einige Sonderprobleme

Ergebnis der bisherigen Überlegungen ist, daß sich mit der beruflichen Handlungslogik der quantitativ orientierten Sozialwissenschaften im allgemeinen kein Interesse an einer Deanonymisierung von anonymisierten Mikrodaten verbindet. Der wichtigste Grund ist: Eine Deanonymisierung brächte im Normalfall keinen Nutzen. Im Gegenteil: Eine mißbräuchliche Verwendung von Daten wäre neben der Gefahr rechtlicher Sanktionen für einen Wissenschaftler mit den Risiken beruflicher Nachteile verbunden.

Verstöße durch die Wissenschaft gegen das hier relevante Deanonymisierungsverbot bei Mikrodaten aus amtlichen Erhebungen sind nicht bekannt geworden, obwohl die Daten in früherer Zeit ohne weitreichende Anonymisierungsmaßnahmen an die Forschung übergeben wurden. Ähnliches gilt für Erfahrungen in anderen Ländern, bei denen zum Teil weniger restriktive Datenschutzregelungen als in der Bundesrepublik Deutschland gelten (vgl. Paaß/Wauschkuhn 1985:36), oder in denen die Datenweitergabepraxis aus anderen Gründen weniger restriktiv gehandhabt wird.

Im Zusammenhang quantitativer Methoden ist ein Personenbezug nur für die Zwecke der Erhebung und Zusammenführung von Daten wichtig. Damit könnte sich jedoch das Problem stellen, daß Daten der amtlichen Statistik bei eben dieser Datenerhebung bzw. -zusammenführung nützlich sein könnten.

In diesem Fall wäre zwar die allgemeine Aussage nach wie vor gültig, daß die Forschung nicht auf identifizierbare Personen, sondern auf verallgemeinerbare Ergebnisse ausgerichtet ist. Jedoch könnte bei einem Zwischenschritt im Forschungsprozeß unter Umständen ein Interesse an einer verbotenen Deanonymisierung entstehen. Im folgenden wollen wir dieses am Beispiel von drei hypothetischen Situationen untersuchen, in denen dies der Fall sein könnte.

5.1.4.1 Verwendung zur Stichprobenziehung

Der erste dieser möglichen Fälle bezieht sich darauf, daß ein deanonymisierter Mikrozensus helfen könnte, verbesserte Stichproben für die Umfrageforschung zu erhalten. Angenommen, ein Wissenschaftler interessiert sich für eine sehr spezielle Gruppe der Bevölkerung, die nur mit großem Aufwand lokalisiert werden kann. Wenn die fraglichen Merkmale im Melderegister oder in einem

anderen zugänglichen Register nicht enthalten sind, können diese für eine Stichprobenziehung nicht verwendet werden. Dann ist es vorstellbar, daß eine amtliche Erhebung, die in Form von Einzeldaten vorliegt, nützlich sein könnte, um eine Stichprobe zu gewinnen.

Wie wahrscheinlich ist es nun, daß ein Forscher aus dieser indirekten, aber möglicherweise doch wirksamen Motivation heraus die Deanonymisierung eines ihm zugänglichen Mikrodatensatzes der amtlichen Statistik versucht? Die folgenden Überlegungen zeigen, daß der beschriebene Fall sehr unwahrscheinlich ist.

1) Der amtliche Mikrodatensatz müßte die Merkmale enthalten, die zur Lokalisierung der vom Forscher anvisierten Stichprobe erforderlich sind. Dieses kann keineswegs als selbstverständlich unterstellt werden.

2) Wenn der betreffende Wissenschaftler die Vorgehensweise, mit der er zu seiner Stichprobe gelangt ist, zu verheimlichen sucht, kann er bei der Veröffentlichung seiner Befunde und in der Diskussion mit seinen Kollegen nicht bekanntgeben, auf welche Weise er seine Stichprobe erhalten hat. Dies hätte aber zur Folge, daß die Ergebnisse mit Skepsis beurteilt werden würden. Gibt der Wissenschaftler jedoch die Einzelheiten der Stichprobenziehung bekannt, müßte er z.B. mit gesetzlich vorgesehenen Strafen und der Mißbilligung durch die "Scientific community" rechnen.

3) Schließlich ist zu fragen, ob überhaupt praktisch oder potentiell verfügbare Deanonymisierungsverfahren existieren, die leistungsfähig genug sind, damit sie für den Zweck der Identifikation nicht nur einzelner Personen, sondern einer ganzen Stichprobe eingesetzt werden könnten.

Auf die zuletzt gestellte Frage wird im Zuge der empirischen Reidentifikationsexperimente eingegangen. Eine einfache Überschlagsrechnung zeigt jedoch, daß die unterstellte Konstellation für einen Deanonymisierungsversuch unrealistisch ist. Voraussetzung eines solchen Versuchs ist, daß der betreffende Wissenschaftler über Zusatzwissen in Form eines Identifikationsfile verfügt.

Nehmen wir an, der Angreifer rechnet damit, daß er mit einem geeigneten Verfahren jede zehnte Person, deren Daten sowohl in seinem Zusatzwissen als auch in seinem Mikrodatenfile enthalten sind, identifizieren kann. Ist der Mikrozensus mit einem Auswahlsatz von einem Prozent Ziel des Angreifers und

unterstellt man ein derartiges Ergebnis[18], kann demnach nur jeder 1000. Fall des Identifikationsfile auch in dem Mikrodatenfile aufgefunden werden. Um nur eine sehr kleine Stichprobe von 100 Fällen zu erhalten, müßte der Datenangreifer also über ein Identifikationsfile von nicht weniger als 100 000 Datensätzen verfügen.

Weiterhin ist zu berücksichtigen, daß in der konstruierten Situation davon ausgegangen wird, daß eine Stichprobe von Personen mit sehr spezifischen Merkmalen anvisiert wird. Wird angenommen, daß vielleicht jede zehnte reidentifizierte Person die benötigten Merkmale aufweist, müßte das Identifikationsfile nicht weniger als eine Million Fälle enthalten, damit eine vom Wissenschaftler intendierte Stichprobe von 100 Personen zustande käme.

Es erscheint abwegig, daß jemand den Aufwand auf sich nehmen könnte, eine derartige Menge an Daten für sein Identifikationsfile zusammenzustellen, statt unmittelbar - d. h. ohne den Umweg über das Mikrodatenfile - die für die Stichprobenziehung benötigten Informationen zu sammeln.[19] Im Sinne des Rationalkriteriums der faktischen Anonymität wäre dies wohl der kostengünstigere Weg.

Außerdem haben die denkbaren Deanonymisierungsverfahren einen prinzipiellen Nachteil, der sie unter Umständen zur Gewinnung von Material für eine Stichprobenziehung als gänzlich ungeeignet erscheinen läßt. In allen Fällen, bei denen nicht alle interessierenden Personen identifiziert werden können, wird das gewonnene Datenmaterial gravierende Verzerrungen aufweisen: Personen, die in irgendeiner Weise ungewöhnliche Ausprägungskombinationen aufweisen, sind leichter zu identifizieren als der 'unauffällige Durchschnittsbürger', der nur schwer von 'statistischen Doppelgängern' mit der gleichen (oder fast der gleichen) Ausprägungskombination unterschieden werden kann.[20] Selbst wenn Ziel des Forschers von vornherein eine Gruppe mit speziellen, also vielleicht ungewöhnlichen Merkmalskombinationen ist, wird innerhalb der Gruppe noch einmal eine unkontrollierbare Auswahl durch das Deanonymisierungsverfahren

---

18 Die empirischen Zuordnungsexperimente erbrachten Resultate, die vom Standpunkt des Angreifers ungünstiger waren. Hier wird also eine Abschätzung nach der 'sicheren Seite' gemacht.

19 Darüber hinaus ist es, selbst wenn jemand diesen Versuch unternehmen würde - wie in Kapitel 6 aufgezeigt wird - äußerst fraglich, ob es mit dem im Bereich der Sozialwissenschaften verfügbaren bzw. beschaffbaren Zusatzwissen auch nur annähernd möglich ist, ein Potential an gemeinsamen Überschneidungsmerkmalen für mehr als 100 000 Personen bereitzustellen.

20 Zum Informationsgehalt von Merkmalen und dem Auftreten von Doppelgängern vgl. Paaß/Wauschkuhn (1985:157ff.).

getroffen, die das deanonymisierte Datenmaterial für den angenommenen Zweck im Regelfall unbrauchbar machen wird.

5.1.4.2 Synthese von Einzeldaten aus einem Mikrodatenfile und anderen Datenquellen

Etwas anders gelagert ist der zweite Fall, bei dem ein Interesse von Sozialwissenschaftlern an einer Deanonymisierung von Einzeldaten der amtlichen Statistik konstruiert werden kann: Hier wird angenommen, daß ein Forscher versucht, eine eigene Erhebung, oder Daten die öffentlich zugänglich sind, mit dem Mikrozensus oder der Einkommens- und Verbrauchsstichprobe (EVS) zu kombinieren, um dadurch längere Datensätze mit mehr Merkmalen zu erhalten. Speziell die EVS wäre für ein solches Vorhaben besonders interessant, weil sie eine Vielzahl von Merkmalen enthält, die nur mit beträchtlichem Aufwand erhoben werden können.

Da vorstellbar ist, daß eine Umfrage, deren Daten mit jenen aus der amtlichen Statistik kombiniert werden sollen, schon im Hinblick auf diese Aufgabe konzipiert wird, können potentiell sehr viele Überschneidungsmerkmale auftreten.

Doch selbst wenn im Extremfall jede Person, die sowohl in der Erhebung der amtlichen Statistik als auch in jener aus dem Bereich der Sozialwissenschaft interviewt wurde, als solche mit Sicherheit erkannt werden kann, ist der Effekt für den Datenangreifer enttäuschend: Nur für einen sehr kleinen Prozentsatz der Befragten, die in der sozialwissenschaftlichen Stichprobe enthalten sind, kann das Ziel einer Anreicherung des Umfragedatensatzes mit den Informationen des Mikrodatensatzes erreicht werden. Geht man wieder vom Mikrozensus als Einprozentstichprobe der Bevölkerung aus, werden sich im Durchschnitt die Daten nur von jeder hundertsten Person in beiden Datensätzen befinden. Damit kann jedoch keine Verbesserung der sozialwissenschaftlichen Erhebung erreicht werden. Wenn von beispielsweise 2000 befragten Personen - eine für die Umfrageforschung durchaus typische Stichprobengröße - für 20 ein verlängerter Datensatz vorliegt, so vermag diese Zusatzinformation nicht dem angegebenen Zweck zu dienen. Ein derart kleiner Datenbestand ist uninteressant.

Unterschiedlich zu beurteilen ist dagegen der Fall, bei dem in der Form einer öffentlichen Datenquelle für praktisch alle Angehörigen einer bestimmten regionalen Gruppe Informationen vorliegen, die für die Forschung verwertbar sind (z.B. für eine bestimmte Berufsgruppe). Wenn hinreichend viele Über-

schneidungsmerkmale mit einem Mikrodatenfile vorhanden sind, könnte der Versuch gemacht werden, die in der öffentlichen Datenquelle vorhandenen Informationen durch Zusammenführung mit dem Mikrodatenfile dafür zu nutzen, dieses anzureichern. Wenn die öffentliche Datenquelle mehr oder weniger eine Totalerhebung darstellt und je nach Erfolgschancen der Zusammenführung könnte unter Umständen ein erheblicher Teil der relevanten Mikrodatensätze auf diese Weise erweitert und dann für weiterführende Analysen genutzt werden. Dabei ist allerdings zu berücksichtigen, daß die Verknüpfung der beiden Informationsquellen keine Identifikation darstellt, da diese Verknüpfung anonym vor sich gehen kann und auch die erweiterten Mikrodatensätze anonymisiert bleiben können. Die Verknüpfung kann im übrigen auch im oben erörterten Fall der Erweiterung von Daten einer sozialwissenschaftlichen Erhebung anonym erfolgen. Auch in diesem Fall sind für die Synthese der Datensätze keine personenbezogenen Daten erforderlich.[21]

5.1.4.3 Methodenvergleich [22]

Ausgangspunkt des dritten zu diskutierenden Falls ist, daß ein Wissenschaftler die Erhebungsmethode, die für die amtliche Statistik verwendet wird, mit einer eigenen vergleichen möchte. Vorstellbar ist in diesem Zusammenhang ein Vergleich unterschiedlicher Frageformulierungen, die Gegenüberstellung einer Pflicht- und einer freiwilligen Erhebung oder auch ein Vergleich verschiedener Abgrenzungen der Kategorien in den jeweiligen Fragebögen.

Ein hypothetischer Forscher stellt beim Vergleich von Ergebnissen seiner Untersuchungen mit denen der amtlichen Statistik Abweichungen fest, die nicht allein durch Zufallsfehler im Zusammenhang mit den unterschiedlichen verwendeten Stichproben erklärbar sind. Er will die Instrumente, die er bei seinen Forschungen verwendet, überprüfen und gegebenfalls verbessern. Ein solches

---

21 Die Bildung synthetischer Datensätze mit Daten verschiedener amtlicher Erhebungen ohne Rücksicht darauf, ob die verknüpften Daten auch von einer Person stammen, ist für verschiedene Fragestellungen sinnvoll. Auf diese Weise werden z. B. Datensätze gewonnen, die für Mikrosimulationen verwendet werden können (vgl. für eine Synthese aus Daten der EVS und des Mikrozensus Brennecke 1980, Kortmann/Schmaus 1976, Kortmann 1982 sowie generell zur Methode Paaß 1986). Dies ist jedoch ebensowenig ein Datenanonymitätsproblem wie die Gewichtung von Stichproben der Umfrageforschung mit aggregierten Eckdaten aus amtlichen Erhebungen (vgl. Rothe/Wiedenbeck 1987).

22 Für einen Hinweis auf diese Problematik sei Frau Dr. Preißl (DIW Berlin) gedankt.

Motiv kann im vorliegenden Zusammenhang bei folgenden methodischen Problemen eine Rolle spielen:

1) Bei Befragungen zeigt sich immer wieder, daß dem Inhalt nach äquivalente alternative Formulierungen einer bestimmten Frage unterschiedliches Antwortverhalten provozieren. Diesen Effekt könnte man am besten durch eine Wiederholungsbefragung bei den gleichen Personen überprüfen. Kann man den Mikrozensus für die erste Erhebungswelle benutzen, so entfällt der Aufwand für eine eigene Erhebung.

2) Sehr eng mit dem soeben beschriebenen Fall verwandt, aber doch davon zu unterscheiden, ist die Vorgabe von inhaltlich geringfügig unterschiedlichen Antwortkategorien in den Fragen zweier Erhebungswellen. Dabei ist wiederum vorstellbar, daß eine dieser Erhebungen im Rahmen der amtlichen Statistik durchgeführt wird.

3) Der Dritte in diesem Zusammenhang mögliche Fall ist der Vergleich einer Pflicht- mit einer freiwilligen Erhebung. Ein wichtiges Problem der empirischen Sozialforschung ist der Antwortausfall bei Befragungen (vgl. Porst 1985:91f, Schnell/Hill/Esser 1988:282ff, Erbslöh/Koch 1988, Esser u.a. 1989): Ein Teil der Personen, die für eine sozialwissenschaftliche Umfrage über ein geeignetes Verfahren ausgewählt wurden, antwortet nicht. Dabei spielen zwei unterscheidbare Ursachen eine Rolle: Der eine Teil der Personen ist aus verschiedenen Gründen von den Interviewern nicht auffindbar, der andere verweigert die Antwort. Damit stellt sich die Frage nach der Repräsentativität der betreffenden Umfrage. Über die Zusammensetzung des Personenkreises, der ursprünglich zur Befragung ausgewählt wurde, deren Daten jedoch nicht zur Verfügung stehen, ist in der Regel (fast) nichts bekannt. Im Mikrozensus, der eine Pflichterhebung ist, liegen jedoch über einen wesentlich größeren Anteil der Zielpopulation als in einer freiwilligen sozialwissenschaftlichen Erhebung Daten vor. Über einen Vergleich der Angaben jener Personen, die in der amtlichen Erhebung antworten mußten, aber in der sozialwissenschaftlichen Umfrage die Antwort verweigerten, mit den Daten der Personen, die aus beiden Erhebungen zur Verfügung stehen, kann die Qualität, d. h. in erster Linie die Repräsentativität der Umfrage beurteilt werden.

Auch in diesem Fall wird jedoch die Stichprobeneigenschaft der amtlichen Erhebung zum Hindernis, die einen Deanonymisierungsversuch zum Zweck der Methodenforschung als uninteressant erscheinen läßt. Die Argumente zur Zusammenführung von Umfrage- und amtlichen Mikrodaten lassen sich analog

übertragen. Selbst wenn dann von der völlig unrealistischen Annahme ausgegangen wird, daß alle Personen reidentifiziert werden können, deren Angaben sowohl im Mikrodatenfile als auch in der sozialwissenschaftlichen Stichprobe enthalten sind, müßten im Fall des Mikrozensus in der sozialwissenschaftlichen Stichprobe 100 mal mehr Personen befragt werden, als tatsächlich vergleichbare Datensätze erhalten werden. Das angestrebte Ziel wird also durch eine Parallelerhebung eher erreicht. Sollen etwa die Frageformulierungen auf diese Weise überprüft werden, müßten lediglich die Fragen der Parallelerhebung entsprechend der amtlichen Erhebung formuliert werden.

Allein der Unterschied zwischen Pflicht- und freiwilliger Erhebung kann auf diese Weise nur eingeschränkt untersucht werden. Aber dies ist eher ein Forschungsproblem der amtlichen Statistik als (z. B.) der universitären Sozialforschung (vgl. dazu das in Kooperation mit der amtlichen Statistik durchgeführte Projekt des wissenschaftlichen Beirats für Mikrozensus und Volkszählung, Esser et al. 1989).[23]

Weitere Gegenüberstellungen von Umfragen und amtlichen Erhebungen sind bereits mit anonymen Einzeldaten, unter Umständen sogar mit Aggregatdaten möglich und werden auch ständig durchgeführt, um die Qualität von Stichproben der Umfrageforschung zu beurteilen.[24]

5.1.5 Fazit

Die Untersuchung der Handlungslogik der Sozialwissenschaften zeigte, daß mit den beruflichen Motiven der Forscher in der Regel keine Interessen an deanonymisierten Daten aus der amtlichen Statistik verbunden sind. Selbst dort, wo man in bestimmten Grenzfällen zunächst meinte, ein solches Interesse feststellen zu können, kann dieses nach einer genaueren Analyse nicht mehr konstatiert werden. Dies gilt für die Situationen "Methodenvergleich" und "Synthese von Umfrage- und amtlichen Daten". Die Beurteilung der Möglichkeit der Zusammenführung von amtlichen Mikrodaten mit anderen öffentlich verfüg-

---

23 Bei der Untersuchung wurde eine Teilstichprobe des Mikrozensus, die vom Statistischen Bundesamt im Unterschied zur sonst üblichen Praxis mit freiwilliger Befragung erhoben wurde, mit der Stichprobe verglichen, die auf der Pflichterhebung beruht. Zu den Unterschieden von Pflichterhebungen der amtlichen Statistik gegenüber solchen, die auf freiwilliger Basis durchgeführt werden, vgl. auch Herberger (1985:40ff.).

24 Für ein Beispiel vgl. Blossfeld (1987).

baren Daten, etwa aus Registern, hängt im wesentlichen von den dort enthaltenen, und im nächsten Kapitel näher zu untersuchenden Überschneidungsmerkmalen ab. Im Fall der Situation "Gewinnung von Stichproben" wäre zwar ein Interesse da, es erscheint jedoch ohne praktische Bedeutung. Das allgemeine Ergebnis der Analyse ist, daß quantitativ orientierte Sozialwissenschaftler durch ihre berufliche Handlungslogik zwar an Einzeldaten interessiert sind, aber keinen Nutzen von einem eventuellen Personenbezug haben. Demgemäß sind auf der Grundlage von beruflichen Motiven Angriffsszenarien kaum plausibel konstruierbar (vgl. Krupp/Preißl 1989).

Der einzige gefundene Fall, bei dem ein vermitteltes Interesse an deanonymisierten Daten unterstellt werden konnte, betrifft die Verwendung z. B. des Mikrozensus nach Art eines im Informationsgehalt wesentlich reicheren Melderegisters für Zwecke der Stichprobenbildung. Das Interesse an einer solchen Verwendung kann zwar nur unter unwahrscheinlichen Bedingungen auftreten. Es kann jedoch nicht vollständig ausgeschlossen werden. Praktisch relevant kann es jedoch nur dann werden, wenn ein entsprechend leistungsfähiges Verfahren zur Deanonymisierung in Verbindung mit einem dazu ebenfalls notwendigen außerordentlich umfangreichen Zusatzwissen verfügbar ist. In der bisherigen Analyse wurde unterstellt, daß beide Bedingungen sehr unplausibel sind. Um den Sachverhalt gründlicher abzuklären, werden wir dieses Problem in den Kapiteln 8 und 10 als spezielles Szenario aufgreifen und im Detail untersuchen.

5.2 Nicht an der sozialwissenschaftlichen Handlungslogik orientierte Deanonymisierungsmotive

Neben den aus der Handlungslogik der Sozialwissenschaften ableitbaren Motiven sind Motive denkbar, die nicht an der wissenschaftlichen Verwertung der deanonymisierten Daten orientiert sind. Solche Motive sollen hier als außerberufliche bezeichnet werden.

Man kann geteilter Meinung sein, ob eine Berücksichtigung dieser Motive im Kontext der Weitergabe von Einzelangaben zum Zweck wissenschaftlicher Forschung an einen eingegrenzten und besonders verpflichteten Personenkreis überhaupt erforderlich ist. Selbst das Bundesverfassungsgericht hat es in seinem Volkszählungsurteil nicht für nötig empfunden, auf diesen Punkt überhaupt hinzuweisen.

Die Diskussion dieser Motive erfolgt hier nicht deshalb, weil unterstellt würde, daß sie bei Wissenschaftlern eine besondere Rolle spielen würden, sondern hauptsächlich aus dem Grunde, möglichst alle denkbaren Fälle von Datenmißbrauch, so unwahrscheinlich sie auch sein mögen, in der Analyse zu berücksichtigen. Auch in der öffentlichen Diskussion wird in der Regel nicht davon ausgegangen, daß Gefährdungen aus dem Handeln des Wissenschaftlers in seiner beruflichen Eigenschaft gegeben sind, sondern es werden - wenn überhaupt - eher Gefährdungen angenommen, die aus Motivlagen des allgemein menschlichen Handelns resultieren. Wenn deshalb im folgenden entsprechende Motive näher untersucht werden, so hauptsächlich aus dem Grund, daß nicht ein Aspekt vernachlässigt wird, der Vertretern dieser Position als entscheidende Lücke erscheinen müßte.

In den Ausführungen zur Handlungslogik wurde deutlich, daß in den Sozialwissenschaften berufliche Deanonymisierungsmotive - wenn überhaupt von solchen ausgegangen werden kann - Massenidentifikationen bzw. Massenfischzüge notwendig machen würden. Die Erweiterung der Analyse auf *nichtberufliche* Motive bewirkt, daß auch die beiden Deanonymisierungsstrategien "*Gezielte Suche*" und "*Einzelfischzug*", die sich auf die Reidentifikation von einzelnen Datensätzen beziehen, einbezogen werden müssen. Dies ist insofern hilfreich, als damit die Notwendigkeit besteht, die Forschungsbemühungen auch zum Zwecke verbesserter Kenntnisse in diesem Bereich voranzutreiben.

5.2.1 Der Stellenwert der außerberuflichen Motive

Wenn damit die grundsätzliche Einbeziehung außerberuflicher Motive naheliegt, so soll dennoch betont werden, daß sie nicht als Hebel benutzt werden dürfen, um eine Übermittlung von Einzeldatensätzen an die Wissenschaft unter akzeptablen Bedingungen zu verhindern. Mit anderen Worten, es wäre sicherlich nicht im Sinne von § 16 Abs.6 BStatG, eine Situation zu konstruieren, bei der ein Wissenschaftler durch ein als kriminell zu bezeichnendes Verhalten, beispielsweise für Zwecke der Erpressung, einen bestimmten Einzeldatensatz deanonymisiert, um dann daraus abzuleiten, daß prophylaktisch ein entsprechend hoher Anonymisierungsgrad gewährleistet sein muß, der ein derartiges Vorgehen unmöglich macht. Dies wäre diskreditierend und mit den Absichten des Gesetzgebers nicht konform. Die unterstellten Motive müssen daher einen realistischen "Hintergrund" beibehalten und aus dem Untersuchungsgegenstand ableitbar sein. Das rein numerische Übergewicht von drei außerberuflichen Motiven zu zwei auf beruflichen Motiven basierenden

Situationen, die an späterer Stelle einer Kosten-Nutzen-Analyse unterzogen werden, entstand, weil zum einen zu wenige berufliche Motive vorliegen und zum anderen aus der Vielzahl der hypothetisch denkbaren außerberuflichen, einige häufig genannte aufgegriffen werden sollten, auch stellvertretend für eine Reihe anderer, ähnlicher Motive. Außerdem sollte grundsätzlich das ganze Spektrum möglicher Deanonymisierungsstrategien abgedeckt werden.

5.2.2 Charakterisierung der außerberuflichen Motive

Im Gegensatz zu den beruflichen, bieten die nichtwissenschaftlichen ein weites Feld hypothetisch denkbarer Motive. Der Phantasie sind im Prinzip keine Grenzen gesetzt. Das Spektrum reicht von reiner Neugierde über eigene monetäre Vorteile oder wirtschaftliche Schädigung Dritter bis hin zu eindeutig kriminellen Aktivitäten. Es erscheint wenig ergiebig, an dieser Stelle eine mehr oder weniger vollständige Aufzählung derartiger Gründe zu versuchen; vielmehr soll auf einige Besonderheiten und auf Unterschiede zu den beruflichen Motiven eingegangen werden.

Bei den Verwertungsmöglichkeiten deanonymisierter Daten für persönliche (außerberufliche) Zwecke kann zwischen interner (z.B. Hacker-Ambitionen, das heißt, Freude über das Gelingen einer Deanonymisierung) und externer Verwendung (z.B. Verunglimpfung Dritter) unterschieden werden. Bei interner Verwendung erfolgt im Prinzip keine Offenlegung der Herkunft der gewonnenen Information. Eine Bestrafung oder sonstige Sanktionierung der Deanonymisierung ist in diesen Fällen nur möglich, wenn durch eingebaute technische Sicherungsmaßnahmen der Mißbrauch der Daten nachgewiesen werden kann: D.h., es müßten Sicherungsmaßnahmen vorliegen, die Zugangsberechtigten eine derartige Zweckentfremdung erschweren. Dann müßte der Angreifer auch bei interner Verwendung Nachteile ins Kalkül stellen. Bei der externen Verwendung, von der unmittelbar Dritte betroffen werden, gibt es solche Formen, bei denen die Quelle nicht offengelegt wird. Dann ist die Gefahr einer Sanktionierung gering. Dies gilt z.B. für die anonyme Publikation von Informationen, die etwa aus Datensätzen von "Prominenten" gewonnen werden. Darüber hinaus gibt es Formen, die - möglicherweise im Verlauf eines Strafverfahrens - zu einer Offenlegung führen könnten, z.B. bei Verunglimpfung, Rufmord oder merklicher wirtschaftlicher Schädigung. Mit der Notwendigkeit zur Offenlegung bzw. der steigenden Gefahr, daß die Quelle der Informationen ermittelt wird, wächst für den Angreifer das Risiko einer (zusätzlichen) Bestrafung und sonstiger Sanktionen. Mit anderen Worten: Die Deanonymisierung ist zwar immer illegal, aber

bei beruflichen Motiven ist die Nutzungsabsicht - die wissenschaftliche Forschung - legal. Bei den außerberuflichen Motiven dagegen kann die Nutzungsabsicht legal oder illegal sein. Die Befriedigung der reinen Neugierde und die Erpressung können als (extreme) Beispiele genannt werden. Bei illegaler Verwendung deanonymisierter Daten besteht für den Angreifer das Risiko, nicht nur für die Deanonymisierung als solche, sondern zusätzlich für die Verwendung sanktioniert zu werden. Je nach Art der Verwendung könnte der damit zusammenhängende Gesetzesbruch sogar schwerwiegender sein als der Akt der Deanonymisierung selbst. Damit sind jedoch Extremsituationen angesprochen, die sehr hypothetisch sind und hier insofern nicht weiter verfolgt werden müssen, als der Gesetzgeber die Datenweitergabe zum Zweck wissenschaftlicher Forschung nicht von der Sicherheit des Ausschlusses krimineller Akte abhängig gemacht hat, für die ein Unverhältnismäßigkeitskriterium in keiner vernünftigen Weise formuliert werden kann.

Diskrepanzen zwischen den beiden Motivkategorien gibt es damit auch beim Versuch der Nutzenermittlung einer Reidentifikation. Können die materiellen und immateriellen Vorteile im wissenschaftlichen Bereich zumindest beschrieben werden, ist dies bei den außerberuflichen Motiven nicht mehr möglich, sondern hängt stark vom jeweiligen Motiv ab. Neben materiellen Beweggründen, z.B. durch Adressenverkauf, können Antriebskräfte wie Neugier, Rache, Haß u.s.w. eine wesentliche Rolle spielen. Eine allgemeine Beschreibung oder Bewertung des Nutzens ist bei persönlichen Gründen daher ein aussichtsloses Unterfangen und nur abgestellt auf eine spezielle Situation in groben Zügen möglich. Die Analyse der Vor- und Nachteile, bzw. der Nutzen und Kosten einer Deanonymisierung ist deshalb fast ausschließlich auf die Kostenseite angewiesen und besonders auf einen Kostenvergleich alternativer Datenbeschaffungsmöglichkeiten. Derartige Alternativen können zwar auch nur fallbezogen dargestellt werden, aber als möglicherweise billigere, bessere und legalere Informationsquellen bieten sich für die einigermaßen realistischen Motive persönliche Recherchen bei Kollegen oder Nachbarn, Anfragen bei zugänglichen Auskunfteien (z.B. der SCHUFA), Einschalten von Privatdetektiven u.s.w. geradezu an. Es kann - wie die weitere detaillierte Analyse zeigen wird - davon ausgegangen werden, daß das aus außerberuflichen Motiven resultierende Risiko einer Deanonymisierung aus Kostengründen im allgemeinen unwahrscheinlich erscheint.

5.2.3 Ausgewählte außerberufliche Motive

Auf der Basis der grundsätzlichen Überlegungen wurden für eine detaillierte Untersuchung der außerberuflichen Motive drei Situationen ausgewählt, welche die vorgesehene lückenlose Einbeziehung aller Deanonymisierungsstrategien gewährleisten und sich im Rahmen der realistischen und hier einzubeziehenden Motive bewegen. Ferner werden sie in der Diskussion oft genannt und decken auch das Spektrum der geschilderten Charakteristika weitgehend ab. Die erste Situation stellt darauf ab, daß aus persönlicher Neugier die bereits vorhandene Information über einen Bekannten, im unterstellten Fall einen Arbeitskollegen, erweitert werden soll. Dabei handelt es sich um eine Verwendung der Daten, die durch die Strategie einer "Gezielten Suche" gewonnen werden sollen, ohne daß die Offenlegung der Quelle notwendig ist. Ein anderes Motiv, das ein gleiches Vorgehen nach sich ziehen würde, wäre z.B. die Informationsbeschaffung über einen Marktkonkurrenten, zur Gewinnung von Marktvorteilen. Auch Verwendungen wie Herabwürdigung des Ansehens einer bestimmten Person aus Rachegefühlen, könnten als Anlaß für eine gezielte Suche angeführt werden. Der Übergang von der legalen zur illegalen Verwendung ist fließend. Je nach unterstelltem Motiv ergeben sich bei der Kosten-Nutzen-Analyse gewisse Modifikationen.

Als Beispiel für einen persönlichen Grund, der zu einer Reidentifikation mittels Einzelfischzug führt, kann auf die Gewinnung von Informationen über einen "Prominenten" mit dem Ziel der anonymen Publikation hingewiesen werden. Dabei liegt eine externe, in das Illegale tendierende Verwendung vor, bei der jedoch das Risiko, "entlarvt" zu werden gering ist.

Die dritte Situation stellt schließlich auf das Motiv der Erzielung eines ökonomischen Vorteils ab. Dabei wird unterstellt, daß in Zusammenarbeit mit einem Adressenhändler über eine(n) Massenidentifikation/-fischzug versucht wird, Informationen aus dem Mikrozensus zu gewinnen, die dann an den Adressenhändler weiterverkauft werden können. Bei dieser Verwendung muß die Offenlegung der Quelle gegenüber dem potentiellen Käufer angenommen werden. Die Art der Datenverwendung selbst ist aber wiederum im legalen Bereich angesiedelt.

Die Ergebnisse der Analyse dieser drei konkreten Situationen bieten eine ausreichende Basis für die Beurteilung der außerberuflichen Motive als potentielle Gefahrenquelle einer Deanonymisierung und die daraus zu ziehenden Folgerungen für die Anwendung von Anonymisierungsverfahren.

5.2.4 Der Nachweis der prinzipiellen Deanonymisierbarkeit eines Mikrodatenfile als eigenständiges wissenschaftliches Motiv

Im Zusammenhang mit der Diskussion um die Volkszählung 1987 wurden im Fachbereich Informatik der Universität Hamburg Reidentifikationsexperimente mit auf Volkszählungsmerkmale abgestellten synthetischen Daten simuliert. Sie hatten das Ziel zu untersuchen, "ob die vom Bundesverfassungsgericht geforderte faktischen Anonymität der Volkszählungsdaten tatsächlich gewährleistet ist".[25] Es ist daher denkbar, daß seitens der Wissenschaft ähnliche Versuche mit von der amtlichen Statistik übermittelten faktisch anonymisierten "echten" Daten unternommen werden. Ein solches Motiv kann jedoch nicht Grundlage für die vorzunehmenden Anonymisierungsmaßnahmen sein, sondern muß als verbleibendes Restrisiko getragen werden (vgl. dazu Brennecke 1980). Darin liegt letztlich der Unterschied zwischen der faktischen und der absoluten Anonymität. Die Formulierung des § 16 Abs.6 BStatG macht deutlich, daß eine Deanonymisierung nicht gänzlich ausgeschlossen sein muß, bzw. daß sie prinzipiell möglich ist, aber eben nur unter ganz bestimmten Voraussetzungen. Das verbleibende Risiko ist demnach durch § 16 Abs.6 BStatG abgedeckt.

Das Motiv des Nachweises einer prinzipiellen Deanonymisierbarkeit stellt einen Spezialfall dar, bei dem das vom Gesetzgeber formulierte Unverhältnismäßigkeitskriterium nicht anwendbar ist, weil hier der Kostenaspekt selbst nicht in einer üblichen Weise ins Kalkül einbezogen werden muß. Ein Wissenschaftler, z.B. aus dem Bereich der Informatik, könnte seine gesamten beruflichen Forschungsanstrengungen, die ihm zur Verfügung stehende apparative Ausstattung und die Mitarbeiter seines Instituts dafür einsetzen, neue Deanonymisierungsroutinen zu entwickeln. Der Nachweis der prinzipiellen Machbarkeit wäre demzufolge das Ziel seiner Forschungsarbeiten. Das Unverhältnismäßigkeitskriterium wird deshalb unterlaufen, weil es mit einem solchen Forschungsziel sinnvoll sein kann, Deanonymisierungsversuche durchzuführen, wie hoch auch immer die Kosten dafür sein mögen. Der Angreifer selbst müßte für diese Kosten selbst nicht aufkommen, da sie voraussichtlich von seinem Institut getragen würden.

Bei dem Versuch des Nachweises der prinzipiellen Machbarkeit von Deanonymisierungen können auch Zielsetzungen verfolgt werden, mit denen die Diskreditierung der Regelung einer Freigabe von Mikrodaten angestrebt wird oder auf Intentionen eines verstärkten Datenschutzes abgestellt wird. Zwar müßte

---

25 Fischer-Hübner (1986): Vorwort von Brunnstein, S.II.

der Angreifer seine Vorgehensweise offenlegen und damit grundsätzlich erhebliche Strafen und sonstige Sanktionen befürchten. Es ist aber auch denkbar, daß durch die Verbindung des Vorhabens mit Anliegen des Datenschutzes der Schaden begrenzt werden kann. Von einer unter anderem daraus resultierenden Aussetzung oder Erschwerung der Weitergabe faktischer anonymisierter Daten wäre der Angreifer selbst nicht betroffen, da er nicht auf dem entsprechenden Wissenschaftsgebiet tätig ist.

Der Nutzen aus einer erfolgreichen Deanonymisierung ist dagegen nicht abzuschätzen. Ein Angreifer kann unter Umständen erhebliche berufliche Anerkennung gewinnen. Der Nutzen wäre dann als durchaus hoch einzuschätzen. Auf der anderen Seite geht der Wissenschaftler trotz der geschilderten Schadensbegrenzungsstrategie auch ein unkalkulierbares Risiko ein, weil er mit der Deanonymisierung einen illegalen Akt begeht. Aus diesen unberechenbaren Konstellationen ist keine eindeutige Nutzenfunktion abzuleiten.

Die Einbeziehung des Motivs des Nachweises der prinzipiellen Machbarkeit der Deanonymisierung würde die vom Gesetzgeber mit dem §16 Abs.6 BStatG verfolgte Intention geradezu torpedieren. Mit der Regelung der *faktischen* Anonymität hat der Gesetzgeber ja implizit anerkannt, daß unter bestimmten Voraussetzungen Deanonymisierungen möglich sind und er bereit ist, solche Fälle als verbleibendes Risiko hinzunehmen, wenn sie nur unter Einsatz unverhältnismäßiger Mittel zustande kommen können. Der hier untersuchte Fall kann insofern darunter subsumiert werden, als der Angreifer gerade bereit ist, unverhältnismäßige Mittel einzusetzen. Sie rechtfertigen sich aus dem Nachweis der prinzipiellen Deanonymisierbarkeit, bei der eine völlig andere Kosten-Nutzen-Rationalität bestehen kann.

6 Das Zusatzwissen für Deanonymisierungsvorhaben im Bereich der Wissenschaft*

Einer der wichtigsten Bestimmungsfaktoren des Risikos einer Reidentifikation ist das sogenannte Zusatzwissen, über das ein Datenangreifer verfügt. Darunter verstehen wir die Informationen, die dem Datenangreifer über einzelne namentlich identifizierbare Personen bekannt sind und die als Überschneidungsmerkmale auch im Mikrodatenfile enthalten sind. Das allgemeine Vorgehen bei einem Deanonymisierungsversuch besteht darin, durch den Abgleich dieser Überschneidungsmerkmale Fälle aus dem Mikrodatenfile, die der persönlichen Identität nach unbekannt sind, den namentlich bekannten Fällen des Zusatzwissens (bzw. des Identifikationsfile) zuzuweisen und damit zu identifizieren.[2]

Eine Untersuchung des Reidentifikationsrisikos muß sich deshalb besonders gründlich mit der Verfügbarkeit und Nutzbarkeit solchen Zusatzwissens befassen. Da es im Zusammenhang dieser Untersuchung um faktische Anonymität von Mikrodaten geht, die an die Wissenschaft übermittelt werden, müssen im Mittelpunkt des Interesses Situationen stehen, die für den Bereich wissenschaftlicher Institutionen typisch sind, und zwar sowohl was die Motive für einen Deanonymisierungsversuch, als auch was das im wissenschaftlichen Bereich typischerweise zur Verfügung stehende Zusatzwissen betrifft.

Als potentiell wissenschaftliche Deanonymisierungsmotive wurden im vorausgegangenen Kapitel die Gewinnung einer Stichprobe oder die Ergänzung eines eigenen Datenfile mit Informationen aus einem Mikrodatenfile herausgearbeitet. Bei einem solchen Deanonymisierungsszenario müßte eine große Zahl von

---

* Autorin: Petra Beckmann

2 Abweichend von dieser allgemeinen Problemlage ist es bei einem sehr kleinen Kreis von Personen nicht erforderlich , daß Namen und Adressen im Zusatzwissen vorhanden sind, da diese wenigen Personen bereits aufgrund ganz bestimmter mittelbarer Identifikatoren deanonymisierbar sind. Dieses Problem wird in die Betrachtung mit eingeschlossen, aber nicht gesondert behandelt. Als Beispiel wäre hier an eine Sozialdatenerhebung für die Stadt Stuttgart zu denken, die einen männlichen Kommunalbeamten im Alter von über 50 Jahren mit einem Einkommen von über 100 000 DM ausweist, so daß klar erkennbar ist, daß es sich hier um den Oberbürgermeister der Stadt Stuttgart handeln muß. Das Beispiel verdeutlicht jedoch, daß auf diese Weise immer nur Einzelfälle identifiziert werden können, die in der vorliegenden Analyse nicht im Mittelpunkt des Interesses stehen (vgl. Kapitel 5).

Fällen reidentifiziert werden, um überhaupt dem Zweck des Datenangriffs gerecht werden zu können, d.h. es muß ein Datenangriff in Form einer Massenidentifikation oder eines Massenfischzuges vorausgesetzt werden.

Was das verfügbare Zusatzwissen betrifft, ist zunächst darauf hinzuweisen, daß es für den wissenschaftlichen Bereich geradezu charakteristisch ist, daß dort in aller Regel wenig Informationen über namentlich bekannte Individuen vorliegen, ganz im Unterschied etwa zu Eingriffsbehörden oder auch Institutionen des privatwirtschaftlichen Sektors. Solche Institutionen müssen, zum Teil zum Zweck des Erfüllens ihrer Aufgaben, Daten in expliziter Verbindung mit identifizierbaren Personen anlegen und aufbewahren. Ein Sparer ist in der Regel daran interessiert, daß die Einzahlungen, die er tätigt, auf einem eindeutig seiner Person zuzuordnenden Konto verbucht werden, in ähnlicher Weise ein Versicherungsnehmer oder ein Steuerzahler. Wie schon die Untersuchung der Handlungslogik der Sozialwissenschaften deutlich gemacht hat, ist das wissenschaftliche Interesse dagegen nicht auf die Erkenntnis über namentlich bekannte Individuen, Haushalte oder Institutionen ausgerichtet, sondern lediglich auf solche Einheiten als anonyme Fälle und Merkmalsträger. Deshalb sind sozialwissenschaftliche Forschungsstellen auch nicht Einrichtungen, in denen sich große Informationsmengen über namentlich bekannte Personen anhäufen.

Im wesentlichen sind es zwei Arten von Informationen, die im sozialwissenschaftlichen Bereich vorhanden oder zugänglich sind und die als Zusatzwissen verwandt werden können. In der Literatur wird im Hinblick auf das Zusatzwissen immer wieder auf Wissen hingewiesen, welches aus Daten stammt, die "*allgemeinkundig*" ("offenkundig") bzw. aus "*allgemein zugänglichen Quellen*" beschafft oder beschaffbar sind (Hammerbacher 1982:287). Bei diesen Daten handelt es sich in erster Linie um Angaben, die aus öffentlichen Registern, prozeßduzierten Datensammlungen (Bick/Müller 1977) und jedermann zugänglichen privaten Datensammlungen (Adreßbücher, Telefonbuch etc.) zu entnehmen sind.

Der Wissenschaftler verfügt zusätzlich zu den genannten Datenquellen, die jedem Bürger zugänglich sind, professionsgemäß noch über die Daten, die er in *eigenen empirischen Untersuchungen* erhoben hat. Auch diese Daten sind - soweit sie selbst personenbezogen sind - Bestandteil des Zusatzwissens, das dazu verwendet werden könnte, die anonymisierten Einzeldaten, die dem Wissenschaftler von seiten der statistischen Ämter übermittelt werden, zu deanonymisieren.

In diesem Kapitel soll im einzelnen untersucht werden, welches Zusatzwissen in den Sozialwissenschaften typischerweise vorhanden ist und welche Voraussetzungen für seine Nutzung zu einer Deanonymisierung von anonymisierten Einzeldaten gegeben sind. Im dritten Teil dieses Kapitels werden dabei wichtige, einschlägige öffentliche und private Register zunächst hinsichtlich ihres rechtlichen Zugangs, ihres Informationsgehalts, ihrer Aktualität sowie ihrer Überschneidungsmerkmale mit den Daten des Mikrozensus und der EVS dargestellt. In einem Unterpunkt erfolgt eine detailliertere Auseinandersetzung mit dem Problem, wie sich die Zugangsmöglichkeiten zu Daten aus öffentlichen Registern speziell für die Wissenschaft in der Praxis darstellen und welche Konsequenzen sich daraus für das unterstellte Szenario ergeben.

Die in den Sozialwissenschaften zusätzlich verfügbaren Daten, die vom Forscher selbst erhoben werden, sollen im vierten Teil des Kapitels unter den gleichen Gesichtspunkten wie die Daten aus öffentlichen Registern untersucht werden.

Abschließend sollen die Ergebnisse, die die Analysen des Zusatzwissens hinsichtlich seiner Verwendbarkeit im Deanonymisierungsprozeß erbracht haben, kurz resümiert werden.

Bevor jedoch der Informationsgehalt der in den Registern und privaten Dateien vorhandenen Überschneidungsmerkmale genauer analysiert wird, sollen vorab zwei Sachverhalte nochmals kurz in Erinnerung gerufen werden, von denen die Nutzbarkeit eines Zusatzwissens für Reidentifikationen wesentlich abhängt. Sie wurden bereits in Kapitel 4 ausführlich diskutiert. Zum einen geht es um die mögliche Fehlerhaftigkeit des Zusatzwissens und um die Kompatibilität des Zusatzwissens mit den Informationen im Mikrodatenfile. Zum anderen geht es um die Frage des Erfassungsgrades einer Population im Zusatzwissen. Handelt es sich um eine vollständige Erfassung einer durch bestimmte Kriterien abgegrenzten Population oder um Stichproben? Welches sind die Kriterien, nach denen ein Populationsmitglied im Zusatzwissen enthalten ist oder nicht?

6.1 Kompatibilität und Umfang von Mikrodatenfile und Identifikationsfile

6.1.1 Kompatibilität

Wie schon in Abschnitt 4.3 ausgeführt und wie insbesondere bei den empirischen Untersuchungen in Kapitel 9 deutlich werden wird, ist einer der wesentlichsten Bestimmungsgründe dafür, ob es zu Reidentifikationen kommen kann, daß die Überschneidungsmerkmale im Zusatzwissen (Identifikationsfile) in gleicher Weise abgebildet sind wie im Mikrodatenfile. Wie schon diskutiert, kann es viele Gründe geben, weshalb es zu Inkompatibilitäten in der Abbildung eines Sachverhaltes im Zusatzwissen und im Mikrodatenfile kommen kann. An dieser Stelle soll nicht die Kompatibilitätsproblematik insgesamt nochmals diskutiert werden, sondern nur auf einzelne Punkte hingewiesen werden, die zu berücksichtigen sind, wenn die Nutzbarkeit einzelner Arten von Zusatzwissen für Reidentifikationszwecke einzuschätzen ist. Neben eher banalen Ursachen für Inkompatibilitäten (Übermittlungs- und Kodierfehler, Fehlinterpretation von Merkmalsdefinitionen etc.) ist besonders auf Inkompatibilitäten zu achten, die sich aus spezifischen Zwecksetzungen und den spezifischen operationalen Prozeduren ergeben, unter denen ein Zusatzwissen entsteht. Dies ist vor allem bedeutsam, wenn es sich beim Zusatzwissen um eine Datenbasis handelt, die in Form eines Registers für einen spezifischen Zweck zustande gekommen ist. Als Faktoren, die die Kompatibilität beeinträchtigen können, sind unter anderem wichtig:

- Das bewußte oder unbewußte Interesse der Personen oder Institutionen, die Informationen für die Quelle des Zusatzwissens liefern, sich bei der informationssammelnden Stelle in einer bestimmten Weise darzustellen. Dieser Aspekt ist dann von besonderer Bedeutung, wenn mit einer bestimmten Falldarstellung Vor- oder Nachteile verbunden sein können. Ein solches Interesse kann bei vielen Datenbasen insbesondere des Verwaltungsvollzuges, aber auch bei privaten Informationssammelstellen, bei denen es etwa um eine günstige werbende Darstellung geht, angenommen werden. Ein ähnliches Interesse kann man analog auch bei informationssammelnden Stellen annehmen.

- Der zeitliche Bezug der Informationen im Zusatzwissen und seine Übereinstimmung mit dem Zeitbezug der Informationen des zu deanonymisierenden Mikrodatenfile. Insbesondere bei Registern hängt der Zeitbezug mit den Anlässen zusammen, die zu einem Registereintrag oder zu einer

Veränderung der Registrierung führen. So beziehen sich beispielsweise die Befragungen im Rahmen des Mikrozensus in der Regel auf einen bestimmten, einheitlichen Erhebungszeitpunkt. Den Angaben in öffentlichen und privaten Datenquellen liegt dagegen in der Regel ein personenspezifischer und damit variabler Erhebungszeitpunkt zugrunde, wie etwa der Tag, an dem jemand einem Verein beigetreten ist, geheiratet oder einen Telefonanschluß beantragt hat. Dieser Tatbestand kann zur Folge haben, daß die Angaben zur Person völlig veraltet sind. Jemand kann bereits seit 20 Jahren Mitglied eines Vereins sein, ebensolange verheiratet sein, und auch der Telefonanschluß muß sich in diesem Zeitraum nicht geändert haben. Da die Register in der Regel nur dann aktualisiert werden, wenn bezüglich der betreffenden Person registererhebliche Veränderungen eintreten, dürfte es nahezu unmöglich sein, einen völlig übereinstimmenden inhaltlichen Bezug zwischen den Daten in den Registern und den Daten im Mikrodatenfile herzustellen. Beispielsweise kann die Berufsangabe im Melderegister nur dann geändert werden, wenn der Betreffende dem Einwohnermeldeamt seinen Auszug aus einer bzw. Einzug in eine Wohnung mitgeteilt hat.

- Die Definition der Einheiten des Zusatzwissens. Während die Einheiten der hier interessierenden Mikrodatenfiles in der Regel Haushalte oder individuelle Personen sind, gibt es viele Register, deren Erhebungseinheiten Vereine, Genossenschaften, Betriebe, Berufe, Grundstücke etc. sind. Der konkrete Personenbezug muß demnach in vielen Fällen erst hergestellt werden.

- Die örtliche Lokalisierung der im Zusatzwissen erfaßten Informationen und die Kompatibilität dieser Lokalisierung mit den Informationen im Mikrodatenfile. Während der lokale Bezug der Personen im Mikrodatenfile in der Regel der Wohnort ist, ist der Ortsbezug des Zusatzwissens oftmals die Betriebsstätte oder der Ort, an dem ein von einer bestimmten Person erworbenes Grundstück steht.

Während bei allen diesen Punkten das Zusatzwissen korrekt im Sinne der Zwecksetzung einer Informationssammlung sein kann, können die spezifischen Konstruktionsprinzipien aber dennoch zu weitreichenden Inkompatibilitäten mit der Abbildung von Sachverhalten im Mikrodatenfile führen. Wie weitreichend solche Inkompatibilitäten sind, müßte in jedem einzelnen Fall gesondert untersucht werden. Neben Inkompatibilitäten, die auch bei sachgerechter Abbildung einzig wegen des unterschiedlichen Zweckes der Abbildung zustande kommen können, ist aber auch auf Inkompatibilitäten hinzuweisen, die aus fehlerhaften -

getroffen, die das deanonymisierte Datenmaterial für den angenommenen Zweck im Regelfall unbrauchbar machen wird.

5.1.4.2 Synthese von Einzeldaten aus einem Mikrodatenfile und anderen Datenquellen

Etwas anders gelagert ist der zweite Fall, bei dem ein Interesse von Sozialwissenschaftlern an einer Deanonymisierung von Einzeldaten der amtlichen Statistik konstruiert werden kann: Hier wird angenommen, daß ein Forscher versucht, eine eigene Erhebung, oder Daten die öffentlich zugänglich sind, mit dem Mikrozensus oder der Einkommens- und Verbrauchsstichprobe (EVS) zu kombinieren, um dadurch längere Datensätze mit mehr Merkmalen zu erhalten. Speziell die EVS wäre für ein solches Vorhaben besonders interessant, weil sie eine Vielzahl von Merkmalen enthält, die nur mit beträchtlichem Aufwand erhoben werden können.

Da vorstellbar ist, daß eine Umfrage, deren Daten mit jenen aus der amtlichen Statistik kombiniert werden sollen, schon im Hinblick auf diese Aufgabe konzipiert wird, können potentiell sehr viele Überschneidungsmerkmale auftreten.

Doch selbst wenn im Extremfall jede Person, die sowohl in der Erhebung der amtlichen Statistik als auch in jener aus dem Bereich der Sozialwissenschaft interviewt wurde, als solche mit Sicherheit erkannt werden kann, ist der Effekt für den Datenangreifer enttäuschend: Nur für einen sehr kleinen Prozentsatz der Befragten, die in der sozialwissenschaftlichen Stichprobe enthalten sind, kann das Ziel einer Anreicherung des Umfragedatensatzes mit den Informationen des Mikrodatensatzes erreicht werden. Geht man wieder vom Mikrozensus als Einprozentstichprobe der Bevölkerung aus, werden sich im Durchschnitt die Daten nur von jeder hundertsten Person in beiden Datensätzen befinden. Damit kann jedoch keine Verbesserung der sozialwissenschaftlichen Erhebung erreicht werden. Wenn von beispielsweise 2000 befragten Personen - eine für die Umfrageforschung durchaus typische Stichprobengröße - für 20 ein verlängerter Datensatz vorliegt, so vermag diese Zusatzinformation nicht dem angegebenen Zweck zu dienen. Ein derart kleiner Datenbestand ist uninteressant.

Unterschiedlich zu beurteilen ist dagegen der Fall, bei dem in der Form einer öffentlichen Datenquelle für praktisch alle Angehörigen einer bestimmten regionalen Gruppe Informationen vorliegen, die für die Forschung verwertbar sind (z.B. für eine bestimmte Berufsgruppe). Wenn hinreichend viele Über-

schneidungsmerkmale mit einem Mikrodatenfile vorhanden sind, könnte der Versuch gemacht werden, die in der öffentlichen Datenquelle vorhandenen Informationen durch Zusammenführung mit dem Mikrodatenfile dafür zu nutzen, dieses anzureichern. Wenn die öffentliche Datenquelle mehr oder weniger eine Totalerhebung darstellt und je nach Erfolgschancen der Zusammenführung könnte unter Umständen ein erheblicher Teil der relevanten Mikrodatensätze auf diese Weise erweitert und dann für weiterführende Analysen genutzt werden. Dabei ist allerdings zu berücksichtigen, daß die Verknüpfung der beiden Informationsquellen keine Identifikation darstellt, da diese Verknüpfung anonym vor sich gehen kann und auch die erweiterten Mikrodatensätze anonymisiert bleiben können. Die Verknüpfung kann im übrigen auch im oben erörterten Fall der Erweiterung von Daten einer sozialwissenschaftlichen Erhebung anonym erfolgen. Auch in diesem Fall sind für die Synthese der Datensätze keine personenbezogenen Daten erforderlich.[21]

5.1.4.3 Methodenvergleich [22]

Ausgangspunkt des dritten zu diskutierenden Falls ist, daß ein Wissenschaftler die Erhebungsmethode, die für die amtliche Statistik verwendet wird, mit einer eigenen vergleichen möchte. Vorstellbar ist in diesem Zusammenhang ein Vergleich unterschiedlicher Frageformulierungen, die Gegenüberstellung einer Pflicht- und einer freiwilligen Erhebung oder auch ein Vergleich verschiedener Abgrenzungen der Kategorien in den jeweiligen Fragebögen.

Ein hypothetischer Forscher stellt beim Vergleich von Ergebnissen seiner Untersuchungen mit denen der amtlichen Statistik Abweichungen fest, die nicht allein durch Zufallsfehler im Zusammenhang mit den unterschiedlichen verwendeten Stichproben erklärbar sind. Er will die Instrumente, die er bei seinen Forschungen verwendet, überprüfen und gegebenfalls verbessern. Ein solches

---

21 Die Bildung synthetischer Datensätze mit Daten verschiedener amtlicher Erhebungen ohne Rücksicht darauf, ob die verknüpften Daten auch von einer Person stammen, ist für verschiedene Fragestellungen sinnvoll. Auf diese Weise werden z. B. Datensätze gewonnen, die für Mikrosimulationen verwendet werden können (vgl. für eine Synthese aus Daten der EVS und des Mikrozensus Brennecke 1980, Kortmann/Schmaus 1976, Kortmann 1982 sowie generell zur Methode Paaß 1986). Dies ist jedoch ebensowenig ein Datenanonymitätsproblem wie die Gewichtung von Stichproben der Umfrageforschung mit aggregierten Eckdaten aus amtlichen Erhebungen (vgl. Rothe/Wiedenbeck 1987).

22 Für einen Hinweis auf diese Problematik sei Frau Dr. Preißl (DIW Berlin) gedankt.

Motiv kann im vorliegenden Zusammenhang bei folgenden methodischen Problemen eine Rolle spielen:

1) Bei Befragungen zeigt sich immer wieder, daß dem Inhalt nach äquivalente alternative Formulierungen einer bestimmten Frage unterschiedliches Antwortverhalten provozieren. Diesen Effekt könnte man am besten durch eine Wiederholungsbefragung bei den gleichen Personen überprüfen. Kann man den Mikrozensus für die erste Erhebungswelle benutzen, so entfällt der Aufwand für eine eigene Erhebung.

2) Sehr eng mit dem soeben beschriebenen Fall verwandt, aber doch davon zu unterscheiden, ist die Vorgabe von inhaltlich geringfügig unterschiedlichen Antwortkategorien in den Fragen zweier Erhebungswellen. Dabei ist wiederum vorstellbar, daß eine dieser Erhebungen im Rahmen der amtlichen Statistik durchgeführt wird.

3) Der Dritte in diesem Zusammenhang mögliche Fall ist der Vergleich einer Pflicht- mit einer freiwilligen Erhebung. Ein wichtiges Problem der empirischen Sozialforschung ist der Antwortausfall bei Befragungen (vgl. Porst 1985:91f, Schnell/Hill/Esser 1988:282ff, Erbslöh/Koch 1988, Esser u.a. 1989): Ein Teil der Personen, die für eine sozialwissenschaftliche Umfrage über ein geeignetes Verfahren ausgewählt wurden, antwortet nicht. Dabei spielen zwei unterscheidbare Ursachen eine Rolle: Der eine Teil der Personen ist aus verschiedenen Gründen von den Interviewern nicht auffindbar, der andere verweigert die Antwort. Damit stellt sich die Frage nach der Repräsentativität der betreffenden Umfrage. Über die Zusammensetzung des Personenkreises, der ursprünglich zur Befragung ausgewählt wurde, deren Daten jedoch nicht zur Verfügung stehen, ist in der Regel (fast) nichts bekannt. Im Mikrozensus, der eine Pflichterhebung ist, liegen jedoch über einen wesentlich größeren Anteil der Zielpopulation als in einer freiwilligen sozialwissenschaftlichen Erhebung Daten vor. Über einen Vergleich der Angaben jener Personen, die in der amtlichen Erhebung antworten mußten, aber in der sozialwissenschaftlichen Umfrage die Antwort verweigerten, mit den Daten der Personen, die aus beiden Erhebungen zur Verfügung stehen, kann die Qualität, d. h. in erster Linie die Repräsentativität der Umfrage beurteilt werden.

Auch in diesem Fall wird jedoch die Stichprobeneigenschaft der amtlichen Erhebung zum Hindernis, die einen Deanonymisierungsversuch zum Zweck der Methodenforschung als uninteressant erscheinen läßt. Die Argumente zur Zusammenführung von Umfrage- und amtlichen Mikrodaten lassen sich analog

übertragen. Selbst wenn dann von der völlig unrealistischen Annahme ausgegangen wird, daß alle Personen reidentifiziert werden können, deren Angaben sowohl im Mikrodatenfile als auch in der sozialwissenschaftlichen Stichprobe enthalten sind, müßten im Fall des Mikrozensus in der sozialwissenschaftlichen Stichprobe 100 mal mehr Personen befragt werden, als tatsächlich vergleichbare Datensätze erhalten werden. Das angestrebte Ziel wird also durch eine Parallelerhebung eher erreicht. Sollen etwa die Frageformulierungen auf diese Weise überprüft werden, müßten lediglich die Fragen der Parallelerhebung entsprechend der amtlichen Erhebung formuliert werden.

Allein der Unterschied zwischen Pflicht- und freiwilliger Erhebung kann auf diese Weise nur eingeschränkt untersucht werden. Aber dies ist eher ein Forschungsproblem der amtlichen Statistik als (z. B.) der universitären Sozialforschung (vgl. dazu das in Kooperation mit der amtlichen Statistik durchgeführte Projekt des wissenschaftlichen Beirats für Mikrozensus und Volkszählung, Esser et al. 1989).[23]

Weitere Gegenüberstellungen von Umfragen und amtlichen Erhebungen sind bereits mit anonymen Einzeldaten, unter Umständen sogar mit Aggregatdaten möglich und werden auch ständig durchgeführt, um die Qualität von Stichproben der Umfrageforschung zu beurteilen.[24]

5.1.5 Fazit

Die Untersuchung der Handlungslogik der Sozialwissenschaften zeigte, daß mit den beruflichen Motiven der Forscher in der Regel keine Interessen an deanonymisierten Daten aus der amtlichen Statistik verbunden sind. Selbst dort, wo man in bestimmten Grenzfällen zunächst meinte, ein solches Interesse feststellen zu können, kann dieses nach einer genaueren Analyse nicht mehr konstatiert werden. Dies gilt für die Situationen "Methodenvergleich" und "Synthese von Umfrage- und amtlichen Daten". Die Beurteilung der Möglichkeit der Zusammenführung von amtlichen Mikrodaten mit anderen öffentlich verfüg-

---

23 Bei der Untersuchung wurde eine Teilstichprobe des Mikrozensus, die vom Statistischen Bundesamt im Unterschied zur sonst üblichen Praxis mit freiwilliger Befragung erhoben wurde, mit der Stichprobe verglichen, die auf der Pflichterhebung beruht. Zu den Unterschieden von Pflichterhebungen der amtlichen Statistik gegenüber solchen, die auf freiwilliger Basis durchgeführt werden, vgl. auch Herberger (1985:40ff.).

24 Für ein Beispiel vgl. Blossfeld (1987).

baren Daten, etwa aus Registern, hängt im wesentlichen von den dort enthaltenen, und im nächsten Kapitel näher zu untersuchenden Überschneidungsmerkmalen ab. Im Fall der Situation "Gewinnung von Stichproben" wäre zwar ein Interesse da, es erscheint jedoch ohne praktische Bedeutung. Das allgemeine Ergebnis der Analyse ist, daß quantitativ orientierte Sozialwissenschaftler durch ihre berufliche Handlungslogik zwar an Einzeldaten interessiert sind, aber keinen Nutzen von einem eventuellen Personenbezug haben. Demgemäß sind auf der Grundlage von beruflichen Motiven Angriffsszenarien kaum plausibel konstruierbar (vgl. Krupp/Preißl 1989).

Der einzige gefundene Fall, bei dem ein vermitteltes Interesse an deanonymisierten Daten unterstellt werden konnte, betrifft die Verwendung z. B. des Mikrozensus nach Art eines im Informationsgehalt wesentlich reicheren Melderegisters für Zwecke der Stichprobenbildung. Das Interesse an einer solchen Verwendung kann zwar nur unter unwahrscheinlichen Bedingungen auftreten. Es kann jedoch nicht vollständig ausgeschlossen werden. Praktisch relevant kann es jedoch nur dann werden, wenn ein entsprechend leistungsfähiges Verfahren zur Deanonymisierung in Verbindung mit einem dazu ebenfalls notwendigen außerordentlich umfangreichen Zusatzwissen verfügbar ist. In der bisherigen Analyse wurde unterstellt, daß beide Bedingungen sehr unplausibel sind. Um den Sachverhalt gründlicher abzuklären, werden wir dieses Problem in den Kapiteln 8 und 10 als spezielles Szenario aufgreifen und im Detail untersuchen.

5.2 Nicht an der sozialwissenschaftlichen Handlungslogik orientierte Deanonymisierungsmotive

Neben den aus der Handlungslogik der Sozialwissenschaften ableitbaren Motiven sind Motive denkbar, die nicht an der wissenschaftlichen Verwertung der deanonymisierten Daten orientiert sind. Solche Motive sollen hier als außerberufliche bezeichnet werden.

Man kann geteilter Meinung sein, ob eine Berücksichtigung dieser Motive im Kontext der Weitergabe von Einzelangaben zum Zweck wissenschaftlicher Forschung an einen eingegrenzten und besonders verpflichteten Personenkreis überhaupt erforderlich ist. Selbst das Bundesverfassungsgericht hat es in seinem Volkszählungsurteil nicht für nötig empfunden, auf diesen Punkt überhaupt hinzuweisen.

Die Diskussion dieser Motive erfolgt hier nicht deshalb, weil unterstellt würde, daß sie bei Wissenschaftlern eine besondere Rolle spielen würden, sondern hauptsächlich aus dem Grunde, möglichst alle denkbaren Fälle von Datenmißbrauch, so unwahrscheinlich sie auch sein mögen, in der Analyse zu berücksichtigen. Auch in der öffentlichen Diskussion wird in der Regel nicht davon ausgegangen, daß Gefährdungen aus dem Handeln des Wissenschaftlers in seiner beruflichen Eigenschaft gegeben sind, sondern es werden - wenn überhaupt - eher Gefährdungen angenommen, die aus Motivlagen des allgemein menschlichen Handelns resultieren. Wenn deshalb im folgenden entsprechende Motive näher untersucht werden, so hauptsächlich aus dem Grund, daß nicht ein Aspekt vernachlässigt wird, der Vertretern dieser Position als entscheidende Lücke erscheinen müßte.

In den Ausführungen zur Handlungslogik wurde deutlich, daß in den Sozialwissenschaften berufliche Deanonymisierungsmotive - wenn überhaupt von solchen ausgegangen werden kann - Massenidentifikationen bzw. Massenfischzüge notwendig machen würden. Die Erweiterung der Analyse auf *nichtberufliche* Motive bewirkt, daß auch die beiden Deanonymisierungsstrategien "*Gezielte Suche*" und "*Einzelfischzug*", die sich auf die Reidentifikation von einzelnen Datensätzen beziehen, einbezogen werden müssen. Dies ist insofern hilfreich, als damit die Notwendigkeit besteht, die Forschungsbemühungen auch zum Zwecke verbesserter Kenntnisse in diesem Bereich voranzutreiben.

5.2.1 Der Stellenwert der außerberuflichen Motive

Wenn damit die grundsätzliche Einbeziehung außerberuflicher Motive naheliegt, so soll dennoch betont werden, daß sie nicht als Hebel benutzt werden dürfen, um eine Übermittlung von Einzeldatensätzen an die Wissenschaft unter akzeptablen Bedingungen zu verhindern. Mit anderen Worten, es wäre sicherlich nicht im Sinne von § 16 Abs.6 BStatG, eine Situation zu konstruieren, bei der ein Wissenschaftler durch ein als kriminell zu bezeichnendes Verhalten, beispielsweise für Zwecke der Erpressung, einen bestimmten Einzeldatensatz deanonymisiert, um dann daraus abzuleiten, daß prophylaktisch ein entsprechend hoher Anonymisierungsgrad gewährleistet sein muß, der ein derartiges Vorgehen unmöglich macht. Dies wäre diskreditierend und mit den Absichten des Gesetzgebers nicht konform. Die unterstellten Motive müssen daher einen realistischen "Hintergrund" beibehalten und aus dem Untersuchungsgegenstand ableitbar sein. Das rein numerische Übergewicht von drei außerberuflichen Motiven zu zwei auf beruflichen Motiven basierenden

Situationen, die an späterer Stelle einer Kosten-Nutzen-Analyse unterzogen werden, entstand, weil zum einen zu wenige berufliche Motive vorliegen und zum anderen aus der Vielzahl der hypothetisch denkbaren außerberuflichen, einige häufig genannte aufgegriffen werden sollten, auch stellvertretend für eine Reihe anderer, ähnlicher Motive. Außerdem sollte grundsätzlich das ganze Spektrum möglicher Deanonymisierungsstrategien abgedeckt werden.

5.2.2 Charakterisierung der außerberuflichen Motive

Im Gegensatz zu den beruflichen, bieten die nichtwissenschaftlichen ein weites Feld hypothetisch denkbarer Motive. Der Phantasie sind im Prinzip keine Grenzen gesetzt. Das Spektrum reicht von reiner Neugierde über eigene monetäre Vorteile oder wirtschaftliche Schädigung Dritter bis hin zu eindeutig kriminellen Aktivitäten. Es erscheint wenig ergiebig, an dieser Stelle eine mehr oder weniger vollständige Aufzählung derartiger Gründe zu versuchen; vielmehr soll auf einige Besonderheiten und auf Unterschiede zu den beruflichen Motiven eingegangen werden.

Bei den Verwertungsmöglichkeiten deanonymisierter Daten für persönliche (außerberufliche) Zwecke kann zwischen interner (z.B. Hacker-Ambitionen, das heißt, Freude über das Gelingen einer Deanonymisierung) und externer Verwendung (z.B. Verunglimpfung Dritter) unterschieden werden. Bei interner Verwendung erfolgt im Prinzip keine Offenlegung der Herkunft der gewonnenen Information. Eine Bestrafung oder sonstige Sanktionierung der Deanonymisierung ist in diesen Fällen nur möglich, wenn durch eingebaute technische Sicherungsmaßnahmen der Mißbrauch der Daten nachgewiesen werden kann: D.h., es müßten Sicherungsmaßnahmen vorliegen, die Zugangsberechtigten eine derartige Zweckentfremdung erschweren. Dann müßte der Angreifer auch bei interner Verwendung Nachteile ins Kalkül stellen. Bei der externen Verwendung, von der unmittelbar Dritte betroffen werden, gibt es solche Formen, bei denen die Quelle nicht offengelegt wird. Dann ist die Gefahr einer Sanktionierung gering. Dies gilt z.B. für die anonyme Publikation von Informationen, die etwa aus Datensätzen von "Prominenten" gewonnen werden. Darüber hinaus gibt es Formen, die - möglicherweise im Verlauf eines Strafverfahrens - zu einer Offenlegung führen könnten, z.B. bei Verunglimpfung, Rufmord oder merklicher wirtschaftlicher Schädigung. Mit der Notwendigkeit zur Offenlegung bzw. der steigenden Gefahr, daß die Quelle der Informationen ermittelt wird, wächst für den Angreifer das Risiko einer (zusätzlichen) Bestrafung und sonstiger Sanktionen. Mit anderen Worten: Die Deanonymisierung ist zwar immer illegal, aber

bei beruflichen Motiven ist die Nutzungsabsicht - die wissenschaftliche Forschung - legal. Bei den außerberuflichen Motiven dagegen kann die Nutzungsabsicht legal oder illegal sein. Die Befriedigung der reinen Neugierde und die Erpressung können als (extreme) Beispiele genannt werden. Bei illegaler Verwendung deanonymisierter Daten besteht für den Angreifer das Risiko, nicht nur für die Deanonymisierung als solche, sondern zusätzlich für die Verwendung sanktioniert zu werden. Je nach Art der Verwendung könnte der damit zusammenhängende Gesetzesbruch sogar schwerwiegender sein als der Akt der Deanonymisierung selbst. Damit sind jedoch Extremsituationen angesprochen, die sehr hypothetisch sind und hier insofern nicht weiter verfolgt werden müssen, als der Gesetzgeber die Datenweitergabe zum Zweck wissenschaftlicher Forschung nicht von der Sicherheit des Ausschlusses krimineller Akte abhängig gemacht hat, für die ein Unverhältnismäßigkeitskriterium in keiner vernünftigen Weise formuliert werden kann.

Diskrepanzen zwischen den beiden Motivkategorien gibt es damit auch beim Versuch der Nutzenermittlung einer Reidentifikation. Können die materiellen und immateriellen Vorteile im wissenschaftlichen Bereich zumindest beschrieben werden, ist dies bei den außerberuflichen Motiven nicht mehr möglich, sondern hängt stark vom jeweiligen Motiv ab. Neben materiellen Beweggründen, z.B. durch Adressenverkauf, können Antriebskräfte wie Neugier, Rache, Haß u.s.w. eine wesentliche Rolle spielen. Eine allgemeine Beschreibung oder Bewertung des Nutzens ist bei persönlichen Gründen daher ein aussichtsloses Unterfangen und nur abgestellt auf eine spezielle Situation in groben Zügen möglich. Die Analyse der Vor- und Nachteile, bzw. der Nutzen und Kosten einer Deanonymisierung ist deshalb fast ausschließlich auf die Kostenseite angewiesen und besonders auf einen Kostenvergleich alternativer Datenbeschaffungsmöglichkeiten. Derartige Alternativen können zwar auch nur fallbezogen dargestellt werden, aber als möglicherweise billigere, bessere und legalere Informationsquellen bieten sich für die einigermaßen realistischen Motive persönliche Recherchen bei Kollegen oder Nachbarn, Anfragen bei zugänglichen Auskunfteien (z.B. der SCHUFA), Einschalten von Privatdetektiven u.s.w. geradezu an. Es kann - wie die weitere detaillierte Analyse zeigen wird - davon ausgegangen werden, daß das aus außerberuflichen Motiven resultierende Risiko einer Deanonymisierung aus Kostengründen im allgemeinen unwahrscheinlich erscheint.

5.2.3 Ausgewählte außerberufliche Motive

Auf der Basis der grundsätzlichen Überlegungen wurden für eine detaillierte Untersuchung der außerberuflichen Motive drei Situationen ausgewählt, welche die vorgesehene lückenlose Einbeziehung aller Deanonymisierungsstrategien gewährleisten und sich im Rahmen der realistischen und hier einzubeziehenden Motive bewegen. Ferner werden sie in der Diskussion oft genannt und decken auch das Spektrum der geschilderten Charakteristika weitgehend ab. Die erste Situation stellt darauf ab, daß aus persönlicher Neugier die bereits vorhandene Information über einen Bekannten, im unterstellten Fall einen Arbeitskollegen, erweitert werden soll. Dabei handelt es sich um eine Verwendung der Daten, die durch die Strategie einer "Gezielten Suche" gewonnen werden sollen, ohne daß die Offenlegung der Quelle notwendig ist. Ein anderes Motiv, das ein gleiches Vorgehen nach sich ziehen würde, wäre z.B. die Informationsbeschaffung über einen Marktkonkurrenten, zur Gewinnung von Marktvorteilen. Auch Verwendungen wie Herabwürdigung des Ansehens einer bestimmten Person aus Rachegefühlen, könnten als Anlaß für eine gezielte Suche angeführt werden. Der Übergang von der legalen zur illegalen Verwendung ist fließend. Je nach unterstelltem Motiv ergeben sich bei der Kosten-Nutzen-Analyse gewisse Modifikationen.

Als Beispiel für einen persönlichen Grund, der zu einer Reidentifikation mittels Einzelfischzug führt, kann auf die Gewinnung von Informationen über einen "Prominenten" mit dem Ziel der anonymen Publikation hingewiesen werden. Dabei liegt eine externe, in das Illegale tendierende Verwendung vor, bei der jedoch das Risiko, "entlarvt" zu werden gering ist.

Die dritte Situation stellt schließlich auf das Motiv der Erzielung eines ökonomischen Vorteils ab. Dabei wird unterstellt, daß in Zusammenarbeit mit einem Adressenhändler über eine(n) Massenidentifikation/-fischzug versucht wird, Informationen aus dem Mikrozensus zu gewinnen, die dann an den Adressenhändler weiterverkauft werden können. Bei dieser Verwendung muß die Offenlegung der Quelle gegenüber dem potentiellen Käufer angenommen werden. Die Art der Datenverwendung selbst ist aber wiederum im legalen Bereich angesiedelt.

Die Ergebnisse der Analyse dieser drei konkreten Situationen bieten eine ausreichende Basis für die Beurteilung der außerberuflichen Motive als potentielle Gefahrenquelle einer Deanonymisierung und die daraus zu ziehenden Folgerungen für die Anwendung von Anonymisierungsverfahren.

5.2.4 Der Nachweis der prinzipiellen Deanonymisierbarkeit eines Mikrodatenfile als eigenständiges wissenschaftliches Motiv

Im Zusammenhang mit der Diskussion um die Volkszählung 1987 wurden im Fachbereich Informatik der Universität Hamburg Reidentifikationsexperimente mit auf Volkszählungsmerkmale abgestellten synthetischen Daten simuliert. Sie hatten das Ziel zu untersuchen, "ob die vom Bundesverfassungsgericht geforderte faktischen Anonymität der Volkszählungsdaten tatsächlich gewährleistet ist".[25] Es ist daher denkbar, daß seitens der Wissenschaft ähnliche Versuche mit von der amtlichen Statistik übermittelten faktisch anonymisierten "echten" Daten unternommen werden. Ein solches Motiv kann jedoch nicht Grundlage für die vorzunehmenden Anonymisierungsmaßnahmen sein, sondern muß als verbleibendes Restrisiko getragen werden (vgl. dazu Brennecke 1980). Darin liegt letztlich der Unterschied zwischen der faktischen und der absoluten Anonymität. Die Formulierung des § 16 Abs.6 BStatG macht deutlich, daß eine Deanonymisierung nicht gänzlich ausgeschlossen sein muß, bzw. daß sie prinzipiell möglich ist, aber eben nur unter ganz bestimmten Voraussetzungen. Das verbleibende Risiko ist demnach durch § 16 Abs.6 BStatG abgedeckt.

Das Motiv des Nachweises einer prinzipiellen Deanonymisierbarkeit stellt einen Spezialfall dar, bei dem das vom Gesetzgeber formulierte Unverhältnismäßigkeitskriterium nicht anwendbar ist, weil hier der Kostenaspekt selbst nicht in einer üblichen Weise ins Kalkül einbezogen werden muß. Ein Wissenschaftler, z.B. aus dem Bereich der Informatik, könnte seine gesamten beruflichen Forschungsanstrengungen, die ihm zur Verfügung stehende apparative Ausstattung und die Mitarbeiter seines Instituts dafür einsetzen, neue Deanonymisierungsroutinen zu entwickeln. Der Nachweis der prinzipiellen Machbarkeit wäre demzufolge das Ziel seiner Forschungsarbeiten. Das Unverhältnismäßigkeitskriterium wird deshalb unterlaufen, weil es mit einem solchen Forschungsziel sinnvoll sein kann, Deanonymisierungsversuche durchzuführen, wie hoch auch immer die Kosten dafür sein mögen. Der Angreifer selbst müßte für diese Kosten selbst nicht aufkommen, da sie voraussichtlich von seinem Institut getragen würden.

Bei dem Versuch des Nachweises der prinzipiellen Machbarkeit von Deanonymisierungen können auch Zielsetzungen verfolgt werden, mit denen die Diskreditierung der Regelung einer Freigabe von Mikrodaten angestrebt wird oder auf Intentionen eines verstärkten Datenschutzes abgestellt wird. Zwar müßte

---

25 Fischer-Hübner (1986): Vorwort von Brunnstein, S.II.

der Angreifer seine Vorgehensweise offenlegen und damit grundsätzlich erhebliche Strafen und sonstige Sanktionen befürchten. Es ist aber auch denkbar, daß durch die Verbindung des Vorhabens mit Anliegen des Datenschutzes der Schaden begrenzt werden kann. Von einer unter anderem daraus resultierenden Aussetzung oder Erschwerung der Weitergabe faktischer anonymisierter Daten wäre der Angreifer selbst nicht betroffen, da er nicht auf dem entsprechenden Wissenschaftsgebiet tätig ist.

Der Nutzen aus einer erfolgreichen Deanonymisierung ist dagegen nicht abzuschätzen. Ein Angreifer kann unter Umständen erhebliche berufliche Anerkennung gewinnen. Der Nutzen wäre dann als durchaus hoch einzuschätzen. Auf der anderen Seite geht der Wissenschaftler trotz der geschilderten Schadensbegrenzungsstrategie auch ein unkalkulierbares Risiko ein, weil er mit der Deanonymisierung einen illegalen Akt begeht. Aus diesen unberechenbaren Konstellationen ist keine eindeutige Nutzenfunktion abzuleiten.

Die Einbeziehung des Motivs des Nachweises der prinzipiellen Machbarkeit der Deanonymisierung würde die vom Gesetzgeber mit dem §16 Abs.6 BStatG verfolgte Intention geradezu torpedieren. Mit der Regelung der *faktischen* Anonymität hat der Gesetzgeber ja implizit anerkannt, daß unter bestimmten Voraussetzungen Deanonymisierungen möglich sind und er bereit ist, solche Fälle als verbleibendes Risiko hinzunehmen, wenn sie nur unter Einsatz unverhältnismäßiger Mittel zustande kommen können. Der hier untersuchte Fall kann insofern darunter subsumiert werden, als der Angreifer gerade bereit ist, unverhältnismäßige Mittel einzusetzen. Sie rechtfertigen sich aus dem Nachweis der prinzipiellen Deanonymisierbarkeit, bei der eine völlig andere Kosten-Nutzen-Rationalität bestehen kann.

6 Das Zusatzwissen für Deanonymisierungsvorhaben im Bereich der Wissenschaft*

Einer der wichtigsten Bestimmungsfaktoren des Risikos einer Reidentifikation ist das sogenannte Zusatzwissen, über das ein Datenangreifer verfügt. Darunter verstehen wir die Informationen, die dem Datenangreifer über einzelne namentlich identifizierbare Personen bekannt sind und die als Überschneidungsmerkmale auch im Mikrodatenfile enthalten sind. Das allgemeine Vorgehen bei einem Deanonymisierungsversuch besteht darin, durch den Abgleich dieser Überschneidungsmerkmale Fälle aus dem Mikrodatenfile, die der persönlichen Identität nach unbekannt sind, den namentlich bekannten Fällen des Zusatzwissens (bzw. des Identifikationsfile) zuzuweisen und damit zu identifizieren.[2]

Eine Untersuchung des Reidentifikationsrisikos muß sich deshalb besonders gründlich mit der Verfügbarkeit und Nutzbarkeit solchen Zusatzwissens befassen. Da es im Zusammenhang dieser Untersuchung um faktische Anonymität von Mikrodaten geht, die an die Wissenschaft übermittelt werden, müssen im Mittelpunkt des Interesses Situationen stehen, die für den Bereich wissenschaftlicher Institutionen typisch sind, und zwar sowohl was die Motive für einen Deanonymisierungsversuch, als auch was das im wissenschaftlichen Bereich typischerweise zur Verfügung stehende Zusatzwissen betrifft.

Als potentiell wissenschaftliche Deanonymisierungsmotive wurden im vorausgegangenen Kapitel die Gewinnung einer Stichprobe oder die Ergänzung eines eigenen Datenfile mit Informationen aus einem Mikrodatenfile herausgearbeitet. Bei einem solchen Deanonymisierungsszenario müßte eine große Zahl von

---

* Autorin: Petra Beckmann

2 Abweichend von dieser allgemeinen Problemlage ist es bei einem sehr kleinen Kreis von Personen nicht erforderlich , daß Namen und Adressen im Zusatzwissen vorhanden sind, da diese wenigen Personen bereits aufgrund ganz bestimmter mittelbarer Identifikatoren deanonymisierbar sind. Dieses Problem wird in die Betrachtung mit eingeschlossen, aber nicht gesondert behandelt. Als Beispiel wäre hier an eine Sozialdatenerhebung für die Stadt Stuttgart zu denken, die einen männlichen Kommunalbeamten im Alter von über 50 Jahren mit einem Einkommen von über 100 000 DM ausweist, so daß klar erkennbar ist, daß es sich hier um den Oberbürgermeister der Stadt Stuttgart handeln muß. Das Beispiel verdeutlicht jedoch, daß auf diese Weise immer nur Einzelfälle identifiziert werden können, die in der vorliegenden Analyse nicht im Mittelpunkt des Interesses stehen (vgl. Kapitel 5).

Fällen reidentifiziert werden, um überhaupt dem Zweck des Datenangriffs gerecht werden zu können, d.h. es muß ein Datenangriff in Form einer Massenidentifikation oder eines Massenfischzuges vorausgesetzt werden.

Was das verfügbare Zusatzwissen betrifft, ist zunächst darauf hinzuweisen, daß es für den wissenschaftlichen Bereich geradezu charakteristisch ist, daß dort in aller Regel wenig Informationen über namentlich bekannte Individuen vorliegen, ganz im Unterschied etwa zu Eingriffsbehörden oder auch Institutionen des privatwirtschaftlichen Sektors. Solche Institutionen müssen, zum Teil zum Zweck des Erfüllens ihrer Aufgaben, Daten in expliziter Verbindung mit identifizierbaren Personen anlegen und aufbewahren. Ein Sparer ist in der Regel daran interessiert, daß die Einzahlungen, die er tätigt, auf einem eindeutig seiner Person zuzuordnenden Konto verbucht werden, in ähnlicher Weise ein Versicherungsnehmer oder ein Steuerzahler. Wie schon die Untersuchung der Handlungslogik der Sozialwissenschaften deutlich gemacht hat, ist das wissenschaftliche Interesse dagegen nicht auf die Erkenntnis über namentlich bekannte Individuen, Haushalte oder Institutionen ausgerichtet, sondern lediglich auf solche Einheiten als anonyme Fälle und Merkmalsträger. Deshalb sind sozialwissenschaftliche Forschungsstellen auch nicht Einrichtungen, in denen sich große Informationsmengen über namentlich bekannte Personen anhäufen.

Im wesentlichen sind es zwei Arten von Informationen, die im sozialwissenschaftlichen Bereich vorhanden oder zugänglich sind und die als Zusatzwissen verwandt werden können. In der Literatur wird im Hinblick auf das Zusatzwissen immer wieder auf Wissen hingewiesen, welches aus Daten stammt, die "*allgemeinkundig*" ("offenkundig") bzw. aus "*allgemein zugänglichen Quellen*" beschafft oder beschaffbar sind (Hammerbacher 1982:287). Bei diesen Daten handelt es sich in erster Linie um Angaben, die aus öffentlichen Registern, prozeßduzierten Datensammlungen (Bick/Müller 1977) und jedermann zugänglichen privaten Datensammlungen (Adreßbücher, Telefonbuch etc.) zu entnehmen sind.

Der Wissenschaftler verfügt zusätzlich zu den genannten Datenquellen, die jedem Bürger zugänglich sind, professionsgemäß noch über die Daten, die er in *eigenen empirischen Untersuchungen* erhoben hat. Auch diese Daten sind - soweit sie selbst personenbezogen sind - Bestandteil des Zusatzwissens, das dazu verwendet werden könnte, die anonymisierten Einzeldaten, die dem Wissenschaftler von seiten der statistischen Ämter übermittelt werden, zu deanonymisieren.

In diesem Kapitel soll im einzelnen untersucht werden, welches Zusatzwissen in den Sozialwissenschaften typischerweise vorhanden ist und welche Voraussetzungen für seine Nutzung zu einer Deanonymisierung von anonymisierten Einzeldaten gegeben sind. Im dritten Teil dieses Kapitels werden dabei wichtige, einschlägige öffentliche und private Register zunächst hinsichtlich ihres rechtlichen Zugangs, ihres Informationsgehalts, ihrer Aktualität sowie ihrer Überschneidungsmerkmale mit den Daten des Mikrozensus und der EVS dargestellt. In einem Unterpunkt erfolgt eine detailliertere Auseinandersetzung mit dem Problem, wie sich die Zugangsmöglichkeiten zu Daten aus öffentlichen Registern speziell für die Wissenschaft in der Praxis darstellen und welche Konsequenzen sich daraus für das unterstellte Szenario ergeben.

Die in den Sozialwissenschaften zusätzlich verfügbaren Daten, die vom Forscher selbst erhoben werden, sollen im vierten Teil des Kapitels unter den gleichen Gesichtspunkten wie die Daten aus öffentlichen Registern untersucht werden.

Abschließend sollen die Ergebnisse, die die Analysen des Zusatzwissens hinsichtlich seiner Verwendbarkeit im Deanonymisierungsprozeß erbracht haben, kurz resümiert werden.

Bevor jedoch der Informationsgehalt der in den Registern und privaten Dateien vorhandenen Überschneidungsmerkmale genauer analysiert wird, sollen vorab zwei Sachverhalte nochmals kurz in Erinnerung gerufen werden, von denen die Nutzbarkeit eines Zusatzwissens für Reidentifikationen wesentlich abhängt. Sie wurden bereits in Kapitel 4 ausführlich diskutiert. Zum einen geht es um die mögliche Fehlerhaftigkeit des Zusatzwissens und um die Kompatibilität des Zusatzwissens mit den Informationen im Mikrodatenfile. Zum anderen geht es um die Frage des Erfassungsgrades einer Population im Zusatzwissen. Handelt es sich um eine vollständige Erfassung einer durch bestimmte Kriterien abgegrenzten Population oder um Stichproben? Welches sind die Kriterien, nach denen ein Populationsmitglied im Zusatzwissen enthalten ist oder nicht?

6.1 Kompatibilität und Umfang von Mikrodatenfile und Identifikationsfile

6.1.1 Kompatibilität

Wie schon in Abschnitt 4.3 ausgeführt und wie insbesondere bei den empirischen Untersuchungen in Kapitel 9 deutlich werden wird, ist einer der wesentlichsten Bestimmungsgründe dafür, ob es zu Reidentifikationen kommen kann, daß die Überschneidungsmerkmale im Zusatzwissen (Identifikationsfile) in gleicher Weise abgebildet sind wie im Mikrodatenfile. Wie schon diskutiert, kann es viele Gründe geben, weshalb es zu Inkompatibilitäten in der Abbildung eines Sachverhaltes im Zusatzwissen und im Mikrodatenfile kommen kann. An dieser Stelle soll nicht die Kompatibilitätsproblematik insgesamt nochmals diskutiert werden, sondern nur auf einzelne Punkte hingewiesen werden, die zu berücksichtigen sind, wenn die Nutzbarkeit einzelner Arten von Zusatzwissen für Reidentifikationszwecke einzuschätzen ist. Neben eher banalen Ursachen für Inkompatibilitäten (Übermittlungs- und Kodierfehler, Fehlinterpretation von Merkmalsdefinitionen etc.) ist besonders auf Inkompatibilitäten zu achten, die sich aus spezifischen Zwecksetzungen und den spezifischen operationalen Prozeduren ergeben, unter denen ein Zusatzwissen entsteht. Dies ist vor allem bedeutsam, wenn es sich beim Zusatzwissen um eine Datenbasis handelt, die in Form eines Registers für einen spezifischen Zweck zustande gekommen ist. Als Faktoren, die die Kompatibilität beeinträchtigen können, sind unter anderem wichtig:

- Das bewußte oder unbewußte Interesse der Personen oder Institutionen, die Informationen für die Quelle des Zusatzwissens liefern, sich bei der informationssammelnden Stelle in einer bestimmten Weise darzustellen. Dieser Aspekt ist dann von besonderer Bedeutung, wenn mit einer bestimmten Falldarstellung Vor- oder Nachteile verbunden sein können. Ein solches Interesse kann bei vielen Datenbasen insbesondere des Verwaltungsvollzuges, aber auch bei privaten Informationssammelstellen, bei denen es etwa um eine günstige werbende Darstellung geht, angenommen werden. Ein ähnliches Interesse kann man analog auch bei informationssammelnden Stellen annehmen.

- Der zeitliche Bezug der Informationen im Zusatzwissen und seine Übereinstimmung mit dem Zeitbezug der Informationen des zu deanonymisierenden Mikrodatenfile. Insbesondere bei Registern hängt der Zeitbezug mit den Anlässen zusammen, die zu einem Registereintrag oder zu einer

Veränderung der Registrierung führen. So beziehen sich beispielsweise die Befragungen im Rahmen des Mikrozensus in der Regel auf einen bestimmten, einheitlichen Erhebungszeitpunkt. Den Angaben in öffentlichen und privaten Datenquellen liegt dagegen in der Regel ein personenspezifischer und damit variabler Erhebungszeitpunkt zugrunde, wie etwa der Tag, an dem jemand einem Verein beigetreten ist, geheiratet oder einen Telefonanschluß beantragt hat. Dieser Tatbestand kann zur Folge haben, daß die Angaben zur Person völlig veraltet sind. Jemand kann bereits seit 20 Jahren Mitglied eines Vereins sein, ebensolange verheiratet sein, und auch der Telefonanschluß muß sich in diesem Zeitraum nicht geändert haben. Da die Register in der Regel nur dann aktualisiert werden, wenn bezüglich der betreffenden Person registererhebliche Veränderungen eintreten, dürfte es nahezu unmöglich sein, einen völlig übereinstimmenden inhaltlichen Bezug zwischen den Daten in den Registern und den Daten im Mikrodatenfile herzustellen. Beispielsweise kann die Berufsangabe im Melderegister nur dann geändert werden, wenn der Betreffende dem Einwohnermeldeamt seinen Auszug aus einer bzw. Einzug in eine Wohnung mitgeteilt hat.

- Die Definition der Einheiten des Zusatzwissens. Während die Einheiten der hier interessierenden Mikrodatenfiles in der Regel Haushalte oder individuelle Personen sind, gibt es viele Register, deren Erhebungseinheiten Vereine, Genossenschaften, Betriebe, Berufe, Grundstücke etc. sind. Der konkrete Personenbezug muß demnach in vielen Fällen erst hergestellt werden.

- Die örtliche Lokalisierung der im Zusatzwissen erfaßten Informationen und die Kompatibilität dieser Lokalisierung mit den Informationen im Mikrodatenfile. Während der lokale Bezug der Personen im Mikrodatenfile in der Regel der Wohnort ist, ist der Ortsbezug des Zusatzwissens oftmals die Betriebsstätte oder der Ort, an dem ein von einer bestimmten Person erworbenes Grundstück steht.

Während bei allen diesen Punkten das Zusatzwissen korrekt im Sinne der Zwecksetzung einer Informationssammlung sein kann, können die spezifischen Konstruktionsprinzipien aber dennoch zu weitreichenden Inkompatibilitäten mit der Abbildung von Sachverhalten im Mikrodatenfile führen. Wie weitreichend solche Inkompatibilitäten sind, müßte in jedem einzelnen Fall gesondert untersucht werden. Neben Inkompatibilitäten, die auch bei sachgerechter Abbildung einzig wegen des unterschiedlichen Zweckes der Abbildung zustande kommen können, ist aber auch auf Inkompatibilitäten hinzuweisen, die aus fehlerhaften -

diesem regional begrenzten Gebiet zu reidentifizieren. Betont werden muß aber, daß auch dieses Beispiel wieder von der Grundannahme der hier durchgeführten Analyse abweicht und daher ausgeklammert werden muß.

Für die Suchstrategie der Sozialwissenschaftler resultiert aus oben Genanntem, daß ein zur Verfügung stehender Identifikationsfile aus öffentlichen Registern in keinem Fall so groß sein wird, daß bei den gegebenen Überschneidungsmerkmalen eine erfolgreiche Deanonymisierung für mehr als einen Zufallstreffer möglich erscheint (vgl. hierzu auch Abschnitt 6.1.2).

An dieser Stelle sei im Vorgriff auf spätere Teile der Studie ein Hinweis auf die Kostenseite angefügt. Der Versuch, über öffentliche Register ein Zusatzwissen aufzubauen, mit dem vielleicht eine größere Zahl von Reidentifikationen möglich würde, wäre mit einem ganz erheblichen Aufwand verbunden. Vor allem bei Registern, die dezentral geführt werden, wären ganz unabhängig von den technischen Arbeiten der Reidentifizierung mit Sicherheit ungewisse und langwierige Verhandlungen erforderlich, um den entsprechenden Datenzugang zu erreichen. Schon dies dürfte abschreckend wirken.

In bezug auf die Problematik des privaten Adressenhandels und die Frage nach dem dort verfügbaren Zusatzwissen ergab die Rücksprache mit einigen Anbietern, daß das dort erhältliche Zusatzwissen in der Regel aus der Nutzung von öffentlich zugänglichen Datenquellen besteht, wie sie in Tabelle B.I aufgelistet sind. So werden etwa Anschriften, Telefonnummern und Berufsbezeichnungen aus den Angaben in öffentlichen Telefonbüchern und Branchenfernsprechbüchern bezogen; die Daten über Selbständige stammen z.T. aus für diesen Bereich erhältlicher spezifischer Fachliteratur (z.B. Veröffentlichungen der Ärztekammer). Als weitere Quellen, auch für den internationalen Raum, wurden genannt: Firmen ABC, Veröffentlichungen der deutschen Botschaften im Ausland, Geschäftsberichte von Firmen und Publikationen von Universitäten.

Die meisten dieser Datenquellen bieten, abgesehen von der freiwilligen Berufsangabe, den Regionalangaben und dem zum Teil indirekt enthaltenen Wirtschaftszweig, kaum weitere personenspezifische Merkmale, die für eine Deanonymisierung des Mikrozensus von Belang sind. Altersangaben sind z.B. nur für ganz wenige Personen erhältlich, die in der Regel aus freiwilligen Angaben in Preisausschreiben oder ähnlichem von einzelnen Firmen gewonnen wurden. Diese Adressen, die die jeweiligen Firmen "besitzen" und den Adressenhändlern zum Weiterverkauf anbieten, werden als sog. "Broker-Adressen" gehandelt. Das bedeutet, daß ein potentieller Käufer dieser Adressen das

Material aus Konkurrenzausschlußgründen nicht selbst in die Hand bekommt, sondern daß sich z.B. die Abwicklung einer Werbekampagne direkt über den Adressenhändler vollzieht.

Eine Auflistung von "kaufkräftigen Privatpersonen", die von vielen Adressenhändlern angeboten wird, basiert ebenfalls auf den Angaben zum Beruf und evtl. zur Stellung im Beruf. Dementsprechend wird eine willkürliche Zuordnung dieser Personen zur Gruppe der "kaufkräftigen Privatpersonen" vorgenommen und nicht etwa, wie ebenfalls zu vermuten wäre, aufgrund von vorhandenen Angaben zum Einkommen.

Von den befragten Adressenhändlern wurde darüber hinaus noch auf zwei wichtige Probleme bezüglich der Datenqualität hingewiesen: (1) die Repräsentativität und (2) die Aktualität der Angaben.

Hinsichtlich der Repräsentativität kann man davon ausgehen, daß maximal 85 bis 95 Prozent der Angehörigen einer Berufsgruppe erfaßt sind. Die Aktualität der Daten ist insofern problematisch, als sich allein 35 bis 40 Prozent der Eintragungen in einem Telefonbuch innerhalb eines Jahres ändern.

Zusammenfassend muß man das Angebot der Adressenhändler als schnell verfügbar und relativ kostengünstig (ca. 130 DM/1 000 Adressen) bewerten. Für den Zweck einer massenhaften Deanonymisierung des Mikrozensus reichen die so erhältlichen personenbezogenen Merkmale jedoch nicht aus.

Die bisherigen Ergebnisse zur Verwendbarkeit des Zusatzwissens aus öffentlichen Registern und privaten Datenquellen im Deanonymisierungsprozeß lassen sich folgendermaßen resümieren:

1) Die Register und Dateien enthalten die unterschiedlichsten Arten von Dateninkompatibilitäten, die das Risiko einer falschen Zuordnung von Angaben des Identifikations- zu Angaben des Mikrodatenfile erhöhen.

2) Handelt es sich sowohl beim Identifikationsfile als auch beim Mikrodatenfile um eine Stichprobe, werden die Chancen einer korrekten eindeutigen Zuordnung weiter herabgesetzt.

3) Auch aufgrund der vergleichsweise geringen Anzahl der vorhandenen Überschneidungsmerkmale, ihres meist eher niedrigen Informationsgehaltes sowie der oftmals geringen Aktualität der Informationen, besteht in den öffentlich zugänglichen Registern und privaten Datenquellen kein Zusatzwissen, aus dem sich hohe Reidentifikationsrisiken ergeben.

4) Bei den Registern, die eventuell für eine Reidentifikation verwendbar wären, schließt die restriktive Zugangsregelung eine Massenidentifikation/einen Massenfischzug aus.

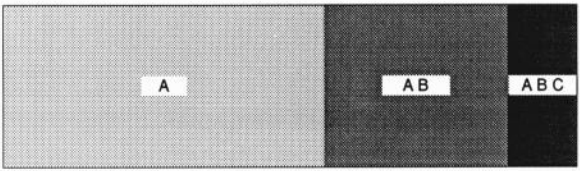
6.2.3 Kumulation von Informationen aus verschiedenen Registern/privaten Datenquellen

Nach diesen Ergebnissen besteht allenfalls für eine sehr begrenzte Zahl und einen spezifischen Typus von Registern eine gewisse Wahrscheinlichkeit, daß mit Hilfe der Informationen eines einzelnen, dem Wissenschaftler zugänglichen Registers, aussichtsreiche Deanonymisierungsversuche möglich sein könnten. Ergibt sich nun ein anderes Bild durch Versuche, personenbezogene Daten, die aus unterschiedlichen Registern zu Forschungszwecken zur Verfügung gestellt werden, miteinander zu kombinieren? Durch die Kumulation von Daten besteht die Möglichkeit, ein bereits existierendes Identifikationsfile um zusätzliches Wissen zu erweitern und den Datensatz damit zu verlängern. Infolge dieser vergrößerten Informationsbasis könnte sich auch das Risiko erhöhen, Datensätze eines Mikrodatenfile deanonymisieren zu können. Inwiefern es sich hierbei um eine realistische Annahme handelt, soll im folgenden detailliert diskutiert werden.

Die Kumulation verschiedener Register setzt zunächst voraus, daß ein Forscher bereits für die Datenbeschaffung einen erheblichen Aufwand an Organisation, Zeit und finanziellen Kosten tragen muß. Dieser Aufwand steigt mit der Anzahl der zu vergleichenden Register. Außerdem ist zu berücksichtigen, daß die meisten Verwaltungsdateien verstreut und ohne Verbindung zueinander geführt werden. Da eine einheitliche Personenkennummer in der Bundesrepublik Deutschland nicht vorhanden ist, dürfte eine Zusammenführung oder ein Abgleich verschiedener Register mit einem erheblichen organisatorischen Aufwand verbunden sein, einmal abgesehen von den dabei bestehenden rechtlichen Problemen. Die Tatsache, daß beispielsweise die meisten Einwohner- und Gebäudedateien dezentral geführt werden, bedeutet eine zusätzliche Erschwernis (Redfern 1986, Benker 1985, Bretz/Wedel 1987).

Ein zweites zentrales Problem entsteht aus der Tatsache, daß immer nur ein Bruchteil der Personen, für die Daten aus dem Register A vorliegen, in einem Vergleichsregister B enthalten sind, so daß sich der Umfang des an Informationszuwachs gewinnenden Identifikationsfile in Abhängigkeit von der Anzahl

der zum Vergleich herangezogenen Register fortlaufend reduziert. Schematisch läßt sich dieser Vorgang etwa wie folgt darstellen:



Personen, die in Register A enthalten sind

Personen, die in Register A und B enthalten sind

Personen, die in Register A, B und C enthalten sind

Neben den oben genannten Schwierigkeiten tritt noch ein drittes, bedeutendes Problem auf, welches die Kumulation von Fehlern beim Vergleich veschiedener Register betrifft. Man muß - wie oben ausgeführt - davon ausgehen, daß auch in öffentlichen Registern Fehler enthalten sind. Ihr Ausmaß dürfte von Datei zu Datei variieren und läßt sich somit zahlenmäßig nicht genau angeben. Im Vergleich zwischen dem Identifikationsfile (Register) und dem Mikrodatenfile ergeben sich darüber hinaus weitere Inkonsistenzen, die aus unterschiedlichen Erhebungszeitpunkten, Erhebungseinheiten und der uneinheitlichen Definition von Merkmalen resultieren. Diese Fehler kumulieren mit der wachsenden Zahl an Vergleichsregistern, so daß die Richtigkeit der erweiterten Daten zum Teil stark angezweifelt werden muß.

Zieht man nun zur Konkretisierung des Problems ein *fiktives Beispiel* heran, in dem ein Forscher versucht, möglichst viele Ärzte im Mikrozensus zu reidentifizieren, so könnte er als Identifikationsfile z.B. auf die Kombination von Melderegister, Arztregister und Familienbuch zurückgreifen. Daß ein Zugang zu den Daten der genannten Register in der Weise, wie er in dem Beispiel vorausgesetzt wird, real nicht besteht, hebt noch einmal hervor, daß es sich in dem gewählten Beispiel um eine reine Fiktion handelt. Aus Tabelle A.II (im Anhang) geht hervor, daß das Melderegister insgesamt acht Überschneidungsmerkmale mit dem Mikrozensus aufweist, das Arztregister zwar insgesamt fünf Über-

schneidungsmerkmale mit dem Mikrozensus beinhaltet, der tatsächliche Zugewinn an Informationen vom Melderegister zum Arztregister jedoch nur in Form von zwei Variablen (Beruf, Stellung im Beruf) besteht. Das gleiche gilt für das dritte Register, das Familienbuch, das sechs Überschneidungsmerkmale mit dem Mikrozensus aufweist, aber im Vergleich mit den anderen beiden Registern nur zwei zusätzliche Variablen (Eheschließungsjahr, Kinder des Haushaltsvorstands) enthält. Die Gesamtzahl der so gewonnenen Überschneidungsmerkmale mit dem Mikrozensus beträgt also zwölf. Es ist offensichtlich, daß ein sehr hoher Aufwand betrieben werden müßte, um durch die Zusammenführung der Informationen aus den drei Einzelregistern ein reichhaltiges kumuliertes Register zu entwickeln, ganz abgesehen von dem Umstand, daß massenweise Auszüge aus den Einzelregistern, inklusive dem Familienregister, nicht zur Verfügung gestellt würden.

Insgesamt läßt sich die Gewinnung eines Identifikationsfile mit Hilfe eines Abgleichs verschiedener Register als ein zu aufwendiges, kostenintensives, fehleranfälliges und wenig gewinnbringendes Verfahren beschreiben, das daher in der Praxis wohl kaum Anwendung finden dürfte.

6.3 Zusatzwissen aus Umfragen in den Sozialwissenschaften

In den Sozialwissenschaften ist zunächst zu unterscheiden zwischen zwei Arten von Daten, die der Forschung zur Verfügung stehen: erstens Primärdaten und zweitens Sekundärdaten.

Primärdaten entstammen Untersuchungen, die der Forscher selbst durchgeführt hat. Bei solchen Untersuchungen verfügt der Forscher teilweise über Namen und Anschriften der untersuchten Personen. Sekundärdaten hingegen stammen aus Datenerhebungen, die von Dritten durchgeführt wurden. Diese Daten liegen dem Forscher in aller Regel nur in anonymisierter Form vor.

Diese generelle Unterscheidung ist vor dem Hintergrund einer potentiellen Deanonymisierung von Mikrodaten insofern wichtig, als Sekundärdaten in aller Regel von vornherein ausschließen, daß der Forscher über personenbezogene Daten verfügt. Von Bedeutung ist in diesem Zusammenhang daher auch die in der sozialwissenschaftlichen Literatur oft erwähnte Einschätzung, daß bei zunehmender Knappheit an Forschungsgeldern und den damit verbundenen geringen Chancen zur Durchführung eigener Erhebungen der Rückgriff auf

Daten, die von Dritten erhoben wurden, in Zukunft noch häufiger erfolgen wird (vgl. z.B. Porst 1985:37).

Im folgenden sollen zunächst Datenbestände kurz charakterisiert werden, bei denen der Forscher selbst Zugang zu personenbezogenen Daten erhält. Daran anschließend soll am Beispiel der "Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) eine typische Art häufig genutzter Daten dargestellt werden, bei denen für den Forscher, der diese Erhebungen nutzt, kein Personenbezug mit Hilfe von Namen und Adressen herstellbar ist.

6.3.1 Umfragen, bei denen Zugang zu pesonenbezogenen Daten besteht

6.3.1.1 Pretests

Pretests sind Voruntersuchungen, "in denen die erstellten Meßinstrumente daraufhin überprüft werden, ob sie gültige (valide) und zuverlässige (reliable) Messungen ermöglichen" (Schnell/Hill/Esser 1988:114). Ihre Durchführung gilt in allen umfangreichen empirischen Forschungsprojekten als unabdingbare Voraussetzung für die Vorbereitung einer Haupterhebung.

Pretests werden mit kleinen Fallzahlen oft vom Forscher selbst durchgeführt, um Erfahrungen mit dem verwendeten Erhebungsinstrument zu sammeln. Als Richtlinie für die Anzahl solcher Probeerhebungen, z.B. Interviews, ist von einer Zahl zwischen 20 und 30 (Friedrichs 1973:221), maximal 50 befragten Personen auszugehen.

Selbst wenn man davon ausgeht, daß Daten mehrerer Pretests verfügbar sind, muß dieser Datentyp schon von der Fallzahl her als ungeeignet für Versuche betrachtet werden, Mikrodaten zu deanonymisieren. Dies ist schon deshalb der Fall, weil es wegen der Stichprobeneigenschaft von Mikrozensus und EVS in höchstem Maße unwahrscheinlich ist, daß in einem dieser Mikrodatenfiles überhaupt eine in einer eigenen Erhebung befragte Person enthalten ist.

6.3.1.2 Eigene Haupterhebungen

Erhebungen, die Wissenschaftler in Forschungsinstituten von Universitäten oder sonstigen Einrichtungen unabhängiger Forschung in eigener Regie durchführen, haben in aller Regel ebenfalls einen kleinen Umfang mit regionalem Bezug. Solche Einrichtungen verfügen über keine eigenen Interviewerstäbe. Man kann von Stichproben mit einer Nettoausschöpfungsquote von 600-800 Fällen ausgehen. Es dürfte in der Bundesrepublik - wenn überhaupt - nur ganz wenige Institutionen der unabhängigen Forschung geben, die mehr als etwa eine solche Studie pro Jahr durchführen. Selbst wenn man davon ausgeht, daß die Überschneidungsmerkmale zu einem Mikrodatenfile so zahlreich, differenziert und fehlerfrei sind, daß eine hundertprozentige Reidentifikation möglich wäre, könnten mit Hilfe einer Stichprobe von so geringem Umfang nicht einmal zehn Personen im Mikrozensus reidentifiziert werden. Damit erscheint das Zusatzwissen aus selbst durchgeführten Befragungen herkömmlichen Ausmaßes als für Deanonymisierungsvorhaben ebenfalls vernachlässigbar.

In einigen nicht-universitären Einrichtungen (Sonderforschungsbereiche, Deutsches Jugendinstitut, Max-Planck-Institute, Deutsches Institut für Wirtschaftsforschung), werden mit fortschreitender Entwicklung in den Umfragemethoden häufiger auch Befragungen mit größeren Stichproben durchgeführt. So liegt die Größenordnung für eine Unternehmensbefragung des Deutschen Instituts für Wirtschaftsforschung (DIW) bei 1 500 bis 6 000 Befragten, bei einer Umfrage unter Berufsanfängern wurden sogar 6 000 Haushalte und 6 000 Personen schriftlich interviewt.[13] Anzumerken ist, daß es sich zum einen zwar in der Regel um regionale Befragungen handelt (im oben genannten Fall Berlin), zum anderen aber kaum Merkmalsüberschneidungen mit dem Mikrozensus vorliegen, so daß auch diese Umfragen kaum als riskantes Zusatzwissen zu werten sind.

In der forschungsüblichen Befragung werden neben einer Reihe von soziodemographischen Merkmalen, die für die Analyse wichtig sind (z.B. Geschlecht, Alter, Beruf, Schulbildung) solche Variablen erhoben, die für die konkrete Fragestellung des Forschungsvorhabens von Belang sind. Was die soziodemographischen Merkmale angeht, so ergeben sich in der Regel eine Reihe von Überschneidungen mit den Datensätzen des Mikrozensus oder der EVS; der

[13] Bei noch größeren Stichproben werden in letzter Zeit meist Telefoninterviews durchgeführt, wobei aufgrund der mangelnden technischen Ausstattung der jeweiligen sozialwissenschaftlichen Institute mit der Feldarbeit meistens spezielle Umfrageinstitute beauftragt werden. Dieses Problem wird im nächsten Abschnitt aufgegriffen.

Überschneidungsgrad aller anderen Merkmale ist jedoch in den meisten Fällen eher gering. Wäre die Überschneidung groß, könnte man direkt die Ergebnisse aus dem Mikrozensus oder der EVS verwenden und müßte keine eigene Erhebung durchführen.

Als Faktoren, die die Reidentifikationschancen herabsetzen, sind Dateninkompatibilitäten in Form von unterschiedlichen Erhebungszeitpunkten, Erhebungseinheiten sowie die unterschiedliche Definition von Merkmalen im Mikrodaten- und Identifikationsfile zu berücksichtigen.

Im Hinblick auf eine(n) angestrebte(n) Massenidentifikation/Massen- fischzug muß die Möglichkeit einer Reidentifikation mit Hilfe von Befragungen, wie sie oben beschrieben wurden, aus den genannten Gründen wohl als sehr unwahrscheinlich betrachtet werden.

6.3.2 Umfragen, bei denen kein Zugang zu personenbezogenen Daten besteht

Dem Sozialwissenschaftler stehen neben den Daten, die er in eigener Regie erhoben hat, sogenannte Sekundärdaten zur Verfügung, die aus Untersuchungen stammen, die von Umfrageinstituten durchgeführt wurden. Dabei kann es sich um einmalige oder mehrmalige Befragungen der gleichen Personen (Panel) handeln. Bei repräsentativen Bevölkerungsumfragen variieren die Stichprobengrößen zwischen 1 000 und 3 000 Fällen. Ausnahmsweise können die Stichproben auch größer sein wie z.B. beim Sozio-ökonomischen Panel (SÖP), in dem alle Personen aus ca. 5 000 Haushalten befragt werden.

Die Allgemeine Bevölkerungsumfrage der Sozialwissenschaft (ALLBUS) ist ein typisches Beispiel einer solchen Erhebung, an dem die Verwendbarkeit der Daten für den Versuch einer Deanonymisierung von Mikrozensus oder EVS näher erläutert werden kann. Diese Erhebung wird alle zwei Jahre bei einer jeweils wechselnden repräsentativen Stichprobe von ca. 3 000 Personen durchgeführt. Es besteht ein vergleichsweise hoher Grad an Übereinstimmung der im ALLBUS und im Mikrozensus oder der EVS erfaßten Variablen der Standarddemographie. So verteilen sich auf die Bereiche "Allgemeine demographische Merkmale", "Familiengeschichte", "Ausbildung", "Erwerbstätigkeit", "Einkommen", "Haushaltszusammensetzung", "Pendelwanderung" 24 Überschneidungsmerkmale zwischen dem ALLBUS und dem Mikrozensus und elf Überschneidungsmerkmale zwischen dem ALLBUS und der EVS. Allerdings

sind die Variablen nicht in jedem Fall in gleicher Weise erfaßt und verkodet. Im ALLBUS werden die Berufe z.B. nach dem ISCO-Code (International Standard Classification of Occupations) verschlüsselt, der erhebliche Abweichungen von der Berufsklassifikation des Mikrozensus aufweist. Auch die verwandte Klassifikation der Wirtschaftszweige ist nicht völlig identisch. Dennoch erscheint aufgrund der Anzahl und des Differenzierungsgrades der Überschneidungsmerkmale des ALLBUS mit dem Mikrozensus und der EVS nach den Ergebnissen der GMD eine Deanonymisierung von einzelnen Fällen durchaus möglich. Genauer geprüft werden müßte, welche Inkompatibilitäten in der Verkodung der für die Deanonymisierung besonders geeigneten Variablen bestehen.

Es gibt nun aber eine institutionelle Barriere, die solche Reidentifikationen im Wissenschaftsbereich faktisch unmöglich macht, selbst wenn der Nachweis gelingt, daß ein Datensatz aus dem ALLBUS oder dem SÖP von der gleichen Person stammt wie ein Datensatz vom Mikrozensus oder der EVS. Die Daten des ALLBUS und des SÖP werden von Umfrageinstituten erhoben, und die Namen und Anschriften verbleiben gemäß Vertrag bei den Datenerhebungsinstituten. Weder die Mitarbeiter der mit dem ALLBUS befaßten Projektgruppe noch die Mitarbeiter des SÖP haben selbst Zugang zu personenbezogenen Daten. Ganz auszuschließen ist demzufolge die Vorstellung, daß ein Dritter, also ein Forscher oder ein Institut, das mit den Daten des ALLBUS im Rahmen einer Sekundäranalyse arbeitet, auf diese Identifikationsmerkmale zugreifen kann. So gesehen fungieren die Umfrageinstitute in diesem Prozeß als Datentreuhänder.

Diese Vorkehrung bedeutet, daß ein Wissenschaftler, der die Daten des ALLBUS oder des SÖP dafür benutzen möchte, einen Mikrodatenfile der amtlichen Statistik zu reidentifizieren, zunächst selbst den ALLBUS oder das SÖP Panel mit all dem Aufwand und den Ungewißheiten, die ein solcher Versuch beinhaltet, reidentifizieren müßte. Da der ALLBUS und das SÖP vergleichsweise kleine Stichproben sind (im ALLBUS wird ca. einer von 20 000 Einwohnern, im SÖP einer von 6 000 Einwohnern befragt), werden bei diesen Datensätzen nur vergleichsweise wenig Anonymisierungsmaßnahmen getroffen. Neben dem Weglassen von Namen und Anschrift werden nur die kleinräumigen Regionalinformationen gelöscht. Sowohl im SÖP als auch im ALLBUS ist der kleinste regionale Bezug der Regierungsbezirk. Es wird davon ausgegangen, daß der geringe Umfang der Stichprobe einen ausreichenden Schutz vor Deanonymisierung darstellt.

Die für den ALLBUS und das SÖP beschriebenen Bedingungen gelten in gleicher Weise auch für andere Daten, über die Sozialwissenschaftler durch

Kooperation mit Umfrageinstituten verfügen. In den regelmäßig durchgeführten Wahlstudien, dem Wohlfahrtssurvey und anderen Bevölkerungsumfragen haben die Forscher selbst ebenfalls nur Zugang zu anonymisierten Einzelangaben. Eine Verbindung zu den befragten Personen können für eine begrenzte Zeit allenfalls die Befragungsinstitute herstellen.

In der Regel haben auch in dem Umfrageinstitut, welches die Befragung durchführt, nur wenige Personen Zugang zu den Namen und Adressen der Befragten, die auf einem Magnetband zusammen mit der Identifikationsnummer des Fragebogens gespeichert sind. Die Adressen werden nach Abschluß der Untersuchung innerhalb festgesetzter Fristen vernichtet.[14]

6.4 Fazit

Insgesamt führt die Diskussion der wichtigsten Arten personenbezogener oder personenbeziehbarer Daten, wie sie in den Sozialwissenschaften vorliegen oder zugänglich sind, zu der Einschätzung, daß damit keineswegs ein Zusatzwissen vorhanden ist, von dem eine bedeutende Gefahr der Reidentifikation von Mikrodaten aus dem Mikrozensus oder der EVS ausgehen könnte.

Vor dem Hintergrund einer Massenidentifikation/eines Massenfischzugs ergab die Untersuchung des vorhandenen Zusatzwissens aus öffentlichen Registern, daß zum einen der Zugang zu den erforderlichen großen Datenmengen für keines der Register gewährleistet ist. Selbst wenn dem nicht so wäre, müßte sich der Datenangreifer in aller Regel zunächst eine eigene Datei aus den verfügbaren Daten aufbauen, da sich das Material, das als potentielles Zusatzwissen zur Verfügung steht, häufig sowohl in seiner physikalischen Form (elektronische Speicherung, Kartei, Liste) als auch in der Definition der möglichen statistischen Einheiten/Merkmale von den zu deanonymisierenden Mikrodaten unterscheiden wird. Daß diese Prozedur einen enormen Zeitaufwand erfordert und damit hohe Kosten verursacht, dürfte offensichtlich sein. Ein entsprechender Versuch wird in Kapitel 9 in allen diesen Implikationen beschrieben.

Zum anderen gab es für keines der Register offensichtliche Anhaltspunkte dafür, daß es die Anforderungen, die hinsichtlich der Überschneidungs-

---

14 Eine Übersicht über die genaue Praxis aller Institute kann hier nicht gegeben werden.

merkmale an einen Identifikationsfile gestellt werden müssen, erfüllen würde. Für eine sichere Datenzuordnung reichte weder die Anzahl der vorhandenen Überschneidungsmerkmale aus, noch waren die meisten Überschneidungsmerkmale hinreichend differenziert, um von Nutzen sein zu können. Hinzu kommt, daß die Aktualität der Angaben in den öffentlichen Registern - abhängig vom Zeitpunkt ihrer Erhebung - oft als veraltet eingestuft werden müssen. Auch die Kombination verschiedener Register zum Zweck der Kumulation von Informationen erwies sich als zu aufwendig und zu wenig nutzbringend, als daß diese Möglichkeit der Datengewinnung weiter berücksichtigt werden müßte .

Am ehesten erwies sich noch für das Einwohnermelderegister sowie für Informationsbücher über besonders herausgehobene Personengruppen (z.B. Gelehrtenkalender, Who is Who), daß in ihnen teilweise Informationen verfügbar sind, die für umfangreichere Deanonymisierungsversuche verwendet werden könnten. Für das Melderegister ergab sich die Vermutung, daß mit Hilfe der im Register enthaltenen Informationen zu Geschlecht, Familienstand, Geburtsdatum und Staatsangehörigkeit, Mitglieder einer bestimmten Personengruppe, nämlich ausländische Staatsangehörige, soweit eindeutig bestimmt werden könnten, daß eine Reidentifikation mit Hilfe dieser Informationen möglich erscheint. Auch bei den näher betrachteten beiden privaten, jedermann zugänglichen Datenquellen, Who is Who in Germany und Kürschners Gelehrtenkalender könnte unter Umständen eine kleine Bevölkerungsgruppe aufgrund der Angaben zu Beruf und Geburtsdatum vergleichsweise eindeutig abgegrenzt werden und somit einem erhöhten Deanonymisierungsrisiko ausgesetzt sein.

Was die Datenlage in den eigenen sozialwissenschaftlichen Erhebungen angeht, wurde festgestellt, daß die zur Verfügung stehende Stichprobengröße in einer forschungsüblichen Befragung allenfalls ausreicht, um vereinzelt zufällige Zuordnungen vornehmen zu können, eine Massenidentifikation bzw. ein Massenfischzug jedoch ausgeschlossen ist.

Auch bei den größeren Befragungen, die von Umfrageinstituten durchgeführt werden, kann man nicht von einer hinreichenden Stichprobengröße sprechen. Außerdem hat der Forscher in diesen Fällen keinen Zugang zu den Identifikationsmerkmalen. Insgesamt gesehen schafft das typische Zusatzwissen der Sozialwissenschaften demnach keine besonderen Risiken für eine Deanonymisierung von Mikrodaten.

Näher zu prüfen wäre hier jedoch, ob oder inwiefern sich aus der großen Zahl der in einer sozialwissenschaftlichen Umfrage vorhandenen Überschneidungs-

merkmale eine potentiell erfolgversprechende Ausgangslage für einen Deanonymisierungsversuch ergibt.

In den in Kapitel 8 entwickelten Szenarien und in späteren empirischen Versuchen werden diese hier nicht mit Sicherheit abzuklärenden, noch am ehesten als riskant erscheinenden Fälle, wieder aufgegriffen werden.

Anhang zu Kapitel 6

A. Öffentliche Register[1]

A.I Solche öffentlichen Bücher und Register, die jedermann ohne weitere Voraussetzungen zugänglich sind

(1) Jedermann kann persönlich Einsicht nehmen, sowie beglaubigte und unbeglaubigte Abschriften und Negativtest verlangen, d.h. eine Bescheinigung darüber, daß überhaupt keine oder keine weiteren als bestimmt angegebenen Eintragungen vorhanden sind.

(2) Bücher und Register mit der gleichen Öffentlichkeit, jedoch unter Ausschluß des Negativtests.

(3) Es besteht nur ein Recht auf Einsicht und auf Erteilung von Auskunft über das Bestehen oder Nichtbestehen von Eintragungen.

(4) Öffentliche Bücher, die das Gesetz lediglich als öffentlich, d.h. zugänglich bezeichnet, ohne den Umfang der Zugänglichkeit näher anzugeben

---

[1] Die Tabellen wurden anhand folgender Literatur zusammengestellt: Brand/Marowski (1956), Burkert (1979, 1980), Karl (1976), Leue (1986), Werhahn (1911).

TABELLE A.I: REGISTER, DIE JEDERMANN, OHNE WEITERE VORAUSSETZUNGEN ZUGÄNGLICH SIND

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
I.1 Vereinsregister	Name Anschrift (der Vorstands-mitglieder)	Name, Sitz des Vereins Datum der Errichtung der Satzung Beruf der Vorstandsmitglieder Änderungen d. Vorstands u. der Satzung Auflösung des Vereins Entziehung der Rechtsfähigkeit Konkurs Liquidation Verbot des Vereins	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf	Geschlecht Land der BRD Regierungsbezirk Gemeindegrößenklasse Bundesraumordnungsregion Beruf Stellung im Beruf	§ 79 S.1 BGB	jedermann, § 79 BGB, § 162 FGG öffentliche Bekanntgabe
I.2 Handelsregister A	Name Anschrift (der Einzelkaufmänner, persönl. haftenden Gesellschafter, Prokuristen, Kommandi-tisten, Vorstandsmit-glieder, Geschäfts-führer)	Firma Ort der Handelsniederlassung Sitz der Gesellschaft Beruf der unter (a) genannten Personen Prokura sonstige Rechtsverhältnisse Betrag der Einlage Tag der Eintragung	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Branche Beruf Stellung im Beruf	§§ 29, 106, 162 Abs. 1 HGB	jedermann, § 6 HGB öffentliche Bekanntgabe
Handelsregister B	Name Anschrift (der Vorstandsmitglie-der, persönl. haften-den Gesellschafter, Geschäftsführer, Abwickler, Prokuristen)	Firma Ort der Niederlassung Sitz der Gesellschaft Gegenstand des Unternehmens Grund- oder Stammkapital Beruf der unter (a) genannten Personen sonstige Rechtsverhältnisse Tag der Eintragung	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Beruf Stellung im Beruf	§§ 125-158 FGG Handelsregister-verfügung	

Fortsetzung Tabelle A.I:

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
I.3 Genossenschafts-register	Name Anschrift (der Genossen, des Vorstands, der Liquidatoren)	Genossen: Tag der Eintragung Beruf weitere Geschäftsanteile Grund, Tag des Ausscheidens Register: Firma Sitz der Genossenschaft Gegenstand des Unternehmens Höhe der Haftsumme höchste Zahl der Geschäftsanteil Beruf der Vorstandsmitglieder, Liquidatoren Datum des Status Form der Bekanntmachung Zeitdauer Geschäftsjahr Form für die Willenserklärung des Vorstands und der Liquidatoren Vertretungsbefugnis weitere Geschäftsanteile Auflösung, Konkurs, Fortsetzung, Nichtigkeit, Geschäftsnummer Tag der Eintragung	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Branche Beruf Stellung im Beruf	Verordnung §§ 125-158 FGG	jedermann, § 9 HGB, § 156 GenG, § 26 der Verordnung öffentliche Bekanntgabe
I.4 Geschmacks-muster-register	Name Anschrift (des Anmeldenden)	Firma des Anmeldenden Tag und Stunde der Anmeldung Bezeichnung des angemeldeten Musters oder Modells Angabe, ob das Muster für Flächenerzeugnisse oder für plastische Erzeugnisse bestimmt ist Schutzfrist Verlängerung der Schutzfrist Akten über das Musterregister	Geschlecht Land Gemeindegrößenklasse	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Branche	§ 13 Geschmacks mustergesetz	jedermann, § 11 GeschmMG öffentliche Bekanntgabe

Fortsetzung Tabelle A.I:

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikationsmerkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
I.5 Gebrauchsmusterrolle	Name Anschrift (des Anmeldenden)	Zeitpunkt der Anmeldung	Geschlecht Land Gemeindegrößenklasse	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion	Gebrauchsmustergesetz	jedermann, § 3 Abs. 5 GebrMustG
I.6 Urheberrolle	Name (des Urheberrechtsinhabers)	Geburtsdatum Geburtsort ggfs. Sterbejahr Titel, unter dem das Werk veröffentlicht ist Zeitpunkt u. Form der ersten Veröffentlichung des Werkes	Geschlecht Geburtsjahr	Geschlecht Geburtsmonat Geburtsjahr	Verordnung	jedermann, § 138 der Verordnung
I.7 Sortenschutzrolle	Name Anschrift (des Ursprungszüchters, Sortenschutzinhabers, Inhabers eines ausschließlichen Nutzungsrechts	Art u. Sortenbezeichnung Zeitpunkt des Beginns und der Beendigung des Sortenschutzes, Beendigungsgrund Zeitpunkt des Beginns u. der Beendigung eines ausschließlichen Nutzungsrechts Jedermannserlaubnis Zwangserlaubnis	Geschlecht Land Gemeindegrößenklasse	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion	Sortenschutzgesetz	jedermann, §§ 30 u. 31 des Sortenschutzgesetzes
I.8 Patentrolle	Name Anschrift (des Patentinhabers)	Gegenstand und Dauer des erteilten Patents Anfang, Ablauf, Erlöschung, Anordnung der Beschränkung, Erklärung der Nichtigkeit u. Zurücknahme der Patente	Geschlecht Land Gemeindegrößenklasse	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion	Patentgesetz	Einsicht in Rolle u. Akten: jedermann, § 24 Abs. 3 PatG Einsicht in Modelle u. Produkte: soweit berechtigtes Interesse glaubhaft gemacht wird §§ 1, 24 Patentgesetz öffentliche Bekanntgabe

Fortsetzung Tabelle A.I:

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikationsmerkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
I.9 Warenzeichenrolle	Name Anschrift (des Zeicheninhabers)	Zeitpunkt der Anmeldung Bezeichnung des Geschäftsbetriebs Verzeichnis der Waren, für die das Zeichen bestimmt ist Verlängerung der Schutzdauer Zeitpunkt der Löschung des Zeichens	Geschlecht Land Gemeindegrößenklasse	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Branche	Warenzeichengesetz	Einsicht in die Zeichenrolle: § 3 des Gesetzes Einsicht in die Akten: nur wenn berechtigtes Interesse glaubhaft gemacht wird
I.10 Güterrechtsregister	Name Anschrift (der Ehegatten)	Beruf Bestehen, Änderung u. Aufhebung von Eheverträgen Änderungen im gesetzl. Güterstand Beschränkung u. Ausschluß der sogenannten Schlüsselgewalt	Geschlecht Land Gemeindegrößenklasse Familienstand Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Familienstand Beruf Stellung im Beruf	§ 1412 Abs. 1, BGB	jedermann, § 1563 BGB, § 162 FGG öffentliche Bekanntgabe
I.11 Schuldnerverzeichnis	Name Anschrift (des Schuldners)	eidesstattliche Versicherung Erzwingungshaft	Geschlecht Land Gemeindegrößenklasse	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion	§ 915 Abs. 1, ZPO	jedermann, § 915 Abs. 3 ZPO
I.12 Liste der Rechtsanwälte	Name Anschrift	Beruf	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Beruf Stellung im Beruf	§§ 31, 32 BRAO	jedermann

Fortsetzung Tabelle A.I:

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
I.13 Liste der Patent anwälte	dto.	dto.	dto.	dto.	§§ 29, 30 der Patentanwaltsord.	jedermann
I.14 Berufsregister der Steuerberater, Steuerberatungs-gesellschaften, Steuerbevollmäch-tigten	dto.	dto.	dto.	dto.	Berufsordnung	jedermann, § 37 Abs. 3 der Verordnung
I.15 Verzeichnis über Lohnsteuerhilfe-vereine					§ 30 des Gesetzes	jedermann
I.16 Berufsregister der Wirtschafts-prüfer, WPgesell-schaften, ver-eidigte Buchprüfer	dto.	dto.	dto.	dto.	WPordung	§§ 37, 130 WPordung

Fortsetzung Tabelle A.I:

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
I.17 Liste der öffentlich bestellten Vermessungsingenieure	dto.	dto.	dto.	dto.	Berufsordnung	jedermann
I.18 Architektenliste	dto.	dto.	dto.	dto.	Berufsordnung	jedermann
I.19 Liste der beratenden Ingenieure	dto.	dto.	dto.	dto.	landesrechtliche Spezialregelungen	öffentl. Bekanntgabe

A.II Öffentliche Bücher und Register, die nur einem gewissen, mehr oder weniger bestimmten Personenkreis zugänglich sind

(1) Diejenigen Personen, denen das Register zugänglich ist, sind im Gesetz erschöpfend aufgezählt.

(2) Die Zugänglichkeit besteht nur für solche Personen, die nach dem Ermessen der Registerbehörden ein rechtliches, d.h. ein durch ein bestehendes Rechtsverhältnis bedingtes Interesse an der Einsicht haben.

(3) Endlich gibt es öffentliche Bücher und Register, deren Zugänglichkeit nur für den Fall besteht, daß ein berechtigtes, sondern auch wirtschaftliches Interesse dargetan wird.

TABELLE A.II: REGISTER, DIE NUR EINEM GEWISSEN, MEHR ODER WENIGER BESTIMMTEN PERSONENKREIS ZUGÄNGLICH SIND

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
II.1 Geburtenbuch	Name Anschrift (der Eltern, der Anzeigenden)	Beruf der unter (a) genannten Personen Religionszugehörigkeit Ort, Tag, Stunde der Geburt Geschlecht u. Vorname des Kindes Bezug zum Familienbuch Randvermerke über den Status des Kindes (Namensänderungen etc.)	Geschlecht Land Gemeindegrößenklasse Geburtsjahr Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Geburtsmonat Geburtsjahr Beruf Stellung im Beruf	Personenstands-gesetz; Verordnung zur Ausführung des PersstG, Dienstanweisung für Standes-beamte	eingetragene Personen: grundsätzlich; Ehegatten, Abkömm-linge; bei besonderem Inter-esse; sonst: rechtliches § 16 Abs. 2 u. 3 PStG
II.2 Heiratsbuch	Name Anschrift (der Eheschließenden, der Zeugen)	Beruf der unter (a) genannten Personen Religionszugehörigkeit Geburtsort Geburtsdatum Alter der Zeugen Randvermerke über Veränderungen des ehelichen Status	Geschlecht Land Gemeindegrößenklasse Geburtsjahr Familienstand Ehedauer Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Geburtsmonat Geburtsjahr Familienstand Eheschließungsjahr Beruf Stellung im Beruf	Personenstands-gesetz; Verordnung zur Ausführung des PersstG, Dienstanweisung für Standes-beamte	eingetragene Personen: grundsätzlich; Ehegatten, Abkömm-linge; bei besonderem Inter-esse; sonst: rechtliches Interesse
II.3 Familienbuch	Name Anschrift (der Eheschließenden, der Eltern)	gemeinsam ehelich geborene, durch Eheschließung legitimierte gemeinschaftlich oder von einem Ehegatten adoptierte Kinder Aufhebung oder Scheidung der Ehe Nichtigerklärung der Ehe Wiederverheiratung sonstige Änderungen des Personenstands	Geschlecht Land Gemeindegrößenklasse Familienstand Ehedauer Kinder des HV	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Familienstand Eheschließungsjahr Kinder des HV	Personenstands-gesetz Verordnung zur Ausführung des PersstG, Dienstanweisung für Standes-beamte	eingetragene Personen: grundsätzlich; Ehegatten, Abkömm-linge: bei besonderem Inter-esse; sonst: rechtliches Interesse

Fortsetzung Tabelle A.II:

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
II.4 Sterbebuch	Name Anschrift (des Verstorbenen, des Ehegatten, d.s Anzeigenden)	Ort und Tag der Geburt Beruf des Verstorbenen und des Anzeigenden Religionszugehörigkeit Ort, Tag, Stunde des Todes	Geschlecht Land Gemeindegrößenklasse Geburtsjahr Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Geburtsmonat Geburtsjahr Beruf Stellung im Beruf	Personenstands-gesetz Verordnung zur Ausführung des PersstG, für Standes-beamte	eingetragene Personen: grundsätzlich; Ehegatten, Abkömm-linge; bei besonderem Inter-esse; sonst: rechtliches Interesse
II.5 Melderegister	Name Anschrift (aller Personen, die eine Wohnung beziehen oder aus einer Wohnung ausziehen)	Geschlecht Geburtsdatum Familienstand Religionszugehörigkeit Staatsangehörigkeit Zweitwohnung	Geschlecht Land Gemeindegrößenklasse Geburtsjahr Familienstand	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Geburtsmonat Geburtsjahr Familienstand Staatsangehörigkeit Zweitwohnung	Landesmelde-gesetze	berechtigtes Interesse
II.6 Kraftfahrzeug-kartei	Name Anschrift (des Halters)	Art u. Hersteller des Fahrzeugs	Geschlecht Land Gemeindegrößenklasse PKW-Besitz Anschaffungsjahr Anschaffungskosten	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion	Straßenverkehrs-zulassungs-ordnung	im Einzelfall auf Antrag an Behörden, bei Darlegung eines berechtigten Inter-esses auch an andere §29f. Abs.2 StVZO

Fortsetzung Tabelle A.II:

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikationsmerkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
II.7 Grundbuch - Bestandsverzeichnis -Abteilungen I-III (auch: Erbbaugrundbuch, Wohnungsgrundbuch)	Name Anschrift (des Eigentümers)	Beruf Geburtsdatum Eigentumsanteil Darlehensbelastung Vormerkungen Widersprüche	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf Geburtsjahr Eigentümer von Grundstücken, Gebäuden u. Eigentumswohnungen Tilgung von Hypotheken, Krediten usw.	Geschlecht Land Gemeindegrößenklasse Beruf Regierungsbezirk Bundesraumordnungsregion Geburtsmonat Geburtsjahr	§§ 873, 883, 899 u.a. BGB	berechtigtes Interesse, § 12 GBO
II.8 Liegenschaftskataster	Name Anschrift (des Grundstückseigentümers)	Gemeinde Gemarkung Flur Flurstücksnummer Straße Veränderungsnummer Flurstücksfläche Abschnittsfläche Gebäudefläche Nutzungsart Gebäudeart Flurkarte	Geschlecht Land Gemeindegrößenklasse Eigentümer von Grundstücken, Gebäuden u. Eigentumswohnungen	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion	Landesrechtliche Vorschriften	berechtigtes Interesse, § 12 GBO
II.9 Baulastenverzeichnis		baurechtliche Verpflichtungen des Grundstückseigentümers			Bauordnung oder Länder	berechtigtes Interesse
II.10 Höferolle					entsprechende Verordnung der Länder	berechtigtes Interesse
II.11 Weinbergrolle					entsprechende Verordnung der Länder	berechtigtes Interesse

Fortsetzung Tabelle A.II:

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
II.12 Register für Pfandrechte an Luftfahrzeugen					Gesetz über Rechte an Luftfahrzeugen	berechtigtes Interesse
II.13 Arztregister	Name Anschrift	Angaben über das Staatsexamen, die Approbation Facharztanerkennung Zulassung als Kassenarzt	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Beruf Stellung im Beruf	§ 368 Abs. 1 RVO für Kassenärzte	eingeschränkt zulässig; §§ 9, 10 der Zulassungsordnung
II.14 Kassenarztregister Bundesarztreg. Zahnarztregister					Zulassungs-ordnung	eingeschränkt zulässig; §§ 9, 10 der Zulassungsordnung
II.15 Lehrlingsrolle		Lehrverhältnisse im Handwerk			Handwerks-ordnung	berechtigtes Interesse
II.16 Handwerksrolle		selbständige Handwerksmeister			Handwerks-ordnung	berechtigtes Interesse §6 Abs.3 Handwerksord.
II.17 Verzeichnis handwerksähn-licher Betriebe					Handwerks-ordnung	berechtigtes Interesse

Fortsetzung Tabelle A.II:

Registerart	Eintragungen im Register				Rechtsgrundlage	Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
II.18 Verzeichnis der Berufsaus-bildungsverträge		Registrierung der Lehrverhältnisse bei: Industrie- u. Handelskammern, Landwirtschaftskammern, Ärzte- u. Zahnärztekammern, Apotheker-kammern, Rechtsanwalts- u. Notarkammern, Steuerberater-kammern, Patentanwaltskammern, Wirtschaftsprüferkammern			§ 31 BerBildG	berechtigtes Interesse
II.19 Immatrikulation an Hochschulen	Name Anschrift (des Studierenden)	Geburtsdatum Studienfach	Geschlecht Land Gemeindegrößenklasse Geburtsjahr	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Geburtsjahr Geburtsmonat Schulbesuch		

A.III Bücher und Register, die nur bestimmten Behörden für bestimmte Zwecke zugänglich sind

TABELLE A.III: REGISTER, DIE NUR BESTIMMTEN BEHÖRDEN FÜR BESTIMMTE ZWECKE ZUGÄNGLICH SIND

Registerart	Eintragungen im Register				Rechtsgrundlage	Auskunftsberechtigung
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
III.1 Bundeszentral-register	Name Anschrift	Geburtsort Geburtsdatum Geschlecht Staatsangehörigkeit alle Eintragungen aus den Strafre-gistern der Bundesländer	Geschlecht Land Gemeindegrößenklasse Geburtsjahr	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Geburtsdatum Geburtsjahr Staatsangehörigkeit	§§ 3, 5 BZRG	§§ 30-40 BZRG Selbstauskunft (Führungszeugnis) §§ 41-43 BZRG unbeschränkte Auskünfte § 44 Auskünfte an Behörden
III.2 Erziehungs-register	Name Anschrift	Geburtsort Geburtsdatum Geschlecht Staatsanghörigkeit Maßnahmen nach § 3 Satz 2 JGG Verfahrenseinstellungen gemäß §§ 45, 47 JGG	Geschlecht Land Gemeindegrößenklasse Geburtsjahr	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Geburtsmonat Geburtsjahr Staatsangehörigkeit	§§ 3, 56 BZRG	Selbstauskunft (Führungszeug-nis); nur bestimmte Behörden für bestimmte Zwecke §§ 39, 42 BZRG

Fortsetzung Tabelle A.III:

Registerart	Eintragungen im Register				Rechtsgrundlage	Auskunftsberechtigung
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ		
III.3 Ausländer-zentralregister	Name Anschrift (ausländischer Personen)	Zahl der Haushaltsmitglieder Geburtsdatum Geschlecht Familienstand Staatsangehörigkeit Erwerbstätigkeit (ja oder nein)	Geschlecht Land Gemeindegrößenklasse Geburtsjahr Familienstand Haushaltsgröße Haushaltstyp Anzahl der Erwerbs-tätigen	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Geburtsmonat Geburtsjahr Familienstand Zahl der Personen im HH Haushaltstyp Staatsangehörigkeit Erwerbstätigkeit in der Berichtswoche Haushaltsgröße Anzahl der Erwerbs-tätigen	AuslG, allgemeine Vorschriften zur Ausführung des AuslG	Institutionen im Einzelfall bei Verdacht auf strafbare Handlungen
III.4 Verkehrs-zentralregister	Name Anschrift	Verkehrsstraftaten, Ordnungswidrig-keiten, Fahrverbote, Entziehung der Fahrerlaubnis			§ 28 StVG	§ 30 StVG nur zur Verfolgung von Verkehrs-straftaten oder Ordnungswidrigkeiten, für Verwaltungs-maßnahmen auf Grund dieses Gesetzes

B. Private Datenquellen[2]

B.I Solche Datenquellen, die jedermann ohne weitere Voraussetzungen zugänglich sind

[2] Die Tabellen wurden anhand folgender Literatur zusammengestellt: Adressbuch der Stadt Mannheim (1987), Kürschners Deutscher Gelehrtenkalender (1976), Taschenbuch des öffentlichen Lebens (1983/84), Taschenbuch des öffentlichen Lebens (1984/84), Teske (1986), Who is Who in Germany (1974).

TABELLE B.I: PRIVATE DATENQUELLEN, DIE JEDERMANN OHNE WEITERE VORAUSSETZUNGEN ZUGÄNGLICH SIND

Registerart	Eintragungen im Register				Einsichtsrechte
	(a) Identifikationsmerkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ	
I.1 Telefonbuch	Name Anschrift (des Teilnehmers)	Telefonnummer Beruf	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf Telefonbesitz	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Beruf Stellung im Beruf	jedermann
I.2 Branchenfernsprechbuch *)	Name Anschrift (des Teilnehmers)	Telefonnummer Beruf Branche	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf Telefonbesitz	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Beruf Stellung im Beruf Branche	jedermann
I.3 Taschenbuch des öffentlichen Lebens	Name Anschrift (von den unter (b) genannten Personen)	Informationen zu folgenden Bereichen: Bund, Länder, Regierungsbezirke, kommunale Verbände, Großstädte, Ausland, Wirtschaft, Sozialpartner u. berufl. Organisationen, Presse, Funk/Fernsehen, Film, Verlag/Buchhandel, Politische Parteien, Religionsgemeinschaften, Wissenschaft, Erziehungs- u. Bildungswesen, Kunst Umwelt- u. Freizeitgestaltung, Verbindungsstellen im Raum Bonn-Köln, Personen-, Sachregister			jedermann

*) Der Vorname ist nicht in allen Fällen angegeben, so daß das Geschlecht des Teilnehmers nicht immer identifiziert werden kann. Des weiteren fehlen oft Angaben zum Straßennamen, zur Hausnummer und zum Stadtteil sowie i.d.R. die Berufsangabe.

Fortsetzung Tabelle B.I:

Registerart	Eintragungen im Register				Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ	
I.4 Adressbücher	Name Anschrift (der Einwohner über 18 Jahre u. Firmen)	Beruf Branche	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Beruf Stellung im Beruf Branche	jedermann
I.5 Who is who in Germany	Name Anschrift (von Personen des öffentlichen Lebens deren Ehefrauen und Eltern)	Telefonnummer Beruf Geburtsdatum Jahr der Eheschließung Geschlecht, Anzahl der Kinder beruflicher Werdegang Mitgliedschaft Hobbys Auszeichnungen Sprachkenntnisse	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf Geburtsjahr Ehedauer Familientyp Telefonbesitz Kinder des HV	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Beruf Stellung im Beruf Geburtsmonat Geburtsjahr Eheschließungsjahr Kinder des HV Familientyp	jedermann
I.6 Kürschners Gelehrtenkalender	Name Anschrift (des Akademikers)	akademische Grade Geburtsdatum Geburtsort akademischer Werdegang jetzige berufliche Stellung wissenschaftl. Auszeichnungen Mitgliedschaften in wissenschaftl. Vereinigungen Publikationen Herausgeberschaften	Geschlecht Land Gemeindegrößenklasse Geburtsjahr Stellung im Beruf	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Geburtsmonat Geburtsjahr Beruf	jedermann

B.II Solche Datenquellen, die nur einem bestimmten Personenkreis zugänglich sind[3]

3 An dieser Stelle sollen nur exemplarisch zwei Dateien aus der Kreditwirtschaft beschrieben werden, die stellvertretend für eine Vielzahl der in der privaten Verwaltung geführten Dateien stehen (z.B. Personaldateien, Krankenversichertendateien).

TABELLE B.II: PRIVATE DATENQUELLEN, DIE NUR EINEM GEWISSEN, MEHR ODER WENIGER BESTIMMTEN BESTIMMTEN PERSONENKREIS ZUGÄNGLICH SIND

Registerart	Eintragungen im Register				Einsichtsrechte
	(a) Identifikations-merkmale	(b) Sonstige Merkmale	(c) Überschneidungsmerkmale mit EVS	MZ	
II.1 Schufa-Datei	Name Anschrift (des Konsumenten)	Angaben zum Giro-Konto Bürgschaften Pfändungen Kredite Mahnbescheide etc.	Geschlecht Land Gemeindegrößenklasse Stellung im Beruf Besitz von Sparbüchern, Wertpapieren etc.	Geschlecht Land Gemeindegrößenklasse Regierungsbezirk Bundesraumordnungsregion Beruf	ca. 30.000 Unternehmen (500.000 Personen): Banken u. Sparkassen, große Versandhandelsunternehmen, Waren- u. Kaufhäuser, Einzelhandelsgeschäfte.
II.2 Schimmelpfeng-Datei	Name Anschrift (der Firma)	Rechtsform des Unternehmens Eintragung in das Handelsregister Gesellschafter Gründung u. Entwicklung des Betriebes Branche u. Produkte Umsatz, Grundeigentum Betriebseinrichtung Warenlager Außenstände Verpflichungen Bonitätsbeurteilung Bankverbindung Kreditfrage			die angeschlossenen Unternehmen

7 Allgemeine Kosten-Nutzen-Überlegungen zu Deanonymisierungsversuchen*

Das Bundesstatistikgesetz definiert die faktische Anonymität anhand der Unverhältnismäßigkeit des Aufwandes an Zeit, Kosten und Arbeitskraft, der erforderlich ist für eine Zuordnung der Einzelangaben zu den Personen, von denen sie stammen. Unverhältnismäßigkeit bedeutet ein Abwägen zwischen zwei Gütern. Im Gesetzestext selbst hat der Gesetzgeber mit dem Hinweis auf den Aufwand nur eine Seite bestimmt: Im Vergleich zu welchem Gut der Aufwand abzuwägen ist, ist nicht ausgeführt. Am naheliegendsten erscheint jedoch, daß er ein Rationalkalkül unterstellt zwischen dem für eine Deanonymisierung zu betreibenden Aufwand und dem Nutzen, den ein Datenangreifer aus einer erfolgreichen Zuordnung der Einzelangaben ziehen kann: Wenn nur vom unverhältnismäßigen Aufwand, den ein Angreifer aufbringen muß, die Rede ist, dann kann als am nächsten liegender Bezug die Überlegung des Angreifers unterstellt werden, ob sich der Aufwand für ihn lohnt.

Alternative Unverhältnismäßigkeitsabwägungen sind jedoch denkbar. Beispielsweise könnte der Aufwand abgewogen werden gegen den Wert des Gutes der Unverletzlichkeit der Privatsphäre der Person, von der die Daten stammen, oder gegen den Wert dieses Prinzips überhaupt. Gegenüber der oben genannten einfachen Nutzenabwägung des Angreifers erscheinen solche Überlegungen jedoch konstruierter und weniger plausibel. Deshalb gehen wir hier von einem Unverhältnismäßigkeitsverständnis im Sinne eines individuellen Kosten-Nutzen-Vergleichs eines potentiellen Datenangreifers aus. Unter der Annahme eines ökonomisch rationalen Verhaltens wird er sich für einen Deanonymisierungsversuch nur entscheiden, falls dessen erwartete Kosten den erwarteten Nutzen nicht übersteigen.

In diesem Kapitel sollen Nutzen- und Kostenkomponenten, die bei einem Deanonymisierungsversuch als relevant erscheinen, allgemein beschrieben werden. Sie können dann später für die einzelnen Deanonymisierungsszenarien als Bewertungsgerüst herangezogen werden. Zuvor sind hier jedoch einige Einschränkungen nötig.

Der Nutzen, den ein Individuum für sich selbst seiner Entscheidungsalternative - hier der Deanaonymisierung - zumißt, kann von Dritten ohne entsprechende

---

* Autor: Thomas Helmcke

Nachfragen oder Untersuchungen bei dieser Person im allgemeinen nicht bestimmt, sondern höchstens mit Plausibilitätsannahmen abgeschätzt werden. Daher sind insbesondere über den hier zu diskutierenden individuellen Nutzen von Deanonymisierungsversuchen keine allgemeingültigen Aussagen möglich.

Die Kosten hängen unter anderem vom jeweiligen Motiv für die Reidentifikation ab, so daß auch für sie kein eindeutiges Ergebnis ableitbar ist. Denkbar sind auch Kostenüberwälzungen vom Angreifer auf andere Stellen. Setzt z.B. ein Wissenschaftler zur Deanonymisierung Rechenanlagen oder gar Personal seines Instituts ein, so können ihm die dadurch verursachten Kosten nicht persönlich zugerechnet werden und somit auch nicht in die Kosten-Nutzen-Analyse eingehen. Der Erfolg solcher Versuche hängt sicher von den technisch-organisatorischen Sicherungsmaßnahmen in der mit dem amtlichen Mikrodatenfile belieferten Forschungseinrichtung ab. Kostenüberwälzungen erhöhen allerdings auch die Wahrscheinlichkeit der Aufdeckung von Deanonymisierungsabsichten.

Im übrigen ist zu beachten, daß selbst dem potentiellen Angreifer die einzelnen Kosten-Nutzen-Komponenten vor einem Deanonymisierungsversuch im allgemeinen nicht genau bekannt sind, sondern daß es sich für ihn um Erwartungsgrößen handelt. Er muß z.B. allein schon deshalb von einem erwarteten statt sicheren Nutzen ausgehen, weil der Erfog seiner Versuche nicht garantiert ist.

7.1 Nutzenkomponenten eines Deanonymisierungsversuchs

Die Höhe des Nutzens einer Deanonymisierung hängt unter anderem vom Wert der durch die Reidentifikation gewonnenen zusätzlichen Informationen im Hinblick auf die Verwertungsinteressen des Angreifers ab, die wiederum durch dessen Motive bestimmt werden. Diese Motive lassen sich in berufliche, d.h. wissenschaftliche, und außerberufliche unterteilen. Entsprechend gibt es Nutzenkomponenten, die sich ein Angreifer aus professionell motivierten Deanonymisierungen versprechen könnte und solche, die auf außerberuflichen Interessen beruhen. Wie schon die Analyse der Deanonymisierungsmotive in Kapitel 5 gezeigt hat, dürften wissenschaftliche Nutzenkomponenten von geringerer Bedeutung sein. Zunächst soll aber mit einer eventuell möglichen Ausgabeneinsparung eine Nutzenkomponente erörtert werden, die sogar unabhängig von Verwertungsinteressen zu berücksichtigen ist.

7.1.1 Eingesparte Ausgaben

Einsparungen sind sinnvollerweise auf einen Vergleich zweier Handlungsalternativen A und B zu beziehen. Entscheidet sich ein potentieller Angreifer z.B. für A, etwa eine Deanonymisierung, so vermeidet er dadurch die unmittelbaren Kosten K_B von B. Die so eingesparten Ausgaben stellen also eine Nutzenkomponente von A dar, die ihrem sonstigen (= unmittelbaren, direkten) Nutzen N_A hinzuzufügen ist. Andererseits resultieren aus dem entgangenen direkten Nutzen N_B von B Opportunitätskosten für A (Diller 1988:261) , die zu den sonstigen Kosten K_A von A addiert werden müssen. Der Gewinn für den potentiellen Angreifer durch seine Entscheidung für A ergibt sich damit aus dem gesamten Nutzen $N_A + K_B$ von A abzüglich der gesamten Kosten $K_A + N_B$ von A zu $N_A + K_B - K_A - N_B$. Bei unterstelltem rationalem Verhalten wird der potentielle Angreifer die Alternative A der Alternative B vorziehen, falls dieser Gewinn positiv ist.

Vergleicht man nun speziell zwei Alternativen, die dem potentiellen Angreifer beide den gleichen direkten Nutzen $N_A = N_B$ bringen, so erhält man als Gewinn bei einer Entscheidung für A einfach $K_B - K_A$, d.h. die Differenz der unmittelbaren Kosten beider Alternativen. Der potentielle Angreifer wählt dann also Alternative A, wenn ihre unmittelbaren Kosten unter denen von B liegen. Der direkte Nutzen spielt damit für die Entscheidung keine Rolle mehr. Eine solche Situation liegt zum Beispiel vor, wenn sich ein potentieller Angreifer zwischen einer Deanonymisierung und einer anderen Methode zur Beschaffung derselben Daten zu entscheiden hat. Der direkte Nutzen besteht dann aus den Verwertungsmöglichkeiten der gewonnenen Daten und ist natürlich für beide Alternativen gleich. Es verbleiben also die unmittelbaren Kosten als alleiniges Entscheidungskriterium.

Solche Vergleiche verschiedener Datenbeschaffungsmethoden können die Beurteilungen von Deanonymisierungen vereinfachen und spielen daher in diesem Zusammenhang eine große Rolle. Liegen nämlich die unmittelbaren Kosten einer Deanonymisierung deutlich über denen einer alternativen Beschaffungsmethode für dieselben Daten oder für Daten gleicher Validität, die denselben Zweck erfüllen, so bedeutet dies, wie erwähnt, einen entsprechend deutlichen Verlust (= negativen Gewinn) in Höhe der Kostendifferenz. Dann stellen die unmittelbaren Deanonymisierungskosten also einen unverhältnismäßig hohen Aufwand für den potentiellen Angreifer dar, so daß er sich für die andere Beschaffungsmethode entscheiden wird. Eine Nutzenanalyse ist in diesem Falle nicht mehr nötig. Dies bedeutet eine gewisse Vereinfachung, da die Abschätzung der individuellen Nutzenkomponenten - insbesondere der nicht-

monetären - im Vergleich zur Diskussion der Kostenkomponenten schwieriger und problematischer ist. Verschiedene Datenbeschaffungsmethoden werden später noch kurz angesprochen.

Die Bedeutung derartiger Kostenvergleiche bringt unter anderem der Bundesbeauftragte für den Datenschutz (1981:50) in seinem Vierten Tätigkeitsbericht zum Ausdruck: "Maßgeblich für die Beurteilung sind deshalb die Sensitivität der Daten, das denkbare Interesse Dritter an den personenbezogenen Daten und der Aufwand, der betrieben werden müßte, um den Personenbezug wieder herzustellen, im Vergleich mit dem notwendigen Aufwand, um dieselbe Information auf andere Weise zu erlangen."

Zusammenfassend ist damit festzuhalten, daß vor einer Gegenüberstellung von Kosten und Nutzen einer Deanonymisierung im Rahmen einzelner Szenarien zu prüfen ist, ob das von einem potentiellen Angreifer angestrebte Ziel statt durch Reidentifikation mit einem deutlich kostengünstigeren Mittel erreicht werden kann. Ist dies der Fall, so erübrigt sich eine Einschätzung des direkten Deanonymisierungsnutzens. Im anderen Fall ist dieser Nutzen dagegen aus ökonomischer Sicht in die Abwägung einzubeziehen.

7.1.2 Nutzenkomponenten bei beruflicher Motivation

In Kapitel 5 zur Handlungslogik von Sozialwissenschaftlern wurde bereits dargelegt, daß kaum eine andere berufliche Motivation für Deanonymisierungen vorstellbar sei als die Gewinnung einer geeigneten Auswahlgrundlage für eigene Erhebungen oder die Ergänzung vorhandener Daten um zusätzliche Informationen. Der Nutzen, der aus einer entsprechenden Deanonymisierung erwachsen könnte, ist nun schwer abzuschätzen. Er kann dem Angreifer aus den wissenschaftlichen Erkenntnissen an sich erwachsen, die als Folge einer Deanonymisierung resultieren. Von ihrer Veröffentlichung könnte sich ein Angreifer weiterhin berufliche Anerkennung und möglicherweise Erlöse in Form von Honoraren für Zeitschriftenaufsätze, Vorträge oder Verlagstantiemen versprechen. Solche Nutzenelemente sind aber für die Sozialwissenschaft wohl als eher unbedeutend einzuschätzen. Für die immateriellen Nutzenelemente, die aus wissenschaftlichen Veröffentlichungen gewonnen werden, gilt zudem, daß sie ja nur zu einem Teil aus den zugrundeliegenden empirischen Daten selbst resultieren, sondern vielmehr aus deren analytischer Verarbeitung und der damit verbundenen intellektuellen Arbeit. Auch hatten die in Kapitel 5 angestellten Erörterungen gezeigt, daß selbst bei den denkbaren beruflichen

Deanonymisierungsmotiven die konkrete Nutzbarkeit deanonymisierter Mikrodaten sehr begrenzt ist. Obwohl berufliche Nutzenkomponenten nicht völlig außer acht gelassen werden sollten, ist zusammenfassend festzuhalten, daß sie wahrscheinlich nur ein geringes Gewicht haben und im Prinzip vernachlässigt werden können.

7.1.3 Nutzenkomponenten bei außerberuflicher Motivation

Außerberufliche Motive für Deanonymisierungen sind auf der anderen Seite keine zentralen Bestandteile von typischen Wissenschaftsszenarien. Dennoch sollen einige entsprechende Nutzenkomponenten erörtert werden, um auch eine solche potentielle Gefährdung nicht unberücksichtigt zu lassen.

7.1.3.1 Publikationserlöse

Eine Möglichkeit, durch Deanonymisierung gewonnene Daten ökonomisch zu verwerten, besteht in Erlösen aus deren Publikation. Solche Erlöse können ein Hauptmotiv für eine Deanonymisierung darstellen oder aber vom Angreifer nur als Nebenprodukt angesehen werden.

In der Regel wird der Angreifer die Publikation wohl nicht selbst vornehmen, sondern versuchen, die Information - z.B. an Massenmedien - zu verkaufen. Dabei dürfte seine Position in Preisverhandlungen für eine nicht anonyme Publikation stärker sein, als wenn er die Wahrung seiner Anonymität zur Bedingung macht. Die Medien könnten im Hinblick auf die Glaubwürdigkeit eher an einer Publikation unter Angabe der Quelle interessiert sein, aber auch um eine "Enthüllungsgeschichte" für das Publikum attraktiver zu gestalten.

Publikationserlöse kommen als Hauptmotiv für einen Angreifer nur in Frage, wenn sie relativ hoch ausfallen. Dies ist wohl am ehesten der Fall, wenn es sich um Daten prominenter Personen handelt. Die Erlöse dürften mit deren Bekanntheitsgrad, der Datenmenge, -aktualität und -sensitivität zunehmen. Die Sensivität der Daten ist jedoch zumindest für den Mikrozensus als relativ begrenzt einzuschätzen.

Besteht dagegen das Hauptmotiv des Angreifers nicht in Publikationserlösen und ist er gleichwohl stark an einer Publikation interessiert, so dürfte er in

Preisverhandlungen eher nachgiebig auftreten. Eine solche Situation läge z.B. vor, wenn die Schädigung bekannter Personen oder etwa der statistischen Ämter das vordringliche Ziel der Publikationen darstellt. Wegen des geringen Stellenwerts von Erlösen wird der Angreifer dann sogar vermutlich eine anonyme Publikation vorziehen, um sein Renommee nicht zu gefährden, und sonstige Sanktionen zu vermeiden.

Eine weitere Situation, in der der Angreifer der Publikation an sich mehr Bedeutung beimißt als den Erlösen daraus, liegt vor, wenn sein Hauptmotiv in der öffentlichen Anerkennung seiner Deanonymisierungsleistung besteht, falls er sich also einen Namen als "Hacker" oder "Enthüller" machen will. Von daher ist wie im zuvor geschilderten Fall eine eher nachgiebige Haltung in Verkaufsverhandlungen zu erwarten. Im Gegensatz zur vorigen Situation gibt es hier jedoch auch einen gegenläufigen Effekt: Zur Erreichung der angestrebten öffentlichen Anerkennung, muß die Publikation zwangsläufig in nichtanonymer Form erfolgen. Dies führt vermutlich wiederum zu Sanktionen gegen den Angreifer. Zur Schadensbegrenzung könnte er daher verstärkt versuchen, einen möglichst hohen Publikationserlös zu erzielen.

Allgemein resultiert jedoch ein erlösbegrenzender Effekt aus der Tatsache, daß etwa Massenmedien Informationsbeschaffung ohnehin professionell betreiben und ihnen dazu neben dem Ankauf deanonymisierter Daten alternative Möglichkeiten zur Verfügung stehen.

7.1.3.2 Weitere monetäre Verwertungsmöglichkeiten

Speziell in Verbindung mit Informationen über Einkommen, Vermögen und Schulden lassen sich weitere monetäre Nutzungsmöglichkeiten konstruieren. So liefert insbesondere die Einkommens- und Verbrauchsstichprobe ein ziemlich differenziertes Bild über die Einkommens- und Vermögenslage der Auskunftgebenden. Der Verkauf deanonymisierter Daten an Adressenhändler, etwa für Marketingwecke, stellt daher eine naheliegende Verwertungsmöglichkeit dar. Vorstellbar ist in diesem Zusammenhang auch die Ausnutzung einer wirtschaftlich schlechten Lage interessierender Personen. So könnte ein Angreifer, der Interesse am Erwerb eines bestimmten Grundstücks hat, versuchen, Informationen über eine Notlage des Anbieters in den Kaufverhandlungen auszunutzen, um sich monetäre Vorteile zu verschaffen.

Unwahrscheinlicher, aber gleichwohl denkbar, erscheinen Motive, die im Bereich rechtswidrigen Verhaltens angesiedelt sind. So könnten Deanonymisierungen Hinweise darauf geben, ob und wo sich Einbrüche lohnen. Aus einer Reidentifikation gewonnene Informationen mögen im Extremfall sogar für Erpressungsversuche genutzt werden: Ein Angreifer könnte prominenten Personen etwa mit der Publikation deanonymisierter Daten über ihre Vermögens- und Schuldenlage drohen. Sind die zugeordneten Personen dagegen in der Öffentlichkeit nicht bekannt, kommt eine Publikation der gewonnenen Daten wohl kaum in Betracht. Der Angreifer kann in diesem Fall jedoch dem Betroffenen unter Umständen mit einer anderweitigen Schädigung drohen. Vorstellbar ist z.B. die Androhung einer Weitergabe von Daten an bestimmte interessierte Kreise, etwa an Mitteilungen über die Vermögensverhältnisse der betroffenen Personen an deren Geschäftspartner, Gläubiger oder Konkurrenten.

Die genannten Motive erfordern im allgemeinen keine Preisgabe der Informationsquelle, also nicht die Offenlegung der Deanonymisierung. Dagegen spielt der Aktualitätsgrad der Daten eine wichtige Rolle. Mit zunehmendem Alter der Informationen wächst die Unsicherheit des Angreifers darüber, ob sie weiterhin gültig sind. Entsprechend niedriger dürfte er den zu erwartenden Nutzen einschätzen.

7.1.3.3 Nichtmonetäre Nutzenelemente

Ein Nutzenelement, das für einen Angreifer bei einer Deanonymisierung bedeutsam sein kann, ihm aber ohnehin in jedem Fall zumindest als Nebenprodukt zukommt, besteht in der Befriedigung der eigenen Neugier durch den Informationszufluß. Denkbar ist z.B., daß ein Mitarbeiter eines sozialwissenschaftlichen Instituts Informationen über einen Arbeitskollegen gewinnen möchte, ohne diese irgendwie verwerten zu wollen. Ebenso kann man sich ein solches Interesse an prominenten Personen vorstellen.

Als weiteres nichtmonetäres Nutzenelement eines Angreifers soll Schadenfreude über die Schädigung anderer Personen oder von Institutionen angesprochen werden. Schädigungen können dabei ihrerseits für den Betroffenen nichtmonetärer, aber auch monetärer Art sein, wie zum Beispiel bei den in den vorherigen Abschnitten genannten rechtswidrigen Verwertungen deanonymisierter Daten. Eine nichtmonetäre Schädigung kann z.B. in der Offenlegung privater Daten bestehen. Opfer können dabei sowohl aus dem eigenen Umfeld

des Angreifers stammen als auch aus dem öffentlichen Leben. Im ersten Fall wären etwa persönliche Neidgefühle mögliche Motive. Im zweiten Fall sind dagegen auch politische Gründe denkbar. Dies gilt auch für eine angestrebte Bloßstellung der amtlichen Statistik als weiteres vorstellbares nichtmonetäres Nutzenelement eines Angreifers. Auch hier können Ärger über den Staat und seine Institutionen oder andere politische Motive zugrunde liegen.

Die verschiedenen Nutzenelemente mögen jeweils als Hauptmotiv für einen Angreifer bedeutsam sein oder als Nebeneffekt auftreten. Einige der genannten Nutzenkomponenten erscheinen eher hypothetisch und konstruiert, insbesondere im Hinblick darauf, daß die Daten an Institutionen der wissenschaftlichen Forschung und an Personen weitergegeben werden, als deren primäres Interesse man die Gewinnung wissenschaftlicher Erkenntnisse unterstellen kann. Mit besonderer Aufmerksamkeit wird man vor allem die Frage prüfen müssen, ob massenhafte Deanonymisierungen mit dem Ziel erfolgversprechend erscheinen, deanonymisierte Daten an Adressenhändler oder an andere professionelle Informationssammler (z.B. Detekteien) mit einem ökonomischen Ertrag zu verkaufen.

Eine besondere Situation stellt der Fall dar, in dem eine Deanonymisierung primär deshalb angestrebt wird, um nachzuweisen, daß man in der Lage ist, Daten zu deanonymisieren. Die Deanonymisierungsleistung und die damit in Teilen der Öffentlichkeit erzielbare Anerkennung kann als umso höher eingestuft werden, je größer die Zahl der deanonymisierten Personen ist. Die öffentliche Aufmerksamkeit dürfte auch mit dem Bekanntheitsgrad der betroffenen Personen steigen, da das öffentliche Interesse im allgemeinen und das der Medien im besonderen an spektakulären Enthüllungen größer ist. Die beschriebene Situation ist deshalb ein Sonderfall, weil hier für einen Angreifer unter Umständen kein Preis zu hoch ist, um den Nachweis der prinzipiellen Machbarkeit einer Deanonymisierung zu erbringen. Es erscheint deshalb unplausibel, diesen Fall unter dem Gesichtspunkt der Unverhältnismäßigkeit der Mittel zu evaluieren.

7.2 Kostenkomponenten eines Deanonymisierungsversuchs

Zu Beginn der Kostendiskussion sei noch einmal auf das bereits in den Vorbemerkungen erwähnte Problem der Kostenüberwälzung hingewiesen. Ein potentieller Angreifer wird nur solche Deanonymisierungskosten in seine Kalkulation einbeziehen, die auch tatsächlich von ihm selbst zu tragen sind. Hat er dagegen zum Beispiel die Möglichkeit, Personal oder Rechenanlagen des wissenschaftlichen Instituts, dem er angehört, für seine Zwecke einzusetzen, sind die dem Institut dadurch entstehenden Kosten für seine persönliche Kosten-Nutzen-Analyse irrelevant.

Kosten zur Deanonymisierung fallen nun im wesentlichen an für die Beschaffung des notwendigen Zusatzwissens, Bearbeitung von Mikrodatenfile oder Identifikationsfile, Erstellung einschließlich Implementierung und Ablauf des Deanonymisierungsalgorithmus' und für die vom Angreifer möglicherweise gewünschte Verschleierung seiner Reidentifikationsversuche zur Vermeidung von Sanktionen. Zu berücksichtigen sind aber auch diese potentiellen Sanktionen selbst und Opportunitätskosten in Form entgangenen Nutzens durch den Verzicht des Angreifers auf andere Handlungsalternativen.

7.2.1 Kosten zur Beschaffung von Zusatzwissen

Kosten zur Beschaffung von Zusatzwissen hängen im allgemeinen wesentlich von der Datenmenge im gewünschten Identifikationsfile ab. Diese Menge wird bestimmt durch die Anzahl der Identifikationsfile-Datensätze und deren Länge, d.h. durch die Anzahl der Überschneidungsmerkmale mit dem amtlichen Mikrodatenfile. Wie in Kapitel 6 bereits dargelegt, stellen diese beiden Größen entscheidende Faktoren für die zu erwartende Anzahl erfolgreicher Deanonymisierungsversuche dar. Die Beschaffungskosten für das Zusatzwissen werden somit stark durch die vom Angreifer angestrebte Zahl an Reidentifikationen beeinflußt. Andererseits nimmt der notwendige Umfang des Identifikationsfile mit wachsendem Informationsgehalt der Überschneidungsmerkmale ab.

Die genannten Einflußgrößen führen zu einer praktisch beliebig großen Spannweite der möglichen Beschaffungskosten. Im Extremfall einer Einzelidentifikation benötigt der Angreifer nur einen einzigen Datensatz als Zusatzwissen, der beispielsweise sogar einfach in Form persönlicher Kenntnisse, etwa über einen Nachbarn ohnehin zur Verfügung steht, so daß dann überhaupt keine

Beschaffungskosten anfallen. Das gleiche gilt für den Fall, daß aus institutseigenen Erhebungen vorliegende Daten als Zusatzwissen dienen sollen. Wäre es dagegen für ein bestimmtes Vorhaben erforderlich, etwa 100 000 Datensätze aus Melderegistern zu erwerben, so sind diese, wie in Abschnitt 6.2.1 dargelegt wurde, wegen der einschlägigen Vorschriften und ihrer restriktiven Anwendung durch die zuständigen Behörden eventuell gar nicht zu erhalten. Die Beschaffungskosten können also in einem solchen Fall als unendlich hoch angesetzt werden.

Zwischen diesen Extremen sind Fälle mit einer großen Streuung der Beschaffungskosten denkbar. Beispielsweise existieren Handbücher über bestimmte Personengruppen, die zwar vergleichsweise wenig Überschneidungsmerkmale, aber immerhin mehrere Tausend Datensätze enthalten. Die reinen Beschaffungskosten für derartiges Zusatzwissen (im wesentlichen die Anschaffungs- oder Ausleihkosten der Handbücher) bleiben relativ niedrig.

Sie erhöhen sich jedoch, wenn eine Kumulation von Daten aus verschiedenen Quellen erforderlich ist. Im einzelnen ist dabei zu denken an Kosten zur Ermittlung der in Frage kommenden Datenquellen, Wegekosten beim Aufsuchen dieser Quellen und Zeitaufwand für deren Durchforstung. Selbst wenn dem Angreifer eine einzige Datensammlung als Zusatzwissen ausreichend erscheint, machen vielleicht Lücken in den darin erfaßten Datensätzen eine Vervollständigung aus anderen Quellen notwendig, so daß dann auch in diesen Fällen eine, wenn auch nicht so umfangreiche, Kumulation stattfindet.

7.2.2 Kosten zur Bearbeitung der Files

7.2.2.1 Übernahme auf maschinenlesbare Datenträger

Die Notwendigkeit zur Übernahme von Daten auf maschinenlesbare Träger hängt zunächst von der Deanonymisierungsstrategie des Angreifers ab. Strebt er zum Beispiel eine Einzelidentifikation an, so ist alternativ auch an die manuelle Anwendung eines einfachen Abgleichverfahrens zu denken. Dann fallen gar keine Kosten für eine Datenübernahme an.

Derartige Kosten entstehen auch nicht für solche Daten, die von vornherein in geeigneter maschinenlesbarer Form, z.B. direkt aus Datenbanken beschafft werden. So ist davon auszugehen, daß das Mikrodatenfile an Forschungseinrichtungen in maschinenlesbarer Form geliefert wird. Auch bei Daten, die als

Identifikationsfile in Betracht kommen, ist eine zunehmende Verbreitung der elektronischen Speicherung beobachtbar. Beispiel dafür sind die Einwohnermelderegister. Aber ebenso im kommerziellen Bereich nimmt die elektronische Speicherung von Einzeldaten zu, so insbesondere bei Auskunfteien oder Adressenhändlern. In diesen Fällen sind die Kosten für die Datenübernahme auf maschinenlesbare Träger bereits in den Beschaffungskosten enthalten. Neben dem käuflichen Erwerb kommt für den Angreifer auch die Nutzung solcher elektronisch gespeicherter Daten in Frage, die seinem Institut zu Forschungszwecken z.B. aus eigenen Erhebungen zur Verfügung stehen.

In den übrigen Fällen sind die beschafften Daten des Zusatzwissens jedoch in der Regel auf maschinenlesbare Träger zu übernehmen. Die Kosten hierfür hängen dann im wesentlichen von der Datenmenge, d.h. von der Anzahl der Datensätze und der Überschneidungsmerkmale, ab. Diese Datenmenge ist für Massenidentifikationen oder Massenfischzüge im allgemeinen wohl so beträchtlich, daß die entsprechende Arbeitszeit für die Datenübernahme ein bedeutendes Gewicht in der Kostenkalkulation eines potentiellen Angreifers darstellt. Die im Rahmen der allgemeinen Überlegungen zum Zusatzwissen herangezogene Annahme einer Reidentifikationsquote von ca. 0,001 bei zehn Überschneidungsmerkmalen verdeutlicht dies: Für zehn Deanonymisierungen wären durchschnittlich ca. 10 000 Datensätze des Zusatzwissens und somit ca. 100 000 Daten zu übertragen.

Zu den reinen Übernahmekosten kommen unter Umständen Kosten für Plausibilitätskontrollen hinzu, die der Angreifer an den aufgenommenen Daten vornimmt, etwa um Übertragungsfehler zu korrigieren.

7.2.2.2 Fehlerabschätzungen

Eine Einflußgröße für die Reidentifikationsquote bilden Datenfehler im Mikrodaten- und Identifikationsfile. Sie sind dem Angreifer in aller Regel unbekannt. Er kann sie dann allenfalls abschätzen. Solche Schätzungen sind jedoch wegen der im allgemeinen mangelnden Kenntnisse über Plan, Verlauf und insbesondere über die Fehlerprozesse bei der Erhebung und die bei nachträglichen Plausibilitätskontrollen vorgenommenen Datenmanipulationen sehr schwierig. Neben vagen Vermutungen bleiben dem Angreifer eventuell Fehlerrechnungen ähnlicher Erhebungen als Anhaltspunkt. Die große Unsicherheit wird ihn aber wohl dazu veranlassen, relativ wenig Zeit auf Fehlerschätzungen zu verwenden, so daß die resultierenden Kosten eher niedrig anzusetzen sind,

insbesondere etwa im Vergleich zu den Kosten für die oben erörterte Datenübernahme auf maschinenlesbare Träger.

7.2.2.3 Anpassung der Überschneidungsmerkmale

Da Deanonymisierungsalgorithmen im Prinzip durch die Suche nach sich entsprechenden Datensätzen im Mikrodatenfile und Identifikationsfile charakterisiert sind, muß die Vergleichbarkeit der Ausprägungen der jeweiligen Überschneidungsmerkmale gewährleistet sein. Eine maschinelle Datenauswertung erfordert etwa, daß das Deanonymisierungsprogramm inhaltlich gleiche Ausprägungen dieser Merkmale in beiden Files erkennt. Dazu sind eventuell unterschiedliche Signierungen oder Kategorienabgrenzungen einander anzupassen.

Nimmt der Angreifer diese Rekodierung für jeden Datensatz einzeln vor, so kann aus dieser Prozedur je nach Fileumfang ein sehr großer Aufwand an Arbeitszeit resultieren. Alternativ wird daher vielleicht versucht, unterschiedliche Schlüssel oder Klassifizierungen mit Hilfe von Formeln ineinander zu überführen und die eigentliche Anpassung für die einzelnen Datensätze dann maschinell durchführen zu lassen. Dies gelingt aber nur, wenn die Signierungen bzw. Abgrenzungen der Überschneidungsmerkmale in beiden Files vollständig bekannt sind, damit alle möglichen Fälle abgedeckt werden können. Anstelle der Arbeitszeit für die manuelle Anpassung der einzelnen Datensätze tritt dann der Zeitaufwand für die Aufstellung der Formeln und deren Programmierung, der je nach Komplexität von Schlüsseln, Gliederungen und Anzahl der zu bearbeitenden Überschneidungsmerkmale mehr oder weniger hoch ausfallen kann.

Zusätzliche Überlegungen und damit Arbeitszeit werden für den Angreifer notwendig, wenn sich die Gliederungen eines Überschneidungsmerkmals in beiden Files überlappen und sich dieses Problem durch Klassenverbreiterung nicht zufriedenstellend lösen läßt. Ein eindeutiger Abgleich unter Verwendung dieses Merkmals ist dann nicht möglich. In diesem Fall muß der Angreifer entscheiden, ob er die Ähnlichkeit von Ausprägungen des Überschneidungsmerkmals in beiden Files unter Inkaufnahme von Fehlzuordnungen trotzdem für einen Abgleich nutzen will oder nicht. Ebenso kann sich die Frage stellen, ob und wie zwei eng korrelierte, aber nicht identische Merkmale in den Files zu einem Überschneidungsmerkmal zusammengefaßt werden sollten, um auf diese Informationen nicht verzichten zu müssen. Auch hier sind zusätzliche Fehlzuordnungen von Datensätzen als negative Effekte zu berücksichtigen.

7.2.3 Programmentwicklung und -ablauf

Kosten für Deanonymisierungssoftware entfallen, wenn der Angreifer auf eine maschinelle Auswertung verzichtet. Dies ist z.B., wie schon erwähnt, bei einer Einzelidentifikation denkbar, bei der der Angreifer als Zusatzwissen lediglich seine persönlichen Kenntnisse über eine einzige Person, also nur einen Datensatz verwendet. Stattdessen erfordert ein manueller Abgleich jedoch Arbeitszeit, die von der Filelänge und der Anzahl der abzugleichenden Überschneidungsmerkmale abhängt.

Bei Massenidentifikationen bzw. -fischzügen läßt sich ein EDV-Einsatz dagegen kaum vermeiden. Dabei ist davon auszugehen, daß einfache Sortierverfahren relativ kostengünstig auf dem Softwaremarkt zu beschaffen und eventuell auch recht einfach auf die konkrete Deanonymisierungssituation zuzuschneiden sind. Dem Angreifer bleibt aber ebenso die Möglichkeit, das Programm selbst zu erstellen oder Experten damit zu beauftragen. Der Arbeitsaufwand bzw. das Auftragsentgelt nehmen dann mit der Komplexität der Software zu. Aus so komplizierten Algorithmen, wie etwa dem im Rahmen dieses Projekts getesteten diskriminanzanalytischen Reidentifikationstechnik von Paaß/Wauschkuhn (1985), können, wie sich gezeigt hat, mehrere Personen-Arbeitsmonate resultieren.

Neben dem Programm selbst führt seine Implementierung in die zu verwendende EDV-Anlage, einschließlich der Durchführung von Probeläufen und Fehlerkorrekturen, zu Kosten. Auch diese bestimmen sich maßgeblich nach der Komplexität der Software.

Schließlich verursacht der Programmablauf an sich, d.h. die eigentliche Deanonymisierung, ebenfalls Kosten. Sie hängen ab von Rechenzeit, beanspruchter Kapazität der Anlagen und damit vom Umfang des Programms sowie der zu verarbeitenden Datenmenge. Wegen des im EDV-Bereich bekanntlich schnellen technischen Fortschritts sind im Laufe der Zeit allerdings entsprechend erhöhte Verarbeitungsgeschwindigkeiten sowie erweiterte Speichermöglichkeiten und daher eine Abnahme dieser Kosten zu erwarten.

Maschinenkosten dürften im übrigen wohl noch am ehesten überwälzbar sein, da in wissenschaftlichen Institutionen EDV-Ressourcen in der Regel zur Verfügung stehen, ohne daß eventuell in jedem Einzelfall des Zugriffs auf solche Ressourcen geprüft wird, wofür genau sie verwendet werden.

7.2.4 Kosten zur Verschleierung von Deanonymisierungsversuchen

Ein Angreifer muß zunächst darauf bedacht sein, über Verwendung oder Verbleib der Datenfiles Dokumentationen zu verhindern, die auf eine Deanonymisierung hinweisen können. Zeit und Mühe, die er dafür aufzuwenden hat, hängen davon ab, wie umfassend und sicher das diesbezügliche Kontroll- und speziell das Protokollierungssystem seines Forschungsinstituts ausgestaltet ist. Bei Benutzung von Institutskapazitäten sind zum einen etwaige automatische Aufzeichnungen der Rechenanlage über verwendete Datenquellen und Programme, Rechenzeit oder benötigten Speicherplatz zu umgehen oder nachträglich zu löschen. Zum anderen muß sich der Angreifer über eventuelle Vorschriften bezüglich schriftlicher Dokumentationen hinwegsetzen oder diese im nachhinein fälschen.

Zusätzliche Verschleierungskosten z.B. in Form von Gefälligkeiten oder gar "Schweigegeldern" könnten anfallen, wenn andere Personen über die Deanonymisierung zumindest teilweise informiert sind oder eine solche vermuten. Als Mitwisser kommen dabei z.B. Verhandlungspartner beim Verkauf der deanonymisierten Daten oder Institutspersonal, dem zufällig oder aufgrund von Kontrollen die Aktivitäten des Angreifers aufgefallen sind, in Frage.

Versuche des Angreifers, Teile der in den vorigen Abschnitten angesprochenen Kosten auf die Forschungseinrichtung zu überwälzen, erhöhen die Wahrscheinlichkeit für die Aufdeckung der Reidentifikationsversuche. Um dies zu vermeiden, wären dann also erhöhte Verschleierungsanstrengungen erforderlich.

Im übrigen muß der Angreifer selbst bei Deanonymisierungsmotiven, die eine nachträgliche Preisgabe seiner Methoden erfordern, während des Reidentifikationsprozesses dessen Aufdeckung verhindern, um sein Vorhaben nicht zu gefährden.

7.2.5 Rechtliche Sanktionen

Eine wesentliche Kostenkomponente bilden für einen Angreifer Sanktionen, die im Fall eines offensichtlichen Bruchs des Reidentifikationsverbots über ihn verhängt werden können. Dabei ist an strafrechtliche, privatrechtliche und disziplinarische Konsequenzen zu denken.

Für die Kalkulation des Angreifers ist es wichtig, ob er von vornherein eine Offenlegung der Reidentifikation plant, etwa bei einer nicht anonymen Veröffentlichung, oder ob die Deanonymisierung nur entgegen seiner Absicht publik werden könnte. Im ersten Fall wird er ziemlich sicher mit Sanktionen rechnen, im zweiten dagegen vielleicht nur mit einer eher geringen Wahrscheinlichkeit. Entscheidend sind also die vom Angreifer zu erwartenden Kosten, d.h. etwa deren Ausmaß gewichtet mit der Wahrscheinlichkeit, daß sie überhaupt zum Tragen kommen.

Damit der Angreifer die Höhe von Sanktionen jedoch in sein Kalkül einbeziehen kann, muß er sie kennen. Entscheidend ist es daher, daß einem potentiellen Angreifer die aus einer Reidentifikation resultierenden rechtlichen Konsequenzen mitgeteilt werden. Dies kann im Zuge der Lieferung eines Mikrodatenfile an ein Forschungsinstitut durch hinreichende Verdeutlichung von seiten der amtlichen Statistik, aber auch durch das Institut selbst geschehen.

Die zweite oben genannte Komponente der zu erwartenden Kosten, nämlich die Wahrscheinlichkeit der Aufdeckung, wird dagegen wesentlich durch Umfang und Qualität von Kontrollmaßnahmen in der Forschungseinrichtung bestimmt, sowie von den Verschleierungsanstrengungen des Angreifers.

7.2.6 Entgangener Nutzen

Wie schon im Zusammenhang mit den "Eingesparten Ausgaben" in Abschnitt 7.1.1 ausführlich erörtert, ist beim Vergleich zweier Handlungsalternativen der unmittelbare Nutzen der nicht gewählten Alternative den unmittelbaren Kosten der gewählten Alternative hinzuzurechnen. Dabei sei auf das Opportunitätskostenprinzip verwiesen.

Als alternative Handlungsmöglichkeit anstelle der Deanonymisierung kann hier nun sinnvollerweise eine solche zum Vergleich herangezogen werden, die eine unbeeinträchtigte Fortsetzung der beruflichen und wissenschaftlichen Laufbahn des potentiellen Angreifers beinhaltet. Entscheidet sich der Wissenschaftler für eine Deanonymisierung, entgeht ihm dann also bei ihrer Aufdeckung oder Offenlegung insbesondere der Nutzen, den er im anderen Fall aus seiner zukünftigen wissenschaftlichen Arbeit ziehen könnte.

Bei der Bewertung des zu erwartenden Nutzenentgangs spielen auch diesmal die zugehörige Wahrscheinlichkeit und damit die Effektivität von Kontroll-

maßnahmen eine große Rolle, aber auch wieder die Verschleierungsanstrengungen des Angreifers.

Der Nutzenentgang selbst umfaßt zunächst als monetäre Komponente die mögliche, auf Beeinträchtigungen seiner weiteren wissenschaftlichen Arbeit basierende Verminderung des zukünftig an ihn gezahlten Entgelts für seine berufliche Tätigkeit. Der Wert hängt unter anderem von der verbleibenden Lebensarbeitszeit ab. Er ist zwar wegen des im allgemeinen eher langfristigen Bezugs mit einer relativ großen Unsicherheit behaftet, kann aber aus demselben Grund recht beträchtlich ausfallen. Das nichtmonetäre Element des entgangenen Nutzens resultiert wohl hauptsächlich aus einem möglichen Reputationsverlust des Angreifers.

7.3 Alternative Datenbeschaffungsmethoden

Im Zusammenhang mit den "Eingesparten Ausgaben" wurde in Abschnitt 7.1.1 formal die plausible Aussage abgeleitet, daß sich der Aufwand einer Deanonymisierung für einen potentiellen Angreifer nicht lohnt, wenn eine alternative Beschaffungsmethode annähernd gleichwertige Daten mit deutlich niedrigerem Aufwand liefert. Dabei ist nicht unbedingt die Beschaffung identischer Daten notwendig. Beispielsweise kann eine Alternative zu einem Fischzug Daten anderer Personen hervorbringen, die für den Angreifer aber denselben Zweck erfüllen. Nachstehend sollen daher noch kurz alternative Datenbeschaffungsmöglichkeiten und ihre Kosten angesprochen werden.

Zuvor ist allerdings festzuhalten, daß es Situationen gibt, in denen sich Alternativen zur Deanonymisierung geradezu verbieten. Dies ist dann der Fall, wenn das Motiv des Angreifers die Offenlegung seiner Methoden einschließt, wie z.B. beim Versuch des Nachweises der prinzipiellen Machbarkeit von Deanonymisierungen und bei einer angestrebten öffentlichen Anerkennung solcher "Hackerleistungen".

Als alternative Datenbeschaffungsmöglichkeit stellen eigene Recherchen des potentiellen Angreifers sicher eine naheliegende Vorgehensweise dar. Darunter soll im folgenden eine große Spannweite von Möglichkeiten verstanden werden, die so unterschiedliche Methoden wie eigene wissenschaftliche Erhebungen und schlichtes Ausfragen von Personen aus seinem eigenen Umfeld umfaßt.

Eine weitere plausible Möglichkeit ist die Nutzung fremder Recherchen, auf die hier aber nicht näher eingegangen wird. Zu denken ist dabei etwa an die Erteilung eines Auftrages an eine Auskunftei oder Detektei.

Eine wichtige Rolle für die bei einem reinen Kostenvergleich notwendige sonstige Gleichwertigkeit zweier Beschaffungsalternativen spielt nun sicher die Wahrscheinlichkeit, mit der die jeweilige Methode eine bestimmte, gleiche Anzahl von Datensätzen liefert. In dieser Hinsicht werden im folgenden die Deanonymisierungsstrategien "Identifikation" und "Fischzug" den eigenen Recherchen getrennt gegenübergestellt.

7.3.1 Eigene Recherchen versus Identifikation

Geht man zur Vereinfachung von der gleichen Datenqualität beider Beschaffungsmethoden aus, so hängt deren Beurteilung sicherlich vom Verhältnis zwischen der Erfolgswahrscheinlichkeit eigener Recherchen und der Deanonymisierungswahrscheinlichkeit jeweils für jeden der gesuchten, bei Identifikationen stets vorgegebenen Datensätze ab. In Anlehnung an den Sprachgebrauch bei Erhebungen soll die genannte Erfolgswahrscheinlichkeit im folgenden als Antwortwahrscheinlichkeit bezeichnet werden.

Zur Verdeutlichung der Bedeutung dieser Größen wird einerseits vereinfachend angenommen, daß sowohl die Deanonymisierungswahrscheinlichkeit P_D als auch die Antwortwahrscheinlichkeit P_A für alle gezielt gesuchten Datensätze gleich ist, und andererseits, daß für die einzelnen Datensätze sowohl die Deanonymisierungsversuche unabhängig voneinander verlaufen als auch die einzelnen Recherchen. In diesem Fall ergibt sich die Wahrscheinlichkeit, alle n gesuchten Datensätze zu reidentifizieren, zu P_D^n und die Wahrscheinlichkeit, diese n Datensätze durch eigene Recherchen zu gewinnen, zu P_A^n. Die Erfolgswahrscheinlichkeit für die gesamte Identifikation und damit wohl ebenso der vom Angreifer zu erwartende Nutzen, ist für beide Methoden also nur gleich, wenn auch die entsprechenden Einzelwahrscheinlichkeiten P_D und P_A gleich sind.

Geht man dagegen etwa von der in Kapitel 5 angenommenen Größenordnung $P_D = 0{,}001$ aus, wenn der Angreifer Personen im Mikrozensus sucht, ohne zu wissen, ob diese daran teilgenommen haben, und unterstellt eine eher niedrige Antwortwahrscheinlichkeit von $P_A=0{,}1$, so würde er bei nur einem gezielt gesuchten Datensatz den erwarteten Nutzen eigener Recherchen bei ent-

sprechender Gewichtung hundertmal höher ansetzen als den eines Deanonymisierungsversuchs. Handelt es sich gar um eine Identifikation von n = 100 Datensätzen, so beträgt die Erfolgswahrscheinlichkeit für den gesamten Deanonymisierungsversuch nur noch $P_D^n = 0{,}001^{100} = 10^{-300}$ und für die gesamten eigenen Recherchen nur $P_A^n = 0{,}1^{100} = 10^{-100}$, so daß deren erwarteter Nutzen jetzt sogar 10^{200} mal höher liegt. Abgesehen davon, daß solch geringe Wahrscheinlichkeiten den erwarteten Nutzen für beide Alternativen ohnehin beträchtlich verringern, wenn dieser, wie hier unterstellt, aus der Gewichtung des bei Erfolg möglichen Nutzens mit eben diesen Wahrscheinlichkeiten P_D^n bzw. P_A^n resultiert, zeigen die großen Unterschiede im erwarteten Nutzen zwischen beiden Methoden, daß unter diesen Voraussetzungen nicht von zwei gleichwertigen Beschaffungsalternativen gesprochen werden kann. Unter solchen Bedingungen wäre ein reiner Kostenvergleich also nicht sinnvoll.

Während man bei der gezielten Suche nach Personen, die dem Angreifer unbekannt sind, annehmen kann, daß er, wie eben vorausgesetzt, im allgemeinen nichts von deren möglicher Beteiligung an amtlichen Stichproben weiß, ist im eigenen Umfeld des Angreifers durchaus damit zu rechnen, daß ihm die eventuelle Teilnahme bekannt ist. Die Deanonymisierungswahrscheinlichkeit P_D liegt dann, wie in Abschnitt 4.1 erläutert, zwar im allgemeinen deutlich höher; andererseits wächst aber wohl auch die Antwortwahrscheinlichkeit P_A bei eigenen Recherchen wegen der persönlichen Kenntnisse und guten Gelegenheiten zur Ausforschung in bekannter Umgebung. Der Angreifer könnte P_A daher recht nahe bei 1 ansetzen, beispielsweise bei $P_A = 0{,}75$. Geht man gleichzeitig etwa von einer ebenfalls hohen, aber dennoch niedrigeren Deanonymisierungswahrscheinlichkeit $P_D = 0{,}5$ aus, so übersteigt der erwartete Nutzen der eigenen Recherchen denjenigen der Deanonymisierung bei einem einzigen gesuchten Datensatz nur um den Faktor 1,5. Sollen dagegen etwa zehn Personen identifiziert werden, erhöht sich der Faktor allerdings schon auf $1{,}5^{10} = 58$, so daß auch dann keine Gleichwertigkeit vorläge.

Dazu wäre, wie schon festgehalten, die Gleichheit $P_A = P_D$ der Erfolgswahrscheinlichkeiten nötig. Exemplarisch soll ein solcher Fall, etwa $P_A = P_D = 0{,}5$ betrachtet werden. Da die eigenen Recherchen im wesentlichen durch Ausfragen und Ausforschen innerhalb des eigenen Umfeldes eventuell "nebenbei" vonstatten gehen, können die Kosten dafür unter günstigen Umständen sogar fast bei 0 angesiedelt werden. Auf der anderen Seite entfallen von den in Abschnitt 7.2 diskutierten Kostenkomponenten eines Deanonymisierungsvorhabens diejenigen zur Beschaffung und Bearbeitung des Zusatzwissens, da dieses bereits in Form der persönlichen Kenntnisse vorhanden ist. Sollte das

Mikrodatenfile schon maschinenlesbar vorliegen, entstehen ferner keine Kosten für dessen bei maschineller Auswertung sonst notwendigen Übernahme. Es verblieben dann Kosten für Entwicklung und Ablauf des Programms oder, bei manuellem Abgleich, in Form von Arbeitszeit. Außerdem lägen Verschleierungskosten vor und, jeweils etwa gewichtet mit der Aufdeckungswahrscheinlichkeit für das Deanonymisierungsvorhaben, Sanktionen sowie der Nutzenentgang aus dem Verzicht des Angreifers auf die eigenen Recherchen als Handlungsalternative und damit auch die unbeeinträchtigte Fortsetzung seiner beruflichen Laufbahn.

Bei Einbeziehung aller genannten Kostenkomponenten einer Deanonymisierung erscheint es somit sehr wahrscheinlich, daß eigene Recherchen in einem wie hier gelagerten Fall eine deutlich kostengünstigere Alternative zu den Reidentifikationen darstellen. Beschränkt man sich dagegen bei den Deanonymisierungskosten, aus welchen Gründen auch immer, auf die nicht mit der Aufdeckungswahrscheinlichkeit gewichteten Komponenten, vernachlässigt man also Sanktionen und Nutzenentgang, ist eine solche Aussage nicht mehr ohne weiteres möglich, da dann auch für die Reidentifikation eventuell nur noch ein recht geringer Aufwand verbleibt.

Versteht man im übrigen den Begriff Identifikation und speziell Massenidentifikation nicht in dem eben implizit unterstellten engen Sinn, daß ein Erfolg nur vorliegt, wenn alle Personen aus einer vorgegebenen Menge richtig zugeordnet werden, sondern in dem weiteren Sinn, daß schon die korrekte Zuordnung nur einiger dieser vorgegebenen Personen zumindest einen Teilerfolg darstellt, so lassen sich entsprechende Überlegungen formal wie bei einem Fischzug anstellen. Dazu sei auf den folgenden Abschnitt verwiesen.

7.3.2 Eigene Recherchen versus Fischzug

Eigene Recherchen können etwa dann als gleichwertige Datenbeschaffungsalternative zu einem Fischzug angesehen werden, wenn die durchschnittlich zu erwartende Anzahl an selbst "erhobenen" bzw. deanonymisierten Datensätzen ungefähr gleich ist und man bei jedem möglichen zu erlangenden Datensatz vereinfachend einen ungefähr gleichen Nutzen für den Angreifer unterstellt. Zur Vereinfachung und größenordnungsmäßigen Abschätzung wird ferner neben gleicher Datenqualität beider Methoden wieder angenommen, daß die Antwortwahrscheinlichkeit P_A und die Deanonymisierungswahrscheinlichkeit P_D jeweils für alle n_A bzw. n_D potentiell gewinnbaren Datensätze gleich ist. In diesem Fall

beträgt die im Durchschnitt zu erwartende Anzahl an Erfolgen gerade $E_A = n_A P_A$ bzw. $E_D = n_D P_D$, wenn man wieder für die einzelnen in Frage kommenden Datensätze voneinander unabhängige Reidentifikationsversuche bzw. voneinander unabhängige eigene Recherchen voraussetzt.

Für den Fall der Unkenntnis des Angreifers über die Mikrozensus-Teilnahme irgendwelcher Identifikationsfile-Personen sei nun ein weiteres Mal zur Illustration die bereits mehrfach unterstellte Wahrscheinlichkeit $P_D = 0{,}001$. Zur Reidentifikation von im Durchschnitt nur $E_D = 10$ Datensätzen müßte der Angreifer wegen $E_D = n_D P_D$ dann also $n_D = 10\,000$ Versuche einkalkulieren. Im Falle einer eher niedrigen Antwortwahrscheinlichkeit bei den eigenen Recherchen von $P_A = 0{,}1$ sind dagegen $E_A = 10$ Erfolge bereits bei $n_A = 100$ Versuchen zu erwarten.

Setzt man im Fall einer eigenen Erhebung des Angreifers die in den Sozialwissenschaften realistische Größenordnung von ca. 150 DM an Kosten pro zustandegekommenem "Interview" an, so resultieren insgesamt 1 500 DM. Ohne den im folgenden Kapitel dargestellten Ergebnissen der Zuordnungsexperimente vorzugreifen, liegt gleichwohl bereits jetzt die Vermutung nahe, daß für die alternativ notwendigen 10 000 Deanonymisierungsversuche deutlich höhere Kosten allein zur Beschaffung und Bearbeitung des Zusatzwissens einschließlich seiner Übernahme auf maschinenlesbare Datenträger und den eigentlichen Abgleich mit Hilfe eines einfachen Verfahrens anfielen. In derartigen Fällen stellen eigene Recherchen somit wieder eine im Vergleich zu Reidentifikationen kostengünstigere Datenbeschaffungsmethode dar.

Ist dagegen, etwa ausgehend von markanten Mikrodatenfile-Datensätzen, die sich durch seltene Ausprägungen in den Überschneidungsmerkmalen von den übrigen Mikrodatenfile-Datensätzen abheben, schnell und einfach eine enge Eingrenzung der möglichen zugehörigen Personen realisierbar, so müßte der Angreifer auch nur ein entsprechend kleineres Identifikationsfile aufbauen, um die vorgegebene zu erwartende Anzahl an Reidentifikationen zu erreichen. Die Kosten für die Beschaffung des Zusatzwissens, seine Bearbeitung, Entwicklung und Ablauf des Deanonymisierungsalgorithmus' sowie Verschleierungskosten könnten dann zusammen ebenfalls recht gering ausfallen. Würde man jetzt die übrigen, mit der Aufdeckungswahrscheinlichkeit gewichteten Kostenelemente in der Kalkulation vernachlässigen, d.h. Sanktionen und Nutzenentgang aus dem Verzicht auf eigene Recherchen als Handlungsalternative, so ist keine eindeutige Aussage mehr über die kostengünstigere Methode möglich. Bei Einbeziehung der zuletzt genannten, verbleibenden Kostenkomponenten erscheinen jedoch wieder eigene Recherchen ökonomisch sinnvoller.

Am Ende der Gegenüberstellungen von Deanonymisierung und eigenen Recherchen als alternativer Datenbeschaffungsmethode sei noch einmal darauf hingewiesen, daß mit fremden Recherchen noch weitere grundsätzliche Möglichkeiten in Frage kommen. Der Vergleich mit Reidentifikationen verliefe aber wohl ähnlich zur hier vorgenommenen Erörterung in der Regel zugunsten eigener Recherchen, so daß darauf verzichtet werden soll.

Nach den in diesem Kapitel dargestellten Kosten-Nutzen-Überlegungen allgemeiner Art, wird im Rahmen der folgenden Diskussionen einzelner ausgewählter Deanonymisierungssituationen untersucht, in welchem Maße hier erörterte Nutzen- und Kostenkomponenten dort zum Tragen kommen. Diese Konkretisierung soll damit insbesondere eine genauere Quantifizierung einzelner bisher nur grob angesprochener Elemente beinhalten.

8 Entwicklung konkreter Angriffsszenarien für die Operationalisierung der faktischen Anonymität und Datenbasis für die empirische Überprüfung einzelner Szenarien*

Wie die vorangegangenen Analysen der wichtigsten Bestimmungsfaktoren des Reidentifikationsrisikos anonymisierter Einzelangaben gezeigt haben, ist dieses Risiko von einer Vielzahl unterschiedlicher Faktoren abhängig. Hierzu zählen sowohl die Eigenschaften des Zusatzwissens und des Mikrodatenfile (beispielsweise Umfang und Repräsentativität), der Informationsgehalt bzw. Auflösungsgrad der Überschneidungsmerkmale, der Grad der Kompatibilität bzw. Inkompatibilität der verwendeten Daten sowie die zur Verfügung stehenden Reidentifikationstechniken. Schon die Bestimmung der Bedeutung jedes einzelnen dieser Faktoren für das Reidentifikationsrisiko erweist sich als schwierig. Die Auswirkungen der kombinierten Effekte der einzelnen Komponenten auf das Reidentifikationsrisiko eindeutig festzulegen, muß daher auch bei einer sehr sorgfältigen Analyse als in hohem Maße spekulativ erscheinen. Eine lediglich analytische Bestimmung eines potentiellen Reidentifikationsrisikos erscheint aus diesem Grund nicht ausreichend.

Ausgehend von diesen Überlegungen wird in dieser Studie daher ein vergleichsweise neuer Weg beschritten, indem potentielle Reidentifikationsrisiken von existierenden Mikrodatenfiles (Mikrozensus) unter Verwendung von Identifikationsfiles, die ebenfalls empirische Daten enthalten, überprüft werden.[2]

---

* Autoren: Peter Knoche, Heike Wirth (8.1)
Uwe Blien, Peter Knoche, Walter Müller, Heike Wirth (8.2)

2 In der bereits mehrfach erwähnten Untersuchung der GMD wurde zwar auf ein reales Mikrodatenfile zurückgegriffen, das Identifikationsfile wurde jedoch aus dem Mikrodatenfile synthetisch generiert. Die Untersuchung von Brunnstein und Fischer-Hübner wurde vollständig mit einem künstlich erzeugten Datenfile unter nicht näher konkretisierten Randbedingungen durchgeführt. Ein zur vorliegenden Studie vergleichbares Untersuchungsdesign, d.h. die Heranziehung von zwei empirischen Datenfiles liegt unseres Wissens bislang nur der von Block und Olsson (1976) durchgeführten Studie zugrunde. Hierbei wurden Angaben des schwedischen Einwohnerregisters mit Interviewdaten abgeglichen. Da in dieser Studie als eine der wichtigsten Randbedingungen allerdings von einem amtsinternen Datenangriff ausgegangen wurde, sind die Resultate nicht unmittelbar für die Bewertung eines Reidentifikationsrisikos bei einem externen Datenangriff, wie von §16 Abs.6 BStatG vorausgesetzt, übertragbar.

Während in theoretischen Risikoberechnungen oder Simulationsstudien in der Regel vom reinen Wirken einzelner Faktoren ausgegangen und per Modellannahmen der Einfluß störender Faktoren vielfach ausgeschaltet wird, gehen wir einen fast konträren Weg. Im Mittelpunkt des Interesses steht das erzielbare Endergebnis von realistischen Angriffsversuchen, bei denen all die Schwierigkeiten und Störfaktoren die in der Unreinheit der Empirie stets gegeben sind, nicht künstlich weggedacht, sondern in einer Weise wirksam werden, wie dieses in einem tatsächlichen Angriffsversuch ebenfalls gegeben wäre. Wenn es uns dabei auch nicht möglich ist, die Bedeutung einzelner Faktoren säuberlich zu isolieren, so kann jedoch im Gegenzug der Aufwand realistischer Angriffsversuche ermittelt und mit seinem Ertrag im Sinne erfolgreicher Deanonymisierungen verglichen werden.

Hierfür werden unter Rückgriff auf die Szenario-Technik (Berekoven et al. 1987), die in Kapitel 5 erarbeiteten wissenschaftlich begründbaren und wissenschaftsfremden Deanonymisierungsmotive wieder aufgegriffen und mit den in Kapitel 6 als potentiell riskant eingestuften Zusatzwissensquellen sowie den unterschiedlichen Angriffsstrategien (Einzelsuche/-fischzug; Massenidentifikation/-fischzug) kombiniert. Aus diesen Elementen werden fünf Angriffsszenarien entwickelt. Sie sollen möglichst die Obergrenze bestehender Risiken abdecken. Deshalb sind sie so konstruiert, daß sie Gefährdungssituationen abbilden, mit denen im jeweils schlimmsten anzunehmenden Fall zu rechnen wäre.

Eine empirische Analyse aller fünf Angriffsszenarien ist aus zeitlichen wie auch finanziellen Gründen nicht möglich; daher bleibt die empirische Überprüfung des Reidentifikationsrisikos auf zwei Subszenarien beschränkt. Diese werden allerdings in verschiedenen Varianten und mit unterschiedlichen Reidentifikationstechniken überprüft. Die verbleibenden drei Subszenarien werden unter Zugrundelegung von plausiblen Randbedingungen sowie den Ergebnissen der empirischen Überprüfung einer argumentativen Kosten-Nutzen-Analyse unterzogen.

Im folgenden sollen zunächst nochmals kurz die wichtigsten Befunde zu möglichen Deanonymisierungsmotiven und Zusatzwissen skizziert werden, die den im anschließenden Kapitel beschriebenen Szenarien zugrunde liegen. An die Darstellung der fünf Angriffsszenarien schließt sich eine ausführliche Beschreibung der für die empirischen Analysen verwandten Datenbasis an.

8.1 Deanonymisierungsmotive und Zusatzwissen als Grundlage für die Entwicklung von Angriffsszenarien im Wissenschaftsbetrieb

Die Prüfung einer Vielzahl unterschiedlicher Bedingungskonstellationen der sozialwissenschaftlichen Datennutzung in Kapitel 5 konnte aufzeigen, daß die in der Wissenschaft benutzten Datenanalysemethoden zwar vielfach die Verwendung von Einzelangaben voraussetzen. Ein Personenbezug ist hierbei jedoch weder notwendig noch in irgendeiner Form für die Analyseziele nützlich. Nach der vorliegenden Analyse erscheint es allenfalls in Grenzfällen vorstellbar, daß deanonymisierte Daten in einer Zwischenphase des Forschungsprozesses in Verbindung mit einer eigenen Erhebung von Nutzen sein könnten. Als mögliche wissenschaftlich begründbare Motive wurden in Kapitel 5 genannt:

- die Verwendung von deanonymisierten Mikrodaten als Basis für eine eigene Stichprobenziehung;

- die Nutzung deanonymisierter Mikrodaten, um einen Methodenvergleich zwischen amtlicher Statistik und empirischer Sozialforschung durch die Befragung eines identischen Personenkreises vorzunehmen;

- Erweiterung der eigenen Datenbasis mit den deanoymisierten Datensätzen eines Mikrodatenfile, die zum Zwecke der Zusammenführung deanonymisiert werden müssen.

Für den Bereich von Deanonymisierungsmotiven (vgl. Abschnitt 5.2), die nicht auf der Handlungslogik der Wissenschaft beruhen, sondern auf persönlich-privaten Interessenlagen eines Wissenschaftlers oder einer sonstigen Person, die im Wissenschaftsbereich Zugang zu Einzelangaben haben könnte, sind dagegen eine Vielzahl hypothetischer Motive vorstellbar. Diese reichen von persönlicher Neugier über ökonomisch motivierte Deanonymisierungsversuche (beispielsweise dem Verkauf von Adressen) bis hin zu eindeutig krimineller Motivation, bei welcher deanonymisierte Daten etwa für Erpressungsversuche herangezogen werden könnten. Da einzelne dieser nicht-wissenschaftlichen Motive auf die Reidentifikation einer oder nur weniger Personen abzielen, könnte in diesen Fällen der Aufwand für die Beschaffung des Zusatzwissens unter Umständen gering sein. Wissenschaftsfremde Deanonymisierungsversuche könnten damit die untere Grenze des zu leistenden Aufwands darstellen. Deshalb werden sie bei der Operationalisierung der faktischen Anonymität und damit in den Subszenarien ebenfalls berücksichtigt, obwohl sie nicht

den Handlungsmotivationen entsprechen, die man einem Wissenschaftler unterstellen darf.

Die eingehende Analyse der in den Sozialwissenschaften vorhandenen oder beschaffbaren - und als Zusatzwissen nutzbaren - Datenquellen in Kapitel 6 wiederum hat aufgezeigt, daß zwei Typen von Zusatzwissen auf ihr Gefährdungspotential für amtliche Mikrodaten näher untersucht werden müssen:

- Für den Bereich von öffentlich oder für einen beschränkten Personenkreis zugänglichen Registern oder privater Datenquellen wurden insbesondere berufsgruppenspezifische Adreßbücher als möglicherweise riskant charakterisiert. In diesen Adreßbüchern finden sich zwar in der Regel nur relativ wenige Überschneidungsmerkmale zu amtlichen Daten; hierbei handelt es sich jedoch zum Teil um detaillierte berufsbezogene Angaben in Verbindung mit Regionalangaben für teilweise deutlich von der Durchschnittsbevölkerung abgegrenzte Subpopulationen. Da in derartigen Handbüchern jeweils eine Vollerfassung der entsprechenden Berufsgruppen angestrebt wird, könnte für diese Gruppen ein im Vergleich zur Durchschnittsbevölkerung erhöhtes Reidentifikationsrisiko bestehen, das einer näheren empirischen Untersuchung unterzogen werden sollte.

  Ein vergleichbares Risiko spezifisch für ausländische Mitbürger wurde für das Einwohnermelderegister festgestellt. Auch hier wird eine unter Umständen vergleichsweise kleine Subpopulation - nun durch das Merkmal "Staatsangehörigkeit" - von der Durchschnittsbevölkerung abgegrenzt. Ein wesentlicher Unterschied zu den oben erwähnten allgemein zugänglichen Informationsquellen besteht allerdings darin, daß das Einwohnermelderegister erheblichen Zugangsbeschränkungen unterliegt (vgl. Kapitel 6), die seine Nutzung für Reidentifikationszwecke erschweren.[3]

- Im Kontext der Sozialwissenschaft professionsgemäß zur Verfügung stehende Daten, die im wesentlichen aus eigenen Erhebungen stammen. Im Unterschied zu öffentlich zugänglichen Informationsquellen handelt es sich bei sozialwissenschaftlichen Erhebungen vorwiegend um Stichproben. Der Stichprobenumfang ist meistens relativ klein, kann aber in

[3] Entsprechend schwierig würde sich auch die Verwendung des Melderegisters für die empirische Überprüfung von Reidentifikationsrisiken gestalten. Die von dieser Datenquelle ausgehende spezifische Gefährdungssituation wird daher später einer argumentativen Aufwandsanalyse unterzogen. Ebenfalls in die argumentative Analyse einbezogen, werden die Informationsquellen "Adressenhändler" sowie sonstige öffentliche und private Datenquellen.

Einzelfällen auch einige tausend Personen umfassen. Von den Teilnehmern dieser Erhebungen wird z.T. ein sehr umfangreicher Merkmalskatalog erfaßt. Die Stichprobeneigenschaft stellt zwar eine generelle Schranke für Reidentifikationsversuche dar. Ferner verfügen gerade bei umfangreichen Datenfiles die Wissenschaftler in der Regel nicht über die Adressen, da die Erhebungen durch spezialisierte Umfrageinstitute durchgeführt werden und diese die Daten ohne Adressen an die Wissenschaftler weitergeben. Dennoch liegt die Annahme nahe, daß beim Vorliegen einer eigenen, personenbezogenen Erhebung der Informationsgehalt der Überschneidungsmerkmale so hoch sein könnte, daß dem Versuch, gezielt einzelne Personen zu reidentifizieren, Erfolgsschancen zugebilligt werden können.

Auf der Basis dieser Befunde wurden die fünf hypothetischen Angriffsszenarien entwickelt, die als repräsentativ gelten können für verschiedene - sowohl wissenschaftliche als auch wissenschaftsfremde - Angriffsmotive und Deanonymisierungsstrategien.

Im einzelnen untersucht werden zwei beruflich motivierte Angriffsszenarien:

- Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe aus dem Mikrozensus;

- Nutzung des Mikrozensus für die Erweiterung der Informationsbasis einer eigenen Untersuchung, analysiert am Beispiel einer Studie über Wissenschaftler;

und drei nicht-wissenschaftlich motivierte Angriffsszenarien:

- Erweiterung der vorhandenen Information über einen Bekannten zur Befriedigung der persönlichen Neugier;

- Gewinnung von Informationen über eine Persönlichkeit des öffentlichen Lebens mit dem Ziel der anonymen Publikation;

- Gewinnung ökonomisch verwertbarer Informationen aus dem Mikrozensus mit dem Ziel des Verkaufs an einen Adressenhändler.

Im folgenden soll ein Überblick über die Datenbasis, Angriffsstrategien und -motive der einzelnen Angriffsszenarien gegeben werden. Eine zusammenfassende Darstellung dieser Randbedingungen gibt Übersicht 8.1.

8.1.1 Angriffsszenario 1: Erweiterung der Informationsbasis einer wissenschaftlichen Untersuchung

In diesem Angriffsszenario wird eine Situation angenommen, in welcher Wissenschaftler im Rahmen eines Forschungsprojekts eine Untersuchung über die soziale, ökonomische und berufliche Situation einer spezifischen Berufsgruppe (hier: "Gelehrte") durchführen. Zu diesem Zweck werden Daten aus Kürschners Deutschem Gelehrtenkalender (Identifikationsfile) aufbereitet. Hypothetisch wird unterstellt, daß sich erst nach Aufbereitung dieser Datenquelle zeigt, daß die enthaltenen Informationen nicht ausreichen, um die Fragestellung vollständig zu bearbeiten. Die Forscher benötigen daher zusätzliche Daten über die soziale und ökonomische Situation der betrachteten Berufsgruppe. Es wird daher angestrebt, unter Heranziehung eines Mikrozensus Zusatzanalysen durchzuführen. Bei einer näheren Betrachtung der beiden Datenquellen (aufbereitetes Handbuch und Mikrozensus) wird deutlich, daß die Fragestellung des Forschungsprojektes dann am einfachsten zu bearbeiten wäre, wenn es gelänge, die Datensätze des Gelehrtenkalenders auf individueller Ebene mit den Datensätzen des Mikrozensus zu verknüpfen und die Analysen auf der so erweiterten Informationsbasis durchzuführen.

Wie erwähnt handelt es sich hierbei um ein hypothetisches - für Analysezwecke konstruiertes - Angriffsmotiv, denn eine derartige Verknüpfung von Individualdaten erfordert nicht die Herstellung eines Personenbezuges und damit auch keine Reidentifikation von Einzeldatensätzen, sondern nur deren Matching. Dabei könnte ihre Anonymität erhalten bleiben. Dennoch wird in diesem Szenario nun unterstellt, daß die Forscher mangels Sensibilität gegenüber Datenschutzaspekten den Versuch unternehmen, über einen *Massenfischzug* möglichst viele der sie interessierenden Datensätze aus dem *Mikrozensus* (Mikrodatenfile) zu deanonymisieren.

Durch die allgemeine Zugänglichkeit sowie die Vollständigkeit derartiger berufsgruppenspezifischer Handbücher handelt es sich - abgesehen von der Motivlage - um ein realistisches und riskantes Angriffsszenario. Ein besonders hohes Gefährdungspotential ergibt sich nach den Ausführungen in Kapitel 4 aus der Anwendung eines Massenfischzugs in einer Situation, in der das Identifikationsfile nahezu einer Totalerhebung der Population gleichkommt. Nach der Analyse des Zusatzwissens muß in diesem Fall auch mit einem hohen Auflösungsgrad der Überschneidungsmerkmale gerechnet werden. Deshalb wird diese spezifische Angriffskonstellation später einer *empirischen Überprüfung* sowohl mit einer *einfachen Abgleichtechnik* als auch mit der *diskriminanzanalytischen Reidentifikationstechnik* unterzogen.

8.1.2 Angriffsszenario 2: Erweiterung der Information über einen Bekannten zur Befriedigung der persönlichen Neugier

In diesem Angriffsszenario wird hypothetisch unterstellt, daß ein Wissenschaftler aus persönlicher Neugier Informationen über einen Arbeitskollegen, Bekannten oder Nachbarn gewinnen will. Zu diesem Zweck versucht, er über eine *Einzelsuche* diese Person in einem *Mikrozensus* (Mikrodatenfile) zu deanonymisieren.

Dieses oder ähnlich formulierte Motive werden in der Diskussion über die Notwendigkeit von Deanonymisierungsmaßnahmen häufig aufgegriffen. Letztlich können als hypothetische Ausgangsmotive für derartige Deanonymisierungsversuche, wie in Kapitel 5 gezeigt wurde, alle bekannten "menschlichen" Schwächen angeführt werden. Derartige Angriffsmotive können zwar naheliegend sein, allerdings werden sie in der Regel keine zu starken Antriebskräfte hinsichtlich der Investionen von Zeit-, Kosten- und Arbeitsaufwand freisetzen, da sich alternative (insbesondere legale) Wege der Informationsbeschaffung (beispielsweise durch direkte Befragung des Betroffenen) oft unmittelbar anbieten.

Ebenso steht jedoch Alltags- bzw. Anschauungswissen über bestimmte Personen in aller Regel ohne größeren Aufwand zur Verfügung und könnte daher für einen solchen Reidentifikationsversuch problemlos eingesetzt werden. Aus diesem Grund soll das hier dargestellte Subszenario ebenfalls einer *empirischen Überprüfung* mit einer *Abgleichtechnik* unterzogen werden, wobei es unter folgenden Bedingungen überprüft wird:

- Es steht kein 'response knowledge' zur Verfügung, d.h. es ist unbekannt, ob die gesuchte Person an der Mikrozensuserhebung teilgenommen hat;

- es steht 'response knowledge' zur Verfügung, d.h es ist bekannt, daß die gesuchte Person in der Stichprobe enthalten ist.

Als Identifikationsfile dient eine sozialwissenschaftliche Erhebung, die eine Vielzahl von Informationen enthält, die auch als Alltags- bzw. Anschauungswissen über einen Bekannten oder Arbeitskollegen vorliegen könnten. Auf diese Weise kann zugleich auch ein von sozialwissenschaftlichen Stichproben möglicherweise ausgehendes Gefährdungspotential näher betrachtet werden.

8.1.3 Angriffsszenario 3: Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe

Nach den Ausführungen zum Zusatzwissen muß in bezug auf das Melderegister ein erhöhtes Reidentifikationspotential für ausländische Mitbürger angenommen werden. Daher wird im dritten Angriffsszenario hypothetisch unterstellt, daß ein Wissenschaftler auf der Basis des *Einwohnermelderegisters* (Identifikationsfile) versucht, möglichst viele Einzeldatensätze von ausländischen Mitbürgern im Mikrozensus (Mikrodatenfile) zu deanonymisieren. Das Ziel dieses *Massenfischzugs* besteht darin, die reidentifizierten Datensätze als Auswahlgrundlage für eine eigene Ausländererhebung zu verwenden.

Eine solche Vorgehensweise ist - wie den Ausführungen in Kapitel 5 entnommen werden kann - allerdings höchstens bei einem Methodenvergleich zwischen Mikrozensus und der Ausländerstichprobe als angestrebtem Forschungsziel sinnvoll, oder wenn durch die deanonymisierten Datensätze eine Verbesserung des Stichprobenplans möglich ist. Dies könnte etwa dann der Fall sein, wenn die Merkmale des Mikrozensus eng mit jenen Merkmalen korreliert sind, die der Wissenschaftler bei seiner Stichprobe erheben will, da dann eine Schichtung der Auswahlgrundlage auf der Basis der Mikrozensus-Merkmale die Zuverlässigkeit der auf der Stichprobe beruhenden Schätzung unter Umständen erhöhen könnte.

Da keine realistische Aussicht besteht, ein entsprechendes Reidentifikationsexperiment empirisch auf der Basis des Einwohnermelderegisters durchführen zu können, wird zu diesem Szenario eine rein argumentative Kosten-Nutzen-Analyse in Abschnitt 10.1 erfolgen.

8.1.4 Angriffsszenario 4: Gewinnung von Informationen über eine Persönlichkeit des öffentlichen Lebens

Für dieses Szenario wird hypothetisch unterstellt, daß ein Wissenschaftler versucht, aus einem *Mikrozensus* (Mikrodatenfile), Datensätze von Persönlichkeiten des öffentlichen Lebens herauszufischen, um für diese Datensätze nachträglich einen Personenbezug über *öffentlich zugängliche Informationsquellen* (Identifikationsfile) herzustellen.

Einer solchen Vorgehensweise können - wie in Kapitel 5 gezeigt wurde - eine Reihe von Motiven zugrunde liegen. Beispielsweise könnte ein Forscher daran

interessiert sein, die Machbarkeit einer Reidentifikation prinzipiell nachzuweisen und damit das Ziel verfolgen, die Anonymisierungsmaßnahmen der amtlichen Statistik zu diskreditieren. Für das hier zu untersuchende Subszenario wird allerdings angenommen, daß der Wissenschaftler daran interessiert ist, die gewonnenen Erkenntnisse über eine prominente Person anonym zu veröffentlichen, um hieraus in irgendeiner Form Nutzen zu ziehen.

Als Reidentifikationsstrategie wird der sogenannte *Einzelfischzug* angenommen, bei welchem das Mikrodatenfile zunächst nach "interessanten" Datensätzen durchkämmt wird. Erst im Anschluß daran, wird spezifisch für die interessierenden Datensätze nach Zusatzwissen gesucht, welches eine Reidentifikation ermöglicht. Unter dieser Bedingung verliert die Stichprobeneigenschaft des Mikrodatenfile ihre Schutzwirkung, da die Wahrscheinlichkeit, daß der gesuchte Fall im Mikrodatenfile ist, mit eins angesetzt werden muß. Unter den in Abschnitt 4.1.2 erläuterten Bedingungen könnte daher unter Umständen ein erhöhtes Reidentifikationsrisiko vorliegen.

Eine empirische Überprüfung dieses Szenarios erscheint allerdings nicht realisierbar; daher wird es - unter Heranziehung der Ergebnisse des Gelehrtenszenarios, bei welchem die Erfolgschancen des Massenfischzugs empirisch überprüft werden, in Abschnitt 10.2 einer argumentativen Kosten-Nutzen-Analyse unterzogen.

8.1.5 Angriffsszenario 5: Gewinnung ökonomisch verwertbarer Informationen

In der Diskussion über Datenschutzrisiken wird gelegentlich die potentielle Möglichkeit einer Verwendung von deanonymisierten Mikrodaten für gewerbliche Zwecke angesprochen. In dem hier zu untersuchenden Szenario wird daher hypothetisch unterstellt, daß ein Wissenschaftler versucht, Mikrodaten zu deanonymisieren, um die gewonnenen Informationen an einen Adressenhändler weiterzugeben.

Grundsätzlich kann hier als Deanonymisierungsstrategie sowohl die *Massenidentifikation* als auch der *Massenfischzug* in Frage kommen. Am plausibelsten kann man sich dieses Szenario in einer Weise vorstellen, daß Adressenhändler und Wissenschaftler in dem Sinn zusammenarbeiten, daß ein Adressenhändler einem Wissenschaftler eigenes Datenmaterial (als Identifikationsfile) zur Verfügung stellt und dieser gegen Entgeld versucht, z.B. über bestimmte Berufs-

gruppen Informationen aus dem *Mikrozensus* (Mikrodatenfile) zu gewinnen. Ein Adressenhändler könnte sich durch ein derartiges Vorgehen möglicherweise eine bessere Durchführung einer Marketingaktion seiner Kunden versprechen. Beispielsweise verfügen Adressenhändler über sogenannte "Zielgruppen-Pools", etwa zu "Entscheidungsträgern der Wirtschaft" und sonstigen Personengruppen, die mit Informationen aus dem Mikrozensus erweitert werden könnten. Möglich wäre auch, daß z.B. beim Aufbau derartiger Pools erwartet wird, unter Vorgabe bestimmter Ausprägungen von Mikrozensus-Merkmalen, wie beispielsweise höchste Einkommensklasse, durch einen Massenfischzug zusätzliche wichtige Informationen zu gewinnen. Eine solche Konstellation soll in Abschnitt 10.3 einer detaillierten argumentativen Analyse unterzogen werden.

Übersicht 8.1: Randbedingungen der im einzelnen untersuchten Szenarien

Szenario:	Nutzung des Mikrozensus für die Erweiterung der Informationsbasis einer eigenen Untersuchung	Erweiterung der vorhandenen Informationen über einen Bekannten zur Befriedigung persönlicher Neugier	Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe aus dem Mikrozensus	Gewinnung von Informationen über eine Persönlichkeit des öffentlichen Lebens mit dem Ziel der anonymen Publikation	Gewinnung ökonomisch verwertbarer Informationen aus dem MZ mit dem Ziel des Verkaufs an Adressenhändler
Analyse erfolgt:	empirisch	empirisch	argumentativ	argumentativ	argumentativ
Angriffsmotivation	beruflich	außerberuflich	beruflich	außerberuflich	außerberuflich
Angriffsstrategie	Massenfischzug	Einzelsuche	Massensuche/ Massenfischzug	Einzelfischzug	Massensuche/ Massenfischzug
Mikrodatenfile	Mikrozensus NRW	Mikrozensus NRW	Mikrozensus	Mikrozensus	Mikrozensus
Identifikationsfile	Kürschners Gelehrtenkalender (bezogen auf NRW)	sozialwissenschaftliche Erhebung (bezogen auf NRW)	Melderegister	öffentlich zugängliche Daten	Datenmaterial v. Adressenhändlern
Anzahl Überschneidungsmerkmale*)	10	35	8	10/12	5

*) Eine detaillierte Auflistung der einzelnen Überschneidungsmerkmale befindet sich bei der Darstellung der einzelnen Szenarien.

8.1.6 Zusammenfassung

Zusammenfassend sind die für die einzelnen Szenarien unterstellten Annahmen über Angriffsmotivation, Angriffsstrategie sowie die verwandten Mikrodatenfiles und Identifikationsfiles in Übersicht 8.1 aufgeführt. Wie bei der Ausarbeitung der einzelnen Szenarien angeklungen ist, sind hier durchaus noch Variationen vorstellbar. Die Betrachtung aller denkbar möglichen Szenarien ist jedoch nicht nur unökonomisch, sondern auch unnötig. Denn für jedes der ausgewählten Szenarien wurde angestrebt, eine nach den bisherigen Erkenntnissen hochriskante Angriffskonstellation abzudecken. Damit stellen die fünf Subszenarien als "worst cases" jeweils die Eckpunkte eines Gefährdungsrahmens bei Übermittlung von amtlichen Mikrodaten an die Wissenschaft dar. Die große Zahl realistischer Angriffssituationen dürfte daher in den Aufwands- und Ertragsbedingungen einem dieser stellvertretend ausgewählten Szenarien nahe kommen. Ausgehend von den Ergebnissen der Überprüfungen können daher allgemeinere Aussagen für Reidentifikationsrisiken bezogen auf den Wissenschaftsbereich getroffen werden und hierauf aufbauend Empfehlungen für Maßnahmen, die die faktische Anonymität von Mikrodaten sicherstellen, entwickelt werden.

8.2 Datenbasis der empirischen Zuordnungsexperimente

Für zwei der fünf Angriffsszenarien sollen - wie bereits erwähnt - Experimente auf einer realen Datenbasis durchgeführt werden. Die Nutzung von Realdaten für Reidentifikationsexperimente stößt auf ein prinzipielles Problem: Nach den gesetzlichen Vorschriften ist jede Zusammenführung von Angaben der amtlichen Statistik mit anderen Informationen mit dem Ziel, einen Personenbezug herzustellen, untersagt. Bevor daher die Datenbasis für die Experimente beschrieben wird, soll zuvor ein wichtiges Detail des Untersuchungsdesigns kurz erläutert werden, mit dem eine datenschutzgerechte Durchführung der Deanonymisierungsexperimente versucht wurde.

8.2.1 Doppelt-Blind-Verfahren

In den allgemeinen Ausführungen über die Bestimmungsfaktoren eines Reidentifikationsrisikos sowie bei der Darstellung unterschiedlicher leistungsfähiger Reidentifikationstechniken wurde deutlich, daß unter empirischen Bedingungen

eine eindeutige Zuordnung von identischen oder ähnlichen Datensätzen aus Mikrodaten- und Identifikationsfile nicht gleichbedeutend mit einer Reidentifikation sein muß. Konkrete Aussagen über die vermuteten Einflüsse unterschiedlichster Art im Hinblick auf die Erfolgswahrscheinlichkeit von Reidentifikationsversuchen erfordern jedoch, daß das Verhältnis von korrekten, falschen und Nicht-Zuordnungen ebenso bekannt ist wie das Gesamtpotential der hypothetisch möglichen korrekten Zuordnungen. Dies macht die strikte Kontrolle der empirischen Ergebnisse notwendig. Bei den in der Vergangenheit durchgeführten Experimenten von anderen Forschergruppen war mit ganz oder teilweise synthetisch generierten Daten gearbeitet worden. Dabei war die Kontrolle der Ergebnisse einfach gewesen, da bekannt war, welche Datensätze in dem Identifikations- und dem Mikrodatenfile jeweils von einer einzigen wirklichen oder hypothetischen Person abstammen.

Im vorliegenden Fall, wo ausschließlich mit empirischen Daten gearbeitet wird, liegt eine problematischere Situation vor. Bei den Experimenten muß stets strikt auf die Gewährleistung der Anonymität der amtlichen Mikrodaten geachtet werden. Der Forscher darf nicht selbst zum "Täter" werden, d.h. eine verbotene Deanonymisierung vornehmen. Aus diesem Grund müssen auch die verwendeten Identifikationsfiles ausschließlich aus anonymisierten Daten bestehen.

Wen nun im Verlauf der Experimente zwei anonyme Datensätze zugeordnet werden, ist unbekannt, ob dieses Matching korrekt ist oder nicht. Um dennoch eine Überprüfung zu ermöglichen, wurde eine spezielle Prozedur entwickelt, das sogenannte *Doppelt-Blind Verfahren*.[4] Es gewährleistet, daß zu keiner Zeit eine verbotene Verknüpfung von Daten und Personen hergestellt wird.

Eine Voraussetzung zur Anwendung der Prozedur ist die strikte *personelle* und *institutionelle Trennung* zwischen Datenlieferanten bzw. Adressenbesitzern und den mit der Durchführung der Experimente betrauten Forschern. Außerdem wird eine dritte Person eingeschaltet, ein sogenannter *Datentreuhänder*.[5] Für die Zuordnungsexperimente werden die Datenbestände - formal anonymisiert - mit durchnumerierten Datensätzen zur Verfügung gestellt. Die Forscher geben nur die Nummern jener Datensätze, die aufgrund der Experimente als von ein und derselben Person stammend ermittelt wurden, an den Treuhänder weiter.

4 Über ähnliche Prozeduren bei Datenübertragungen zwischen Administrationen berichten Litz/Litowatz (1986).

5 Vgl. u.a. auch Dittrich/Schlörer (1987) zum allgemeinen Kontext von Treuhändern.

Zugleich erhält der Treuhänder von den Adressenbesitzern sukzessive Anschriftenbestandteile (Postleitzahlen, Straßen, Hausnummern, Namen mit der jeweiligen Nummer im Datenfile), so daß ihm in einer letzten Stufe schließlich nur die Namen und Nummern für jene Personen bekannt sind, die sowohl in dem jeweiligen Zusatzwissen als auch in dem entsprechenden Mikrozensus gemeinsam enthalten sind.[6] Über die Personen selbst erhält er keinerlei weitere Informationen. Anhand des Resultats dieses Anschriftenabgleichs teilt der Treuhänder den Forschern nach Beendigung der Reidentifikationsexperimente mit, wie viele Zuordnungen maximal möglich gewesen wären, wie viele der vorgenommenen Zuordnungen korrekt und wie viele falsch waren.

Der Treuhänder kann also durch den Adressenvergleich feststellen, welche Nummern der Datensätze des Mikrodatenfile und des Identifikationsfile von der gleichen Person stammen. Er ist jedoch gegenüber sämtlichen Informationen in den Datensätzen blind. Der Forscher dagegen kennt die Information in den Datensätzen, ist jedoch blind im Hinblick auf die zu den Datensätzen gehörigen Namen und Adressen.

Die Realisierung eines solchen Doppelt-Blind-Verfahrens ist mit einem erheblichen organisatorischen Aufwand verbunden, da für die jeweiligen Einzelschritte des Anschriftenvergleichs sowie für die spätere Kontrolle der Ergebnisse eine enge Zusammenarbeit zwischen dem Treuhänder und den Adressenbesitzern, den Forschern und den Datenbesitzern sowie den Forschern und dem Treuhänder erforderlich ist. Zugleich muß diese Zusammenarbeit so gestaltet werden, daß die jeweiligen Zuständigkeiten und Informationsabgrenzungen exakt eingehalten und nicht vermischt werden. Die Notwendigkeit, die hierbei anfallenden Interaktionen in einem überschaubaren und vor allem koordinierbaren und kostenmäßig bewältigbaren Rahmen zu halten, machte die Nutzung eines bundesweiten Mikrozensus oder von Mikrozensen unterschiedlicher Bundesländer unmöglich. Es war daher eine landesspezifische Auswahl zu treffen. Glücklicherweise konnte das Landesamt für Statistik und Datenverarbeitung Nordrhein-Westfalen für die Mitwirkung an diesem Vorhaben gewonnen werden. Mit Daten aus dem Mikrozensus dieses bevölkerungsreichsten Landes war ein Mikrodatenfile gegeben, auf dessen

---

6 Da es sich hierbei um eine äußerst aufwendige Vorgehensweise handelt, wurde sie nur für die sozialwissenschaftliche Stichprobe verwandt. Im Falle des Gelehrtenkalenders sind die personenbezogenen Angaben jedoch jederzeit öffentlich zugänglich, die Anschriften wurden dem Treuhänder daher in einem Block zur Verfügung gestellt.

Basis Aussagen für einen möglichst großen Teil der Gesamtbevölkerung der Bundesrepublik Deutschland möglich sind.

Wie sich die Datenbasis im Hinblick auf die Identifikationsfiles im einzelnen gestaltete, wird in den folgenden Abschnitten dargestellt.

8.2.2 Datenbasis von Angriffsszenario 1: Erweiterung der Informationsbasis einer wissenschaftlichen Untersuchung

Grundlage der hier zu untersuchenden Angriffssituation ist das in Abschnitt 8.1.1 dargestellte Subszenario 1, in welchem hypothetisch unterstellt wird, daß Wissenschaftler im Rahmen eines Forschungsvorhabens den Versuch unternehmen, auf der Basis einer öffentlich zugänglichen Informationsquelle möglichst viele beliebige Fälle eines Mikrozensus zu deanonymisieren (Massenfischzug), um eine eigene Datenbasis für weitergehende Analysen zu ergänzen. Für die empirische Umsetzung dieses Szenarios dient als Mikrodatenfile der Mikrozensus von Nordrhein-Westfalen 1987, als Zusatzwissen bzw. Identifikationsfile - repräsentativ für öffentlich zugängliche Informationsquellen - Kürschners Deutscher Gelehrtenkalender 1987.

8.2.2.1 Beschreibung der Datenbasis

Kürschners Deutscher Gelehrtenkalender bietet sich - stellvertretend für eine Reihe ähnlicher Handbücher - für die empirische Überprüfung des Reidentifikationsrisikos spezifischer Subpopulationen im Mikrozensus insofern an, als er bestimmte Charakteristika aufweist, die in der Datenschutzdiskussion als besonders riskant eingestuft werden.

Die Annahme, daß von Handbüchern über spezifische Berufsgruppen ein erhöhtes Gefährdungspotential für amtliche Daten ausgehen könnte, beruht weniger auf der Anzahl der hier in der Regel zur Verfügung stehenden Überschneidungsmerkmale. Diese ist vergleichsweise gering, wie die vorangegangene allgemeine Analyse des Zusatzwissens gezeigt hat. Es wird vielmehr vermutet, daß sich aus der Zusammensetzung der Überschneidungsmerkmale und ihres Informationsgehalts ein erhöhtes Reidentifikationsrisiko ergibt:

- Die erfaßte Subpopulation wird in der Regel durch ein sehr präzises Merkmal, die Berufsangabe, gekennzeichnet. Berufsgruppen, für die Handbücher bestehen, umfassen in der Regel nur vergleichsweise kleine Zahlen von Berufsangehörigen. Sie nehmen in der Gesellschaft oft auch eine herausgehobene Position ein. Die Handbücher existieren vorwiegend für akademische Berufe und auch hier nur für ausgewählte Berufsgruppen. In der Mehrzahl der anderen Datenquellen, die als Zusatzwissen in Betracht gezogen werden könnten, ist eine derartige Eingrenzung einer spezifischen Population nicht möglich;

- Die Handbücher enthalten in der Regel vollständige Anschriften der befragten Personen (ohne diese könnten die Handbücher keine Verwendung als Zusatzwissen finden);

- Die Handbücher enthalten demographische Variablen. Diese variieren zwar von Handbuch zu Handbuch, in den überwiegenden Fällen ist jedoch immer das Geburtsdatum der befragten Person und damit ein hoch differenzierendes Überschneidungsmerkmal enthalten.

Während die eindeutige Abgrenzung der spezifischen Berufsgruppe es ermöglicht, bereits eine in der Regel kleine und markante Subpopulation in einem Mikrodatenfile zu isolieren, bewirken die zumeist vorhandenen Regionalangaben insbesondere in Verbindung mit dem Merkmal Geburtsjahr sowie einigen wenigen sonstigen Merkmalen eine starke Differenzierung zwischen den einzelnen Fällen innerhalb der selektierten Gruppe. Der Gesamtzahl der zur Verfügung stehenden Überschneidungsmerkmale kommt unter diesen Bedingungen daher eine eher untergeordnete Rolle zu. Der hier beschriebene Effekt ist umso stärker, je kleiner die betrachtete Subpopulation ist, da sich in diesem Fall die Zahl der potentiell möglichen statistischen Doppelgänger reduziert (vgl. Kapitel 4).

Für Kürschners Gelehrtenkalender kann aufgrund dieser Überlegungen ein im Vergleich zu ähnlichen Handbüchern insofern erhöhtes Gefährdungspotential unterstellt werden, als die spezifische Berufsgruppe der Wissenschaftler, im Vergleich zu anderen Berufsgruppen, wie beispielsweise der Juristen oder Ärzte, eine relativ kleine Population darstellt. Durch die Fachzugehörigkeit kann sie zugleich noch in eine große Zahl von Subgruppen aufgegliedert werden.

Ein weiteres wichtiges Kriterium, das den Gelehrtenkalender als riskantes Zusatzwissen kennzeichnet, ergibt sich daraus, daß es sich hierbei um das umfassendste Verzeichnis von Wissenschaftlern und Forschern im deutschsprachigen Raum handelt. Die Ausgabe von 1987 umfaßt insgesamt etwa 45 000 Gelehrte, die sich auf folgende fünf Berufsgruppen verteilen:

a) Professoren mit einer Tätigkeit in der wissenschaftlichen Lehre und Forschung an einer Hochschule oder einer Institution gleichen Ranges (z.B. Max-Planck-Institut).

b) Alle habilitierten Wissenschaftler, soweit nicht unter a) genannt.

c) Wissenschaftliche Räte.

d) Wissenschaftliche Mitarbeiter an Forschungseinrichtungen außerhalb der Hochschule (z.B. Einrichtungen des Bundes, der Länder, der Industrie, Stiftungen), die wissenschaftlich publiziert haben.

e) Angehörige akademischer Berufe, die innerhalb dieses Berufes durch wissenschaftliche Publikationen besonders hervorgetreten sind.

Da die Erfassung der Gruppen a) bis c) nach Auskunft der Redaktion des Handbuches über die Hochschulverzeichnisse erfolgt, dürften die im Hochschulbereich tätigen Wissenschaftler - soweit sie der Zielpopulation angehören - nahezu vollständig enthalten sein. Bei nicht im Hochschulbereich tätigen Forschern können hingegen Lücken auftreten. Gemessen an dem Kriterium 'Vollständigkeit des Zusatzwissens in bezug auf die Grundgesamtheit der betrachteten Zielpopulation' beinhaltet der Gelehrtenkalender, nach den Überlegungen in Abschnitt 4.1.1, ein sehr hohes Reidentifikationspotential.

Ein weiterer Aspekt, der für ein von dem Gelehrtenkalender ausgehendes erhöhtes Gefährdungspotential spricht, ergibt sich daraus, daß dieses Datenmaterial nur relativ wenig fehlerbehaftet sein dürfte. Sowohl die Vorgehensweise der Handbuchredaktion als auch die Spezifika der befragten Subpopulation stellen diesbezüglich günstige Randbedingungen dar. Die Angaben der Personen, die für die Veröffentlichung im Gelehrtenkalender Verwendung finden sollen, werden von dem Verlag über einen von der betreffenden Person selbst auszufüllenden Fragebogen erhoben bzw. auch den jeweiligen Hochschulverzeichnissen entnommen. Unabhängig von der Informationsquelle wird den erfaßten Personen die Möglichkeit geboten, die aufgenommenen Angaben nochmals zu überprüfen, bevor sie endgültig in Druck gehen. Das individuelle Interesse der Befragten, daß publizierte personen- und berufsspezifische Informationen, die sowohl für den Kollegenkreis als auch einer breiten Öffentlichkeit zugänglich sind, nicht fehlerbehaftet sind, ist als relativ hoch anzusehen. Im Vergleich zu Erhebungen, bei welchen diese Kontrollmechanismen durch die Befragten nicht gegeben sind, kann angenommen werden, daß die in dem Gelehrtenkalender veröffentlichten Angaben einen höheren Wahrheitsgehalt aufweisen.

8.2.2.2 Datenaufbereitung

Die Nutzung des Gelehrtenkalenders als Zusatzwissen erfordert umfangreiche Vorarbeiten, da entsprechend der Zielsetzung dieses Handbuchs die Angaben der Befragten nur im Klartext und in gedruckter Form vorliegen. Für die Durchführung der Zuordnungsexperimente ist es jedoch unabdingbar, daß die Überschneidungsmerkmale entsprechend der Mikrozensuskonvention verschlüsselt waren und maschinenlesbar zur Verfügung stehen. Die hierfür notwendige Umsetzung der Klartextangaben in die Verschlüsselung der amtlichen Statistik und die Aufnahme dieser Angaben auf Datenträger erfolgt im *Statistischen Bundesamt*. Mit dieser Vorgehensweise wird zum einen eine dem Doppelt-Blind-Verfahren entsprechende personelle und institutionelle Trennung zwischen der Aufbereitung der personenbezogenen Daten und der Durchführung der Experimente gewährleistet. Zum anderen kann durch die Aufbereitung des Datenmaterials durch Personal, das mit der Mikrozensusverkodung vertraut ist, hohe Vergleichbarkeit der Verschlüsselung des Gelehrtenkalenders mit der Mikrozensusverkodung gewährleistet werden.

Neben bibliographischen Daten zu wissenschaftlichen Publikationen stehen im Gelehrtenkalender folgende biographische Daten der Befragten zur Verfügung:

> Name und Vorname; Geburtsort, -jahr, -monat, -tag; Postanschrift und Telephonanschluß; Akademische Grade, Titel, jetzige berufliche Stellung; wissenschaftliche Laufbahn; wissenschaftliche Auszeichnungen und Preise, Mitgliedschaften in wissenschaftlichen Akademien und entsprechenden Gesellschaften, Hauptarbeitsgebiete.

Im Rahmen der empirischen Überprüfung sind dabei nur jene Angaben von Relevanz, die sich, bezogen auf den Mikrozensus, direkt als Überschneidungsmerkmale eignen oder aber eine Basis für die Rekonstruktion von Überschneidungsmerkmalen darstellen. Die Analyse dieser Angaben ergibt folgende Überschneidungsmerkmale des Gelehrtenkalenders zum Mikrozensus:

1) Gemeindegrößenklasse
2) Geschlecht
3) Geburtsjahr
4) Geburtshalbjahr
5) Wirtschaftszweig
6) Stellung im Beruf
7) Beruf

8) Überwiegend ausgeübte Tätigkeit
9) Hauptfachrichtung des letzten Hochschulabschlusses
10) Bundesraumordnungsregion[7]

Während die Ermittlung der Überschneidungsmerkmale relativ problemlos ist, zeigt sich, daß bei einer empirischen Analyse eines Reidentifikationsrisikos auf der Grundlage einer empirischen Datenbasis, anders als bei einem lediglich analytisch bestimmten oder auf der Grundlage von synthetisch generierten Daten ermittelten Reidentifikationspotential, Probleme auftreten, die analog auch bei einem realen Datenangriff zu erwarten sind.

So ist eine genaue und vollständige Umsetzung der Klartextangaben in die vorgegebene amtliche Verschlüsselung unter empirischen Bedingungen nicht in allen Fällen möglich. Es muß vielmehr mit Unschärfen, Informationsverlusten und Lücken gerechnet werden, die sich als "Fehler" in dem Datenmaterial niederschlagen und in Form von Dateninkompatibilitäten Reidentifikationsversuche entscheidend behindern können. Derartige Fehlerquellen werden zwar in einem geringeren Ausmaß auftreten, wenn die Aufbereitungsarbeiten von Experten durchgeführt werden, aber auch in diesem Fall können sie nicht vollständig vermieden werden.

Als ein erstes Problem erweist sich, daß die Angaben im Gelehrtenkalender vor allem in bezug auf das Geburtsdatum nicht immer vollständig sind. Um dennoch nahezu lückenlose Informationen für die einzelnen Personen zu erlangen, wird bei fehlenden Angaben auf zusätzliche Informationsquellen (beispielsweise Who is Who in Germany) zurückgegriffen. Trotz sorgfältiger Recherchen ist es dennoch unmöglich, alle fehlenden Werte durch weitere Informationsquellen zu ersetzen. Allerdings weisen bei acht der zehn Überschneidungsmerkmale über 99 Prozent der Befragten einen gültigen Wert auf, so daß sich die fehlenden Angaben hier kaum als Störfaktor auf die Reidentifikationssimulationen auswirken dürften. Problematischer hingegen ist, daß bei etwa einem Drittel der befragten Personen keine gültigen Angaben zum Geburtsdatum (Geburtsjahr/-monat) ermittelbar sind. Dies könnte sich - sofern gesuchte Personen davon betroffen sind - insbesondere bei Reidentifikationsversuchen durch einfache Abgleichtechniken als Störfaktor erweisen. Da beim Mikrozensus eine Antwortpflicht besteht, werden die Überschneidungsmerkmale zum Geburtsdatum in der Regel einen gültigen Wert aufweisen. Steht bei einer gesuchten Person

7 Da sich die Daten ausschließlich auf Nordrhein-Westfalen beziehen, ist implizit auch das Regionalmerkmal "Bundesland" als elftes Überschneidungsmerkmal enthalten.

einer gültigen Angabe im Mikrodatenfile jedoch ein fehlender Wert [8] im Identifikationsfile gegenüber, so können diese Datensätze mit einer einfachen Abgleichtechnik nicht zugeordnet werden. Das Kriterium Identität in den Merkmalsausprägungen ist nicht erfüllt, die Datensätze werden als nicht passend aussortiert. Ein Verzicht auf diese Überschneidungsmerkmale erscheint allerdings insofern nicht angebracht, als dem Merkmal Geburtsjahr eine stark diskriminierende Eigenschaft zukommt und immerhin 70 Prozent der Befragten hier einen gültigen Wert aufweisen.

Als ebenfalls erweist sich, daß bei einer Vielzahl der erfaßten Personen die im Gelehrtenkalender enthaltenen Angaben für die Merkmale "Beruf", "Wirtschaftszweig", "ausgeübte Tätigkeit" und "Fachrichtung des Hochschulabschlusses" nicht immer eindeutig einer spezifischen Kategorie zuordenbar sind. So ist es durchaus üblich, daß ein Hochschulprofessor im Fach Medizin zugleich auch als Arzt tätig ist. Aus den im Gelehrtenkalender enthaltenen Angaben ist es in solchen Fällen jedoch nicht immer unmittelbar ersichtlich, ob es sich hierbei um alternative Erwerbstätigkeiten oder um eine Haupt- und Nebenerwerbstätigkeit handelt. Spezifisch für die im Gelehrtenkalender enthaltene Population ist auch die Mehrdeutigkeit für das Merkmal "ausgeübte Tätigkeit". Die Aufgabenbeschreibung eines an der Hochschule tätigen Wissenschaftlers beinhaltet sowohl Forschung als auch Lehre. Die Selbsteinstufung eines Wissenschaftlers in eine dieser beiden Kategorien kann den Angaben im Gelehrtenkalender nicht entnommen werden, ist aber entscheidend dafür, wie diese Frage im Mikrozensus beantwortet wird.

Eine kategorische Festlegung, daß jeweils nur die am plausibelsten erscheinende Alternative aufgenommen wird, würde zwar die Datenaufbereitung erheblich vereinfachen, käme aber zum einen einer Einschränkung des in dem Gelehrtenkalender enthaltenen Informationspotentials gleich. Zum anderen könnte aufgrund von möglichen falschen Einschätzungen eine unkontrollierbare Fehlerquote in dem Datenmaterial auftreten, die zu einer erheblichen Verzerrung in den Ergebnissen der Simulationsrechnungen beitragen könnte. Um derartige Fehlerquellen möglichst auszuschließen, wird daher für die vier Überschneidungsmerkmale (Beruf, Wirtschaftszweig, ausgeübte Tätigkeit, Fach-

---

8 Um keine Information zu verschenken, werden die "fehlenden" Werte ebenfalls als gültige Angaben behandelt. Denn sofern eine gesuchte Person sowohl im Mikrodatenfile als auch im Identifikationsfile einen fehlenden Wert aufweisen würde, ist - sofern die restlichen Überschneidungsmerkmale ebenfalls kompatibel abgebildet sind - eine Zuordnung möglich.

richtung) eine Verschlüsselung nach je zwei alternativen Möglichkeiten vorgenommen.

Da aus den genannten Gründen der Mikrozensus (1987) von Nordrhein-Westfalen als Mikrodatenfile für die Experimente dient, beschränkt sich die Aufnahme und Aufbereitung des Gelehrtenkalenders auf jene Personen, die aufgrund der Anschrift eindeutig dem Bundesland Nordrhein-Westfalen zugeordnet werden können. Bedingt durch dieses Selektionskriterium enthält das Identifikationsfile 7 983[9] für die Reidentifikationssimulationen verwertbare Fälle. Da der Mikrozensus eine Einprozent-Stichprobe der Gesamtbevölkerung darstellt, ist zu erwarten, daß die Daten von rund 80 Personen in beiden Datenfiles enthalten sind. Diese Zahl entspricht daher bei grober Abschätzung der maximal zu erwartenden Zahl von Fällen, die bei einem Versuch korrekt reidentifiziert werden können.

Der als Zusatzwissen genutzte Gelehrtenkalender wurde zwar 1987 herausgegeben, die Daten beziehen sich jedoch vorwiegend auf den Stand des Sommersemesters 1985. Deshalb wäre der Mikrozensus 1985 für die Simulationsrechnungen vorzuziehen gewesen. Aufgrund einer zeitlich festgelegten Löschungspflicht von Hilfsmerkmalen bei amtlichen Erhebungen würde dann jedoch kein Material für die Kontrolle der Simulationsergebnisse durch den Treuhänder zur Verfügung stehen. Eine empirische Beurteilung des Reidentifikationsrisikos wäre unter diesen Bedingungen nicht möglich. Da die vorliegenden Überschneidungsmerkmale zum Teil keinen zeitlichen Veränderungen unterliegen und die Mehrzahl der flexiblen Merkmale für die betrachtete Population als eher stabil eingeschätzt werden, werden die aufgrund der zeitlichen Differenz eventuell auftretenden Dateninkompatibilitäten als das geringere Übel in Kauf genommen. Mit entsprechenden zeitlichen Inkompatibilitäten zwischen Informationen des Zusatzwissens und Informationen im Mikrodatenfile müßte in der Regel auch bei wirklichen Reidentifikationsversuchen gerechnet werden.

Die Reidentifikationsexperimente haben ausschließlich das Ziel, ein von allgemein zugänglichen Informationsquellen ausgehendes Gefährdungspotential für den Mikrozensus zu untersuchen. Es wird also nicht - wie bei realen Deanonymisierungsversuchen - angestrebt, das Mikrodatenfile heranzuziehen,

---

9 Insgesamt enthielt Kürschners Gelehrtenkalender Angaben für 8 000 Personen aus Nordrhein-Westfalen. 17 Fälle enthielten jedoch keinerlei als Überschneidungsmerkmal verwertbare Informationen, weshalb sie nicht in das Abgleichverfahren miteinbezogen werden konnten.

um zusätzliche Informationen (sogenannte Nutzmerkmale) für die zugeordneten Fälle zu erlangen. Daher werden für das Mikrodatenfile nur jene Merkmale benötigt, die aufgrund der Auswertung des Gelehrtenkalenders als Überschneidungsmerkmale dienen sollen. Es wäre allerdings denkbar, drei der vier alternativ verschlüsselten Überschneidungsmerkmale (Beruf, Branche und ausgeübte Tätigkeit) als zusätzliche Variablen zu betrachten. Denn bei der Mikrozensuserhebung werden auch Angaben über eine zweite Erwerbstätigkeit abgefragt, worin die genannten Merkmale ebenfalls enthalten sind. Da bereits die Weitergabe der zehn unbedingt erforderlichen Überschneidungsmerkmale bei dem zuständigen Datenschutzbeauftragten aufgrund ihres hohen Informationsgehalts auf erhebliche datenschutzrechtliche Bedenken gestoßen ist, wird auf die Anforderung dieser drei Alternativvariablen für den Mikrozensus verzichtet. Dieser Verzicht ist insofern unproblematisch, als sich das Fehlen dieser Angaben nicht negativ auf die Reidentikationsexperimente auswirken dürfte: sofern eine gesuchte Person, deren Angaben im Mikrozensus im Gelehrtenkalender alternativ verschlüsselt sind, auch bei der Mikrozensuserhebung eine Erst- und Zweiterwerbstätigkeit angegeben hat, kann dies durch entsprechende Variationen der alternativen Variablen im Zusatzwissen berücksichtigt werden. Sind die Angaben kompatibel, sind auch die entsprechenden Datensätze aus Mikrodaten- und Identifikationsfile dieser Person zuordenbar.

Für die empirische Überprüfung stehen somit die in Übersicht 8.2 dargestellten Überschneidungsmerkmale zur Verfügung:

Übersicht 8.2: Überschneidungsmerkmale des Angriffsszenarios 1

Gelehrtenkalender[10] N=7983	Anzahl Merkmals-ausprägungen im Gelehrtenkalender	Mikrozensus N=169.368
Gemeindegrößenklasse	9	Gemeindegrößenklasse
Geschlecht	2	Geschlecht
Geburtsjahr	61	Geburtsjahr
Geburtsmonat	2	Geburtsmonat
Wirtschaftszweig1	15	Wirtschaftszweig
Wirtschaftszweig2	40	
Stellung im Beruf	4	Stellung im Beruf
Beruf1	26	Beruf
Beruf2	47	
Ausgeübte Tätigkeit1	5	Ausgeübte Tätigkeit
Ausgeübte Tätigkeit2	7	
Fachricht. Hochschulabschluß1	71	Fachricht.
Fachricht. Hochschulabschluß2	68	Hochschulabschluß
Bundesraumordnungsregion	16	Bundesraumordnungs-region

Die mit der Durchführung der Experimente betrauten Forscher erhalten den Mikrozensus ebenso wie den Gelehrtenkalender in einer anonymisierten Form, wobei die Einzeldatensätze für die spätere Kontrolle der Ergebnisse mit einer eindeutigen Kennziffer versehen waren. An dem Datenmaterial selbst werden keine weitergehenden Anonymisierungsvorkehrungen vorgenommen: Der Mikrozensus kann in vollem Umfang genutzt werden, d.h. es stehen insgesamt 169 368 verwertbare Einzeldatensätze zur Verfügung. Auch die zehn Überschneidungsmerkmale werden in keiner Weise modifiziert. Aufgrund dieses äußerst konstruktiven Entgegenkommens des Landesamts für Datenverarbeitung und Statistik Nordrhein-Westfalen steht damit für die Experimente ein Datenmaterial zur Verfügung, das für das zu untersuchende Subszenario ein höchstmögliches Risiko beinhaltete.

[10] Da hier die spezifische Subpopulation der Gelehrten betrachtet wird, die in aller Regel eindeutig durch Beruf und Branche gekennzeichnet ist, bedeutet dies für die Anzahl der Ausprägungen von Wirtschaftszweig 1 und Beruf 1, daß knapp 99 Prozent der erfaßten Fälle hier jeweils in eine Kategorie fallen. Die verbleibenden Kategorien werden von "Ausreißern" eingenommen und sind nur schwach besetzt. Demgegenüber erklärt sich die erhöhte Anzahl von Ausprägungen bei Wirtschaftszweig 2 und Beruf 2 damit, daß hier jeweils breiter gestreute Alternativen zu der Tätigkeit der erfaßten Personen aufgenommen wurden.

Die Resultate der auf dieser Datenbasis durchgeführten Zuordnungsexperimente werden in Abschnitt 9.1 (einfache Abgleichtechnik) sowie in Abschnitt 9.2 (diskriminanzanalytische Reidentifikationstechnik) dargestellt und erörtert.

8.2.3 Datenbasis von Angriffsszenario 2: Erweiterung der Informationen über einen Bekannten zur Befriedigung der persönlichen Neugier

Der Überprüfungsaspekt des zweiten empirisch zu untersuchenden Szenarios richtet sich auf das von sozialwissenschaftlichen Erhebungen ausgehende Gefährdungspotential für amtliche Datenbestände. Auch wenn man es hier in der Regel nur mit relativ kleinen Stichproben zu tun hat, kann der Informationsgehalt der Überschneidungsmerkmale so hoch sein, daß bei dem Versuch, gezielt einzelne Personen zu reidentifizieren, Erfolge zu erwarten sind. Hypothetisch wird für dieses Angriffsszenario unterstellt, daß ein Wissenschaftler aufgrund persönlicher Motive versucht, einen Arbeitskollegen oder Bekannten in einem Mikrodatenfile zu reidentifizieren, ohne jedoch zu wissen, ob die gesuchte Person an der amtlichen Stichprobe teilgenommen hat.

Die in der sozialwissenschaftlichen Erhebung enthaltenen Informationen können als typische Beispiele für vorhandenes Alltags- bzw. Anschauungswissen über andere Personen, beispielsweise Nachbarn, Bekannte oder Arbeitskollegen gelten. Jeder in einer solchen Erhebung enthaltenen Fälle steht für entsprechendes Wissen, das über einen Dritten vorhanden sein kann. Damit soll für mehrere tausend unterschiedliche, aber immer realistische Fälle überprüft werden, welches Risiko besteht, eine beliebige Person in einem Mikrodatenfile zu deanonymisieren. Zusätzlich kann auf diese Weise das Gefährdungspotential eingegrenzt werden, das sich aus der Verfügbarkeit von großen personenbezogenen Datenbeständen in den Sozialwissenschaften ergibt. Als Mikrodatenfile für die Experimente dient wiederum eine spezielle Datei mit Mikrozensusdaten von Nordrhein-Westfalen.

8.2.3.1 Beschreibung der Datenbasis

Eine sozialwissenschaftliche Stichprobe ist für Zuordnungsexperimente zunächst nur dann geeignet, wenn personenbezogene Informationen vor-

handen sind. Im allgemeinen werden jedoch Namen und Adressen der Befragten nach der Erhebungsphase vernichtet (vgl. Kapitel 6). Eine Ausnahme sind lediglich Wiederholungs-Befragungen (sog. Panel-Befragungen). In diesem Fall müssen die Identifikatoren aufgehoben werden, damit die Personen bei späteren Befragungs-"Wellen" wieder ausfindig gemacht werden können.

Eine weitere Schwierigkeit ist, daß nur Erhebungen mit ungewöhnlich großen Stichproben in Frage kommen. Allein in Nordrhein-Westfalen sollten mehrere tausend Befragte zur Verfügung stehen. Andernfalls wäre die Chance zu klein, überhaupt auf Personen zu stoßen, deren Daten in beiden Files enthalten sind. Weiterhin dürfen keine spezifischen Zusicherungen an die Befragten gegeben worden sein, die eine Datenübermittlung ausschließen. Dies war bei einigen Erhebungen der Fall, die ansonsten als gut geeignet eingeschätzt wurden.[11] Schließlich erwies es sich als schwierig, die für die Erhebungen verantwortlichen Institutionen für eine Kooperation zu gewinnen. Aus eingehenden Forschungen ist bekannt, daß die Antwortbereitschaft vieler Befragter sehr sensibel auf vielfältige Einflüsse reagiert (Erbslöh/Koch 1988:39ff, Rendtel 1990, Singer 1978). Es wurde befürchtet, daß diese Bereitschaft durch Mißverständnisse zerstört werden könnte, wenn öffentlich bekannt würde, daß die Daten für Reidentifikationsexperimente verwendet werden. Aus diesem Grund wurde z.B. die Verwendung von Daten einer gut geeigneten Panelerhebung verweigert.[12] Schließlich gelang es jedoch, eine geeignete Untersuchung zu finden und die Bereitschaft nicht nur des Auftraggebers und des zuständigen sozialwissenschaftlichen Instituts, sondern auch des Umfrageinstituts, das die Erhebung durchführte, zu erhalten.[13]

Die gewählte Untersuchung erfüllt alle Bedingungen, die für ein brauchbares Identifikationsfile erforderlich sind. Die Stichprobe ist hinreichend groß: Bezogen auf Nordrhein-Westfalen stehen insgesamt 2 685 Befragte zur Verfügung. Da der Mikrozensus eine Einprozent-Stichprobe der Bevölkerung darstellt, ist zu erwarten, daß rund 27 Personen in beiden Files enthalten sind.

11 Für die Verwendung einer speziellen Umfrage wurde ein Rechtsgutachten eingeholt, das negativ ausfiel.

12 So hinderlich diese Schwierigkeiten für das vorliegende Vorhaben auch waren, so zeigen sie doch, daß die Datenschutzbelange in der Sozialwissenschaft - nicht zuletzt aus eigenem Interesse - ernst genommen werden.

13 Damit den betroffenen Institutionen durch möglicherweise unsachgemäße Berichterstattung seitens Dritter kein Schaden entstehen kann, wurde vereinbart, in Abweichung von der üblichen Praxis, den Namen der Erhebung und der beteiligten Institutionen nicht zu nennen.

Allerdings ist dieser Schätzwert mit einer relativ hohen Unsicherheit belastet. Der Mikrozensus 1987 ist eine Flächenstichprobe mit "Klumpen" von ca. 20 (in größeren Städten 30) Haushalten (vgl. Herberger 1985). Die sozialwissenschaftliche Stichprobe basiert auf einem komplizierten mehrstufigen Zufallsauswahlverfahren, dessen Details hier nur insoweit interessieren, als sogenannte "Samplepoints" zufällig ausgewählt werden. Dies sind Stimmbezirke zur Bundestagswahl, die mindestens 400 wahlberechtigte Bürger enthalten. Aus diesen Stimmbezirken werden dann jeweils etwa zehn Personen bzw. Haushalte[14] zufällig für Interviews bestimmt. Je nachdem, ob sich ein oder mehrere "Klumpen" des Mikrozensus mit den gezogenen Samplepoints der sozialwissenschaftlichen Stichprobe decken, kann die Anzahl jener Personen variieren, die in beiden Samples enthalten sind. Es kann vorab nicht ausgeschlossen werden, daß diese "Schnittmenge" von Personen leer ist, d.h. daß niemand mit seinen Daten in beiden Files enthalten ist. Allerdings ist dieser Fall sehr unwahrscheinlich.

Der zeitliche Abstand zum Mikrozensus 1987 ist ein weiterer Punkt für die Eignung der sozialwissenschaftlichen Stichprobe. Der Abstand sollte möglichst klein sein. Die Erhebungsphase der verwendeten sozialwissenschaftlichen Stichprobe dauerte von August bis Dezember 1988. Der durchschnittliche zeitliche Abstand zum Mikrozensus, dessen Stichtag in den März 1987 fiel, beträgt also 20 Monate.

Bei der verwendeten Erhebung stehen eine große Zahl von Überschneidungsmerkmalen zum Mikrozensus zur Verfügung (vgl. Übersicht 8.3). Zwar wurden in der benutzten sozialwissenschaftlichen Erhebung die wegen ihrer Differenzierung wichtigen Variablen Beruf und Branche nicht erfragt. Außer dem implizit enthaltenen Merkmal Bundesland liegen auch keine Regionalinformationen vor. Neben einigen demographischen Merkmalen wird jedoch der Erwerbsstatus differenziert erfaßt. Es liegt sogar eine Biographie des Erwerbsverlaufs vor, die, anders als bei "Kürschner", eine genaue Bezugnahme auf den Erhebungszeitpunkt des Mikrozensus ermöglicht. Vor allem sind jedoch viele Merkmale zum Familien- und Haushaltskontext der Befragten enthalten. Die mit der sozialwissenschaftlichen Stichprobe durchgeführten Experimente stellen somit eine sinnvolle Ergänzung zu jenen mit den Daten des Gelehrtenkalenders dar.

---

14 Die Stichprobe ergab sich durch Zusammenführen einer separaten Haushalts- und einer Personenstichprobe. Die Haushaltsstichprobe wurde auf der Grundlage des sogenannten ADM-Mastersamples gezogen (vgl. dazu Kirschner 1984:118ff., Porst 1985:74ff.).

Bei letzteren sind vor allem der Beruf und damit in Zusammenhang stehende Variablen wichtig, es stehen jedoch keine Haushaltsmerkmale zur Verfügung.

Übersicht 8.3: Überschneidungsmerkmale der sozialwissenschaftlichen Erhebung zum Mikrozensus

Überschneidungsmerkmale	Anzahl Ausprägungen[15]	Überschneidungsmerkmale	Anzahl Ausprägungen
PERSONENSPEZIFISCHE MERKMALE		DETAILLIERTE HAUSHALTSINFORMATIONEN	
Geschlecht	2	Partner: Stellung i. Beruf	11
Geburtsjahr	38	Partner: Arbeitszeit (i.Stunden)	8
Familienstand	4	Partner: Erwerbstätigkeit	3
Schulische Ausbildung	6	Partner: Arbeitslos	3
Berufsausbildung	8	Partner: Arbeitssuche	3
Stellung im Beruf	11	Geburtsjahr Kind1	33
Erwerbstätigkeit	3	Geburtsjahr Kind2	27
Arbeitslosigkeit	3	Geburtsjahr Kind3	25
wöchentliche Arbeitszeit	7	Geburtsjahr Kind4	18
Arbeitssuche	3	Geburtsjahr Kind5	10
Art des Arbeitsvertrags	4	Geschlecht Kind1	2
Ende d. Erwerbstätigkeit	8	Geschlecht Kind2	2
persönliches Nettoeinkommen	9	Geschlecht Kind3	2
		Geschlecht Kind4	2
ALLGEMEINE HAUSHALTSINFORMATIONEN		Geschlecht Kind5	2
		Ausbildung Kind1	4
		Ausbildung Kind2	4
Zahl Kinder im Hhlt unt. 3 Jahren	3	Ausbildung Kind3	4
dto von 3 bis unter 6 Jahren	3	Ausbildung Kind4	4
dto von 6 bis unter 10 Jahren	3	Ausbildung Kind5	2
dto von 10 bis unter 15 Jahren	3	Vorwieg. Unterhalt des Haushalts	7
dto von 15 bis unter 18 Jahren	3		
dto von 18 bis unter 28 Jahren	5		
dto über 28 Jahre	2		
Haushaltsnettoeinkommen	9		
Gesamtzahl der Überschneidungmerkmale:	35		

[15] Diese Angaben beziehen sich auf die sozialwissenschaftliche Erhebung, da nur die dort auftretenden Merkmalsausprägungen für die Zuordnungen mit einer einfachen Reidentifikationsmethode relevant sind.

8.2.3.2 Datenaufbereitung

Während bei "Kürschners Deutschem Gelehrtenkalender" die Informationen im Klartext vorhanden sind und gemäß den Mikrozensuskategorien verkodet werden müssen, liegen die Daten der sozialwissenschaftlichen Stichprobe fertig verkodet in einer gegenüber dem Mikrozensus z.T. abweichenden Systematik vor. Dazu sei als Beispiel die Variable "Stellung im Beruf" herangezogen. Diese hat im Mikrozensus elf, in der sozialwissenschaftlichen Erhebung 25 Ausprägungen, die in Übersicht 8.4 gegenübergestellt sind. Abgleichtechniken, die auf Identität der Merkmale beim Vergleich zweier Fälle reagieren, setzen voraus, daß die Merkmalsausprägungen so umkodiert werden, daß zwei deckungsgleiche Skalen von Kategorien entstehen. Dies ist demnach eine Arbeit, die ein potentieller Datenangreifer stets zu leisten hat.

Die für die Experimente gewählte Lösung ist in Übersicht 8.5 dargestellt. Sie vermag nicht voll zu befriedigen. Bei Selbständigen erfolgt die Abgrenzung unter anderem durch die Anzahl der Beschäftigten. Während im Mikrozensus die Grenze zwischen zwei Kategorien gemäß dem Vorhandensein von abhängig Beschäftigten gezogen wird, faßt die sozialwissenschaftliche Erhebung Selbständige ohne und mit lediglich einem Beschäftigen in einer Kategorie zusammen. Ein anderes Problem ist die Kategorie "freie Berufe" in der sozialwissenschaftlichen Erhebung. Diese liegt quer zu den Kategorien des Mikrozensus, da in ihr sowohl Personen mit als auch ohne Beschäftigte eingestuft sind. Weitere Probleme werden bei genauerer Betrachtung der Tabelle deutlich. Hier sei nur noch darauf verwiesen, daß die sozialwissenschaftliche Erhebung Auszubildende im Gegensatz zum Mikrozensus nicht als erwerbstätig betrachtet. Die genannten Probleme könnten zum größten Teil durch eine weitere Zusammenfassung von Kategorien vermieden werden. Dies würde jedoch auf Kosten des Informationsgehalts der Variablen gehen.

Übersicht 8.4: Stellung im Beruf ("Tätigkeit wird ausgeübt als...")

Mikrozensus		Sozialwissenschaftliche Erhebung	
Nr.	Bedeutung	Nr.	Bedeutung
0	Selbständiger ohne Beschäftigte	Selbstständige	
1	Selbständiger mit Beschäftigte	10	Landwirt
2	Mithelfender Fam.angehöriger	11	Akademischer freier Beruf
3	Beamter, Richter	Selbstständige in Handel, Gewerbe und Industrie	
4	Angestellter	20	Allein oder mit 1 Mitarbeiter
5	Arbeiter, Heimarbeiter	21	Mit bis zu 9 Mitarbeitern
6	kaufm./techn. Auszubildender	22	Mit 10 und mehr Mitarbeitern
7	gewerbl. Auszubildender	30	Mithelfender Fam.angehöriger
K	Zeit-/Berufssoldat	Beamte	
L	Grundwehrdienstleistender	40	Einfacher Dienst
M	Zivildienstleistender	41	Mittlerer Dienst
		42	Gehobener Dienst
		43	Höherer Dienst, Richter
		Angestellte	
		50	Industrie- und Werkmeister im Angestellten verhältnis
		51	Angestellter mit einfacher Tätigkeit
		52	Angestellter mit schwieriger Tätigkeit
		53	Angestellter mit verantwortungsvoller Tätigkeit
		54	Angestellter mit Führungsaufgaben
		Arbeiter	
		60	Ungelernte Arbeiter
		61	Angelernte Arbeiter
		62	Gelernte und Facharbeiter
		63	Vorarbeiter
		64	Meister
		Sonstige Berufstätige	
		74	Praktikanten/Volontäre
		76	Haushalts-/diakonisches Jahr
		77	Wehrdienst
		78	Zivildienst
		79	Soziales Jahr

Übersicht 8.5: Rekodierungen für Stellung im Beruf

Mikrozensus	gemeinsames Rekodierungsschema		sozialwissenschaftliche Erhebung
Nr.	Nr.	Bezeichnung	Nr.
0	0	Selbständige ohne Besch.	10, 20
1	1	Selbständige mit Besch.	21, 22, 11
2	2	Mithelfender Familienangehöriger	30
3	3	Beamter, Richter	40 - 43
4	4	Angestellter	50 - 54, 64
5	5	Arbeiter	60 - 63
6, 7	6	Auszubildender	74, 79*
K, L	7	Wehrdienst	77
M	8	Zivildienst	78

* Zur Kennzeichnung der Auszubildenden wurden bei der sozialwissenschaftlichen Erhebung Informationen aus zwei anderen Fragen herangezogen.

Die sozialwissenschaftliche Erhebung beansprucht, die erwachsene Bevölkerung von Nordrhein-Westfalen repräsentativ abzubilden. Ein Vergleich der Häufigkeitsabbildungen mit jenen des Mikrozensus gibt Hinweise auf mögliche Probleme mit der Repräsentativität der verwendeten Stichproben und/oder der Kompatibilität der betreffenden Merkmale. In Tabelle 8.1 ist die Häufigkeitsverteilung für Variable "beruflicher Ausbildungsabschluß" als Beispiel aufgetragen.

Tabelle 8.1: Häufigkeitsverteilung beim Merkmal "beruflicher Ausbildungsabschluß"

	Mikrozensus %	sozialwissenschaftliche Erhebung %
in Ausbildung	13,0	5,7
ohne Ausbildung	22,6	4,4
Lehre/Fachschule	49,8	57,6
Praktikum	1,2	1,9
Meister	4,9	7,2
Fachhochschule	2,7	3,9
Universität	5,7	6,2
Keine Angabe	0,7	13,1

In diesem Fall sind nur relativ unproblematische Rekodierungen erforderlich. So gliedert die sozialwissenschaftliche Erhebung den Lehrabschluß stärker auf. Dennoch zeigen sich in der Verteilung der Bildungsabschlüsse erhebliche Abweichungen. Personen in und ohne Ausbildung sind in der sozialwissenschaftlichen Stichprobe unterrepräsentiert. Diese Unterschiede können auf zwei Weisen erklärt werden:

- Die beiden Stichproben bilden die Grundgesamtheit unterschiedlich ab. Dies würde bedeuten, daß die sozialwissenschaftliche Erhebung nicht vollständig repräsentativ ist, da angenommen werden kann, daß der Mikrozensus die Verteilung der Bevölkerung genauer wiedergibt.

- Es wird der gleiche Personenkreis befragt, aber es treten Unterschiede im Antwortverhalten auf.

Ist die erste Erklärung korrekt, funktioniert ein einfaches Abgleichverfahren ohne Probleme. Die Gültigkeit der zweiten Erklärung impliziert hingegen eine Störungsquelle. Eine Gewichtung der beiden Hypothesen oder eine Entscheidung für eine von beiden ist an Hand der Daten mit einer Ausnahme unmöglich. Die Ausnahme betrifft die Häufigkeit der Antwortverweigerungen. Hier ist anzunehmen, daß ein echter Verhaltensunterschied zutage tritt. In der amtlichen Erhebung besteht eine Verpflichtung zur Antwort; darum sind hier die Nonresponse Raten viel niedriger als bei der sozialwissenschaftlichen Erhebung. Der Unterschied zwischen den Antwortverweigerungen von 12,4 Prozent gibt eine untere Grenze für Dateninkompatibilitäten bei diesem Material an. Im Hin-

blick auf die bekannten Verzerrungen bei sozialwissenschaftlichen Stichproben (vgl. Esser u.a. 1989, Hartmann 1990, Schimpl-Neimanns 1991) ist anzunehmen, daß zumindest ein Teil der Unterschiede in den Randverteilungen auf einen systematischen Non-Response-Effekt zurückgeführt werden kann.

Bei den Zuordnungen wird die Kategorie "keine Angabe" wie eine "normale" Antwortmöglichkeit betrachtet. Sie wird nicht von den inhaltlich interpretierbaren Kategorien unterschieden. Personen mit einem fehlenden Wert in einer Variablen werden nicht aus der Analyse ausgeschlossen. Alternative Vorgehensweisen wären:

- Die Schätzung der "wahren" Merkmale mit einem statistischen Modell (vgl. dazu Little, Rubin 1987, Rubin 1987 und verschiedene Beiträge in Madow et al. 1983).

- Ausschluß aller Fälle mit fehlenden Werten bei einer oder mehreren Variablen.

Der Weg einer statistischen Schätzung wird nicht beschritten, weil dabei Wahrscheinlichkeiten zugrundegelegt werden müßten, d.h. die Ergebnisse unsicher sind. Außerdem ist das Kriterium für die ausschließlich angewendete Zuordnungsmethode (Abgleichverfahren) ihre Einfachheit. Es wäre bei aufwendigen und komplizierten Schätzungen nicht mehr erfüllt. Der Ausschluß von Fällen würde Information verschenken, da Zuordnungen bei Datensätzen mit fehlenden Werten im Mikrozensus und in der Stichprobe nicht mehr möglich wären.

Probleme mit unterschiedlichen Merkmalsausprägungen und mit fehlenden Werten treten bei verschiedenen Variablen auf. Sie sind typisch für die Verwendung verschiedener Datenquellen, die für jeweils eigene Zwecke erhoben wurden. Die Rekodierung der Merkmale ist verhältnismäßig aufwendig. In einigen Fällen müssen Filter für mehrere Variablen gesetzt werden, um zu identischen Skalen zu gelangen. Die Arbeiten betreffen beide Datenquellen. Soweit möglich, wird der Mikrozensus als Basis genommen.

8.2.4 Zusammenfassung

Insgesamt zeigt die hier beschriebene Datenaufbereitungsphase, daß eine problemlose Umsetzung von Datenquellen als Zusatzwissen für Reidentifikationsversuche nicht möglich ist. Die Aufbereitung des Gelehrtenkalenders hat verdeutlicht, daß die Umsetzung von Klartextangaben entsprechend der amtlichen Verschlüsselung mit erheblichen Unschärfen und z.T. auch Lücken einhergeht. Eindeutige Zuweisungen von Angaben der Befragten in vorgegebene Kategorien sind in der Regel nur bei fixen und klar definierten Merkmalen wie Geburtsdatum oder Geschlecht möglich.

Aber auch wenn die Merkmale bereits verschlüsselt und maschinenlesbar vorliegen, wie in der Regel bei sozialwissenschaftlichen Erhebungen, sind sie nicht schon per Knopfdruck für einen Reidentifikationsversuch verwendbar. Da in den Sozialwissenschaften bislang keine Anpassung an die amtliche Standarddemographie stattgefunden hat, müssen diese Merkmale ebenso wie Klartextangaben äußerst sorgfältig und daher sehr zeitintensiv aufbereitet werden. Der scheinbare Vorteil des geringeren Aufwands bei verschlüsselt und bereits maschinenlesbar vorliegendem Datenmaterial wird dann zum Nachteil, wenn die Abgrenzung der Kategorien nicht nach einem identischen Schema vorgenommen wurde, da hier der Spielraum der Klartextangaben fehlt. Eine Verkodung ist daher nur noch innerhalb des vorgegebenen Rahmens möglich, was sich unter Umständen negativ auf einen Reidentifikationsversuch auswirken kann.

9 Empirische Überprüfung des Reidentifikationsrisikos*

Die oben dargestellte Datenbasis ist im folgenden Grundlage für die experimentelle Überprüfung der faktischen Anonymität von Mikrodaten. Als Angriffstechniken werden sowohl das einfache Zuordnungsverfahren als auch die diskriminanzanalytische Reidentifikationstechnik von Paaß/Wauschkuhn eingesetzt. Da diese Techniken jeweils Extreme hinsichtlich ihrer technischen Komplexität und Leistungsfähigkeit darstellen, kann auf diese Weise das Kontinuum möglicher Reidentifikationsrisiken, sowohl in Hinblick auf die Erfolgschancen als auch auf den damit verbundenen Aufwand, ermittelt werden.

Die Nutzung von ausschließlich empirischen Daten ermöglicht es, einen Datenangriff unter realistischen Bedingungen nachzuvollziehen. Einer der wesentlichsten Unterschiede zwischen einem realen Angriffsszenario und den hier durchgeführten Experimenten besteht aber darin, daß die Datenbestände der Forscher keinerlei personenbezogene Informationen, sondern nur durchnumerierte Datensätze enthalten. Weiterhin stehen keine Nutzmerkmale, sondern nur die für die Experimente benötigten Überschneidungsmerkmale zur Verfügung.

Hiervon abgesehen, ist das Wissen der Forscher - vor und während der Experimente - dem eines Angreifers vergleichbar. Es ist unbekannt, welche im Zusatzwissen bzw. im Identifikationsfile (Gelehrtenkalender/Sozialwissenschaftliche Stichprobe) erfaßten Personen auch an der amtlichen Erhebung teilgenommen haben. Damit bleibt ebenfalls unbekannt, welche Datensätze im Zusatzwissen und Mikrodatenfile von einer Person stammen und wie hoch das reale Reidentifikationspotential ist. Ebenso wenig wie ein Angreifer wissen die Forscher, in welchem Ausmaß die Daten inkompatibel und inwieweit die gesuchten Personen von eventuell auftretenden Inkompatibilitäten betroffen sind. Auf diese Weise sind einem realen Datenangriff vergleichbare Unsicherheitsfaktoren gegeben. Im Unterschied zu einem realen Angriffsszenario werden die Resultate der Experimente jedoch durch einen Treuhänder überprüft und ermöglichen somit eindeutige Feststellungen, welche Zuordnungen richtig und welche falsch vorgenommen wurden.

---

* Autoren: Heike Wirth (9.1)
Uwe Blien, Michael Müller (9.2)
Heike Wirth, Uwe Blien (9.3)

Das erste empirisch zu überprüfende Subszenario geht von einem beruflich motivierten Datenangriff aus. Es wird hypothetisch unterstellt, daß ein Wissenschaftler versucht, möglichst viele Datensätze eines Mikrozensus zu deanonymisieren (*Massenfischzug*), um eine erweiterte Informationsbasis für eine Untersuchung über Wissenschaftler zu erhalten. Als Datenbasis für die Experimente wird - repräsentativ für öffentlich zugängliche Informationsquellen - der in Abschnitt 8.2 dargestellte Kürschners Deutscher Gelehrtenkalender (1987) herangezogen, das Mikrodatenfile wird durch den Mikrozensus Nordrhein-Westfalen (1987) repräsentiert.

Für die empirische Überprüfung dieses Szenarios wird sowohl die einfache als auch die diskriminanzanalytische Reidentifikationstechnik herangezogen. Im folgenden Abschnitt werden zunächst die Ergebnisse dargestellt, die bei der Verwendung einfacher Abgleichtechniken erzielt wurden.

9.1 Zuordnungen zwischen Mikrozensus und "Kürschners Deutschem Gelehrtenkalender" mit einfacher Abgleichtechnik

Die Diskussion der Faktoren, die das Reidentifikationsrisiko beeinflussen sowie der Fähigkeit unterschiedlicher Reidentifikationstechniken, diesen Faktoren gerecht zu werden, hat gezeigt, daß für ein vergleichsweise geringes Gefährdungspotential vermutet werden kann. Diese Reidentifikationstechniken berücksichtigen nicht, daß einzelne Merkmale im Mikrodatenfile oder Identifikationsfile inkompatibel oder fehlerbehaftet sein können. Ist dies der Fall, kann es zu Falsch- und/oder Nichtzuordnungen kommen, sofern nicht beide Datenbestände exakt die gleichen Fehler enthalten. Diese Verfahren stellen außer Ausprägungsidentität keine weiteren Kriterien zur Verfügung, um die Vornahme einer Zuordnung zu entscheiden. Weiterhin bieten einfache Reidentifikationstechniken keine angemessene Lösung der Doppelgängerproblematik. Das Ausmaß von Falsch- und Nichtzuordnungen ist daher durch diese Reidentifikationstechnik nicht kontrollierbar. Da diese einfache Methode jedoch bei einem Datenangriff die unterste Aufwandsgrenze markiert, ist es für die Operationalisierung der faktischen Anonymität unabdingbar zu überprüfen, wie die Erfolgschancen dieser "naiven" Techniken unter empirischen Bedingungen zu beurteilen sind.

Im folgenden wird skizziert werden, wie die einfache Abgleichtechnik technisch umgesetzt wurde (Abschnitt 9.1.1), und in welcher Weise die Alternativ-

verschlüsselung des Gelehrtenkalenders berücksichtigt wurde (Abschnitt 9.1.2). In den Abschnitten 9.1.3 und 9.1.4 werden die Zuordnungsexperimente sowie die Überprüfung der Ergebnisse durch den Treuhänder dargestellt. In Abschnitt 9.1.5 wird dieses Resultat kurz diskutiert. In Abschnitt 9.1.6 schließlich wird eine Kosten-Nutzen-Analyse im Sinne des Unverhältnismäßigkeitskriteriums geleistet.

9.1.1 Praktische Umsetzung der einfachen Reidentifikationstechnik

Für einen Reidentifikationsversuch mittels eines einfachen Zuordnungsverfahrens findet sich eine breite Palette möglicher Vorgehensweisen und verwendbarer EDV-Programme. Im Prinzip können alle Statistik- und Datenbankprogramme angewendet werden, die einen sukzessiven Abgleich jedes einzelnen Datensatzes aus dem Mikrodatenfile mit den Datensätzen des Identifikationsfile oder umgekehrt ermöglichen. Welche Vorgehensweise ein potentieller Angreifer verwenden würde, dürfte letztendlich von den ihm zur Verfügung stehenden Programmen und seinen EDV-Kenntnissen abhängen. Für die praktische Umsetzung einer einfachen Abgleichtechnik wird im folgenden auf ein in den Sozialwissenschaften weit verbreitetes Anwenderprogramm (SPSS^x) zurückgegriffen. Die Anwendung dieses Verfahrens erfordert keine tiefergehenden Programmierkenntnisse und steht in aller Regel jedem Sozialwissenschaftler zur Verfügung.

Das Verfahren basiert darauf, daß die Einzeldatensätze von zwei oder mehreren Datenfiles[2] im Hinblick auf bestimmte Schlüsselvariablen, hier repräsentiert durch die Überschneidungsmerkmale des Mikrodatenfile und des Identifikationsfile, auf Identität abgeglichen werden. Das Ergebnis dieses Abgleichs wird in einem neuen File, einem Matchfile, gespeichert (vgl. Block/Olsson 1975).

Die erfolgreiche Durchführung der Matching-Prozedur setzt voraus

- daß jeder individuelle Datensatz mit einer eindeutigen Kennziffer gekennzeichnet ist;

---

2 Mit dem SPSS^x-Programm können auf diese Weise maximal 50 Datenfiles zusammen gespielt werden.

- daß die Datensätze nach ihren Ausprägungskombinationen in aufsteigender Reihenfolge sortiert vorliegen. Eine hypothetische Datenbasis für eine Matching-Prozedur wird beispielhaft in Übersicht 9.1 dargestellt.

Übersicht 9.1: Hypothetische Datenbasis für eine einfache Reidentifikationstechnik

Mikrodatenfile(MF)				Identifikationsfile(IF)			
Kennung	Ü1	Ü2	Ü3	Kennung	Ü1	Ü2	Ü3
MF1	0	2	2	IF1	0	2	1
MF2	0	2	3	IF2	0	2	2
MF3	0	2	8	IF3	0	4	8
MF4	0	2	8	IF4	0	5	3
MF5	0	3	3	IF5	0	5	8
MF6	0	3	3	IF6	0	7	4
MF7	0	7	4	IF7	0	7	4
MF8	0	7	4	IF8	1	2	4
MF9	1	3	2	IF9	1	3	2
MF10	1	4	3	IF10	1	3	2
MF11	1	4	8	IF11	1	3	7
MF12	1	4	8	IF12	1	5	2
MF13	1	5	3	IF13	1	7	3
MF14	1	5	4	IF14	1	7	8
MF15	1	7	8	IF15	1	7	8

Die Verbindung der Einzeldatensätze in einem Matchfile erfolgt nach entsprechender Aufbereitung des Datenmaterials durch unterschiedliche Programmroutinen nach folgendem Prinzip:

Der Datensatz mit der niedrigsten Wertekombination für beide Datenbestände (im Beispiel IF 1 mit den Werten '0 2 1'), wird mit seiner Kennziffer in dem Matchfile gespeichert (vgl. Übersicht 9.2). Das Programm überprüft nun, inwieweit weitere Datensätze mit dieser spezifischen Ausprägungskombination auftreten. Im einfachsten Fall verläuft der Suchprozeß negativ, d.h. eine Zuordnung ist nicht möglich, da keine weiteren Datensätze mit dieser Ausprägungskombination ermittelt werden können. Das Programm kennzeichnet diesen

Befund durch einen spezifischen Missing-Code (in Übersicht 9.2 durch '*' gekennzeichnet).

Positiv verläuft der Suchprozeß dann, wenn eine spezifische Ausprägungskombination sowohl im Mikrodatenfile als auch im Identifikationsfile auftritt. Hierbei kommen verschiedene Möglichkeiten in Betracht: findet sich für eine spezifische Ausprägungskombination jeweils nur ein Datensatz in beiden Datenfiles (1:1 Relation), so wird diese Kombination mit der jeweiligen Kennziffer der beiden Datensätze in das Matchfile eingetragen. Sind in den Daten statistische Doppelgänger enthalten, wird bei einer 1:n (bzw. n:1) Relation die Einser-Kombination der ersten n-Kombination zugeordnet. Die verbleibenden statistischen Doppelgänger werden sukzessive als identisch in das Matchfile eingetragen, bekommen jedoch keinen Partnerdatensatz des anderen Datenbestands zugewiesen (vgl. Übersicht 9.2). Liegt bei den statistischen Doppelgängern eine m:n Relation vor, so werden die Kennziffern entsprechend ihrer Reihenfolge vermerkt.

Liegt dieses Matchfile vor, kann durch eine programmspezifische Subroutine ermittelt werden, wie viele Zuordnungen maximal möglich sind, d.h. wie viele unterschiedlich identische Ausprägungskombinationen zwischen Mikrodatenfile und Identifikationsfile auftreten. Aussagen über die Zuordnungsrelation, also ob eine 1:1 und damit eine eindeutige Zuordnung vorgenommen wurde, oder ob eine 1:n, n:1 oder m:n Zuordnung vorliegt, sind hierbei noch nicht möglich. Für diese Überprüfung stellt das Programm keine Subroutinen zur Verfügung. Dies kann nur durch eine relativ aufwendige manuelle Eingabe der jeweiligen Ausprägungskombinationen erfolgen oder durch die Entwicklung eines eigenen kleinen Programms, welches den Suchprozeß routinisiert. Da der Anspruch für die empirische Überprüfung darin bestand, die nach Möglichkeit einfachste Reidentifikationstechnik nachzuvollziehen, wurde kein eigenes Programm entwickelt, sondern eine Einzelüberprüfung aller zugeordneten Datensätze vorgenommen.

Übersicht 9.2: Hypothetisches Matchfile bei einer einfachen Reidentifikationstechnik

Ü1	Ü2	Ü3	Kennziffern MF	If
0	2	1	*	IF1
0	2	2	MF1	IF2
0	2	3	MF2	*
0	2	8	MF3	*
0	2	8	MF4	*
0	3	3	MF5	*
0	3	3	MF6	*
0	4	8	*	IF3
0	5	3	*	IF4
0	5	8	*	IF5
0	7	4	MF7	IF6
0	7	4	MF8	IF7
1	2	4	*	IF8
1	3	2	MF9	IF9
1	3	2	*	IF10
1	3	7	*	IF11
1	4	3	MF10	*
1	4	8	MF11	*
1	4	8	MF12	*
1	5	2	*	IF12
1	5	3	MF13	*
1	5	4	MF14	*
1	7	3	*	IF13
1	7	8	MF15	IF14
1	7	8	*	IF15

9.1.2 Berücksichtigung der alternativen Verkodung einzelner Merkmale im Gelehrtenkalender

Für das erste hier empirisch zu untersuchende Subszenario dient als Identifikationsfile Kürschners Deutscher Gelehrtenkalender, als Mikrodatenfile der Mikrozensus. Wie bereits bei der Datenbeschreibung erwähnt, konnten bei der Datenaufbereitung die Angaben für die Merkmale "Beruf", "Branche", "ausgeübte Tätigkeit" und "Fachrichtung des letzten Hochschulabschlusses" nicht in allen Fällen eindeutig einer spezifischen Kategorie der amtlichen Ver-

schlüsselung zugewiesen werden.[3] Um sich hieraus ergebende Fehlerquellen weitgehend auszuschließen, wurde für diese vier Merkmale eine Verschlüsselung nach je zwei alternativen Möglichkeiten vorgenommen (im folgenden als Alternativverschlüsselung bezeichnet).

Ein Abgleichverfahren in seiner einfachsten Form bietet keinen Ansatz, eine solche Alternativverschlüsselung angemessen einzubeziehen. Ein erweiterter Ansatz könnte jedoch in einer Zweiteilung des Zusatzwissens bestehen: In einer ersten Zuordnungsphase wären nur die ersten Alternativen zu berücksichtigen; in einer zweiten Phase könnten dann die zweiten Alternativen berücksichtigt werden.

Bei dieser Vorgehensweise würde allerdings das Informationspotential des Zusatzwissens nicht voll ausgeschöpft. Auch wenn sich die Datenaufbereitung an Plausibilitätskriterien orientiert hat, d.h. die plausibleren Alternativen jeweils an erste Stelle gestellt wurden, spiegelt die Alternativverschlüsselung ja gerade einen Unsicherheitsfaktor wider, der sich für die betroffenen Personen hinsichtlich dieser Angaben ergeben hat. Beispielsweise kann ein Hochschullehrer seine Tätigkeit sowohl als "Forschen" wie auch als "Lehren" betrachten. Wenn zusätzlich auch die Angabe zur "Hauptfachrichtung des Hochschulabschlusses" nicht eindeutig einer Kategorie zugeordnet werden konnte, kann anhand von Plausibilitätskriterien nicht mehr entschieden werden, in welcher Kombination eine betroffene Person in einer amtlichen Erhebung antworten würde. Werden daher zunächst ausschließlich die als Alternative 1 verschlüsselten Überschneidungsmerkmale und im folgenden nur die als Alternative 2 angegebenen Merkmale berücksichtigt, wäre eine gemischte Konstellation, in welcher eine Person im Mikrozensus bei ihrer Tätigkeit Alternative 1, für den Hochschulabschluß jedoch Alternative 2 angegeben hat, aus der Analyse ausgeschlossen.

Um diesen möglichen Informationsverlust zu vermeiden, müssen daher alle sich potentiell für eine Person ergebenden Antwortmöglichkeiten berücksichtigt werden. Bei vier (n) Überschneidungsmerkmalen mit jeweils zwei (k) Alternativen ergeben sich für jede im Zusatzwissen enthaltene Person theoretisch 16

---

[3] Zur Erinnerung seien nochmals die Überschneidungsmerkmale aufgeführt: Geschlecht, Geburtsjahr, Geburtsjahreshälfte, Beruf, Branche, Stellung im Beruf, ausgeübte Tätigkeit, Hauptfachrichtung des Hochschulabschlusses, Bundesraumordnungsregion und Gemeindegrößenklasse. Da sich die Daten auf Nordrhein-Westfalen beziehen, ist implizit das Bundesland als weiteres Überschneidungsmerkmal enthalten.

(k^n) Antwortvariationen im Mikrozensus.[4] Die im Mikrozensus hypothetisch für jeden Befragten existierenden 16 Variationsmöglichkeiten werden in den Zuordnungsexperimenten getrennt berücksichtigt.[5] Hierfür werden aus den Originalangaben des Gelehrtenkalenders - entsprechend Übersicht 9.3 - 16 unterschiedliche Identifikationsfiles gebildet. Jedes Identifikationsfile enthält jeweils die vollständige Population des Gelehrtenkalenders (n=7 983). Die eindeutigen Überschneidungsmerkmale werden in jedem Identifikationsfile unverändert übernommen, die mehrdeutigen Angaben werden variiert. Auf diese Weise können die im Gelehrtenkalender enthaltenen Informationen ohne Reibungsverluste für die Reidentifikationsexperimente genutzt werden.

Übersicht 9.3: Aus dem Original-Identifikationsfile erzeugte Versionen[6]

Version1 :	W1	B1	T1	H1	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version2 :	W1	B1	T1	H2	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version3 :	W1	B1	T2	H2	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version4 :	W1	B2	T2	H2	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version5 :	W1	B2	T2	H1	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version6 :	W1	B2	T1	H2	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version7 :	W1	B2	T1	H1	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version8 :	W1	B1	T2	H1	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version9 :	W2	B2	T2	H2	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version10 :	W2	B2	T2	H1	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version11 :	W2	B2	T1	H1	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version12 :	W2	B2	T1	H2	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version13 :	W2	B1	T2	H2	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version14 :	W2	B1	T1	H2	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version15 :	W2	B1	T1	H1	Gmgr	Gschl	Gbj	Gbh	Stib	Raum
Version16 :	W2	B1	T2	H1	Gmgr	Gschl	Gbj	Gbh	Stib	Raum

4 Da die Merkmale Beruf und Branche stark korrelieren, ist eine Antwortvariation mit Beruf 1 und Branche 2 bzw. umgekehrt äußerst unwahrscheinlich. Diese Variationen könnten daher unberücksichtigt bleiben. Um jedoch das Informationspotential des Zusatzwissens lückenlos auszuschöpfen, wurden auch diese Variationen in die Zuordnungsexperimente einbezogen.

5 Alternativ hätten die unterschiedlichen Variationsmöglichkeiten auch durch Entweder-Oder-Abfragen berücksichtigt werden können. Da die Matching-Subroutine von SPSS^x diese Möglichkeit nicht vorsieht, wurden die 16 Variationsmöglichkeiten in 16 unterschiedlichen Identifikationsfiles abgebildet.

6 W1/W2: Wirtschaftszweig1/2, B1/B2: Beruf1/2, T1/2: ausgeübte Tätigkeit1/2, H1/2: Hauptfachrichtung des Hochschulabschlusses1/2, Gmgr: Gemeindegrößenklasse, Gschl: Geschlecht, Gbj: Geburtsjahr, Gbh: Geburtsjahreshälfte, StiB: Stellung im Beruf, Raum: Bundesraumordnungsregion.

Insgesamt stehen für die Durchführung des Massenfischzugs damit 16 unterschiedlichen Identifikationsfiles mit jeweils knapp 8 000 Fällen (bezogen auf Nordrhein-Westfalen) etwa 170 000 Einzeldatensätze im Mikrozensus gegenüber.

9.1.3 Ergebnisse des Zuordnungsexperiments

Bei dem so entstandenen Datenmaterial muß im Hinblick auf das anzunehmende Gefährdungspotential nach den Annahmen des Uniqueness-Konzeptes von einem sehr hohen Reidentifikationsrisiko ausgegangen werden. Das Zusatzwissen umfaßt eine durch ihre Erwerbstätigkeit deutlich von der Durchschnittsbevölkerung abgegrenzte Subpopulation, die zudem stark in sich differenziert ist. Dies kommt unter anderem dadurch zum Ausdruck, daß in Abhängigkeit davon, welche der Alternativvariablen berücksichtigt werden, zwischen 45 und 65 Prozent der Datensätze des Gelehrtenkalenders einzigartige Ausprägungskombinationen aufweisen. Der Gesamt-Mikrozensus weist in der Aufgliederung nach allen zehn Überschneidungsmerkmalen und allen Merkmalsausprägungen zwar nur eine Einzelfallquote von 38,7 Prozent auf. Berücksichtigt man jedoch nur die für die Reidentifikationsexperimente relevante Zielpopulation[7], so stehen im Mikrozensus noch 3 099 Fälle zur Verfügung. Hiervon weisen dann 79 Prozent eine einzigartige Ausprägungskombination auf. Der sehr hohe Anteile von Fällen mit einmaligen Ausprägungskombinationen belegt, daß eine Untersuchungsanlage gewählt wurde, die nach den herkömmlichen Erwartungen Bedingungen enthält, bei welchen ein sehr hohes Reidentifikationsrisiko gegeben ist (vgl. z.B. Brunnstein 1987, Keller et al. 1990). Dies wird noch dadurch verstärkt, daß das Zusatzwissen nahezu eine Vollerhebung (vgl. Abschnitt 4.1) der spezifischen Gelehrten-

---

7 Die Zielpopulation definiert sich über die im Gelehrtenkalender auftretenden Merkmalsausprägungen, da nur diese bei einer einfachen Abgleichtechnik Berücksichtigung finden können. Die Berechnung der Einzelfallquote für die Zielpopulation im Mikrozensus erfolgt daher nur auf der Basis der Datensätze, die aufgrund ihrer Merkmalsausprägungen auch im Gelehrtenkalender auftreten können. Eine spezifischere Eingrenzung der Zielpopulation, beispielsweise ausschließlich auf die Kategorie der Gelehrten, erschien angesichts der partiell auftretenden mehrdeutigen Angaben nicht gerechtfertigt.

population darstellt. Nach dem Uniqueness-Konzept müßte daher ein sehr hohes Reidentifikationsrisiko gegeben sein.

Diese Erwartungen werden durch die Ergebnisse nicht bestätigt. Von den 7 983 Fällen des Gelehrtenkalenders können bei Berücksichtigung aller 16 unterschiedlichen Verkodungsvarianten lediglich 14 Datensätze eindeutig - d.h. in einer eins-zu-eins Relation - 14 Datensätzen des Mikrozensus zugeordnet[8] werden. Darüber hinaus finden sich weitere 11 identische Ausprägungskombinationen zwischen Mikrozensus und Gelehrtenkalender, bei welchen 15 Datensätze des Mikrozensus 25 Datensätzen des Gelehrtenkalenders in einer 1:n, n:1 bzw. m:n Relation, also mehrdeutig zugeordnet werden. Da für diese mehrdeutigen Zuordnungen mit dem zur Verfügung stehenden Datenmaterial keine weitere Eingrenzung möglich ist und ein Angreifer ebenfalls vor diesem Problem stehen würde, werden diese zunächst aus den weiteren Betrachtungen ausgeklammert.

Tabelle 9.1: Prozentualer Anteil der einzigartigen Ausprägungskombinationen in den und die Verteilung der eindeutigen Zuordnungen auf die einzelnen Datenfiles

	Identifikationsfile: Version Nr.																
	gesamt	red.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
N	169368	3099	7983	7983	7983	7983	7983	7983	7983	7983	7983	7983	7983	7983	7983	7983	7983	7989
a)	% 38,6	% 79,3	% 62,7	% 45,1	% 45,2	% 51,5	% 64,5	% 51,5	% 64,5	% 62,8	% 51,4	% 64,4	% 64,4	% 51,4	% 50,8	% 50,8	% 64,1	% 64,1
b)	14		1	-	3	-	-	-	-	7	-	3	-	-	-	-	-	-

a) Anteil der einzigartigen Ausprägungskombinationen in dem entsprechenden Datenfile.
b) Anzahl der eindeutigen Zuordnungen in dem entsprechenden Datenfile.

8 Der Begriff der Zuordnung besagt nicht, daß hier schon eine Reidentifikation vorliegt, sondern nur, daß sich in Mikrodatenfile und Zusatzwissen Datensätze mit identischen (bei einfachen Zuordnungsverfahren) und/oder ähnlichen Ausprägungskombinationen (bei der diskriminanzanalytischen Reidentifikationstechnik) finden. Diese Zuordnungen können sowohl in eindeutiger (1:1 Relation) als auch in mehrdeutiger (1:n, n:1, m:n) Relation vorliegen.

Wie aus Tabelle 9.1 hervorgeht, wird die Zahl der eindeutigen Zuordnungen hierbei - entgegen den Annahmen des Uniqueness-Konzepts - nicht durch den Anteil der Datensätze mit einzigartigen Merkmalskombinationen in Mikrodatenfile und Identifikationsfile bestimmt[9], sondern davon, in welcher Variation die alternativ verschlüsselten Überschneidungsmerkmale in den Identifikationsfiles berücksichtigt werden. Denn ungeachtet des relativ hohen Anteils von Einzelfällen im Mikrozensus sowie in den verschiedenen Identifikationsfile-Versionen ist in 12 der 16 Identifikationsfile keine Zuordnung möglich. Die Wahrscheinlichkeit einer Zuordnung - die noch nicht gleichbedeutend mit der Reidentifikationswahrscheinlichkeit ist - wird daher maßgeblich davon beeinflußt, wie die in den Klartextangaben - bedingt durch mehrdeutige Angaben - enthaltenen Unschärfen durch die Alternativverschlüsselung kompensiert werden konnten.

Die Verteilung der eindeutigen Zuordnungen auf die vier Identifikationsfile-Versionen überrascht insofern, als sieben der 14 zwischen Zusatzwissen und Mikrozensus als identisch sowie eindeutig ermittelten Ausprägungskombinationen auf die Identifikationsfile-Version 8 entfallen. Da sich die Verschlüsselung der Alternativ-Variablen an Plausibilitätskriterien orientierte, wäre eine vermehrte Zahl von Zuordnungen eher für Identifikationsfile-Version 1 zu erwarten gewesen. Auf diese Identifikationsfile-Variante entfällt jedoch nur eine der 14 eindeutigen Zuordnungen, d.h. wären die Angaben des Gelehrtenkalenders nur einfach verschlüsselt worden, wäre von 7 983 nur ein Fall eindeutig einem der 170 000 Fälle des Mikrozensus zuordenbar gewesen.

Da sich die Verkodungsvariante 1 des Gelehrtenkalenders von Variante 8 nur hinsichtlich des Merkmals "ausgeübte Tätigkeit" unterscheidet - in Variante 1 wurde die Alternative 1, in Variante 8 die Alternative 2 berücksichtigt - sind die Differenzen in den Zuordnungszahlen auf die für die Subpopulation der Gelehrten bereits erwähnte Mehrdeutigkeit hinsichtlich der ausgeübten Tätigkeit zurückzuführen. Bei der Datenaufbereitung wurde angenommen, daß die Gelehrten ihre eigene Tätigkeit eher im Sinne von "Forschen" als im Sinne von "Lehren" verstehen. Dementsprechend wurde bei den betroffenen Personen die ausgeübte Tätigkeit mit "Forschen" als Alternative 1 verschlüsselt, "Lehren" hingegen mit untergeordneter Priorität als Alternative 2 aufgenommen. Die

---

9 Dieses Ergebnis ist insofern nicht überraschend, als bereits in den Ausführungen zum Zusatzwissen und zu den Dateninkompatibilitäten deutlich wurde, daß die Einzigartigkeit von Datensätzen in einem Mikrodatenfile keineswegs die Folgerung zuläßt, daß diese Fälle auch eindeutig zugeordnet werden können und in der Konsequenz auch reidentifizierbar sind.

vermehrte Zuordnungschance bei Identifikationsfile-Variante 8 zeigt nun, daß die Hochschullehrer in ihrer Eigenwahrnehmung dazu tendieren, ihre Tätigkeit als "Lehren" aufzufassen. Dieser Eindruck wird durch die Zuordnungen in Identifikationsfile-Version 3 und 10 verstärkt; auch hier wurde jeweils Alternative 2, d.h. Lehren als ausgeübte Tätigkeit berücksichtigt.[10]

Gemessen an dem theoretisch zu erwartenden Reidentifikationspotential von etwa 80 Personen[11], ist die Zahl der tatsächlich erfolgten Zuordnungen überraschend gering. Selbst wenn man die mehrdeutigen Zuordnungen ebenfalls berücksichtigen würde, finden sich maximal nur 25 übereinstimmende Ausprägungskombinationen zwischen Mikrozensus und Gelehrtenkalender.

Ein Angreifer könnte diesen Befund nun folgendermaßen interpretieren:

- Die gesuchten Fälle sind zum Teil aufgrund von Dateninkompatibilitäten in Mikrodatenfile und Identifikationsfile unterschiedlich abgebildet und können daher mit einem einfachen Zuordnungsverfahren nicht ermittelt werden;

- die hohe Quote einmaliger Ausprägungskombinationen einerseits und die geringe Zahl der Zuordnungen andererseits deutet darauf hin, daß die 10 Überschneidungsmerkmale bezogen auf die spezifische Subpopulation der Gelehrten einen sehr hohen Informationsgehalt aufweisen. Die einzelnen Fälle nehmen daher im Merkmalsraum eine vergleichsweise eindeutige Position ein. Analog zu der Argumentation des Uniqueness-Konzeptes ist daher zu folgern, daß unter dieser Bedingung zufällige Zuordnungen aufgrund von statistischen Doppelgängern eher unwahrscheinlich sind und daher die eindeutigen Zuordnungen mit hoher Wahrscheinlichkeit auch korrekt sind (vgl. Bethlehem/Keller/Pannekoek 1990).

---

10 Dies bedeutet noch nicht, daß die ermittelten Zuordnungen korrekt sind, sondern nur, daß nach diesen Ergebnissen im Mikrozensus mehr Gelehrte mit der Ausprägungskombination "Lehren" als mit der Ausprägung "Forschen" ermittelt werden können.

11 Da der Mikrozensus eine Einprozent-Stichprobe der Gesamtbevölkerung darstellt und das Zusatzwissen 7 983 Fälle umfaßt, ergibt sich ein theoretischer Erwartungswert von etwa 80 Personen, die an beiden Erhebungen teilgenommen haben und damit potentiell zuordenbar sein müßten. Da dieser wahrscheinlichkeitstheoretische Wert jedoch Zufallsschwankungen nach oben und unten unterliegen kann, kann diese Zahl nur als Anhaltspunkt dienen.

Inwieweit diese Vermutungen der Realität entsprechen, würde sich der Kenntnis eines Angreifers allerdings entziehen, da er keine Möglichkeit hat, seine Ergebnisse in irgendeiner Form zu überprüfen.

9.1.4 Überprüfung der Ergebnisse durch den Treuhänder

Die Überprüfung der Ergebnisse durch den Treuhänder zeigt, daß eine eindeutige - auf der Basis einer identischen Ausprägungskombination ermittelte - Zuordnung zweier Datensätze aus Mikrodatenfile und Identifikationsfile nicht mit einer Reidentifikation gleichgesetzt werden kann.

Von den 14 mit dem einfachen Zuordnungsverfahren eindeutig zugeordneten Datensätzen sind lediglich vier korrekt, d.h. nur vier Datensätze stammen auch in der Realität von ein und derselben Person. Bezogen auf das Identifikationsfile entspricht dies einer Reidentifikationswahrscheinlichkeit von 0.0005. Diese kommt allerdings erst dadurch zustande, daß die 7 983 Fälle des Gelehrtenkalenders in den 16 unterschiedlichen Verkodungsvarianten auf identische Ausprägungskombinationen geprüft wurden. Zugleich liegt die Wahrscheinlichkeit einer Falschzuordnung bei der Nutzung einer einfachen Reidentifikationstechnik mit 0.71 wesentlich höher als die Wahrscheinlichkeit einer korrekten Zuordnung (vgl. Tab. 9.2).

Tabelle 9.2: Überprüfung der eindeutigen Zuordnungen durch den Treuhänder

Eindeutige Zuordnungen	absolut	%
Ingesamt	14	100
korrekt	4	29
falsch	10	71

Diese hohe Irrtumswahrscheinlichkeit bei einem Reidentifikationsversuch mit einer einfachen Abgleichtechnik wird von einer noch weitaus höheren "Nichtzuordnungs"-Wahrscheinlichkeit begleitet. Nach den Ermittlungen des Treuhänders sind 53 Fälle aus dem Identifikationsfile auch im Mikrozensus ent-

halten. Hiervon sind nach dem Kriterium der eindeutigen Zuordnung 92 Prozent nicht zuordenbar. Die Wahrscheinlichkeit, daß durch die einfache Abgleichtechnik Fälle, die tatsächlich sowohl im Mikrodatenfile als auch im Identifikationsfile enthalten sind, auch zugeordnet werden können, erhöht sich etwas, wenn man nicht nur die eindeutigen, sondern auch die *mehrdeutigen* Zuordnungen berücksichtigt. Von den 15 Fällen des Mikrodatenfile, die identische Ausprägungskombinationen zu 25 Fällen des Identifikationsfile in einer 1:n-, n:1- oder m:n-Konstellation aufweisen, sind sechs Fälle nach den Überprüfungen des Treuhänders tatsächlich in beiden Datenfiles enthalten. Mit den im Gelehrtenkalender enthaltenen Informationen ist in diesen Fällen aber keine eindeutige Zuordnung und damit keine Reidentifikation möglich. Ein Angreifer könnte allenfalls versuchen, die mehrdeutigen Zuordnungen eindeutig zu machen. Würde ihm dies gelingen, könnte er von den 53, in beiden Datenfiles enthaltenen Fällen, maximal zehn zuordnen. 43 Fälle bzw. 81 Prozent der in beiden Datenfiles enthaltenen Fälle können mit der einfachen Abgleichtechnik auch durch die Beschaffung weiterer Informationen *definitiv nicht* zugeordnet werden, da sie in einem oder mehreren der berücksichtigten Merkmale unterschiedlich abgebildet sind. Diese 43 Fälle hätten daher auch dann nicht zugeordnet werden können, wenn beliebig viele zusätzliche Überschneidungsmerkmale zur Verfügung gestanden hätten (vgl. Tab. 9.3). Die Begründung hierfür ergibt sich aus der Logik einfacher Abgleichtechniken, die in Abschnitt 3.2 dargestellt ist.

Tabelle 9.3: Verteilung von ein-, mehrdeutigen Zuordnungen sowie Nichtzuordnungen auf die zwischen Kürschners Gelehrtenkalender und Mikrozensus insgesamt möglichen Zuordnungen

	absolut	%
Insgesamt mögliche Zuordnungen	53	100
nicht zugeordnet	43	81
eindeutig zugeordnet	4	8
mehrdeutig zugeordnet	6	11

Nach den Ausführungen über die Schwächen einfacher Reidentifikationstechniken war zu erwarten, daß Datensätze von einer bestimmten Person, auch wenn sie in Zusatzwissen und Mikrodatenfile nur geringfügig unterschiedlich abgebildet sind, mit dieser Technik nicht mehr zugeordnet werden können. Ebenso war zu erwarten, daß vorgenommene Zuordnungen auch fehlerhaft sein können, sei es dadurch, daß statistische Doppelgänger nicht erkannt werden können oder dadurch, daß Zuordnungen zustande kommen, die nicht von derselben Person stammen. Allerdings ist das Ausmaß des Versagens unerwartet hoch, weshalb nochmals kurz mögliche Ursachen für die Falsch- und Nichtzuordnungen erörtert werden.

9.1.5 Mögliche Ursachen für das Versagen der einfachen Abgleichtechnik

Die Doppelgängerproblematik sollte unter den im untersuchten Szenario gegebenen Bedingungen nur eine geringe Rolle spielen, da der Gelehrtenkalender nahezu eine Vollerhebung der spezifischen Berufsgruppe der Hochschullehrer darstellt. Die entsprechende Berufskategorie im Mikrozensus umfaßt allerdings nicht nur Hochschullehrer, sondern auch Hochschulassistenten, wissenschaftliche Mitarbeiter und Lehrkräfte für besondere Aufgaben. Es wäre daher möglich, daß beispielsweise ein Hochschulassistent, der an der Mikrozensuserhebung teilgenommen hat, aber nicht im Gelehrtenkalender enthalten ist, die gleichen Merkmalsausprägungen wie ein Hochschullehrer aufweist, dessen Angaben zwar in Kürschners Deutschem Gelehrtenkalender enthalten sind, der aber kein Mikrozensusteilnehmer ist. Diese Möglichkeit kann allerdings als eher unwahrscheinlich eingestuft werden, da zwischen diesen Berufsgruppen i.d.R. Altersunterschiede bestehen und das Merkmal Alter als Überschneidungsmerkmal berücksichtigt wurde.

Angesichts der Tatsache, daß 81 Prozent der Fälle nicht zugeordnet werden konnten, obwohl sie sowohl der Population des Mikrozensus als auch des Gelehrtenkalenders angehörten, ist es wahrscheinlicher, daß die Falschzuordnungen ebenso wie die Nichtzuordnungen vor allem auf Dateninkompatibilitäten beruhen. Allerdings können hier nur Annahmen über die möglichen Quellen der Unstimmigkeiten getroffen werden. Exakte Aussagen würden eine Recherche bei den Betroffenen voraussetzen und sind damit ausgeschlossen.

Als Quelle derartiger Inkompatibilitäten ist zunächst der zeitliche Abstand zwischen der Erhebung des Gelehrtenkalenders (Sommersemester 1985) und des Mikrozensus (Frühjahr 1987) zu erwägen. Allerdings handelt es sich bei den zur Verfügung stehenden Überschneidungsmerkmalen vorwiegend um fixe Angaben. So sind weder beim Geburtsjahr und der Geburtsjahreshälfte noch beim Geschlecht oder Hochschulabschluß zeitlich bedingte Veränderungen zu erwarten. Auch bei den Merkmalen zur Erwerbstätigkeit (Wirtschaftszweig, Beruf, Stellung im Beruf und ausgeübte Tätigkeit) ist aufgrund der betrachteten Population anzunehmen, daß sich die Veränderungen nur in einem minimalen Rahmen bewegen. Da die Ermittlung der 53 Fälle umfassenden Schnittmenge zwischen den beiden Datenfiles von dem Treuhänder durch einen Anschriftenvergleich vorgenommen wurde, sind jene Personen, bei welchen sich innerhalb der unterschiedlichen Erhebungszeitpunkte regionale Veränderungen (z.B. durch Umzug) ergeben, in dieser Schnittmenge nicht enthalten. Daher können die nicht erfolgten Zuordnungen auch nicht auf Inkompatibilitäten in den Regionalmerkmalen beruhen. Als Quelle von Inkompatibilitäten verbleiben daher neben dem unterschiedlichen Antwortverhalten der befragten Personen vor allem Inkompatibilitäten, die sich aus dem unterschiedlichen Zweck der jeweiligen Erhebung ergeben können, sowie Kodierungs- und Übertragungsinkompatibilitäten. Dabei ist es unerheblich, ob die Inkompatibilitäten durch die Angaben im Mikrodatenfile oder im Identifikationsfile vorliegen. Denn wenn sie im Mikrodatenfile vorliegen, hat ein Angreifer bei der Verwendung einer einfachen Abgleichtechnik keinerlei Möglichkeiten, hierauf Einfluß zu nehmen. Liegen sie im Identifikationsfile vor, so ist angesichts der Tatsache, daß dieses File von mit der Mikrozensusverkodung vertrauten Fachkräften aufbereitet wurde, davon auszugehen, daß bei einer Verkodung durch einen Angreifer noch wesentlich mehr Abweichungen zu erwarten und die Erfolgschancen daher als noch geringer einzuschätzen sind. Es ist zu berücksichtigen, daß eine einfache Reidentifikationstechnik schon dann scheitert, wenn nur bei einem der zehn Überschneidungsmerkmale entweder bei Mikrodatenfile oder Identifikationsfile nicht exakt nach den gleichen Konventionen verkodet wurde.

9.1.6 Folgerungen für die faktische Anonymität

Bei der untersuchten Situation ergab sich für die Umsetzung der Klartext-Angaben des Gelehrtenkalenders in den Mikrozensuskode, die Ergänzung fehlender Angaben durch andere Informationsquellen, die maschinelle Aufnahme der Daten und Plausiblitätskontrollen ein Aufwand in Höhe von ca. 44 000 DM. Diese Arbeit wurde von Mitarbeitern des Statistischen Bundesamtes durch-

geführt, d.h. von Fachkräften, die mit der Mikrozensusverkodung vertraut sind. Wäre diese Arbeit, wie bei dem Szenario unterstellt, von den Wissenschaftlern selbst geleistet worden, wäre ein höherer Betrag zu veranschlagen, da die Aufbereitung des Gelehrtenkalenders ausgehend von der angenommenen Situation zunächst in der für die Sozialwissenschaften üblichen Verkodung vorgenommen worden wäre. Für die Verkodung von Berufsangaben wird beispielsweise üblicherweise der ISCO-Code verwendet. Für die Nutzung des Gelehrtenkalenders wären in diesem Fall zusätzliche Rekodierungsarbeiten und damit höhere Kosten angefallen. Weiterhin kann davon ausgegangen werden, daß wegen der fehlenden Praxis mit den Mikrozensus-Konventionen für die gleiche Leistung ein wesentlich höherer Aufwand an Zeit und Arbeitskraft angefallen wäre.

Für die Durchführung des Zuordnungsexperimentes entstanden zusätzliche Kosten in Höhe von über 17 000 DM. Diese beinhalten den für die Entwicklung des Konzeptes sowie für die Durchführung des Verfahrens erbrachten Arbeitsaufwand sowie die konkreten Rechenkosten. Nicht einbezogen wurden die für die Testläufe angefallenen Rechenkosten, da diese individuell beträchtlich variieren können und man auf diese Weise etwa die Kosten bei einer routinisierten Reidentifikationstechnik erhält. Der damit für die Ermittlung von 14 eindeutigen Zuordnungen[12] angefallene Betrag in Höhe von 62 000 DM bei cirka 1 000 Arbeitsstunden stellt eine untere Grenze des zu erbringenden Aufwands dar, denn für die Ermittlung der vier korrekten Zuordnungen - sofern dies überhaupt möglich ist - wäre ein zusätzlicher Aufwand zu erbringen.

Bei einer Alternativbeschaffung dieser Informationen beispielsweise durch eine Erhebung eines professionellen Umfrageinstituts kostet ein Interview im Schnitt zwischen 120 und 150 DM. Selbst unter der Bedingung, daß man die obere Grenze von 150 DM pro Interview zugrunde legt, hätte die Reidentifikationsquote in der untersuchten Situation noch um den Faktor 100 höher sein

12 Wie aus den allgemeinen Ausführungen über die Zuordnungsproblematik deutlich wurde, kann beim Auftreten von mehrdeutigen Zuordnungen ein eindeutiger Personenbezug nur dann hergestellt werden, wenn ein Angreifer in der Lage ist, zusätzliche Informationen zu erlangen und demnach einen zusätzlichen Aufwand erbringt. Eine Abschätzung derartiger Zusatzkosten ist hier allerdings nicht möglich und für die Bewertung des Unverhältnismäßigkeitskriteriums nach §16 Abs.6 BStatG auch nicht zwingend notwendig. Es ist jedoch abzusehen, daß der Einbezug der mehrdeutigen Zuordnungen die entstehenden Kosten nochmals beträchtlich erhöhen würde. Denn in diesem Fall würde ein Angreifer nicht nur zusätzliche Informationen benötigen, um mehrdeutige Zuordnungen eindeutig zu machen, sondern auch dann wäre es ihm noch nicht möglich, zwischen den falschen und korrekten Zuordnungen zu entscheiden.

können und die *faktische Anonymität* der Einzelangaben wäre noch gegeben gewesen.

Aus der Perspektive eines Angreifers erweisen sich einfache Abgleichtechniken damit als kein geeignetes Instrument für eine massenhafte Deanonymisierung von Einzeldatensätzen. Zwar ist die Durchführung des Verfahrens tatsächlich relativ problemlos und wenig kostenintensiv (der Kostenschwerpunkt liegt eindeutig bei der Aufbereitung des Gelehrtenkalenders), die unter empirischen Bedingungen allerdings immer zu erwartenden Inkompatibilitäten und statistische Doppelgänger stören diese Techniken so nachhaltig, daß auch nur annähernd sichere Zuordnungen nicht möglich sind:

- 81 Prozent der Fälle, die sowohl im Mikrozensus als auch im Gelehrtenkalender enthalten sind, konnten nicht zugeordnet werden, da die Angaben für diese Fälle im Mikrozensus anders abgebildet waren als im Gelehrtenkalender.

- Von den zehn identisch abgebildeten und damit auch mit der einfachen Abgleichtechnik potentiell korrekt zuzuordnenden Fällen konnten lediglich vier Datensätze auch eindeutig zugeordnet werden, weil die übrigen Datensätze in einem der Files mehrdeutig waren.

- Die Wahrscheinlichkeit, daß eine eindeutige Zuordnung falsch war, lag mit 0,71 wesentlich höher als die einer korrekten Zuordnung.

- Würde man - abweichend vom bisherigen Szenario - eine Einzelsuche unterstellen und zusätzlich annehmen, daß dem Angreifer die Mikrozensusteilnahme der oben genannten 53 Fälle bekannt ist (response knowledge), könnte der Angreifer das Identifikationsfile auf diese 53 Fälle beschränken. Somit entfallen störende statistische Doppelgänger im Identifikationsfile und fünf der sechs bislang mehrdeutig (korrekt) zugeordneten Datensätze könnten unter dieser Annahme ebenfalls eindeutig zugeordnet werden. Unter dieser Bedingung können also neun der 53 Fälle korrekt zugeordnet werden. Falschzuordnungen würden in diesem Fall nicht mehr vorkommen. Hierbei ist allerdings ebenso wie bei den anderen Ergebnissen vorausgesetzt, daß diese 53 Fälle in allen 16 Verkodungsvarianten überprüft werden. Ein solches Szenario, in welchem der Angreifer über alle Fälle Kenntnis hat, die gleichzeitig im Mikrodatenfile und im Identifikationsfile enthalten sind, ist in der Realität allerdings äußerst unwahrscheinlich.

Angesichts dieser unter empirischen Bedingungen so deutlich zutage tretenden Defizite einfacher Abgleichtechniken kann man darauf gespannt sein, wie die von Paaß/Wauschkuhn entwickelte diskriminanzanalytische Reidentifikationstechnik mit diesen Schwierigkeiten umgeht.

9.2 Empirischer Test der Anonymität des Mikrozensus. Verwendung der diskriminanzanalytischen Reidentifikationstechnik von Paaß/Wauschkuhn und "Kürschners Deutschem Gelehrtenkalender" als Identifikationsfile

Im vorausgehenden Abschnitt wurden die Ergebnisse empirischer Tests dargestellt, die auf einer Anwendung von einfachen Zuordnungsverfahren beruhten. Nun wenden wir uns den Experimenten zu, bei denen die diskriminanzanalytische Reidentifikationstechnik verwendet wird.[13] In der Literatur wird allgemein anerkannt, daß diese von der GMD entwickelte Methode dem beim vorliegenden Problem erreichbaren Optimum an methodischer Raffinesse nahekommt (vgl. z. B. Bethlehem et al. 1990).

Entscheidender Vorzug der diskriminanzanalytischen Reidentifikationstechnik von Paaß/Wauschkuhn ist, daß sie nicht nur in einem ersten Schritt eine Zuordnung eines Datensatzes des Identifikationsfile zu einem Datensatz des Mikrodatenfile gemäß maximaler "Ähnlichkeit" erlaubt (Prinzip der Distanzminimierung). Im zweiten Schritt wird vielmehr eine Berechnung der Wahrscheinlichkeit vorgenommen, daß die beiden Datensätze von der gleichen Person abstammen. Die Technik erlaubt eine Zuordnung trotz Datenfehler und beansprucht auch das Problem der Verwechslung mit identischen Datensätzen (Problem "statistischer Doppelgänger") zu lösen.

Der Unterschied des hier zugrunde gelegten Konzeptes für die Experimente gegenüber jenem der GMD ist, daß bei den im Rahmen des Anonymisierungsprojektes durchgeführten Tests ein Identifikationsfile mit unabhängig erhobenen empirischen Daten verwendet werden kann. Die GMD hatte für die Zuordnungen ein synthetisch aus dem Mikrodatenfile produziertes Identifikationsfile benutzt. Bei den in diesem Abschnitt dargestellten Tests wird dagegen die

---

[13] Vgl. den Forschungsbericht von Paaß und Wauschkuhn (1985) sowie Paaß (1985a), (1985b), (1987) und (1988). Die Reidentifikationstechnik wurde ausführlich in Abschnitt 3.4 dargestellt.

gleiche Datenbasis und die gleiche Untersuchungsanlage benutzt, die bereits im vorausgehenden Abschnitt bei der einfachen Reidentifikationstechnik zugrunde lag; das heißt Kürschners Deutscher Gelehrtenkalender dient als Identifikationsfile und der Mikrozensus von Nordrhein-Westfalen bildet das Mikrodatenfile.

Der Einsatz der diskriminanzanalytischen Reidentifikationstechnik ermöglicht somit einen unmittelbaren Vergleich der Schwierigkeit und Leistungsfähigkeit beider Verfahren. Um die Vergleichbarkeit zu gewährleisten, werden jeweils identische Annahmen über die Rahmensituation, Motive des Angreifers etc. zugrundegelegt. Ausgegangen wird von der Annahme eines beruflich motivierten Datenangriffs. Es wird unterstellt, daß der Angreifer ein Sozialwissenschaftler ist, der über einen Massenfischzug möglichst viele Datensätze des Mikrozensus identifizieren will, um eine erweiterte Datenbasis für eine Untersuchung über Wissenschaftler zu erhalten.

Es wird ebenfalls unterstellt, daß der Angreifer keine Kenntnis hat, welche der Personen, über deren Daten er im Identifikationsfile verfügt, auch an der Mikrozensusbefragung teilgenommen haben. Dies entspricht der Ausgangssituation eines Massenfischzugs. Eine Variation dieser Annahme ist mit der diskriminanzanalytischen Reidentifikationstechnik nicht möglich, da unbekannt ist, welche Datensätze im Mikrozensus zu Personen gehören, die auch bei der Erhebung des Gelehrtenkalenders befragt wurden.

9.2.1 Vorbereitung der Zuordnungsexperimente

Die diskriminanzanalytische Reidentifikationstechnik setzt voraus, daß drei verschiedene Quellen von Information als Input zur Verfügung stehen: Neben dem Mikrodatenfile und dem Identifikationsfile muß der Angreifer Angaben zu einem sogenannten "Fehlerprozeß" spezifizieren. In diese Spezifikationen gehen seine Erwartungen über Datenfehler und andere Abweichungen zwischen den beiden verwendeten Files ein. Die Anwendung der diskriminanzanalytischen Reidentifikationstechnik auf die vorhandene Datenbasis hat das Problem zu berücksichtigen, daß bei der Verkodung des Gelehrtenkalenders ein Teil der Personen bei vier Variablen nicht eindeutig bestimmten Kategorien zugewiesen werden konnte (vgl. Abschnitt 8.2.2.2). Aus diesem Grund wurden für diese vier Variablen Alternativausprägungen kodiert.

9.2.1.1 Berücksichtigung der Alternativvariablen

Die diskriminanzanalytische Reidentifikationstechnik ist nicht für die Aufnahme von Alternativvariablen konstruiert. Als Behelf wird eine ähnliche Vorgehensweise wie bei den einfachen Zuordnungsverfahren gewählt. Von den dort gebildeten sechzehn unterschiedlichen Identifikationsfiles werden zwei für unabhängig voneinander durchgeführte Experimente ausgewählt. Neben der plausibelsten Kodierung wurde diejenige Kodierung verwendet, die bei den Experimenten mit der einfachen Abgleichtechnik die größte Zahl an Zuordnungen erbracht hatte. Außerdem wurden die Unsicherheiten in der Spezifikation des Fehlerprozesses berücksichtigt. Im einzelnen ist das gewählte Vorgehen im folgenden begründet:

Eine Voraussetzung der diskriminanzanalytischen Reidentifikationstechnik ist, daß das Identifikationsfile und das Mikrodatenfile die gleiche Anzahl von Überschneidungsmerkmalen aufweisen. Aus diesem Grund können die zusätzlichen Variablen mit den Alternativausprägungen nicht gleichzeitig mit den übrigen Überschneidungsmerkmalen berücksichtigt werden.

Die Bildung von Alternativvariablen ist Ausdruck einer Unsicherheit über die Ausprägung eines Werts im Mikrodatenfile. Kann diese nicht bei der Spezifikation des Fehlerprozesses berücksichtigt werden? Die Bildung der Alternativvariablen basiert jedoch nicht auf allgemeinen Abschätzungen, sondern auf Informationen, die sich nur auf einen einzigen Fall beziehen.[14] So ist z. B. aus dem Gelehrtenkalender im Fall eines ganz bestimmten Professors die Fachrichtung seines Hochschulabschlusses eindeutig zu entnehmen. In einem anderen Fall ist keine Entscheidung zwischen zwei Abschlüssen möglich, z. B. weil der betreffende Forscher fächerübergreifend tätig ist und das Fach, in dem er einmal geprüft wurde, nicht angegeben ist.

Um die Alternativvariablen angemessen berücksichtigen zu können, werden zwei Maßnahmen gewählt:

- Eine indirekte Vorgehensweise, die einen Teil des Problems abdeckt, ist die Berücksichtigung von gehäuft auftretenden Doppeldeutigkeiten, die mit anderen Merkmalen korreliert sind. Eine bereits erwähnte typische Konstellation betrifft den Hochschulabschluß "Medizin". Hier kann oft nicht zwischen den Berufskodes Hochschullehrer und Arzt entschieden werden. Zur Berücksichtigung des "Fehlers", der möglicherweise gemacht wird, wenn eine spezifische Ausprägung gewählt wird, muß bei der Generierung des Fehlerfile jeweils die alternative Ausprägung angesetzt werden. Dann findet die Korrelation zwischen dem Auftreten dieses spezifischen "Fehlers" und der Ausprägung "Medizin" Eingang in die Abschätzung von

[14] Zu einer formalen Fassung des Problems vgl. den Anhang zum Abschnitt 9.2

Fehlerwahrscheinlichkeiten. Die diskriminanzanalytische Reidentifikationstechnik vermag derartige Korrelationen zu berücksichtigen, da in die dabei verwendeten Mahalanobisdistanz die Varianz-Kovarianz-Matrix der Fehler eingeht. Die gewählte Strategie zur Berücksichtigung von Alternativvariablen erfolgt also lediglich in Form ihres Ansetzens im Fehlerfile.

- Zusätzlich wird bei der Variablen mit der höchsten Anzahl von Doppeldeutigkeiten ("Tätigkeitsschwerpunkt") vorgegangen wie bei den einfachen Zuordnungsverfahren, d.h. es wurden beide Varianten als eigene Teil-Identifikationsfiles verwendet. Beim ersten Teil-Identifikationsfile wird die erste Alternative berücksichtigt ("Forschen"), beim zweiten Teil-Identifikationsfile die zweite ("Lehren").

Die Beschränkung auf zwei Teil-Identifikationsfiles ist notwendig, da der Rechenaufwand zu hoch wäre, um alle 16 Kombinationen für die vier doppeldeutigen Variablen durchzutesten, wie bei den Experimenten mit einfachen Zuordnungsverfahren. In der dort verwendeten Numerierung werden die Identifikationsfiles Nr. 1 und Nr. 8 für die diskriminanzanalytische Reidentifikationstechnik verwendet. Das File Nr. 1 enthält für die betreffenden vier Variablen jeweils die erste Alternative, die bei der Datenaufnahme als die plausibelste gegolten hatte.

Nach Durchführung der einfachen Zuordnungsverfahren zeigte sich, daß das Teil-Identifikationsfile mit der Nr. 8 die meisten eindeutigen Zuordnungen erbracht hatte. Dies war ein weiterer Grund, dieses auch für die Experimente mit der diskriminanzanalytischen Reidentifikationstechnik zu verwenden. Zum Zeitpunkt, als die Entscheidung für die Files Nr. 1 und Nr. 8 gefällt wurde, lagen die Ergebnisse der Überprüfung durch den Treuhänder noch nicht vor. Dies entspricht der Situation des Angreifers, der keinen Treuhänder fragen kann, ob seine Vorgehensweise richtig ist. So wird eine pragmatische Vorgehensweise des Angreifers simuliert, der als ersten Schritt einfache Zuordnungsverfahren anwendet, die von geringerem Aufwand sind. Im zweiten Schritt überprüft und ergänzt er die zunächst erhaltenen Ergebnisse mit einer komplizierteren Methode.

Auch wenn sich der Angreifer auf als plausibel anzusehende Verkodungskombinationen konzentriert, so beinhaltet seine Vorgehensweise eine Verletzung der Voraussetzungen der diskriminanzanalytischen Reidentifikationstechnik. Die Unsicherheit der vorliegenden Informationen, die durch das Vorhandensein von Alternativvariablen angezeigt wird, kann nicht mehr vollständig adäquat abgebildet werden. Dies läßt Konsequenzen für die Ergebnisse des Verfahrens erwarten: Es ist anzunehmen, daß der Anteil an Fehlzuordnungen gegenüber den Ergebnissen der GMD steigen wird. Vor Durchführung der Experimente ist jedoch nicht abzusehen, ob dies eine wichtige Einschränkung der Funktionsfähigkeit der Methode ist. Ihre von Paaß und Wauschkuhn (1985:12) berichtete Robustheit kann zunächst als Indikator dafür genommen werden, daß die Ergebnisse interpretationsfähig bleiben.

Schließlich ist festzuhalten, daß nur eine Minderheit der Fälle bei den drei Variablen, deren Alternativausprägungen nicht durch die Aufnahme der entsprechenden Teil-Identi-

fikationsfiles abgedeckt sind, gültige Werte aufweist. Immerhin 42 Prozent konnten hinsichtlich aller drei Variablen eindeutig bestimmten Ausprägungen zugewiesen werden. Für diese 42 Prozent hätten also bei den fraglichen drei Variablen Wirtschaftszweig, Beruf und Hochschulabschluß keine Alternativvariablen gebildet werden müssen. Zumindest dieser Teil der Fälle stellt das Potential dar, bei dem ein Deanonymisierungsverfahren von Alternativausprägungen unbehindert ansetzen kann.

9.2.1.2 Vorbereitung eines Massenfischzugs

Kennzeichen eines "Fischzugs" ist, daß ein Angreifer möglichst viele beliebige Datensätze eines Mikrodatenfile deanonymisieren will. Es kommt ihm auf die Zahl seiner 'Erfolge' an. Welche Personen er letztlich erhält, ist ihm gleichgültig.[15]

Bei einfachen Zuordnungsverfahren ergibt sich in der Vorgehensweise des Angreifers kein Unterschied zwischen einem Massenfischzug und einer Massenidentifikation. Dies ist bei der diskriminanzanalytischen Reidentifikationstechnik durchaus der Fall. Je nach Umfang der Datenfiles ist es zweckmäßig, die Rolle von Identifikationsfile und Mikrodatenfile umzukehren: Während bei einer Massenidentifikation zur Abschätzung verschiedener Eigenschaften der zugrundegelegten Population, die für die Beurteilung der Wahrscheinlichkeit von Doppelgängern und von Datenfehlern erforderlich ist, normalerweise das Mikrodatenfile verwendet wird, erfolgt die Schätzung beim Massenfischzug ausgehend vom Identifikationsfile.

Auch der eigentliche Prozeß der Identifikation läuft in umgekehrter Weise ab. Bei der Massenidentifikation werden die Datensätze des Identifikationsfile sukzessiv einzeln eingelesen, und es wird nach ähnlichen im Mikrodatenfile, das die Zieldatei bildet, gesucht. Da beim Massenfischzug beliebige im Mikrodatenfile enthaltene Fälle reidentifiziert werden sollen, übernimmt das Mikrodatenfile die Rolle des Identifikationsfile. Hier bildet letzteres die Zieldatei.

Dieser Austausch ist insofern von Bedeutung, als es vorkommen kann, daß einem Fall der Zieldatei mehrere Datensätze zugeordnet werden (vgl. Abschnitt 3.4.1.6). In der vorliegenden Situation eines Massenfischzugs ist also nicht ausgeschlossen, daß mehreren Fällen des Mikrozensus ein und derselbe Fall aus Kürschners Deutschem Gelehrtenkalender zugewiesen wird.

---

[15] Die Besonderheiten von Fischzugsszenarien wurden in Abschnitt 4.1.2 ausführlich erörtert.

9.2.1.3 Eingrenzung der Population

Die von der GMD eingesetzten Identifikationsfiles waren vergleichsweise sehr klein gewesen. Sie hatten zumeist nur 100 zufällig ausgewählte Fälle umfaßt. Bei der Durchführung der Rechnungen im Rahmen des Kripo-Szenarios wurden für 100 Fälle ca. 10 Minuten CPU-Zeit auf einer IBM 3083 benötigt (Paaß/Wauschkuhn 1985:167). Obwohl für die vorliegende Studie ungewöhnlich große EDV-Kapazitäten zur Verfügung stehen[16], erscheint es in Anbetracht des genannten Werts für die Rechenzeit von vornherein unmöglich, alle 160 000 Fälle des Mikrozensus von Nordrhein-Westfalen für die Experimente zu benutzen.

Zur Beschränkung der Rechenzeit wird darum eine Vorauswahl aus dem Mikrozensus getroffen. Es werden alle Fälle ausgeschlossen, die bei einer der relevanten Variablen eine Ausprägung aufweisen, die im Gelehrtenkalender nicht vorkommt. In Abhängigkeit von der Variablenkombination des verwendeten Teil-Identifikationsfile (siehe oben) wird auf diese Weise die Größe des relevanten Mikrodatenfile auf 2 633 (Teil-Identifikationsfile Nr. 1) bzw. 3 038 (Teil-Identifikationsfile Nr. 8) Fälle reduziert. Damit wird notgedrungen in Kauf genommen, daß potentielle Kandidaten für korrekte Reidentifikationen nicht mehr zugeordnet werden können. Dies ist z.B. dann der Fall, wenn durch einen Kodierungsfehler der Berufskode für "Hochschullehrer" in den Kode für "Schlosser" umgesetzt wird, da der Wert für "Schlosser" im Gelehrtenkalender nicht vorkommt. Eine der großen Stärken der diskriminanzanalytischen Reidentifikationstechnik, ihre Fehlertoleranz, wird erhalten, wenn eine Änderung zu einem Wert führt, der bei einem Datensatz im Gelehrtenkalender auftritt. Es ist zu erwarten, daß diese Art Dateninkompatibilität die bei weitem häufigere ist. Kodierfehler, die völlig unsystematisch erfolgen, sind vergleichsweise selten. Zudem liegt der bei weitem größte Teil der ausgeschlossenen Fälle des Mikrozensus im durch die Überschneidungsmerkmale aufgespannten Merkmalsraum (vgl. Abbildung 3.3, Situation 4) so weit von den Wissenschaftlern entfernt, deren Daten im korrespondierenden File enthalten sind, daß sie für eine Zuordnung ohnehin nicht in Frage kommen. Die durch die beschränkten EDV-Kapa-

---

16 Nach einer Übereinkunft mit dem Universitätsrechenzentrum Heidelberg werden die Rechenaufträge des Projekts mit besonders hohen Prioritäten ausgestattet. Nur so ist es möglich, die sehr rechenzeitintensiven Experimente durchzuführen. Um für die umfangreichen Dateien mit Zwischenergebnissen, die das AIMIPH-Programm automatisch angelegt, über entsprechenden Plattenspeicherplatz zu verfügen, wird eine komplette Magnetplatte (ca. ein Achtel der Kapazität des Rechenzentrums) allein für das Projekt reserviert.

zitäten erzwungene Vorauswahl der Fälle im Mikrozensus auf die für eine korrekte Zuordnung aussichtsreichsten Kandidaten erscheint daher vertretbar.

9.2.2 Generierung der Fehlerverteilungen[17]

Die diskriminanzanalytische Reidentifikationstechnik setzt die Eingabe von Informationen über Datenfehler und andere Inkompatibilitäten voraus. In dem Forschungsbericht des AIMIPH-Projekts wird in erster Linie auf echte Fehler abgestellt (vgl. auch Duncan/Lambert 1989:210). Im folgenden wird die dabei gewählte Begrifflichkeit beibehalten, indem von "Fehlerprozeß", "Fehlerfile" etc. die Rede ist. Diese Terminologie wird jedoch nur zum Zweck der Abkürzung verwendet. Andere Inkompatibilitäten sind eingeschlossen, wenn nicht ausdrücklich von spezielleren Fehlerquellen die Rede ist.

Welche Entscheidungen bei der Spezifikation des Fehlerprozesses zu treffen waren, ist in den folgenden drei Abschnitten im einzelnen dargestellt.

9.2.2.1 Nutzung vorhandener Informationsquellen

Die GMD hatte bei ihren Experimenten verschiedene Fehlerprozesse mit großer Sorgfalt spezifiziert. Um spezialisierte sozialwissenschaftliche Erkenntnisse verwenden zu können war auf das Fachwissen externer Experten zurückgegriffen worden, die mit der Erhebung der jeweils verwendeten Daten vertraut waren (vgl. Paaß/Wauschkuhn 1985:172ff.). Bei der im vorliegenden Zusammenhang wichtigen Untersuchung eines sozialwissenschaftlichen Szenarios muß von maximalen Kenntnissen über die Datenerhebung und die Reliabilität von Instrumenten ausgegangen werden.

Anhaltspunkte über das Auftreten von Fehlern und Problemen mit der Zuverlässigkeit eines Erhebungsinstruments können empirischen Analysen entnommen werden. Untersuchungen zur Reliabilität sind oft als Test-Retest angelegt. Aus Veränderungen einer Variablen zwischen zwei Panelwellen lassen sich globale Änderungswahrscheinlichkeiten errechnen. Im vorliegenden Fall wird wiederum die ALLBUS Test-Retest Studie (1984) herangezogen. Dabei handelte es sich um eine dreiwellige Panelbefragung mit jeweils gleichem Fragebogen. Der zeitliche Abstand zwischen den Wellen betrug jeweils ca. einen Monat. (vgl. hierzu auch Abschnitt 4.3).

Die Überschneidungsmerkmale von Gelehrtenkalender und Mikrozensus werden dabei größtenteils erfragt, so daß Schätzwerte empirisch ermittelt werden können. Nach einer Angleichung der Variablenstruktur hinsichtlich Ausprägungen und Wertebereich des

---

17 Vgl. hierzu insbesondere Müller (1991).

ALLBUS an den Mikrozensus werden für jede Variable die Veränderungen zwischen den Wellen bestimmt. Die dadurch erhaltenen Häufigkeiten von Abweichungen sind in Tabelle 9.4 in der ersten Spalte eingetragen. Diese Werte können zur Abschätzung der Bedeutung von Problemen mit der Reliabilität eines Erhebungsinstrumentes, einschließlich der enthaltenen Kodier- und sonstigen Datenfehler, verwendet werden.

Darüber hinaus sind zu den Abschätzungen auf der Basis der ALLBUS- Auswertungen noch andere als die genannten Formen von Dateninkompatibilitäten hinzuzufügen: Es fehlen die Auswirkungen der völlig unterschiedlichen Erhebungssituationen. Einerseits ist nicht bekannt, ob sich die Befragten bei der Erhebung zum Gelehrtenkalender die gleiche Mühe geben werden, wie bei einer Befragung durch die amtliche Statistik. Andererseits werden die Angaben im Gelehrtenkalender in personenbezogener Form veröffentlicht, so daß von einem gewissen Interesse der Befragten ausgegangen werden kann, korrekte Informationen zu geben. Es kann davon ausgegangen werden, daß derartige Informationsquellen wiederum eine eigene Fehlerstruktur haben, die von der einer schriftlichen Befragung abweicht.

Ein weiterer wichtiger Unterschied zwischen dem Gelehrtenkalender und dem Mikrozensus ist die Verpflichtung zur wahrheitsgemäßen Antwort beim Mikrozensus.

Wie bereits erwähnt, kann der Einfluß der jeweils spezifischen Erhebungssituation in seinen Auswirkungen kaum abgeschätzt werden. Ergeben sich in den beiden Datenquellen unterschiedliche Werte für einzelne Variablen, so sind die Abweichungen nicht einfach Fehler in einem üblichen Sinne, sondern teilweise darauf zurückzuführen, daß die Prozeduren der Datengenerierung selbst jeweils unterschiedliche soziale Prozesse sind.

Als eine ebenfalls mögliche Quelle von Abweichungen sind schließlich Veränderungen anzuführen, die durch den Abstand von ca. zwei Jahren zwischen den Erhebungszeitpunkten bedingt sind. Über ihre Höhe liegen keine Anhaltspunkte vor. Durchschnittswerte aus Panelanalysen für die gesamte Bevölkerung sind nicht verwendbar, da die hier relevante Population eine möglicherweise völlig davon abweichende Mobilität zeigt.

9.2.2.2 Alternative Annahmen über den Fehlerprozeß

Die Spezifikation der Fehler und anderer Inkompatibilitäten als Eingabe in das Programm der GMD ist angesichts der erwähnten Unsicherheiten nur auf der Basis von Annahmen möglich. Um den Einfluß solcher Annahmen auszutesten, werden vier verschiedene Varianten berücksichtigt. Sie sind in Tabelle 9.4 dargestellt. Die angegebenen Prozentwerte ergeben sich durch Addition

- der angenommenen Fehler im Mikrozensus,
- der angenommenen Fehler im Gelehrtenkalender und
- der angenommenen sonstigen Inkompatibilitäten, die aus dem zeitlichen Abstand der Files, der unterschiedlichen Fragesituation etc. entstehen.

Die Prozentwerte in Tabelle 9.4 drücken jeweils aus, welcher Anteil der Fälle bei der betreffenden Variablen von einer Abweichung betroffen ist.

Tabelle 9.4: Quantifizierung der Fehlerwahrscheinlichkeiten

Variable	Änderungsrate gemäß ALLBUS*	Fehlerprozeß Nr. 1	Fehlerprozeß Nr. 2	Fehlerprozeß Nr. 3	Fehlerprozeß Nr. 4
Gemeindegröße	2.6	5.0	5.0	2.6	5.0
Geschlecht	0.0	0.1	0.1	0.0	0.2
Alter	2.6	10.0	10.0	2.6	10.0
Geburtsmonat	-	10.0	10.0	2.6	37.5
Wirtschaftszweig	11.0	20.0	30.6	11.0	15.0
Stellung im Beruf	6.5	15.0	15.0	6.5	5.0
Beruf	7.8	15.0	26.2	7.8	15.0
Tätigkeit	-	15.0	57.5	5.0	5.0
Hochschulabschluß**	7.8	10.0	26.5	7.8	10.0
Raumordnungsregion	-	5.0	5.0	2.0	5.0

Angaben in Prozent

* Quelle für Spalte 1: Eigene Berechnungen mit Daten der ALLBUS Test-Retest Studie.

** Die Variable Fachrichtung des Hochschulabschlusses wurde im ALLBUS nicht explizit erfaßt. Der angegebene Wert bezieht sich auf den allgemeinen Schulabschluß.

Die Tabelle zeigt, daß die für die ALLBUS Test-Retest erhaltenen Werte beim Fehlerprozeß Nr. 1 als Anhaltspunkte dienten. Sie wurden gemäß Plausibilitätsüberlegungen nach oben verändert, um die bei der Studie zur Reliabilität von Umfragen nicht erfaßten Quellen von Dateninkompatibilitäten abzudecken.

Beim Fehlerprozeß Nr. 2 wird versucht, die im Gelehrtenkalender enthaltene Information der Alternativvariablen explizit als Fehlergröße zu interpretieren. Dabei wird davon ausgegangen, daß immer dann, wenn eine Person nicht eindeutig einer bestimmten Kategorie einer Variablen zugewiesen werden kann, mit einer Wahrscheinlichkeit von 0.5 (statt der ansonsten zugrunde gelegten Alternative 1) die Alternative 2 anzusetzen ist. Zusätzlich sind dann noch die üblicherweise angenommenen Fehler zu berücksichtigen. Aus der Kombination der beiden Fehlerquellen resultieren relativ hohe Fehlerwahrscheinlichkeiten. Die betreffenden Variablen werden für den Identifikationsprozeß dennoch nicht vollständig wertlos. Der Anspruch der diskriminanzanalytischen Reidentifikationstechnik ist vielmehr, die Modellierung von Inkompatibilitäten in großer Differenziertheit zu ermöglichen. Dabei wird auch jene Information berücksichtigt, die in der Umsetzung von bestimmten Kategorien auf bestimmte andere enthalten ist. Die Berechnung der numerischen Werte für die zugrundegelegten Fehlerwahrscheinlichkeiten bei den vier

Variablen mit Alternativausprägungen, wie sie in Tabelle 9.4, Fehlerprozeß Nr. 2, dargestellt sind, ist aus dem Anhang zu Abschnitt 9.2 zu ersehen.

In einem weiteren Fehlerprozeß (Nr. 3) werden besonders niedrige Fehler simuliert. Zur Vereinfachung wurden die Werte für die ALLBUS Test-Retest Studie berücksichtigt und für dort fehlende Variablen plausible Werte eingesetzt.

Die Generierung einer Fehlerverteilung setzt stets die Definition von Übergangswahrscheinlichkeiten zwischen Kategorien voraus. Hat jemand z.B. einen Hochschulabschluß im Fachgebiet der Volkswirtschaftslehre erworben, so ist die Wahrscheinlichkeit, daß er fälschlicherweise als Betriebswirt eingestuft wird, höher als die Wahrscheinlichkeit, daß er für einen Arzt gehalten wird. Bei allen generierten Fehlerprozessen wird mit plausiblen Werten für diese Übergangswahrscheinlichkeiten gearbeitet. Um auch hier zu alternativen Strukturen zu gelangen und um einige weitere Hypothesen über die absolute Höhe von Fehlern zu testen, wird ein weiterer Fehlerprozeß (Nr. 4) spezifiziert. Bei ihm werden für die relevanten Variablen mit Alternativausprägungen wiederum die betreffenden Alternativwerte, wie bei Fehlerprozeß Nr. 2, angesetzt. Allerdings werden die dabei auftretenden Übergangswahrscheinlichkeiten deutlich nach unten korrigiert.

Die vier Varianten für die Höhe und Struktur von Fehlern und anderen Inkompatibilitäten decken ein breites Spektrum an Annahmen ab. Die Information der Alternativvariablen wird bei zwei Fehlerprozessen berücksichtigt.

9.2.2.3 Generierung der Fehlerfiles

Die Angaben zum Fehlerprozeß werden von einem Teilmodul des AIMIPH-Programms verarbeitet.[18] Ausgegangen wird normalerweise vom größeren der beiden als Dateninput verwendeten Files. Die Spezifikationen zum jeweiligen Fehlerprozeß werden benutzt, um ein neues File, das sogenannte "Fehlerfile" (FF), zu generieren. Dazu werden an den Variablenwerten des Identifikationsfile Änderungen gemäß eines Zufallsprozesses vorgenommen. Der Zufallsprozeß hat definierte Eigenschaften, die exakt festzulegen sind.

Welches der beiden Inputfiles, Identifikationsfile oder Mikrodatenfile, sollte zur Generierung des Fehlerfile verwendet werden? Bei der Behandlung der Datenkonstellation eines Massenfischzuges wurde erläutert, daß der Mikrozensus nicht verwendet werden sollte. Die Subpopulation der "Gelehrten" ist im Mikrozensus nur schwach vertreten, da dieser eine Einprozentstichprobe darstellt. Viele im Gelehrtenkalender vorkommenden Ausprägungskombinationen sind im Mikrozenus unbesetzt. Eine Schätzung der Fehlerhäufigkeiten auf der Grundlage des Mikrodatenfile wäre zu ungenau.

Aber auch die Verwendung von "Kürschners Deutschem Gelehrtenkalender" hat Nachteile. Dieser ist zwar nach seinen eigenen Kriterien nahezu eine Totalerhebung und

[18] Lediglich für Fehlerprozeß Nr. 4 wird ein spezielles SPSS^X-Programm verwendet.

insofern für die Berechnung beliebiger Wahrscheinlichkeiten ideal geeignet. Die Abgrenzung der Basispopulation des Gelehrtenkalenders ist jedoch nicht trennscharf nach den Definitionen des Mikrozensus. In der Mikrozensuskategorie "Hochschullehrer" sind 199 Fälle enthalten. Dies wären hochgerechnet ca. 20 000 Personen, deutlich mehr als die knapp 8 000 "Gelehrten" des Gelehrtenkalenders. Demnach gibt es Hochschullehrer, die im Sinne des Kalenders keine "Gelehrte" darstellen. Dies wird bereits durch die Bestimmung des Begriffs im Kalender selbst indiziert. Nicht aufgenommen wurden z.B. wissenschaftliche Angestellte, die noch nicht durch Veröffentlichungen hervorgetreten sind. Im Gelehrtenkalender sind andererseits Personen enthalten, die nach Mikrozensuskategorien keine Hochschullehrer sind. Bei Zugrundelegung des Kategoriensystems der amtlichen Statistik ist das Handbuch also keine Totalerhebung. Es ist aber auch keine Zufallsstichprobe, da die Auswahl systematisch erfolgt. Letztlich folgt also das nur scheinbar paradoxe Resultat, daß eine Totalerhebung, eben "Kürschners Deutscher Gelehrtenkalender"[19], nicht nur keine Verwendung als Totalerhebung finden kann, sondern auch noch zusätzliche Probleme aufwirft, weil es sich bei ihm nicht einmal um eine Zufallsstichprobe handelt.

Das skizzierte Problem ist nicht nur für Handbücher typisch, sondern auch für Spezialstichproben der Sozialforschung, in denen die Merkmale anders definiert sind als im Mikrozensus.[20] Es weist über die Abschätzung der Fehlerverteilungen hinaus. Bei der Berechung der verschiedenen Wahrscheinlichkeitsdichten im AIMIPH-Programm wird vorausgesetzt, daß die Daten entweder Totalerhebungen oder Zufallsstichproben sind. In Abwägung der dünnen Besetzung im Mikrozensus mit dem Problem der systematischen anstatt zufälligen Auswahl im Gelehrtenkalender, erfolgt die Entscheidung doch dafür, letzteren zur Grundlage für die Generierung des Fehlerfile und in der Folge für die Dichteschätzungen zu verwenden, da andernfalls zu große Schwierigkeiten bei der Durchführung der Rechnungen die Konsequenz gewesen wären.

Bei der Durchführung der Experimente zeigt sich dann indirekt, daß diese Entscheidung richtig gewesen ist. Obwohl der Gelehrtenkalender die Subpopulation weitgehend komplett enthält, tauchen immer wieder technische Schwierigkeiten mit verschiedenen Schätzern auf. Zu geringe Besetzungen an den "Rändern" der Verteilung führen dazu, daß nicht mehr zuverlässig geschätzt werden kann und die Abänderung des Programms sowie die Einfügung zusätzlicher Routinen zur Fehlerbehandlung notwendig wird. Mit dem Mikrozensus könnte nicht in der gleichen Weise gearbeitet werden. Die diskriminanzanalytische Reidentifikationstechnik war von seinen Autoren nicht an Spezialpopulationen, sondern an großen Bevölkerungsstichproben mit metrischen Variablen getestet worden, bei denen viele Fälle relativ gleichmäßig nebeneinander liegen und die Verteilungen einigermaßen kontinuierlich ausfallen.

Damit wird erneut deutlich, daß die Verwendung des Gelehrtenkalenders für die Tests eine relativ günstige Angriffssituation unterstellt, weil zusätzlich zum Mikrodatenfile ein File mit einem sehr hohen Anteil der Population für die Schätzungen zur Verfügung

---

19 Hier kann nicht geprüft werden, welche Lücken der Gelehrtenkalender aufweist, wenn man ihn an seinen eigenen Ansprüchen mißt.

20 Zur Problematik der Definition von Grundgesamtheiten und ihrer Umsetzung in der empirischen Sozialforschung vgl. Schnell (1991).

steht. Besäße ein Angreifer nur einzelne Datensätze mit Informationen über Wissenschaftler, könnte er das Programm bei der vorliegenden Datenstruktur nicht in adäquater Weise anwenden.

Da das Identifikationsfile für die Fehler- und Dichteschätzungen verwendet wird, können die Wahrscheinlichkeiten, mit der die Angaben des Identifikationsfile geändert werden um das Fehlerfile zu erzeugen, Tabelle 9.4 entnommen werden. Die Art und Weise, in der die Veränderung der Werte erfolgt, wird durch den Typ und die Parameter der angenommenen Fehlerverteilung festgelegt. Die im Mikrozensus bzw. Gelehrtenkalender enthaltenen Variablen sind ganz überwiegend von ordinalem bzw. nominalem Skalenniveau. Dieser Variablentyp wird im AIMIPH-Fehlerprogrammodul separat berücksichtigt. Für die Variablen Geschlecht, Geburtsmonat, Gemeindegröße, Stellung im Beruf, Raumordnungsregion und Tätigkeitsschwerpunkt werden Fehlerverteilungen auf die entsprechende Weise generiert. Das AIMIPH-Programm erlaubt die Angabe separater Abgangswahrscheinlichkeiten für die einzelnen Kategorien. Bei den Fehlerprozessen Nr. 1 - 3 werden Werte angenommen, die sich soweit wie möglich an den Auswertungen der ALLBUS Test-Retest Daten orientieren. Da die ALLBUS-Studie für die hier untersuchte Population nicht sehr viel weiterhilft, müssen wiederum Annahmen gemäß Plausibilitätsüberlegungen getroffen werden, die im Fehlerprozeß Nr. 4 variiert werden.

Die einzige Variable mit metrischem Skalenniveau ist das Alter. Hier wird eine Normalverteilung für die Fehler zugrunde gelegt. Bei den Variablen Wirtschaftszweig, Beruf und Hauptfachrichtung des letzten Hochschulabschlusses wird ebenfalls angenommen, daß die Fehler normalverteilt sind. Dies ist möglich, da die Ausprägungen so angeordnet wurden, daß 'ähnliche' nebeneinander liegen. Die betreffenden Variablen haben derart viele Ausprägungen, daß aufgrund der vorliegenden spärlichen Informationen eine Spezifikation von einzelnen Übergangswahrscheinlichkeiten nicht möglich ist.

Da bei Anwendung der linearen Diskriminanzanalyse ordinale und nominale Variablen zweckmäßig verarbeitet werden können, wenn diese durch dichotome Variablen ersetzt werden, erfolgt als nächster Arbeitsschritt die Dichotomisierung der betreffenden Variablen. Dies geschieht, indem für eine Variable mit I Ausprägungen, I - 1 Dummyvariablen gebildet werden.

Die Differenzen zwischen dem Identifikationsfile und dem Fehlerfile sind genau die durch die Spezifikation des entsprechenden Fehlerprozesses zufällig erzeugten Fehler. Für die Durchführung der diskriminanzanalytischen Reidentifikationstechnik ist die Ermittlung der Varianz-Kovarianz-Matrix der Fehler notwendig. Diese geht in die Ermittlung der a priori Wahrscheinlichkeiten ein, daß zwei Fälle zusammengehören (vgl. die Bayessche Formel in Abschnitt 3.4.1.1). Sie wird aus den Differenzen zwischen Identifikationsfile und Fehlerfile geschätzt.[21]

---

[21] Bei der Spezifikation des Fehlerprozesses könnten beliebige Abhängigkeiten zwischen den Variablen festgelegt werden. Darauf wurde jedoch verzichtet, da wiederum die Informationen fehlten. Die GMD nahm solche Abhängigkeiten bei ihren eigenen Experimenten nur dort an, wo für das fehlerhafte Entfallen von Haushaltsmitgliedern Annahmen entwickelt wurden (Paaß/Wauschkuhn 1985:172ff.). Man könnte annehmen, daß in der Varianz-Kovarianz-Matrix alle nicht in der

9.2.3 Durchführung und Ergebnisse der Rechnungen

9.2.3.1 Zuordnung durch Distanzminimierung

Die Berechnung der Varianz-Kovarianz-Matrix der Fehler ist bereits Teil des eigentlichen Deanonymisierungsprogramms. Die Matrix wird benötigt, um ein spezielles Distanzmaß, die sogenannte Mahalanobisdistanz, zu berechnen. Dieses Maß gibt an, wie (un-)ähnlich sich zwei Fälle - im vorliegenden Zusammenhang ein Fall aus dem Identifikationsfile und ein Fall aus dem Mikrodatenfile - sind. Das Eingehen der Varianz-Kovarianz-Matrix der Fehler (genauer: der Inversen dieser Matrix) in dieses Maß führt dazu, daß eine Gewichtung der einzelnen Variablen gemäß Fehlerhäufigkeiten und Korrelationen mit anderen Variablen vorgenommen wird (vgl. Abschnitt 3.4.1.4).

Das AIMIPH-Programm berechnet automatisch die Distanz des aktuell betrachteten Falles des Mikrodatenfile zu allen Fällen des Identifikationsfile und sucht die jeweils nächstgelegenen Nachbarn heraus. Eine solche Tabelle ist beispielhaft im Tabellenanhang (Tabelle 9A1) wiedergegeben. Sie enthält in den Zeilen einzelne Fälle des Mikrodatenfile und in den Spalten die jeweils fünf nächsten Nachbarn, die dazu im Identifikationsfile enthalten sind. Programmintern wird mit den 500 nächsten Nachbarn weitergerechnet.

Die Distanztabelle mit den fünf nächsten Nachbarn ist das erste inhaltlich interpretierbare Ergebnis der Programmanwendung. Würde der Angreifer an dieser Stelle abbrechen, z.B. weil er den mit der Implementation des Rest-Programms verbundenen Aufwand scheut, könnte er einen 'Datenangriff durch Distanzminimierung' versuchen. Hinter diesem Stichwort verbirgt sich ein Verfahren, das in der Schwierigkeit der Anwendung eine Mittelstellung zwischen der diskriminanzanalytischen Reidentifikationstechnik und einfachen Abgleichverfahren einnimmt (Abschnitt 3.3). Bei dieser Vorgehensweise sucht sich der Angreifer besonders "nahe" zusammenliegende Fälle y_i (aus dem Mikrodatenfile) und z_j (aus dem Identifikationsfile) heraus. Unterschreitet die Distanz einen bestimmten Schwellenwert $\hat{e}$, so betrachtet er die betreffenden Datensätze als zusammengehörig, d.h. von der gleichen Person stammend. Er nimmt dann an,

---

Hauptdiagonale stehenden Elemente nahe Null sind, da keine Abhängigkeiten zwischen den Fehlern verschiedener Variablen spezifiziert wurden. Solche Elemente können jedoch von Null verschieden sein, weil durch die Dichotomisierung Abhängigkeiten bei den Fehlern entstehen. Außerdem können bei den Fehlerprozessen Nr. 2 und 4 empirische Abhängigkeiten zwischen dem Auftreten von Alternativvariablen auf die Varianz-Kovarianz-Matrix der Fehler durchschlagen.

daß die vorhandenen Unterschiede in den Daten allein durch Fehler (oder andere Inkompatibilitäten) bedingt sind.

Für die Wahl eines Wertes ê steht jedoch kein auf irgendeine Weise begründbares Kriterium zur Verfügung. Da bereits die Wahl der Distanz Null, die der Durchführung eines Abgleichs auf identische Merkmale in einem einfachen Zuordnungsverfahren entspricht, zu einem hohen Anteil von Falschzuordnungen geführt hat (vgl. Abschnitt 9.1), sind noch mehr Falschzuordnungen zu erwarten, wenn eine Distanz größer Null und damit ein weicheres Kriterium für die Zuordnungen gewählt wird. Paaß/Wauschkuhn vertraten bei der Entwicklung ihrer Methode den umgekehrten Standpunkt: Sie erwarteten das Auftreten von Fehlern, von Abweichungen, und wollte Kriterien dafür berechnen, welche Fälle trotz Distanzen größer Null zugeordnet werden sollten.

9.2.3.2 Darstellung der Ergebnisse der Zuordnungsexperimente

Der Hauptteil des Programms der GMD dient zur Schätzung der unbedingten Wahrscheinlichkeit verschiedener Ausprägungskombinationen (vgl. Abschnitt 3.4.2). Diese wird zweckmäßigerweise mit den Daten des größeren File vorgenommen, im vorliegenden Fall eines Massenfischzugs also mit den Daten des Identifikationsfile. Die dabei entstehende Problematik wurde bereits im Zusammmenhang mit der Behandlung der Generierung von Fehlerfiles diskutiert. Das Verfahren verlangt die Eingabe des Auswahlsatzes der betreffenden Stichprobe. Nun ist der Gelehrtenkalender keine Stichprobe, sondern eine Totalerhebung. Allerdings ist der Gelehrtenkalender zwar nach den Kategorien des Mikrozensus verkodet, die Abgrenzung der in ihm enthaltenen Subpopulation erfolgte jedoch nach ganz anderen Kriterien. Näherungsweise wird ein Auswahlsatz von 30 Prozent angenommen. Dies ist eine Abschätzung nach der sicheren Seite. Je kleiner der Auswahlsatz angenommen wird, umso höhere Maßstäbe werden an die Ähnlichkeit von Fällen angelegt. Wird trotz niedrigem Auswahlsatz zugeordnet, kann auf dieses Ergebnis umso eher vertraut werden.

In dem ersten Zuordnungsexperiment wurde zur Generierung eines Fehlerfile der Fehlerprozeß Nr. 1 benutzt. Die Ergebnisse der Rechnungen[22] sind in

[22] Eine umfassende Darstellung der Zuordnungsergebnisse findet sich bei Müller (1991).

Tabelle 9.5 festgehalten, in der alle Zuordnungen mit einer Wahrscheinlichkeit von über 95 Prozent eingetragen sind. Die Werte für die Zuordnungswahrscheinlichkeit sind in Spalte 3 angegeben. Spalte 1 enthält eine Kennummer des Identifikationsfile und Spalte 2 eine solche des Mikrodatenfile. Zum zusätzlichen Schutz der Identität der Personen, von denen die Daten stammen, wird eine durchlaufende Numerierung der Fälle benutzt.[23] In Spalte 4 steht jeweils ein Stern, wenn für das betreffende Paar von Datensätzen eine Zuordnungswahrscheinlichkeit von über 99 Prozent berechnet wird und es eindeutig ist. Als Mehrfachzuordnung, die keinen Stern erhielt, wurde ein Fall bereits dann gezählt, wenn ein anderer mit der Wahrscheinlichkeit von mindestens 95 Prozent ebenfalls zugeordnet wurde. Spalte 5 enthält schließlich die aus der Transformation der Mahalanobisdistanz (vgl. Abschnitt 3.4) zwischen den zugeordneten Fällen berechnete euklidische Distanz.

Die Tabelle zeigt erwartete und unerwartete Effekte. Die einzige Assoziation bei einer Distanz von Null (Identifikationsfile-Fall Nr. 4512) war auch schon mit einfachen Abgleichverfahren festgestellt worden. Für sie wird eine sehr hohe Wahrscheinlichkeit knapp unter dem Maximalwert von 1 berechnet. Dies kann damit erklärt werden, daß in relativ großer Distanz zu dem betreffenden Fall im Identifikationsfile weitere liegen. Eine Verwechslung kann somit nicht mit vollständiger Sicherheit ausgeschlossen werden, ist aber sehr unwahrscheinlich.

---

23 Die Nummern, mit denen die beiden Files ursprünglich von ihren Besitzern ausgestattet worden waren, wurden nicht verwendet. Sie dienten nur zur Kommunikation mit dem Treuhänder.

Tabelle 9.5: Zuordnungen in Situation 1

IF	MF	Zuordnungs-wahrschein-lichkeit	99% + eindeutig	Dist.
1	2	3	4	5
58	132376	0.9938	*	22
1322	149243	0.9714		191
1991	130034	0.9769		47
2095	64527	0.9928	*	11
2200	59772	0.9995		490
	66964	0.9999		390
	102403	0.9844		520
	146750	0.9999		463
2637	61419	0.9783		235
2649	68506	0.9987		419
	97754	0.9985		430
	158558	0.9999		190
3052	17936	0.9998	*	16
3606	155415	0.9961	*	152
4512	75606	0.9995	*	0
6625	153005	0.9572		28
6713	30406	0.9949	*	31
6848	23488	0.9938		3347
	45469	0.9957		3345
	47136	0.9752		3310
	106486	0.9924		3297
	132581	0.9945		3487
7034	64094	1		14
	74724	1		14
7452	24666	0.9692		390
		Summe	6	

Ansonsten überraschen vor allem zwei Eigenschaften des Verfahrens, die aus Tabelle 9.5 deutlich werden. Die letzte Spalte der Tabelle zeigt, daß Zuordnungen mit hohen Wahrscheinlichkeitswerten selbst bei extrem hohen Distanzen stattfinden. Hier scheint eine Fehlfunktion des Verfahrens vorzuliegen. In engem Zusammenhang damit steht die zweite Eigentümlichkeit. Durch das Verfahren wird nämlich eine größere Anzahl von mehrdeutigen Zuordnungen hergestellt. Wie bereits bei der Behandlung der Methode in Abschnitt 3.4.1.6 dargestellt, ist durch die Konstruktion des Verfahrens nicht prinzipiell ausgeschlossen, daß ein Fall des Identifikationsfile mehreren Fällen des Mikrodatenfile zugeordnet wird; lediglich die umgekehrte Relation kann nicht vor-

kommen. Derart viele mehrdeutige Zuordnungen waren jedoch nicht erwartet worden. Dies kann jedoch durch Rekurs auf die erste Eigentümlichkeit des Verfahrens erklärt werden: Wenn Zuordnungen auch bei großen Distanzen mit hohen Wahrscheinlichkeiten vorgenommen werden, liegen entsprechend viele Fälle noch innerhalb des kritischen Bereichs und es kommt zu mehrdeutigen Zuordnungen, die demgemäß vor allem bei unplausibel hohen Distanzen auftreten. Bevor dieses Problem interpretiert wird, sollen die Ergebnisse der anderen Experimente im Überblick dargestellt werden. Dies erfolgt in Tabelle 9.6 Eine detaillierte Darstellung der einzelnen Ergebnisse der Experimente 2 - 4 kann den Tabellen 9A2 bis 9A4 im Tabellenanhang zu Abschnitt 9.2 entnommen werden. Auf eine entsprechende Darstellung des Experiments 5 wurde aus Raumgründen verzichtet, da in diesem Fall besonders viele Mehrfachzuordnungen auftraten.

Tabelle 9.6: Zuordnungen mit der diskriminanzanalytischen Reidentifikationstechnik

Teil-IF Nr.	1		8		
Situation	1	2	3	4	5
angenommener Fehler	mittel	hoch	mittel	niedrig	mittel
Nummer des Fehlerprozesses	1	2	1	3	4
Anzahl eindeutig zugeordneter Fälle (Wahrscheinlichkeit größer 95%)	6	4	11	9	11
Anzahl mehrdeutig zugeordneter Fälle (Mikrodatenfile)	19	21	8	12	57

Tabelle 9.6 zeigt erneut, daß bei den fünf Reidentifikationsexperimenten mit der diskriminanzanalytischen Reidentifikationstechnik und dem Mikrozensus bzw. "Kürschners Deutschem Gelehrtenkalender" als Datenbasis zwei Randbedingungen variiert wurden, die zum einen den Fehlerprozeß, zum anderen das verwendete Teil-Identifikationsfile betrafen. Die angegebenen Nummern der Teil-Identifikationsfiles (1 und 8) beziehen sich auf die entsprechenden Files

beim einfachen Zuordnungsverfahren (vgl. Abschnitt 3.2).[24] Die Experimente erbringen jeweils deutlich mehr eindeutige Zuordnungen als mit einfachen Sortierverfahren hergestellt werden können.[25] Das Verfahren scheint deshalb die hochgesteckten Erwartungen zu erfüllen. Wenn es einem Angreifer gelänge, in einem Durchgang elf Personen zu identifizieren - dies ist die Anzahl an Zuordnungen in Situation 3 - so könnte er dies zumindest als Teilerfolg betrachten. Er könnte sich seiner Sache relativ sicher sein, da er eine hohe Sicherheitsschwelle gewählt hat.

9.2.3.3. Die Überprüfung der Ergebnisse durch den Treuhänder

Die Überprüfung der Ergebnisse durch den Treuhänder wirkt desillusionierend (vgl. Tabelle 9.7). Bei den ersten beiden Experimenten (mit dem Teil-Identifikationsfile Nr. 1) ist von den mit mindestens 99 Prozent Wahrscheinlichkeit vorgenommenen Zuordnungen keine einzige richtig, die Quote der "Treffer" also Null. Bei den drei Experimenten, für die das Teil-Identifikationsfile Nr. 8 als Datenbasis verwendet worden war, werden jeweils zwei richtige Treffer festgestellt. Dabei handelt es sich allerdings um insgesamt nur drei verschiedene korrekte Paare von Datensätzen, die z.T. in mehreren Experimenten gefunden werden. Alle drei waren bereits mit einfachen Zuordnungsverfahren festgestellt worden, weil die betreffenden Merkmalsausprägungen in dem Teil-Identifikationsfile und im Mikrodatenfile jeweils identisch abgebildet sind. Die empirische "Trefferrate"[26], d.h. der Anteil der korrekten Zuordnungen an allen eindeutigen beträgt 18 Prozent (Situationen 3 und 5), bzw. 22 Prozent (Situation 4). Mit anderen Worten: Der weitaus größte Teil der Zuordnungen ist falsch.

---

24 Genaue Angaben über den Fehlerprozeß sind Tabelle 9.4 zu entnehmen.

25 Zur Erinnerung: Die 14 eindeutigen Zuordnungen bei Nutzung eines einfachen Zuordnungsverfahrens kamen erst durch die Überprüfung aller 16 unterschiedlichen Identifikationsfile-Versionen zustande.

26 Zu verschiedenen Definitionen von Fehlerraten oder "Fehlklassifikationswahrscheinlichkeiten" vgl. Fahrmeir/Häusler/Tutz (1984:304).

Tabelle 9.7: Überprüfung der experimentell ermittelten Zuordnungen

Teil-IF Nr.	1			8	
Situation	1	2	3	4	5
angenommener Fehler	mittel	hoch	mittel	niedrig	mittel
Nummer des Fehlerprozesses	1	2	1	3	4
Anzahl eindeutiger Zuordnungen	6	4	11	9	11
davon richtig	-	-	2	2	2
Trefferrate in %	0	0	18	22	18
zum Vergleich: eindeutige Zuordnungen Abgleichverfahren	1			7	
davon richtig Abgleichverfahren	-			3	

Wie die richtig ermittelten Fälle, werden auch die falsch zugeordneten teilweise in mehreren Situationen gefunden. Deshalb ergiebt sich insgesamt eine Zahl von 29 zugeordneten Fällen über alle fünf Situationen hinweg, von denen - wie erwähnt - drei korrekt sind. Dies ergibt eine "Gesamttrefferrate" von 10,3 Prozent. Ein anderer, ebenfalls zweckmäßiger Weg, eine solche "Gesamtrate" zu ermitteln, ist einfach die Durchschnittsbildung über die Trefferraten der einzelnen Situationen, wie sie in Tabelle 9.7 eingetragen sind, hinweg. Danach ergibt sich ein etwas höherer Wert von 11,6 Prozent, der immer noch deutlich niedriger ist, als der Wert von 29 Prozent, der mit Abgleichtechniken erreicht worden war. Unabhängig davon, wie man die Trefferquote berechnet, bleibt die Tatsache bestehen, daß mit der diskriminanzanalytischen Reidentifikationstechnik keine nicht schon mit dem einfachen Abgleichverfahren ermittelten Zuordnungen vorgenommen werden können obwohl dies potentiell möglich gewesen wäre, da - wie oben ausgeführt - insgesamt 53 Fälle sowohl in Kürschners Deutschem Gelehrtenkalender als auch im Mikrozensus enthalten sind. Die Differenz von vier (einfaches Zuordnungsverfahren) zu drei

(diskriminanzanalytische Reidentifikationstechnik) eindeutigen korrekten Zuordnungen ist darauf zurückzuführen, daß bei der einfachen Abgleichtechnik weitere Versionen des Gelehrtenkalenders überprüft wurden. Hieraus ergibt sich auch die mit 0.0004 (bezogen auf das Identifikationsfile) etwas niedrigere Quote korrekter Zuordnungen der diskriminanzanalytischen Reidentifikationstechnik. Damit erweist sich dieses Verfahren - entgegen den ursprünglichen Erwartungen - als nicht leistungsfähiger als die einfache Reidentifikationstechnik. Aus der Perspektive eines Angreifers sind die Ergebnisse sogar schlechter. Denn was sich zunächst als Vorzug dieser Methode darstellt, nämlich die größere Anzahl der Zuordnungen, entpuppt sich jetzt als Nachteil, da lediglich der Anteil der Falschzuordnungen ansteigt. Drückt die vom Angreifer gewählte Sicherheitsschwelle von 99 Prozent seine Risikoaversion tatsächlich auch nur annähernd adäquat aus, wird er sein Verfahren als nicht zweckgemäß einstufen. Wie bei dem Abgleichverfahren bieten ihm die Ergebnisse keine Möglichkeit zwischen korrekten und falschen Zuordnungen zu unterscheiden. Trotz der ausgewiesen hohen Sicherheitsschwelle von 99 Prozent hat er keinen Anhaltspunkt bei welchen der Zuordnungen es sich um ein und dieselbe Person handelt.

9.2.3.4 Erklärung der erhaltenen Ergebnisse (I)

Ein bemerkenswertes Teilresultat der Experimente ist, daß die eigentlich typische Konstellation, für die die Reidentifikationstechnik konstruiert worden war, nicht zum Tragen kommt. Diese Konstellation war dadurch gekennzeichnet, daß sich - bedingt durch kleine Inkompatibilitäten - die Variablenwerte einer Person im Mikrodatenfile und im Identifikationsfile geringfügig voneinander unterscheiden. Mit der Methode werden jedoch nur Fälle korrekt zugeordnet, die eine Distanz von Null, also keine Inkompatibilitäten, aufweisen und demgemäß bereits beim einfachen Abgleich auf identische Ausprägungen zugeordnet worden waren.

Umgekehrt werden in keiner Situation alle drei mit der Sortiermethode korrekt ermittelten Fälle zugeordnet. In jeder der Situationen 3 bis 5 lag jeweils ein Fall knapp unter der Wahrscheinlichkeit von 99 Prozent. Dies kann erklärt werden, indem an die Funktionsweise der Reidentifikationstechnik erinnert wird. Bei der Spezifikation des Fehlerprozesses findet implizit eine Bewertung der Dimensionen des Merkmalsraums statt, in dem die Fälle in größerer oder geringerer Entfernung (gemessen an der Mahalanobisdistanz) liegen. Unterschiede in den einzelnen Variablen werden - je nach ihrer Fehlerhäufigkeit - als mehr oder

weniger bedeutsam eingestuft. So hängt es von der konkreten Spezifikation des Fehlerprozesses ab, ob zwei identische Fälle nah oder fern von anderen liegen, mit denen möglicherweise eine Verwechslung stattfinden könnte. Der Ausschluß von Fällen, bei denen Verwechslungen wahrscheinlich sind, ist eine vernünftige Strategie, die bei der vorliegenden Datenlage jedoch in einigen Fällen zum Ausschluß der richtigen Datensätze führt.

Bereits vor der Überprüfung der Ergebnisse durch den Treuhänder war klar geworden, daß die Anwendung des Verfahrens mit den vorliegenden Daten auf Probleme stößt. Es wurde schon erwähnt, daß in allen Situationen mehrdeutige Zuordnungen bei relativ hohen Distanzen auftreten. Um die Ursache dieses Phänomens verstehen zu können, muß erneut kurz auf die Funktionsweise der Reidentifikationstechnik eingegangen werden. Für diese ist die Bayessche Formel zentral (vgl. Abschnitt 3.4.1.1), weil mit ihr die Zuordnungswahrscheinlichkeiten berechnet werden. Etwas umgeformt, hat der Ausdruck folgende Gestalt:

$$(1) \qquad p(I|y_i) = \frac{p(y_i|I)\, p(I)}{p(y_i)}$$

Dabei sind:

y: Merkmalsausprägungen eines Falles im Mikrodatenfile
I: Ein bestimmter Fall im Identifikationsfile
p: Bedingte bzw. unbedingte Wahrscheinlichkeiten

Die Formel unterscheidet sich insofern von Formel (2) in Abschnitt 3.4, als in ihr entsprechend der hier angewandten Suchstrategie Massenfischzug die Rollen von Mikrodatenfile und Identifikationsfile vertauscht sind. Außerdem läuft i von 1 bis n, wobei n hier die Zahl der Fälle bezeichnet, die aus dem Mikrodatenfile für die Experimente herausgezogen worden waren.

Im Zähler von Ausdruck (1) steht eine bedingte Wahrscheinlichkeit. Sie gibt an, wie wahrscheinlich es ist, daß der betreffende Fall des Mikrodatenfile ein 'Abkömmling' des aktuell betrachteten Falles des Identifikationsfile ist. Dieser Wert wird aus der Distanz der beiden Fälle in Identifikationsfile und Mikrodatenfile berechnet. Im Nenner steht die unbedingte Wahrscheinlichkeit für das Auf-

treten des jeweiligen Falls im Identifikationsfile. Die Bedeutung des Nennerausdrucks kann intuitiv verstanden werden, wenn man sich verdeutlicht, daß Verwechslungen mit Doppelgängern um so unwahrscheinlicher sind, je unwahrscheinlicher die Ausprägungskombination des betreffenden Falles ist.

Die Berechnung des Zählerausdrucks führt in den fraglichen Fällen zu keinen ungewöhnlichen Ergebnissen. Die Schätzung der dazugehörigen unbedingten Wahrscheinlichkeit, die in der Bayesschen Formel im Nenner steht, ist jedoch in irgendeiner Weise so nach unten hin verzerrt, daß der Gesamtausdruck für die Zuordnungswahrscheinlichkeit, wie er sich aus der Bayesschen Formel ergibt, unplausibel hoch ausfällt.

Die Ursachen für dieses spezielle Ergebnis sind darin zu suchen, daß im Mikrozensus trotz der vorgenommenen Eingrenzung Fälle auftreten, die außerhalb der speziellen Subpopulation der "Gelehrten" liegen, die zur Basis der Wahrscheinlichkeitsabschätzung $p(y_i)$ verwendet wurde. Derartige Fälle weisen dann Ausprägungskombinationen auf, die nach der Logik dieser Abschätzung so 'unmöglich' sind, daß für sie derart verschwindend geringe Wahrscheinlichkeiten $p(y_i)$ errechnet werden, daß sich in der weiteren Konsequenz derart hohe Werte für den gesamten Ausdruck, d.h. für $p(l|y_i)$, ergeben. Die Zuordnungen bei hohen Distanzen sind demnach nicht Konsequenz einer Fehlfunktion des Verfahrens, sondern eine Folge dessen, daß sich die Erhebungen, denen Identifikationsfile und Mikrodatenfile zugrunde liegen, auf unterschiedlich definierte Populationen beziehen.

Diese Interpretation der Zuordnungen bei hohen Distanzen kann sich auf eine visuelle Inspektion eines Teils dieser 'perversen' Paare stützen. Neben Fällen, deren Randstellung in der Population nicht ohne statistische Abschätzung festzustellen ist, finden sich auch Extrembeispiele wie ein selbständiger Theologe. Die zugeordneten Fälle weisen oft nur eine Minderheit an übereinstimmenden Merkmalen auf.

Alle Fälle, für die hohe Distanzen festzustellen sind, und die dennoch zugeordnet werden, können gemäß der vorgetragenen Überlegung ohne Schaden aus der Betrachtung ausgeschlossen werden. Allerdings stellt sich die Frage dann erneut, welche Distanzen als 'hoch' einzustufen sind. So kann es umgekehrt bei dem ohnehin verwendeten Kriterium mehrdeutige Zuordung bleiben. Diese werden nicht als "Reidentifikation" gezählt. Eindeutige Zuordnungen mit Distanzen größer als 1000 sind nur bei der Situation 5 betroffen, wo drei Fälle auftreten. Wertet ein Angreifer derartige Paare nicht als Reidentifikationen, sinkt

die Gesamtzahl der eindeutigen Zuordnungen entsprechend und die Rate der korrekten 'Treffer' steigt geringfügig an.

Die beschriebenen Probleme sind Ausdruck dessen, daß die von Paaß und Wauschkuhn (1985:137ff) bei den modellmäßigen Überlegungen zur Durchführung der Dichteschätzung angenommene Äquivalenz von Mikrodatenfile und Identifikationsfile in der Praxis nicht bei jeder Datengrundlage gegeben ist. Welche weiteren Ursachen kann es für die niedrige Trefferrate geben? Folgende Möglichkeiten sind in Betracht zu ziehen:

1) Das Verfahren funktioniert nicht richtig, weil bei seiner Implementation Fehler gemacht wurden, die auf Grund der hohen Komplexität des Programms nicht endeckt wurden.

2) Das Verfahren ist nicht zum Einsatz bei der vorliegenden Datenstruktur geeignet, z. B. weil es auf der linearen Diskriminanzanalyse beruht und diese bei kategorialen Variablen (wie im vorliegenden Fall) zwar robust reagiert, in erster Linie jedoch für metrische Variablen vorgesehen ist.

3) Die beiden Datenquellen, also Mikrozensus und Gelehrtenkalender, unterscheiden sich in solchem Grade, daß Abschätzungen der Wahrscheinlichkeitsverteilungen (bzw. -dichten), wie sie bei einer Anwendung der diskriminanzanalytischen Reidentifikationstechnik vorgenommen werden, nicht von einem Datenbestand auf den anderen übertragen werden können. Dies könnte wegen völlig unterschiedlicher Definitionen und Realitätsbezügen der Daten der Fall sein.

4) Die bei der Anwendung des Verfahrens vorgenommenen Dichteschätzungen sind falsch, weil sie am Gelehrtenkalender vorgenommen werden mußten, der keine Zufallsstichprobe ist.

5) Der Fehlerprozeß wurde völlig falsch spezifiziert, weil die in der Realität auftretenden Inkompatibilitäten von unbekannter und unerwarteter Art sind.

6) Das Vorliegen von Alternativausprägungen bei einem Teil der Überschneidungsmerkmale verhindert Zuordnungen, weil für die Untersuchung die "falschen" Alternativen bei den Variablen mit doppelten Ausprägungen gewählt wurden. Im Extremfall wiesen nur die gefundenen drei Fälle die

entsprechenden Ausprägungen auf, die anderen 50 der Schnittmenge zwischen Gelehrtenkalender und Mikrozensus gehörten einem der Teil-Identifikationsfiles an, die nicht untersucht wurden.

Die Erklärung Nr. 6 kann von vornherein ausgeschieden werden. Mit ihr sind die vielen Falschzuordnungen nicht zu erklären. Außerdem wurde zu Beginn ausgeführt, daß 42 Prozent der Fälle hinsichtlich der nicht durch entsprechende Teil-Identifikationsfiles abgedeckten drei Variablen nur fehlende Werte aufweisen. Somit ist extrem unwahrscheinlich, daß lediglich die drei gefundenen Fälle (oder auch nur einige wenige mehr) in den untersuchten Teil-Identifikationsfiles enthalten waren. Ein weiterer Einwand gegen diese Erklärung ist, daß zusätzlich versucht wurde, das Vorliegen von Alternativvariablen über die Modellierung des Fehlerprozesses zu berücksichtigen.

Damit verbleiben fünf Erklärungsmöglichkeiten. Um den Bereich der Erklärungen stärker einzuengen, wird ein zusätzliches Experiment durchgeführt, dessen Design enger an das der GMD-Simulationen angelehnt wird. Auf diese Weise können Zweifel am korrekten Funktionieren des GMD-Programms entweder bestätigt oder ausgeräumt werden.

9.2.3.5 Vergleich mit den Ergebnissen einer Testrechnung bei Verwendung eines synthetischen Identifikationsfile

Wie bei den vorher durchgeführten Experimenten wird Kürschners Deutscher Gelehrtenkalender als Zieldatei verwendet. Aus diesem File wird auf eine zum Vorgehen der GMD analoge Weise ein zweiter Datenbestand synthetisch erzeugt. Die Annahmen über den Fehlerprozeß sind die gleichen wie in der Situation 5 (vgl. Tabelle 9.4, Fehlerprozeß 4). Es wird angenommen, daß der Angreifer diese Fehlerverteilung kennt[27] und demgemäß die gleiche Verteilung als Fehlerprozeß spezifiziert. Wiederum in Analogie zur Vorgehensweise von Paaß/Wauschkuhn werden 500 Fälle aus dem synthetischen Identifikationsfile zufällig ausgewählt und für Zuordnungsversuche verwendet.

---

27 Die GMD-Experimente zeigten, daß das Verfahren kaum andere Ergebnisse erbringt, wenn angenommen wird, daß der Angreifer die Fehlerverteilungen nicht kennt (vgl. Paaß/Wauschkuhn 1985:184).

Tabelle 9.8: Zuordnungen in der Testkonstellation: Mikrodatenfile und synthetisches Identifikationsfile aus den Daten von "Kürschners Deutschem Gelehrtenkalender" erzeugt, bei einer Stichprobe des Identifikationsfile von 500:

Zuordnungen im MF (p > 99%)	absolut	in %
	145	100.0
falsch	22	15.2
richtig	123	84.8
Nur korrekte Zuordnungen davon:	123	100.0
Merkmalsausprägungen:		
identisch	71	57.7
abweichend	52	42.2

Tabelle 9.8 zeigt das Ergebnis dieses Experiments. Es fällt wiederum völlig überraschend aus. Die diskriminanzanalytische Reidentifikationstechnik kann unter der gegebenen Konstellation derart 'erfolgreich' angewendet werden, daß das Ergebnis unter anderen Bedingungen, wenn keine Experimente mit zwei verschiedenen empirischen Datenquellen durchgeführt worden wären, möglicherweise als Beleg dafür gewertet worden wäre, daß ein Massenfischzug zwar technisch schwierig und aufwendig, aber letztlich doch erfolgreich durchführbar ist.

Mit der angenommenen Sicherheitsschwelle (p) von 99 Prozent werden 145 von 500 Datensätzen zugeordnet. Dies ist ein Anteil von 29 Prozent. Von diesen zugeordneten Datensätzen wird der überwiegende Teil (rund 85 Prozent) dem richtigen Partnerdatensatz zugewiesen. In der Testrechnung findet sich demnach ein deutlich höherer Anteil an Falschzuordnungen als bei den GMD-Experimenten, wo der entsprechende Wert ca. 1 Prozent betragen hatte. Unter der vorliegenden Datenkonstellation wäre die Methode demnach nur anwendbar von Angreifern, die mit einer relativ niedrigen Sicherheitsschwelle (z. B. 80 Prozent) auskämen. Allerdings ist diese Überlegung rein hypothetisch, da bei Verwendung einer eigenen, unabhängigen Datenquelle für das Identifikationsfile die bekannten, viel ungünstigeren Ergebnisse folgen.

Trotz des - verglichen mit den Simulationen der GMD - höheren Anteils an Falschzuordnungen zeigen die Ergebnisse der Testrechnung die beträchtliche Leistungsfähigkeit der diskriminanzanalytischen Reidentifikationstechnik. Obwohl Kürschners Deutscher Gelehrtenkalender eine völlig andere Variablenstruktur aufweist als die Einkommens- und Verbrauchsstichprobe, für die die Methode entwickelt worden war, wird ein hoher Anteil an korrekten Zuordnungen erreicht. Damit wird die 'Robustheit' und Leistungsfähigkeit des Verfahrens in eindeutiger Weise demonstriert. Entscheidend ist auch, daß Zuordnungen nicht nur vorgenommen werden, wenn die Fälle jeweils identische Merkmalsausprägungen aufweisen, sondern auch dann, wenn durch die Fehlerüberlagerung Änderungen in den Variablenwerten erzeugt worden sind. Die Methode verhält sich auch in dieser Hinsicht den Erwartungen entsprechend.

Der relativ hohe Anteil an (korrekten und falschen) Zuordnungen ist auf den (im Sinne der mathematischen Informationstheorie) relativ hohen Informationsgehalt der Überschneidungsmerkmale zurückzuführen. Das Ergebnis verdeutlicht noch einmal, daß die zugrunde gelegte Datenbasis Datenangriffe ganz besonders begünstigt.

9.2.3.6 Erklärung der erhaltenen Ergebnisse (II)

Oben waren sechs Hypothesen als Erklärungen für die (vom Standpunkt des Angreifers) 'schlechten' Ergebnisse bei Nutzung von Mikrozensus und Gelehrtenkalender als Mikrodatenfile und Identifikationsfile angeführt worden. Mit den Ergebnissen der Testrechnung können zwei davon definitv ausgeschieden werden.

Die Implementation der diskriminanzanalytischen Reidentifikationstechnik ist - anders, als in Erklärung Nr. 1 angenommen wurde - offensichtlich nicht mißglückt.

Ebensowenig können Probleme mit dem Skalenniveau der Überschneidungsmerkmale (Erklärung Nr. 2) herangezogen werden. Die Testrechnung zeigt, daß das Verfahren vergleichsweise wenig dadurch gestört wird, daß die meisten Variablen nur ordinales bzw. nominales Skalenniveau aufweisen und nicht auf metrischem Niveau gemessen wurden wie in den Experimenten der GMD. Das Verfahren ist unter technischem Aspekt ein eindrucksvolles Beispiel für die Leistungsfähigkeit der statistischen Wissenschaft.

Zur Erklärung dieser Ergebnisse verbleiben somit nur die Hypothesen drei bis fünf, die ohnehin im Zusammenhang zu sehen sind. Nach Erklärung Nr. 3 scheitert ein Reidentifikationsversuch, weil die Abschätzung von Wahrscheinlichkeitsverteilungen (bzw. -dichten) nicht genau genug von dem einen Datenbestand auf den anderen übertragen werden kann. Die empirischen Ergebnisse zeigen, daß dies zumindest für jene Fälle des Mikrozensus gilt, die außerhalb der durch den Gelehrtenkalender definierten Subpopulation liegen, denn sie dürfte es nach der Logik des Verfahrens 'gar nicht geben'. Darüber hinaus ist zu erwarten, daß beim Vorliegen von starken Inkompatibilitäten auch die Abschätzung der für die Bayessche Formel benötigten Dichten leidet. Dies könnte die Falschzuordnungen teilweise erklären.

Als andere wichtige Erklärung (Nr. 5) ist anzuführen, daß den Inkompatibilitäten zwischen den Datenquellen eine wichtigere Rolle zukommt als a priori anzunehmen war. Dabei spielt nicht die Höhe der Fehler die entscheidende Rolle. Diese wird in den Experimenten variiert, ohne daß die Ergebnisse jeweils grundsätzlich anders ausfallen.

Maßgeblich für das erhaltene Resultat ist vielmehr die Struktur der Fehler. Diese folgt nicht den einfachen Verteilungen, wie bei den Experimenten für die Fehlerprozesse angenommen. Die Analyse von Dateninkompatibilitäten (vgl. Abschnitt 4.3), die sich auf sozialwissenschaftliche Arbeiten zur Reliabilität, zu unterschiedlichen Erhebungssituationen, zu Kontexteffekten etc. stützen konnte, zeigte vielmehr, daß derartige Abweichungen zwischen zwei Datenbeständen nicht regellos erfolgen, sondern durch ein kompliziertes Wirkungsgeflecht bestimmt sind. Die Häufigkeit von spezifischen Abweichungen hängt zusätzlich von Eigenschaften der befragten Personen ab.

Die konkrete Beschaffenheit der Inkompatibilitäten ist jedoch unbekannt. Es existiert keine geschlossene soziologische oder psychologische Theorie, aus der sie ohne zusätzliche empirische Vermittlungsglieder abgeleitet werden könnte. Die Häufigkeit und Struktur der Inkompatibilitäten müßte bei jedem Datenangriff von neuem experimentell für die beiden beteiligten Erhebungen bestimmt werden. Ein solches Experiment ist jedoch für einen Datenangreifer undurchführbar: Er hat keine Möglichkeit, eine amtliche Befragung zu simulieren.

Vor der Durchführung der Experimente waren die Auswirkungen der Abweichungen der Fehler von den angenommenen einfachen Verteilungen unbekannt. Es konnte zumindest nicht ausgeschlossen werden, daß sie vernachlässigbar sind. Jetzt zeigt sich jedoch, daß sie die zentrale Ursache dafür sind,

daß die diskriminanzanalytische Reidentifikationstechnik bei einer realistischen Untersuchungsanlage nicht die in dieses Verfahren gesetzten Erwartungen erfüllt.

9.2.4 Folgerungen für die faktische Anonymität von Daten

Das Ergebnis der Tests der diskriminanzanalytischen Reidentifikationstechnik mit empirischen Daten ist, daß diese unter den getesteten Bedingungen keine Gefährdung für die Anonymität amtlicher Mikrodaten darstellt. Bei einem Datenangriff sind die wenigen korrekten Zuordnungen nicht von den viel häufiger auftretenden Falschzuordnungen zu unterscheiden. In den durchgeführten Experimenten konnten 50 (94 Prozent) jener 53 Fälle, die in beiden Files enthalten sind, nicht zugeordet werden. Die in der Literatur (vgl. Bethlehem et al. 1990 und Dittrich/Schlörer 1987) vielfach mit der diskriminanzanalytischen Reidentifikationstechnik verbundene Erwartung, sie wäre ein ideales Instrument in der Hand eines Angreifers, kann nicht bestätigt werden.

Vom Standpunkt eines hypothetischen Angreifers fiel das Resultat insgesamt ungünstiger aus als bei der Anwendung des Abgleichverfahrens. Dort hatte der Anteil der richtigen an allen Zuordnungen 29 Prozent betragen, hier beträgt er, gemittelt über alle fünf relevanten Experimente, rund 10 Prozent.

Das Ergebnis kann also die viel höheren Kosten der diskriminanzanalytischen Reidentifikationstechnik nicht rechtfertigen. Im vorliegenden Fall, bei dem auf den Ergebnissen der GMD aufgebaut werden konnte, war dennoch rund ein Jahr Arbeit erforderlich, um den Algorithmus zu rekonstruieren und die notwendigen Anpassungen an dem Programm vorzunehmen, die die Implementation auf dem vorhandenen Rechner und die Umstellung an die geänderte Datenstruktur erforderlich machte. Erst dann konnte mit der Generierung der Fehlerverteilungen und der Durchführung der eigentlichen Experimente begonnen werden. Die Rechenkosten waren sehr hoch, da der Algorithmus der diskriminanzanalytischen Reidentifikationstechnik sehr viel CPU-Zeit und Speicherplatz erfordert. Die Gesamtkosten belaufen sich auf insgesamt 266 000 DM, wie Tabelle 9.9 entnommen werden kann. Es sind jeweils Marktpreise und Selbstkosten angesetzt, die jedoch eher niedrig geschätzt sind.

Tabelle 9.9: Kosten der Anwendung der diskriminanzanalytischen Reidentifikationstechnik (gerundete Werte)

Aufnahme und Rekodierung der Datenbestände	45 000 DM
Lohnkosten für Anpassung des Verfahrens und für die Durchführung der Experimente	81 000 DM
Marktpreis der Rechenkosten	135 000 DM
Gesamtkosten	261 000 DM

Die enormen Kosten der Reidentifikationstechnik lassen es als ausgeschlossen erscheinen, daß - unter den zugrunde gelegten Annahmen des sozialwissenschaftlichen Szenarios - ein Wissenschaftler versuchen könnte, das Verfahren für einen tatsächlichen Datenangriff zu mißbrauchen. Seine Kosten lägen wesentlich höher als angegeben, da das Programm der GMD nicht zur Verfügung stünde und die Methode erst neu entwickelt werden müßte.

Die Ausgangsfragestellung der für das ganze Experiment grundlegenden Deanonymisierungssituation "Gewinnung einer erweiterten Datengrundlage für sozialwissenschaftliche Untersuchungen" näher zu prüfen, ist überflüssig, da die Quote der sicheren Reidentifikationen bei Null lag. Soweit es die angewandte Methode anbetrifft und Unkenntnis der Teilnahme an der Erhebung unterstellt wird, können Mikrozensusdaten mit den zehn zu berücksichtigten Merkmalen, also einschließlich der Regionalmerkmale, als faktisch anonym gelten.

9.2.5 Ausblick

Die Experimente mit der diskriminanzanalytischen Reidentifikationstechnik wurden unternommen, um das leistungsfähigste Verfahren, das überhaupt zur Verfügung steht, zu testen. Allerdings liegt ein Einwand zur Relativierung nahe: Zum heutigen Zeitpunkt mag diese Methode die am weitesten entwickelte sein. In der Statistik, numerischen Mathematik und Informatik vollziehen sich jedoch

Neuentwicklungen mit derart hoher Geschwindigkeit, daß bereits in naher Zukunft eine Methode - unter Umständen gestützt auf noch leistungsfähigere Computer - verfügbar sein könnte, die das adäquate Instrument für den Angreifer darstellt. Aus diesem Grund sei kurz auf eine Entwicklungsperspektive eingegangen, die sich in den letzten Jahren abgezeichnet hat.

In der Informatik werden Probleme, die dem vorliegenden eng verwandt sind, unter dem Titel "Mustererkennung" behandelt (Niemann 1981, 1983, vgl. die Beiträge in Devijver/Kittler 1987, insbesondere Jain 1987). Dabei ist jeweils das Ziel, aus gegebener Information, die transformiert oder gestört sein kann, bestimmte "Muster" zu erkennen. Im vorliegenden Fall werden die zu erkennenden "Muster" durch die realen Merkmale von Personen konstituiert. Die Information, aus der sie erkannt werden sollen, besteht aus den gegebenen Datensätzen.

In typischen Mustererkennungsproblemen ist eine vergleichsweise kleine Anzahl von Mustern zu identifizieren. Eine gewöhnlich viel größere Anzahl von "Bildern" soll diesen Mustern zugeordnet werden. Dazu wurden spezialisierte Verfahren entwickelt. Für diese ist zumeist kennzeichnend, daß vor der eigentlichen Anwendungsphase eine Lernphase stattfindet, in der der betreffende Algorithmus trainiert wird, indem ihm bei der Präsentation eines Bildes jeweils mitgeteilt wird, zu welchem Muster es gehört. Geschieht dies oft genug, können in der Anwendungsphase - aufgrund einer eingebauten Fähigkeit zur Generalisierung - Bilder mit unbekannter Zugehörigkeit zu spezifischen Mustern klassifiziert werden (vgl. Jain 1987:9).

Diese Problemstellung entspricht in ihrem Kern jener, für die die Diskriminanzanalyse entwickelt wurde (vgl. Young/Calvert 1974). Aus diesem Grunde wurden verschiedene Varianten der Diskriminanzanalyse für Mustererkennungsprobleme verwendet. Betrachtet man das durch die Reidentifikationsabsicht entstehende Zuordnungsproblem von einem etwas abstrakteren Standpunkt, ist das von Paaß/Wauschkuhn entwickelte Verfahren nur ein besonders komplexer Anwendungsfall der genannten größeren Klasse von Mustererkennungsverfahren.

Nun sind jedoch in jüngerer Zeit Ansätze bekannt geworden, die im Hinblick auf die gleiche Aufgabenstellung eine wesentlich größere Leistungsfähigkeit versprechen als die herkömmlichen Algorithmen. Diese Ansätze werden gewöhnlich unter dem Titel "Neuronale Netze" abgehandelt, weil sie beanspruchen, Verschaltungen von Neuronen in biologischen Hirnen nachzuahmen. Dies soll entweder durch eine entsprechend gestaltete Computerarchitektur, die völlig

von herkömmlichen Entwürfen abweicht, oder durch geeignete Softwaresimulationen geschehen. Versprochen werden Computer, die "Denken nach Menschenart" (Titel eines Artikels in "Bild der Wissenschaft", 1989) oder wenigstens bestimmte von der Informatik bisher nicht zufriedenstellend beherrschte Leistungen anbieten - eben u. a. die Erkennung von Mustern (vgl. Bridle 1987 und Fougelman/Soulie et al 1987).

Auf dem Gebiet der Neuronalen Netze wird international in großem Maßstab geforscht.[28] Einzelne Ergebnisse sind bereits in die mit hohen Auflagen vertriebenen populären Computerzeitschriften eingedrungen (vgl. z.B. Computer Persönlich 1991/11:116ff.), Computerwoche 1991/12:50ff.), "c't" 1991/5:182ff.). Programme zur Simulation solcher Netze sind nicht nur von kommerziellen Herstellern, sondern auch als Toolsammlungen (Kruse et al. 1991), als Shareware oder als Diskettenbeilagen zu Büchern erhältlich (so bei McClelland/Rumelhart 1988, Schöneburg et al. 1990).

Kann die obige Argumentation aufrechterhalten werden, nach der ein Programm von der Leistungsfähigkeit des AIMIPH-Programms eines großen Entwicklungsaufwands bedarf, wenn möglicherweise ähnlich leistungsfähige Methoden schlicht über den Buchhandel bezogen werden können? Liegt hierin nicht ein großes Gefahrenpotential für die Sicherheit amtlicher Daten?

Um beurteilen zu können, ob eine Gefahr droht, wird die Beratung durch Experten gesucht und einige Tests mit einem der genannten Software-Werkzeuge durchgeführt. Zwei spezielle Algorithmen, die in einem auf DOS-PCs einsetzbaren Programmpaket von McClelland und Rumelhart implementiert sind, werden getestet. Verwendet werden der "Interactive Activation and

---

28 Dies ist nicht zuletzt auf die erhebliche militärische Bedeutung der von den Finanziers erwarteten Ergebnisse zurückzuführen. Viele Artikel enthalten Dankadressen an militärische Institutionen (vgl. z.B. die Beiträge in Grossberg 1988). Neben militärischen Anwendungen, die z.B. die Steuerung von Raketen gemäß einer zu erkennenden Landschaft betreffen, sind oft genannte Entwicklungsziele (vgl. Schöneburg et al. 1990) die Auswahl zuverlässiger Kreditnehmer, die Aktienkursprognose und die Zuordnung von Labormeßwerten zu bestimmten Typen.
Am anderen Ende des Spektrums zwischen Anwendungs- und Grundlagenforschung stehen Arbeiten, in denen die statistisch-mathematischen Eigenschaften der verwendeten Algorithmen untersucht werden. Ein Teil der Verfahren weist enge Verwandtschaftsbeziehungen zu loglinearen Modellen auf (Paaß 1989a), da Entropieoptimierungsmethoden verwendet werden (Blien/Graef 1991, Paaß 1989b und allgemein Wilson 1970), bzw. läßt sich unter die Modellklasse subsumieren, für die latente Variablen kennzeichnend sind (White 1989).

Competition Algorithm" und die sogenannte Boltzmann-Maschine[29] für einige einfache Zuordnungsprobleme. Die Funktionsweise eines solchen Algorithmus kann im wesentlichen als einer nichtlinearen Diskriminanzanalyse entsprechend beschrieben werden, deren statistische Eigenschaften allerdings nicht derart klar sind, wie bei den herkömmlichen Verfahren. Bei der Anwendung zeigt sich schnell, daß bereits einige hundert "Knoten" in einem solchen Netz, die nicht nur Fälle, sondern auch Variablen symbolisieren, die Grenze der Leistungsfähigkeit der verwendeten Hard- und Softwarekombination bedeuten. Die Anpassungsarbeiten für diese extrem kleine Kombination betragen schon einige Tage.

Um einen Datenangriff auf einen Mikozensus starten zu können, wären Programme und Rechner erforderlich, die Hunderttausende oder Millionen von Knoten verarbeiten können. Von kommerziellen Anbietern werden spezielle Hardwarezusatzausrüstungen angeboten, die eine derartige Leistungsfähigkeit aufweisen sollen (z. B. die sogenannte ANZA- Karte für PCs und Workstations, vgl. Schöneburg et al. 1990:205ff.), aber teuer sind. Damit entfällt das Argument von der leichten Beschaffbarkeit des Verfahrens für den Angreifer.

Derartig große Netze sind zudem nur sehr aufwendig zu konfigurieren. Ihre Ergebnisse sind sehr schwer interpretierbar. Bisher sind trotz des hohen Aufwands an Forschungskapazität nur wenige Beispiele für erfolgreiche praktische Anwendungen bekannt geworden. Zur Zeit ist noch nicht klar, ob diese Schwierigkeiten lediglich 'Kinderkrankheiten' eines in der Entwicklung befindlichen Produkts oder prinzipielle Schranken darstellen, die den Ansatz grundlegend betreffen.

Das entscheidende Argument gegen die Anwendbarkeit Neuronaler Netze für das hier interessierende Zuordnungsproblem wurde jedoch implizit bereits genannt: Das Mustererkennungsproblem, für das die Netze entworfen sind, setzt - wie die übliche Diskriminanzanalyse auch - eine Trainingsmöglichkeit voraus. Diese ist hier jedoch nicht vorhanden, da keine Informationen darüber vorliegen, welche systematischen Abweichungen das Antwortverhalten im Rahmen einer Mikrozensuserhebung gegenüber der Erhebung für Kürschners Deutschem Gelehrtenkalender auszeichnet. Bei der diskriminanzanalytischen Reidentifikationstechnik war dies der Grund für die aufwendigen Dichte-

---

29 Eine Darstellung beider Methoden enthält McClelland/Rumelhart (1988). Insbesondere zur komplizierteren Boltzmann-Maschine gibt es eine umfangreiche Literatur (vgl. z.B. Aarts/Korst 1989, Aarts/Laarhoven 1987 und Hinton/Sejnovski 1986).

schätzungen auf der Basis eines simulierten Fehlerprozesses gewesen, die sich als der entscheidende Schwachpunkt des ganzen Zuordnungsversuchs gezeigt hatten. Da eine Trainingsmöglichkeit für das betreffende Neuronale Netz ebenso fehlt wie die Lernstichprobe für die Diskriminanzanalyse, stellt sich das Problem bei den neuen Algorithmen in der gleichen Weise. Der Angreifer wäre wiederum auf Annahmen verwiesen. Das Netz könnte ganz enorme Verbesserungen in der Leistungsfähigkeit gegenüber der klassischen linearen Diskriminanzanalyse bringen, es würde aber nichts nützen, da die für die Zuordnung notwendige Information nicht vorhanden ist oder nur falsch geschätzt werden könnte.

Die Leistungsfähigkeit des betreffenden Algorithmus ist in der vorliegenden Anwendung nicht das entscheidende Problem. Die Testrechnung mit einem synthetischen Identifikationsfile hat vielmehr noch einmal die beträchtliche Performanz des verwendeten Verfahrens bestätigt. Damit wird deutlich, daß nur geringe Gefahren von Verbesserungen der potentiellen Werkzeuge ausgehen können. Die technische Seite des Problems wurde von der GMD bereits gelöst, die entscheidende Hürde ist jedoch inhaltlicher Natur. Hier liegt die Barriere, die für das Scheitern der Zuordnungsversuche maßgeblich ist.

Tabellenanhang Abschnitt 9.2

Tabelle 9 A1: Distanzen zwischen benachbarten Fällen

IOO: Nummer eines Falls im Mikrozensus
IOON1 - IOON5: Die fünf nächsten Fälle des Gelehrtenkalenders
DOON1 - DOON5: Die zugehörigen Distanzen

OBS	IA2	IOO	EXACT	DOON1	IOON1	DOON2	IOON2	DOON3	IOON3	DOON4	IOON4	DOON5	IOON5
.													
.													
.													
56	56	2572	0	1069.79	6837	1104.88	3083	1289.59	6505	1300.25	5156	1309.19	5549
57	57	2573	0	231.86	5043	364.53	5498	469.47	7300	478.02	4222	483.41	6735
58	58	2574	0	414.98	3664	430.16	3342	437.28	5799	449.12	4002	454.91	3390
59	59	2575	0	200.76	7470	211.02	119	211.02	4712	211.74	551	211.74	7500
60	60	2576	0	200.76	7470	211.02	119	211.02	4712	211.74	551	211.74	7500
61	61	2577	0	394.44	4411	697.14	7428	704.58	477	704.58	1571	704.58	1892
62	62	2578	0	1403.89	7439	1775.32	6837	1980.26	6505	1987.52	5156	1996.23	5549
63	63	2579	0	49.49	6641	52.20	222	62.87	917	62.87	7362	63.54	2414
64	64	2580	0	257.83	5517	277.06	7470	281.69	2242	281.69	6296	282.33	119
65	65	2581	1	0.01	6456	0.74	2242	0.74	6296	0.74	3227	2.93	119
66	66	2582	0	186.67	3452	186.68	3786	191.05	2392	198.24	3767	198.24	6198
67	67	2583	0	176.00	5977	177.66	127	177.66	7581	188.39	2373	188.54	943
68	68	2584	0	104.35	2653	146.43	6239	242.75	2850	244.82	93	244.97	5837
69	69	2585	0	104.35	2653	146.43	6239	242.75	2850	244.82	93	244.97	5837
70	70	2586	0	1727.32	2010	1734.75	2419	1752.48	2038	1760.65	4453	1760.91	7640
.													
.													
.													

Tabelle 9 A2: Zuordnungen in Situation 2

IF	MF	Zuordnungswahrscheinlichkeit	99% und eindeutig	Distanz
67	64246	0.9738		15
1624	21129	0.9649		506
1835	161625	0.9521		387
1882	75637	0.9793		1125
	164656	0.9532		1158
	165511	0.9667		1209
2200	66964	0.9998	*	364
2649	8688	0.9987		1009
	13442	0.9870		1098
	22121	0.9807		1101
	23685	0.9999		894
	26284	0.9991		1021
	27088	0.9991		1076
	59298	1		529
	63144	0.9980		1055
	63393	0.9999		791
	65963	0.9999		744
	105114	0.9991		1057
	108298	0.9996		979
	109029	0.9999		788
	113611	1		587
	132335	0.9999		630
	151497	1		543
	158558	0.9999		292
	161036	0.9907		1070
	170888	0.9998		897
2808	126531	0.9995		135
	130396	0.9788		203
2913	20046	0.9986		332
	125678	0.9973		310
	131570	0.9923		191
	131759	0.9883		224
	160347	0.9792		397
3052	82305	0.9924	*	17
4512	5606	0.9999	*	0
6284	126367	0.9540		1459
	127981	0.9797		425
6713	30406	0.9998	*	32
7034	64101	0.9766		198
	64108	0.9849		141
7452	24666	0.9984		188
	74136	0.9727		416
	75018	0.9678		421
Summe			4	

Tabelle 9 A3: Zuordnungen in Situation 3

IF	MF	Zuordnungswahrscheinlichkeit	99% und eindeutig	Distanz
262	74980	0.9789		0
364	20615	1		0
704	8041	0.9992		202
	21293	0.9992		189
1598	156268	0.9990	*	0.7
1991	130034	0.9918	*	40
2200	66964	0.9998	*	323
2406	65094	0.9925	*	0
2586	72772	0.9967	*	0
2719	74680	1	*	0
2727	160072	0.9999	*	0
2897	66493	0.9969	*	0
2913	113885	0.9932		361
	131570	0.9940		186
6108	160066	0.9862		0.7
6746	67812	0.9999	*	0
6948	67639	0.9827		0.7
7568	74640	0.9994	*	0.7
Summe			11	

Tabelle 9 A4: Zuordnungen in Situation 4

IF	MF	Zuordnungswahrscheinlichkeit	99% und eindeutig	Distanz
253	161066	0.9731		45
262	74980	0.9982	*	0
364	20615	1	*	0
1253	74242	0.9816		376
2200	69279	0.9920		1151
	96280	0.9703		1243
	146750	0.9956		1011
2406	65094	0.9998	*	0
2586	72772	0.9989	*	0
2719	74680	0.9998	*	0
2727	160072	0.9994	*	0
2913	113885	0.9627		1309
	131759	0.9543		521
	151850	0.9818		1192
3003	68259	0.9906	*	186
4433	109670	1	*	0
6746	67812	0.9999	*	0
7452	8336	0.9779		1046
	13970	0.9999		741
	54878	1		464
Summe			9	

Anhang zu 9.2.2.1:

Formalisierte Darstellung zweier Typen von Fehlern und anderen Inkompatibilitäten bei Umfragedaten

1. Der im Rahmen des AIMIPH-Projekts angenommene Fehlerprozeß hat folgende Gestalt:

$$y_i = f_M(x_i, \theta_M)$$

$$z_i = f_K(x_i, \theta_K)$$

dabei sind:

x: ein Vektor an Werten für die Überschneidungsmerkmale, wie er 'in der Wirklichkeit' vorliegt,
y: ein Vektor an Werten im Mikrodatenfile, der bereits durch Datenfehler, Veränderungen in der Zeit, etc. modifiziert ist,
z: ein ebenso veränderter Vektor an Werten im Identifikationsfile,
f: die Wahrscheinlichkeitsfunktion (bzw. Dichtefunktion) dieser Fehler,
θ: ein Vektor an Parametern, der die konkrete Gestalt der Funktion bestimmt,
i: der Laufindex für einzelne Fälle (Haushalte, Personen)
M,K: der Index für das Mikrodatenfile bzw das Identifikationsfile.

Es wird angenommen, daß der Funktionstyp f und der Parametervektor θ nicht von der Person i abhängen.

2. Das Vorliegen von Alternativvariablen kann in folgender Weise beschrieben werden:

$$y_i = f_M(x_i, \theta_M)$$

$$z_i = g_i(f_K(x_i, \theta_K))$$

g ist dabei keine Wahrscheinlichkeits- (oder anders zu bestimmende) Funktion, sondern eine Relation, genauer: eine Korrespondenz, die einigen Komponenten von x jeweils zwei Komponenten von z zuordnet. g ist für jeden Fall i separat anzugeben.

Im vorliegenden Fall ist z ein Vektor mit mehr Komponenten als y, da z für alle Variablen, die mit Alternativen vorhanden sind, zwei Komponenten enthält (ist m die Anzahl der Komponenten von x und a die Anzahl der Variablen mit doppeldeutigen Ausprägungen, so hat z h = (m + a) Komponenten).

Während in der von der GMD unterstellten ersten Situation Mikrodatenfile und Identifikationsfile jeweils m Überschneidungsmerkmale enthalten, umfaßt das Identifikationsfile in der zweiten Situation deren h = m + a, weshalb bei den einfachen Zuordnungsverfahren 2^a Teil-Identifikationsfiles gebildet wurden.

Anhang zu 9.2.2.2:

Berechnung der Fehlerwahrscheinlichkeiten im Fall des Fehlerprozesses Nr. 2 bei Variablen mit Alternativausprägungen

Die Höhe der Fehlerrate bei den vier Variablen mit alternativen Ausprägungen kann nach den Regeln der elementaren Wahrscheinlichkeitsrechnung ermittelt werden. Der Fehlerprozeß Nr. 2 ist zusammengesetzt aus dem Fehlerprozeß Nr. 1 und der Berücksichtigung von Alternativausprägungen. Da zu berechnen ist, wann ein Fall von einem Fehler gemäß Prozeß Nr. 1 (im folgenden als Ereignis A benannt) oder von einer Abänderung gemäß der Alternativausprägung betroffen ist (Ereignis B), ist folgende Formel heranzuziehen (P(A) bzw. P(B) stehen dabei für die Wahrscheinlichkeiten der Ereignisse A bzw. B):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Sind A und B unabhängig, gilt weiterhin:

$$P(A \cap B) = P(A)\ P(B)$$

wird für P(A) := a_j und für P(B) := 0.5 (1 - m_j) gesetzt, so gilt:

$$G_j = a_j + 0.5\ (1 - m_j)\ (1 - a_j)$$

dabei sind:

G:	die effektive Gesamtfehlerwahrscheinlichkeit, wie sie in Tabelle 2 dargestellt ist.
a:	der Fehleranteil, wie er sich auf Grund von Plausibilitätsüberlegungen gemäß Fehlerprozeß Nr. 1 ergibt
m:	der Anteil an "Missing Values" bei der zweiten Alternative (vgl. Tabelle 2)
j:	der Laufindex für die Variablen mit Alternativausprägungen
(1 - m_j):	die Wahrscheinlichkeit, daß für die Variable j zwei alternative Ausprägungen A1 und A2 vorliegen. Diese Wahrscheinlichkeit wird mit 0,5 multipliziert, da angenommen wird, daß beide Ausprägungen A1 und A2 gleich wahrscheinlich sind.

9.3 Zuordnung zwischen Mikrozensus und einer sozialwissenschaftlichen Stichprobe mit einer einfachen Ableichtechnik

Die Analyse jener Datenquellen, die im Wissenschaftsbereich vorhanden oder verfügbar sind (vgl. Kapitel 6), hatte ergeben, daß im wesentlichen zwei Arten von Informationsquellen als Zusatzwissen für einen Deanonymisierungsversuch in Frage kommen können. Hierzu gehören zum einen öffentlich zugängliche Informationsquellen in Form von Registern oder privaten Datenquellen, für welche in den vorangegangenen Abschnitten Kürschners Deutscher Gelehrtenkalender exemplarisch einer empirischen Überprüfung unterzogen wurde und die auch später bei der argumentativen Analyse einzelner Subszenarien nochmals aufgegriffen werden. Zum anderen verfügen die Sozialwissenschaften auch über Datenbestände, die entweder in eigenen Untersuchungen (Primärdaten) oder von Dritten (Sekundärdaten) erhoben werden. Soweit diese Daten personenbezogen sind, bilden sie einen Bestandteil des Zusatzwissens und könnten daher unter Umständen für Reidentifikationsversuche zweckentfremdet werden.

Wie den Ausführungen zum Zusatzwissen entnommen werden kann, liegt ein solcher Personenbezug - von wenigen Ausnahmen abgesehen - jedoch nur bei Primärdaten vor. Schon aus finanziellen Gründen basieren derartige Eigenerhebungen in der Regel auf Stichproben. Ihr Umfang ist meist relativ klein, jedenfalls gemessen am Umfang amtlicher Erhebungen. Im Vergleich zu berufsspezifischen Handbüchern, wie beispielsweise Kürschners Deutschem Gelehrtenkalender, die eine eingegrenzte Subpopulation zum Teil nahezu vollständig erfassen, wirkt die Stichprobeneigenschaft von sozialwissenschaftlichen Erhebungen daher als Schutz vor der Verwendung der Daten zu Reidentifikationszwecken; je kleiner der Auswahlsatz eines Datenfile ist, desto geringer ist die Wahrscheinlichkeit, daß sich Informationen über eine beliebige Person sowohl in diesem Datenfile als auch in einem Mikrodatenfile finden.

Die Schutzwirkung von Stichproben könnte jedoch durch den umfangreichen Merkmalskatalog immunisiert werden, der sozialwissenschaftlichen Erhebungen oftmals eigen ist. Bei einer Auswahl von manchmal mehreren hundert Variablen finden sich in aller Regel mehr direkte und indirekte Überschneidungsmerkmale

zu amtlichen Datenbeständen, als dies bei Handbüchern der Fall ist.[30] Damit liegt die Annahme nahe, daß diese Datenbestände - zumal bei solchen eines größeren Stichprobenumfangs - einen ausreichend hohen Informationsgehalt aufweisen, um als Zusatzwissen für Reidentifikationszwecke nutzbar zu sein. Ihre Tauglichkeit zu diesem Zweck wird in diesem Abschnitt einer näheren Überprüfung unterzogen.

Zur Erinnerung sei nochmals die schon in Abschnitt 8.1 ausführlich dargestellte Versuchsanlage skizziert. Als Angriffsszenario wird unterstellt, daß ein Wissenschaftler aus persönlicher Neugier Informationen über einen Arbeitskollegen, Bekannten oder Nachbarn gewinnen will. Hypothetisch wird nun unterstellt, daß er zu diesem Zweck versucht, durch eine *gezielte Suche* diese Person in einem Mikrozensus zu deanonymisieren, wobei er allerdings nicht weiß, ob die gesuchte Person auch tatsächlich an der Mikrozensus-Erhebung teilgenommen hat. Während als Mikrodatenfile wiederum der Mikrozensus Nordrhein-Westfalen Verwendung findet, werden als Identifikationsfile Daten aus einer für sozialwissenschaftliche Verhältnisse außergewöhnlich umfangreichen Stichprobe[31] verwandt, in welcher allein für Nordrhein-Westfalen 2 685 Personen enthalten sind. Die im Datenfile enthaltenen Informationen können als typische Beispiele für vorhandenes Alltags- beziehungsweise Anschauungswissen über andere Personen, wie z.B. Nachbarn, Bekannte oder Arbeitskollegen gelten. Jeder der über 2 600 in der sozialwissenschaftlichen Stichprobe enthaltenen Fälle steht für entsprechendes Wissen, das über einen Dritten vorhanden sein kann. Auf diese Weise kann

- zum einen überprüft werden, welches Gefährdungspotential von sozialwissenschaftlichen Stichproben mit einem sehr umfangreichen Merkmalskatalog ausgeht;

- zum anderen überprüft werden, welches Risiko besteht, eine beliebige Person anhand von Alltags- beziehungsweise Anschauungswissen im Mikrodatenfile zu deanonymisieren.

---

30 Die Nutzung der in sozialwissenschaftlichen Stichproben enthaltenen Informationen für Reidentifikationszwecke ist jedoch ebenso wenig problemlos in die Praxis umzusetzen, wie dies bei Handbüchern möglich ist (vgl. Abschnitt 8.2).

31 Hierbei handelt es sich allerdings um Sekundärdaten, die Daten selbst weisen keinen Personenbezug auf. Die Nutzung der Stichprobe für die Reidentifikationsexperimente wurde durch das oben beschriebene Doppel-Blind-Verfahren möglich.

9.3.1 Untersuchungsdesign

Die für die Zuordnungsexperimente zur Verfügung stehende sozialwissenschaftliche Stichprobe enthält 35 personen- und haushaltsspezifische Überschneidungsmerkmale für (bezogen auf Nordrhein-Westfalen) 2 685 Fälle. Während sich die Datenkonstellation des Gelehrtenszenarios durch den spezifischen Bezug auf eine deutlich eingegrenzte Subpopulation sowie deren nahezu vollständiger Erfassung idealerweise für die Überprüfung eines Massenfischzugs anbot, scheidet diese Möglichkeit für die sozialwissenschaftliche Erhebung aufgrund des geringen Auswahlsatzes aus.[32] Die 35 zum Teil sehr detailliert ausgewiesenen Überschneidungsmerkmale bieten jedoch gute Bedingungen, um die Strategie der gezielten Suche beliebiger Personen (Einzelsuche) zu überprüfen.

Die Datenlage ist für die Überprüfung eines solchen Angriffsszenarios insofern besonders aufschlußreich, da neben den personenspezifischen Angaben der Befragten auch differenzierte Merkmale der im Haushalt lebenden Partner und Kinder enthalten sind. Spezifisch bei Haushaltskontextmerkmalen im Zusatzwissen wird in der Literatur häufig von einem erhöhten Gefährdungspotential ausgegangen (Bethlehem et al. 1990). Vorausgesetzt ist dabei wiederum das sogenannte "Uniqueness-Konzept", bei dem sich das Reidentifizierungsrisiko allein durch den Anteil der Einzelfälle in der Grundgesamtheit beziehungsweise - davon abgeleitet - in der Stichprobe bestimmt.[33] Auf der Grundlage von Häufigkeitsabschätzungen kann gezeigt werden, daß insbesondere haushaltsbezogene Merkmale relativ stark zwischen einzelnen Personen diskriminieren. Bereits wenige Variablen können zu einzigartigen Kombinationen bestimmter Ausprägungen in einem Datenfile führen (Brunnstein 1987, Fischer-Hübner 1986, Greenberg 1990). Da im Alltagswissen Informationen über Haushalte von Arbeitskollegen, Nachbarn oder sonstigen Bekannten in einem beschränkten

---

32 Wie den Ausführungen zur Handlungslogik und zum Zusatzwissen entnommen werden kann, handelt es sich hierbei nicht um ein Spezifikum der ausgewählten Studie. Massenfischzüge erfordern vielmehr, daß das Zusatzwissen des Angreifers einen sehr hohen Auswahlsatz aufweist, der bei der überwiegenden Mehrheit sozialwissenschaftlicher Stichproben nicht gegeben ist.

33 Wie in Kapitel 4 ausgeführt, kann mit statistischen Methoden auf der Basis der Stichprobeninformation abgeschätzt werden, wie hoch der Anteil eindeutiger Fälle in der Grundgesamtheit ist und hieraus wiederum der Anteil echter eindeutiger Fälle in einem Mikrodatenfile geschätzt werden. Die Bereitstellung von Methoden, mit denen dies geleistet werden kann, ist einer der wesentlichen Beiträge, die von Vertretern des "Uniqueness-Konzepts" zu der Datenschutzdiskussion erbracht wurde und hebt die Arbeiten von Dittrich/Schlörer (1985) oder Bethlehem et al. (1990) über die naive Anwendung einfacher Sortierverfahren heraus.

Umfang entweder bekannt oder vergleichsweise einfach beschaffbar sind, wird weiter gefolgert, daß das Reidentifikationsrisiko deutlich ansteigt, wenn detaillierte Angaben über den Haushaltszusammenhang in einem Mikrodatenfile enthalten sind.

Um diese vorwiegend wahrscheinlichkeitstheoretisch begründete Argumentation empirisch präzisieren zu können, wurden die Reidentifikationsexperimente in drei Phasen durchgeführt.

- In einer ersten Phase werden lediglich jene Überschneidungsmerkmale einbezogen, die sich direkt auf die befragte Person in der sozialwissenschaftlichen Erhebung beziehen (*personenspezifische Überschneidungsmerkmale*). Hierzu gehören unter anderem das Geburtsjahr, die Ausbildung oder das Geschlecht des Befragten. Insgesamt stehen 13 dieser personenspezifischen Überschneidungsmerkmale zur Verfügung. Da in der sozialwissenschaftlichen Erhebung nur Personen im Alter von 18 bis 55 Jahren befragt wurden, umfaßt die relevante Zielpopulation im Mikrozensus 94 747 Personen.

- In einer zweiten Phase werden die personenspezifischen Merkmale durch allgemeine Informationen über den Haushaltskontext der Befragten ergänzt. Konkret handelt es sich hier um Angaben über die Anzahl der Kinder im Haushalt - differenziert nach Alter - und das Haushaltsnettoeinkommen. Diese Merkmale liegen im Mikrozensus als sogenannte "Bandsatzerweiterungen"[34] vor und sind bereits in die individuellen Datensätze eingetragen. Für diese Phase liegen 21 Überschneidungsmerkmale für wiederum 94 747 Personen vor.

- In der letzten Phase wird eine nochmals wesentlich erweiterte Informationsbasis über den jeweiligen Haushaltskontext der Befragten für die Reidentifikationsexperimente berücksichtigt.[35] Hierunter fallen präzise

---

34 Bei Bandsatzerweiterung handelt es sich um Merkmale, die nicht direkt bei dem Befragten erhoben werden, sondern nachträglich aus den direkt erhobenen Angaben abgeleitet werden.

35 Diese Angaben sind in den beiden Datenfiles allerdings in unterschiedlicher Weise erfaßt. In der sozialwissenschaftlichen Erhebung gelangte nur jeweils eine Person pro Haushalt als Untersuchungseinheit in die Stichprobe. Haushaltsmerkmale werden als Kontext dieser Personen erfragt (Proxy-Interview). Im Mikrozensus hingegen werden alle Haushaltsmitglieder in die Stichprobe aufgenommen. Die Haushaltsinformationen ergeben sich hier größtenteils erst aus der haushaltsbezogenen Auswertung der Angaben der individuellen Haushaltsmitglieder. Die Berück-

Angaben zum Geburtsjahr und zur Ausbildung der im Haushalt lebenden Kinder[36], wie auch Informationen über die Erwerbstätigkeit eines im Haushalt lebenden Partners. Diese Informationen bilden damit in etwa das Wissen über einen Haushalt ab, über welches beispielsweise ein sehr gut informierter Bekannter oder ein Arbeitskollege verfügen könnte. Das Zusatzwissen enthält in dieser Phase 35 Überschneidungsmerkmale zum Mikrozensus. Da sich diese zusätzlichen Informationen nur auf Mehrpersonenhaushalte beziehen, können Einpersonenhaushalte in dieser Phase unberücksichtigt bleiben.[37] Durch dieses Selektionskriterium umfaßt die Zielpopulation im Mikrozensus nun 53 441 Fälle.

Im einzelnen sind die in den drei Zuordnungsphasen als Überschneidungsmerkmale genutzten Informationen in Übersicht 9.4 aufgeführt.

Die Reidentifikationsexperimente erfolgen ausschließlich auf der Basis einer einfachen Abgleichtechnik.[38] Die Vorgehensweise bei einer einfachen Abgleichstechnik, die Zuordnungen nur auf der Basis von Ausprägungsidentität vornimmt, wurde bereits in Abschnitt 3.2 detailliert beschrieben. Im vorliegenden Szenario wird analog mit dem Statistik-Programm SPSS^{x} gearbeitet. Dieses Programm wird zunächst benutzt, um die Datenbestände entsprechend den vorliegenden Merkmalsausprägungen zu sortieren und anschließend über die Funktion "match files" Datensätze mit identischen Merkmalsausprägungen in den beiden Files zuzuordnen.

---

sichtigung der haushaltsspezifischen Merkmale bei den Reidentifikationsexperimenten erfordert - bedingt durch die unterschiedliche Datenstruktur von Mikrozensus und Identifikationsfile - sehr aufwendige Vorarbeiten, da die verwendeten Zuordnungstechniken voraussetzen, daß Identifikationsfile und Mikrodatenfile in ihren Überschneidungsmerkmalen eine identische Struktur aufweisen.

36 Nach Brunnstein (1987) wirken diese Haushaltsmerkmale sehr selektiv und führen schnell zu einzigartigen Merkmalsausprägungen in einem Mikrodatenfile.

37 Das Gefährdungspotential bei alleinstehenden Personen wird in den beiden ersten Zuordnungsphasen überprüft, Einpersonenhaushalte können daher ohne Informationsverlust aus der Analyse ausgeschlossen werden.

38 Auf eine weitere Anwendung der diskriminanzanalytischen Reidentifikationstechnik wurde angesichts der Ergebnisse des Gelehrtenszenarios sowie des mit dieser Reidentifikationstechnik verbundenen Aufwands verzichtet.

Übersicht 9.4: Angriffsszenario 2: Überschneidungsmerkmale in Zuordnungsphase1 bis 3

Überschneidungsmerkmale	Anzahl Auspräg. *	Zuordnungsphase: 1	2	3
PERSONENSPEZIFISCHE MERKMALE				
Geschlecht	2	x	x	x
Geburtsjahr	38	x	x	x
Familienstand	4	x	x	x
Schulische Ausbildung	6	x	x	x
Berufsausbildung	8	x	x	x
Stellung im Beruf	11	x	x	x
Erwerbstätigkeit	3	x	x	x
Arbeitslosigkeit	3	x	x	x
wöchentliche Arbeitszeit	7	x	x	x
Arbeitssuche	3	x	x	x
Art des Arbeitsvertrags	4	x	x	x
Ende d. Erwerbstätigkeit	8	x	x	x
persönliches Nettoeinkommen	9	x	x	x
ALLGEMEINE HAUSHALTSINFORMATIONEN				
Zahl Kinder im Hhlt unt. 3 Jahren	3		x	
dto von 3 bis unter 6 Jahren	3		x	
dto von 6 bis unter 10 Jahren	3		x	
dto von 10 bis unter 15 Jahren	3		x	
dto von 15 bis unter 18 Jahren	3		x	
dto von 18 bis unter 28 Jahren	5		x	
dto über 28 Jahre	2		x	
Haushaltsnettoeinkommen	9		x	x
DETAILLIERTE HAUSHALTSINFORMATIONEN				
Partner: Stellung i. Beruf	11			x
-"- Arbeitszeit (i.Stunden)	8			x
-"- Erwerbstätigkeit	3			x
-"- Arbeitslos	3			x
-"- Arbeitssuche	3			x
Geburtsjahr Kind1	33			x
Geburtsjahr Kind2	27			x
Geburtsjahr Kind3	25			x
Geburtsjahr Kind4	18			x
Geburtsjahr Kind5	10			x
Geschlecht Kind1	2			x
Geschlecht Kind2	2			x
Geschlecht Kind3	2			x
Geschlecht Kind4	2			x
Geschlecht Kind5	2			x
Ausbildung Kind1	4			x
Ausbildung Kind2	4			x
Ausbildung Kind3	4			x
Ausbildung Kind4	4			x
Ausbildung Kind5	2			x
Vorwieg. Unterhalt des Haushalts	7			x
Gesamtzahl Überschneidungsmerkmale		13	21	35
N (Mikrozensus)		94.747	94.747	53.441
N (Sowi. Erhebung)		2.685	2.685	2.685

* Diese Angaben beziehen sich auf die sozialwissenschaftliche Erhebung, da nur die dort auftretenden Merkmalsausprägungen für die Zuordnungen mit einer einfachen Reidentifikationsmethode relevant sind.

9.3.2 Ergebnisse des Zuordnungsexperiments

Die Analyse des Mikrozensus im Hinblick auf jene Datensätze, die eine einmalig auftretende Ausprägungskombination (Einzelfälle) aufweisen, zeigt, daß Haushaltsmerkmale, wie erwartet, erheblich zwischen den einzelnen Personen der Stichprobe diskriminieren (vgl. Tabelle 9.10). Werden nur die personenspezifischen Merkmale berücksichtigt, sind lediglich 20 Prozent der Fälle im Mikrozensus einzigartig. Die Zahl der Einzelfälle steigt erheblich, wenn zusätzlich Angaben über das Haushaltsnettoeinkommen sowie die Anzahl der Kinder im Haushalt - klassifiziert nach Alter - (Zuordnungsphase 2) zur Verfügung stehen. Von den 94 747 Datensätzen des Mikrozensus sind hierfür mehr als drei Viertel (79 Prozent) in ihren Ausprägungskombinationen einmalig.[39] Sind die Haushaltsinformationen - wie in Zuordnungsphase 3 - noch präziser und beinhalten Angaben zur Erwerbstätigkeit eines im Haushalt lebenden Partners sowie detaillierte Informationen über die Kinder, erhöht sich der Anteil der einzigartig auftretenden Ausprägungskombinationen auf 84 Prozent.

Da der Stichprobenumfang in der sozialwissenschaftlichen Erhebung wesentlich kleiner ist, ergeben sich dort schon mit den Merkmalen der ersten Zuordnungsphase 77 Prozent einzigartige Merkmalskonstellationen. In den Phasen 2 und 3 steigen diese Anteile auf 97 beziehungsweise 99 Prozent. Die ermittelten Quoten einzigartiger Ausprägungskombinationen sowohl für das Mikrodatenwie für das Identifikationsfile machen deutlich, daß auch bei diesem Experiment eine Untersuchungsanlage gewählt wurde, die nach dem Uniqueness-Konzept als in höchstem Maße riskant zu bezeichnen ist.

Das von den Haushaltsmerkmalen ausgehende Diskriminierungspotential (gemessen an der Quote der Einzelfälle) spiegelt sich auch in der von Zuordnungsphase 1 zu Zuordnungsphase 3 deutlich reduzierten Zahl der Zuordnungen wider. Je höher der Anteil der Einzelfälle wird, desto niedriger wird zugleich die Zahl der Fälle, die im Mikrozensus eine zu der sozialwissenschaftlichen Stichprobe identische Ausprägungskombination aufweisen. Bei der vergleichsweise wenig differenzierten Aufgliederung in Zuordnungsphase 1 weist noch ein großer Teil der Datensätze der sozialwissenschaftlichen Erhebung eine übereinstimmende Ausprägungskombination mit einem - beziehungsweise vorwiegend mehreren Datensätzen - des Mikrozensus auf. Die Zahl der Zuord-

[39] In Abänderung zur Vorgehensweise in Zuordnungsphase 1 wurde die wöchentliche Arbeitszeit in Zuordnungsphase 2 mit 61 Ausprägungen wesentlich differenzierter bei den Reidentifikationsexperimenten berücksichtigt. Wäre die Arbeitszeit in Zuordnungsphase 2 ebenfalls nur stark vergröbert (sieben Ausprägungen) berücksichtigt worden, würde sich eine Einzelfallquote von 66 Prozent ergeben.

nungen nimmt dann in Zuordnungsphase 2 und 3 deutlich ab. Die Abnahme vollzieht sich jedoch nicht in dem Umfang, wie durch den Anstieg der Einzelfallquote entsprechend den Annahmen des Uniqueness-Konzepts zu erwarten wäre (vgl. Tabelle 9.10).

Tabelle 9.10: Anzahl der auf der Grundlage von identischen Merkmalsausprägungen möglichen (ein- und mehrdeutigen) Zuordnungen in Abhängigkeit von den einbezogenen Überschneidungsmerkmalen

Phase	ÜM	Einzigartige Ausprägungskombinationen MZ %	Sowi %	Zahl der Zuordnungen: insg.	eindeutig	mehrdeutig[40]
1	13	20	77	1107	nicht überprüft	
2	21	79	97	298	nicht überprüft	
3	35	84	99	74	35	39

Bedingt durch den Auswahlsatz des Mikrozensus (ein Prozent der Gesamtbevölkerung) ergibt sich, da die sozialwissenschaftliche Stichprobe 2685 Fälle umfaßt, ein Erwartungswert von etwa 27 Personen, die an beiden Erhebungen teilgenommen haben und damit das Reidentifikationspotential bilden. Gemessen an diesem Erwartungswert ist die Zahl der aufgrund von Ausprägungsidentität ermittelten Zuordnungen in allen drei Zuordnungsphasen zu hoch: In der ersten Zuordnungsphase werden 1 107 Datensätze, d.h. 41 Prozent aller Fälle der sozialwissenschaftliche Stichprobe, ein- oder mehrdeutig Fällen des Mikrozensus zugeordnet. In Zuordnungsphase 2 (21 Überschneidungsmerkmale) findet sich noch immer für 298 Fälle der sozialwissenschaftlichen Erhebung mindestens ein identischer Fall im Mikrozensus. Diese Zahl ist im Vergleich zum Erwartungswert - ungeachtet der differenzierteren Überschneidungsmerkmale und der dadurch bedingten Einzelfallquote von 79 Prozent im Mikrozensus - noch immer um den Faktor 10 zu hoch. Dieser Über-

[40] Mit der gewählten Methode wird jeweils ein Fall des Mikrozensus einem Fall der sozialwissenschaftlichen Erhebung zugeordnet. Sind im Mikrozensus weitere Fälle mit identischen Merkmalsausprägungen enthalten, so müssen diese in separaten Abfragen gesucht werden. Darauf wurde in Zuordnungsphase 1 und 2 angesichts der Klarheit des Ergebnisses verzichtet.

hang an Zuordnungen kann dadurch erklärt werden, daß sich im Mikrozensus eine Vielzahl von Datensätzen findet, die in ihren Ausprägungskombinationen identisch mit Datensätzen der sozialwissenschaftlichen Stichprobe sind, ohne daß diese Datensätze von ein und derselben Person stammen. Es handelt sich demnach um statistische Doppelgänger. Dies würde bedeuten, daß die 21 Überschneidungsmerkmale in bezug auf die Gesamtbevölkerung bei weitem nicht in dem Ausmaß diskriminieren, wie dies - legt man die formalen Modelle zur Schätzung von Populationseinzigartigkeit zugrunde (vgl. Kapitel 4) - bei einer Einzelfallquote von 79 Prozent im Mikrozensus anzunehmen wäre.

Aufgrund der noch höheren Einzelfallquote in Zuordnungsphase 3 (Mikrozensus: 84 Prozent, sozialwissenschaftliche Stichprobe: 99 Prozent), wäre entsprechend den Annahmen des Uniqueness-Konzeptes anzunehmen, daß das Diskriminierungspotential der präzisen Haushaltsinformationen ausreichend hoch ist, um vielen Datensätzen im Merkmalsraum eindeutige Positionen zuzuweisen. Die Wahrscheinlichkeit einer zufälligen Übereinstimmung der Merkmalsausprägungen zweier beliebiger Datensätze müßte in diesem Falle sinken. In der Folge wäre daher zu erwarten, daß sich die Zahl der empirisch möglichen Zuordnungen an den Erwartungswert von 27 Fällen annähert.

Diese Annahme wird durch die Ergebnisse der Zuordnungsexperimente allerdings nicht bestätigt. Bei 2 685 Einzelsuchen, wobei jeweils ein Fall mit allen 53 441 Fällen des Mikrozensus abgeglichen wurde, können zwar 35 Fälle in einer eins-zu-eins Entsprechung, d.h. eindeutig zugeordnet werden. Zusätzlich finden sich jedoch für weitere 39 Fälle im Schnitt sechs statistische Doppelgänger im Mikrozensus.[41] Im Vergleich zu den vorhergehenden Zuordnungsphasen reduziert sich damit zwar die Zahl der Zuordnungen beträchtlich, mit 74 möglichen Zuordnungen liegt sie jedoch noch immer um den Faktor drei höher als der theoretische Erwartungswert. Dieser Befund deutet zum einen darauf hin, daß die Haushaltskontextmerkmale - bezogen auf die Population - offensichtlich bei weitem nicht in dem Ausmaß differenzieren, wie dies aufgrund der Einzelfallquote zu vermuten wäre.[42] Zum anderen ist anzunehmen, daß eine

41 Konkret werden diesen 39 Fällen des Identifikationsfile 250 Fälle des Mikrozensus aufgrund von Ausprägungsidentität zugeordnet.

42 Hält man sich vor Augen, daß es in der alten Bundesrepublik über 26 Millionen Haushalte gibt (Wedel 1989) wird deutlich, daß es einer gleichzeitigen und sehr tiefen Sach- und Regionaldifferenzierung eines Mikrodatenfile bedarf, damit statistische Doppelgänger in der Population unwahrscheinlich werden.

Vielzahl der hier - auch eindeutig - erfolgten Zuordnungen auf Verwechslungen mit statistischen Doppelgängern beruhen und daher falsch sind.

9.3.3 Überprüfung der Ergebnisse durch den Treuhänder

Die Überprüfung der vorgenommenen Zuordnungen durch den Treuhänder ist überraschend und zeigt, daß bei einem realen Angriffsszenario mit Unsicherheitsfaktoren gerechnet werden muß, die bei einer ausschließlich wahrscheinlichkeitstheoretischen Analyse des Reidentifikationsrisikos nicht in hinreichendem Maße berücksichtigt werden können und daher zu einer erheblichen Überschätzung eines realen Risikos führen können:

- Entgegen der wahrscheinlichkeitstheoretisch als plausibel anzunehmenden Zahl von etwa 27 Personen, die an beiden Erhebungen teilgenommen haben, finden sich in der Realität nur zehn Fälle, die sowohl im Mikrozensus als auch in der sozialwissenschaftlichen Stichprobe enthalten sind (im folgenden als Schnittmenge bezeichnet). Diese Diskrepanz ist sehr wahrscheinlich auf das unterschiedliche Stichprobendesign bei der Mikrozensus-Erhebung zurückzuführen. Hierdurch kann die Zahl der Personen, die in der Überschneidungsmenge zu allen potentiellen Stichproben enthalten ist, relativ stark - sowohl nach oben als auch nach unten - streuen. Für die Beurteilung der Erfolgschancen eines Reidentifikationsversuchs ist dieser Erwartungswert, wenn sowohl das Mikrodatenfile als auch das Zusatzwissen nur eine Stichprobe darstellen, nur ein sehr grober Indikator.

- Die Datensätze der zehn in beiden Stichproben enthaltenen Fälle können in keiner Zuordnungsphase korrekt zugeordnet werden. Alle empirisch (auf der Basis von Ausprägungsidentität) ermittelten Zuordnungen sind falsch, d.h. sie beruhen auf einer zufälligen Ausprägungsgleichheit der zugeordneten Datensätze. Der Anteil der Falschzuordnungen liegt damit für das hier untersuchte Szenario bei 100 Prozent. Die Quote der korrekten Zuordnungen ist gleich Null (vgl. Tabelle 9.11).

Tabelle 9.11: Anteil der korrekten und falschen Zuordnungen an der Gesamtzahl der Zuordnungen in den einzelnen Zuordnungsphasen des untersuchten Subszenarios

Zuordnungs-phase:	Anzahl der Überschneid-ungsmerkmale	Zuordnungen:	Hiervon: korrekt	falsch
1	13	1107	0%	100%
2	21	298	0%	100%
3	35	79	0%	100%

Während die *Falschzuordnungen* sowohl auf Verwechslungen aufgrund von Dateninkompatibilitäten als auch auf echte statistische Doppelgänger zurückführbar sein können, muß die *Nichtzuordnung* der zehn gesuchten Fälle auf Dateninkompatibilitäten beruhen. Das heißt, die betreffenden Datensätze sind in den beiden Stichproben unterschiedlich abgebildet und können daher nicht nach dem Kriterium der Ausprägungsidentität zugeordnet werden. Dieser Befund weist damit in die gleiche Richtung wie die Ergebnisse des Gelehrtenszenarios: Dateninkompatibilität kommt unter empirischen Bedingungen ebenso wie der Stichprobeneigenschaft von Daten eine außerordentlich wichtige Schutzfunktion bei Reidentifikationsversuchen zu.

Überraschend ist allerdings das Ausmaß, in welchem Inkompatibilitäten die Reidentifikationsversuche stören. Ungeachtet der Vielzahl von Zuordnungen - in Zuordnungsphase 1 werden 41 Prozent aller Fälle der sozialwissenschaftlichen Erhebung ein- oder mehrdeutig Datensätzen des Mikrozensus zugeordnet - kann in keiner Phase auch nur ein gesuchter Fall korrekt zugeordnet werden.

Da - wie in Abschnitt 8.2 dargestellt - die Datenanpassung für die Reidentifikationsexperimente umfangreiche Datenmodifikationen vorausgesetzt hatte, liegt die Annahme nahe, die aufgetretenen Abweichungen könnten eventuell erst in dieser Phase in das Datenmaterial eingebracht worden sein. Diese Mög

lichkeit kann jedoch nach einem Vergleich[43] der zehn unveränderten Datensätze aus Mikrozensus und sozialwissenschaftlicher Erhebung ausgeschlossen werden. Danach bestehen die Abweichungen bereits in den Originaldaten und ziehen sich infolgedessen wie ein roter Faden durch alle Zuordnungsphasen.

9.3.4 Dateninkompatibilitäten als natürlicher Schutz vor Reidentifikationsversuchen

Der zentrale Befund dieser Analyse ist also, daß Dateninkompatibilitäten einen impliziten Schutz vor der Möglichkeit der Reidentifikation anonymisierter Daten darstellen. Für das Verständnis dieser überraschenden These ist es instruktiv, im einzelnen zu verfolgen, wie es zu diesem Ergebnis kommt. Im folgenden werden daher jene zehn Fälle genauer untersucht, die sowohl in den Datensätzen der sozialwissenschaftlichen Stichprobe wie auch im Mikrozensus abgebildet sind.

Zur Illustration der Wirkungsweise von Inkompatibilitäten sollen zunächst nur die für die zehn Fälle bei den 13 personenspezifischen Merkmalen von Zuordnungsphase 1 aufgetretenen Dateninkompatibilitäten betrachtet werden. Wie Übersicht 9.5 zeigt, sind die Angaben bei keinem der zehn Fälle über alle 13 Merkmale vollständig kompatibel.

Für lediglich fünf Merkmale (Geschlecht, Geburtsjahr, Familienstand, Erwerbstätigkeit und Art des Arbeitsvertrags) sind alle zehn Fälle in Mikrozensus und sozialwissenschaftlicher Stichprobe identisch abgebildet. Bei den verbleibenden acht personenspezifischen Merkmalen sind pro Merkmal im Schnitt etwa 2,8 Fälle im Mikrozensus in anderer Weise abgebildet als in der sozialwissenschaftlichen Erhebung. Die einzelnen Datensätze sind hierbei ganz unterschiedlich von Inkompatibilitäten betroffen: Vier Fälle sind lediglich in einem Merkmal inkompatibel, zwei Fälle weisen jeweils in zwei, ein Fall in drei und drei Fälle in vier Merkmalen Inkompatibilitäten auf. Da die einfache Abgleichtechnik jedoch nur dann korrekt zuordnen kann, wenn die abgebildeten Informationen für einen gesuchten Fall vollständig identisch sind, ist es bereits in der ersten Zuordnungsphase nicht möglich, auch nur eine korrekte Zuordnung vorzunehmen.

---

43 Dieser Vergleich ist durch den Treuhänder möglich, der die jeweiligen Datensatznummern der zehn gesuchten Personen kennt (vgl. Abschnitt 8.2.1). Anhand

Übersicht 9.5: Verteilung der Inkompatibilitäten für die 10 in Mikrozensus und sozialwissenschaftlicher Erhebung gemeinsam enthaltenen Fälle auf die 13 personenspezifischen Überschneidungsmerkmale[44]

Fall	Personenspezifische Überschneidungsmerkmale[1)]													
Nr	1	2	3	4	5	6	7	8	9	10	11	12	13	Summe
1)	k	k	k	k	k	k	k	k	k	k	*	k	k	1
2)	k	k	k	k	k	k	*	k	k	k	k	k	*	2
3)	k	k	k	k	k	k	*	*	k	k	k	k	*	3
4)	k	k	k	k	k	k	k	k	k	k	*	k	*	2
5)	k	k	k	k	k	k	k	k	k	k	k	k	*	1
6)	k	k	k	k	k	k	*	k	k	k	k	k	k	1
7)	k	k	k	k	k	k	*	k	*	*	k	k	*	4
8)	k	k	k	k	k	*	*	k	*	k	k	*	k	4
9)	k	k	k	k	k	*	*	*	k	*	k	k	k	4
10)	k	k	k	k	k	k	k	k	k	k	k	k	*	1
Su:	0	0	0	0	0	2	6	2	2	2	2	1	6	23

1) k: Das Überschneidungsmerkmal ist für den betreffenden Fall kompatibel
*: Das Überschneidungsmerkmal ist für den betreffenden Fall inkompatibel

Es stellt sich nun die Frage, inwieweit sich die Reidentifikationschancen erhöhen würden, wenn ein Angreifer zufällig oder aufgrund von Plausibilitätsüberlegungen bei jenen vier Fällen, die nur in einem Merkmal abweichen, exakt jenes abweichende Merkmal bei seinem Reidentifikationsversuch ausschließen würde. Diese interessante Frage soll im Moment allerdings noch zurückgestellt werden, um dann in Abschnitt 9.3.5.2 wieder aufgegriffen zu werden. Im folgenden interessiert vielmehr, inwieweit auch die Haushaltskontextmerkmale von Inkompatibilitäten betroffen sind.

Tabelle 9.12 listet auf, wie viele der insgesamt 35 Überschneidungsmerkmale für alle zehn Fälle kompatibel sind. Hiernach weichen bei zwei Drittel aller Merkmale die Angaben in der sozialwissenschaftlichen Erhebung in mindestens einem Fall von den Angaben im Mikrozensus ab. Lediglich 12 der 35 Merkmale

dieser Nummernpaare können die entsprechenden Datensätze - unter Wahrung der Anonymität der dahinter stehenden Personen - verglichen werden.

44 Personenspezifische Überschneidungsmerkmale: 1. Geschlecht, 2. Geburtsjahr, 3. Familienstand, 4. Erwerbstätigkeit, 5. Art des Arbeitsvertrags, 6. Schulische Ausbildung, 7. Berufsausbildung, 8. Ende d. Erwerbstätigkeit, 9. Arbeitslosigkeit, 10. Arbeitssuche, 11. wöchentliche Arbeitszeit, 12. Stellung im Beruf, 13. persönliches Nettoeinkommen

sind über alle zehn Fälle kompatibel. Von den Abweichungen sind auch die Haushaltskontext-Merkmale betroffen: Keines der fünf Partner-Merkmale war für alle zehn Datensätze kompatibel. Bei den 15 Kinder-Merkmalen erwiesen sich sechs Merkmale für alle zehn Datensätze als identisch abgebildet. Von den zwei Variablen mit allgemeinen Haushaltsinformationen wies eines stets kompatible Angaben auf.

Tabelle 9.12: Anzahl der über alle zehn Fälle kompatiblen und der in mindestens einem Fall inkompatiblen Merkmale

Merkmalsbereich	über alle zehn Fälle kompatibel	für mindestens einem Fall inkompatibel	Summe
Personenspezifisch	5	8	13
Partner	0	5	5
Kinder	6	9	15
allg. Hhltangaben	1	1	2
Summe	12	23	35

Übersicht 9.6 kann entnommen werden, daß die einzelnen Merkmale unterschiedlich häufig von Inkompatibilitäten betroffen sind. Besonders häufig erweist sich die Angabe zur "Arbeitszeit des Partners" als inkompatibel. Sieben von zehn Fällen weisen hierfür im Mikrozensus andere Ausprägungen auf als in der sozialwissenschaftlichen Erhebung.[45] Jeweils sechs von zehn Fällen sind in den Merkmalen "persönliches Nettoeinkommen", "Haushaltsnettoeinkommen" und "beruflicher Ausbildungsabschluß" im Mikrozensus anders abgebildet als in der sozialwissenschaftlichen Stichprobe. Die Inkompatibilitäten bei den verbleibenden Merkmalen betreffen jeweils nur einige wenige Fälle. Sie tragen dennoch dazu bei, daß die entsprechenden Datensätze nicht als identisch erkannt und daher nicht beziehungsweise nur falsch zugeordnet werden können.

[45] Die im Mikrozensus und in der sozialwissenschaftlichen Erhebung enthaltenen Angaben differieren so stark, daß die Abweichungen auch dann noch auftreten, wenn die Merkmalsausprägungen - wie für Zuordnungsphasen 1 und 3 geschehen - von über 80 auf knapp neun Ausprägungen zusammengefaßt werden.

Übersicht 9.6: Anzahl der zwischen Mikrozensus und sozialwissenschaftlicher Erhebung inkompatiblen Angaben. Basis: 10 Fälle, die gleichzeitig in Mikrozensus und sozialwissenschaftlicher Erhebung enthalten sind

Überschneidungs-merkmale	Inkompatible Angaben	Überschneidungs-merkmale	Inkompatible Angaben
Geschlecht	0	Schulabschluß	2
Geburtsjahr	0	Ende d. Berufstät.	2
Familienstand	0	Arbeitslos	2
Erwerbstätigkeit	0	Arbeitssuche	2
Art d. Arbeitsvertr.	0	Arbeitszeit	2
Geburtsjahr Kind3	0	Arbeitslos (*Partner*)	2
Geburtsjahr Kind5	0	Erwerbstätig. (*Partner*)	2
Geschlecht Kind3	0	Arbeitssuche (*Partner*)	2
Ausbildung Kind3	0	Stell. Beruf (*Partner*)	2
Ausbildung Kind5	0	Ausbildung Kind2	2
Vorwieg. Unterhalt	0	Geschlecht Kind1	2
Stellung i. Beruf	1	Ausbildung Kind1	4
Geburtsjahr Kind1	1	Nettohhlt.Einkommen	6
Geburtsjahr Kind2	1	Ausbildungsabschluß	6
Geschlecht Kind2	1	Nettoprivateinkommen	6
Ausbildung Kind4	1	Arbeitszeit (*Partner*)	7
Geschlecht Kind4	1		
Geburtsjahr Kind4	1		

Im Hinblick auf die Schutzwirkung von Inkompatibilitäten ist der folgende Sachverhalt bedeutsam: Je mehr Merkmale benötigt werden, um in einem Mikrodatenfile zu einzigartigen Ausprägungskombinationen zu gelangen, umso größer ist die Wahrscheinlichkeit, daß zumindest eines der benötigten Merkmale inkompatibel abgebildet ist. Eine korrekte Zuordnung kann jedoch nur erfolgen, wenn bei keinem einzigen Merkmal eine Inkompatibilität vorliegt. Die Konsequenzen dieser Grundvoraussetzung werden aus der eingangs dargestellten Übersicht 9.5 sehr deutlich. Wenn auch bei den zehn im Mikrozensus und in der sozialwissenschaftlichen Erhebung gleichzeitig enthaltenen Fällen die Abweichungen bei jeweils anderen Merkmalen auftreten, über die 13 Merkmale ist keine Person vollständig kompatibel in beiden Files abgebildet und deshalb ist eine korrekte Zuordnung nicht möglich.

9.3.4.1 Mögliche Ursachen für die aufgetretenen Dateninkompatibilitäten

Eine detaillierte Analyse der Ursachen für die aufgetretenen Inkompatibilitäten kann nicht geleistet werden. Dies würde eine Recherche bei den zehn Personen erfordern und hierfür müßte deren Identität bekannt sein. Wie aus den allgemeinen Ausführungen zu Ursachen, Häufigkeit und Wirkung von Inkompatibilitäten in Abschnitt 4.3 hervorgeht, können neben Datenfehlern im klassischen Sinne (Rekodierung, Aufnahme etc.) auch unterschiedliche sachliche und zeitliche Bezüge zu Abweichungen führen.

Als eine mögliche Ursache für die ermittelten Inkompatibilitäten kommt demnach der im Schnitt um 20 Monate differierende Erhebungszeitpunkt der beiden Stichproben in Betracht. Diese zeitliche Differenz könnte sich unter anderem in der Einkommenshöhe oder der wöchentlichen Arbeitszeit bemerkbar machen. Auch die beim Geschlecht und Geburtsjahr der Kinder festgestellten Abweichungen können auf zeitlichen Differenzen beruhen: Der Mikrozensus spiegelt die Haushaltszusammensetzung zu einem bestimmten Zeitpunkt wider. Veränderungen im Zeitablauf werden in der Regel nicht erfaßt: Verläßt ein Kind den Haushalt, kann es durch einen Nachrückeffekt der anderen Kinder zu Verschiebungen in der Alters- oder Geschlechtsstruktur der Kinder kommen. Da zwischen dem Erhebungszeitpunkt des Mikrozensus und der frühestmöglichen Weitergabe - bedingt durch umfangreiche Aufbereitungsarbeiten - mehrere Monate liegen[46], besteht bei der Verwendung von nicht fixen Merkmalen als Zusatzwissen immer das Problem, daß die Informationen des Angreifers aktueller sein können als die im Mikrozensus enthaltenen Angaben.

Neben dem Zeitfaktor können die Abweichungen auch durch Unsicherheiten der Befragten in Hinblick auf die "wahre" Angabe hervorgerufen werden. So weiß zwar jede Person ungefähr, wieviel Geld ihr monatlich zur Verfügung steht. Der präzise Betrag des persönlichen wie auch des Haushaltseinkommens ist jedoch oft nicht bekannt. Erfahrungsgemäß wird das Haushaltseinkommen bei amtlichen Erhebungen von den Befragten jedoch eher unter- als überschätzt (Statistisches Bundesamt 1989). Es ist daher nicht verwunderlich, wenn insbesondere Angaben zur Höhe der Einkünfte beträchtliche Schwankungen aufweisen.

---

46 Im Schnitt dauert die Aufbereitung des Mikrozensus mit Erstellung der Basis-Tabellen sechs bis neun Monate (Herberger 1985).

Auch in bezug auf die zu einem Haushalt zählenden Personen kann es zu Unschärfen beispielsweise durch Mißverständnisse kommen: Zählen Kinder, die an einem anderen Ort leben und studieren noch zum Haushalt oder nicht? Die Angaben, die die Befragten zu solchen Fragen in einer amtlichen Erhebung machen, können sich beträchtlich von den Antworten in einer anderen Erhebung beziehungsweise von dem Eindruck eines Außenstehenden unterscheiden.

> Interessant in diesem Zusammenhang ist allerdings die Richtung der Abweichungen für die Bildungsqualifikation. Bei der Beschreibung der Datenbasis ist festgestellt worden, daß in den univariaten Randverteilungen in bezug auf die Berufsausbildung erhebliche Unterschiede zwischen dem Mikrozensus und der sozialwissenschaftlichen Erhebung auftraten. Der Anteil der Personen in und ohne Ausbildung war im Mikrozensus deutlich höher als in der sozialwissenschaftlichen Stichprobe. Zugleich lag der Anteil der Personen mit einem Lehr- oder Fachschulabschluß in der sozialwissenschaftlichen Erhebung um etwa acht Prozent höher als im Mikrozensus.
>
> Bei diesen Abweichungen handelt es sich nicht um ein spezifisches Problem dieser sozialwissenschaftlichen Erhebung, sondern um typische Verzerrungen, die in der Literatur als "Mittelschichtsbias der Umfrageforschung" bekannt sind (Esser et al. 1989). Der Mittelschichtsbias besagt unter anderem, daß Personen mit weiterführender Bildungsqualifikation in sozialwissenschaftlichen Erhebungen überrepräsentiert, Personen mit niedrigen Bildungsabschlüssen eher unterrepräsentiert sind. Im allgemeinen wird dieser Bias darauf zurückgeführt, daß durch die freiwillige Teilnahme an Erhebungen spezifische Selektionsmechanismen (als eine Kombination von Erreichbarkeit und Kooperationsbereitschaft) bei der Zielpopulation einsetzen, die systematische Ausfälle bewirken (Hartmann 1990).
>
> Die Überprüfung der zehn in Mikrodatenfile und Identifikationsfile gemeinsam enthaltenen Personen zeigt, daß die in den univariaten Randverteilungen aufgetretenen Verzerrungen auch auf individueller Ebene bestehen, d.h. in bezug auf die zehn in Mikrozensus und sozialwissenschaftlicher Erhebung gemeinsam enthaltenen Fälle: Zwei dieser zehn Fälle sind in bezug auf den Schulabschluß, sechs von zehn in bezug auf den Ausbildungsabschluß inkompatibel. Diese Inkompatibilitäten werden dadurch verursacht, daß Befragte, die beim Mikrozensus keine(n) oder eine(n) niedrige(n) Schulbildung oder Ausbildungsabschluß angegeben haben, in der sozialwissenschaftlichen Erhebung eine höhere Qualifikation angeben, beziehungsweise zum Teil keine Angabe machen. Es kann - aufgrund der Hintergrundvariablen in der sozialwissenschaftlichen Erhebung - ausgeschlossen werden, daß die betreffenden Personen zwischen den beiden Erhebungszeitpunkten eine formale Weiterqualifikation erwor-

ben haben. Es handelt sich demnach um unterschiedliches Antwortverhalten in Abhängigkeit von dem Erhebungskontext.

Gegen die hier vorgetragene Argumentation der "natürlichen" Schutzfunktion von Dateninkompatibilitäten mag eingewandt werden, daß sie im wesentlichen auf einer empirischen Illustration durch nur zehn Fälle beruht und deshalb wenig überzeugend sei. Die Befunde stimmen zwar mit den theoretischen Erwartungen überein, aber es wäre in der Tat wünschenswert, wenn sie auf breiterer Basis abgesichert werden könnten. Dieses ist mit den durchgeführten Zuordnungsexperimenten nicht möglich, weil die Schnittmenge der im Mikrozensus und in der sozialwissenschaftlichen Stichprobe gemeinsam enthaltenen Datensätze nicht nachträglich erhöht werden kann. Durch eine Reanalyse der Allbus Test-Retest-Studie kann jedoch ein äquivalentes Experiment auf einer quantitativ breiteren Basis nachgestellt werden. In dieser von ZUMA 1984 durchgeführten Studie wurden einer Zufallsauswahl der ALLBUS-Stichprobe im Monatsabstand dreimal dieselben Fragebögen vorgelegt, um die Reliabilität von Umfragedaten zu untersuchen (Porst/Zeifang 1987). Als Ergebnis dieser Studie zeigt sich, daß zwar die soziodemographischen Merkmale relativ hohe Antwortstabilitäten aufweisen, mit der Ausnahme der Variable Geschlecht jedoch alle untersuchten Merkmale mit einer gewissen Wahrscheinlichkeit von Abweichungen betroffen sind.[47] Hiervon ist - wie den allgemeinen Ausführungen zu Dateninkompatibilitäten in Abschnitt 4.3 entnommen werden kann - auch das Geburtsjahr nicht ausgenommen. In weitaus höherem Maße zeigen jedoch die Merkmale zur Erwerbstätigkeit Abweichungen. So unterscheiden sich zwischen Welle 1 und Welle 3

- 45 Prozent der Befragten hinsichtlich der Angaben zur Einkommenshöhe;
- 18 Prozent hinsichtlich der geleisteten Arbeitswochenstunden;
- 16 Prozent hinsichtlich des angegebenen beruflichen Ausbildungsabschlusses;
- 13 Prozent hinsichtlich ihrer überwiegenden Einkünfte.

Auf individueller Ebene zeigen sich von Welle 1 zu Welle 2 sowie von Welle 1 zu Welle 3 ähnlich hohe Abweichungen wie zwischen dem hier untersuchten Mikrozensus und der sozialwissenschaftlichen Stichprobe. Aus Tabelle 9.13 geht hervor, daß bei Berücksichtigung von elf Merkmalen lediglich 22 Prozent der Befragten zwischen Welle 1 und 2 vollständig gleiche Angaben machen. Knapp 39 Prozent unterscheiden sich in einer Angabe, 28 Prozent in zwei Angaben und etwa neun Prozent in drei Angaben. Eine ähnliche Abweichungsstruktur zeigt sich auch zwischen Welle 1 und Welle 3, wobei hier allerdings

47 Ähnliche Befunde aus anderen Untersuchungen wurden bereits in Abschnitt 4.3 bei den Ausführungen zu Dateninkompatibilitäten erwähnt.

prozentual mehr Fälle zwei und mehr Abweichungen aufweisen als zwischen Welle 1 und 2.

Tabelle 9.13: Allbus Test-Retest-Studie 1984: Fallspezifische Häufigkeiten von aufgetretenen Inkompatibilitäten bei elf berücksichtigten Merkmalen zwischen Welle 1 und Welle 2 sowie Welle 1 und Welle 3

Von 11 Merkmalen*) waren inkompatibel	Anteil der betroffenen Fälle von: Welle 1 zu Welle 2 %	Welle 1 zu Welle 3 %
0	22.1	20.8
1	38.7	33.1
2	28.2	31.2
3	8.8	14.3
4	2.2	.6
N	181	154

Quelle: eigene Berechungen auf Basis der Allbus Test-Retest-Studie 1984

*) Berücksichtigte Variablen (in Klammern: Anzahl der Merkmalsausprägungen):

- Geschlecht (2)
- Alter (58)
- Familienstand (5)
- Allgemeiner Schulabschluß (5)
- Berufl. Ausbildungsabschluß (5)
- Berufl. Erwerbstätigkeit derz.(22)
- Arbeitslos (2)
- Berufliche Stellung derz. (22)
- Berufliche Stellung früher (20)
- Arbeitswochenstunden (74)
- Monatliches Netto-Einkommen (56)

Wenn aber schon Erhebungen, die explizit mit dem Ziel durchgeführt werden, die Stabilität des Antwortverhaltens zu messen, eine zum Teil beträchtliche Inkompatibilität zwischen einzelnen Merkmalen von Erhebungswelle zu Erhebungswelle aufweisen, ist nicht weiter erstaunlich, daß bei Datenquellen,

denen ein vollständig anderer Erhebungskontext zugrunde liegt, ähnliches zu beobachten ist.

9.3.5 Weitergehende Überlegungen zu dem untersuchten Szenario

Angesichts der scheinbar unvermeidbaren Dateninkompatibilitäten könnte ein Angreifer den Versuch unternehmen, sich bei seinen Reidentifikationsversuchen auf jene Merkmale zu beschränken, die er als vergleichsweise valide einschätzt. Bei der vorliegenden Datenbasis beträfe dies die personenspezifischen Merkmale "Geschlecht, Geburtsjahr, Familienstand, Erwerbstätigkeit und Art des Arbeitsvertrags".[48] Wie bereits aus der Vielzahl von zufälligen Zuordnungen bei Berücksichtigung von 13 personenspezifischen Merkmalen geschlossen werden kann, sind diese fünf Merkmale bei weitem nicht ausreichend, um die zehn gesuchten Datensätze eindeutig im Mikrozensus abzugrenzen. Die Einzelfallquote des Mikrozensus liegt bei Berücksichtigung nur dieser fünf Merkmale bei 0,1 Prozent. Entsprechend finden sich für die spezifischen Merkmalsausprägungen der zehn Fälle zwischen 100 und 881 Fälle im Mikrozensus, die exakt diese Kombinationen aufweisen (vgl. Tabelle 9.14).

Mit jedem zusätzlich berücksichtigten Merkmal erhöht sich zwar das Differenzierungspotential der Merkmale, und der Anteil der einmaligen Ausprägungskombinationen steigt an. Zugleich wächst jedoch auch die Wahrscheinlichkeit, daß ein zusätzlich berücksichtigtes Merkmal in den beiden Vergleichsdatensätzen inkompatibel abgebildet ist. Dies kann entweder zu einer Nichtzuordnung oder aber zu Falschzuordnungen aufgrund von Verwechslungen führen. Je mehr Merkmale benötigt werden, um zu einzigartigen Ausprägungskombinationen zu gelangen, umso eher wird ein Datenangreifer zu Merkmalen greifen, die in aller Regel nicht mit so hoher Reliabilität wie Geburtsjahr, Geschlecht und Familienstand erfaßt werden können. Mit einfachen Techniken des Datenabgleichs gibt es zwischen der Einzigartigkeit fordernden Skylla und den - wenn auch nur mit kleinen Fehlerspitzen geschärften - Klippen von Charybdis kein Entkommen.

---

48 Eigene Berechnungen mit der Allbus-Test-Retest-Studie zeigen, daß auch bei diesen Merkmalen inkompatible Fälle auftreten, wenn eine größere Zahl von Personen untersucht wird.

Tabelle 9.14: Anzahl der Datensätze im Mikrozensus, die bei Berücksichtigung von fünf kompatiblen Überschneidungsmerkmalen jeweils identische Ausprägungskombinationen zu einem der zehn Fälle aufweisen, die auch in der sozialwissenschaftlichen Stichprobe enthalten sind.

Sozialwissenschaftliche Stichprobe Fall Nr.:	Mikrozensus Anzahl der Datensätze mit identischer Ausprägungskombination
1	100
2	193
3	382
4	393
5	395
6	521
7	522
8	592
9	678
10	881

9.3.5.1 Regionalinformationen

Mit Ausnahme des implizit enthaltenden Bundeslandes (die Daten bezogen sich ausschließlich auf das Land Nordrhein-Westfalen) standen in den Datenfiles keine Regionalinformationen zur Verfügung. Da die gesuchten Datensätze jedoch von Inkompatibilitäten betroffen waren, wären sie entsprechend der Logik einfacher Abgleichtechniken auch dann nicht korrekt zuordenbar gewesen, wenn das Zusatzwissen tief gegliederte Regionalinformationen enthalten hätte. Es stellt sich allerdings die Frage, inwieweit sich die Erfolgschancen eines Reidentifikationsversuchs verbessert hätten, wenn sich ein Angreifer auf jene fünf über alle zehn Fälle kompatiblen Überschneidungsmerkmale beschränken würde und seine Datenquelle zusätzlich tief gegliederte Regionalinformationen, beispielsweise in der Größenordnung von 100 000 Einwohnern pro Regionaleinheit, aufweisen würde.

Der Effekt einer solchen zusätzlichen Information kann auf folgende Weise abgeschätzt werden: Nordrhein-Westfalen hat über 16 Millionen Einwohner. Unter der Annahme, daß die Ausprägungen der Merkmale zufällig über den Raum

streuen, wäre bei 100 000 Einwohnern pro Regionaleinheit davon auszugehen, daß Ausprägungskombinationen, die ohne Regionalinformationen weniger als 160fach in einem Mikrodatenfile auftreten, bei Berücksichtigung der Regionalangaben nur noch einfach besetzt sind.

Wie aus Tabelle 9.14 hervorgeht, würde dies für einen der zehn Fälle in dem untersuchten Szenario zutreffen. Allerdings wäre auch unter diesen spezifischen Randbedingungen noch keine sichere Reidentifikation möglich. Bedingt durch den Auswahlsatz des Mikrozensus von einem hundertstel der Gesamtbevölkerung ist anzunehmen, daß die entsprechende Ausprägungskombination in der Grundgesamtheit für etwa weitere hundert Personen zutrifft. Sofern der Angreifer also nicht weiß, ob die gesuchte Person an der amtlichen Erhebung teilgenommen hat, kann Sicherheit über die korrekte Identität dieser Person nur dann hergestellt werden, wenn diese hundert Personen in der Grundgesamtheit ermittelt werden und anhand weiterer Informationen überprüft wird, welche Person tatsächlich im Mikrodatenfile enthalten ist.

9.3.5.2 Erfolgschancen bei Wissen um Teilnahme

Ein stark erhöhtes Reidentifikationsrisiko wird in der Datenschutzdiskussion dann angenommen, wenn ein Angreifer weiß, daß eine gesuchte Person an der amtlichen Erhebung teilgenommen hat.[49] Die Brisanz einer Angriffssituation mit "Teilnahmekenntnis" (Bethlehem et al. 1990) wird abgeleitet aus der Sicherheit des Angreifers, daß die Daten der gesuchten Person in dem Mikrodatenfile enthalten sein müssen (vgl. Kapitel 4). Entspricht dem gesuchten Fall exakt ein Datensatz im Mikrodatenfile, so sind statistische Doppelgänger in der Grundgesamtheit ohne Belang. Denn unter der Bedingung, daß das Mikrodatenfile in vollem Umfang zur Verfügung steht, kann es sich bei einer eindeutigen Zuordnung der Ausprägungskombination dieser gesuchten Person zu einem Datensatz des Mikrodatenfile nur um den Datensatz dieser Person handeln (vgl. unter anderem Bethlehem et al. 1990). Dieser Argumentation zufolge kann daher bei "response knowledge" für eine gesuchte Person eine eindeutige Zuordnung mit

[49] In der Untersuchung von Paaß/Wauschkuhn an teilsynthetischem Datenmaterial führte unter der Randbedingung "Teilnahmekenntnis" über die Hälfte der Reidentifikationsversuche zu einem Erfolg (1985:13).

an Sicherheit grenzender Wahrscheinlichkeit einer Reidentifikation gleichgesetzt werden.[50]

Der empirische Befund der Reidentifikationsexperimente unter der modifizierten Angriffsstrategie "Einzelsuche bei Teilnahmekenntnis"[51] zeigt jedoch, daß auch bei dieser spezifischen Konstellation eine erhebliche Schutzwirkung von Dateninkompatibilitäten ausgehen kann, deren Ausmaß bislang bei weitem unterschätzt wurde. Die unterschiedliche Abbildung von Informationen, die den gleichen Sachverhalt für eine spezifische Person wiedergeben, kann nämlich - unabhängig von der Angriffsstrategie - sowohl zu Nichtzuordnungen wie auch zu Falschzuordnungen führen.

Für die Angriffsstrategie "Teilnahmekenntnis" muß das Untersuchungsdesign leicht modifiziert werden. Aus der sozialwissenschaftlichen Erhebung werden nur jene zehn Fälle für die Experimente berücksichtigt, für welche bekannt ist, daß die datengebenden Personen auch an der Mikrozensuserhebung teilgenommen haben. Auf diese Weise ist sichergestellt, daß sich im Mikrozensus für jeden einzelnen dieser Fälle exakt ein dazugehöriger Datensatz finden lassen müßte. Diese Vorgehensweise simuliert das Wissen des Angreifers, daß die von ihm gesuchte Person an der Erhebung teilgenommen hat und daher in dem Mikrodatenfile enthalten sein muß.

Um den Zuordnungsprozeß und die daraus resultierenden Ergebnisse zu verdeutlichen, wird als Angriffstaktik die sogenannte Hintertreppenidentifikation (Block/Olsson 1976) benutzt. Hierbei wird mittels einer sukzessiven Erweiterung der einbezogenen Überschneidungsmerkmale versucht, die für den Angreifer relevante Zielpopulation einzugrenzen (Dittrich/Schlörer 1987, Bethlehem et al. 1990), wobei als Zuordnungskriterium abermals die Identität in den Merkmalsausprägungen für die gesuchten Fälle gilt. In Phase I werden wiederum jene fünf personenspezifischen Überschneidungsmerkmale berücksichtigt, die über alle zehn Fälle kompatibel sind (vgl. Abschnitt 9.3.5.1). In Phase II werden

50 Eine der "Teilnahmekenntnis" vergleichbare Brisanz wird von den Vertretern des Uniqueness-Konzepts darin gesehen, daß ein Angreifer weiß, daß eine gesuchte Person in ihren Merkmalsausprägungen in der Grundgesamtheit einzigartig ist. Dies könnte beispielsweise dann eintreten, wenn das Mikrodatenfile sehr tiefgegliederte Regionalangaben enthalten würde. Auch in diesem Fall wäre bei einer eindeutigen Zuordnung einer gesuchten Person zu einem einzelnen Datensatz eines Mikrodatenfile mit einer sehr hohen Wahrscheinlichkeit von einer korrekten Zuordnung auszugehen (Greenberg/Voshell 1990, Mokken/Pannekoek/Willenborg 1989).

51 Diese modifizierte Angriffsstrategie kann überprüft werden, da durch den Treuhänder bekannt ist, welche Fälle der sozialwissenschaftlichen Erhebung auch an der Mikrozensuserhebung teilgenommen haben.

fallspezifisch nur jene personenspezifischen Merkmale berücksichtigt, die für den jeweils gesuchten Fall im Mikrozensus und in der sozialwissenschaftlichen Erhebung kompatibel sind. In den weiteren Phasen werden analog zu den vorangegangenen Experimenten zunächst 13 und schließlich 21 Überschneidungsmerkmale in das Zuordnungsverfahren einbezogen. Die Ergebnisse können Übersicht 9.7 entnommen werden.

Übersicht 9.7: Zahl der Fälle im Mikrozensus mit identischer Ausprägungskombination zu dem jeweils gesuchten Fall aus der sozialwissenschaftlichen Stichprobe in Abhängigkeit von der Zahl der berücksichtigten Überschneidungsmerkmale

Phase	I	II	III	IV
Sowi. Erheb. Fall Nr:	5	Anzahl der berücksichtigten Überschneidungsmerkmale in Klammern	13	21
1)	100	1 (12)	0	0
2)	193	55 (11)	2	0
3)	382	22 (10)	0	0
4)	393	137 (11)	0	0
5)	395	53 (12)	12	0
6)	521	235 (12)	159	1
7)	522	192 (9)	0	0
8)	592	25 (9)	0	0
9)	678	588 (9)	0	0
10)	881	27 (12)	5	0

Werden zunächst nur jene fünf personenspezifischen Merkmale berücksichtigt, die für alle zehn gesuchten Personen kompatibel sind, finden sich im Mikrozensus pro gesuchter Person jeweils über 100 Datensätze mit exakt den gleichen Merkmalsausprägungen. Eine eins-zu-eins Zuordnung ist unter dieser Bedingung für keine der gesuchten Personen möglich. Ein Angreifer ist daher gezwungen, weitere Überschneidungsmerkmale zu berücksichtigen. Unter der - völlig hypothetischen und äußerst unrealistischen - Annahme, der Angreifer wüßte exakt, welche der personenspezifischen Merkmale für den einzelnen Fall jeweils kompatibel sind (Phase II), wäre es ihm in einem Fall möglich, eine eindeutige Zuordnung vorzunehmen. In neun von zehn Fällen wären diese Informationen allerdings noch nicht für eine eindeutige Zuordnung ausreichend, es würden weitere Angaben benötigt werden.

Da die Datensätze der zehn Fälle in den verbleibenden personenspezifischen Merkmalen zwischen Mikrozensus und sozialwissenschaftlicher Erhebung abweichende Werte aufweisen, können die zusammengehörigen Datensätze bei Einbezug dieser Merkmale nicht mehr nach dem Kriterium der Ausprägungsidentität zugeordnet werden. Vor diesem Hintergrund sind die Ergebnisse der weiteren Zuordnungsphasen für die Beurteilung der Erfolgschancen von Reidentifikationsversuchen mit "Teilnahmekenntnis" von großer Bedeutung.

Denn bei der Berücksichtigung aller 13 personenspezifischen Merkmale (Phase III) findet sich erwartungsgemäß für sechs von zehn gesuchten Fällen kein identischer Datensatz im Mikrozensus. Vier der zehn Fälle können jedoch nach wie vor aufgrund von Ausprägungsidentität entsprechenden Datensätzen im Mikrozensus zugeordnet werden. Da die Angaben der gesuchten Datensätze im Mikrozensus abweichend von denjenigen in der sozialwissenschaftlichen Stichprobe abgebildet sind, sind sie nicht mehr in der eingekreisten Subpopulation enthalten. Der Suchprozeß führt in eine völlig falsche Richtung, da es aufgrund der Inkompatibilitäten zu Verwechslungen kommt. Dieser Sachverhalt bleibt einem Angreifer jedoch verborgen. Da er weiß, daß die von ihm gesuchten Personen in dem Mikrozensusfile enthalten sein müssen, besteht für ihn kein Grund, an der Qualität der vorgenommenen Zuordnungen zu zweifeln. Denn solange es gelingt, für eine gesuchte Person einen entsprechenden Datensatz auf der Basis identischer Ausprägungen zuzuordnen, enthalten die Daten kein Indiz für ein mögliches Auftreten von Dateninkompatibilitäten.

Noch deutlicher tritt dieser Sachverhalt in Phase IV zutage. Bei Berücksichtigung von 21 Überschneidungsmerkmalen wird von den in Phase III zugeordneten vier Fällen nur noch einer zugeordnet, dieser nun jedoch eindeutig. Diese Situation entspricht der oben charakterisierten Konstellation mit Teilnahmekenntnis: Ein Angreifer weiß, daß die von ihm gesuchte Person im Mikrozensus enthalten sein muß; er verfügt über sehr präzise Informationen über diese Person und deren Haushaltskontext (beispielsweise wohnen zwei Kinder im Haushalt); und er findet im Mikrozensus einen einzigartigen Datensatz, auf welchen diese Informationen exakt zutreffen. Folgt man der obigen Argumentation, müßte es sich mit sehr hoher Wahrscheinlichkeit um eine korrekte Zuordnung handeln.

Auch diese eindeutige Zuordnung ist jedoch falsch. Sie beruht - bedingt durch Inkompatibilitäten - wiederum auf einer zufälligen Übereinstimmung eines beliebigen Datensatzes des Mikrozensus mit den Informationen des Angreifers über eine von ihm gesuchte Person. Der Datensatz der gesuchten Person ist zwar im Mikrozensus enthalten, aber nicht mit dem Zusatzwissen des An-

greifers kompatibel. Da der Angreifer jedoch scheinbar erfolgreich zuordnen kann, würde ihm diese Verwechslung entgehen.

9.3.6 Folgerungen für die faktische Anonymität von Mikrodaten

Die hier ermittelten Resultate ergänzen die Befunde aus dem Gelehrtenszenario und sind von zentraler Bedeutung für die Diskussion über Reidentifikationsrisiken. Die Zuordnungsexperimente haben gezeigt, daß die untersuchten Einzelangaben faktisch anonym waren.

Der Datensatz einer beliebigen Person erwies sich als faktisch anonym auch unter der Bedingung, daß er nur insoweit anonymisiert war, als neben Name und Adresse Regionalinformationen unterhalb der Ebene der Bundesländer weggelassen wurden.

Hierfür war ausschlaggebend, daß es zum einen nahezu unmöglich ist, eine beliebige Person anhand einiger weniger Merkmale eindeutig zuzuordnen. Denn solange die gesuchte Person nicht eine in der Grundgesamtheit äußerst seltene Ausprägungskombination in diesen Merkmalen aufweist, werden sich in der Grundgesamtheit und daher auch im Mikrodatenfile eine Vielzahl von statistischen Doppelgängern finden, die den Reidentifikationsprozeß nachhaltig stören. Um in diesem Fall zu eindeutigen Ausprägungskombinationen zu kommen, müßte das Zusatzwissen des Angreifers eine Vielzahl von Überschneidungsmerkmalen enthalten.

Je mehr Überschneidungsmerkmale zum anderen jedoch benötigt werden, um zu einer eindeutigen Ausprägungskombination zu kommen, desto wahrscheinlicher wird das Auftreten von Dateninkompatibilitäten. Der Umfang des Zusatzwissens wirkt damit in paradoxer Weise auf den Deanonymisierungsprozeß. Je mehr Überschneidungsmerkmale das Zusatzwissen enthält, desto höher ist der Anteil der Datensätze mit einzigartigen Ausprägungskombinationen und damit das potentielle Reidentifikationsrisiko. Mit jedem zusätzlichen Überschneidungsmerkmal erhöht sich jedoch zugleich auch die Wahrscheinlichkeit, daß ein gesuchter Fall in diesem Merkmal inkompatibel abgebildet ist, woraus Falsch- beziehungsweise Nichtzuordnungen resultieren können. Wie aus der Auflistung der aufgetretenen Dateninkompatibilitäten in Abschnitt 9.3.4 hervorgeht, sind Datensätze daher umso mehr gegen eine Reidentifikation geschützt,

je mehr Merkmale benötigt werden, um zu einzigartigen Ausprägungskombinationen zu kommen.[52]

Weiterhin konnte gezeigt werden, daß Dateninkompatibilitäten in einem unerwartet hohen Ausmaß auch in der bislang als höchst riskanten Angriffskonstellation der "Teilnahmekenntnis" eine immanent wichtige Schutzfunktion einnehmen. Unter der Annahme, ein Angreifer hätte gewußt, daß die zehn gesuchten Fälle an der Mikrozensus-Erhebung teilgenommen haben, hätte er in sechs von zehn Fällen - aufgrund von Inkompatibilitäten - keine Zuordnung vornehmen können. In vier von zehn Fällen wäre es jedoch - ebenfalls durch Inkompatibilitäten bedingt - zu Verwechslungen und damit Falschzuordnungen gekommen. Diese Ergebnisse schließen nicht aus, daß es unter der spezifischen Risikokonstellation "Teilnahmekenntnis" mit einer gewissen Wahrscheinlichkeit möglich ist, eine gesuchte Person in einer eindeutigen Weise einem Datensatz in einer amtlichen Erhebung zuzuordnen. Solange der Angreifer seine Zuordnungen jedoch nicht verifizieren kann, und dies wäre letztendlich nur über einen Anschriftenvergleich möglich, kann er auch bei dieser spezifischen Risikokonstellation nicht davon ausgehen, daß die zugeordneten Datensätze von ein und derselben Person stammen. Die Unsicherheit, inwieweit die zugeordneten Datensätze tatsächlich dieselbe Person betreffen, bleibt daher auch dann bestehen, wenn der Angreifer weiß, daß die gesuchte Person im Mikrodatenfile enthalten sein muß.

Eine direkte Übertragung des im Rahmen dieses Szenarios real geleisteten Aufwands auf die unterstellte Angriffssituation der Einzelsuche ist nicht möglich. Die sozialwissenschaftliche Stichprobe lag bereits maschinenlesbar vor. Bei einem realen Datenangriff müßte dieser Aufwand zunächst erbracht werden. Hierbei hatte sich im Gelehrtenszenario gezeigt, daß in Abhängigkeit von der Anzahl der Merkmale und dem Umfang des Identifikationsfile zum Teil beträchtliche Kosten anfallen. Da diese Arbeiten hier nicht vorgenommen werden mußten, können nur Aussagen darüber getroffen werden, welcher Aufwand beim Einsatz eines bereits maschinell erfaßten Datenfile für einen Reidentifikationsversuch entsteht.

Für die Entwicklung des Konzepts, die Anpassung (Rekodierung) der Überschneidungsmerkmale und die Angleichung der Datenstruktur von Identi-

52 Marsh et al. (1991) berichten von analogen Mechanismen in bezug auf einzelne Merkmale. Ihre Analysen zeigen, daß mit zunehmender Differenzierung eines Merkmals die Wahrscheinlichkeit von inkompatiblen Abbildungen einzelner Ausprägungen steigt (vgl. Kapitel 4)

fikationsfile und Mikrodatenfile entstanden Arbeitskosten in Höhe von ca. 15 000 DM (9,5 Wochen BAT IIa). Obwohl beide Datenfiles maschinenlesbar vorlagen, ergaben sich aufgrund der umfangreichen Rekodierungsarbeiten einschließlich der Durchführung der Zuordnungsexperimente Rechenkosten in Höhe von etwa 13 000 DM. Der gesamte finanzielle Aufwand betrug damit 28 000 DM. Da die Quote der korrekten Zuordnungen unabhängig von der gewählten Angriffsstrategie bei Null lag, wäre im vorliegenden Szenario eine Alternativbeschaffung der gewünschten Informationen in jedem Fall die kostengünstigere Alternative gewesen.

10 Argumentative Analyse einzelner Szenarien*

Wie schon in Kapitel 8 festgehalten, kann die Untersuchung der dort bereits kurz beschriebenen Angriffsszenarien 3, 4 und 5 nicht ebenfalls wie die Szenarien 1 und 2 durch eigene empirische Überprüfungen erfolgen. Eine wesentliche Grundlage bei der argumentativen Analyse dieser Szenarien bilden in den folgenden Abschnitten gleichwohl die quantitativen, durch die Zuordnungsexperimente empirisch für die Szenarien 1 und 2 ermittelten Ergebnisse aus Kapitel 9. Sie kommen unter Beachtung der für die Situationen 3, 4 und 5 unterschiedlichen Randbedingungen im Rahmen von Plausibilitätsüberlegungen und Wertungen zur Anwendung. Daher erheben natürlich nicht alle Aussagen in den Erörterungen dieser Situationen einen Anspruch auf Allgemeingültigkeit.

Die Diskussion erfolgt für jede Situation nach dem gleichen Schema: Zunächst werden in vier Unterabschnitten die Ausgangslage, die zugrundegelegte Datenbasis, die Fehler- und Abgrenzungsproblematik sowie das konkrete Vorgehen für das jeweilige Szenario kurz vorgestellt. Den Hauptteil bildet danach die ökonomische Beurteilung der Konstellation. Sie enthält in jedem Fall eine Erörterung der Deanonymisierungskosten, der Kosten für eine alternative Datenbeschaffungsmethode und schließlich ein auf die faktische Anonymisierung ausgerichtetes Fazit. Außer auf die genannten empirischen Ergebnisse für die Szenarien 1 und 2 wird bei diesen ökonomischen Beurteilungen auch auf allgemeine Kosten-Nutzen-Überlegungen aus Kapitel 7 zurückgegriffen.

Im übrigen werden dabei in den folgenden Diskussionen unter den zur Deanonymisierung notwendigen Kosten nur diejenigen betrachtet, die bis zur Zuordnung der Datensätze durch den vom Angreifer verwendeten Algorithmus entstehen, nicht aber der Aufwand zur eventuellen Überprüfung dieser vorgenommenen Zuordnungen auf ihre Korrektheit durch weitere Recherchen. Ebenso sind mögliche Kosten eines Angreifers aufgrund von Fehlzuordnungen ausgeklammert. Damit bilden die hier lediglich im engeren Sinne betrachteten Deanonymisierungskosten jeweils nur eine untere Schranke für einen wegen der genannten zusätzlichen Komponenten möglicherweise umfassenderen Aufwand eines Angreifers.

---

* Autor: Thomas Helmcke

10.1 Angriffsszenario 3: Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe

10.1.1 Beschreibung des Szenarios

Aus den allgemeinen Erörterungen der typischen Handlungslogik von Sozialwissenschaftlern in Kapitel 5 ging als wesentliches Ergebnis hervor, daß eine mögliche wissenschaftliche Motivation für Deanonymisierungen wohl höchstens darin bestehen könnte, die reidentifizierten Datensätze als Auswahlgrundlage für eine eigene Stichprobenerhebung zu verwenden. Eine solche Methode zur Beschaffung einer Grundgesamtheit kann für einen Wissenschaftler allerdings nur dann sinnvoll sein, wenn er sich von den im Mikrodatenfile enthaltenen Daten eine Verbesserung oder einen sonstigen Vorteil für seine Erhebung verspricht.

Als naheliegendes Szenario kommt hier in Betracht, daß der Angreifer seine Erhebung von vornherein auf Datensätze mit ganz bestimmten Mikrodatenfile-Ausprägungen beschränken will. Die Nutzung eines Mikrodatenfile für ein solches Vorhaben wäre plausibel, wenn er damit die Erwartung verbindet, eine gute Stichprobe von Personen mit bestimmten Merkmalen zu gewinnen, die er auf andere Weise entweder überhaupt nicht oder nur mit deutlich höherem Aufwand in vergleichbarer Qualität erzielen kann. Als weitere Möglichkeit ist denkbar, daß ein Wissenschaftler ein Mikrodatenfile als Basis für die Konstruktion einer eigenen Erhebung benutzt, indem er Informationen des Mikrodatenfile etwa für die Bildung einer nach bestimmten Merkmalen geschichteten Stichprobe heranzieht.

Als ein Nebeneffekt bräuchte der Forscher in beiden Fällen in seiner eigenen Erhebung die bereits im Mikrodatenfile enthaltenen und ihn interessierenden Informationen im allgemeinen nicht mehr selbst zu ermitteln. Die eigene Erhebung könnte er allerdings auch speziell nutzen, um z.B. Veränderungen in der Zeit zwischen der Mikrodatenfile-Erhebung und der eigenen Erhebung zu analysieren, möglicherweise auch um Methodenvergleiche anzustellen zwischen Erhebungen mit und ohne Auskunftspflicht, etwa im Hinblick auf Reliabilitätsmaße.

Im folgenden soll als Deanonymisierungsszenario die Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe näher betrachtet werden, wobei der Mikrozensus als Mikrodatenfile diene und das Identifikationsfile aus dem Melderegister zu beschaffen sei. Die Beschränkung auf die spezielle

Personengruppe der Ausländer in Verbindung mit dem relativ informativen Überschneidungsmerkmal "Staatsangehörigkeit" läßt eine vergleichsweise hohe Reidentifikationsquote erwarten, so daß dieses Szenario im hier diskutierten Rahmen zu denjenigen zählt, in denen noch am ehesten mit einer Gefährdung der faktischen Anonymität zu rechnen wäre. Die Nutzung des Mikrodatenfile zur Gewinnung einer Auswahlgrundlage könnte in diesem Fall etwa dadurch motiviert sein, eine Stichprobe aus einer bestimmten Ausländergruppe, z. B. mit bestimmten Berufen, zu ziehen.

10.1.2 Identifikations- und Mikrodatenfile

Wie bereits erwähnt, soll als Mikrodatenfile der Mikrozensus herangezogen werden. Das Identifikationsfile bestehe aus Ausländer-Datensätzen, die aus dem Melderegister zu gewinnen seien.

Im Rahmen der allgemeinen Diskussion über das Zusatzwissen wurde dargelegt, daß die Datenbeschaffung aus dem Melderegister mit wachsender Anzahl der gewünschten Auskünfte wegen der restriktiven Handhabung der einschlägigen Vorschriften durch die zuständigen Behörden schwieriger wird und oberhalb gewisser Grenzen nicht mehr möglich ist. Um diese Widerstände zu mildern, könnte der Angreifer sein Identifikationsfile aus verschiedenen Melderegistern kumulieren, so daß er pro einzelnem Register weniger Datensätze benötigt. Andererseits ist es denkbar, daß der Wissenschaftler aufgrund seiner Kenntnisse regionaler Gegebenheiten Informationen über Mikrozensus-Teilnahmen besitzt und deshalb mit dem Ziel einer höheren Reidentifikationsquote Melderegister-Daten eingegrenzt auf bestimmte Regionen zu beschaffen wünscht. In diesem Fall wäre die Anzahl der involvierten Melderegister relativ niedrig, die durchschnittliche Anzahl der pro Register benötigten Datensätze aber entsprechend hoch, was die Informationsbeschaffung dort erschweren dürfte. Unter Umständen muß auch damit gerechnet werden, daß die Meldebehörden wegen der Konzentration auf eine spezifische Bevölkerungsgruppe, nämlich die der Ausländer, besonders strenge Maßstäbe anlegen. In jedem Fall sind je nach gewünschtem Identifikationsfile-Umfang mehr oder weniger große Hindernisse zu überwinden.

An Überschneidungsmerkmalen enthalten der Mikrozensus und das Melderegister

- Geschlecht
- Geburtsjahr einschließlich Geburtsjahreshälfte
- Familienstand
- Staatsangehörigkeit
- Existenz einer Zweitwohnung
- Regionalmerkmal,

wobei sich das Regionalmerkmal aus der Kombination von Bundesraumordnungsregion, Regierungsbezirk und Gemeindegrößenklasse zusammensetzt.

Während die Anzahl der Überschneidungsmerkmale ziemlich niedrig ausfällt, ist der Informationsgehalt insgesamt relativ hoch einzuschätzen. Im einzelnen trifft dies zunächst auf die Geburtsangaben zu. Aber auch die Staatsangehörigkeit leistet selbst in vergröberter Form mit ihren knapp 30 möglichen Ausprägungen sicher einen wesentlichen Informationsbeitrag, zumal sich bei einem Ausländerfile die Suche nur auf die vergleichsweise kleine Population dieser Bevölkerungsgruppe bezieht. Schließlich spielen die Regionalangaben eine wichtige Rolle. Die Kombination von Regierungsbezirk und Gemeindegrößenklasse erlaubt nicht nur oft eine Eingrenzung von Gemeinden der obersten, sondern manchmal sogar der vierthöchsten Gemeindegrößenklasse. Bei Verwendung der Bundesraumordnungsregion zur weiteren Eingrenzung lassen sich auch Gemeinden noch kleinerer Kategorien bis zur untersten Größenklasse eindeutig bestimmen.

10.1.3 Abgrenzungs- und Fehlerproblematik

Aufgrund der Einfachheit und eindeutigen Definition der genannten Überschneidungsmerkmale, aber auch wegen der eindeutigen und im wesentlichen identischen Gliederung ihrer Ausprägungen im Mikrozensus und im Melderegister kann kaum von einer großen Abgrenzungsproblematik gesprochen werden. Die übereinstimmende Kodierung des jeweiligen Überschneidungsmerkmals sollte somit kaum Schwierigkeiten bereiten.

Auch die Aktualität der Daten scheint in ausreichendem Maße gewährleistet zu sein, zumal sich die Ausprägungen der meisten Überschneidungsmerkmale ohnehin entweder gar nicht oder nur relativ selten ändern. Das Sonderproblem

des eventuellen Verbleibs von Datensätzen ausgewanderter Personen im Melderegister fällt wohl kaum ins Gewicht, da diese in der Regel in dem zum Abgleich verwendeten Mikrozensus nicht enthalten sind.

Wegen der aufgeführten Gründe ist die Quote fehlerhafter Daten als relativ gering einzuschätzen. Etwas vereinfachend sei daher für den Angreifer die Erwartung unterstellt, daß die Sachverhalte im Mikrozensus und im Melderegister weitgehend identisch abgebildet sind.

10.1.4 Praktische Durchführung

Obwohl zur Beurteilung des hier betrachteten Deanonymisierungsszenarios keine Simulationen vorgenommen werden und stattdessen die argumentative Erörterung im Vordergrund steht, werden dabei auch quantitative Überlegungen angestellt. Insbesondere soll die Beziehung zwischen dem vom Angreifer gewünschten Stichprobenumfang, dem Umfang des Identifikationsfile sowie des benötigten Melderegister-Auszugs und der Reidentifikationsquote die Grundlage der Kostendiskussion bilden, bevor dann auch Simulationsergebnisse für Szenario 1 aus Kapitel 9 herangezogen werden. Als Bestimmungsgrund für den Nutzen der Deanonymisierungen wird die statistische Qualität der vom Angreifer geplanten Stichprobenerhebung diskutiert.

10.1.5 Kosten-Nutzen-Analyse

Im Rahmen der allgemeinen Kosten-Nutzen-Überlegungen wurde dargelegt, daß vor einer möglichen Gegenüberstellung von Deanonymisierungskosten und -nutzen in einem konkreten Szenario, also auch jetzt, zunächst geprüft werden sollte, ob eine gleichwertige, aber deutlich günstigere Datenbeschaffungsalternative, existiert. Unabhängig vom Ergebnis dieses Kostenvergleichs wird hier danach aber zusätzlich auch der Deanonymisierungsnutzen erörtert.

10.1.5.1 Deanonymisierungskosten

Bei Betrachtung der Elemente des allgemeinen Kostengerüsts aus Kapitel 7 erkennt man, daß einige der dort aufgeführten Komponenten mehr oder weniger stark vom Umfang des Identifikationsfile abhängen, d.h. in diesem Fall von der Anzahl der Ausländer im Melderegisterauszug und damit von der Größe des gesamten Auszugs. Wesentlich ist die genannte Abhängigkeit insbesondere für die Kosten der Übernahme des Zusatzwissens auf geeignete maschinenlesbare Datenträger, falls der Auszug nicht bereits in einer adäquaten solchen Form geliefert wird. Einen großen Einfluß hat der Auszugsumfang auch auf die Kosten der Beschaffung des Zusatzwissens.

Es soll daher nun zuerst die notwendige Größe des Melderegisterauszugs in Abhängigkeit von ihren Bestimmungsfaktoren erörtert werden. Bezeichnet man diesen Umfang mit R, den vom Angreifer gewünschten Umfang seiner geplanten Stichprobenerhebung mit n, die Anzahl der deanonymisierten Datensätze, denen diese Stichprobe entnommen werden soll, mit N, den Auswahlsatz n/N der Stichprobe mit f, den Anteil der Ausländer im Registerauszug mit a und schließlich die auf die Anzahl A der Ausländer in R, d. h. auf den Identifikationsfile-Umfang bezogene Reidentifikationsquote mit p, so gilt

$$n = fN = fpA = fpaR$$

und damit

$$(*) \qquad R = \frac{n}{fap} \quad .$$

Diese Gleichung beschreibt formal den gesuchten Zusammenhang zwischen Auszugsumfang R und den ihn bestimmenden Variablen n, f, a sowie p und stellt natürlich auch inhaltlich eine plausible Aussage dar: Ein größerer Stichprobenumfang n erfordert bei gleichbleibendem Auswahlsatz f einen höheren Umfang N der Auswahlgrundlage, bei konstanter Reidentifikationsquote p daher einen höheren Identifikationsfile-Umfang A und somit bei konstantem Ausländeranteil a auch einen größeren Auszugsumfang R. Ebenso leuchtet ein, daß R mit ceteris paribus wachsendem f, a oder p jeweils abnimmt:

Ein zunehmender Auswahlsatz bedeutet einen geringeren Überhang von N über n und damit bei konstantem n ein kleineres N, was letztlich wiederum den notwendigen Auszugsumfang vermindert; ferner reicht bei einer größeren Ausländerquote ein geringerer Gesamtumfang des Melderegisterauszugs aus; eine höhere Reidentifikationsquote erfordert schließlich ein kleineres Identifikationsfile und somit bei konstantem a einen geringeren Auszugsumfang. Zu erwähnen ist noch, daß Gleichung (*) nicht gilt, wenn f, a oder p den Wert 0 annehmen. Natürlich ist in solchen Fällen allerdings auch keine Stichprobenziehung bzw. Deanonymisierung durchführbar, so daß sich dieses formale Problem nicht stellt.

Von den Bestimmungsgrößen des Auszugsumfangs wird sich der Angreifer gemäß seinen angestrebten wissenschaftlichen Zielen Stichprobenumfang und Auswahlsatz vorgeben. Die Ausländerquote hängt dagegen von der Weitergabepraxis der Meldebehörden ab. Der für den Angreifer günstigste Fall liegt vor, wenn er einen reinen Ausländerauszug erhält, so daß ihm hier keine Kosten für "unbrauchbare" Datensätze entstehen. Dann wäre also das Identifikationsfile mit dem Auszug identisch, d. h. a = 1 und A = R. Liefern die Behörden als Auszug jedoch nur eine unabhängig von der Staatsangehörigkeit aus dem Register gezogene Stichprobe, so ist im Identifikationsfile eine Ausländerquote zu erwarten, die derjenigen im gesamten Melderegister entspricht. Die Reidentifikationsquote schließlich ist - wie früheren Ausführungen entnommen werden kann - insbesondere abhängig von Anzahl und Informationsgehalt der Überschneidungsmerkmale, Dateninkompatibilitäten, der Anzahl der in Mikrozensus und Identifikationsfile gleichzeitig enthaltenen Personen sowie vom Kenntnisstand des Angreifers über die Teilnahme am Mikrozensus.

Zur Beurteilung der möglichen Größenordnung des benötigten Auszugsumfangs R werde nun in (*) zunächst die Reidentifkationsquote p betrachtet und dabei auf die Ergebnisse der Zuordnungssimulation mit "Kürschners Deutschem Gelehrtenkalender 1987" aus Szenario 1 (s. Kapitel 9) zurückgegriffen.

Ein Vergleich zwischen den Überschneidungsmerkmalen des hier diskutierten Szenarios und Szenario 1, die nicht beiden gemeinsam sind, zeigt, daß den jetzigen Merkmalen Familienstand, Staatsangehörigkeit und Zweitwohnung in Kürschners Gelehrtenkalender als insgesamt wohl weitaus informativere etwa Beruf, Wirtschaftszweig und Fach des letzten Hochschulabschlusses gegenüberstehen. Zudem wurden dort auch noch die Stellung im Beruf und die überwiegend ausgeübte Tätigkeit verwendet. Demzufolge wird die dortige

Quote der korrekten Zuordnungen von 0,0005 wohl deutlich über der jetzigen liegen, wenn man hier ebenso den Normalfall der Unkenntnis des Angreifers über Mikrozensus-Teilnahmen von Identifikationsfile-Personen unterstellt. Gleichwohl sei im folgenden eine zehnmal höhere Quote $p = 0{,}005$ vorausgesetzt. Man nehme hierzu beispielsweise doch eine - wenn auch eher unwahrscheinliche - Kenntnis über die Mikrozensus-Teilnahme hinreichend vieler Identifikationsfile-Personen an.

Ferner werde ein für sozialwissenschaftliche Untersuchungen realistischer Stichprobenumfang $n = 1000$ vorgegeben. Unter diesen Voraussetzungen geht (*) über in

$$R = \frac{200\,000}{fa} .$$

Eine weitere entscheidende Größe ist nun f, der Auswahlsatz der Stichprobe aus den deanonymisierten Fällen. Dieser Auswahlsatz muß als deutlich kleiner als 1 unterstellt werden, denn die Nutzung eines reidentifizierten Mikrodatenfile macht nur Sinn, wenn die Stichprobenziehung tatsächlich selektiv nach darin enthaltenen Informationen erfolgt. Wenn wir nun mit einer sehr hoch angesetzten Obergrenze davon ausgehen, daß der Angreifer plant, ca. die Hälfte der deanonymisierten Datensätze in seine Stichprobe aufzunehmen, daß er also $f = 0{,}5$ ansetzt, so benötigt er in dem für ihn äußerst günstigen, aber wohl unwahrscheinlichen Fall der Übermittlung eines reinen Ausländerauszugs ($a = 1$) durch die Meldeämter demnach den Auszugsumfang $R = 400\,000$. Aber selbst bei Verwendung fast aller reidentifizierter Datensätze müßte ein reiner Ausländerauszug - unter Beachtung des rechnerischen Grenzfalls $f=1$ - immerhin mindestens den Umfang $R = 200\,000$ aufweisen. Eine Halbierung des Stichprobenumfangs auf $n = 500$ wiederum würde gemäß (*) auch den Auszugsumfang reduzieren, im zuletzt genannten Fall also auf $R = 100\,000$.

Im Zusammenhang mit dem notwendigen Umfang des Registerauszugs soll im übrigen ein weiterer, bisher ausgeklammerter Aspekt wenigstens noch angesprochen werden: die Ausschöpfung der Stichprobe. Ein Phänomen stellt in der empirischen Sozialwissenschaft die im Vergleich zum Bevölkerungsdurchschnitt geringere Erreichbarkeit von Ausländern bei Umfragen dar. Man muß davon ausgehen, bestenfalls von 50 Prozent der in eine Stichprobe einbezogenen Ausländer tatsächlich ein Interview zu erhalten. Um eine gewünschte Anzahl von Antworten im Rahmen seiner Erhebung zu erzielen,

müßte der Angreifer also eine entsprechend größere Anzahl von Personen auswählen, was einen größeren Umfang N der Auswahlgrundlage, also ceteris paribus einen kleineren Auswahlsatz f und somit gemäß Gleichung (*) auch einen höheren Umfang R erfordert: er müßte aufgrund einschlägiger Erfahrungen vermutlich mindestens verdoppelt werden.

Insgesamt kann man damit wohl von wenigstens einigen Hunderttausend benötigten Datensätzen aus den Melderegistern ausgehen, wobei nochmals auf die unterstellten günstigen Rahmenbedingungen für den Angreifer hingewiesen sei. Der auch erwähnte Umfang R = 100 000 stellt dann also eine recht tief liegende untere Schranke dar.

Für die Kostendiskussion werden im folgenden exemplarisch die oben genannten R = 400 000 benötigten Melderegister-Datensätze als Ziel vorausgesetzt. Der dort dabei unterstellte vom Angreifer angestrebte Stichprobenumfang n = 1 000 beim Auswahlsatz f = 0,5 erfordert N = 2 000 Deanonymisierungen.

Zunächst erscheint es jedoch unabhängig von irgendwelchen Kostenüberlegungen sehr zweifelhaft, ob der Angreifer überhaupt eine derart hohe Zahl an Datensätzen von den Meldebehörden erhält, selbst wenn er seine Anfragen auf viele Ämter verteilt. Dabei sei noch einmal erwähnt, daß der Angreifer zur Ausnutzung seiner etwaigen regionalen Kenntnisse über Mikrozensus-Teilnahmen zudem sogar einen Schwerpunkt auf bestimmte Gemeinden legen möchte. Hinzuweisen ist nun wieder auf die Ausführungen über Melderegister in Kapitel 6: Danach würden zum einen an Daten über eine Vielzahl namentlich nicht bezeichneter Personen (allgemeine Gruppenauskünfte) in der Regel nur Name und Anschrift weitergegeben, nicht aber alle in 10.1.2 aufgeführten Überschneidungsmerkmale. Ferner müsse den Behörden ein öffentliches Interesse an dem Forschungsvorhaben nachgewiesen werden. Es dürfte aber äußerst schwierig sein zu belegen, daß für eine sozialwissenschaftliche Untersuchung ein Melderegisterauszug von 400 000 Fällen erforderlich ist. Dies widerspricht aller üblicher Praxis. Schließlich wurde in Kapitel 6 noch festgehalten, daß sich Meldeämter bei Anfragen über bestimmte abgrenzbare Gruppen, hier also über Ausländer, allgemein besonders restriktiv verhielten.

Insgesamt erscheint es also ziemlich wahrscheinlich, daß die Deanonymisierungsaktion unabhängig von ihren Kosten bereits an dem zu großen benötigten Registerauszug scheitert. Trotzdem sollen die einzelnen Kostenkomponenten angesprochen werden.

Für die grobe Abschätzung der Kosten der Beschaffung des Zusatzwissens liefert ein sozialwissenschaftliches Projekt aus dem Jahre 1985 einen Anhaltspunkt. Dort werden für die Übermittlung von 3 300 Adressen Jugendlicher aus Melderegistern von 200 Gemeinden ca. 30 000 DM angegeben (Allerbeck/Hoag 1985:11,18). Die proportionale Umrechnung auf den hier zugrunde gelegten reinen Ausländerauszug ergibt eine Größenordnung von 3,6 Millionen DM. Geht man wegen des großen Umfangs der zu liefernden Adressen von einer Kostendegression aus und setzt die damit erzielbaren Einsparungen mit 50 Prozent hoch an, so würden die reinen Kosten für die Registerauszüge 1,8 Mio. DM betragen.

Zur Einschätzung der Kosten für die Anpassung der Überschneidungsmerkmale beider Files und für die Datenübernahme auf geeignete maschinenlesbare Träger in geeigneter Weise wird auf die empirischen Ergebnisse aus Szenario 1 zurückgegriffen. Dort resultierten dafür bei ca. 8 000 Identifikationsfile-Datensätzen etwa 44 000 DM. Während aber jetzt die Anzahl der Identifikationsfile- Datensätze mit 400 000 ungefähr 50-mal höher ausfällt, kann man nun für die Anzahl der Überschneidungsmerkmale etwa nur 80 Prozent gegenüber Szenario 1 ansetzen. Mit dem Gesamtfaktor 50 x 0,8 = 40 führte dann die proportionale Umrechnung diesmal zu Kosten in Höhe von 40 x 44 000 DM = 1,76 Millionen DM. Die Übernahmekosten könnten sich jedoch in dem Maße verringern, in dem Meldebehörden Auszüge von vorneherein bereits in geeigneter maschinenlesbarer Form lieferten. Gleichwohl verblieben auch in solchen Fällen im allgemeinen noch Rekodierungskosten. Diese wiederum können abnehmen, wenn auch Umsignierungen maschinell durchführbar sind. Insgesamt werden sich aber allein schon wegen der unterschiedlichen Weitergabepraxis der einzelnen Meldeämter die Aufnahme- und Anpassungskosten nicht allzu weit reduzieren lassen.

Auf Überlegungen zu Datenfehlern in den beiden Files und mögliche Korrekturmaßnahmen wird der Angreifer angesichts mangelnder Information und mit Blick auf die relative Einfachheit der hier verwendeten Überschneidungsmerkmale kaum Zeit verwenden, so daß daraus resultierende Kosten wohl zu vernachlässigen sind.

Aussagen zu den Programmierungskosten sind im Prinzip aus deren allgemeiner Erörterung zu übernehmen, da sich für das jetzige Szenario keine wesentlichen speziellen Aspekte ergeben: Sie hängen von der Komplexität des verwendeten Deanonymisierungsalgorithmus ab und mögen daher, etwa bei einfachen Abgleichverfahren, nur schwach ins Gewicht fallen oder bis zu einer Höhe von mehreren Personen-Arbeitsmonaten anfallen, wie bei der

Programmierung des Verfahrens von Paaß/Wauschkuhn (1985). Angesichts der in Kapitel 9 dargestellten empirischen Ergebnisse zur Erfolgsquote beider Verfahren sei hier jedoch die Verwendung des einfachen Algorithmus, der dort bessere Resultate erzielt hatte, unterstellt, was also eher geringe Programmierungskosten bedeutet.

Zur Einbeziehung der Anwendung des Programms, d.h. seiner Implementation und seines Ablaufs, in die Kostendiskussion betrachte man ebenfalls zunächst wieder das Ergebnis für Szenario 1 in Kapitel 9. Dort fielen die Kosten für die Entwicklung des Programms und seine Anwendung zwar deutlich niedriger aus als diejenigen zur Aufnahme und Rekodierung des Zusatzwissens in Höhe von ca. 44 000 DM, konnten aber mit immerhin rund 17 000 DM gleichfalls nicht vernachlässigt werden. Letzteres gilt dann sicher auch jetzt, wo mit 400 000 etwa 50mal so viele, wenn auch etwas kürzere Identifikationsfile-Datensätze zu verarbeiten sind.

Verschleierungskosten sind ebenso wegen der in dem jetzigen Szenario großen Datenmenge zu beachten. Falls technisch-organisatorische Sicherungsmaßnahmen bestehen, können solche Kosten erheblich sein, wenn der Angreifer die Deanonymisierung auf Rechenanlagen seines Forschungsinstituts vornehmen will. Aber auch Maßnahmen etwa gegen die Entwendung von Datenträgern zur Bearbeitung auf einer privaten Anlage können wohl im allgemeinen soweit ausgebaut werden, daß daraus resultierende Verschleierungskosten zur Umgehung der Vorkehrungen im Kalkül des Angreifers ein nicht zu geringes Gewicht erhalten.

Wir wollen diese schwer quantifizierbaren Aspekte wie im übrigen auch potentielle Sanktionen und möglichen Nutzenentgang im Fall der Aufdeckung hier jedoch nicht weiter berücksichtigen, da die absehbaren Kosten auch ohne diese Komponenten schon eine beachtliche Größenordnung insbesondere im Vergleich zu alternativen Datenbeschaffungsmethoden (vgl. den folgenden Abschnitt) erreichen.

Schließlich sei noch darauf hingewiesen, daß Versuche des Wissenschaftlers, Deanonymisierungskosten auf sein Forschungsinstitut zu überwälzen, hier als eher unwahrscheinlich einzuschätzen sind, da die große zu verarbeitende Datenmenge sonst wohl zu einer für den Angreifer zu großen Aufdeckungsgefahr führte.

10.1.5.2 Alternative Datenbeschaffung

Als alternative Beschaffungsmöglichkeit für die Auswahlgrundlage der geplanten Ausländerstichprobe bietet sich eine eigene Erhebung des potentiellen Angreifers an, wenn man von speziellen Forschungsmotiven, wie etwa dem oben genannten Methodenvergleich zwischen Erhebungen mit und ohne Auskunftspflicht absieht, für die amtliche Mikrodaten als notwendig unterstellt werden können. Legt man nun etwa die in den Sozialwissenschaften wohl realistische Größenordnung von ca. 150 DM als Kosten für ein Interview zugrunde, so resultieren für die als Auswahlgrundlage gewünschten 2 000 Datensätze ungefähr 300 000 DM. Ein Vergleich mit der im vorigen Abschnitt begründeten, eher niedrig eingeschätzten Größenordnung von 1,8 Millionen DM allein zur Beschaffung des zur Deanonymisierung benötigten Zusatzwissens zeigt, daß dann die faktische Anonymität des Mikrozensus gewährleistet ist. Dies gilt natürlich erst recht bei Hinzurechnung der ebenfalls im vorigen Abschnitt begründeten möglichen Größenordnung von ca. 1,75 Millionen DM an Kosten für die Rekodierung des Zusatzwissens und seine adäquate Übernahme auf maschinenlesbare Datenträger. Schon diese, bisher genannten Komponenten bedeuten dann nämlich über drei Millionen DM an Deanonymisierungskosten. Deren verbleibende Komponenten wie Programmentwicklungs- und -anwendungskosten, Verschleierungs-, Sanktions- und Opportunitätskosten sind bei dieser Gegenüberstellung überhaupt nicht berücksichtigt.

10.1.5.3 Deanonymisierungsnutzen

Sieht man von den genannten speziellen Forschungsmotiven wie einem Methodenvergleich zwischen Pflichterhebung und freiwilliger Auskunftserteilung ab, besteht somit wegen des deutlichen Kostennachteils von Reidentifikationen gegenüber alternativen Datenbeschaffungsmöglichkeiten kein Bedarf mehr an der Erörterung des Deanonymisierungsnutzens.

Für die angesprochenen speziellen Forschungsmotive, für die amtliche Mikrodaten als notwendig unterstellt sind, ist darauf hinzuweisen, daß sich wegen starker Verzerrungen, mit denen bei einer Deanonymisierung gerechnet werden müßte, eine so gewonnene Stichprobe wohl ohnehin kaum für derartige Forschungsarbeiten eignen würde. Die als Auswahlgrundlage dienenden deanonymisierten Datensätze bilden nämlich zum einen keine repräsentative Stichprobe aus dem Identifikationsfile, da z.B. Personen mit herausragenden Ausprägungen in den Überschneidungsmerkmalen überproportional

reidentifiziert werden könnten. Damit stellt natürlich ebenso die eigentliche Stichprobenerhebung des Angreifers bezogen auf die interessierende Grundgesamtheit keine repräsentative Auswahl dar. Die Verzerrungen würden zum anderen noch verstärkt, wenn sich der Wissenschaftler mit dem Ziel einer möglichst hohen Reidentifikationsquote beim Aufbau seines Identifikationsfile vorwiegend auf solche Fälle konzentrierte, deren Mikrozensus-Teilnahme ihm bekannt ist. Insgesamt ist wohl von einer so starken Verzerrung auszugehen, daß der potentielle Angreifer aus fachlichen Gründen auf die Deanonymisierung verzichten wird.

10.1.5.4 Fazit

Nach den oben aufgeführten Begründungen für einen Verzicht eines möglichen Angreifers auf Reidentifikationsversuche bei der angestrebten Gewinnung einer Auswahlgrundlage für eine Ausländerstichprobe läßt sich eine allgemeine Schlußfolgerung ziehen: Da die Gewinnung einer Auswahlgrundlage, wie in 10.1.1 erwähnt, nahezu die einzige denkbare sozialwissenschaftliche Motivation für Deanonymisierungen bildet[2] und hier eine solche spezielle Situation mit relativ guten Reidentifikationsbedingungen betrachtet wurde, kann man davon ausgehen, daß ein solcher Verzicht eines potentiellen Angreifers für einen großen Teil, wenn nicht gar für alle sozialwissenschaftlich motivierten Deanonymisierungsszenarien gewährleistet ist.

Während die Ablehnung für die angesprochenen speziellen Motive, für die also amtliche Mikrodaten als notwendig unterstellt waren, wie bei einem angestrebten Methodenvergleich zwischen einer Pflichterhebung und freiwilliger Auskunftserteilung, im vorigen Abschnitt mit fachlichen Erwägungen eines möglichen Angreifers begründet wurde, war sie sonst bei einer allgemeinen

---

2 Ein weiteres denkbares wissenschaftliches Deanonymisierungsmotiv, die Ergänzung eigener Daten mit Informationen aus einem Mikrodatenfile, wurde schon im Kapitel über das Zusatzwissen kurz angesprochen. Auch dieses Motiv stellte sich jedoch als untauglich heraus wegen zu geringer Reidentifikationsquoten, die durch die Zuordnungssimulationen bekräftigt wurden; solche Zusatzinformationen könnten allenfalls für einen so extrem kleinen Bruchteil der Untersuchungspersonen gewonnen werden, daß ein derartiges Vorhaben von vorneherein wissenschaftlich als unergiebig und nutzlos zu betrachten ist. Im übrigen wäre es für eine solche Informationsergänzung gar nicht nötig, Namen oder Adressen von Identifikationsfile-Personen amtlichen Mikrodaten zuzuordnen. Ausreichend wäre vielmehr schon die Zuordnung amtlicher Mikrodaten zu den um die Identifikatoren verkürzten Identifikationsfile-Datensätzen.

Verbesserung oder Einschränkung einer Auswahlgrundlage allein schon auf deutlich günstigere alternative Datenbeschaffungsmethoden zurückzuführen.

10.2 Angriffsszenario 4: Gewinnung von Informationen über eine Persönlichkeit des öffentlichen Lebens mit dem Ziel der anonymen Publikation

10.2.1 Beschreibung des Szenarios

In diesem Szenario wird folgende Ausgangslage angenommen: Ein Mitarbeiter eines sozial-wissenschaftlichen Instituts hat sich zum Ziel gesetzt, aus dem übermittelten Mikrodatenfile der amtlichen Statistik Datensätze über eine prominente Person herauszufischen, dann zu deanonymisieren und schließlich die gewonnenen Erkenntnisse anonym zu veröffentlichen. Das hinter diesen Absichten stehende Motiv kann etwa in der Schadenfreude über die Offenlegung oder der Enthüllung privater Daten solcher Personen bestehen, die einer vom Angreifer möglicherweise besonders kritisch betrachteten Gruppe der Gesellschaft angehören, oder ähnlichem.

Als geeignete Reidentifikationsstrategie dafür erscheint der Einzelfischzug, bei dem das Mikrodatenfile zunächst nach "interessanten" Datensätzen durchsucht und dann versucht wird, wieder einen Personenbezug herzustellen. Es kommt also nicht darauf an, ganz bestimmte prominente Personen zu reidentifizieren. Ziel ist vielmehr, daß sich überhaupt ein Erfolg einstellt (Paaß/Wauschkuhn 1985:46).

Der Einzelfischzug ist wegen des Verzichts auf Forderungen nach mehreren Deanonymisierungserfolgen und nach Reidentifikationen vorgegebener Personen die erfolgversprechendste Deanonymisierungsstrategie, wenn das Mikrodatenfile Merkmale bzw. Merkmalsausprägungen enthält, die es ermöglichen, aus dem üblichen Rahmen herausfallende Personen oder Haushalte herauszufiltern. Darum eignet sich die Strategie auch am ehesten für den Versuch, nachzuweisen, daß die Anonymisierungsmaßnahmen der amtlichen Statistik prinzipiell nicht ausreichend seien. Hinsichtlich der Deanonymisierungsstrategie und der dabei entstehenden Kosten ist dieses Motiv daher unter dem hier angesprochenen Sachverhalt subsumierbar.

10.2.2 Identifikations- und Mikrodatenfile

Ein wichtiger Grund, der gegen die Durchführung eigener empirischer Experimente für dieses Szenario spricht, sind die Kosten für den Aufbau des dafür benötigten Identifikationsfile. Wenn aus den Variablen des Mikrodatenfile nicht relativ leicht spezielle kleine Gruppen isoliert werden können, z. B. bekannte Politiker, Schriftsteller oder Schauspieler, so ist ein hoher Aufwand zum Aufbau des Zusatzwissens notwendig. Beim Mikrozensus liegt dieses Problem vor, da seine Variablen die Beschränkung auf eine klar umrissene und zahlenmäßig eng begrenzte spezielle Gruppe prominenter Personen, z. B. bekannte Filmschauspieler, im allgemeinen nicht zulassen. Der Aufbau einer EDV-gerechten Datei durch Sammlung aus öffentlich zugänglichem Material, wie Handbüchern, Registern, Fachzeitschriften, Presseartikeln usw. war bei den gegebenen zeitlichen und finanziellen Restriktionen im Rahmen des Projekts jedoch unmöglich.

Unterstellt seien nun folgende Überschneidungsmerkmale zwischen Mikrozensus und dem aus den genannten Quellen sowie ohnehin schon vorhandenen Vorkenntnissen des Angreifers zusammengestellten Identifkationsfile:

- Geschlecht
- Familienstand
- Staatsangehörigkeit
- Geburtsjahr einschließlich Geburtsjahreshälfte
- Eheschließungsjahr
- Beruf
- Wirtschaftszweig
- Stellung im Beruf
- Überwiegend ausgeübte Tätigkeit.

Zusätzlich soll noch die mögliche Lieferung eines Mikrozensus-Regionalmerkmals berücksichtigt werden, das sich aus Gemeindegrößenklasse und Bundesraumordnungsregion zusammensetzt.

10.2.3 Abgrenzungs- und Fehlerproblematik

Wird ausgehend von bestimmten "exotischen" Ausprägungen im Mikrodatenfile versucht, ein passendes Identifikationsfile aus öffentlich zugänglichen Informationen aufzubauen, so muß das Material also im allgemeinen aus den verschiedensten, exemplarisch bereits aufgeführten Quellen zusammengestellt werden. Entsprechend groß sind Diskrepanzen in Merkmalsabgrenzungen und Unterschiede in den Zeitpunkten, auf die sich die einzelnen Informationen beziehen. Auch Art und Höhe der Fehler sind daher recht unterschiedlich, und der Angreifer verfügt üblicherweise wohl kaum über verwertbare Erkenntnisse darüber. Günstiger wäre es für ihn dagegen, wenn aus dem Mikrodatenfile spezielle kleine Personengruppen, zum Beispiel Bundestagsabgeordnete, herausgefiltert werden können, für die Spezialinformationen bereits gesammelt vorliegen, etwa in Form von Handbüchern. Dann besteht eine Ähnlichkeit zur Abgrenzungs- und Fehlerproblematik beim ersten Szenario.

10.2.4 Praktische Durchführung

Das Szenario wird auf argumentativer Basis dargestellt und beurteilt. Anhaltspunkte sind dabei die Ergebnisse der Zuordnungsexperimente aus der ersten Situation mit "Kürschners Deutschem Gelehrtenkalender 1987" als Identifikationsfile. Im übrigen wurde der Mikrozensus hinsichtlich der Frage durchforstet, ob mit seinen Variablen die genannte Fischzug-Strategie überhaupt anwendbar erscheint.

10.2.5 Kosten-Nutzen-Analyse

10.2.5.1 Mögliche Vorgehensweise des Angreifers

Das hier zu diskutierende Deanonymisierungsszenario weist zweifellos Parallelen zum Journalisten-Szenario von Paaß/Wauschkuhn (1985:197f.) auf, da auch dort ein Einzelfischzug aus "extremen" Haushalten Gegenstand der Diskussion war.

Dazu wurden dort aus der als Mikrodatenfile dienenden Einkommens- und Verbrauchsstichprobe diejenigen 25 Datensätze herausgesucht, deren Merkmalsausprägungen nach gewissen Kriterien insgesamt am stärksten von

Durchschnittswerten abwichen. Die so gewonnenen 25 Haushalte bildeten dann das Identifikationsfile. Der Grund für die Konzentration auf "extreme" Haushalte lag in den höheren Deanonymisierungschancen (vgl. Kapitel 4).

Auch der Angreifer könnte nun in ähnlicher Weise versuchen, zunächst das Mikrodatenfile, in diesem Fall also den Mikrozensus, nach prominenten Personen zu durchforsten und diesen dann Datensätze aus dem noch aufzubauenden Identifikationsfile zuzuordnen. Diese Methode hätte den Vorteil, daß Gewißheit über die Mikrozensus-Teilnahme der gesuchten Personen besteht. Je nach dem Grad der Vorinformation des Angreifers würde sich dadurch die Erstellung des Identifikationsfile vereinfachen, da eine Konzentration auf die "Mikrozensus-Prominenten" möglich wäre. Der Schutz durch die Stichprobeneigenschaft des Mikrodatenfile nähme in entsprechendem Maße ab.

Zu fragen ist also, inwieweit das Erhebungsprogramm des Mikrozensus die Herausfilterung prominenter Personen zuläßt. Die einzige Möglichkeit besteht für den Angreifer wohl darin, das Merkmal Beruf auf solche Ausprägungen hin zu untersuchen, wie sie für derartige Personen typisch sind, z. B. "Filmschauspieler" oder "Bundestagsabgeordneter". Für den künstlerischen Bereich weist die Mikrozensus-Berufsklassifizierung nur eine recht grobe Merkmalsgliederung aus. Die Ausprägung "Darstellende Künstler" umfaßt z. B. so unterschiedliche Berufe wie Spielleiter, Schauspieler, Gesangssolisten, Chorsänger und Gruppentänzer, so daß daraus kaum auf den Bekanntheitsgrad oder gar bestimmte Personen zu schließen ist. Eine für den Angreifer interessantere Abgrenzung bildet dagegen die Ausprägung "Abgeordnete, Minister, Wahlbeamte"[3], da diese Personen im Prinzip alle als mehr oder weniger prominent angesehen werden können. In Verbindung mit einer nicht allzu weit gefaßten Mikrozensus-Regionalangabe fielen Reidentifizierungen etwa unter Zuhilfenahme von Abgeordneten-Handbüchern vielleicht nicht mehr allzu schwer.

Fehlt dagegen im Mikrozensus der Regionalbezug, so vergrößert sich der Kreis möglicher zugehöriger Personen deutlich und damit auch der Umfang des benötigten Identifikationsfile. Eventuell wird der Angreifer nun ohne Berücksichtigung einzelner Mikrozensus-Datensätze versuchen, Informationen über bekannte Personen zu erhalten, vielleicht besonders zu solchen, über die

---

3 Unter diese Ausprägung fallen Abgeordnete, Präsidenten, Minister, Senatoren, Wahlbeamte und politische Beamte. Der Mikrozensus (1987) für Nordrhein-Westfalen enthält in dieser Kategorie zwölf Fälle.

er ohnehin schon Vorkenntnisse besitzt oder von denen er sich relativ hohe Reidentifikationswahrscheinlichkeiten verspricht. In der Regel entfällt dann allerdings auch die Information über eine mögliche Mikrozensus-Teilnahme. Kenntnisse darüber könnten dem Angreifer zwar etwa aus seiner Eigenschaft als Regionalwissenschaftler erwachsen, z. B. durch Hinweise von Mikrozensus-Teilnehmern selbst oder von Gruppierungen, die sich die Durchleuchtung, Behinderung oder Diskreditierung der amtlichen Statistik zum Ziel gesetzt haben. Solche Informationen dürften aber recht mager ausfallen, da dem Angreifer ja nicht mit irgendwelchen, sondern nur mit prominenten Personen gedient ist und diese natürlich auch unter den Mikrozensus-Teilnehmern entsprechend selten vertreten sind.

10.2.5.2 Deanonymisierungskosten

Eine wichtige Rolle für die Deanonymisierungskosten spielt auch diesmal wieder der notwendige Umfang des Identifikationsfile und damit die darauf bezogene Reidentifikationsquote. Zu ihrer groben Abschätzung werde wieder auf die Ergebnisse des ersten Szenarios mit "Kürschners Deutschem Gelehrtenkalender 1987" als Identifikationsfile zurückgegriffen.

Man betrachte zunächst den Fall, daß die amtliche Statistik dem Angreifer das in Abschnitt 10.2.2 aufgeführte Regionalmerkmal diesmal nicht mitliefert. An nicht gemeinsamen Überschneidungsmerkmalen weist dann die hier betrachtete Situation Familienstand, Staatsangehörigkeit und Eheschließungsjahr auf, Situation 1 demgegenüber das Fach des letzten Hochschulabschlusses und das genannte Regionalmerkmal. Dabei erscheint neben dem Familienstand auch die Staatsangehörigkeit relativ wenig informativ, wenn der Angreifer nicht von bestimmten prominenten Ausländern weiß, daß sie am Mikrozensus teilgenommen haben. Im Fall der Unkenntnis besteht nämlich neben dem Mikrozensus dann auch das vom Angreifer aufzubauende Identifikationsfile wohl überwiegend aus Personen deutscher Staatsangehörigkeit. Die Beschaffung der Angaben über das Eheschließungsjahr schließlich dürfte in der Praxis oft sehr schwierig und mühsam sein, so daß in vielen Fällen vielleicht darauf verzichtet würde. Insgesamt wird wohl der Beitrag von Familienstand, Staatsangehörigkeit und Eheschließungsjahr durch den Informationsgehalt der Gelehrten-Merkmale 'Fach des letzten Hochschulabschlusses' sowie 'Bundesraumordnungsregion' und 'Gemeindegrößenklasse' deutlich übertroffen.

Im Normalfall der Unkenntnis des Angreifers über Mikrozensus-Teilnahmen ihm bekannter Personen stellt damit die Quote der korrekten Zuordnungen im Gelehrtenszenario in Höhe von etwa 0,0005 auch im jetzigen Szenario eine obere Schranke dar. Legt man sie zugrunde, so werden zur Deanonymisierung einer, wenngleich beliebigen, prominenten Person im Durchschnitt 2 000 Datensätze im Identifikationsfile benötigt.

Während im Gelehrten-Szenario praktisch keine Beschaffungskosten für das Zusatzwissen anfielen, muß jetzt ein Teil des Identifikationsfile aus verschiedenen Quellen, etwa auch Pressearchiven zusammengestellt werden, da die Daten im Gegensatz zu dort nicht alle in einem Handbuch vorliegen, sondern teilweise Einzelinformationen darstellen. Verwiesen sei dazu insbesondere auf die Merkmale Familienstand, Eheschließungsjahr und Geburtsjahr mit Geburtsjahreshälfte. Geht man also etwa davon aus, daß von den bei neun Überschneidungsmerkmalen und 2 000 Datensätzen insgesamt 9 x 2 000 = 18 000 Einzeldaten ca. 3 x 2 000 = 6 000 auf diese Weise beschafft werden müssen, und setzt pro Stunde einschließlich Reisezeit beispielsweise den recht hohen Wert von sechs Sucherfolgen an, so erfordert die Beschaffung zusammen 6 000/6 = 1 000 Stunden. Bei einer Bewertung von 50 DM/Stunde resultieren daraus dann 50 000 DM. Da ferner Signierung und Übernahme auf maschinenlesbare Träger für die 8 000 Gelehrten-Datensätze ca. 44 000 DM erforderten, können jetzt dafür als Untergrenze mindestens 10 000 DM angesetzt werden.

Außerdem betrugen in dem Gelehrten-Szenario die Kosten zur Erstellung bzw. Beschaffung und Implementierung eines einfachen Abgleichverfahrens, die wohl nur unterproportional mit dem Umfang des Identifikationsfile wachsen, sowie die eigentlichen Rechenkosten zusammen ca. 17 000 DM. Daher erscheint jetzt etwa eine Größenordnung von 10 000 DM plausibel. Somit ergeben die bisherigen Kostenkomponenten insgesamt ca. 70 000 DM.

Zusätzlich fallen sicher auch diesmal wegen der großen - wenngleich etwa gegenüber Szenario 3 geringeren - Datenmenge nicht zu vernachlässigende Verschleierungskosten an. Zu beachten ist dabei, daß der Angreifer bei dem hier unterstellten Motiv der anonymen Publikation durch die Verschleierung nicht nur die Vereitelung seines Vorhabens verhindern will, sondern dieses auch nach dessen Abschluß geheimhalten möchte.

Entgangener Nutzen aufgrund der Beeinträchtigung der zukünftigen beruflichen Laufbahn des Wissenschaftlers sowie Sanktionen gegen ihn jeweils im Fall der Aufdeckung seines Vorhabens als weitere - etwa mit der Auf-

deckungswahrscheinlichkeit gewichtete - Kostenkomponenten könnten diesmal wegen des Deanonymisierungsmotivs der Schädigung höher ausfallen als bei der wissenschaftlichen Verwertung im vorangegangenen Szenario. Ferner ist jetzt der auch recht unwahrscheinliche Fall auszuschließen, daß der Angreifer diese Kosten aus Unkenntnis über die Rechtswidrigkeit seines Vorhabens nicht in sein Kalkül zieht.

Kostenüberwälzungen auf sein Forschungsinstitut wird der Angreifer im übrigen wohl auch diesmal nicht riskieren. Zwar liegt die abzugleichende Datenmenge wegen des geringeren Identifikationsfile-Umfanges unter derjenigen im vorigen Szenario, aber immer noch so hoch, daß eine Aufdeckung des Vorhabens bei hinreichenden Kontrollmaßnahmen dann ziemlich wahrscheinlich erscheint. Verwiesen sei dabei noch einmal auf das jetzige Deanonymisierungsmotiv, das eine Geheimhaltung auch noch nach Verwertung der reidentifizierten Datensätze erfordert.

Nach dieser bisherigen Kostendiskussion betrachte man nun noch den Fall, daß auch das erwähnte Regionalmerkmal im amtlichen Mikrodatenfile zur Verfügung steht. Dann kann der Angreifer durch Konzentration auf Politiker aus dem Mikrozensus und dem eingegrenzten Gebiet, wie beschrieben, die Reidentifikationsquote deutlich erhöhen, wenn er das Identifikationsfile so aufbaut, daß der Anteil der Mikrozensus-Teilnehmer darin möglichst hoch ausfällt. Zu beachten ist dabei, daß allein schon die Kombination von Gemeindegrößenklasse und Regierungsbezirk oft nicht nur die Eingrenzung von einzelnen Gemeinden der ersten, sondern z. B. auch der vierten Kategorie, d.h. mit 50 000 bis 100 000 Einwohnern erlaubt. Dies gilt natürlich erst recht bei Mitlieferung der Bundesraumordnungsregion, wobei dann häufig noch kleinere Gemeinden eindeutig bestimmbar sind.

Geht man also etwa von einer unter solchen Umständen vorstellbaren hohen Reidentifikationsquote in der Größenordnung von 0,25 aus, so wäre hier wegen des nur mindestens einen angestrebten Deanonymisierungserfolgs lediglich ein Identifkationsfile mit vier Datensätzen aufzubauen. In diesem Fall können sich die meisten Kostenkomponenten eventuell so stark reduzieren, daß sie unter Umständen einfach zu vernachlässigen sind. Dies gilt zunächst für die Kosten zur Beschaffung des Zusatzwissens, wobei besonders an regionale oder lokale Informationsquellen über Politiker zu denken ist, aber auch für die Anpassung der gewonnenen Identifikationsfile-Informationen an die Mikrozensus-Definitionen und -Abgrenzungen sowie für die Übernahme auf geeignete maschinenlesbare Datenträger. Dagegen nehmen die Kosten zur Programmerstellung bzw. -beschaffung und -implementierung durch den von

2 000 auf vier Datensätze verkleinerten Umfang des Identifikationsfile wohl nur unterproportional ab und sind für ein einfaches Abgleichverfahren aber dennoch eher niedrig anzusetzen. Ferner verursacht vermutlich der eigentliche Programmablauf wiederum lediglich deutlich reduzierte Kosten gegenüber dem zuvor betrachteten Fall. Im übrigen ist bei den hier unterstellten lediglich vier benötigten Identifikationsfile-Datensätzen auch ein manueller Abgleich denkbar, der zwar zusätzlich Arbeitszeit, dafür aber keine Programm- oder Maschinenkosten erfordert.

Deutlich niedriger als zuvor fallen wahrscheinlich unter sonst gleichen Bedingungen auch die Verschleierungskosten aus. Durch hinreichende Kontroll- und Sicherungsmaßnahmen ließen sie sich aber wohl soweit erhöhen, daß sie nicht zu vernachlässigen wären. Unverändert gegenüber dem Fehlen des Regionalmerkmals bleiben dagegen die potentiell möglichen Sanktionen gegen den Angreifer sowie der ihm potentiell entgehende Nutzen durch Beeinträchtigung seiner beruflichen Laufbahn jeweils bei Aufdeckung des Deanonymisierungsvorhabens. Allerdings ist wegen des äußerst kleinen Umfangs des Identifikationsfile auch mit einer deutlich verminderten Wahrscheinlichkeit dafür zu rechnen, so daß der Wissenschaftler diese potentiellen Kosten entsprechend geringer gewichten könnte.

Im übrigen erscheinen Versuche, Deanonymisierungskosten auf das Forschungsinstitut abzuwälzen auch jetzt wieder eher abwegig. Zwar verringert sich durch den deutlich kleineren Umfang des Identifikationsfile die Wahrscheinlichkeit der Aufdeckung des Vorhabens; andererseits reduzieren sich aber auch, wie dargestellt, die gesamten Deanonymisierungskosten in hohem Maße, was wiederum die Bereitschaft des Angreifers, durch Überwälzungsversuche das Aufdeckungsrisiko zu erhöhen, entsprechend vermindert.

Zusammenfassend ist festzuhalten, daß die Deanonymisierungskosten entscheidend davon abhängen, ob das Regionalmerkmal in Verbindung mit einem sehr differenzierten Berufsmerkmal im Mikrozensus zur Verfügung steht, da dann im zuletzt genannten Fall im Prinzip alle Komponenten bis auf vielleicht diejenigen zur Erstellung bzw. Beschaffung und Implementierung des Reidentifikationsalgorithmus deutlich abnehmen.

10.2.5.3 Alternative Datenbeschaffung und Fazit

Bei dem hier betrachteten Deanonymisierungsmotiv der anonymen Publikation ist besonders wegen der nichtmonetären Verwertungsabsicht eine externe Beurteilung des dem Angreifer erwachsenden Nutzens zum Vergleich mit den Reidentifikationskosten kaum möglich, obwohl der Wissenschaftler seinerseits bei unterstelltem ökonomisch-rationalen Verhalten eine solche Abwägung vornähme.

Eine allgemeine Bewertung der Situation erscheint vielmehr nur durch eine Gegenüberstellung von Deanonymisierungen und alternativen Datenbeschaffungsmethoden sinnvoll. Unterstellt man dazu ein weiteres Mal das schlichte Befragen der interessierenden Personen als eine solche Alternative und setzt dabei wieder die für sozialwissenschaftliche statistische Erhebungen realistische Größenordnung von 150 DM pro Interview an, so ergeben sich, da ein Antworterfolg hier ausreicht, bei vereinfachend als konstant angenommenen Durchschnittskosten auch insgesamt nur Kosten in dieser Höhe.

Diese lägen deutlich unter den im vorigen Abschnitt erläuterten Deanonymisierungskosten, wenn das Regionalmerkmal aus 10.2.2 im Mikrozensus nicht ausgewiesen wird. Dies gilt auch für den Fall, daß man potentielle Sanktionen gegen den Angreifer und den ihm möglicherweise entgehenden Nutzen nicht berücksichtigt. Denn selbst unter diesen Voraussetzungen ergäben sich noch Deanonymisierungskosten in einer Größenordnung von ca. 70 000 DM. Auch bei einem höheren Ansatz als den oben genannten 150 DM für eine alternative Datenbeschaffung verbleibt damit noch ein so großer Spielraum bis zur Erreichung dieser Reidentifikationskosten, daß die faktische Anonymität hier gesichert erscheint.

Steht dem Angreifer dagegen das Regionalmerkmal im Mikrozensus zur Verfügung, sind solch eindeutige Aussagen nicht mehr möglich. Zwar werden die Deanonymisierungkosten bei Einbeziehung potentieller Sanktionen gegen den Angreifer und des ihm vielleicht entgehenden Nutzens vermutlich weiterhin über den genannten 150 DM für eine Befragung liegen. Nimmt man diese Komponenten aber aus der Gegenüberstellung heraus, verbleiben, wie im vorigen Abschnitt begründet, eventuell nur noch recht geringe Reidentifikationskosten. Zumindest liegen diese dann möglicherweise nicht mehr deutlich über den Kosten der Alternative, zumal die hierfür mehrfach genannten 150 DM keine Obergrenze darstellen sondern als Größenordnung lediglich einen Anhaltspunkt liefern können.

Allerdings ist zu beachten, daß der Angreifer wohl gerade bei dem hier betrachteten Deanonymisierungsmotiv Fehlzuordnungen möglichst vermeiden möchte. Je nach seiner Einschätzung von Inkompatibilitäten in den Überschneidungsmerkmalen zwischen den beiden Files wird er daher nach dem eigentlichen Abgleich noch weitere, kostenverursachende Nachforschungen anstellen, um die vorgenommenen Zuordnungen auf ihre Richtigkeit hin zu überprüfen.

10.3 Angriffsszenario 5: Gewinnung ökonomisch verwertbarer Informationen aus dem Mikrozensus zum Verkauf an einen Adressenhändler

10.3.1 Beschreibung des Szenarios

Ausgangslage ist bei dieser Situation ein ökonomisches, außerberufliches Interesse: Die Möglichkeit einer monetären Verwertung von Informationen aus dem Mikrozensus, speziell durch den Verkauf an Adressenhändler, wird häufig als Anstoß für Deanonymisierungsversuche genannt.

Grundsätzlich kann dabei als Strategie sowohl die Massenidentifkation als auch der Massenfischzug in Frage kommen. Wenig wahrscheinlich und auch kaum sinnvoll weiter zu erörtern wäre jedoch die Situation, daß ein Wissenschaftler mehr oder wengier wahllos Deanonymisierungsversuche nur mit dem Ziel startet, möglichst viele Personen aus dem Mikrodatenfile zu reidentifizieren, um die so gewonnenen Erkenntnisse Adressenhändlern zum Kauf anzubieten. Noch eher realistisch ist eine Zusammenarbeit in dem Sinne, daß der Adressenhändler den Wissenschaftler mit seinem Material versorgt und dieser gegen Entgelt versucht, z. B. über bestimmte Berufsgruppen Informationen aus dem Mikrozensus zu gewinnen. Der Adressenhändler könnte sich durch ein derartiges Vorgehen möglicherweise eine effizientere Durchführung eines Marketingauftrags versprechen. Aus den Unterlagen eines Adressenhändlers geht etwa hervor, daß er über ein sog. "Zielgruppen-Pool" verfügt, das die Adressen von 50 000 "Entscheidungsträgern der Wirtschaft" enthält. Möglich wäre auch, daß zum Aufbau derartiger Pools unter Vorgabe bestimmter Ausprägungen von Mikrozensus-Variablen, wie z.B. höchste Einkommensklasse, aus einem Massenfischzug zusätzlich wichtige Informationen erwartet werden.

Hier wird im weiteren unterstellt, daß ein Forscher versucht, Datensätze von Angehörigen "selbständiger und freier Berufe" zu deanonymisieren, die ein Adressenhändler dann auf dem Markt möglichen Interessenten für deren Marketingzwecke anbietet.

10.3.2 Identifikations- und Mikrodatenfile

Als Identifikationsfile dient das Informationsangebot von Adressenhändlern und als Mikrodatenfile wiederum der Mikrozensus.

Das kommerzielle Angebot von Adressen für eine Vielzahl spezifischer Zielgruppen eignet sich relativ gut für Deanonymisierungsversuche auf Massenbasis, da von den Anbietern stets Vollständigkeit angestrebt wird, was wiederum zu Informationsgewinnen über mögliche Mikrozensus-Teilnahmen führen kann. Eine Analyse des Werbematerials von Adressenhändlern zeigt, daß die Lieferung nicht nur in Form von Adressenausdrucken und Selbstklebe-Etiketten sondern auch auf Disketten oder anderen EDV-Trägern erfolgen kann. Ein Vorteil liegt ferner in der recht weitgehenden Tiefenselektion und regionalen Sortierung. Informationsquellen der Anbieter sind Telefonbücher, Handelsregistereintragungen, Messekataloge, Geschäftsberichte, Tages- und Wirtschaftszeitungen, schriftliche und telefonische Recherchen sowie weitere allgemein zugängliche Informationen. Die Aufzählung zeigt, daß es sich überwiegend um Wirtschaftsdaten handelt, wodurch sich nur sehr wenige Berührungspunkte zu personen- und haushaltsbezogenen Erhebungen wie dem Mikrozensus ergeben.

Für das hier zu analysierende Angriffsszenario können daher im Prinzip nur die Überschneidungsmerkmale

- Geschlecht
- Beruf
- Wirtschaftszweig
- Bundesraumordnungsregion
- Gemeindegrößenklasse

zugrunde gelegt werden.

Als "Vorbild" dient dabei das konkrete Angebot eines Händlers über ca. 500 000 Adressen von Angehörigen "selbständiger und freier Berufe". Zwar handelt es sich dort um Firmensitze; aber es ist anzunehmen, daß dann auch eine Liste der entsprechenden Privatanschriften erstellbar ist, zumal bei diesen häufig eher kleinen "Betrieben" beide Adressen wohl oft zusammenfallen.

Der geringen Anzahl an Überschneidungsmerkmalen steht im übrigen deren - außer bei der Geschlechtsangabe - hoher Informationsgehalt gegenüber. Von daher wurde wieder eine Konstellation unterstellt, in der noch am ehesten Deanonymisierungserfolge zu erwarten sind.

10.3.3 Abgrenzungs- und Fehlerproblematik

Aktualitäts- und Qualitätsprobleme bei über kommerzielle Anbieter bezogenem Adressenmaterial wurden bereits in den Ausführungen zum Zusatzwissen erwähnt. Händler selbst konstatieren Fehler und gewisse Zeitverzögerungen, da z.B. fünf bis zehn Prozent aller Adressen in den neuesten Telefonbüchern und anderen Publikationen beim Erscheinen bereits veraltet seien. Aus einem Untersuchungsbericht über einen bestimmten Markt im Dienstleistungsbereich etwa ist zu entnehmen, daß das die Analyse durchführende Institut bei einem Adreßverlag den umfassendsten am Markt erhältlichen Datenpool mit ca. 3 400 Adressen erstanden hatte. Eine Überprüfung des Materials anhand einer Stichprobe ergab dann, daß nur 53 Prozent der gekauften Adressen vielleicht der zu untersuchenden Branche zugerechnet werden konnten. Wenn dies auch nur ein Einzelbeispiel ist, das nicht verallgemeinert werden sollte, liefert es dennoch einen Hinweis auf die Datenbeschaffungsproblematik. Im übrigen gibt es sicher auch Abgrenzungsprobleme zwischen den beiden abzugleichenden Files bei der Wirtschaftszweig- und Berufsgliederung.

10.3.4 Praktische Durchführung

In der folgenden argumentativen Diskussion von Chancen, Kosten und Nutzen möglicher Reidentifikationen für die beschriebene Konstellation werden wiederum die Ergebnisse des Szenarios 1 berücksichtigt, aber auch Resultate der von Paaß/Wauschkuhn (1985) durchgeführten Simulationen herangezogen.

10.3.5 Kosten-Nutzen-Analyse

10.3.5.1 Deanonymisierungskosten

In der vorliegenden Angriffssituation wird sich ein Angreifer vermutlich keinen festen Wert für die Anzahl erfolgreicher Deanonymisierungsversuche zum Ziel setzen, sondern versuchen, auf der Basis des von dem Adressenhändler gelieferten Datenmaterials möglichst viele Fälle aus dem Mikrozensus zu deanonymisieren.

Aus diesem Grund fällt im Vergleich zu anderen, bereits erörterten Szenarien der Umfang des benötigten Identifikationsfile und damit die Reidentifikationsquote als Einflußgröße für die Deanonymisierungskosten in dem hier zu diskutierenden Fall nicht so sehr ins Gewicht. Dies gilt besonders dann, wenn der Adressenhändler das Zusatzwissen dem Angreifer bereits in maschinenlesbarer Form zur Verfügung stellt. Dann entfallen hier völlig die - sonst mit dem Identifikationsfile-Umfang zunehmenden und oft bedeutenden - Kosten für die Beschaffung des Zusatzwissens sowie für dessen Übernahme auf maschinenlesbare Träger.

Auch die Kosten für Fehlerabschätzungen können im Prinzip vernachlässigt werden. Auf der einen Seite besitzt der Angreifer nämlich in der Regel keine Informationen über Fehler im Mikrodatenfile. Auf der anderen Seite wird der Adressenhändler im eigenen Interesse selbst die bestmögliche Qualität seiner Daten anstreben, so daß für diesbezügliche zusätzliche Überlegungen kaum Spielraum bleibt. Im übrigen führen etwa Falschzuordnungen bei dem hier betrachteten Deanonymisierungsmotiv wohl ohnehin nur zu einem eher geringen Schaden. Auch von daher erscheinen Anstrengungen zur Berücksichtigung von Fehlern hier weniger lohnend.

Kosten entstehen aber nun durch die Anpassung der Überschneidungsmerkmale der beiden Files. Dabei ist davon auszugehen, daß dem Angreifer für jedes Überschneidungsmerkmal sowohl eine Auflistung aller möglichen Mikrodatenfile-Ausprägungen vorliegt als auch eine vom Adressenhändler mitgelieferte Zusammenstellung aller möglichen Identifikationsfile-Ausprägungen. Nach der Erstellung geeigneter Überführungsformeln kann die Anpassung dann maschinell erfolgen. Arbeitszeit fällt im wesentlichen lediglich für die Entwicklung eines solchen Algorithmus an. Aufwand verursachen dabei speziell die eventuell zahlreichen, wenn auch unproblematischen Transformationen der Regionalangabe. Das gleiche gilt in der Regel für den

Wirtschaftszweig, während beim Beruf wohl zusätzlich noch Abgrenzungsschwierigkeiten auftreten. Dieser Aufwand reduziert sich allerdings, wenn das Identifikationsfile von vorneherein auf wenige "klassische" freie Berufe, wie z. B. Arzt, Rechtsanwalt oder Architekt beschränkt ist.

Zur Einschätzung der Kosten für Erstellung, einschließlich Implementierung, und Ablauf des Reidentifikationsprogramms greife man wieder auf die Ergebnisse aus dem ersten Deanonymisierungsszenario zurück. Aus dem dortigen Identifikationsfile mußten ca. 8 000 Datensätze mit je zehn Merkmalsausprägungen, d.h. ca. 80 000 Daten verarbeitet werden. Diesmal sind dagegen, wie oben festgehalten, ungefähr 500 000 Datensätze mit je fünf Ausprägungen, d.h. etwa 2,5 Millionen zu verarbeitende Daten und damit über dreißigmal so viele unterstellt. Die dort ermittelte Arbeitszeit und Rechenkosten in Höhe von ca. 17 000 DM werden jetzt also wohl deutlich überschritten, wenngleich der Aufwand zur Programmerstellung und-implementierung vermutlich nicht proportional zur Datenmenge zunimmt. Gleichwohl ist diesmal noch die erwähnte mögliche maschinelle Anpassung der Überschneidungsmerkmale von Mikrozensus und Identifikationsfile ins Kalkül zu ziehen. Insgesamt stellt deshalb etwa ein Größenordnung von 20 000 DM an Kosten für das Programm und seinen Ablauf einen eher niedrigen Ansatz dar, der aber dennoch hier zugrunde gelegt werden soll.

Verschleierungskosten sind auch diesmal nicht zu vernachlässigen. Dies gilt zum einen wegen der äußerst großen, gegenüber den anderen Situationen deutlich höheren zu verarbeitenden Anzahl von 2,5 Millionen Daten des Identifikationsfile. Zum anderen existiert mit dem Adressenhändler ein Mitwisser an den Deanonymisierungsversuchen. Zu deren Geheimhaltung auch von dieser Seite können damit Kosten etwa in Form späterer Zahlungen an ihn anfallen. Natürlich bleibt der Wissenschaftler wiederum an der Geheimhaltung der Deanonymisierungsversuche auch nach deren Abschluß interessiert.

Rechtliche Sanktionen schließlich und Nutzentgang des Angreifers aufgrund der Beeinträchtigung seiner beruflichen Laufbahn können etwa wieder mit der Wahrscheinlichkeit für die Aufdeckung der Deanonymisierungen gewichtet werden. Diese Wahrscheinlichkeit ist wegen der zu verarbeitenden großen Datenmenge und der Mitwisserschaft des Adressenhändlers relativ hoch anzusetzen, hängt aber natürlich auch jetzt von den Verschleierungsanstrengungen ab. Ferner könnten die drohenden Sanktionen an sich wieder stärker ausfallen als etwa bei der wissenschaftlichen Verwertung der gewonnenen Daten in Szenario 3.

Zusammenfassend ist festzuhalten, daß diesmal neben rechtlichen Sanktionen, Nutzenentgang und Verschleierungsanstrengungen im wesentlichen nur noch Kosten für die Erstellung der Überführungsformeln zur Anpassung der Überschneidungsmerkmale der beiden Files und die hier mit ca. 20 000 DM angesetzten Kosten für das Deanonymisierungsprogramm und seinen Ablauf anfallen. Dagegen entfallen die sonst im allgemeinen bedeutenden Kosten zur Beschaffung des Zusatzwissens und seine Übernahme auf maschinenlesbare Datenträger.

Für Versuche des Angreifers, Deanonymisierungskosten auf sein Forschungsinstitut zu überwälzen, verbleibt dann im Prinzip nur der genannte Aufwand für Erstellung und Ablauf des Programms sowie für die genannte Erstellung der Überführungsformeln. Der Wissenschaftler wird aber wohl auch jetzt auf derartige Bemühungen eher verzichten. Bei der Erarbeitung des Reidentifikationsalgorithmus durch Mitarbeiter etwa würde das Vorhaben des Angreifers vermutlich sofort offensichtlich. Aber auch der Programmablauf auf einer institutseigenen Anlage würde bei ausreichenden technisch-organisatorischen Sicherungsmaßnahmen wegen der großen zu verarbeitenden Datenmenge zu einer für den Angreifer wohl zu hohen Wahrscheinlichkeit der Aufdeckung seiner Versuche führen. Hingewiesen sei dabei wieder auf das Interesse an deren Geheimhaltung auch nach dem Abschluß des Vorhabens.

10.3.5.2 Alternative Datenbeschaffungsmethoden

Der Ausgangspunkt für die Diskussion alternativer Datenbeschaffungsmethoden unterscheidet sich hier in formaler Weise gegenüber den bisher betrachteten Szenarien. Es ist davon auszugehen, daß ein Adressenhändler als professioneller Datenbeschaffer vor der Erteilung seines Auftrags an den Angreifer im Prinzip schon selbst alle alternativen Möglichkeiten geprüft und eine Deanonymisierung als kostengünstigste Methode befunden hat. Gäbe es also für den Angreifer eine bessere Alternative, wäre diese wohl unter Umgehung dieses "Zwischenhandels" bereits vom Auftraggeber selbst wahrgenommen worden. Im Gegensatz zu den bisherigen Szenarien werden verschiedene Beschaffungsmethoden hier somit schon von einer anderen Person miteinander verglichen. Inhaltlich erscheint es zur Beurteilung der faktischen Anonymität jedoch unerheblich, ob der Angreifer selbst oder eine andere Person diesen Vergleich anstellt.

Die Abwägung von Datenbeschaffungsalternativen gegen Deanonymisierungen erfordert zunächst wieder eine Einschätzung der Anzahl erfolgreicher Reidentifikationsversuche. Diese Anzahl wiederum wird bei gegebenem Umfang des Identifikationsfile durch die auf diesen bezogene Reidentifikationsquote bestimmt. Da für das jetzige Szenario keine Simulationsrechnungen durchgeführt werden, sollen zur groben Abschätzung dieser Quote ein weiteres Mal Ergebnisse des Gelehrtenkalender-Szenarios, aber auch aus dem GMD-Bericht (Paaß/Wauschkuhn 1985) als Ausgangspunkt dienen. Im dortigen Konzern-Szenario werden zusätzlich zu den in 10.3.2 aufgeführten acht weitere Überschneidungsmerkmale verwendet, darunter neben der Wochenarbeitszeit die informativen Auskünfte über Alter und Tätigkeitsmerkmale. Dagegen entfallen die Regionalangaben einschließlich der Gemeindegrößenklasse (Paaß/Wauschkuhn 1985:206).

Den fünf Überschneidungsmerkmalen in der hier zu diskutierenden Situation stehen also elf im Rahmen des GMD-Experiments gegenüber. Dabei kann der Informationsgehalt der Regionalangabe Bundesraumordnungsregion einschließlich Gemeindegrößenklasse denjenigen von Alter und Tätigkeitsmerkmal überschreiten - allerdings wohl nicht in einem Ausmaß, welches die Überzahl der GMD-Überschneidungsmerkmale ausgleicht. Da nun aber bereits die GMD-Simulation für den Normalfall fehlender Information über die Mikrozensus-Teilnahme keine einzige Deanonymisierung erbrachte (Paaß/Wauschkuhn 1985:207), ist damit erst recht in dem jetzigen Szenario mit einer Reidentifikationsquote 0 zu rechnen. Dafür spricht zusätzlich, daß die Quote in Höhe von ca. 0,0005 im Gelehrtenszenario unter derjenigen für ähnliche Konstellationen im GMD-Bericht lag.

Andererseits könnte der Adressenhändler theoretisch über eine vollständige Kartei der interessierenden Gruppe, hier also der selbständigen und freien Berufe, verfügen. Stellt er dem Angreifer diesen Gesamtbestand als Zusatzwissen zur Verfügung, so wüßte dieser, daß jedes Gruppenmitglied im Mikrodatenfile auch im Identifikationsfile enthalten ist. Damit würde, wie schon früher mehrfach festgehalten, die Reidentifikationsquote erheblich zunehmen, sofern die beiden Datenbestände vollständig kompatibel sind. In der Praxis ist die genannte Vollständigkeit allerdings kaum zu irgendeinem Zeitpunkt erreichbar, da Bestandsänderungen in der Regel erst mit einer gewissen Verzögerung vom Adressenhändler registriert und verarbeitet werden können. Verwiesen sei dazu beispielhaft auf die schon zu Anfang erwähnte Liste eines Adressenhändlers über Unternehmen eines bestimmten Wirtschaftsbereiches,

von denen gemäß einer Überprüfung aktuell jedoch nur ca. 50 Prozent wirklich in dieser Gruppierung existierten.

Insgesamt stellen Deanonymisierungen für das hier unterstellte Motiv somit schon an sich kein taugliches Mittel zur Datenbeschaffung dar. Zu analytischen Zwecken soll im folgenden aber angenommen werden, daß ein Prozent der in beiden Files gleichzeitig enthaltenen Datensätze deanonymisiert werden können. Dann ergibt sich unter Berücksichtigung des Mikrozensauswahlsatzes von ebenfalls ein Prozent die auf das Identifikationsfile bezogene Reidentifikationsquote 0,0001. Unter Zugrundelegung des Identifikationsfile mit ca. 500 000 Datensätzen Angehöriger selbständiger und freier Berufsgruppen wären damit ca. 500 000 x 0,0001 = 50 Deanonymisierungen zu erwarten.

Eine Beschaffungsalternative zur Deanonymisierung sollte also ungefähr genauso viele Datensätze liefern. Deren Länge braucht aber nicht derjenigen im Mikrozensus zu entsprechen, da die Kunden des Adressenhändlers zur beabsichtigten Verbesserung ihres Marketing bei weitem nicht alle dort enthaltenen Merkmale sinnvoll einsetzen können. Vielmehr erscheint etwa zur Vermeidung von Streuverlusten nur die Konzentration auf wenige zusätzliche Informationen lohnend. Zu denken ist dabei beispielsweise an Mikrozensus-Angaben zu Einkommen, Alter, Haushaltszusammensetzung oder Versicherung.

Betrachtet man jetzt als Alternative gegenüber der Deanonymisierung eine eigene Umfrage etwa des Adressenhändlers zur Beschaffung von rund 50 solcher verkürzten Datensätze und geht wieder von der für sozialwissenschaftliche Erhebungen realistischen Größenordnung von 150 DM an Kosten pro Interview aus, so müßte er dafür ca. 50 x 150 DM = 7 500 DM aufwenden. Dem stehen bei der Deanonymisierung zunächst die im vorigen Abschnitt eher niedrig angesetzten 20 000 DM allein an Kosten für Erstellung und Ablauf des dortigen Reidentifikationsprogramms gegenüber. Selbst wenn man dann zusätzlich nur noch die Verschleierungskosten einbezieht, ergäbe sich damit wohl ein deutlicher Vorteil für die eigene Umfrage. Dies gilt natürlich erst recht bei der Berücksichtigung drohender Sanktionen und des möglicherweise entgehenden Nutzens durch die Beeinträchtigung der beruflichen Laufbahn im Fall der Aufdeckung der Reidentifikationsversuche.

10.3.5.3 Deanonymisierungsnutzen

Obwohl im vorigen Abschnitt bereits festgestellt wurde, daß Deanonymisierungen bei dem hier unterstellten Motiv wohl kein taugliches Mittel zur Datenbeschaffung darstellen, so daß eine Nutzendiskussion von daher überflüssig wäre, sei im folgenden, wie schon dort, dennoch angenommen, daß immerhin etwa 50 von den 500 000 Identifikationsfile-Personen identifiziert werden können. Die Erörterung des Deanonymisierungsnutzens erscheint bei der jetzt betrachteten, in der öffentlichen Diskussion relativ häufig angeführten speziellen Datenverwertungsmöglichkeit nämlich gleichwohl von Interesse.

Aus der Deanonymisierung erwächst dem Angreifer Nutzen in Form des Verkaufserlöses für die an den Adressenhändler gelieferten reidentifizierten Datensätze. Seine Höhe hängt ab von deren Anzahl und dem vom Adressenhändler gezahlten Preis für einen einzelnen Datensatz.

Eine obere Schranke findet dieser Preis sicher in dem Erlös, der dem Adressenhändler aus der Vermarktung eines solchen Datensatzes zufließt. Dabei ist zu beachten, daß er diesen natürlich nicht nur einmal, sondern mehrfach am Markt anbieten kann. Außerdem ist den Kunden oft nur die einmalige Verwendung der zur Verfügung gestellten Adressen gestattet bzw. ein jeweils weiterer Einsatz nur gegen entsprechende Mehrfachzahlung. Im übrigen werden etwa Werbeversandaktionen häufig im Kundenauftrag von den Adresenhändlern selbst durchgeführt, so daß von daher die pro Zahlung nur einmalige Adressenverwendung gesichert ist. Der Erlös des Adressenhändlers ergibt sich also aus der Anzahl der Einsätze des Datensatzes und dem Preis für jeden solchen Einsatz.

Zunächst aber stellt sich generell die Frage, ob ein Adressenhändler überhaupt an zusätzlichen Mikrozensus-Merkmalen interessiert ist, um sie für Marketingzwecke seiner Kunden anzubieten. Im vorigen Abschnitt wurde bereits darauf hingewiesen, daß einige Mikrozensus-Merkmale sich durchaus z.B. zur Verringerung von Streuverlusten eignen könnten. In Frage kommen etwa Angaben zu Einkommen, Alter, Versicherung oder Haushaltszusammensetzung. Der Einsatz allzu vieler solcher Merkmale für derartige Zwecke ist aber wohl als unverhältnismäßig aufwendig einzuschätzen.

Um nun eine Vorstellung über die Größenordnung des möglichen Preises zu gewinnen, den der Adressenhändler für den einmaligen Einsatz eines Datensatzes erzielen kann, ist sicherlich keine umfangreiche Marktanalyse notwendig. Vielmehr soll exemplarisch die Preisliste des Händlers, dessen

Adressenkartei selbständiger und freier Berufe hier als Identifikationsfile dient, den Ausgangspunkt bilden. Er bietet Adressen "gängiger" und damit auch im Mikrodatenfile relativ häufig vorkommender Berufsgruppen je nach Bestellmenge zu Preisen um 100 DM für 1 000 Stück bei einmaliger Verwendung an, d.h. für ca. zehn Pf pro Stück und Einsatz. Diese Adressen stellen Datensätze dar, die außer den Identifikationsmerkmalen implizit natürlich Beruf, Wirtschaftszweig und Geschlecht ausweisen.

Geht man nun beispielsweise davon aus, daß der Adressenhändler einen Datensatz, der genug interessante zusätzliche Mikrozensus-Merkmale enthält, zum doppelten Preis, also für 20 Pf statt 10 Pf verkaufen kann und wird etwa ein hundertmaliger Verkauf unterstellt, so beträgt der Erlös aus dieser Vermarktung 100 x 0,2 DM = 20 DM. Dies wäre dann die erwähnte obere Schranke des Preises, der dem Angreifer für einen deanonymisierten Datensatz zu bieten ist. Andererseits ist den Kunden des Adressenhändlers und damit ihm selbst bei den hier relevanten gängigen Berufsgruppen für Marketingzwecke mit einer geringen Anzahl von Adressen, wie den unterstellten ca. 50 Stück, wohl nur wenig gedient.

Aber nicht nur deshalb wird der dem Angreifer gezahlte Preis vermutlich deutlich unter den zugrundegelegten 20 DM liegen. Zum einen muß der Adressenhändler natürlich außer dieser Vorleistung des Angreifers noch andere Kosten kalkulieren, die etwa aus Verwaltung, Pflege und Vertrieb der Daten resultieren. Zum anderen wird er in der Regel außerdem einen Gewinn anstreben.

Unterstellt man beispielsweise 16 DM als Preis für einen vom Angreifer gelieferten deanonymisierten Datensatz, so erbringen ihm die 50 Reidentifikationen schließlich Erlöse von 800 DM.

Folgende Gleichung und Abbildung 10.1 sollen das Zustandekommen und die Elemente des Angreifererlöses E in dem zugrundegelegten Modell noch einmal allgemein verdeutlichen. Es ist

(*) $E = p_1 d = (p_2 - g) z d = (p_2 - g) z q R,$

wobei

p_1 = Vom Adressenhändler an den Angreifer gezahlter Preis für einen deanonymisierten Datensatz,

d = Anzahl der deanonymisierten Datensätze,

z = Durchschnittliche Anzahl der Einsätze eines solchen Datensatzes,

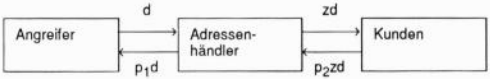
p_2 = Von den Kunden des Adressenhändlers zu zahlender Preis für einen solchen Einsatz,

g = Kosten, die dem Adressenhändler pro Einsatz eines deanonymisierten Datensatzes außer durch dessen Ankauf noch entstehen, zuzüglich eines Gewinnzuschlags,

R = Umfang des Identifikationsfile,

q = d/R = Auf diesen Umfang bezogene Reidentifikationsquote.

Abbildung 10.1: Gelieferte Datensätze (⟶) und resultierende Erlöse (⟵)



In dem oben angenommenen Beispiel betragen $p_1 = 16$ DM, $d = 50$, $z = 100$, $p_2 = 0{,}20$ DM, $g = 0{,}04$ DM, $R = 500\,000$, d.h. $q = 0{,}0001$, und somit $E = 800$ DM.

Kurz angesprochen werden soll noch die Möglichkeit des Angreifers, nicht nur einen, sondern mehrere Adressenhändler zu beliefern. Sie erscheint aber eher unwahrscheinlich, da sich dadurch die Zahl der Mitwissenden erhöht und damit wiederum die Aufdeckungsgefahr zunimmt. Außerdem würde die fehlende Exklusivität den an den Angreifer gezahlten Preis wohl deutlich vermindern. Ferner ist denkbar, daß der Adressenhändler von vorneherein seine ausschließliche Belieferung zur Bedingung für das Geschäft macht und für den anderen Fall mit Konsequenzen droht.

Zum Schluß sei noch einmal betont, daß diese exemplarische Kalkulation nur grobe Anhaltspunkte für den Nutzen des Angreifers liefern kann. Andere Annahmen über die betrachteten Einflußgrößen 'Umfang des Identifikationsfile',

'Reidentifikationsquote', 'Anzahl der Einsätze eines deanonymisierten Datensatzes', 'Preis für einen solchen Einsatz an den Adressenhändler' und 'Preis für einen reidentifizierten Datensatz an den Angreifer' führen natürlich zu einem entsprechend veränderten Erlös.

Dabei ist jedoch noch einmal auf die wohl unrealistisch hohe, zur Ermöglichung der obigen Nutzendiskussion aber dennoch unterstellte Anzahl von 50 reidentifizierten Datensätzen hinzuweisen. Die im vorigen Abschnitt begründete prinzipielle Untauglichkeit von Deanonymisierungen als Datenbeschaffungsmethode beim hier betrachteten Reidentifikationsmotiv ist dementsprechend am ehesten als Nutzen in Höhe von Null zu interpretieren.

10.3.5.4 Fazit

Die genannte Untauglichkeit von Deanonymisierungen für die hier unterstellte Datenverwendung bildet auch einen wesentlichen Bestandteil des zu ziehenden Fazits: Wegen der wenigen Überschneidungsmerkmale sind praktisch gar keine Erfolge zu erwarten. Selbst bei ausnahmsweiser Kenntnis des Angreifers über einige, wenige Mikrozensus-Teilnahmen bliebe die Anzahl der Reidentifikationen wohl deutlich unter derjenigen, die für einen massenweisen Adressenverkauf zu Marketingzwecken notwendig wäre. Damit kann der Mikrozensus für die jetzige Situation als faktisch anonym bewertet werden.

Aber selbst bei Unterstellung einer Quote von immerhin etwa ein Prozent deanonymiserter unter den in beiden Files gleichzeitig enthaltenen Datensätzen ergibt sich wohl ein deutlicher Kostenvorteil für eine eigene Umfrage als alternative Datenbeschaffungsmethode. Allein die Kosten für Erstellung und Ablauf des Reidentifikationsalgorithmus erreichten in 10.3.5.1 eine Größenordnung von ca. 20 000 DM. Dagegen wurden in 10.3.5.2 für eine entsprechende Anzahl von Interviews nur etwa 7 500 DM, also deutlich weniger begründet. Dieser Vorteil erhöht sich natürlich noch, wie dort erwähnt, bei Einbeziehung von Verschleierungskosten, möglichen Sanktionen und eventuellem Nutzentgang durch die drohende Beeinträchtigung der beruflichen Laufbahn des Angreifers.

Die Gegenüberstellung von Deanonymisierungskosten und -nutzen schließlich ergibt bei Zugrundelegung der Größenordnungen aus 10.3.5.1 und dem vorigen Abschnitt einen deutlichen Überhang der Kosten: Wiederum bei Beschränkung auf den Aufwand für Erstellung und Ablauf des Reidentifikationsalgorithmus in Höhe von ca. 20 000 DM fällt der den Nutzen repräsentierende Erlös mit etwa

800 DM um den Faktor 25 niedriger aus. Auch dieser Unterschied vergrößert sich natürlich bei Einbeziehung von Verschleierungskosten, drohenden Sanktionen und möglichem Nutzenentgang durch Beeinträchtigung der beruflichen Laufbahn des Angreifers.

11 Anonymisierungsmaßnahmen*

Für alle im vorausgegangenen Kapitel untersuchten Szenarien war das Kriterium des unverhältnismäßig hohen Aufwands mit Sicherheit erfüllt und damit die faktische Anonymität der Daten gegeben. Die empirische Überprüfung von Subszenario 1 und 2 auf der Basis von empirischen Daten konnte darüber hinaus aufzeigen, daß unter allen untersuchten Randbedingungen die Wahrscheinlichkeit einer eindeutigen und zugleich korrekten Zuordnung mit Werten zwischen 0 und 0,0005 empirisch wesentlich niedriger anzusetzen ist, als dies nach den - auf der Grundlage von synthetisch generiertem Zusatzwissen - ermittelten Befunden der GMD der Fall war. Sie haben auch gezeigt, daß aus einer hohen Quote von Fällen mit einzigartigen Ausprägungskombinationen noch keineswegs auf ein hohes Reidentifikationsrisiko geschlossen werden kann.

Eine wesentliche Erkenntnis war hierbei, daß das Ausmaß der von *Inkompatibilitäten* ausgehenden Schutzwirkung vor erfolgreichen Reidentifikationsversuchen bislang erheblich unterschätzt wurde (vgl. u.a. Bethlehem et al. 1990, Dalenius 1986, Brunnstein 1987). Während aufgrund der theoretischen Vorüberlegungen zu den Zuordnungsmechanismen einfacher Abgleichtechniken bereits deutlich wurde, daß diese Algorithmen beim Auftreten von Inkompatibilitäten mit "Nicht-" und/oder "Falschzuordnungen" reagieren, überraschen die Ergebnisse bei Verwendung der diskriminanzanalytischen Reidentifikationstechnik. Obwohl dieser Algorithmus sowohl die Möglichkeit von statistischen Doppelgängern in der Grundgesamtheit wie auch das Auftreten von Dateninkompatibilitäten berücksichtigt, steigt die Anzahl von korrekten Zuordnungen im Vergleich zu einfachen Abgleichtechniken nicht an. Ebenso wenig war es mit dieser Reidentifikationstechnik möglich, anhand der Zuordnungswahrscheinlichkeiten auch nur annähernd zwischen falschen und korrekten Zuordnungen zu trennen.

Bei der theoretischen Erörterung der verschiedenen Komponenten des Reidentifikationsrisikos hatte sich in Kapitel 4 die Schlußfolgerung ergeben, daß beim Mikrozensus und der EVS mit erhöhten Risiken gerechnet werden muß, wenn unter der Voraussetzung einer hohen Kompatiblität und eines hohen Auflösungsgrades der Überschneidungsmerkmale entweder

---

* Autoren: Uwe Blien, Heike Wirth

- bei einem Fischzugsszenario ein zugängliches Identifikationsfile existiert, oder aufgebaut werden kann, in dem die Angehörigen der Gesamtbevölkerung oder eines Teils der Bevölkerung, der durch spezifische - im Mikrodatenfile enhaltene - Merkmalsausprägungen abgrenzbar ist, vollständig oder weitgehend vollständig enthalten sind, oder

- ein Angreifer Kenntnis darüber hat, daß eine bestimmte Person im Mikrodatenfile enthalten ist (response knowledge).

Der Fall des Fischzuges wurde mit Szenario 1 überprüft, bei dem der Gelehrtenkalender eine sehr spezifische Teilpopulation weitgehend vollständig erzielt und bei dem durch regionale und sachliche Informationen ein hochauflösliches Überschneidungswissen zur Verfügung stand. Bei dieser riskanten Situation war die faktische Anonymität zweifelsfrei gegeben.

Bei der Unterstellung von Teilnahmekenntnis ergaben sich dagegen Anhaltspunkte für eine *Risikokonstellation*, bei denen unter Umständen in Einzelfällen eine erfolgreiche Reidentifikation mit vergleichsweise niedrigem Aufwand möglich erscheint. Dieser - allerdings äußerst seltene Fall, der in Angriffsszenario 4 in Abschnitt 10.2.5 diskutiert wurde, setzt das Zusammentreffen sehr spezifischer Risikofaktoren voraus und kann allgemein wie folgt charakterisiert werden:[2]

- Eine im Mikrodatenfile gesuchte Person gehört einer sehr kleinen, durch ein spezifisches Merkmal eingrenzbaren *Subpopulation* an, z.B. einer bestimmten Berufsgruppe oder Nationalität (*sachliche Tiefengliederung*);

- das Mikrodatenfile enthält *tiefgegliederte Regionalinformationen*, so daß in den jeweiligen Regionaleinheiten nur wenige Angehörige dieser spezifischen Subpopulation leben (*regionale Tiefengliederung*);

- ein Forscher, der Zugang zu den Einzelangaben des Mikrodatenfile hat, kann sich Kenntnisse über einen Angehörigen dieser spezifischen Subpopulation beschaffen und weiß, daß diese Person an der Mikrodatenerhebung, über deren Daten er verfügt, teilgenommen hat (*Teilnahmekenntnis*).

---

2 Damit es zu einem Datenangriff kommt, muß außerdem vorausgesetzt werden, daß ein Datenangreifer ein subjektives Interesse daran hat, das die denkbaren Kosten der Konsequenzen des Angriffs (Reputationsverlust, Vertragsstrafen, gesetzliche Strafen) übersteigt.

- die Merkmale der Person sind genau in der Weise im Mikrodatenfile erfaßt, wie es der Forscher vermutet (*Kompatibilität*).

Beim Zusammentreffen dieser spezifischen Bedingungen erscheint die Möglichkeit der Reidentifikation eines *Einzelfalls* ohne großen Aufwand als gegeben. Es ist wichtig darauf hinzuweisen, daß alle vier Bedingungen *gleichzeitig* erfüllt sein müssen. Bereits wenn eine der Bedingungen nicht gegeben ist, kann eine sichere Reidentifikation ohne den Aufwand unverhältnismäßig hoher Kosten nach den durchgeführten Experimenten als äußerst gering betrachtet werden. Das gleichzeitige Zusammentreffen aller Bedingungen kann bei Stichprobenerhebungen als außergewöhnlich seltenes Ereignis betrachtet werden. Dennoch sollten bei der Datenübermittlung Vorkehrungen getroffen werden, daß auch eine solche (unwahrscheinliche) Risikokonstellation ausgeschlossen ist.

Im folgenden soll daher untersucht werden, mit welchen datenorientierten Schutzmaßnahmen die oben charakterisierte Risikokonstellation entschärft und so das Risiko einer Deanonymisierung weiter verringert werden kann.

Hierfür werden in einem ersten Abschnitt zunächst allgemeine Kriterien für eine zu treffende Auswahl von Anonymisierungsmaßnahmen dargestellt. In einem weiteren Abschnitt werden ausgewählte Anonymisierungsmaßnahmen in ihrer Schutzwirkung und der damit einhergehenden Auswirkung auf das Analysepotential der Daten diskutiert. In einem dritten Abschnitt schließlich soll die von einer Substichprobenziehung und Ausprägungsvergröberung ausgehende Schutzwirkung exemplarisch, anhand des zur Verfügung stehenden Datenmaterials, überprüft werden.

11.1 Kriterien für die Auswahl von Anonymisierungsmaßnahmen

Eine Darstellung sämtlicher in der Literatur diskutierten Anonymisierungsmaßnahmen ist nicht möglich. Daher muß eine Vorauswahl getroffen werden, die sich im wesentlichen an folgenden Kriterien orientiert:

- Es müssen insbesondere einfache Abgleichtechniken durch die gewählten Maßnahmen gestört werden.

Legt man den "unverhältnismäßig hohen Aufwand" für die Herstellung eines Personenbezugs als Auswahlkriterium für die Anonymisierungsmaßnahmen zugrunde, so müssen mit diesen Maßnahmen insbesondere einfache Abgleichtechniken gestört werden. Bislang sind keine Reidentifikationstechniken bekannt, die eine sichere Reidentifikation ermöglichen. Zugleich hat die praktische Anwendung der diskriminanzanalytischen Reidentifikationstechnik gezeigt, daß bereits die Anpassung eines differenzierten statistischen Reidentifikationsalgorithmus an eine spezifische Datenbasis mit einem extrem hohen Zeit-, Kosten- und Arbeitsaufwand verbunden ist. Unabhängig von den Ergebnissen war das Kriterium der faktischen Anonymität hierbei schon erfüllt, bevor überhaupt ein Reidentifikationsversuch stattfinden konnte. Da die Entwicklung der diskriminanzanalytischen Reidentifikationstechnik ebenfalls mit einer mehrjährigen Forschungsarbeit verbunden war, ist zur Zeit nicht davon auszugehen, daß die Entwicklung neuer, ähnlich komplexer Techniken einen geringeren Arbeitsaufwand erfordern würde.

Einfache Abgleichtechniken hingegen sind technisch problemlos umzusetzen. Ihre Anwendung liegt unterhalb der Schwelle, die durch das Kriterium der Unverhältnismäßigkeit für die faktische Anonymität angezeigt wird. Daher sollen Anonymisierungsmaßnahmen im folgenden vor allem daran gemessen werden, ob sie einfache Abgleichtechniken stören.

- Der statistische Gehalt der Daten sollte durch die gewählten Maßnahmen möglichst wenig beeinträchtigt werden.

Die in den Sozialwissenschaften untersuchten Fragestellungen und theoretischen Modelle werden immer komplexer. Hierbei sind Analysen auf der Basis von aggregierten Daten für viele Zwecke zu wenig differenziert, ungenau und damit unbrauchbar (Brennecke/Schneider 1977). Um der vielfältigen gesellschaftlichen Realität Rechnung zu tragen, in der sich Ursache-Wirkungs-Zusammenhänge in immer neuen Konstellationen überlagern und kombinieren, sind multivariate statistische Methoden entwickelt worden, die es ermöglichen, die relative Bedeutsamkeit der unterschiedlichen Kausaleinflüsse zu erkennen. Derartige statistische Verfahren sind auf differenziertes und möglichst wenig verzerrtes Datenmaterial angewiesen, auf Material also, das sehr viel Informationen enthält.[3]

---

[3] Der Informationsbedarf der Wissenschaft ist wegen der grundsätzlichen Offenheit von Forschungsprozessen theoretisch unbegrenzt. Damit sind Interessen und Zielkonflikte zwischen Forschung und Datenschutz in der Materie selbst angelegt.

Die sich hieraus ergebenden Ansprüche der Wissenschaft an anonymisiertes Datenmaterial wurden - wie in Kapitel 1 ausgeführt - bereits vielfach formuliert, so daß hier nur zwei wesentliche Punkte nochmals aufgegriffen werden sollen:

- Der Verlust an Individualinformationen sollte möglichst gering gehalten werden, da viele Analyseverfahren in den Sozialwissenschaften die Verwendung von Individualdaten erfordern.

- Die Zusammenhänge zwischen den Variablen sollten möglichst unverändert erhalten bleiben. Die Konsequenz von Veränderungen sind in vielen Fällen unkontrollierbar und enthalten letztlich die Gefahr von Fehlurteilen. Einem wissenschaftlichen Nutzen laufen Anonymisierungsmaßnahmen, die Zusammenhänge zwischen Variablen verzerren, entgegen.

Die Schutzwirkung von Maßnahmen zur Datenanonymisierung beruht letztlich zwar immer auf einer Reduktion des statistischen Informationsgehalts der Daten und geht nahezu immer mit einem Verlust an Analysepotential einher. Allerdings gibt es eine ganze Skala von Möglichkeiten im Hinblick auf die Weise, in der dies geschieht. Man kann hier unterscheiden zwischen:

- Methoden, deren Schutzwirkung auf dem bewußten Einbringen von falschen Angaben in das Datenmaterial beruht, z.B. durch Überlagerung der Daten mit Zufallszahlen. Hierunter fallen auch Maßnahmen, die auf die Bildung von synthetischen Datensätzen hinauslaufen.

- Maßnahmen, bei welchen die Daten verändert, aber nicht verfälscht werden. Hierzu würde beispielweise die Vergröberung von Merkmalsausprägungen durch die Bildung von Ausprägungsklassen zählen.

- Maßnahmen, die ohne eine direkte Modifikation der Daten auskommen. Beispiele wären etwa die Weitergabe von "veralteten" Daten oder die Ziehung von zufälligen Substichproben.

(Krupp/Preißl 1989:124, vgl. auch Müller/Hauser 1987 sowie Blien/Papastefanou 1988).

11.2 Darstellung ausgewählter Anonymisierungsmaßnahmen

11.2.1 Maßnahmen, die das Einbringen von falschen Angaben in das Datenmaterial vorsehen

Eine Möglichkeit, der oben charakterisierten Risikokonstellation zu begegnen, würde darin bestehen, die Kompatibilität zwischen Mikrodatenfile und potentiellem Zusatzwissen bewußt zu reduzieren, indem das originale Datenmaterial des Mikrodatenfile mit "falschen" Angaben überlagert wird.[4]

Eine vergleichsweise einfache Variante dieser auch als "perturbation" bzw. "contamination" bezeichneten Anonymisierungstechniken (vgl. u.a. Skinner et al. 1990, Marsh et al. 1991) sieht die Einführung von Zufallsfehlern (Zufallsrauschen) in das Datenmaterial vor. In Anlehnung an Paaß/Wauschkuhn (1985:55f) wäre hierbei zu nennen:

- Die Addition von oder Multiplikation mit normalverteilten Zufallszahlen zu kontinuierlichen Merkmalen;
- das zufällige Vertauschen von benachbarten Merkmalsausprägungen bei kontinuierlichen Merkmalen;
- das zufällige Runden von Merkmalsausprägungen von kontinuierlichen Merkmalen;
- die zufällige Veränderung von Merkmalsausprägungen in einem Teil der Datensätze bei diskreten Merkmalen.

Weitergehende Ansätze (vgl. Abschnitt 2.1) sehen die Erzeugung künstlicher Datensätze aus den Originaldaten vor. So wird beispielsweise bei der Mikroaggregation aus mehreren Einzeldatensätzen eine Art Durchschnittsdatensatz gebildet (vgl. u.a. Cox et al. 1985, Wolf 1988). Bei der statistischen Verknüpfung von Datensätzen (Statistical Record Linkage) schließlich wird aus Ausprägungssegmenten unterschiedlicher Originaldatensätze ein neuer synthetischer Datensatz gebildet, dessen statistische Eigenschaften dem Original möglichst nahe kommen sollen (vgl. Paaß/Wauschkuhn 1985:246ff. und Paaß 1988b).

---

4 Eine aktuelle deutschsprachige Übersicht über den Forschungsstand derartiger Anonymisierungstechniken geben Paaß/Wauschkuhn (1985).

Es erübrigt sich, die unterschiedlichen Varianten von Fehlerüberlagerungen detailliert zu erörtern, da es evident ist, daß von derartigen Techniken eine mindestens gleich hohe Schutzwirkung ausgeht wie von "natürlichen" Dateninkompatibilitäten. Einfache Abgleichtechniken werden durch diese Techniken daher nachhaltig gestört. Eine Reidentifikation auf der Basis von identischen Ausprägungskombinationen ist bei derartig modifizierten Daten selbst dann nicht mehr möglich, wenn eine Person eine einzigartige Ausprägungskombination aufweist und einem Angreifer bekannt ist, daß diese Person an der amtlichen Erhebung teilgenommen hat.

Allerdings sind diese Maßnahmen mit einer erheblichen Beeinträchtigung in Hinsicht auf die Verwendungsmöglichkeit der so anonymisierten Daten verbunden. Bereits bei der Überlagerung mit Zufallsfehlern - als einfachster "Perturbations"-Technik - zeigt ein Vergleich zwischen Originaldaten und anonymisierten Daten zum Teil beträchtliche Abweichungen in Hinblick auf Mittelwerte, Standardabweichungen, Korrelations- und Regressionskoeffizienten.[5] Ähnliche Verzerrungen werden auch für andere Formen von "Perturbation" erwähnt (Paaß/Wauschkuhn 1985, Greenberg 1990). So führt die Mikroaggregation zu erheblichen Verzerrungen der Zusammenhänge zwischen den Variablen. Nach den von Paaß/Wauschkuhn durchgeführten Untersuchungen, in welchen die Ergebnisse von statistischen Analysen der Originaldaten mit solchen von anonymisierten Daten verglichen wurden, zeigten sich bei letzteren für deskriptive Kennzahlen erhebliche Verfälschungen, welche die Verwendbarkeit der Daten drastisch reduzieren. Insgesamt, so wurde festgestellt, wiesen sowohl die diskreten als auch die kontinuierlichen Merkmale im Vergleich zu den Originaldaten derart starke Verzerrungen auf, daß sie z.B. in multivariaten Modellen zu signifikant verschiedenen Ergebnissen führten.[6] Ein derart anonymisiertes Datenfile kann damit zwar noch für die Erstellung von Tabellen verwendet werden - sofern die Zellen der Tabelle durch die

---

5 Vgl. hierzu die Ergebnisse von Spruill (1983).

6 Diese Probleme sind nicht bzw. nur bei sehr kleiner Zellbesetzung gegeben, wenn sich die Zufallsüberlagerung nicht auf einzelne Merkmale, sondern auf die Fallzahlen einer Tabelle bezieht, wie dies beispielsweise im Statistischen Informationssystem des Bundes (STATIS-BUND) gehandhabt wird (Kühn/Pfrommer/Schrey 1984, für multivariate Analysen: Heer/Schimpl-Neimanns 1991). Die Zufallsüberlagerung der Fallzahlen hat allerdings den Nachteil, daß Zellen unter Umständen negativ besetzt sind. Weiterhin kann es durch eine Zufallsüberlagerung der Fallzahlen dann zu Verzerrungen kommen, wenn z.B. kleine spezifische Gruppen aus dem Mikrodatenfile für Analysen betrachtet werden sollen.

Gruppierungsmerkmale festgelegt sind -, für komplexere Analysen ist es jedoch unbrauchbar.[7]

In der letzten Zeit wurden Anstrengungen unternommen, diese Techniken zu verbessern (Kim 1990): Sofern ein Forscher genau spezifizieren kann, welche Merkmale mit welchen Ausprägungen er mit welchen Methoden bearbeiten möchte, ist eine präzise Abstimmung der Zufallsüberlagerungen möglich, so daß keine gravierende Abweichungen zwischen den Analysen von Original- und anonymisierten Daten auftreten. Allerdings sind diese anonymisierten Daten dann nur für diese spezifische Fragestellung verwendbar. Für jedes andere Forschungsproblem sind sie wertlos, eine multiple Verwendbarkeit dieser Daten ist nicht gegeben (Greenberg 1990, Marsh et al. 1991).

Ganz abgesehen davon, daß ein Forschungvorhaben einem Entwicklungsprozeß gleichkommt, dessen spezifische Charakteristika nur in Ausnahmefällen bereits im Vorfeld in allen Details feststehen, wäre die Anwendung dieser Technik auch mit einem extrem hohen Arbeitsaufwand für die anonymisierende Stelle verbunden. Denn für jedes Forschungsproblem müßte ein spezifisches Datenfile mit jeweils spezifischen Zufallsüberlagerungen erzeugt werden, was letztendlich mit enormen Kosten für den jeweiligen Nutzer verbunden wäre.

Aus den genannten Gründen und da die Reidentifikationssimulationen gezeigt haben, daß die Wahrscheinlichkeit, daß ein gesuchter Fall im Mikrodaten- und Identifikationsfile kompatibel abgebildet ist, unter empirischen Bedingungen bereits sehr gering ist, stellen derartige Techniken ein viel zu "schweres Geschütz" dar, das bei der Übermittlung von faktisch anonymen Mikrodaten nicht zur Anwendung kommen sollte.

---

7 Auf das spezifische Problem von fehlerüberlagertem Datenmaterial für Längsschnittanalysen weisen Krupp/Preißl (1989) hin. Bei Beobachtung von Individuen über mehrere Untersuchungszeitpunkte können Mikroverläufe dann nicht mehr beschrieben werden, wenn durch die Fehlerüberlagerung nicht mehr erkennbar ist, ob eine ermittelte Veränderung real ist oder auf die Anonymisierungsmaßnahmen zurückgeführt werden muß.

11.2.2 Maßnahmen mit Datenmodifikation ohne Einbringen von falschen Angaben in das Datenmaterial: Vergröberung von Merkmalen

Ein zweiter wichtiger Risikofaktor der oben charakterisierten Bedingungskonstellation war die gleichzeitige sachliche und regionale Tiefengliederung des Mikrodatenfile. Unter dieser Bedingung ist es - wie bei der argumentativen Analyse von Angriffsszenario 4 ausführlich erläutert - nicht nur möglich, kleine spezifische Subpopulationen bzw. Einzelfälle auszugrenzen, sondern durch die starke regionale Einschränkung auch die Beschaffung von Zusatzwissen zu erleichtern (Südfeld 1987, Greenberg 1990). So ist es weniger aufwendig, in einer Kleinstadt eine 40jährige unverheiratete Apothekerin zu ermitteln, als dies in einer Regionaleinheit mit mehr als 500 000 Einwohnern möglich ist, d.h. eine Person, die bezogen auf eine Gemeinde mit wenigen tausend Einwohnern eine einzigartige Ausprägungskombination aufweist, kann in einer größeren Regionaleinheit ein statistischer Doppelgänger zu anderen Personen sein.

Eine ähnlich riskante Situation ist auch dann gegeben, wenn die Ausprägungen eines Merkmals einfach oder nur sehr schwach besetzt sind. Schwach besetzte Ausprägungen treten bei so umfangreichen Stichproben wie dem Mikrozensus in der Regel nur bei sehr differenziert ausgewiesenen Merkmalen (z.B. Beruf und Branche) auf oder bei den Extremwerten eines Merkmals, die entweder sehr hoch oder sehr niedrig liegen. Als typisches Beispiel hierfür wird meist das Alter (beispielsweise Personen, die über 100 Jahre alt sind) und das Einkommen angeführt. Aber auch "exotische" und daher seltene Merkmalsausprägungen - wie etwa ein dreifacher Bootsbesitzer (EVS) - fallen in diese Risikogruppe.

Das Problem schwach besetzter Ausprägungen ist dem der regionalen und zugleich sachlichen Tiefengliederung vergleichbar, die Beschaffung von Zusatzwissen wird erleichtert. Ist eine Ausprägung nur einfach besetzt, so kann unter Umständen sogar ein Merkmal genügen, um auf eine Person zu schließen (Brennecke 1980).

Eine Strategie, die wesentlich zur Verringerung dieses Risikofaktors beiträgt, ohne die Daten zu verfälschen, ist die Vergröberung von Merkmalsausprägungen. Die Grundform dieser Technik, die im übrigen auch bei statistischen Analysen - allerdings mit anderen Zielen - des öfteren Anwendung findet, besteht in einer Zusammenfassung von sehr differenziert ausgewiesenen Merkmalen in Kategorien bzw. Intervallen. Hierdurch verringert sich der Differenzierungsgrad des Merkmals, da eine gleichbleibende Anzahl von

Fällen N sich auf weniger Ausprägungen K beziehen. Die Wahrscheinlichkeit von einzigartigen Ausprägungskombinationen sinkt, während die Zahl von statistischen Doppelgängern steigt. Die Ausgrenzung einzelner Fälle wird erschwert, das Reidentifikationsrisiko sinkt. Wenn eine Vergröberung von Merkmalsausprägungen auch keine Verfälschung von Daten im engeren Sinne des Wortes bedeutet, so ist damit zweifelsohne ein Verlust an Informationen verbunden. Je heterogener die Kategorien sind, die bei Vergröberung der Merkmalsausprägungen zusammengefaßt werden müssen, umso mehr Differenzierungspotential geht für die Analyse verloren. Eine zwischen differenziert erfaßten Variablen bestehende Beziehung kann sich im Grenzfall durch Vergröberung der Variablenausprägung zu einer Null-Beziehung verändern und damit zu fehlerhaften Schlußfolgerungen führen. So kann bei einer Zusammenfassung differenzierter Altersangaben unter Umständen die Möglichkeit von Kohortenanalysen entscheidend eingeschränkt oder unmöglich gemacht werden. Diese wichtige Art von Analysen ist oft darauf ausgerichtet, den Weg einer Gruppe von Personen mit gleichem Geburtsjahr (einer "Geburtsjahrgangskohorte") durch verschiedene Stationen des sozialen Lebens zu verfolgen (vgl. Müller 1978, Blossfeld 1985:180, Mayer/Blossfeld 1990). Bei Krupp/Preißl (1989:125) findet sich der Hinweis, daß sich eine zu grobe Klassenbildung äußerst negativ auf Längsschnittanalysen auswirken kann, da damit gerechnet werden muß, daß minimale und dennoch bedeutsame Veränderungen nicht zu einem Überschreiten der Klassengrenzen führen und daher nicht berücksichtigt werden können. Eine pauschale Vergröberung von Merkmalen kann daher insbesondere die Verfolgung mikroanalytischer Fragestellungen stark beeinträchtigen.

Die mit der Vergröberung von Merkmalsausprägungen verbundene Beeinträchtigung des Analysepotentials läßt sich insofern minimieren, als die Vergröberungstechnik gezielt gegen die genannten Risikofaktoren "gleichzeitige regionale und sachliche Tiefengliederung" sowie "schwach" besetzte Ausprägungen unter Berücksichtigung von forschungsrelevanten Aspekten eingesetzt werden kann.

11.2.2.1 Vergröberung von Regionalangaben

Sehr kleinräumige Regionalangaben gelten generell als anonymitätsgefährdend. Hinzu kommt, daß Regionalangaben in der Regel vergleichsweise fehlerfrei erfaßt sind. Deshalb gehört die Vergröberung des Regionalrasters international zu den Standardinstrumenten des Anonymitätsschutzes

(Greenberg 1990). Beispielsweise werden in den USA bei der Freigabe von Census-Samples nur Gebietseinheiten mit mindestens 100 000 Einwohnern ausgewiesen (vgl. Cox 1987:31).

Die Vergröberung von Regionalangaben erschwert nicht nur die Beschaffbarkeit von Zusatzwissen, sondern senkt auch den Informationsgehalt der Daten in beträchtlichem Maße. Denn je größer bei gegebener Bevölkerungsdichte die Bezugsbasis P einer Region ist, auf die sich die Daten beziehen, desto schwieriger wird es werden, einzelne Personen oder kleine Subgruppen auszugrenzen. Wie in Kapitel 4 anhand der Untersuchungen von Greenberg/Voshell (1990) erläutert wurde, nimmt die Zahl der Einzelfälle bei wachsender Bezugsbasis P sehr schnell ab. Ausgehend von den empirischen Ergebnissen sollte daher ähnlich wie beim Census-Sample eine Grenze für Gebietseinheiten festgelegt werden, die im allgemeinen bei faktisch anonymen Daten nicht unterschritten werden sollte.

Die Folgen für das Analysepotential der Daten müssen im Hinblick auf die relevanten Fragestellungen beurteilt werden. Für viele Analyseziele sind kleinräumige Regionalangaben nicht wichtig. Für viele Zwecke ist die Typisierung von Regionen (z. B. "Ballungsgebiet", "ländlicher Raum") günstiger als die exakte Lokalisierung einer Region. Auf diese Weise könnten spezielle Eigenschaften des Raumes charakterisiert werden. Der Nachteil dieser Methode besteht jedoch darin, daß keine eigentlichen Regionalanalysen durchführbar sind, sondern nur Eigenschaften der Region als Kontextmerkmale zur Verfügung stehen. Vor allem die Regionalforschung ist jedoch auf detaillierte Regionalinformationen angewiesen. Darum bietet sich als Kompromiß für die Datenweitergabe an, spezielle Mikrodatenfiles für die Regionalwissenschaften zu konzipieren, die zwar detailliertere Regionalinformationen enthalten, deren sachliche Tiefengliederung jedoch reduziert ist. Welche Schwellen hierbei in Frage kommen, wird in Kapitel 14 ausführlich erläutert.

11.2.2.2 Vergröberung von schwach besetzten oder extremen Merkmalswerten

Eine weitere Maßnahme, die ebenfalls einen wirksamen Schutz vor Deanonymisierungen bietet, besteht in der Vergröberung von (univariat) schwach besetzten Ausprägungen. Ausgehend von einem festzulegenden Besetzungsminimum können alle Merkmalsausprägungen, die unterhalb dieser Schwelle liegen, nach inhaltlichen Kriterien mit anderen Ausprägungen zusammengefaßt

werden. Auf diese Weise kann ausgeschlossen werden, daß eine durch ein Merkmal eindeutig charakterisierte Person in dem Datenmaterial enthalten ist (Brennecke 1980, Südfeld 1987). Zu denken ist hier etwa an "exotische" Berufsgruppen, die im Mikrozensus unter Umständen sogar nur einfach besetzt sind. Durch eine pauschale Vergröberung schwach besetzter Ausprägungen entsteht zwar ebenfalls ein gewisser Verlust an Analysepotential. Dieser dürfte sich jedoch zumeist in Grenzen halten, da sich das wissenschaftliche Interesse in der Regel nicht auf seltene Fälle richtet. Die Bedeutung dieser Veränderung hängt davon ab, wie viele derartige Zusammenlegungen vorgenommen werden.

Eine spezielle Form der Kategorisierung von Merkmalen ist das sogenannte Top-Coding und Bottom-Coding. Dabei werden Extremwerte einer Variablen zu oberen (Top-coding) und unteren (Bottom-coding) Flügelklassen zusammengefaßt. Damit sollen z.B. besonders "auffällige" Personen geschützt werden. So wird angenommen, daß Personen mit extrem hohem Einkommen zumindest in vielen Fällen in der Öffentlichkeit bekannt und dadurch leichter identifizierbar sind. Außerdem ist vorstellbar, daß sich bei Prominenten spezielle Interessen mit einer Deanonymisierung verbinden.[8] Das typische Beispiel für Top-Coding ist darum die Einkommenskategorie: Alle Personen mit einem Einkommen oberhalb einer bestimmten Schranke werden zu einer nach oben offenen Einkommenskategorie zusammengefaßt. Ein analoges Beispiel ist die Zusammenfassung der höchsten Altersjahre.

Da derartige Extremwerte nur bei speziellen Fragestellungen von eigenem Interesse sind, hält sich der Verlust an Informationen für die Forschung wichtige durch diese Art der Zusammenfassung im allgemeinen in einem erträglichen Rahmen. Auf jeden Fall sind beim Top-Coding die Einschränkungen geringer, als wenn von einer Variablen durchgehend alle Ausprägungen kategorisiert werden. Die mit dem Top-Coding verbundenen Informationsverluste können auch dadurch in sinnvoller Weise eingeschränkt werden, daß Angaben über die Verbreitung der zusammengefaßten Kategorien zur Verfügung gestellt werden.

---

[8] Man denke etwa an das in Kapitel 6 erwähnte Beispiel eines Stuttgarter Kommunalpolitikers mit einem Jahreseinkommen von mehr als 100 000 DM.

11.2.3 Maßnahmen, die keine direkte Datenmodifikation vorsehen

11.2.3.1 Entfernen einzelner Merkmale

Eine weitere Anonymisierungsmaßnahme besteht in dem gezielten Entfernen (Supression) von Variablen, mit denen ein besonders hohes Deanonymisierungsrisiko verbunden ist. Hierunter fallen immer die personenbezogenen Angaben, d.h. die Namen und Anschriften der Befragten (*formale Anonymisierung*). Ansonsten würden hierfür vor allem Variablen mit hohem Informationsgehalt, also mit stark differenzierten Ausprägungen oder mit extremer Verteilung der Häufigkeiten in Frage kommen oder Merkmale, für die Zusatzwissen besonders leicht beschaffbar oder ohnehin bekannt ist.[9] Da abgesehen von den personenbezogenen Angaben ein spezifisches Gefährdungspotential nie nur durch ein spezifisches Merkmal sondern erst durch das Zusammentreffen mehrerer Risikofaktoren entsteht (Marsh et al. 1991) und die vollständige Entfernung von Merkmalen eine gravierende Einschränkung des Analysepotentials bedeutet, sollte diese Maßnahme ausschließlich auf personenbezogene Angaben beschränkt bleiben.

11.2.3.2 Die Weitergabe veralteter Daten

Eine Möglichkeit, die oben genannte Angriffskonstellation zu entschärfen, bestünde darin, Mikrodatenfiles erst dann an die Wissenschaft weiterzugeben, wenn sie beispielsweise durch eine neue Erhebung überholt sind (Südfeld 1987). Diese Maßnahme, die keine Datenanonymisierungsmaßnahme in einem engeren Sinne ist, würde ein Reidentifikationsrisiko verringern, da zum einen das Interesse eines potentiellen Angreifers eher auf aktuelle Daten gerichtet sein wird. Zum anderen wird mit zunehmendem zeitlichen Abstand zum Erhebungszeitpunkt die Verfügbarkeit von Zusatzwissen unwahrscheinlicher, bzw.

---

9 Ein weiteres vorstellbares Problem entsteht bei Variablen, die sehr "auffällig" sind: So enthält die EVS das Merkmal "Kosten eines Motor- und Segelbootes" (sofern vorhanden). Die einzelnen Merkmalsausprägungen sind sicher sehr dünn besetzt. Da zudem das Wissen über die Art eines Bootes, z.B. im Fall eines Prominenten weit verbreitet sein kann, könnte sich diese Variable für "Fischzüge" eignen. Allerdings ist dieses Merkmal eine Ausnahme im Erhebungskatalog von Mikrozensus und EVS. Andere derart herausstechende Variablen sind in diesen Erhebungen kaum vorhanden.

die Inkompatibilitäten zwischen aktuellem Zusatzwissen und Mikrodatenfile steigen.

Die Sperrung aktueller Daten hat jedoch den entscheidenden Nachteil, daß die Daten dem Wissenschaftler erst dann zur Verfügung stehen, wenn sie bereits als zeitlich überholt angesehen werden müssen. Insofern sozialwissenschaftliche Analysen den aktuellen gesellschaftlichen Wandel abbilden sollen, werden aktuelle Informationen benötigt. Dies zeigen z.B. Müller und Hauser (1987) unter anderem am Beispiel der Armutsforschung auf. Beratung der Politik durch die Wissenschaft kann auf der Basis von veralteten Informationen nicht geleistet werden, weil gesellschaftliche Probleme erst viel zu spät erkannt und kurzfristig einsetzende Entwicklungen anhand älterer Daten nicht prognostiziert werden können.

Schon jetzt liegen zwischen dem Erhebungszeitpunkt und dem Abschluß der Datenaufbereitung mehrere Monate (Herberger 1985). Ein Zeitraum also, in welchem bereits mit "Alterungsfehlern"[10] gerechnet werden muß, die zusätzlich zu sonstigen Inkompatibilitäten die Reidentifikationsversuche bereits beträchtlich stören. Daher kann von einer zusätzlichen zeitlichen Sperrung bei Übermittlung von Einzelangaben abgesehen werden.

11.2.3.3 Systemfreie Sortierung

Ebenfalls in den Bereich der Maßnahmen, die ohne direkte Datenmodifikation wirksam werden, fällt die systemfreie Sortierung[11] der Datensätze in einem Mikrodatenfile. Damit soll verhindert werden, daß ein Angreifer anhand einer systematischen Anordnung das Datenmaterial (etwa nach regionalen Kriterien) eingrenzt und sich für diese eingegrenzte Population gezielt Zusatzwissen besorgt (Südfeld 1987). Diese Maßnahme ist in der Regel nur dann mit einer Einschränkung des Analysepotentials der Daten verbunden, wenn für

10 Im Zusammenhang mit der Volkszählung weist beispielsweise Brunnstein (1987:78f) auf diesen Alterungsprozeß hin. Interessanterweise verknüpft er hiermit allerdings nur den Wertverlust dieser Daten für statistische Analysen, die hiervon zugleich ausgehende Schutzwirkung gegenüber Reidentifikationsversuchen wird von ihm nicht erwähnt.

11 Eine systematische Sortierung wäre beispielsweise dann gegeben, wenn alle in einer Straße befragten Haushalte jeweils in aufeinanderfolgenden Datensätzen abgebildet wären.

bestimmte Untersuchungen - wie etwa bei Kontextanalysen - eine systematische Sortierung der Daten beispielsweise nach Regionalinformationen erforderlich ist. Diese Schutzvorkehrung ist daher durchaus sinnvoll, solange Sonderregelungen für jene Fälle vorgesehen werden, in welchen eine systematische Sortierung benötigt wird.

11.2.3.4 Ziehung von zufälligen Substichproben

Im Verlauf der bisherigen Erörterungen wurde mehrfach deutlich, daß Datenfiles, die aus Stichproben bestehen, in mehrfacher Weise ein wesentlich geringeres Deanonymisierungsrisiko aufweisen. Zum einen setzt jede Reidentifikation voraus, daß sich die Angaben einer gesuchten Person sowohl im Mikrodatenfile als auch im Identifikationsfile befinden. Sofern eine beliebige, in der Grundgesamtheit enthaltene Person in einer Stichprobe gesucht wird, entspricht die maximale Erfolgswahrscheinlichkeit dem Auswahlsatz der Stichprobe.[12] Da der Auswahlsatz sowohl beim Mikrozensus mit 1 Prozent als auch bei der EVS mit 0,2 Prozent der Gesamtbevölkerung vergleichsweise klein ist, sind diese Datenbestände schon von vorneherein sehr gut vor Reidentifikationsversuchen geschützt. Verfügt ein Angreifer ebenfalls nur über eine Stichprobe, die er als Identifikationsfile einsetzt, so ergibt sich die Wahrscheinlichkeit, daß eine beliebige Person in beiden Datensätzen enthalten ist, durch die Multiplikation der Auswahlsätze von beiden Stichproben.

Die Stichprobeneigenschaft der Daten bedeutet damit nicht nur eine prinzipielle Schranke für die Möglichkeit des Reidentifikationserfolgs, sondern stellt auch einen wichtigen Unsicherheitsfaktor bei Reidentifikationsversuchen dar. Wie bereits mehrfach ausgeführt, weiß ein Angreifer auch bei einer eindeutigen Zuordnung nie mit Sicherheit, ob er die richtige Person zuordnet, wenn sein Datenvorrat die Grundgesamtheit nicht vollständig umfaßt. In diesem Fall würde lediglich die Möglichkeit bleiben, die Wahrscheinlichkeit des Auftretens von statistischen Doppelgängern in der Grundgesamtheit abzuschätzen. Sofern diese Wahrscheinlichkeit ausreichend gering ist, könnte der Angreifer annehmen, daß keine Doppelgänger in der Grundgesamtheit existieren und somit eine Zuordnung vornehmen (vgl. Dittrich/Schlörer 1985). Allerdings hat die

12 Vgl. zum folgenden auch Kirschner (1987). Die Schutzwirkung von Stichproben wird im übrigen auch von Brunnstein (1987:79f.) betont.

empirische Überprüfung gezeigt, daß gerade derartige Wahrscheinlichkeitsabschätzungen im Realversuch zu vielen Fehlzuordnungen führen.

Die von der Stichprobeneigenschaft ausgehende Schutzwirkung besteht jedoch dann nicht mehr, wenn dem Angreifer - wie in der oben charakterisierten Risikokonstellation - bekannt ist, daß die von ihm gesuchte Person an der amtlichen Erhebung teilgenommen hat. In diesem Fall ist die Wahrscheinlichkeit, daß die betreffende Person in der Stichprobe enthalten ist, mit eins anzusetzen. Findet sich unter dieser Bedingung und der Annahme, daß die Daten kompatibel sind, im Mikrodatenfile ein Datensatz mit einer einzigartigen übereinstimmenden Ausprägungskombination, kann ein Angreifer sicher sein, eine korrekte Zuordnung vorgenommen zu haben.[13] Das Risiko dieses spezifischen Wissens ist in der Regel sehr klein, da die lokale Umsetzung der Stichprobenpläne geheimgehalten wird. Allerdings kann nicht ausgeschlossen werden, daß die Befragten selbst Dritten mitteilen, daß sie beispielsweise an einer Mikrozensuserhebung teilgenommen haben.

Der sich spezifisch aus dem potentiellen response knowledge eines Angreifers ergebende Risikofaktor kann gezielt entschärft werden, wenn nur nach dem Zufallsprinzip gezogene Substichproben aus einem Mikrodatenfile weitergegeben werden (vgl. u.a. Paaß/Wauschkuhn 1985, Südfeld 1987, Marsh et al. 1991). Die Weitergabe von Substichproben versetzt einen potentiellen Angreifer systematisch in Unkenntnis darüber, ob eine gesuchte Person in der weitergegebenen Substichprobe enthalten ist oder nicht. Ob es sich tatsächlich um den Datensatz der gesuchten Person handelt, kann selbst dann nicht mit Sicherheit gesagt werden, wenn bekannt ist, daß eine bestimmte Person an einer amtlichen Erhebung teilgenommen hat, und in der weitergegebenen Substichprobe des Mikrodatenfile ein einzigartiger Datensatz gefunden wird, der den Merkmalsausprägungen der gesuchten Person entspricht. Denn es besteht die Möglichkeit, daß die gesuchte Person in dem nicht weitergegebenen Teil des Mikrodatenfile enthalten ist, während es sich bei dem im weitergegebenen Teil gefundenen Fall um einen statistischen Doppelgänger handelt.

Da die Ziehung zufälliger Substichproben zugleich auch den erwünschten Nebeneffekt einer allgemeinen Reduzierung des Reidentifikationspotentials mit sich bringt, bietet diese Maßnahme einen sehr wichtigen Schutz vor Reidentifikationsversuchen.

---

13 Sind die Daten hingegen nicht kompatibel, so ist, wie in Abschnitt 9.3.5.2 gezeigt, auch bei Teilnahmekenntnis eine absolut sichere Zuordnung nicht möglich, und der Datenangriff kann unter Umständen in die Irre führen.

Unter dem Gesichtspunkt der Bewahrung des Analysepotentials der Daten ist bei der Weitergabe von zufällig ausgewählten Substichproben daraufhinzuweisen, daß mit dieser Anonymisierungsstrategie keine Modifikationen der Daten (etwa Einbringen falscher Angaben in das Datenmaterial) verbunden ist, die sich negativ - im Sinne von Verzerrungen oder Verfälschungen - auf die statistischen Analysen auswirken können. Die zufällige Auswahl beim Prozeß der Substichprobenziehung ist durch die entwickelten Methoden der modernen Statistik beherrschbar.

Durch die Reduzierung des Stichprobenumfangs erhöhen sich zwar die Schätzfehler. Die großen amtlichen Erhebungen, wie z. B. der Mikrozensus, sind jedoch derart umfangreich, daß die Aussagekraft in aller Regel nicht wesentlich eingeschränkt wird, wenn eine Teilstichprobe von z.B. 80 Prozent des ursprünglichen Gesamtfile weitergegeben wird. Dieses gilt allerdings nicht, wenn sich die Untersuchungen auf spezielle Subpopulationen oder tiefgegliederte Regionaleinheiten beziehen, da hier schon die vollständigen Stichproben meistens nur wenige Fälle enthalten. Deshalb muß bei einer regionalen Tiefengliederung die Substichprobenziehung sehr vorsichtig gehandhabt werden.

11.3 Empirische Überprüfung ausgewählter Anonymisierungsmaßnahmen

Ursprünglich bestand die Absicht, eine breite Palette unterschiedlichster Anonymisierungsmaßnahmen mit den zur Verfügung stehenden Daten empirisch zu überprüfen. Diese Absicht mußte und konnte nach Vorliegen der empirischen Befunde revidiert werden.

Bei Verwendung einer sozialwissenschaftlichen Stichprobe als Identifikationsfile erwies sich der Mikrozensus bereits bei lediglich formaler Anonymisierung - aufgrund der nahezu unvermeidbaren Dateninkompatibilitäten und der Stichprobeneigenschaft des Zusatzwissens - als sicher vor Reidentifikationsversuchen. Diese Datenbasis konnte daher für die weitere Überprüfung von Anonymisierungsmaßnahmen nicht mehr verwendet werden. Somit verblieb für die empirische Überprüfung von Anonymisierungsmaßnahmen das Gelehrtenszenario. Entsprechend der tatsächlich gegebenen Datenlage[14] stehen

[14] Zur Erinnerung sei nochmals kurz die Datenbasis des Gelehrtenszenarios beschrieben. Es standen zehn Überschneidungsmerkmale zur Verfügung. Der

dabei nur relativ wenige Überschneidungsmerkmale zur Verfügung. Zugleich waren auch schon bei den Datenfiles, bei denen neben Weglassen von Name und Anschrift keine weiteren Anonymisierungsvorkehrungen getroffen waren, nur eine verschwindend geringe Zahl korrekter Zuordnungen ermittelt worden. Durch Anonymisierungsmaßnahmen verringert sich diese Zahl zwar weiter. Bei der kleinen Zahl der Fälle sind exakte Angaben der Effektivität von Anonymisierungsmaßnahmen jedoch nicht möglich. Wenn im folgenden dennoch entsprechende Zahlen genannt werden, so ist zu berücksichtigen, daß hierbei Zufallsschwankungen eine erhebliche Rolle spielen können. Anhand der empirischen Überprüfungen kann daher nur das Störpotential für Deanonymisierungsversuche aufgezeigt werden.

Zur Überprüfung kamen entsprechend den obigen Ausführungen die Vergröberung von Merkmalsausprägungen sowie die Ziehung von Stichproben hinzu. Die Datenbasis sowie die Vorgehensweise entspricht den Darstellungen in Kapitel 8 und 9, allerdings mit dem Unterschied, daß die Mikrozensusdaten vor den Reidentifikationssimulationen im Sinne der jeweiligen Anonymisierungsmaßnahmen verändert wurden.

11.3.1 Vergröberung von Merkmalsausprägungen

Zur Verringerung des Reidentifikationsrisikos aufgrund der Zugehörigkeit zu einer spezifischen, zahlenmäßig kleinen Subpopulation oder aufgrund der Existenz von tief gegliederten Regionalinformationen auf seiten des Angreifers, wurde die von einer Vergröberung von Merkmalen ausgehende Schutzwirkung geprüft.

Wie oben dargestellt, bewirkt die Ausprägungsvergröberung einen reduzierten Differenzierungsgrad dieser Merkmale, wodurch der Informationsgehalt des Merkmals unpräziser wird. Durch eine Verringerung der Ausprägungsmöglichkeiten K bei einer gleichbleibenden Zahl von Fällen N sinkt die Wahrscheinlichkeit des Auftretens von einzigartigen Ausprägungskombinationen. Zugleich steigt die Zahl der statistischen Doppelgänger innerhalb des Mikro-

Mikrozensus umfaßte 168 368 Datensätze, der Gelehrtenkalender 7 983 Datensätze. Wenn alle 16 Versionen des Gelehrtenkalenders berücksichtigt wurden, konnten insgesamt 14 Fälle eindeutig zugeordnet werden. Hierbei handelt es sich allerdings nur in vier Fällen tatsächlich um Datensätze von ein und derselben Person. Die verbleibenden zehn Zuordnungen waren falsch.

datenfile.[15] Durch die verringerte Zahl von Merkmalsausprägungen erhöht sich zugleich auch die Wahrscheinlichkeit von übereinstimmenden Ausprägungskombinationen zwischen Mikrodaten- und Identifikationsfile. Hierdurch kann sich die Zahl der auf der Basis von identischen Ausprägungskombinationen vorgenommenen Zuordnungen - und zwar sowohl der ein- als auch der mehrdeutigen - erhöhen. Für einen Angreifer ergibt sich hieraus zum einen eine erhöhte Unsicherheit, da die Zahl der Falschzuordnungen steigt. Zum anderen entsteht ein Mehraufwand, da sich die Relation zwischen falschen und korrekten Zuordnungen zugunsten der falschen verschiebt und eine Herausfilterung von möglichen korrekten Zuordnungen daher aufwendiger wird.

Die Wirkungsweise dieser Anonymisierungsmaßnahme läßt sich durch eine Vergröberung der regionalen Tiefengliederung am Beispiel des Gelehrtenszenarios verdeutlichen. Stellt die Gemeindegrößenklasse die kleinste Regionaleinheit dar, so beträgt der Anteil der einzigartigen Ausprägungskombinationen im Gesamtmikrodatenfile ca. 39 Prozent. Der Anteil der Einzelfälle verringert sich auf 27 Prozent, wenn keine Regionalinformationen unterhalb des Bundeslands zur Verfügung stehen. Dieser Effekt zeigt sich noch deutlicher, wenn man im Mikrodatenfile nur die für die einfachen Zuordnungsverfahren relevante Population betrachtet.[16] Bei einer regionalen Differenzierung auf Bundesraumordnungsebene werden in dieser Subpopulation knapp 80 Prozent der Fälle als einzigartig vertreten ausgewiesen. Auf Bundeslandebene hingegen beträgt der Anteil der einzigartigen Datensätze nur noch 60 Prozent.

In den Reidentifikationssimulationen hat diese Vergröberung zwei gegenläufige Effekte zur Folge:

Einzelne Fälle, die aufgrund der differenzierten Regionalinformationen eindeutig zugeordnet werden können, weisen in den zusammengefaßten Regionaleinheiten statistische Doppelgänger auf. Sie können daher nicht mehr eindeutig zugeordnet werden. Auf diese Weise verringert sich die Zahl der vorher korrekt vorgenommenen eindeutigen Zuordnungen, weil ein Teil der einzigartigen Ausprägungskombinationen entfällt.

Durch die weniger differenzierten Regionalangaben finden sich jedoch auf der anderen Seite neue Fälle, die im Mikrodatenfile und im Identifikationsfile

---

15 Hierdurch erhöht sich natürlich auch die Zahl der statistischen Doppelgänger in der Population.

16 Entsprechend der in Abschnitt 9.1.3 gegebenen Definition handelt es hierbei um die Fälle, die eine Ausprägung aufweisen, die auch im Gelehrtenkalender auftritt.

(Gelehrtenkalender) übereinstimmende einzigartige Ausprägungskombinationen aufweisen. Es können insgesamt mehr eindeutige Zuordnungen vorgenommen werden. Dabei erweisen sich alle zusätzlichen Zuordnungen jedoch als falsch, da diese Zuordnungen ja nur dadurch zustande kommen, weil die vorher zwischen den Fällen differenzierende Rauminformation wegfällt. Eine Person mit einer einzigartigen Ausprägungskombination im Mikrodatenfile, die ein Pendant mit einer gleichen Ausprägungskombination im Identifikationsfile aufweist, die aber dort einer anderen Regionaleinheit als im Mikrodatenfile angehört, wird eindeutig - aber falsch - zugeordnet, wenn die trennende Regionalinformation wegfällt.

Die Wirkung dieser gegenläufiger Prozesse zeigt sich bei der durchgeführten Simulation in folgender Weise. In dem Originalmikrodatenfile standen Bundesland, Bundesraumordnungsregion und Gemeindegrößenklasse als Regionalinformationen zur Verfügung. Hier erbrachten die Simulationen 14 eindeutige Zuordnungen, von welchen vier korrekt waren. Steht die Gemeindegrößenklasse als Regionalinformation nicht mehr zur Verfügung, erhöht sich die Zahl der eindeutigen Zuordnungen von den ursprünglichen 14 auf 20. Gleichzeitig sinkt die Zahl der korrekten Zuordnungen von vier auf zwei, d.h. der Anteil der Falschzuordnungen steigt von 71 auf 90 Prozent. Enthält das Mikrodatenfile keine Regionalinformationen unterhalb der Bundesländerebene, so können 24 eindeutige Zuordnungen ermittelt werden. Allerdings ist keine dieser Zuordnungen mehr korrekt, d.h. die Fehlerquote liegt bei 100 Prozent (vgl. Tabelle 11.1).

Tabelle 11.1: Die regionale Vergröberung als Anonymisierungsmaßnahme

Berücksichtigte Regionalinformationen	Eindeutige Zuordnungen insgesamt	hiervon: falsch	korrekt
BL, BROR, GK*	14	10 (71%)	4 (29%)
BL, BROR	20	18 (90%)	2 (10%)
BL	24	24 (100%)	0 (0%)

*BL: Bundesland; BROR: Bundesraumordnungsregion; GK: Gemeindegrößenklasse

Ähnliche Ergebnisse zeigen sich auch bei einer Vergröberung der differenziert erfaßten Merkmale Hochschulabschluß und Geburtsjahr. So hat die Zusam-

menfassung des Merkmals "Hochschulabschluß" von 91 möglichen Ausprägungen in zehn Fachrichtungsklassen im Mikrozensus zur Folge, daß sich der Anteil der falschen Zuordnungen von ursprünglich 71 Prozent auf 83 Prozent erhöht. Wird das Geburtsjahr in Fünfer-Schritten zusammengefaßt, erhöht sich die Quote der Falschzuordnungen auf 86 Prozent (vgl. Tabelle 11.2).

Die erhöhte Zahl von möglichen Zuordnungen aufgrund der Vergröberung differenzierter Merkmale wirkt sich bei Reidentifikationsversuchen zunächst aufwandserhöhend aus. Es müssen mehr Zuordnungen im Hinblick auf ihre Korrektheit überprüft werden, wobei sich die Quote der korrekten Zuordnungen in der Regel nicht erhöht, sondern sinkt, d.h. bei geringer werdendem Ertrag steigen die Kosten. Für die Verifizierung der erzielten Zuordnungen werden zusätzliche Informationen in Form von Überschneidungsmerkmalen benötigt. Diese Informationen müssen für jede Zuordnung gezielt ermittelt werden. Nach der Analyse des Zusatzwissens ist davon auszugehen, daß diese Informationsbeschaffung in der Regel nicht einfach ist. Mit jedem zusätzlich einbezogenen Merkmal verschärft sich auch das Problem der Inkompatibilität zwischen Zusatzwissen und Mikrodatenfile. Eine Merkmalsvergröberung bewirkt also sowohl eine Erhöhung des Aufwands als auch der Unsicherheit des Angreifers bei gleichzeitig sinkendem Ertrag.

Tabelle 11.2: Vergröberung von Merkmalsausprägungen bei den Variablen Hochschulabschluß und Geburtsjahr

Vergröberung von Überschneidungsmerkmalen	Eindeutige Zuordnungen insgesamt	hiervon: falsch	korrekt
keine	14	10 (71%)	4 (29%)
Hochschulabschluß (10 Fachrichtungsklassen)	18	15 (83%)	3 (17%)
Geburtsjahr (Fünf-Jahresklassen)	22	19 (86%)	3 (14%)

11.3.2 Ziehung zufälliger Substichproben

Für das spezifische Gefährdungspotential, welches daraus resultieren könnte, daß ein Angreifer über eine bestimmte Person weiß, daß diese an einer amt-

lichen Erhebung teilgenommen hat, wurde das Mittel der Substichprobenziehung geprüft. Mit fallender Substichprobengröße verringert sich die Wahrscheinlichkeit, daß eine beliebige Person in den übermittelten Mikrodaten enthalten ist. Zugleich hat die Substichprobenziehung den Effekt, daß ursprünglich mehrdeutige Fälle eindeutig werden können. Dadurch erhöht sich in einem informationstheoretischen Sinn die Unsicherheit des Angreifers über die Korrektheit seiner Zuordnungen beträchtlich. Selbst wenn ein Angreifer weiß, daß eine bestimmte Person an einer amtlichen Erhebung teilgenommen hat, kann er - falls nur eine Substichprobe weitergegeben wird - auch bei einer eindeutigen Zuordnung nicht sicher sein, ob es sich hierbei um die gesuchte Person oder um einen statistischen Doppelgänger handelt. Geht man im Fallbeispiel des Gelehrtenkalenders von dem ursprünglich möglichen Potential eindeutiger Zuordnungen aus und setzt dieses als Maßstab, so zeigt sich, daß mit der Ziehung von zufälligen Substichproben eine deutliche Einschränkung sowohl der absolut möglichen als auch der hierbei korrekten Zuordnungen einhergeht. Wird eine Substichprobe von 75 Prozent des Original-Mikrodatenfile ohne sonstige Anonymisierungsmaßnahmen gezogen, so sind von den eindeutigen Zuordnungen nur noch zwei korrekt. War das Mikrodatenfile eine 50 Prozent-Stichprobe, sinkt die Zahl der korrekten Zuordnungen von ursprünglich vier auf eine, der Anteil der Falschzuordnungen erhöht sich auf 83 Prozent (vgl. Tabelle 11.3).

Tabelle 11.3: Schutzwirkung durch Ziehung von zufälligen Substichproben

Auswahlsatz aus dem Mikrozensus	Eindeutige Zuordnungen insgesamt	hiervon: falsch	korrekt
100% (Original MZ)	14	10 (71%)	4 (29%)
75%	9	7 (78%)	2 (22%)
50%	6	5 (83%)	1 (17%)

Was aus dieser Tabelle nicht hervorgeht, sich aber bei der Analyse der Zuordnungen gezeigt hat, ist, daß es durch die Ziehung der Substichprobe zu einer Verschiebung von ursprünglich mehrdeutigen zu eindeutigen Zuordnungen kommt. In einem Fall wurde hierbei der statistische Doppelgänger eindeutig, während der korrekte Datensatz nicht mehr in der Stichprobe enthalten war. Ein Angreifer würde in diesem Fall eine Falschzuordnung vornehmen. Wie bereits

erwähnt und auch aus den Tabellen offensichtlich, sind die Fallzahlen äußerst gering. Daher kann hier nicht gefolgert werden, daß bei einer Substichprobenziehung mit einem Auswahlsatz von 50 Prozent nur noch ein Viertel des ursprünglich vorhandenen Reidentifikationspotentials verbleibt. Gemäß einer Abschätzung von Wahrscheinlichkeiten ist hier eher mit einer Halbierung des Risikos zu rechnen. Dennoch zeigt sich deutlich, daß die Ziehung von zufälligen Substichproben ein vorhandenes Reidentifikationspotential reduziert und damit durch den erzeugten Unsicherheitsfaktor bezüglich der in der Substichprobe enthaltenen Fälle eine außerordentlich wichtige Schutzwirkung entfaltet.

11.4 Zusammenfassung

Bedingt durch die Beschränkungen, die mit den verwendeten Daten gegeben sind, konnte hier nur eine exemplarische Überprüfung der von einer Merkmalsvergröberung und Substichprobenziehung ausgehenden Schutzwirkung geleistet werden. Bereits diese wenigen Beispiele zeigen jedoch, daß - durch zum Teil geringfügige - Veränderungen in einem Mikrodatenfile einfache Abgleichverfahren nachhaltig gestört werden können. Die damit einhergehende vermehrte Zahl von möglichen Falschzuordnungen erhöht ganz allgemein bei einem realen Datenangriff zum einen die Unsicherheit des Angreifers bezüglich korrekter und falscher Zuordnungen; zum anderen müssen für sichere Reidentifikationen mehr Zuordnungen überprüft werden, was mit einer Aufwandssteigerung verbunden ist. Um der zu Beginn dieses Kapitels charakterisierten Risikokonstellation wirksam entgegenzutreten, ist es wichtig, die verschiedenen in Abschnitt 11.2.2 erörterten Vergröberungsmaßnahmen für Regionalmerkmale, schwach besetzte Zellen und Extremwerte in sinnvoller Weise mit der Substichprobenziehung zu verbinden. Diese Verbindung soll in Kapitel 14 geleistet werden. Eine wesentliche Determinante für die Entwicklung eines allgemeinen Anonymisierungskonzeptes für die faktische Anonymität von Mikrodaten besteht allerdings auch in einer Abschottung der gelieferten Daten vor unbefugtem Zugriff sowohl von externer als auch von interner Seite. Deshalb müssen im folgenden Kapitel zunächst die Bedingungen der Datenverarbeitung innerhalb des Wissenschaftsbetriebs erörtert und jene technisch-organisatorischen Sicherungsmaßnahmen geprüft werden, die es erschweren bzw. verhindern, daß Daten den Wissenschaftsbereich verlassen. Erst dann ist es auf der Grundlage einer zusammenfassenden Darstellung der Einzelergebnisse möglich, Empfehlungen für die zukünftige Übermittlung von faktisch anonymen Mikrodaten zu entwickeln.

12 Technisch-organisatorische Sicherungsmaßnahmen gegen unbefugte Datenzugriffe bei faktisch anonymen Daten*

12.1 Faktische Anonymität und Datensicherung

Der Gesetzgeber hat festgelegt, daß Einzeldaten aus der amtlichen Statistik nur bei Einhaltung definierter Auflagen an die Wissenschaft weitergegeben werden dürfen. Die zentrale Bedingung nach §16 Abs.6 Bundesstatistikgesetz ist, daß die Daten faktisch anonym sind, d. h. einem Betroffenen "nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können."

Außerdem wird in §16 Abs.8 bestimmt, daß die Daten nur für spezifizierte wissenschaftliche Zwecke verwendet und nur berechtigten Personen zugänglich gemacht werden. Im folgenden wird untersucht, wie zu den letztgenannten Voraussetzungen durch technisch-organisatorische Datensicherungsmaßnahmen beigetragen werden kann. Auch wenn diese Maßnahmen primär an §16 Abs.8 orientiert sind, tragen sie zusätzlich zum Schutz der faktischen Anonymität der Daten bei, da der Aufwand für Deanonymisierungsversuche steigt, wenn ein potentieller Datenangreifer für sein Vorhaben zusätzliche Schutzvorkehrungen überwinden muß.

Aus der Diskussion der Maßnahmen werden Folgerungen für einen Sicherheitsstandard abgeleitet, der bei der Weitergabe von anonymen Daten der amtlichen Statistik an die Sozialwissenschaft generell zu beachten sein wird. Die vorhandenen gesetzlichen Grundlagen werden als gegebener Ausgangspunkt betrachtet (zur Interpretation vgl. vor allem Dorer et al. 1988) und selbst nicht hinterfragt. Kriterien für die Konkretisierung der gesetzlichen Vorschriften und der Anwendung von Maßnahmen werden der Datenschutzdiskussion unter Juristen und EDV-Fachleuten entnommen. Allerdings ist für diese Diskussion der Schutz personenbezogener Daten maßgeblich. Hier gilt ein anderer Bezugspunkt, da wichtigster Schutz der Daten, um die es hier geht, ihre Anonymität ist. Die veränderte Aufgabenstellung erlaubt, auf einige bei personenbezogenen Daten wichtige Sicherungsmaßnahmen zu verzichten. Ein

---

* Autor: Uwe Blien

Beispiel von Bedeutung ist ein Teilaspekt der im Bundesdatenschutzgesetz geforderten Speicherkontrolle: Danach muß die Löschung und Veränderung von personenbezogenen Daten kontrolliert werden. Im Falle faktisch anonymer Daten mag eine vergleichbare Kontrolle im Sinne des Dateninhabers sein, tangiert jedoch kein schutzwürdiges Interesse desjenigen, dessen Angaben in der betreffenden Datei enthalten sind.

Die Abwägung der Notwendigkeit bestimmter Maßnahmen ist sorgfältig durchzuführen. Kehrseite der Sicherheit, die durch technisch-organisatorische Schutzvorkehrungen erreicht wird, ist nämlich eine zusätzliche Restriktion für die Durchführung sozialwissenschaftlicher Forschungsprojekte, weil die ohnehin stets knappen finanziellen Mittel für zusätzliche Aufgaben ausgegeben werden müssen, die kein genuiner Teil sozialwissenschaftlicher Forschung sind.[2]

Gemäß der gängigen juristischen Auslegung des Grundgesetzes und der einschlägigen gesetzlichen Regelungen ist eine Interessenabwägung zwischen Datensicherung und unbehinderter Forschung vorzunehmen, die nicht ausschließlich zugunsten einer Seite entschieden werden darf.[3] Die Forschung muß sich darauf einstellen, bestimmten Anforderungen hinsichtlich Datensicherheit zu genügen, für die die entsprechenden Ressourcen bereitzustellen sind. Andererseits dürfen die Auflagen nicht zum Erliegen wichtiger Teilbereiche der Forschung führen.

Grundvoraussetzung der folgenden Analyse ist, daß die direkten Identifikatoren, also Namen und Adressen, nicht in den Datensätzen enthalten sind, da nur anonyme Daten weitergegeben werden dürfen. Zur Sicherung der faktischen Anonymität sind die im vorangegegangenen Kapitel erörterten datenorientierten Schutzmaßnahmen ebenfalls erforderlich, die hier nicht mehr näher betrachtet werden sollen. Allerdings wird angenommen, daß ein sehr geringes Risiko der Deanonymisierung verbleibt. Andernfalls, im Fall "absoluter Anonymität", wären

---

2 Von Heinz wird geschätzt, daß "an Forschungsprojekten, in denen die erforderlichen Datenschutzmaßnahmen realisiert werden, mit einem um 20 bis 30 Prozent erhöhten Zeitaufwand zu rechnen ist" (1987:175). Die Kosten würden sich entsprechend erhöhen. Die Abschätzung von Heinz gilt allerdings für personenbezogene Daten, bei denen der Schutz durch Anonymität entfällt. Sie bezieht sich darüberhinaus in erster Linie auf sein Forschungsgebiet, die Kriminologie. In diesem Zusammenhang werden besonders 'sensible' Daten verwendet. Darum ist dort mit wesentlich härteren Auflagen hinsichtlich der Datensicherheit zu rechnen als bei Daten, die z.B. aus einer Mikrozensuserhebung stammen. Deshalb ist in diesem Fall von niedrigeren Kosten auszugehen.

3 Die Interessenabwägung erfolgt analog der Situation bei personenbezogenen Daten (vgl. Dammann 1981b:609).

Datensicherungsmaßnahmen überflüssig.[4] Der Begriff Datensicherung bezieht sich in diesem Zusammenhang ausschließlich auf Maßnahmen, die die Interessen jener Personen schützen, deren Daten gespeichert sind.[5]

Die Aufgabe derartiger Maßnahmen besteht darin:

1) Generelle Zugangsschranken zu den Daten zu errichten, die nur von berechtigten Personen überwindbar sind;

2) Die Art der mit den Daten durchführbaren Operationen auf einen zulässigen Bereich zu beschränken;

3) Kontrollen der Operationen zu ermöglichen, die ein Benutzer mit den Daten ausführt.

Die Maßnahmen richten sich an zwei unterscheidbare Personenkreise: Einerseits sollen 'Insider', d. h. Personen, die die Verantwortung für die Daten übernommen haben, an unerlaubten Operationen gehindert werden. Zum anderen sollen die Vorkehrungen allen 'Outsidern' generell den Datenzugriff verwehren. Unberechtigte Dritte könnten unkalkulierbar starke Interessen an deanonymisierten Daten haben und über einen umfänglichen Vorrat an "Zusatzwissen" verfügen, der bei einem Reidentifikationsversuch helfen würde.

Schließlich dienen die Maßnahmen zum Schutz vor Datenzugriffen durch Wissenschaftler, die nicht zum Kreis der berechtigten Benutzer gehören, also in diesem Sinne ebenfalls Outsider sind.

---

4 Auch Datensicherungsmaßnahmen bieten keinen absoluten Schutz. Dies wird jedoch selbst bei personenbezogenen Daten nicht verlangt (vgl. auch Steinmüller 1980a:114, Pütter 1981).

5 Vgl. z.B. Szyperski et al. 1982:A 6. Damit zu kontrastieren ist, daß in anderem Zusammenhang als Datensicherung der weite Bereich der "Maßnahmen zur Wahrung von Sicherungs- und Integritätsbedingungen" (Reuter 1987 Aufsatztitel, vgl. Dammann 1981a:306ff., Dittrich 1981:257) z.B. in Datenbanken bezeichnet wird. Damit sind umfängliche, vor allem technische Vorkehrungen gemeint, die der ungehinderten Datenverarbeitung und -speicherung und damit unmittelbar den Interessen der Stelle dienen, die über die Daten verfügt. Die in den letzten Jahren intensivierte Diskussion um den Schutz von Daten vor "Computerviren" fällt ebenfalls unter diesen Begriff, der im folgenden nicht weiter betrachtet wird.

Das Ziel der weiteren Überlegungen ist die Zusammenstellung einer "Standardliste" von Sicherheitsvorkehrungen. Dabei dürfen die verschiedenen Vorkehrungen nicht isoliert behandelt werden: Die Bezugnahme auf eine Liste ist notwendig, weil Querverbindungen zwischen einzelnen Vorkehrungen zu beachten sind:

- Verschiedene Maßnahmen können wechselseitig substitutiv eingesetzt werden. Deshalb wird nicht erforderlich sein, alle zu diskutierenden Vorkehrungen gleichzeitig zu realisieren.

- Umgekehrt gibt es Maßnahmen, die komplementären Charakter haben. In diesem Fall ist eine bestimmte Vorkehrung ohne eine andere sinnlos, weil sie dann leicht umgangen werden kann: Datensicherungsmaßnahmen müssen Systemlösungen sein. Bei der Zusammenstellung der Standardliste dürfen keine Lücken im System entstehen.[6]

In die Liste werden Maßnahmen aufgenommen, wenn ihre Realisierung für die Weitergabe von amtlichen Einzeldaten an die Sozialwissenschaft vorausgesetzt wird. Allerdings wird die vorzuschlagende Standardliste ein bestimmtes Sicherheitsniveau repräsentieren, das zu beachten ist, und nicht einen unter allen Umständen einzuhaltenden Katalog. Die in der Forschung verwendeten Rechnertypen unterscheiden sich ebenso stark wie die lokalen Rahmenbedingungen der Datenverarbeitung. Kann eine Maßnahme auf einem bestimmten Rechner nicht verwirklicht werden, ist nach adäquatem Ersatz zu suchen.

Gemäß dem Grundsatz der Berücksichtigung der Interessen der Forschung wird bei der Zusammenstellung der Liste nach Möglichkeit von vornherein auf die jeweils anfallenden Kosten geachtet: Die zu realisierenden Maßnahmen sollen möglichst wirksam, aber für den legitimen Nutzer der Daten möglichst wenig aufwendig sein. Vorkehrungen, die nur wenig zusätzlichen Schutz, aber bedeutende Kosten für die Datenverarbeitung mit sich bringen, sind durch andere Lösungen zu ersetzen.

---

6 Zur Gewährleistung der Datensicherheit ist stets die Entwicklung einer EDV-Gesamtkonzeption erforderlich (vgl. Droux 1984:214ff., Weise 1981). Hinweise auf die Erstellung eines Sicherheitssystems finden sich unter anderem in Schönberg (1984:123ff.).

Ansonsten ist der Umfang und Inhalt des im folgenden entwickelten Maßnahmepakets durch zwei von Abel et al. (1989:485) genannte allgemein geltende Regeln bestimmt, die auch auf anonyme Daten angewendet werden können:

- Die Maßnahmen der Datensicherheit müssen sich am Schutzbedürfnis der Daten orientieren.
- Aufwand und Schutzzweck müssen in einem vertretbaren Verhältnis zueinander stehen.

12.2 Datensicherungsmaßnahmen, die überwiegend der Abschottung gegenüber externen Personen dienen

Gemäß der angeführten Unterscheidung werden zunächst Maßnahmen behandelt, die eher dem Schutz der Daten vor dem Zugriff externer Personen dienen. Allerdings tragen bestimmte Vorkehrungen zusätzlich zur internen Datensicherheit bei.

Da Massendaten in erster Linie an Großrechnern verarbeitet werden, orientieren sich die im folgenden vorgestellten Maßnahmen hinsichtlich der technischen Seite an den Möglichkeiten der Großrechner jener Typen, die an wissenschaftlichen Rechenzentren verbreitet sind.

Allerdings werden kleinere Rechner - bis hinunter zur Workstation oder zum Personal Computer - zunehmend leistungsfähiger, so daß es durchaus möglich wird, auch einen Mikrozensus auf ihnen zu verarbeiten. Darum ist - wenigstens kurz - auf Schutzmaßnahmen bei kleineren Rechnern einzugehen. Dort sind Lösungen zu suchen, die denen äquivalent sind, die für Großrechner gewählt wurden. Zwar bietet das auf PCs häufig verwendete Betriebssystem MS-DOS nur beschränkte Möglichkeiten. Die meisten der nachfolgend genannten Maßnahmen sind jedoch über den Zukauf von spezieller Sicherheitssoftware ohne Schwierigkeiten zu realisieren.[7] Mit dieser Software kann u. a. die Ver-

[7] Zu der Problematik der Datensicherung auf PCs, vgl. Becker (1987). Datensicherungsmaßnahmen auf Personalcomputern, die bei einer Rechnerkoppelung vorgesehen werden könnten, werden im 15. Tätigkeitsbericht des Hessischen Datenschutzbeauftragten (1986:145ff.) behandelt.

schlüsselung von Daten, die Einrichtung von Paßwortabfragen und das 'Verstecken' von Dateien, so daß sie das Betriebssystem nicht 'findet', realisiert werden. Entsprechende Fachliteratur ist bereits verfügbar (vgl. z.B. Aleanakian/Kühnau 1986) und Fachzeitschriften beschäftigen sich mit dem Problem (vgl. z.B. PC Welt 1/89).

1) Eindeutige Zuordnung der Daten zu berechtigten Benutzeridentifikatoren:

Für alle Datenbestände der hier wichtigen Art muß gesichert sein, daß sie bestimmten Benutzerkennungen eindeutig zugeordnet sind. Dies ist eine der Minimalanforderungen an die Verarbeitung faktisch anonymer amtlicher Mikrodaten.

Durch Vorkehrungen auf der Softwareebene - in der Regel über die Mittel des Betriebssystems - muß gewährleistet sein, daß jene Benutzer der Rechenanlage vom Zugriff auf die Daten ausgeschlossen sind, die nicht zu den berechtigten gehören. Personen, die berechtigt sind, mit den fraglichen Daten zu arbeiten, müssen sich in einem Identifikationsprozeß (siehe Maßnahme 2) ausweisen (vgl. Weck 1984:156ff.). Das System weist diesen Personen dann die notwendigen Betriebsmittel und Daten zu und "autorisiert" sie, mit ihnen gemäß ihren jeweiligen Zwecken zu verfahren.

Die technische Möglichkeit zu dieser Maßnahme ist standardmäßig in jedem modernen Betriebssystem für Großrechner vorhanden. Bei manchen Rechnertypen (z. B. VAX/VMS) ist sogar von vornherein die Angabe einer Hierarchie von Zugriffsberechtigungen möglich. Dann wird der unmittelbaren Benutzeridentifikation zusätzlich eine Projektnummer zugeordnet. Die Wirkung dieser Zuweisung ist, daß allen Benutzern mit der gleichen Projektnummer eine Zugriffsberechtigung auf die entsprechend definierten Projektdaten erteilt wird. Auf diese Weise wird gesichert, daß nur solche Benutzer Befugnisse über einen Datenbestand erhalten, die diesen für ihre Forschungen benötigen. Das Betriebssystem MS-DOS sieht die Möglichkeit der Zuordnung zu Benutzeridentifikatoren standardmäßig nicht vor. In diesem Fall ist mit spezieller Sicherheitssoftware Abhilfe zu schaffen.

2) Kontrolle der Benutzeridentität:

Notwendige Voraussetzung jeder Beschränkung des Datenzugangs auf berechtigte Personen sind Techniken der Authentisierung, d.h. Prozeduren,

über die die Identität der Benutzer festgestellt werden kann.[8] Auf der Ebene der Hard- und Software können verschiedene Maßnahmen ergriffen werden, die durch organisatorische Vorkehrungen (vgl. 9.) zu ergänzen sind.

Eine wichtige Form der Identifikation besteht in der Verwendung von Paßwörtern. Der Benutzer muß sich dabei jeweils bei der Anmeldung eines Auftrags an den Rechner durch die Eingabe eines geheimzuhaltenden Code-Wortes ausweisen. Dieses ist zur Erhöhung der Sicherheit in bestimmten Zeitabständen zu ändern. Diese Art von Schutz gehört zu den ältesten, wirksamsten und billigsten Datensicherungsmaßnahmen (Weyer/Pütter 1983: 96), die noch dazu mit den meisten Rechnern realisiert werden kann.

Die Schwachstelle des Paßwortschutzes besteht in der Notwendigkeit, das Wort selbst geheimzuhalten. Dazu können organisatorische Maßnahmen getroffen werden, die vor allem sichern sollen, daß es nicht von einem potentiellen Datenangreifer erraten werden kann.

3) Sicherung des selektiven Zugriffs auf der Ebene der Datenbestände:

Wiederum als standardmäßig zu realisierende Maßnahme ist vorzusehen, daß Dateien nur von berechtigten Personen gelesen oder verändert werden können. Dies kann durch die Vergabe von Dateipaßwörtern geschehen, die nur den berechtigten Benutzern mitgeteilt werden, oder durch die Angabe entsprechender Kompetenzen in einer sogenannten Autorisierungsmatrix.[9]

Schutzvorkehrungen auf der Dateiebene haben folgenden Sinn:

- Zum einen stellen sie eine zusätzliche Hürde dar. Selbst wenn sich ein unterstellter Datenangreifer die Benutzernummer beschaffen könnte, unter

---

8 Vgl. Dammann (1981a:332ff.), Weck (1984:166ff.), Reuter (1987:344 und 369).

9 Dabei wird vorausgesetzt, daß die Benutzer auch auf der Ebene der in einem Rechner ablaufenden "Prozesse" (auch "Aufträge" genannt) klar getrennt werden. Dies wird jedoch von den entsprechenden Komponenten des Betriebssystems normalerweise ohnehin gewährleistet, braucht also nicht als gesonderte Forderung nach Datensicherheit berücksichtigt werden.
Moderne Großrechner garantieren durch einen hardwaregestützten Umsetzungsmechanismus, daß unter den vielen Benutzern, die sich die Anlage im Wege des "Time-sharing" teilen, nicht einer wissentlich oder unwissentlich die Daten eines anderen Benutzers, die sich im sogenannten virtuellen Adreßraum befinden, ansprechen kann (vgl. Reuter 1987:370, Calingaert 1983:109ff. und Weck 1984:110ff.).

der amtliche Daten verarbeitet werden und er den Identifikationsprozeß des Rechners entsprechend besteht, so fehlte ihm dennoch eine notwendige Voraussetzung für den Datenzugriff.

- Zum anderen kann unter dem Bereich der berechtigten Benutzer genauer differenziert werden, indem bestimmten Personen nur ganz bestimmte Teildatenbestände zugänglich gemacht werden, die sie für ihre Arbeit unmittelbar benötigen.

Die mögliche Unterscheidung zwischen der Berechtigung zum Lese- und zum Schreibzugriff ist allerdings bei anonymen Daten von geringerer Bedeutung als bei personenbezogenen Daten. Wenn die Berechtigung zur Veränderung der Daten nur den Personen erteilt wird, deren Aufgabe es ist, mit ihnen zu arbeiten, so dient dies bei anonymen Daten in erster Linie dem Interesse der datenverarbeitenden Stelle am Schutz vor Zerstörung oder Verfälschung der Daten. Bei personenbezogenen Informationen muß hingegen stets zusätzlich berücksichtigt werden, daß eine Veränderung der Daten auch die Interessen der Personen tangieren kann, über die Informationen gespeichert sind. Dies ist ein wichtiger Unterschied mit erheblichen Konsequenzen für die Einrichtung von Schutzvorkehrungen: Bei personenbezogenen sind gegenüber anonymen Daten zusätzliche Maßnahmen vorzusehen.

Von den beiden Techniken, Schutzvorkehrungen auf der Dateiebene einzurichten, ist die Vergabe von Dateipaßwörtern die z.Zt. häufiger verwendete Methode. Sie erlaubt im Normalfall eine Unterscheidung zwischen dem Recht, Datenbestände nur zu lesen und dem Recht, sie auch zu verändern. in der Regel ist nämlich die Zuordnung verschiedener Paßwörter für das Schreiben und das Lesen der Daten vom bzw. auf das betreffende Speichermedium möglich. Die meisten modernen Betriebssysteme unterstützen die Einrichtung dieser Zugangsschranke, während Paßwortalgorithmen, die auch hier eingesetzt werden können, über spezielle Software einzurichten wären.

Manche Betriebssysteme (z.B. VMS, aber nicht BS2000) erlauben die Angabe einer sogenannten Zugriffs- oder Autorisierungsmatrix. Dies ist die zweite wichtige Grundform des Dateischutzes. In die Zeilen dieser Tabelle werden die Rechte der Benutzer eingetragen und in die Spalten die Dateien (vgl. Weck 1984:205ff.). Damit ist die unmittelbare Zuordnung von Dateien auf fremde Benutzernummern möglich. Jeder Verwalter eines Datenbestandes kann anderen Benutzern die Berechtigung für einen Lese- oder Schreibzugriff zuteilen. Die Einrichtung derartiger Zugriffsmatrizen kann die Vergabe von

Dateipaßwörtern ersetzen, da zwischen den berechtigten und den nicht berechtigten Benutzern unterschieden werden kann.

Moderne Datenbanksysteme (z.B. ORACLE, SIR) bieten ebenfalls die Möglichkeit, derartige Zugriffsbefugnistabellen einzurichten. Werden die Daten in einer solchen Datenbank gehalten, ergeben sich also weitere Schutzmöglichkeiten. Allerdings muß bei Wahl dieser Lösung gewährleistet werden, daß die Barrieren nicht auf der Betriebssystemebene umgangen werden können. Die Datenbank selbst muß also zusätzlich gesichert sein.

4) Sicherungen gegen den Zugang von externen Rechenanlagen:

Ein wichtiges Problem der Datensicherung besteht in der vielfältigen Vernetzung der Rechner (vgl. Steinmüller 1987 und 1988b, Schönberg 1984:95f.). Dadurch entstehen zusätzliche Möglichkeiten für unzulässige Datenzugriffe. Werden Benutzer externer Rechenanlagen in gleicher Weise zugelassen, wie Personen, die mit dem lokalen Computer arbeiten, entstehen Sicherheitsprobleme, für die von der Informatik sehr heterogene Lösungswege angeboten werden (vgl. Weck 1984:270ff.).

Allerdings bietet der Umfang der Daten im Falle amtlicher Stichproben bereits einen beträchtlichen Schutz: Datenübertragungen mittels Fernleitungen dauern wesentlich länger als Ein- und Ausgabe- oder Kopiervorgänge an einem Rechner. Der Transfer eines kompletten Datenbestandes über eine Leitung mit typischer Übertragungsgeschwindigkeit würde mehrere Stunden dauern, wenn ein Benutzer eine solche Leitung für sich hätte. Diese müssen sich jedoch viele Benutzer teilen. Darum erscheint es als unrealistisch, daß jemand über ein Netz einen Mikrozensus 'stiehlt'.

Wenn Einzeldaten aus amtlichen Statistiken jeweils am lokalen Rechner verarbeitet werden, sind die gegebenen technischen Möglichkeiten einzusetzen, um die Daten vor den Benutzern anderer Rechenanlagen zu schützen. Unter BS2000 ist es möglich, bestimmte Dateien für Rechenaufträge zu sperren, die von anderen Rechnern gestartet wurden.

Ist ein Zugriff zu amtlichen Mikrodaten über ein Netz unumgänglich, so sollten zusätzliche Sicherheitsvorkehrungen eingerichtet werden. Diese können in weitergehenden Anonymisierungs- oder Datensicherungsmaßnahmen bestehen.

5) Spezielle Sicherungsmaßnahmen für Datenträger:

In manchen Betriebssystemen (z.B. BS2000, Unix) kann für spezielle Datenträger (Magnetplatten) ein eigener Benutzerkatalog definiert werden, so daß alle Nutzer, die nicht in diesem enthalten sind, vom Zugriff ausgeschlossen werden. In diesem Fall werden alle Dateien, die auf den betreffenden Datenträgern[10] abgespeichert sind, zusätzlich geschützt.

Solche im Software-Bereich lokalisierte Vorkehrungen sind jedoch dann nicht hinreichend, wenn der Zugang zu den Datenträgern auf der physikalischen Ebene nicht ausreichend gesichert ist. Darum ist als Standardmaßnahme dafür zu sorgen, daß Unbefugte keinen Zugang zu den betreffenden Magnetbändern bzw. den Geräten erhalten, die die Daten aufnehmen.

Datenbänder müssen für die Öffentlichkeit unzugänglich aufbewahrt werden. Ist der Rechenbetrieb personell und organisatorisch von der eigentlichen Forschung getrennt, kann durch geeignete Maßnahmen (schriftliche Bestätigung, persönlicher Kontakt) sichergestellt werden, daß nur die Berechtigten die Verarbeitung einer Datei mit Einzelangaben aus der amtlichen Statistik auf einem Magnetband veranlassen können (vgl. Weyer/Pütter 1983: 50, 63ff.).

Eine weitergehende Lösung sieht die räumliche Separierung der Bandgeräte vor, über die die Daten in die Rechenanlage einzulesen sind. Diese Möglichkeit könnte u.U. dann gewählt werden, wenn in einem Rechenzentrum nicht für die vor einem äußeren Zugriff sichere Lagerung der Magnetbänder garantiert werden kann oder wenn der Einlesevorgang selbst aus einem ähnlichen Grund problematisch ist.

Derartige Maßnahmen sind u.U. für die betreffenden Stellen mit einigem Aufwand und Problemen im organisatorischen Ablauf verbunden.[11]

Im Falle der Datenverarbeitung auf PCs sind spezielle Sicherungsmaßnahmen für die verwendeten Disketten nötig. Es muß auf jeden Fall gesichert sein, daß kein Benutzer Daten unbemerkt kopieren und auf Disketten mitnehmen kann.

---

10 Generell sind die Sicherungsmöglichkeiten von Daten, die sich auf externen Speichermedien befinden, in den letzten Jahren ständig verbessert worden (vgl. bereits Miehle/Tubies 1980).

11 Zusätzlich ergeben sich technische Probleme, da die Kapazität vieler Leitungen begrenzt ist.

Um dies auszuschließen, ist z. B. denkbar, daß die Diskettenlaufwerke generell entfernt werden.

6) Datenzugang über dedizierte Terminals (vgl. Heinz 1987):

Um die Zugangskontrolle zu effektivieren und um zu verhindern, daß sich jemand unter Umgehung der Sicherheitsvorkehrungen einen Datenzugang verschafft, kann vorgesehen werden, daß die zu schützenden Dateien nur von speziellen Terminals aus angesprochen werden können, die durch geeignete organisatorische Maßnahmen und räumliche Voraussetzungen zu schützen sind. Wenn die Geräte in einem sonst verschlossenen Raum stehen, zu dem nur berechtigte Personen Zugang und Schlüssel haben, wird eine Kontrolle ausgeübt, deren Wirksamkeit auf verschiedene andere, bereits genannte Schutzmaßnahmen aufbaut. Die Verkoppelung der Daten mit speziellen Terminals geschieht nämlich im Normalfall dadurch, daß eine bestimmte Benutzernummer, von der aus die Daten allein angesprochen werden können, nur über diese dedizierten Terminals angemeldet werden können.

In diesem Fall ist der unbemerkte Datenzugriff durch eine fremde Person, die sich die relevanten Paßwörter hat beschaffen können, nicht mehr möglich. Die Verwendung dedizierter Terminals erschwert, daß Endgeräte angesteuert werden, die einen Abtransport der Daten ermöglichen könnten. In diesem Zusammenhang ist vor allem an die Kopplung des Mainframe-Rechners, auf dem die Daten gehalten werden, mit ungeschützten und nicht überwachten Personal-Computern zu denken. Diese Kopplung könnte ermöglichen, daß Daten auf Disketten übertragen und ungehindert und unregistriert entfernt werden.

Dedizierte Terminals verursachen erhebliche zusätzliche Kosten, da Räume und Geräte benötigt werden, die ausschließlich den berechtigten Personen offenstehen dürfen.

7) Die Verschlüsselung der Daten (Kryptographie):[12]

Diese bei kleineren Datenbeständen gelegentlich angewendete Maßnahme (vgl. Heinz 1987) verhindert, daß Informationen, die sich jemand durch einen

12 Vgl. zum folgenden Weck (1984:283ff.), Carroll (1987:181ff.), Stadler (1980:95ff.), Rihaczek (1980) und (1987b).

Rechtsbruch und unter Umgehung der sonstigen Sicherheitsvorkehrungen verschafft hat, auch genutzt werden können. Der Datenangreifer erhielte lediglich Daten, die nur über einen speziellen Software-Schlüssel gelesen werden können.

Die Verschlüsselungsmethoden sind grundsätzlich die gleichen wie sie für geschriebene Texte schon seit langer Zeit bekannt sind: Über ein bestimmtes Schema, das nur den 'Eingeweihten' bekannt oder zugänglich ist, werden den Zeichen (Buchstaben, Zahlen) eines Klartextes beliebige andere Zeichen zugeordnet (vgl. Waidner et al. 1987).

Allerdings ist diese Maßnahme dann nur von begrenzter Wirksamkeit, wenn die verschlüsselten Daten und der Schlüssel auf dem gleichen Rechner gehalten werden. Unter diesen Bedingungen ist vorstellbar, daß sich der Angreifer auch den Schlüssel verschaffen kann, wenn er schon über einen Zugang zu den Daten verfügt. Demgemäß werden kryptographische Methoden in erster Linie bei Datenübertragungen angewendet, wo ein potentieller Datenangreifer relativ leicht eine Leitung abhören kann, der nicht übertragene Schlüssel sich jedoch im ausschließlichen Besitz des Senders und des Empfängers befindet (vgl. Rihaczek 1980:100).[13] Lediglich abgespeicherte Datenbestände werden durch eine Verschlüsselung nur insofern zusätzlich geschützt, als sich ein potentieller Datenangreifer zwei Paßworte statt einem beschaffen müßte, um den auf der gleichen Rechenanlage wie die Daten gespeicherten Schlüssel zu deren Dechiffrierung benutzen zu können.[14]

Die laufenden Kosten einer Datenverarbeitung mit verschlüsselten Einzeldaten der amtlichen Statistik sind beträchtlich, da die hierbei anfallenden Datenmengen, die bei Auswertungen entschlüsselt werden müssen, die Rechenzeiten stark in die Höhe treiben. Aus diesem Grund und wegen ihrer begrenzten Wirksamkeit wird die Sicherungsmaßnahme Kryptographie nur im Ausnahmefall im Zusammenhang von amtlichen Daten zu empfehlen sein. Sobald diese keinen Engpaß darstellt, sind kryptographische Methoden generell zu empfehlen. Die in jüngster Zeit entwickelten Methoden der Verschlüsselung

---

13 Die Methoden der Kryptographie sind in jüngerer Zeit stark verbessert worden (vgl. Rihaczek 1987b:302, Spies 1985), so daß vielleicht über diesen Weg eine Lösung der Schutzprobleme bei Datenübertragungen erwartet werden kann.

14 Neuentwicklungen werden in Zukunft möglicherweise Abhilfe bringen. Dabei ist besonders an auf Chipkarten gespeicherte Schlüssel zu denken (vgl. Rihaczek 1987b).

sind derart wirksam, daß nicht zu befürchten ist, daß die gesteigerte Leistung der Computer dazu dienen könnte, den Code zu 'knacken'.

8) Löschung nicht mehr benötigter Datenbestände:

Nach dem Auslaufen eines Forschungsvorhabens, für das amtliche Daten übergeben wurden, müssen diese gelöscht werden. Dies ist als obligatorische Maßnahme bereits im Bundesstatistikgesetz vorgeschrieben.

Darüber hinaus müssen im Rahmen des üblichen Ablaufs eines Projekts immer wieder nicht mehr benötigte Datenbestände gelöscht werden, damit den Anforderungen der Datensicherheit entsprochen wird. Oftmals werden Sicherungskopien von Dateien angefertigt, bestimmte Datenbestände selektiert und kopiert, die dann auf separaten Dateien verfügbar sind. Werden diese nicht mehr gebraucht, dürfen die betreffenden Datenträger - z.B. Magnetbänder - nicht einfach freigegeben werden. Eine rein logische 'Löschung', bei der lediglich der Katalogeintrag aus dem Verzeichnis der Dateien entfernt wird, die Information durch geeignete Programme jedoch wieder zugänglich gemacht werden kann,[15] reicht nicht aus (vgl. auch Dorer/Mainusch/Tubies 1988:107).

Die auf den betreffenden Datenträgern enthaltene Information muß vielmehr physikalisch vernichtet, d. h. mit binären Nullen überschrieben werden. Die dafür zu verwendenden Programme sind leicht verfügbar, so daß diese Maßnahme weder Aufwand noch Mittel, sondern nur etwas Sorgfalt beansprucht.

9) Abwicklung der Datenverarbeitung als "Closed-Shop"-Betrieb:

Sofern nicht bereits ohnehin dafür gesorgt wurde, ist die Gewährleistung der räumlichen und personellen Abschottung des jeweiligen Rechenzentrums eine unbedingt zu erledigende Aufgabe der Datensicherung.[16]

Ist das technische Personal für den laufenden Betrieb allein zuständig, wird verhindert, daß sich ein potentieller "Datenangreifer" an der Hardware zu schaffen macht oder unmittelbar am Rechner Datenträger manipuliert. Dies

---

15 Diese Situation ist z.B. bei PCs anzutreffen, wo zahlreiche Softwarewerkzeuge existieren, die den Zweck haben, gelöschte Dateien wieder zugänglich zu machen (vgl.: Der Hessische Datenschutzbeauftragte 1986:143f.).

16 Vgl. Hailer (1980:158f.), Carroll (1987:98ff.), Droux (1984:288ff.), Auernhammer (1981:100).

setzt allerdings voraus, daß die Rechner in separaten Räumen stehen, zu denen Außenstehende nicht zugelassen sind. Zugangs- und Abgangskontrollen von Datenträgern werden auf diese Weise erleichtert.

Durch verschlossene Türen, gegebenenfalls Ausweiskontrollen etc., ist für die Separierung der Datenverarbeitung zu sorgen. Unbefugte können dadurch vom Datenzugriff auf der physikalischen Ebene abgehalten werden. Damit kann gewährleistet werden, daß sich nicht jemand die Kompetenzen eines Systemverwalters anmaßt und die mühsam mit großem Aufwand aufgebauten Kontrollen und Barrieren umgeht.

Die organisatorische und räumliche Abtrennung erfordert weiterhin feste Vereinbarungen zwischen der Forschung und dem jeweiligen Rechenzentrum. Werden solche hinsichtlich Datensicherheitsvorkehrungen getroffen, ist die nachträgliche Verletzung dieser Regeln nicht mehr möglich, ohne daß mehrere Personen verschiedener Teilinstitutionen eingeschaltet werden. Dadurch wird die Überwachung der Einhaltung dieser Regeln verbessert.

Bei der Verwendung von Personal Computern oder auf Workstations erfolgt die Datenverarbeitung in der Regel nicht auf einem Rechner, der getrennt vom eigentlichen Benutzer aufgestellt ist und von zusätzlichem technischen Personal überwacht wird. Dies ist das schwierigste Problem, das sich unter Datensicherungsgesichtspunkten im Zusammenhang der Verwendung von Arbeitsplatzrechnern stellt. In diesem Fall wird ein Ausgleich durch besondere organisatorische Kontrollen und Vorkehrungen sowie mittels extensiver Verwendung von Sicherheitssoftware einschließlich spezieller Maßnahmen zur Protokollierung (siehe unten) zu schaffen sein. Dabei ist auch an Eigenentwicklungen und an Veränderungen der Hardware zu denken.[17]

12.3 Sicherungsmaßnahmen gegenüber berechtigten Personen

Neben den Maßnahmen, die dem Schutz vor den Zugriffen externer "Datenangreifer" dienen, sind Kontrollen und Sicherungsvorkehrungen gegenüber den Personen vorzusehen, die berechtigt sind, die Daten zu nutzen und zu verwenden. Auf diese Weise kann der Aufwand unterstellter Deanonymisie-

[17] Heinz (o. J.) diskutiert ein spezielles System für Standalone Rechner (vgl. auch ders. 1987 und den 15. Tätigkeitsbericht des Hessischen Datenschutzbeauftragten 1986:145ff.).

rungsversuche seitens dieser Personen erhöht werden. Derartige Maßnahmen dienen gleichzeitig der in §16 Abs.8 Bundesstatistikgesetz geforderten Zweckbindung der Verwendung amtlicher Mikrodaten, insofern unzulässige Operationen erschwert oder verhindert werden.

Allerdings besteht ein zentraler Unterschied zu den Vorkehrungen, die sich gegen externe Personen richten. "Wie bei den meisten Betrachtungen über die Informationssicherheit muß man feststellen, daß der Schutz eines Systems nur bis zu einer gewissen Grenze möglich ist. Die eigentliche Schwachstelle ist der Mensch, der über die Schutzmaßnahmen Bescheid weiß oder Teil von ihnen ist" (Kauffels 1985:64, Kerkau 1985:88). Während die Schutzmaßnahmen gegenüber externen Datenangreifern zwar nicht völlig unüberwindbar sind, aber doch eine sehr hohe Hürde darstellen, ist ein ähnlicher Grad an Wirksamkeit gegenüber Insidern nicht möglich.

Leitlinie der hier diskutierten Vorschläge zur Datensicherung ist jedoch nicht das Kriterium der absoluten Sicherheit. Der im vorliegenden Zusammenhang erhobenen Forderung, die Kosten einer Deanonymisierung zu erhöhen, wird sehr wohl entsprochen.[18]

Im einzelnen werden folgende Maßnahmen vorgeschlagen:

10) Verpflichtung der berechtigten Personen auf das Datengeheimnis (vgl. Weyer/Pütter 1983):

Die unmittelbare Notwendigkeit einer solchen Verpflichtung aller nicht unmittelbar zum öffentlichen Dienst zu zählenden Personen ergibt sich schon aus §16 Abs.7 Bundesstatistikgesetz (vgl. Dorer/Mainusch/Tubies 1988:105f.). Bei Verstößen gegen die in der Verpflichtung vorgegebenen Verhaltensnormen drohen wesentlich härtere Strafen als dies ansonsten der Fall wäre. Erfolgt eine Verurteilung nach den §§ 202 (unbefugtes Ausspähen von Daten), 203 (Verletzung von Privatgeheimnissen) oder 204 (Verwertung der so erlangten Angaben) Strafgesetzbuch, so droht eine Freiheitsstrafe von bis zu einem bzw. bis zu drei Jahren.

---

18 Die Forderung von Rihaczek, daß sichere Systeme selbst "Mißbrauch vom ordentlichen Gebrauch unterscheiden und wohlunterschieden reagieren können" (1985:233) erscheint wenigstens heute als noch nicht einmal näherungsweise realisierbar - wenigstens solange derart komplexe Probleme wie statistische Auswertungen zur Diskussion stehen - da die verwendeten Rechner bei weitem noch nicht 'intelligent' genug sind.

In jedem Falle ist die vorgeschriebene Verpflichtung der mit den Daten arbeitenden Personen eine zentrale Schutzmaßnahme. Sie wird durch weitere Vorkehrungen technischer und organisatorischer Art ergänzt und verstärkt.

11) Protokollierung der mit den Daten durchgeführten Operationen. Die Realisierung kann auf zwei verschiedene Weisen erfolgen:

- Zum einen kann verlangt werden, daß die berechtigten Personen über alle Datenzugriffe schriftliche Aufzeichnungen anfertigen.

- Zum anderen können Aufzeichnungen von der datenverarbeitenden Anlage selbst, also automatisch, angefertigt werden.

Im ersten Fall wird nur eine globale Protokollierung ohne exakte Auflistung der durchgeführten Operationen möglich sein. Die Leistungsfähigkeit einer modernen EDV-Anlage ist so beträchtlich, daß bei detaillierter manueller Buchführung die Personen, die für die Konzeptionierung und Konkretisierung der Programme verantwortlich sind, einen großen Teil ihrer Arbeitszeit für die Aufgabe der Dokumentation verwenden müßten. Die manuelle Aufzeichnung der Zwecke, Ziele und der wichtigsten Operationen, die bei einem konkreten Datenzugriff durchgeführt werden, erscheint jedoch als realisierbare Möglichkeit. Eine derartige Protokollierung könnte stets zeigen, mit welchen Absichten die Daten verarbeitet werden.[19]

Dieses Verfahren könnte durch bestimmte Vorkehrungen einer automatischen Aufzeichnung ohne Schwierigkeiten unterstützt werden. So kann jeder Zugriff auf den zu schützenden Datenbestand automatisch aufgezeichnet und zusammen mit einigen wichtigen Charakteristika des betreffenden Auftrags ("jobs") an die DV-Anlage automatisch abgespeichert werden.[20] Solche Kennzeichen könnten z.B. Angaben zu den verbrauchten Betriebsmitteln (Rechenzeit etc.) sein.

---

19 Die manuelle Protokollierung könnte durch ein geeignetes Programm unterstützt werden, das z.B. stets aufgerufen wird, wenn ein zu schützender Datenbestand angesprochen wird. In diesem Fall werden automatisch bestimmte Eingaben angefordert, die die Absicht des Benutzers näher charakterisieren, seine Datenzugriffe dokumentieren etc. (vgl. Aleanakian/Kühnau 1986:82).

Der Nachteil einer solchen Lösung ist, daß ihr nur eine eingeschränkte Beweiskraft hinsichtlich der Vorschriftsmäßigkeit der mit den Daten vollzogenen Operationen zukommt. Da die Aufzeichnungen von demjenigen selbst angefertigt werden, der mit den Daten arbeitet, kann er sie so gestalten, daß sie Nachprüfungen standhalten, auch wenn er sich in irgendeiner Weise unkorrekt verhalten hat. Allerdings setzt diese Vorgehensweise die Absicht der bewußten Täuschung voraus. Insofern man im Normalfall nicht von einer solchen ausgehen kann, bedeutet die Maßnahme der obligatorischen Protokollierung einen zusätzlichen Schutz. Eine weitere Wirkung wird gegen externe "Datenangreifer" erzielt, da beim Zugriff auf ein geschütztes Magnetband die zuständigen Stellen über die automatischen Meldungen erfahren, daß jemand einen Versuch unternommen hat, Daten zu lesen, der dazu nicht berechtigt war.

Dieser Schutz könnte noch intensiviert werden, wenn die automatische Aufzeichnung ausführlicher gestaltet wird. Bei Großrechnern ist realisierbar, daß alle an den Rechner gegebenen Steueranweisungen protokolliert und diese Information über längere Zeit gespeichert oder in ausgedruckter Form verfügbar bleibt. Leider ist diese Maßnahme mit Problemen verbunden. Typischerweise werden im Rahmen sozialwissenschaftlicher Untersuchungen ganze Serien von Auswertungen mit oft nur leicht veränderten Konstellationen von Steueranweisungen an den Rechner durchgeführt. Die dabei produzierten Mengen an Protokolltexten werden rasch sehr unübersichtlich. Wenn also der Zeitpunkt eines unerlaubten "Angriffs" zur Deanonymisierung nicht genauer lokalisiert werden kann, etwa weil er bereits längere Zeit zurückliegt, wird es schwierig sein, ihn in den Protokollen aufzufinden.

Bestimmte, für kommerzielle Zwecke genutzte Software erfüllt die Aufgabe der lückenlosen Aufzeichnung bereits. Im kaufmännischen Bereich verwendete Methoden der Verbuchung sind jedoch gerade zum Zweck der Dokumentation aller Operationen über einen langen Zeitraum entwickelt worden. Die Struktur der Operationen, die aufgezeichnet werden sollen, ist darüberhinaus viel einfacher als bei statistischen Auswertungen. Spezielle Verfahren zur Dokumentation sind erst in der Entwicklung.[21]

---

[20] Diese Form der Protokollierung kann von einem Datenbanksystem unterstützt werden.

[21] Das Datenschutzpaket, das von der Projektgruppe um Heinz für die Sicherung kriminologischer Daten entwickelt wird, soll auch eine lückenlose Protokollierung ermöglichen, wird jedoch stark maschinenspezifisch sein. Es ist für IBM-Rechner konzipiert.

Doch selbst bei Realisierung eines Dokumentationssystems ist das Ziel der Nachprüfbarkeit aller Datenzugriffe noch nicht erreicht. Bei den wichtigsten Betriebssystemen kann die automatische Aufzeichnung durch den Benutzer umgangen werden: Die Protokollierung ist zumeist abschaltbar. Außerdem können Protokollisten nachträglich verändert werden. Dies ist allein bei den Funktionen nicht möglich, über die standardmäßig von dem betreffenden Rechenzentrum Buch geführt wird (z.B. über die verbrauchte Rechenzeit zur Abrechnung der Betriebskosten).

Systeme, in denen die Protokollierung wirklich vom Benutzer nicht beeinflußbar ist, sind konstruierbar,[22] der Entwicklungsaufwand ist jedoch beträchtlich. Er müßte für die wichtigsten Rechnertypen jedenfalls mit mehreren Jahren veranschlagt werden, da die im sozialwissenschaftlichen Bereich benötigten Programmpakete zur statistischen Auswertung und Datenverwaltung erst in diesen Rahmen eingepaßt werden müßten.

Eine automatische Protokollierung aller Rechenoperationen würde unter den meisten Umständen den Aufwand für sozialwissenschaftliche Analysen in unverhältnismäßiger Weise in die Höhe treiben. Sie stellt darum auf absehbare Zeit in der Regel keine realistische Möglichkeit dar, die Datensicherheit im vorliegenden Kontext zu erhöhen. Sie ist nur beim Vorliegen besonderer Bedingungen zu fordern, nämlich dann, wenn ein Sicherheitssystem ansonsten wesentliche Lücken aufweisen würde.

12) Klare Kompetenzzuweisungen bei der Organisation der Datenverarbeitung:

Diese zentral wichtige Maßnahme verbessert aus wenigstens drei Gründen die Datensicherheit:

Verschiedene auf der "SoftStat 89" (einer Konferenz von ZUMA über Statistik-Software) vorgestellte Ansätze einer automatischen Protokollierung von statistischen Operationen, die über eine reine Auflistung hinausgehen, dienen in erster Linie der Orientierung des Benutzers selbst. Systeme, die einen 'intelligenten' Überblick über den Ablauf geben, sind jedoch noch nicht ausgereift (vgl. den Vortrag von Fröschl). Eine Verwendung für den hier wichtigen Zweck ist vorerst nicht denkbar.

22 Das von der Projektgruppe um Heinz aufgebaute System wird diese Eigenschaften haben.

- Die Realisierung von Schutzmaßnahmen setzt voraus, daß die Abwicklung der Datenverarbeitung übersichtlich, d. h. geordnet, erfolgt. Andernfalls besteht die Gefahr, daß Lücken im System das Vorhaben eines potentiellen "Angreifers" erleichtern. Nur bei entsprechender Transparenz ist eine Überprüfung der Datensicherheit möglich (Steinmüller 1980b:235).

- Sind die Aufgaben der einzelnen Mitarbeiter und ihre Zugangsrechte zu faktisch anonymen Datenbeständen genau definiert, fällt in einem Forschungsteam auf, wenn sich ein Mitarbeiter unzulässig verhält, d. h. die Daten zu anderen als den vorgeschriebenen Zwecken oder gar zu Deanonymisierungsexperimenten verwendet.

- Nur bei klarer Aufgabendefinition können die Zugangsrechte auf Datenbestände der amtlichen Statistik beschränkt werden, die zur Erledigung der dienstlichen Aufgaben und persönlichen Forschungsinteressen benötigt werden. Wenn an ein Institut mehrere Erhebungen weitergegeben wurden, wird die Zugriffsberechtigung nicht pauschal für den ganzen Bestand zu erteilen sein. Vielmehr werden die einzelnen Personen nur jene Paßwörter erhalten, die Zugang zu den Dateien eröffnen, die für die aktuelle Arbeit wichtig sind.

Durch die Festlegung der Aufgaben in der Gruppe braucht die Freiheit der Forschung nicht eingeschränkt zu werden. Verlangt wird lediglich eine gewisse Offenheit gegenüber den Kollegen und die explizite Festlegung der Regeln bei der Kooperation. Der geforderten Transparenz dient, daß die wichtigsten Organisationsprinzipien schriftlich festgehalten werden (Weyer/Pütter 1983:47).

Das Prinzip der Klarheit von Organisation und Kompetenzzuweisung setzt schließlich voraus, daß ein Mitarbeiter der betreffenden Institution für die Datensicherheit eindeutig zuständig und verantwortlich ist (Steinmüller 1980b:235).

13) Programmdokumentation:

Wenn für jedes Programm genau dokumentiert wird, welchem Zweck es dient, können Verstöße gegen die Datensicherheit bzw. das Deanonymisierungsverbot leichter festgestellt werden. Für einen Forscher, der sich nicht an die definierten Regeln hält, steigt der Aufwand, sich vor Entdeckung zu schützen.

Insofern allein Standardsoftware verwendet wird, kann sich die Dokumentation der Programme auf eine Kommentierung der verwendeten Stapelprozesse beschränken. Wird hingegen in größerem Maßstab eigene Software - z.B. für spezielle statistische Auswertungsverfahren - entwickelt, sind die üblichen Regeln für ihre Dokumentation einzuhalten. Diese sind ein allgemeines Erfordernis jeder Programmierung und nicht nur für den Schutz von Mikrodaten wichtig. "Eine vollständige und aussagefähige Programmdokumentation sollte wenigstens folgende Bestandteile enthalten:

- Programminhalt und Programmgliederung
- Anwendungsbereich
- Datenflußplan
- Programmablaufplan
- Datei- und Satzaufbau
- Ein- und Ausgabe
- Programmliste
- Bedienungsanleitung
- Protokolle des abschließenden Anwendertests und der Freigabe
- Testdaten
- Name des verantwortlichen Programmierers"

(Weyer/Pütter 1983:48).

Ein derart dokumentiertes Programm erlaubt einem sachverständigen Dritten abzuschätzen, welche Operationen mit ihm durchgeführt werden können. Dadurch wird die Transparenz der Abläufe erhöht. Der Aufwand für jemand, der abweichende Interessen verfolgt, steigt, weil er eher befürchten muß, bei seinem Vorhaben entdeckt zu werden.

12.4 Eine 'Standardliste' an Sicherungsmaßnahmen

Die dargestellten Maßnahmen der Datensicherung wurden gemäß Aufwand und Schutzwirkung beurteilt und insbesondere danach eingestuft, ob ihre Anwendung bei der Weitergabe von amtlichen Daten als erforderlich angesehen werden sollte. Wie erwähnt, können solche Beurteilungen nie abschließenden Charakter haben, da die verwendeten Rechnertypen und die lokalen Bedingungen ihres Einsatzes jeweils zu unterschiedlich sind. Außerdem sind oft Maßnahmen vorstellbar bzw. werden Techniken neu entwickelt, die sich von den angegebenen unterscheiden, aber in der Wirkung äquivalent sind bzw. sogar Verbesserungen bedeuten.

Aus diesen Gründen dient die Zusammenstellung der für notwendig betrachteten Maßnahmen nur als Orientierungshilfe und nicht als Katalog, an dem starr festzuhalten wäre. Die im folgenden dargelegte Liste zeigt das Spektrum der anzuwendenden Methoden und das Niveau an Sicherheit, das anzustreben ist. Umgekehrt gibt sie Aufschluß über die Risiken, die zu berücksichtigen sind. Kann eine bestimmte Maßnahme von einem Institut unter den lokal gegebenen Umständen nicht realisiert werden, so muß eine andere getroffen werden, die sie ersetzt.

Jene Maßnahmen der Liste, die zur standardmäßigen Verwirklichung vorgesehen sind, ergänzen sich. So nützen Schutzpaßwörter für Dateien nichts, wenn die illegale Aneignung der Datenträger möglich ist. Der Katalog ergibt deshalb ein Muster für eine Systemlösung, die organisatorische und technische Maßnahmen umfaßt.

Damit entspricht der Katalog den gestellten Anforderungen. Den Vorschriften von §16 Abs.8 BStatG, die den Ausschluß aller nicht berechtigten Personen erfordern, wird entsprochen. Darüber hinaus tragen die Maßnahmen dazu bei, den Aufwand für Deanonymisierungsversuche zu erhöhen, der nach §16 Abs.6 Kriterium der faktischen Anonymität von Daten ist.

Übersicht 12.1: Standardliste der Datensicherungsmaßnahmen

Nr.	Bezeichnung	Standard-maßnahme	Bemerkung
1.	Zuordnung der Daten zu berechtigten Benutzer-identifikatoren	ja	
2.	Kontrolle der Benutzer-identität	ja	
3.	Sicherung des selektiven Zugriffs auf Dateiebene	ja	zu realisieren durch Dateipaßworte oder Dateizugriffsmatrix
4.	Sicherung gg. Zugang von extern. Rechenanlagen	ja, mit Ein-schränkung	bei Verzicht auf Verwirk-lichung Stärkung anderer Maßnahmen
5.	Sicherungsmaßnahm. für Datenträger	ja	
6.	Datenzugang über dedi-zierte Terminals	ja, mit Ein-schränkung	bei Verzicht auf Verwirk-lichung Stärkung anderer Maßnahmen
7.	Datenverschlüsselung	nein	
8.	physische Löschung nicht mehr benötigter Datenbestände	ja	Vernichtung aller Hilfs-dateien etc.
9.	Abwicklung der Datenver-arbeitung im "Closed-shop-Betrieb"	ja, mit Ein-schränkung	
10.	Verpflichtung der zugangs-berechtigten Personen auf das Datengeheimnis	ja	
11.	Protokollierung der mit den Daten durchgeführten Operationen	nein	eingeschränkte Proto-kollierung möglich
12.	klare Kompetenzzuweisung bei der Organisation der Datenverarbeitung	ja	
13.	Programmdokumentation	ja	

13 Zusammenfassende Bewertung der untersuchten Szenarien und Folgerungen für die Übermittlung faktisch anonymer Daten*

13.1 Zusammenfassung und Bewertung der Untersuchungsergebnisse

Die vorliegende Studie hat eine Reihe wichtiger Ergebnisse erbracht, die hier zusammenfassend nochmals in ihrer Bedeutung für die faktische Anonymität von Einzelangaben bewertet werden sollen.

Ganz allgemein hat sich gezeigt, daß auf der Grundlage von Experimenten mit empirischen Daten das Reidentifikationsrisiko weitaus niedriger abgeschätzt werden muß, als dies gemäß bisher durchgeführter Experimente mit ganz oder teilweise synthetisch produzierten Daten geschah.

Bei allen experimentell untersuchten Szenarien war das Kriterium der faktischen Anonymität mit Sicherheit erfüllt. Diese Szenarien waren so ausgerichtet, daß sie das im wissenschaftlichen Bereich verfügbare und beschaffbare Zusatzwissen wie auch unterschiedliche Angriffsstrategien abdeckten.

Der *Massenfischzug* auf der Basis des Gelehrtenkalenders (vgl. Abschnitt 9.1) hat erwiesen, daß unter realistischen Bedingungen ausgeschlossen werden kann, daß mit leicht verfügbarem Zusatzwissen ohne großen Aufwand eine größere Zahl von Personen reidentifiziert werden kann, wenn Einzelangaben aus einer Untersuchung übermittelt werden, die auf einer Stichprobenerhebung basieren, deren Auswahlsatz so klein wie beispielsweise beim Mikrozensus ist. Um z.B. 100 Personen aus dem Mikrozensus richtig zuzuordnen, müßte ein Angreifer über ein Identifikationsfile mit ca. 200 000 Einzeldatensätzen verfügen, deren Informationsgehalt ebenso differenziert, reichhaltig und sorgfältig aufbereitet sein müßte, wie dies beim Gelehrtenkalender der Fall war. Gleichzeitig würden bei einem solchen Vorhaben 250 weitere Mikrozensusfälle zwar eindeutig einem Fall aus dem Identifikationsfile zugeordnet, diese Zuordnungen wären jedoch falsch. Der Angreifer wüßte nicht, welche der insgesamt 350 zugeordneten Fälle zu den 100 richtigen, welche zu den 250 falschen gehören.

---

* Autoren: Walter Müller, Heike Wirth

Entsprechende Identifikationsfiles existieren nicht, ihr Aufbau wäre mit immensen Kosten verbunden.

Die wesentliche Kostenkomponente bei einem Massenfischzug mit Abgleichtechniken ist nicht die EDV-Software oder die Rechenzeit, sondern die Verfügung über umfangreiche Adressdateien mit differenzierten Überschneidungsmerkmalen, die mit Informationen im Mikrodatenfile exakt und fehlerfrei übereinstimmen. Das Risiko der Erfolgschancen eines Massenfischzugs und die damit verbundenen Kosten werden deshalb nicht von den zukünftigen Entwicklungen in der EDV-Technologie oder den EDV-Kosten bestimmt, sondern allenfalls davon, ob innerhalb der Wissenschaft sehr umfangreiche (mehrere hunderttausend Fälle) personenbezogene Dateien leicht zugänglich werden, die in differenzierten Überschneidungsmerkmalen vollständig mit amtlichen Mikrodaten kompatibel und als EDV-lesbare Dateien erfaßt sind. Wenn ein solcher Fall absehbar wird, müßte man neue Aufwandsuntersuchungen durchführen und eventuell neue Anonymisierungsmaßnahmen entwickeln.

Auch bei Strategien der gezielten Suche haben die Experimente das Ergebnis erbracht, daß die untersuchten Einzelangaben faktisch anonym waren. Der Datensatz einer beliebigen Person erwies sich als faktisch anonym auch unter der Bedingung, daß er nur insoweit anonymisiert war, daß neben Namen und Adresse Regionalinformationen unterhalb der Ebene der Bundesländer weggelassen wurden. In welchem Ausmaß differenziertere Regionalinformationen das Reidentifikationsrisiko erhöhen, konnte nicht empirisch geprüft werden, aber die theoretischen Überlegungen auf Basis der vorliegenden Ergebnisse zeigen, daß das Risiko einer korrekten Reidentifikation auch beim Vorliegen von differenzierteren Regionalinformationen als sehr gering einzustufen ist.

Bei den Experimenten mit der Reidentifikationstechnik von Paaß hat sich erwiesen, daß die diskriminanzanalytische Reidentifikationstechnik bei realen Daten nicht 'leistungsfähiger' ist als einfache Abgleichverfahren (vgl. Abschnitt 9.2). Gleichzeitig haben diese Experimente deutlich gemacht, daß die Anwendung des Verfahrens im Hinblick auf den potentiellen Deanonymisierungsertrag einen prohibitiv hohen Aufwand bedeutet. Es kann auf absehbare Zeit ausgeschlossen werden, daß für Deanonymisierungszwecke dieses oder ein ähnlich komplexes Verfahren eingesetzt wird. Die Studie hat deshalb bei den Experimenten verstärkt die Leistungsfähigkeit und Störanfälligkeit der einfacheren, wesentlich weniger aufwendigen Verfahren untersucht.

Dabei ist eine wesentliche Erkenntnis der durchgeführten Experimente, daß das Vorhandensein einmaliger Ausprägungskombinationen zwar eine notwendige, aber keineswegs schon eine hinreichende Bedingung für eine Reidentifikation bedeutet. Sowohl bei den Experimenten mit dem Gelehrtenkalender als auch mit der sozialwissenschaftlichen Stichprobe wiesen sowohl im Mikrodatenfile wie im Identifikationsfile die überwiegende Zahl der relevanten Fälle einmalige Ausprägungskombinationen auf. Obwohl das Identifikationsfile im Falle des Gelehrtenszenarios einer Vollerhebung nahekam, war die Zahl der mit Sicherheit richtig vorgenommenen Reidentifikationen gleich Null. Der wichtigste Grund dafür ist die praktische Unvermeidbarkeit von Inkompatibilitäten zwischen den Datenfiles, wie bei der Überprüfung des sozialwissenschaftlichen Szenarios deutlich wurde.

Derartige Inkompatibilitäten, die auf Erhebungs-, Übertragungs- und Kodierfehler, unterschiedliche Abgrenzung der Kategorien, differierende Erhebungsverfahren, Erhebungskontexte, Erhebungsziele oder Erhebungszeitpunkte etc. zurückzuführen sind, treten offensichtlich in einem wesentlich höheren Ausmaß auf, als bislang angenommen wurde. Im Ergebnis bewirken solche Inkompatibilitäten zwischen Mikrodatenfile und Identifikationsfile bei Reidentifikationsversuchen eine Vielzahl von Falsch- und Nichtzuordnungen. Aus diesem Grunde führt eine Abschätzung des Reidentifikationsrisikos, die lediglich auf einer Auszählung von einzigartigen Fällen in einem Mikrodatenfile beruht oder den Anteil solcher Fälle mit statistischen Methoden in der Grundgesamtheit abschätzt, zwangsläufig zu einer erheblichen Überschätzung des real bestehenden Risikopotentials.

Die Abschätzung der Häufigkeit von einzigartigen Ausprägungskombinationen erfolgt auf der Basis der zur Verfügung stehenden Überschneidungsmerkmale. Je mehr Überschneidungsmerkmale vorhanden sind, desto höher ist der Anteil der einzigartig vertretenen Fälle und damit das potentielle Reidentifikationsrisiko. Zugleich wird eine Dateninkompatibilität um so wahrscheinlicher, je mehr Variablen benötigt werden, um eine einmalige Ausprägungskombination herzustellen. Da jedes Merkmal mit einer gewissen Wahrscheinlichkeit inkompatibel ist, wächst die Wahrscheinlichkeit, daß zumindest bei einem Merkmal eine Dateninkompatibilität vorliegt, mit jedem zusätzlich einbezogenen Merkmal, sofern die Unstimmigkeiten nicht vollständig korreliert sind. Die Bedeutung dieses Aspektes geht aus der folgenden Übersicht hervor:

Übersicht 13.1: Anteil der kompatiblen Ausprägungskombinationen in Abhängigkeit der einbezogenen Merkmale

Anzahl einbezogener Merkmale	Anteil der vollständig kompatiblen Ausprägungskombinationen
1	.9
2	.81
5	.59
8	.43
10	.35

Bei dieser Aufstellung wird angenommen, daß jedes Merkmal im Vergleich der beiden Erhebungen eine Inkompatibilität von durchschnittlich zehn Prozent aufweist und keine Korrelation zwischen den Fehlern besteht. Eine solche Annahme unterschätzt das bestehende Ausmaß von Dateninkompatibilitäten zwischen verschiedenen Datenquellen oder Erhebungen eher, als daß es diese überschätzt. Dabei kann das Identifikationsfile ebenso wie das Mikrodatenfile von "Fehlern" im weitesten Sinn betroffen sein. Unter dieser Annahme weisen bereits bei einem Abgleich von nur einem Merkmal nur noch 90 Prozent der Datensätze keine Inkompatibilität auf. Bei nur zehn abzugleichenden Merkmalen sind bereits annähernd zwei Drittel der Datensätze inkompatibel. Eine wichtige Konsequenz aus diesem Befund ist die Tatsache, daß allein aufgrund 'natürlicher' Dateninkompatibilitäten Datensätze umso mehr gegen das Risiko der Reidentifikation geschützt sind, je mehr Variablen benötigt werden, um zu einzigartigen Ausprägungskombinationen zu gelangen. Schon bei fünf bis acht benötigten Merkmalen sind Datensätze allein wegen impliziter Dateninkompatibilitäten gegen Deanonymisierung mit Hilfe von Abgleich- und Selektionsroutinen in hohem Maße geschützt.

Nicht ausgeschlossen werden kann jedoch, daß in besonders gelagerten, sehr seltenen Einzelfällen bei einer gezielten Suche unter Umständen mit verhältnismäßig geringem Aufwand die Möglichkeit einer Reidentifikation besteht. Dieser Fall wurde wie folgt charakterisiert:

1) Eine im Mikrodatenfile gesuchte Person gehört einer sehr kleinen, durch ein spezifisches Merkmal eingrenzbaren Subpopulation an, z.B. einer bestimmten Berufsgruppe oder Nationalität (*sachliche Tiefengliederung*);

2) das Mikrodatenfile enthält differenzierte Regionalinformationen, so daß in den jeweiligen Regionaleinheiten nur wenige Angehörige dieser spezifischen Subpopulation leben (*regionale Tiefengliederung*);

3) ein Forscher, der Zugang zu den Einzelangaben des Mikrodatenfile hat, kann sich Kenntnisse über einen Angehörigen dieser spezifischen Subpopulation beschaffen und weiß, daß diese Person an der Mikrodatenerhebung, über deren Daten er verfügt, teilgenommen hat (*Teilnahmekenntnis*);

4) die Merkmale der Person sind genau in der Weise im Mikrodatenfile erfaßt, wie es der Forscher vermutet (*Kompatibilität*).

Es ist wichtig, darauf hinzuweisen, daß alle Bedingungen gleichzeitig erfüllt sein müssen, da sich nach den durchgeführten Experimenten allein dann eine Gefahr für die faktische Anonymität ergibt (vgl. Kapitel 11). Das gleichzeitige Zusammentreffen aller Bedingungen kann bei Stichprobenerhebungen als außergewöhnlich seltenes Ereignis betrachtet werden. Dennoch sollten bei der Datenübermittlung Vorkehrungen getroffen werden, daß auch eine solche (unwahrscheinliche) Risikokonstellation ausgeschlossen ist.

Deshalb wurde in einem weiteren Schritt die Wirkung ausgewählter Anonymisierungsmaßnahmen am Beispiel des Gelehrtenszenarios überprüft.

Die empirische Überprüfung erfolgte im Hinblick auf die oben genannten Bedingungen 1) bis 3). Für jede dieser Bedingungen wurden Maßnahmen getestet, von denen angenommen werden konnte, daß durch sie das Risiko einer Deanonymisierung verringert wird. Zur Verringerung des Reidentifikationsrisikos aufgrund der Zugehörigkeit zu einer spezifischen, zahlenmäßig kleinen Subpopulation oder aufgrund der Existenz von tief gegliederten Regionalinformationen auf seiten des Angreifers wurde die von einer Vergröberung von Merkmalen ausgehende Schutzwirkung geprüft.

Gegen das spezifische Gefährdungspotential, welches daraus resultieren könnte, daß ein Angreifer über eine bestimmte Person weiß, daß diese an einer amtlichen Erhebung teilgenommen hat, wurde das Mittel der Substichproben-

ziehung geprüft. Die Substichprobenziehung hat den Effekt, daß ursprünglich mehrdeutige Fälle eindeutig werden können. Zugleich sinkt mit fallender Substichprobengröße die Wahrscheinlichkeit, daß eine beliebige Person in den übermittelten Mikrodaten enthalten ist. Dadurch erhöht sich in einem informationstheoretischen Sinn die Unsicherheit des Angreifers über die Richtigkeit von Zuordnungen beträchtlich. Selbst wenn ein Angreifer weiß, daß eine bestimmte Person an einer amtlichen Erhebung teilgenommen hat, kann er - falls nur eine Substichprobe weitergegeben wird - auch bei einer eindeutigen Zuordnung nicht sicher sein, ob es sich hierbei um die gesuchte Person oder um einen statistischen Doppelgänger handelt.

13.2 Folgerungen für die Übermittlung faktisch anonymer Daten

Die Untersuchung hatte ergeben, daß die den einzelnen Subszenarien zugrundegelegten Mikrodatenfiles, mit Ausnahme der oben präzisierten Risikokonstellation, für die untersuchten Bedingungskonstellationen faktisch anonym waren. Die zusätzlich überprüften Anonymisierungsmaßnahmen zeigten, daß auch gegen diese Risikokonstellation Schutzvorkehrungen getroffen werden können, die so wirksam sind, daß Deanonymisierungen nur mit einem Aufwand möglich sind, der die Kosten für alternative Beschaffungsmöglichkeiten der Informationen bei weitem übersteigt.

Die bei den Experimenten genutzten Datensätze enthielten nur einen begrenzten Ausschnitt der Merkmale des Mikrozensus. Es waren jedoch diejenigen Merkmale, bei denen ein Deanonymisierungsversuch ansetzen müßte, d.h. die typischen Überschneidungsmerkmale mit dem möglicherweise vorhandenen oder beschaffbaren Zusatzwissen. Die Übermittlung zusätzlicher anderer Merkmale beeinflußt das Reidentifikationsrisiko nicht in entscheidender Weise. Dafür sind vor allem zwei Gründe verantwortlich. Zum einen ist in aller Regel das Zusatzwissen, das für die Nutzung dieser Merkmale zu Deanonymisierungszwecken erforderlich wäre, nicht vorhanden. Ein Deanonymisierungsversuch hätte ja geradezu zum Ziel, Kenntnis über solche Merkmale zu erlangen. Je mehr Merkmale man als Überschneidungsmerkmale in Erwägung zieht, um die Reidentifikationschance zu erhöhen, desto größer sind die Kosten der Ermittlung dieser Merkmale und um so geringer ist der Gewinn, der aus einer potentiellen Deanonymisierung gezogen werden kann. Zum andern ist das folgende Ergebnis dieser Studie zentral: Wegen der unvermeidlichen Inkompatibilität unterschiedlicher Datenbestände sinkt mit jedem zusätzlich

genutzten Merkmal die Wahrscheinlichkeit, daß eine eindeutige Zuordnung auch eine richtige Zordnung ist und wirkt damit dem höheren Informationspotential des Zusatzwissens entgegen. Für die Anonymisierung von Datensätzen mit Einzelangaben sind deshalb maßgeblich nur die Merkmale zu berücksichtigen, für die typischerweise Zusatzwissen vorhanden ist.

Mit Ausnahme der beschriebenen Risikokonstellation waren die Anonymisierungsmaßnahmen, wie sie bei der Datenweitergabe im Rahmen des Anonymisierungsprojektes vorgenommen wurden, bei weitem hinreichend für die faktische Anonymität. Im folgenden werden deshalb Empfehlungen für Maßnahmen zur Sicherstellung der faktischen Anonymität von Einzelangaben entwickelt. Die Vorschläge beziehen sich nur auf den Mikrozensus und die EVS, weil die Untersuchungen dieser Studie auf diese Datenfiles ausgerichtet waren.

Die allgemeinen Schutzvorkehrungen sind weitgehend bereits durch Gesetze vorgegeben und können durch vertragliche Bindung des Empfängers faktisch anonymer Daten konkretisiert und ergänzt werden. Des weiteren wird eine möglichst strikte Geheimhaltung der lokalen Umsetzung der Stichprobenpläne durch die amtliche Statistik empfohlen, sowie eine systemfreie Anordnung der weitergegebenen Daten.

Das Restrisiko einer möglichen Deanonymisierung, das nicht mit Sicherheit ausgeschlossen werden konnte, betrifft eine Konstellation, die durch das gleichzeitige Zusammentreffen mehrerer Bedingungen gekennzeichnet ist. Das Eintreten einer solchen Konstellation erscheint zwar als sehr unwahrscheinlich. Im Sinne eines weitestgehenden Datenschutzes und einer weitestgehenden Risikoausschließung für die datenübermittelnden Institutionen sollten aber dennoch Maßnahmen ergriffen werden, durch die eine solche Konstellation - wie unwahrscheinlich sie auch sein mag - nach allem Ermessen verhindert werden kann.

Bei der Entwicklung von Anonymisierungsmaßnahmen muß jedoch auch berücksichtigt werden, daß das wissenschaftliche Analysepotential der Daten durch solche Maßnahmen möglichst wenig beeinträchtigt wird. Das wissenschaftliche Potential von Mikrodaten liegt in der Präzision von Aussagen über sachlich oder regional tiefgegliederte Bevölkerungsgruppen. Nur mit Mikrodaten können auch zahlenmäßig kleine Bevölkerungsgruppen in ihrer Größe präzise bestimmt und in Veränderungen genau analysiert werden. Das gleiche gilt für regionalspezifische Analysen. Nur aufgrund der außerordentlich großen Stichproben von amtlichen Erhebungen (wie z.B. des Mikrozensus) können aussagefähige Analysen auch regional disaggregiert durchgeführt werden. Es ist

davon auszugehen, daß die Wissenschaft Mikrodaten genau zu diesen Zwecken benötigt.

Die gleichzeitige regionale und sachliche Tiefengliederung ist jedoch ein wesentlicher Faktor in der oben dargestellten Risikokonstellation. Bei bestimmten Analyseverfahren (tabellarische Aufgliederungen) ist die gleichzeitige sachliche und die regionale Tiefengliederung nicht sinnvoll, da Zellenbesetzungszahlen sehr klein werden und wegen großer Zufallsschwankungen wenig aussagefähig sind. Bei diesen Verfahren muß man in der Regel entweder in der regionalen oder sachlichen Analysedimension aus statistischen Gründen Vergröberungen vornehmen. Bei fortgeschritteneren multivariaten Analyseverfahren sind solche Aggregierungen nicht erforderlich und beeinträchtigen das Analysepotential. Dennoch scheint es die sinnvollste Lösung zu sein, aus Datenschutzgründen bei der *Datenweitergabe für den Regelfall zwei unterschiedliche Datenfiles* vorzusehen: ein sogenanntes *Grundfile* und ein sogenanntes *Regionalfile*.

Beim *Grundfile* sind die Regionalinformationen nur in undifferenzierter Form, die übrigen Informationen jedoch in möglichst großer Differenzierung enthalten. Auch der Haushaltszusammenhang der Befragten soll erhalten bleiben. Das *Regionalfile* enthält dagegen stärker differenzierte Regionalinformationen, grenzt dafür die Differenzierungstiefe bei den übrigen Variablen ein. Damit sind zwei Elemente entkoppelt, die in der Verbindung (und nur in der Verbindung) zu der oben dargestellten begrenzten Risikokonstellation führen. Werden die Daten ohne kleinräumigen Regionalbezug übermittelt, dann sind sie - so haben die Experimente gezeigt - bereits durch die Entfernung der personenbezogenen Angaben faktisch anonym. Als zusätzliche Sicherung kann die allgemein empfohlene Festlegung eines Minimums in den univariaten Randverteilungen gelten. Bei differenzierten Regionalinformationen soll und kann die faktische Anonymität durch weitere Merkmalsvergröberung bei den differenziert erfaßten Überschneidungsmerkmalen gesichert werden.

Im Detail werden die bei beiden Datenfiles vorzunehmenden Einzelmaßnahmen in Kapitel 14 dargestellt und im Hinblick auf die Anonymisierungswirkung begründet. Sie sind so gestaltet, daß nach allem Ermessen die faktische Anonymität bereits durch diese Maßnahmen gesichert ist. Um die Sicherheit zusätzlich zu erhöhen, wird auch die Ziehung von Substichproben empfohlen. Sie verhindert, daß ein Datenangreifer mit Sicherheit weiß, ob eine bestimmte Person im übermittelten Mikrodatenfile enthalten ist und erhöht damit den Unsicherheitsfaktor bei einem Reidentifikationsversuch. Zugleich verringert sie prinzipiell die Reidentifikationswahrscheinlichkeit. Die Ziehung von Substich-

proben ist in jedem Fall mit Einschränkungen in der Präzision von Analyseergebnissen verbunden, aber sie wirkt sich im Regionalfile schwerwiegender aus als im Grundfile.

Nach dem neuen Stichprobenplan bilden die unterste Ebene, für die der Mikrozensus regional repräsentative Aussagen zuläßt, Regionaleinheiten in der Größe von ca. 200 000 Einwohnern. Für solche Einheiten enthält der Mikrozensus ca. 2 000 Fälle. Bei nur wenigen weiteren Aufgliederungen sind die Fehlerbereiche bei Stichproben dieses Umfangs schon sehr groß. Jede Substichprobenziehung bedeutet deshalb eine empfindliche Einschränkung des Analysepotentials. Es wird deshalb empfohlen, die Substichprobe für das Regionalfile keinesfalls niedriger als bei 85 Prozent anzusetzen.

Beim Grundfile muß eine Substichprobenziehung ebenfalls bei vielen Analysen als empfindlicher Informationsverlust gewertet werden. Man denke beispielsweise an eine Untersuchung der beruflichen Segregation nach Geschlechtern und ihre Entwicklung in der Kohortenfolge. Da bei einer solchen Analyse die gleichzeitige Berücksichtigung sehr differenzierter Berufsgruppen, von Altersgruppen, Geschlecht und eventuell Wirtschaftszweig erforderlich ist, werden auch hier sehr schnell die Grenzen sichtbar, unterhalb derer eine Substichprobenziehung an der Substanz der Analysemöglichkeiten des Mikrozensus rührt. Es wird deshalb empfohlen, beim Grundfile als unterste Grenze eine Substichprobe von 70 Prozent anzusetzen.

Ergänzend wird die Weitergabe einer Einprozent-Substichprobe aus dem Mikrozensus empfohlen. Sie enthält mit Ausnahme der Regionalangaben sämtliche Merkmale des Mikrozenus ohne weitere Anonymisierungsmaßnahmen. Durch den Ausfall der Regionalinformationen ist die starke Substichprobenziehung eine hinreichende Schutzmaßnahme. Dieses File ist für Analysen gedacht, in welchen eine Massendatenbasis nicht erforderlich ist oder für Wissenschaftler, die eine Massendatenanalyse nicht selbst vornehmen, jedoch an einem kleineren Datenfile geplante Analysen testen wollen.

14 Empfehlungen für die Weitergabe von Einzelangaben aus dem Mikrozensus und der EVS an die Wissenschaft*

Nach den durchgeführten Untersuchungen wird die Anwendung unterschiedlicher Maßnahmen der Anonymisierung für ein wenig regionalisiertes Grundfile und für ein stärker regionalisiertes Regionalfile empfohlen. Wegen unterschiedlicher Datenstrukturen erweisen sich für Mikrozensus und EVS zum Teil unterschiedliche Maßnahmen als adäquat.

Mit Hilfe des Regionalfile soll es möglich sein, Mikrodaten auf einer Ebene von Regionaleinheiten zu analysieren, für die diese Mikrodaten gemäß Stichprobenplan als repräsentativ betrachtet werden können. Bei der EVS sind dies die Bundesländer. Eine weitergehende Regionalisierung der EVS als sie im Grundfile vorgesehen ist, erscheint deshalb nicht sinnvoll. Für die EVS muß deshalb kein gesondertes Regionalfile erstellt werden. Nach dem neuen Stichprobenplan des Mikrozensus bezieht sich die Untergrenze für Regionaleinheiten, für die der Mikrozensus repräsentativ ist, auf Gebiete mit ca. 200 000 Einwohnern. Durch ein Regionalfile für den Mikrozensus sollten deshalb Analysemöglichkeiten geschaffen werden, durch die Regionalisierungen bis zu dieser, durch den Stichprobenplan der Daten bestimmten Ebene möglich sind. Werden nur Regionaltypisierungen vorgenommen, die mit den Konventionen des Grundfile vereinbar sind, so gelten die Regeln für das Grundfile.

Die im folgenden dargestellten und begründeten Empfehlungen sollten als erste Konkretisierung für faktisch anonymisierte Mikrodatenfiles der EVS und des Mikrozensus, die in einem Standardfall angewandt werden können, verstanden werden. Nach einiger Zeit der Praxis und Erfahrungssammlung sollten sie im Hinblick auf ihre Bewährung überprüft und gegebenenfalls revidiert werden. Da insbesondere noch keine genauen Erfahrungen dazu vorliegen, auf welchem Niveau der neue Stichprobenplan des Mikrozensus regional verwertbare Analysen zuläßt, sind vor allem die für das Regionalfile gemachten Vorschläge als Orientierungsgrößen zu betrachten, die möglicherweise schon bald in Abstimmung mit Regionalforschern an neue Erkenntnisse, Erfahrungen und einen neuen Bedarf anzupassen sind.

---

* Autoren: Walter Müller, Uwe Blien, Peter Knoche, Heike Wirth

Mit den Vorschlägen soll zwar eine gewisse Routinisierung und damit auch Aufwandseinsparung bei der Datenweitergabe erreicht werden. Es sollte jedoch weiterhin möglich sein, für spezifische Forschungszwecke durch eine unterschiedliche Ausgestaltung verschiedener Anonymisierungs- und Sicherungsmaßnahmen (z.B. Merkmalsvergröberung, Stichprobenziehung, technisch-organisatorische Maßnahmen) Lösungen vorzusehen, die bei einem vergleichbaren Schutz vor Deanonymisierung speziell auf den Forschungszweck abgestimmt sind.

Im einzelnen werden allgemeine und datenfile-spezifische Schutzvorkehrungen unterschieden. Sie werden im folgenden zunächst kurz beschrieben und danach begründet.

14.1 Allgemeine Schutzvorkehrungen

Hierzu zählt zunächst die vertragliche Bindung des Empfängers faktisch anonymer Daten. In §16 Abs.6 BStatG wird der Wissenschaft mit der Akzeptanz eines verbleibenden Risikos ein spezifisches Nutzungsprivileg zugestanden. Das heißt aber auch, daß eine Reidentifikation nicht mit absoluter Sicherheit ausgeschlossen werden kann. Deshalb muß vertraglich festgehalten werden, daß der Datenempfänger

1) durch geeignete technisch-organisatorische Maßnahmen

- eine interne Kontrollmöglichkeit über die Datennutzung hat, um Deanonymisierungsversuche zu verhindern bzw. rechtzeitig zu erkennen,

- die ausschließliche Nutzung der Daten durch nur die im Vertrag bezeichneten nutzungsberechtigten Amtsträger oder für den öffentlichen Dienst bzw. zur Geheimhaltung besonders Verpflichtete sichergestellt und

- externe Zugriffe verhindert;

2) bei der Verwirkung einer zusätzlich vereinbarten Vertragsstrafe keinen Versuch einer Reidentifikation unternimmt;

3) die übermittelten Daten ausschließlich für das angegebene wissenschaftliche Vorhaben verwendet;

4) die übermittelten Daten nicht an Dritte weitergibt oder gelangen läßt;

5) die Daten nach Durchführung des wissenschaftlichen Vorhabens, für das sie übermittelt wurden, löscht, bzw. sich zur Rückgabe der Originalbänder verpflichtet;

6) sich verpflichtet, jede Vervielfältigung bzw. Herstellung von Auszügen der Originalangaben zu unterlassen. Falls ein solches Vorgehen durch das Forschungsvorhaben jedoch notwendig ist, müssen die Duplikate ebenso wie die Originale behandelt werden. Über die Anzahl und den Verbleib der hergestellten Duplikate muß auf Verlangen bzw. unaufgefordert nach Beendigung des Forschungsvorhabens Auskunft erteilt werden;

7) jede Nachforschung hinsichtlich der lokalen Umsetzung der den Angaben zugrundeliegenden Stichprobenpläne unterläßt.

Von seiten der amtlichen Statistik sollte dafür gesorgt werden, daß die lokale Umsetzung der Stichprobenpläne möglichst geheimgehalten wird, um zu verhindern, daß das Wissen um die Teilnehmer an einer amtlichen Erhebung für Reidentifikationszwecke eingesetzt werden kann.

Ebenfalls in den Bereich der allgemeinen Schutzvorkehrungen sind jene Anonymisierungsmaßnahmen zu zählen, die generell an einem Mikrodatenfile zu treffen sind. Hierunter zählt man zum einen die systemfreie Anordnung der Daten, wobei allerdings der Haushaltszusammenhang der Befragten erhalten bleiben soll. Eine systematische Anordnung (z.B. nach räumlichen Kriterien) der Datensätze erhöht das Reidentifikationsrisiko für alle Datensätze, sofern ein potentieller Angreifer das Organisationsprinzip kennt (Südfeld 1987). Um dieser Gefahr vorzubeugen, sollte das Mikrodatenfile daher systemfrei sortiert werden. Ein nach inhaltlichen Kriterien sortiertes Mikrodatenfile wird jedoch vor allem im Zusammenhang mit Kontextanalysen benötigt. Es sollten deshalb Sonderregelungen vorgesehen werden, falls bestimmte Analyseziele keine systemfreie Sortierung erlauben.

14.2 Datenfile-spezifische Anonymisierungsmaßnahmen

Für das *Grundfile* wird empfohlen:

1) Die Regionalangaben in Mikrozensus und EVS werden nur vergröbert weitergegeben. Sie schließen die Angabe über das Bundesland und (außer für die kleinen Bundesländer Bremen und Saarland) eine Klassifikation des siedlungsstrukturellen Typs ein. Zu empfehlen wären hier Typisierungsmerkmale, wie z.B. die von der BfLR entwickelte Gemeindetypologie. Alternativ kommt eine vergröberte Klassifikation der Gemeindegrößenklasse (oder ein ähnliches Merkmal zur Charakterisierung des Raumtyps) in Betracht. Dabei muß sichergestellt sein, daß

- keine einzelne Gemeinde identifiziert werden kann, die weniger Einwohner als 500 000 umfaßt;

- ein Gemeindetyp (z.B. Gemeindegrößenklasse), dem mehrere Gemeinden zugehören, in keinem Bundesland weniger als 400 000 Einwohner umfaßt.

2) Angaben über Staatszugehörigkeit in Mikrozensus und EVS werden nur so aggregiert weitergegeben, daß eine Nationalität oder identifizierbare Gruppe von Nationalitäten in der Bundesrepublik insgesamt wenigstens 50 000 Einwohner umfaßt.

3) Alle übrigen Variablen sollen beim Mikrozensus - falls erforderlich - so aggregiert werden, daß in den univariaten Randverteilungen jede ausgewiesene Merkmalsausprägung für die Grundgesamtheit der Bundesrepublik mindestens 5 000 Fälle umfaßt. Dies entspricht ca. 50 Fällen im Datensatz des Mikrozensus.

   Bei der EVS werden folgende Maßnahmen der Merkmalsvergröberung empfohlen:

- Bei "sichtbaren" oder über die Zeit vergleichsweise stabilen Merkmalen - wie Geburtsjahr, Stellung im Beruf oder Besitz auffälliger Konsumgüter - Maßnahmen der Merkmalsvergröberung oder -aggregierung in der Weise, daß nur Merkmalsausprägungen ausgewiesen werden, die für die Grundgesamtheit der Bundesrepublik mindestens 5 000 Fälle umfassen. Dies entspricht ca. zehn Fällen im Datensatz der EVS.

- Bei öffentlich wenig bekannten oder über die Zeit wenig stabilen, jedoch differenziert erfaßten Merkmalen - im wesentlichen die nicht-gruppierten Einkommens-, Vermögens und Ausgabenbeträge - sollen die jeweils fünf niedrigsten und fünf höchsten Ausprägungen eines Merkmals nur als Mittelwert dieser Ausprägungen ausgewiesen werden. Die übrigen Ausprägungen im untersten und obersten Dezil der Verteilung eines solchen Merkmals sollen mit einem Zufallsfehler von bis zu plus oder minus ein Prozent des jeweiligen Merkmalswertes überlagert werden.

4) Es wird beim Mikrozensus nur eine Substichprobe von 70 Prozent weitergegeben.

   Bei der EVS wird der Datensatz in Abhängigkeit von der Anzahl der weitergegebenen Erhebungsteile mit folgenden Auswahlsätzen weitergegeben:

   98 Prozent: Haushalts- und Personenmerkmale aus dem Grundinterview + 1 Erhebungsteil[2]
   90 Prozent: Haushalts- und Personenmerkmale aus dem Grundinterview + 2 Erhebungsteile[2]
   80 Prozent: Haushalts- und Personenmerkmale aus dem Grundinterview + 3 Erhebungsteile[2]

Für das *Regionalfile* des Mikrozensus wird empfohlen:

1) Durch die Kombination von Regionalklassifikationen soll keine Regionaleinheit ermittelbar sein, die eine Einwohnerzahl von weniger als 100 000 Personen aufweist.

2) Die Überschneidungsmerkmale Beruf, Wirtschaftszweig, Nationalität und Alter sollen so vergröbert werden, daß keine Merkmale ausgewiesen werden, die

- in der Grundgesamtheit nicht wenigstens 50 000 Einwohner umfassen;

---

2 Erhebungsteile in diesem Sinn sind das Schlußinterview, der Erhebungsteil über die Nahrungs- und Genußmittel sowie der Erhebungsteil über die Jahresrechnung.

- pro übermittelter Regionaleinheit (ohne Substichprobenziehung) nicht mindestens drei Fälle im Mikrodatenfile enthalten. Merkmalsausprägungen, die im Mikrodatenfile nur einen oder zwei Fälle enthalten, werden nur in einer stärker aggregierten Weise ausgewiesen.

3) Alle übrigen Variablen sollen - falls erforderlich - so aggregiert werden, daß jede ausgewiesene Merkmalsausprägung für die Grundgesamtheit der Bundesrepublik mindestens 5 000 Fälle umfaßt.

4) Es wird nur eine Substichprobe weitergegeben. Der Auswahlsatz sollte mindestens 85 Prozent betragen.

14.3 Begründung

Bei der Bewertung dieser Maßnahmen soll das Vorliegen einer hinreichenden Garantie der faktischen Anonymität getrennt geprüft werden im Hinblick auf die beiden wichtigsten Angriffsstrategien, a) der Reidentifikation einer größeren Zahl von Datensätzen durch einen Massenfischzug, und b) der Identifikation einzelner Fälle durch eine gezielte Suche, wobei vor allem der Fall der Risikokonstellation zu berücksichtigen ist, in der der Datenangreifer Kenntnis darüber hat, daß die Daten einer bestimmten Person im Mikrodatenfile enthalten sind.

Als Grundlage für die Bewertung der vorgeschlagenen Anonymisierungsmaßnahmen können die durchgeführten empirischen Untersuchungen beigezogen werden. Dafür kommen insbesondere die Analysen des Gelehrtenkalender-Szenarios in Betracht, da in diesem Szenario für die Überschneidungsmerkmale die Kombination von sehr differenzierten sachlichen Merkmalen und von Regionalangaben untersucht wurde. Die mit der sozialwissenschaftlichen Stichprobe und dem Gelehrtenkalender durchgeführten Untersuchungen haben gezeigt, daß das Risiko einer Reidentifikation vor allem mit der Kenntnis von wenigen sehr differenzierten Merkmalen zunimmt, weniger jedoch mit der Kenntnis einer größeren Zahl wenig differenzierter Merkmale, die vielen Personen gemeinsam sind.

Beim Szenario mit dem Gelehrtenkalender enthielten das Mikrodatenfile und auch das Identifikationsfile als Regionalinformationen für Nordrhein-Westfalen die 16 Bundesraumordnungsregionen und die Gemeindegrößenklasse, die in neun Kategorien erfaßt war. Bei der Kombination dieser Regionalmerkmale gehörten nach den durchgeführten Untersuchungen 35 Prozent der Fälle des

Mikrodatenfile Regionaleinheiten zu, die in der Grundgesamtheit eine Bevölkerungszahl von 400 000 oder mehr Einwohnern aufweisen; 46 Prozent der Fälle gehörten Regionaleinheiten zu, die in der Grundgesamtheit zwischen 200 000 und 400 000 Einwohner aufweisen; 19 Prozent der Fälle gehörten Regionaleinheiten mit weniger als 200 000 Einwohnern in der Grundgesamtheit an. Im Hinblick auf den Differenzierungsgrad der Merkmale waren im Gelehrtenszenario ca. 22 Prozent der Fälle in Regionaleinheiten enthalten, in denen weniger als 100 Gelehrte wohnen; 22 Prozent der Fälle in Regionaleinheiten mit 100 bis 200 Gelehrten und 56 Prozent der Fälle in Regionaleinheiten mit über 200 Gelehrten.

14.3.1 Reidentifikation durch Massenfischzug

Für das *Grundfile* legen die vorgesehenen Weitergaberegeln im Hinblick auf die Regionalinformationen einen deutlich restriktiveren Maßstab an als den, der im Gelehrtenszenario empirisch geprüft wurde. Es wird vorgeschlagen, daß keine Regionalinformationen weitergegeben werden, mit denen es möglich wäre, Regionaleinheiten mit weniger als 400 000 Einwohnern abzugrenzen. Die Grenze von 400 000 Einwohnern gilt nur dann, wenn mehrere, in der Regel nicht benachbarte Gemeinden aggregiert werden müssen, um diese Einwohnerzahl zu erreichen. Es kann davon ausgegangen werden, daß es wesentlich aufwendiger ist, Zusatzwissen zu beschaffen, wenn die einer Regionaleinheit zugehörigen Einzelangaben des Mikrodatenfile aus mehreren Gemeinden aggregiert sind. Es darf keine einzelne Gemeinde identifizierbar sein, die weniger als 500 000 Einwohner enthält. Im Gelehrtenszenario gehörten 94 Prozent der Fälle einer Regionaleinheit mit weniger als 500 000 Einwohnern und 65 Prozent der Fälle einer Regionaleinheit mit weniger als 400 000 Einwohnern an.

Da beim Grundfile die Regionalinformationen weniger differenziert sind als im untersuchten Szenario mit dem Gelehrtenkalender und da bei diesem Szenario die faktische Anonymität bei weitem übererfüllt war - beim Massenfischzug wäre die faktische Anonymität selbst bei einer etwa 100 mal größeren Zahl richtiger Zuordnungen noch erfüllt gewesen - ist es im Hinblick auf einen Massenfischzug gerechtfertigt, von der faktischen Anonymität auszugehen, selbst wenn einzelne Merkmale deutlich seltener in der Grundgesamtheit vorkommen als die empirisch untersuchten Gelehrten. Es soll jedoch verhindert werden, daß Personen aufgrund besonders seltener oder auffälliger Merkmalsausprägungen leicht identifizierbar sind.

Für den Mikrozensus wird deshalb das Erfordernis eines Besetzungsminimums von 5 000 Merkmalsträgern in der Grundgesamtheit vorgeschlagen. Damit kann diese Bedingung als erfüllt betrachtet werden. Diese Empfehlung bedeutet, daß beim Mikrozensus einzelne der sehr differenziert ausgewiesenen Berufs- und Wirtschaftszweigkategorien nach sachlichen Gesichtspunkten zusammengefaßt werden müssen.

Bei der EVS ist im Hinblick auf die Verwendbarkeit als Zusatzwissen zwischen zwei Typen von Merkmalen zu unterscheiden:

1. Merkmale von denen unterstellt werden kann, daß sie in der Öffentlichkeit vergleichsweise bekannt ("sichtbar") oder im Zeitverlauf stabil sind (z.B. Alter, Geburtsjahr, Nationalität, Stellung im Beruf, Besitz auffälliger Konsumgüter). Bei dieser Gruppe von Merkmalen sollten die gleichen Regeln der Merkmalsvergröberung wie beim Mikrozensus gelten.

2. Merkmale, die zwar sehr differenziert erfaßt, jedoch weniger "sichtbar" oder stabil sind. Dies gilt insbesondere für die Angaben über Einkommen, Vermögen oder Ausgaben, die in nicht-gruppierten Beträgen erfaßt sind. Solche Angaben sind in der Regel vergleichsweise wenig reliabel, d.h. die Wahrscheinlichkeit, daß zwei unterschiedliche Informationsquellen exakt übereinstimmen, ist verhältnismäßig klein. Solche Angaben sind deshalb in der Regel nicht als riskantes Zusatzwissen einzustufen. Allenfalls bei den Extremen der Merkmalsverteilungen könnte ein gewisses Risiko bestehen, weil einzelne Beträge dort sehr selten vorkommen. Im mittleren Verteilungsbereich sind die Ausprägungen in der Regel dichter besetzt. Auch aus Gründen der geringen Stabilität und Reliabilität sind die Ausprägungen im mittleren Verteilungsbereich für Deanonymisierungszwecke völlig ungeeignet. Es wird deshalb empfohlen, daß diese Merkmale im mittleren Verteilungsbereich unverändert belassen werden. Für die Verteilungsextreme werden dagegen folgende Maßnahmen vorgeschlagen: Die jeweils fünf niedrigsten und fünf höchsten Ausprägungen eines Merkmals werden nur als Mittelwert dieser Ausprägungen ausgewiesen. Die übrigen Ausprägungen im untersten und obersten Dezil der Verteilung werden mit einem Zufallsfehler von bis zu plus oder minus ein Prozent des jeweiligen Merkmalswertes überlagert.

Da die Staatsangehörigkeit ein in der Regel leicht und kompatibel mit dem Mikrodatenfile erfahrbares Merkmal ist, wird zusätzlich empfohlen, dieses Merkmal nur stark vergröbert weiterzugeben. Informationen über die Staatsangehörigkeit sollen nur in einer Klassifikation weitergeben werden, bei der

jeder Merkmalsausprägung in der Grundgesamtheit der Bundesrepublik wenigstens 50 000 Personen zukommen. Dies bedeutet, daß im Durchschnitt in den kleinsten Regionaleinheiten des Grundfile (400 000 Einwohner) ca. 250 Angehörige einer Nation wohnen werden. Da der Wissenschaft über Ausländer in der Regel kein leicht zugängliches und differenziertes Zusatzwissen zur Verfügung steht, erscheint eine solche Minimalbesetzung zusammen mit den übrigen Maßnahmen als hinreichende Sicherung für die faktische Anonymität bezogen auf die Sondersituation von ausländischen Staatsangehörigen.

Das vorgesehene *Regionalfile* für den Mikrozensus enthält differenziertere Regionalinformationen als das Grundfile. Aufgrund der impliziten Begrenzungen in der Repräsentativität der Mikrozensusstichprobe beziehen sich jedoch auch diese Regionalinformationen auf vergleichsweise große Regionaleinheiten, keinesfalls beispielsweise auf Regionaleinheiten im Umfang von Blockseiten wie dies bei der Volkszählung möglich ist. Im untersuchten Szenario mit dem Gelehrtenkalender gehörten 19 Prozent der Fälle Regionaleinheiten an, die weniger als 200 000 Einwohner umfaßten. Durch die Kombination von sich überschneidenden Regionalinformationen kann es vorkommen, daß Regionen ausgegrenzt werden können, die weniger als 200 000 Personen umfassen. Dies war auch im untersuchten Szenario mit dem Gelehrtenkalender der Fall. Insgesamt kann man also davon ausgehen, daß das hier vorgesehene Regionalfile in den Regionsgrößen dem mit dem Gelehrtenkalender empirisch geprüften Szenario nicht unähnlich ist.

Was den Differenzierungsgrad bei den zentralen, differenziert erfaßten Überschneidungsmerkmalen Beruf, Wirtschaftszweig, Geburtsjahr und Staatsangehörigkeit betrifft, wird vorgeschlagen, daß nur Merkmalsausprägungen ausgewiesen werden, die in der Grundgesamtheit wenigstens 50 000 Einwohner umfassen und in einer ausgewiesenen Regionaleinheit im Mikrodatenfile mit mindestens drei Fällen besetzt sind. Merkmalsausprägungen, die nur mit ein oder zwei Fällen besetzt sind, sollen nur in einer stärker aggregierten Klassifikation ausgewiesen werden. Eine Besetzung mit mindestens drei Fällen im Mikrodatenfile des Mikrozensus bedeutet - bei grober Schätzung - daß die Grundgesamtheit in einer Regionaleinheit im Durchschnitt etwa 300 Personen mit dieser Merkmalsausprägung aufweisen wird. Eine solche Regelung ist deutlich restriktiver als die Bedingungen, die im Szenario mit dem Gelehrtenkalender geprüft wurden. In diesem Szenario gehörten 44 Prozent der Gelehrten Raumeinheiten an, die in der Grundgesamtheit weniger als 200 Gelehrte enthielten. Wenn eine minimale Besetzung von mindestens drei Fällen im Mikrodatenfile pro ausgewiesener Regionaleinheit vorgesehen wird, sind die

Besetzungszahlen also deutlich höher angelegt als im geprüften Gelehrtenszenario.

Da die relative Häufigkeit von Merkmalsbesetzungen nach Regionen zum Teil stark variiert, impliziert die vorgeschlagene Regel der Minimalbesetzung von mindestens drei Fällen, daß je nach Region unterschiedliche Ausprägungen ausgewiesen werden können. Um Vergleichbarkeit zwischen Regionen und Flexibilität für verschiedene Analysezwecke sicherzustellen, sollte für die betroffenen Variablen eine Art Doppelverkodung in einem hierarchisch strukturierten Klassifikationsschema vorgesehen werden. Wenn beispielsweise eine Merkmalsausprägung im dreistelligen Klassifikationsschema der Wirtschaftszweige wegen zu geringer Besetzung nicht ausgewiesen werden kann, sollte sie zumindest im zweistelligen Schema ausgewiesen werden. Alle Ausprägungen des dreistelligen Schemas, die das Besetzungskriterium von mindestens drei Fällen erfüllen, sollten dagegen ausgewiesen werden. Bevor an dieser Stelle konkrete Maßnahmen getroffen werden, sollten unterschiedliche Aggregierungsalternativen geprüft werden.[3] Nach einer Erprobungsphase sollte überprüft werden, ob das gewählte Verfahren dem vielseitigen Bedarf der Forschung gerecht wird und gegebenenfalls revidiert werden.

Im Hinblick auf denkbare Angriffsstrategien, in denen eine größere Zahl von Fällen deanonymisiert werden soll, kann ein nach den vorgeschlagenen Maßnahmen aufbereitetes Mikrodatenfile - sowohl in der Grundversion wie in der Regionalversion - zweifelsfrei als faktisch anonym gelten.

14.3.2 Gezielte Suche

Bei der gezielten Suche ist vor allem zu prüfen, ob ein Mikrodatenfile faktisch anonym ist, selbst wenn ein Datenangreifer Kenntnis darüber hat, daß der Datensatz für eine bestimmte Person im Mikrodatenfile enthalten ist. Unter dieser Bedingung ist das Reidentifikationsrisiko um ein Vielfaches höher als ohne entsprechende Kenntnis. Wir untersuchen deshalb im folgenden nur die Situation mit Teilnahmekenntnis. Wenn die faktische Anonymität unter dieser Bedingung gegeben ist, ist sie a fortiori auch unter der Bedingung ohne Teilnahmekenntnis gegeben. Besonders chancenreich ist eine Reidentifikation

---

3 Dabei wäre auch zu berücksichtigen, daß nicht wegen zufälligen Schwankungen von Jahr zu Jahr unterschiedliche Aggregierungen vorgenommen werden. Eine Lösung würde darin bestehen, daß das Besetzungskriterium im gleitenden Durchschnitt von fünf Jahren erfüllt ist.

weiterhin, wenn es sich bei der gesuchten Person um jemanden handelt, der sich durch ein sehr seltenes Merkmal auszeichnet. Es ist also insbesondere zu prüfen, ob die vorgesehenen Anonymisierungsmaßnahmen auch ausreichend sind, um die faktische Anonymität in der Situation sicherzustellen, die als Risikokonstellation charakterisiert wurde.

Zunächst ist darauf hinzuweisen, daß es sich hierbei um eine Situation handelt, bei der mehrere Bedingungen gleichzeitig realisiert sein müssen. Der Datenangreifer muß über Informationen über eine Person mit einem sehr seltenen Merkmal verfügen. Diese Person muß im Mikrodatenfile enthalten sein. Der Datenangreifer muß wissen, daß diese Person im Mikrodatenfile enthalten ist. Er muß bei dieser Person ein - in der Regel außerberufliches - Motiv haben, ihren Datensatz deanonymisieren zu wollen. Das Eintreffen jeder dieser Einzelbedingungen hat in aller Regel nur einen niedrigen Wahrscheinlichkeitsgrad. Daß alle Bedingungen gleichzeitig zutreffen, ist als äußerst unwahrscheinlich zu betrachten.

Das Wissen, daß der Datensatz für eine bestimmte Person im Mikrodatenfile enthalten ist, kann grundsätzlich auf zwei unterschiedliche Arten gewonnen sein. Dieses ist entscheidend für die Bewertung des Aufwandes, den ein Datenangreifer bei gegebenen Anonymisierungsmaßnahmen betreiben muß. Zum einen kann das Wissen dadurch zustande kommen, daß der Datenangreifer im Mikrodatenfile eine Person mit einem sehr seltenen Merkmal findet. Er weiß dann, daß der Datensatz für eine solche Person im Mikrodatenfile enthalten ist und er muß herausfinden, um welche Person es sich in Wirklichkeit handelt. Davon zu unterscheiden ist die Situation, in der der Datenangreifer von einer konkreten Person mit einem seltenen Merkmal, aus irgendwelchen Gründen, weiß, daß ihr Datensatz im Mikrodatenfile enthalten ist, z.B. weil ihm diese Person erzählt hat, daß sie befragt wurde . Beide Fälle werden im folgenden getrennt diskutiert.

In der ersten Situation, bei dem das Wissen aus der Inspektion des Mikrodatensatzes resultiert, ist ausschlaggebend, welchen Aufwand der Datenangreifer betreiben muß, um herauszufinden, um welche konkrete Person es sich in Wirklichkeit handelt. Um in dieser Situation die faktische Anonymität sicherzustellen, reichen die vorgeschlagenen Maßnahmen der Merkmalsvergröberung aus.

1. Im *Grundfile* schließt die vorgeschlagene minimale Merkmalsbesetzung von 5 000 Fällen die Möglichkeit aus, daß ein Datenangreifer aus dem Alltagswissen vermuten kann, um welche konkrete Person es sich handelt. Er muß systematisch die Personen ermitteln, auf die das Merkmal zutrifft.

Dies ist eindeutig aufwendiger als eine alternative Informationsbeschaffung.[4]

2. Für das *Regionalfile* wird vorgeschlagen, daß die ausgewiesenen Regionaleinheiten bei den Berufs- und Wirtschaftszweigmerkmalen, Nationalität und Geburtsjahrgang eine Besetzungshäufigkeit von wenigstens drei Fällen im Mikrodatenfile aufweisen, das entspricht etwa - grob geschätzt - 300 Fällen in der Grundgesamtheit. Damit wäre der Überprüfungsaufwand noch deutlich höher als im Grundfile. Es kann deshalb auch für das Regionalfile davon ausgegangen werden, daß die vorgesehenen Merkmalsvergröberungen bei den Überschneidungsmerkmalen 'Beruf' und 'Wirtschaftszweig' ausreichend sind, um die faktische Anonymität zu sichern.

In der Situation, in der der Datenangreifer von einer konkreten Person mit einem seltenen Merkmal weiß, daß sie im Mikrodatenfile enthalten ist, ergibt sich die faktische Anonymität aus folgenden Gründen:

Für das Grundfile unterstellen wir zur Vereinfachung, daß die Person mit dem seltenen Merkmal, von der der Datenangreifer die Einbeziehung in das Mikrodatenfile kennt, dort eine einzigartige Merkmalskombination aufweist. Dann könnte der Datenangreifer die Person im Mikrodatenfile reidentifizieren, vorausgesetzt alle Überschneidungsmerkmale wären in völliger Kompatibilität zum Wissen des Datenangreifers im Mikrodatenfile erfaßt. Letzteres kann jedoch

[4] Der geringste Aufwand würde dann entstehen, wenn für die Regionaleinheit, in der der Einzelfall gefunden wird, über die Personen mit dem seltenen Merkmal ein vollständiges Register mit informativen Überschneidungsmerkmalen zum Mikrodatenfile vorliegen würde. Wenn wir annehmen, daß es sich um einen Einzelfall in einer Regionaleinheit mit 400 000 Einwohner (ca. 0,5 Prozent der Gesamtbevölkerung) handelt, müssen die Daten von etwa 25 Personen (=0,5 Prozent von 5 000) nach den Konventionen des Mikrozensus aufbereitet werden. Bei einem Aufwand von 8 DM pro Person - wie er beim Gelehrtenszenario ermittelt wurde- würde allein die Überprüfung eines Einzelfalls die Kosten einer alternativen Datenbeschaffung übersteigen. Hinzu kommt jedoch, daß ein im Mikrodatenfile vorhandener Einzelfall keineswegs auch immer eindeutig einen Fall im Identifikationsfile zugewiesen werden kann. Wegen Dateninkompatibilitäten und möglichen Doppelgängern müßten mehrere Einzelfälle auf diese Weise aufbereitet werden, bis eine sichere Reidentifikation erwartet werden kann. Der Aufwand würde sich vervielfachen. Die Kosten einer Deanonymisierung übersteigen also die Kosten einer alternativen Informationsbeschaffung um ein Mehrfaches. Dabei ist zu betonen, daß für den Deanonymisierungsaufwand hier die allergünstigsten Annahmen getroffen wurden, die in der Wirklichkeit so in der Regel nicht gegeben sein werden.

nicht unterstellt werden. Nach Ergebnissen des Anonymisierungsprojektes benötigt er bei den Bedingungen des Gelehrtenszenarios mindestens sechs Fälle, um eine erfolgreiche Deanonymisierung zu erreichen. Wenn die vorgesehenen Anonymisierungsmaßnahmen getroffen werden (Vergröberung bei den Regionalangaben und insbesondere Substichprobenziehung) sind nach den durchgeführten Untersuchungen wenigstens doppelt so viele Versuche erforderlich. Das heißt: Um einen Fall zu reidentifizieren, müßte der Datenangreifer im Durchschnitt über Informationen zu mindestens zwölf Personen verfügen, von denen alle ein sehr seltenes Merkmal aufweisen und von denen der Datenangreifer weiß, daß sie im Mikrodatenfile enthalten sind. Ein solches Wissen kann nicht in einer realistischen Weise als bei einem Datenangreifer irgendwann gegeben vorausgesetzt werden. Ein Datenangreifer mag aus Zufall eine Person mit einem seltenen Merkmal kennen, von der er weiß, daß sie im Mikrodatenfile enthalten ist, aber es ist mit an Sicherheit grenzender Wahrscheinlichkeit auszuschließen, daß er gleich eine größere Zahl solcher Personen kennt.[5] Der Datenangreifer muß also entsprechendes Wissen sammeln. Für eine größere Zahl von Personen, die einer sehr kleinen Bevölkerungsgruppe angehören, herauszufinden, ob sie im Mikrodatenfile enthalten sind und für diese Personen das Wissen zu den Überschneidungsmerkmalen zu eruieren, ist sicher aufwendiger als für eine Person, über die sich der Datenangreifer im Mikrodatenfile informieren will, die Informationen auf direktem Wege zu beschaffen.

Beim Mikrozensus verringert sich zudem durch die Ziehung einer reduzierten Substichprobe die Gewißheit deutlich, daß ein dem Originalmikrodatenfile zugehöriger Datensatz auch noch im weitergegebenen Mikrodatenfile enthalten ist. Damit ist das Gefährdungspotential, das sich aus dem Wissen um die Mikrozensusteilnahme ergibt, prinzipiell gestört.

Die EVS weist hinsichtlich des im wissenschaftlichen Bereich typischerweise vorhandenen Zusatzwissens weniger differenzierte Überschneidungsmerkmale als der Mikrozensus auf. Bei den möglicherweise riskanten Extremwerten von differenzierten Einkommens-, Vermögens- und Ausgabenangaben sind durch Durchschnittsbildung und Fehlerüberlagerung Maßnahmen getroffen, die ihre Nutzung zu Reidentifikationszwecken verhindern. Die Wahrscheinlichkeit, daß

---

5 Die im neuen Stichprobenplan des Mikrozensus vorgesehene Verkleinerung der Klumpengröße auf durchschnittlich neun Wohnungen pro Klumpen verringert die Wahrscheinlichkeit praktisch auf Null, daß jemand, der einen Teilnehmer am Mikrozensus kennt, in dessen Nachbarschaft eine größere Zahl anderer Teilnehmer mit sehr seltenen Merkmalen finden wird.

ein entsprechender Datensatz im Mikrodatenfile reidentifiziert werden kann, erscheint deshalb bei der EVS eher geringer als beim Mikrozensus. Die EVS ist zudem eine kleinere Stichprobe. Damit ist es bei der EVS noch schwieriger als beim Mikrozensus, Personen mit bestimmten Merkmalen zu kennen oder zu finden, die an der Erhebung beteiligt sind. Deshalb kann bei der EVS die Weitergabe einer Substichprobe mit einem deutlich höheren Auswahlsatz als beim Mikrozensus empfohlen werden. Es wird empfohlen, die Größe der Substichprobe bei der EVS von der Zahl der weitergegebenen Erhebungsteile abhängig zu machen. Damit wird in genereller Weise dem unterschiedlichen Informationspotential der Daten Rechnung getragen.

Für das Regionalfile des Mikrozensus gelten zunächst die gleichen Überlegungen wie für das Grundfile. Selbst wenn ein Datenangreifer von einer Person mit einem seltenen Merkmal weiß, daß sie im Mikrodatenfile enthalten ist, kann er noch nicht davon ausgehen, daß er sie dort auch findet. Um einen Treffer zu erzielen, muß er auch hier im Durchschnitt entsprechendes Wissen über mehrere Personen besitzen. Weil sich im Regionalfile die Reduzierung der Stichprobe sehr einschränkend auf das Analysepotential der Daten auswirkt, wird nur eine leicht reduzierte Substichprobe empfohlen. Im Vergleich zum Grundfile wird die dadurch verringerte Schutzwirkung durch eine wesentlich vergrößerte minimale Merkmalsbesetzung ausgeglichen. Sie ist so groß, daß in der Regel auch für die seltenen Merkmale im Mikrodatenfile pro Regionaleinheit immer Mehrfachbesetzungen vorliegen. Dadurch verringert sich die Chance wesentlich, eine Person im Mikrodatenfile eindeutig zu identifizieren.

LITERATURVERZEICHNIS

Aarts, E.H., Korst, J. (1989): Simulated Annealing and Boltzmann Machines. A Stochastic Approach to Combinatorial Optimization and Neural Computing, Chichester usw.: John Wiley and Sons.

Aarts, E.H., Laarhoven, Peter J. van (1987): "Simulated annealing: a pedestrian review of the theory and some applications", in: Devijver, Kittler (1987)

Abel, H. et al. (1989): Datensicherungsmaßnahmen beim Einsatz von Arbeitsplatzcomputern, Datenschutz und Datensicherung 10.

Adorno, T.W. (1972a): Zur Logik der Sozialwissenschaften, in: Adorno et al. (1972).

Adorno, T.W. (1972b): Einleitung zu Adorno et al. (1972).

Adorno, T.W. et al. (1972): Der Positivismusstreit in der deutschen Soziologie, Darmstadt und Neuwied: Hermann Luchterhand.

Adressbuch der Stadt Mannheim (1987): Südwestdeutsche Verlagsanstalt, Mannheim.

Adressen-Archiv Otto Conrad GmbH (1988): Adressenkatalog Deutschland, Frankfurt.

Aleanakian, G., Kühnau, W. (1986): Datensicherung bei Personal Computern, Würzburg: Vogel Verlag.

Alemann, H. von (1977): Der Forschungsprozeß. Eine Einführung in die Praxis der empirischen Sozialforschung, Stuttgart: Teubner.

Allerbeck, K. (1987): Einleitung zu Statistisches Bundesamt 1987.

Allerbeck, K., Hoag, W.J. (1985): Integrationsbereitschaft der Jugend im sozialen Wandel, Frankfurt.

Arminger, G., Lijphart, N., Müller, W. (1981): Die Verwendung log-linearer Modelle zur Disaggregierung aggregierter Daten, in: Allgemeines Statistisches Archiv, Bd. 65, S. 273ff.

Atteslander, P., Kopp, M. (1987): Befragung, in: Roth.

Auernhammer, H. (1981): Bundesdatenschutzgesetz. Kommentar (2. Auflage), Köln usw.

Backhaus, K. et al. (1989): Multivariate Analysemethoden. Eine anwendungsorientierte Einführung (5. revidierte Auflage), New York/Heidelberg: Springer.

Bayer, O. (1987): Datenschutz und empirische Forschung: Überraschungen bei der Zusammenarbeit mit "Dritten", in: Sozialwissenschaften und Berufspraxis, Jg. 10/2, Juni, S. 93ff.

Becker, E.E. (1987): Die Einflüsse der neueren technologischen Entwicklung auf das Datenschutzrecht, in: Datenschutz und Datensicherung 5, S. 214ff.

Becker, G. (1982): Der ökonomische Ansatz zur Erklärung menschlichen Verhaltens, Tübingen.

Bellmann, L., Gerlach, K. (1984): Einkommensfunktionen für Frauen und Männer mit individuellen und strukturellen Bestimmungsfaktoren, in: Bellmann, Lutz, Gerlach, Knut, Hübler, Olaf (1984): Lohnstruktur in der Bundesrepublik Deutschland. Zur Theorie und Empirie der Arbeitseinkommen, Frankfurt a.M., New York: Campus.

Bender, S. (1990): De-Anonymisierung von Individualdaten bei statistischen Erhebungen. Eine Diskussion des diskriminanzanalytischen Verfahrens von Paaß, Wauschkuhn (Diplomarbeit, Universität Mannheim).

Benker, A. (1985): Stellungnahme zum Fragenkatalog für die öffentliche Anhörung am 25. Februar 1985 zum Entwurf eines Gesetzes zur Durchführung einer Repräsentativstatistik über die Bevölkerung und den Arbeitsmarkt (Mikrozensusgesetz).

Berekoven, L., Eckert, W., Ellenrieder, P. (1987): Marktforschung. Methodische Grundlagen und praktische Anwendungen, 3. Aufl., Gabler, Wiesbaden.

Berg, W. (1987): Informationelle Selbstbestimmung und Forschungsfreiheit: Zum Spannungsverhältnis zwischen zwei in der Verfassung verankerten Rechten, in: Jehle (1987).

Bergmann, B., Eliasson, G., Orcutt, G. (1980) (Eds.): Micro Simulation - Models, Methods and Applications (IUI Conference Reports), Stockholm: Almqvist & Wiksell.

Bergmann, L. (1981): Datenübermittlung aus dem Melderegister an die Wirtschaft und im öffentlichen Bereich, In: Datenschutz und Datensicherung, Heft 2, S. 79-82.

Bethlehem, J.G., Keller, W.J., Pannekoek, J. (1988): Disclosure Control of Microdata, in: Fourth Annual Research Conference, Proceedings 1988, U.S. Department of Commerce, Bureau of the Census, S. 181 ff.

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

Bick, W., Müller, P.J. (1977): Die Buchführung der Verwaltungen als Datenbasis. in: Müller, P.J. (Hrsg.): Die Analyse prozeß-produzierter Daten, S. 42-88, Stuttgart.

Blien, U. (1986): Unternehmensverhalten und Arbeitsmarktstruktur. Eine Systematik und Kritik wichtiger Beiträge zur Arbeitsmarkttheorie (Beiträge zur Arbeitsmarkt- und Berufsforschung Band 103), Nürnberg.

Blien, U., Graef, F. (1991): "ENTROP: A General Purpose Entropy Optimizing Method for the Estimation of Tables, The Weighting of Samples, the Disaggregation of Data, and the Development of Forecasts", paper presented at the SOFTSTAT '91, it will be published in the conference proceedings, Heidelberg

Blien, U., Papastefanou, G. (1988): Der Informationsbedarf in den Sozialwissenschaften im Spannungsverhältnis zu den Anforderungen des Datenschutzes, in: Computer und Recht 11.

Blien, U., Papastefanou, G. (1989): Einkommensmodellierung mit Daten des Mikrozensus 1985. Ein Vergleich hinsichtlich der Daten-Erfassung Auskunftspflicht versus freiwilliger Auskunftserteilung (unveröffentlichtes Diskussionspapier), Mannheim.

Blien, U., Tessaring, M. (1986): Die Bildungsgesamtrechnung des IAB. Konzeption und erste Ergebnisse, in: Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 4.

Blien, U., Tessaring, M. (1988): Ein Kohortenkonzept zur Analyse von Bildung und Erwerbstätigkeit: Die Bildungsgesamtrechnung des IAB, in: Mertens (1988).

Block, H., Olsson, L. (1976): Bakvägsidentifiering (engl. Version: Backwards-identification (unveröffentlicht)), in: Statistiskal Tidskrift 14, S. 133-144.

Blossfeld, H.-P. (1985): Berufseintritt und Berufsverlauf. Eine Kohortenanalyse über die Bedeutung des ersten Berufs für die Erwerbsbiographie, in: MittAB 18/2, S. 177ff.

Blossfeld, H.P. (1987): Zur Repräsentativität der SfB-3-Lebensverlaufsstudie - Ein Vergleich mit Daten aus der amtlichen Statistik, in: Allgemeines Statistisches Archiv, Jg. 71, S. 125ff.

Bohrnstedt, G.W. (1983): Measurement, in: Rossi, P.H., Wright, J.D., Anderson, A.B. (Eds.): Handbook of Survey Research, Orlando, S. 69-121.

Bohrnstedt, G.W., Mohler, P.P., Müller, W. (Eds.) (1987) : An Empirical Study of the Reliability and Stability of Survey Research Items (Sociological Methods & Research 15/3, Special Issue) Newbury Park usw.: Sage.

Borchert, G. (1981): Datenzugang für die Forschung, In: Öffentliche Verwaltung und Datenverarbeitung, Heft 7-8, S. 18-21.

Bortz, J. (1984): Lehrbuch der empirischen Forschung für Sozialwissenschaftler, New York/Heidelberg: Springer.

Boudon, R. (1980): Die Logik gesellschaftlichen Handelns, Darmstadt, Neuwied.

Bower, R.T., de Gasparis, P. (1978): Ethics in Social Research. Protecting the Interests of Human Subjects, New York: Praeger.

Bradburn N., Sudman S. and Associates (1979): Improving Interview Method and Questionnaire Design. San Francisco: Jossey Bass.

Brand, Marowski, V. (1956): Die Registersachen in der gerichtlichen Praxis, New York/Heidelberg: Springer.

Brennecke, R. (1980): A Sociopolitical Decision and Indicator System for the Federal Republic of Germany: An Overview of SPES, in: Bergmann, Eliasson, Orcutt (1980).

Brennecke, R. (1980a): Kriterien zur Operationalisierung der faktischen Anonymisierung, in Kaase et al., S. 158-175.

Brennecke, R. (1980b): Das Public Use File des SPES-Projektes: Ein Beispiel für eine faktische Anonymisierung, in Kaase et al., S. 176-218.

Brennecke, R./Schneider H. (1977): Zur Problematik des Bundesdatenschutzgesetzes für die Forschung. SPES-Arbeitspapier Nr.63, Sozialpolitische Forschungsgruppe Frankfurt/Mannheim.

Bretz, M., Wedel, E. (1987): Zweck und Bedeutung der Volks-, Berufs-, Gebäude- und Wohnungszählung, in: Wirtschaft und Statistik, 3, S. 195-206.

Bridle, J. (1987): "Adaptive networks and speech pattern processing", in: Devijver P.A., Kittler J. (1987).

Brunnstein, K. (1987): Über die Möglichkeit der Re-Identifikation von Personen aus Volkszählungsdaten, in: Appel, R. (Hrsg.): Vorsicht Volkszählung! Köln: Volksblattverlag, (2.Auflage).

Büllesbach, A. (1981): Datenschutz versus Wissenschaftsfreiheit?, in: Öffentliche Verwaltung und Datenverarbeitung, Heft 3, S. 9-14.

Bundesbeauftragter für den Datenschutz (1981): Vierter Tätigkeitsbericht des Bundesbeauftragten für den Datenschutz gemäß § 19 Abs.2 Satz 2 des Bundesdatenschutzgesetzes (BDSG), in: Bundestagsdrucksache 9/1243, 30.12.1981.

Bunge, M. (1967): Scientific Research, Vol. I und Vol. II, New York/Heidelberg: Springer.

Burkert, H. (1979): Die Eingrenzung des Zusatzwissens als Rettung der Anonymisierung? In: Datenverarbeitung im Recht, Bd. 8, Heft 1, S. 63-75.

Burkert, H. (1980): Das Problem des Zusatzwissens, in: Kaase, M., Krupp, H.-J., Pflanz, M., Scheuch, E.K., Simitis, S. (Hrsg.), Datenzugang und Datenschutz. Königstein/Ts.:Athenäum.

Calingaert, P. (1983): Betriebssysteme aus Benutzersicht, München, Wien.

Cantril, H., Rugg, D. (1976): Die Formulierung von Fragen, in: König (1976).

Carlsson, C. (1984): On the Relevance of Fuzzy Sets in Management Science Methodology, in: Zimmermann, Gaines, Zadeh (1984).

Carroll, J.M. (1987): Computer Security, Boston usw.: Butterworths.

Citteur, C.A.W, Willenborg, Leon C.R.J. (1991): Public Use Files: Current Practices at National Statistical Bureaus, Netherlands Central Bureau of Statistics, 1991,Pros. M 1-79-589, BPA no.: 9487-91-M 1.

Clogg , C.C. (1984): Some Statistical Models for Analyzing why Surveys disagree, in: Turner et al, S. 319-366.

Coleman, J. S. (1990): Foundations of Social Theory, Cambridge: Harvard University Press.

Copas, J.B., Hilton, F.J. (1990): Record Linkage: Statistical Models for Matching Computer Records, in: Journal of the Royal Statistical Society, Volume 153, S. 287 ff.

Cox, L.H. (1980): Suppression Methodology and Statistical Disclosure Control, in: Journal of the American Statistical Association, (6), Vol.75.

Cox, L.H. (1987): The Practice of the Bureau of the Census with the Disclosure of Anonymized Microdata, in: Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik, Band 5 der Schriftenreihe Forum der Bundesstatistik, Stuttgart und Mainz: Kohlhammer.

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

Cox, L.H., McDonald, S.-K., Nelson, D. (1986): Confidentiality Issues at the United States Bureau of the Census, in: Journal of Official Statistics, Vol.2, S. 135 ff.

Cox, L.H., Sande G. (1979): Techniques for preserving statistical confidentiality. Proceedings of the 42nd Meetings of the International Statistical Institute. Manila, December.

Dalenius, T. (1977): Towards a Methodology for Statistical Disclosure Control, in: Statistical Review 5/1977, S. 429ff.

Dalenius, T. (1984): Ends and Means of Privacy Protection in censuses and surveys, Final Report 1984, A research Project supported by Stiftung Volkswagenwerk.

Dalenius, T. (1986): Finding a Needle in a Haystack or Identifying Anonymous Census Records, in: Journal of Official Statistics 2, S. 329-336.

Dalenius, T. (1988): Controlling Invasion of Privacy in Surveys, Continuing Education Program, Statistics Sweden.

Dammann, U. (1981a): Paragraph 6 - Technische und organisatorische Maßnahmen, in: Simitis et al. (1981).

Dammann, U. (1981b): Die falsche Front oder Wissenschaft und Datenschutz, in: Deutsche Universitätszeitung, S. 608ff.

Datenschutzkommission, Rheinland-Pfalz (1987): Informationen zum Datenschutz, Heft 3: Datenschutzrechtliche Anforderungen an wissenschaftliche Forschungsvorhaben, Mainz: Druckerei des Landtags.

Davis, W.S. (1983): Operating Systems. A Systematic View (2. Auflage), Reading, Mass.: Addison Wesley.

Deininger, R. (1960): Repräsentativstatistik der Bevölkerung und des Erwerbslebens, in: Statistisches Bundesamt (1960).

Deininger, R. (1971): Automatic Transcription of Data in Official Statistics, in: Bulletin of the International Statistical Institute, Vol.44, Book 1, (Proceedings of the 38th Session - Invited Papers).

DeMaio, T. J. (1984): Social Desirability and Survey Measurement: A Review, in: Turner, C.F., Martin, E. (eds.), Surveying Subjective Phenomena, Vol.2, S. 257-282.

Denham, C.J. (1987): Census Microdata in Great Britain - The Possibilities, in: Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik, Band 5 der Schriftenreihe Forum der Bundesstatistik, Stuttgart und Mainz: Kohlhammer.

Deutscher Bundestag (1986): BT DRS 10/5345.

Devijver, P.A., Kittler, J. (1987) (Eds.): "Pattern Recognition Theory and Applications", New York/Heidelberg: Springer.

Diller, K.-D. (1988): Opportunitätskosten, in: Das Wirtschaftsstudium, 5/1988.

Dillon, W.R., Goldstein, Matthew (1984): Multivariate Analysis. Methods and Applications, New York usw.

Dittrich, K., Schlörer, J. (1987): Anonymisierung von Forschungsdaten und Idendifikation anonymisierter Datensätze, in: Datenschutz und Datensicherung 1, 1987, S. 30ff.

Dorer, P., Mainusch, H., Tubies, H. (1988): Bundesstatistikgesetz. Gesetz über die Statistik für Bundeszwecke mit den Leitsätzen des Volkszählungsurteils, Mikrozensusgesetz und Volkszählungsgesetz. Kommentar, München: C. H. Beck.

Dorer, P., Wingen, M. (1987): Datenschutz in der amtlichen Statistik und Datenzugang für die Wissenschaft, in: Sozialwissenschaften und Berufspraxis, Jg. 10/2, Juni, S. 102ff.

Dörr, E., Schmidt, D. (1991): Neues Bundesdatenschutzgeset: Handkommentar und Arbeitshilfe für Wirtschaft und Verwaltung. Datakontext-Verlag, Köln.

Drew, C.J. (1980): Introduction to Designing and Conducting Research, St. Louis, Toronto, London: C. V. Mosby.

Droux, R. (1984): Physische EDV-Sicherheit, in: Zimmerli, Liebl (1984).

Duncan, G. (1984): Rasch Measurement in Survey Research: Further Examples and Discussion, in: Turner, C.F., Martin, E. (Eds.): Surveying Subjective Phenomena, Vol.2, New York, S. 367-403.

Duncan, G., Lambert, D. (1986): Disclosure Limited Data Dissemination, in: Journal of the American Statistical Association, Vol.81, S. 10 ff.

Duncan, G., Lambert, D. (1987): The Risk of Disclosure for Microdata, in: Third Annual Research Conference, Proceedings, 1987, U.S. Department of Commerce, Bureau of the Census, S. 263 ff.

Duncan, G., Lambert, D. (1989): The Risk of Disclosure for Microdata, in: Journal of Business and Economic Statistics 7/2, S. 207 - 217.

Eberle, C.-E. (1981): Implikationen des Datenschutzes für die empirische Sozialforschung, in: Zeitschrift für Soziologie.

Edwards, A. L., Walker, J. N. (1961): Social desirability and agreement response set. Journal of Abnormal and Social Psychology 62, S. 108-183.

Eichener, V. (1988): Datenschutz und Forschungsbedarf: Weitergabe und Reidentifizierbarkeit von anonymisierten Individualdaten, in: Sozialwissenschaften und Berufpraxis, 11/1 März, S. 63ff.

Elster, J. (1986) (Eds.): Rational Choice, Oxford: Blackwell.

Endruweit, G. (1983): Ethik für Soziologen?, in: Soziologie 1, S. 45ff.

Eppmann, H. (1989): Alternativen zur Volkszählung, - Eine Untersuchung zur Qualität der Melderegister. In: Statistische Rundschau, Nordrhein-Westfalen, Heft 11, S. 598-611.

Erbslöh, B., Koch, A. (1988): Die Non-Response-Studie zum ALLBUS 1986: Problemstellung, Design, erste Ergebnisse, in: ZUMA Nachrichten 22, S. 29ff.

Esser, H. (1984): Determinanten des Interviewer und Befragtenverhaltens: Probleme der theoretischen Erklärung und empirischen Untersuchung von Intervieweffekten. S. 26-71, in: Mayer, K.U., Schmidt, P. (Hrsg.), Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. Frankfurt/Main, New York. Campus Verlag.

Esser, H. (1984a): Fehler bei der Datenerhebung, Kurseinheit 1: Methodologische Probleme bei der empirischen Kritik von Theorien, Fernuniversität Hagen.

Esser, H. (1984b): Fehler bei der Datenerhebung, Kurseinheit 2: Meßfehler bei der Datenerhebung und die Techniken der empirischen Sozialforschung, Fernuniversität Hagen.

Esser, H. (1984c): Fehler bei der Datenerhebung, Kurseinheit 3: Datenerhebung als sozialer Prozeß, Fernuniversität Hagen.

Esser, H. (1984d): Fehler bei der Datenerhebung, Kurseinheit 4: Meßfehler in Kausalmodellen, Fernuniversität Hagen.

Esser, H. (1986): Können Befragte lügen? Zum Konzept des "wahren Wertes" im Rahmen der handlungstheoretischen Erklärung von Situationseinflüssen bei der Befragung. Kölner Zeitschrift für Soziologie und Sozialpsychologie 37, S. 314-336.

Esser, H., Grohmann, H., Müller, W., Schäffer, K.-A. (1989): Mikrozensus im Wandel. Untersuchungen und Empfehlungen zur inhaltlichen und methodischen Gestaltung (Band 11 der Schriftenreihe Forum der Bundesstatistik), Metzler-Poeschel Stuttgart.

Euler, M. (1987): Lieferung anonymisierter Einzelangaben aus der Einkommens- und Verbrauchsstichprobe (EVS), in: Statistisches Bundesamt (1987).

Fahrmeir, L., Hamerle, A. (1984) (Hrsg.): Multivariate statistische Verfahren, Berlin.

Fahrmeir, L., Häußler, W., Tutz, G. (1984): Diskriminanzanalyse, in: Fahrmeir, L., Hamerle, A. (Hrsg.): Multivariate statistische Verfahren, Berlin.

Fälker, G. (1980): Eine praktische Lösung der BDSG-Datensicherungsaufgabe, in: Datenschutz und Datensicherung 1, S. 30ff.

Faulbaum, F. (1987): Intergroup Comparisons of Latent Means Across Waves, in: Sociological Methods & Research, Vol.15, No.3, S. 317 - 335.

Fellegi, J.P. (1972): On the Question of Statistical Confidentiality, in: Journal of the American Statistical Association, Vol.67, 1972, S. 7 ff.

Fellegi, J.P., Phillips, J.L. (1974): Statistical confidentiality: some theory and applications to data dissemination, in: Annals of Economic and Social Measurement 3+2, S. 399ff.

Fienberg, S.E. (1980): The Analysis of Cross-Classified Categorical Data, Cambridge (Mass.): MIT-Press.

Fischer-Hübner, S. (1986): Zur Anonymität und Reidentifizierbarkeit statistischer Daten, Mitteilungen Nr. 143 des Fachbereichs Informatik der Universität Hamburg.

Fougelman S., F. et al. (1987): "Learning and associative memory", in: Devijver, Kittler (1987)

Friedrichs, J. (1980): Methoden empirischer Sozialforschung, Reinbek b. Hamburg: Rowohlt.

Fröschl, K. A. (1990): Automated Protocolling of Statistical Data Analyses, in: Faulbaum, F., Haux, R., Jöckel, K.-H. (Hrsg.): Softstat'89. Fortschritte der Statistik-Software 2, Stuttgart, New York.

Fuchs, W. (1984): Biographische Forschung. Eine Einführung in Praxis und Methoden, Opladen: Westdeutscher Verlag.

Gates, G.W. (1988): Census Bureau Microdata: Providing Useful Research Data While Protecting the Anonymity of Respondents, Proceedings of the Social Section of the American Statistical Association, S. 235 ff.

Graef, F., Blien, U. (1989): ENTROP - Ein allgemein einsetzbares Verfahren zur Gewichtung von Stichproben, zur Disaggregierung von Daten und zur Ermittlung von Tabellen aus heterogenen Informationen, in: Allgemeines Statistisches Archiv 2, S. 122ff.

Greenberg, B. (1989): Discussion, zu den Vorträgen zur Disclosure Avoidance, in: Fifth Annual Research Conference, Proceedings, 1989, U.S. Department of Commerce, Bureau of the Census, S. 456 ff.

Greenberg, B. (1990): Disclosure Avoidance Research at the Census Bureau, paper presented at the 1990 Annual Research Conference, Bureau of the Census, Arlington, Virginia, S. 144 ff.

Greenberg, B., Voshell, L. (1990): The Geographic Component of Disclosure Risk for Mikrodata, Bureau of the Census, Statistical Research Division Report Series, SRD Research Report Number: Census/SRD/RR-90/13.

Greiser, E. (1980): Epidemiologische Forschung und Behinderung des Datenzugangs. in: Kaase et al., S. 77-82.

Grohmann, H. (1989): Wozu brauchen Staat und Öffentlichkeit Daten? in: Aus Politik und Zeitgeschichte, Beilage zur Wochenzeitschrift "Das Parlament", B 19/89, Bonn.

Grossberg, S. (1988) (Eds.): Neural Networks and Natural Intelligence, Cambridge, Mass., London: MIT Press/ Bradford Books.

Guchteneire, P. de, Mochmann, E. (Eds.) (1990): Data protection and data access, 1990, Amsterdam: North-Holland.

Gyarfas, G. (1988): Konstruktion einer synthetischen Mikrodatenbasis für die Simulation der Einkommensteuer. Arbeitspapiere der GMD, Nr. 325, St. Augustin.

Hailer, G. (1980): Datensicherung als Orgware-Aufgabe, in: Datenschutz und Datensicherung, Heft 3, S. 157ff.

Hamacher, B. (1980): Resümee zu Datenschutzmaßnahmen, in Kaase et al., S. 219-224.

Hammerbacher, H. (1982): Die Anonymisierung personenbezogener Daten, - Zusammenhänge, Begriffe, Maßnahmen, Ergebnisse - In: Datenschutz und Datensicherung, Heft 4, S. 284-292.

Hammerbacher, H. (1984): Die zu fordernde "ausreichende" Anonymisierung von Datensätzen, in: Datenschutz und Datensicherung, Heft 3, S. 181ff.

Hartmann, P.H. (1990): Wie repräsentativ sind Bevölkerungsumfragen? Ein Vergleich des ALLBUS und des Mikrozensus. in: ZUMA Nachrichten No. 26, S. 7-30.

Heer, G., Schimpl-Neimanns, B. (1991): Probleme bei der Analyse zufallsüberlagerter Fallzahltabellen am Beispiel von Mikrozensusauswertungen mit STATIS-BUND. Vortrag auf der Softstat '91 in Heidelberg.

Heidenreich, K. (1987): Grundbegriffe der Meß- und Testtheorie, in: Roth (1987).

Heinz, W. (1987): Datenschutz bei automatisierter Datenverarbeitung in der kriminologischen Forschung, in: Jehle (1987).

Heinz, W. (o. J.): Sozialwissenschaftliche Forschung und Datenschutz. Ein Forschungsvorhaben über technische und organisatorische Maßnahmen zur Gewährleistung eines Optimums an Sicherungsmaßnahmen für personenbezogene Daten, o. O.

Herberger, L. (1985): Aktualität und Genauigkeit der repräsentativen Statistik der Bevölkerung und des Erwerbslebens, in: Allgemeines Statistisches Archiv 69, S. 16ff.

Hinton, G.E., Sejnovski, T.J. (1986): Learning and Relearning in Boltzmann Machines, in: Rumelhart, McClelland (1986).

Hippler, H.-J., Schwarz, N., Sudman, S. (1987) (Eds.): Social Information Processing and Survey Methodology, New York/Heidelberg: Springer.

Hohmann, H. (1987) (Hrsg.): Freiheitssicherung durch Datenschutz, Frankfurt a.M.: Suhrkamp.

Iversen, G.R. (1984): Bayesian statistical Inference, Beverly Hills, Sage.

Jagodzinski, W., Kühnel, S. (1987): Estimation of Reliability and Stability in Single-Indicator Multiple-Wave Models, Sociological Methods and Research, Vol.15, No.3, S. 219-258.

Jain, A.K. (1987): Advances in statistical pattern recognition, in: Devijver, Kittler (1987).

Jaynes, E.T. (1979): Where Do We stand on Maximum Entropy?, in: Levine, Tribus (1979).

Jehle, J.-M. (1987) (Hrsg.): Datenzugang und Datenschutz in der kriminologischen Forschung, Wiesbaden: Kriminolog. Zentralstelle e.V.

Jones, D.H., Adam, N.R. (1989): Disclosure Avoidance Using the Bootstrap and other Re-sampling Schemes, in: Fifth Annual Research Conference, Proceedings 1989, U.S. Department of Commerce, Bureau of the Census, S. 446 ff.

Jöreskog, K.G., Sörbom, S. (1979): Advances in Factor Analysis and Structural Equation Models, Cambridge, Mass: Abt. Books.

Kaase, M. (1987): Datenschutz und sozialwissenschaftliche Forschung, in: Jehle (1987).

Kaase, M., Krupp, H.-J., Pflanz, M., Scheuch, E.K., Simitis, S. (Hrsg.) (1980): Datenzugang und Datenschutz. Königstein/Ts.:Athenäum.

Karl, F. (1976): Deutsches Registerrecht, Duncker & Humblot. Berlin.

Kauffels, F.-J. (1985): Schwachstellen der Informationssicherheit in lokalen Netzen, in: Spies (1985).

Keidel, Schmatz, Stöber (1985): Handbuch der Rechtspraxis, Bd. 7: Registerrecht. Beck'sche Verlagsbuchhandlung. München.

Keller, W.J., Bethlehem, J.G. (1987): Disclosure Protection of Microdata, Netherlands Central Bureau of Statistics, Proj M 1-79-589 second version, BPA no.: 1582-87 M 1.

Keller, W.J., Bethlehem, J.G. (1990): Disclosure Protection of Microdata: Problems and Solutions, Netherlands Central Bureau of Statistics, Project M3-SI-Algemeen, BPA no.: 15690-90-M 3.

Kerkau, H.-J. (1985): Hat die Datenverarbeitung das Datenschutzrecht überholt? - Überlegungen aus der Sicht der Kontrollpraxis, in: Spies (1985).

Kim, J. (1990): Masking Microdata for National Opinion Research Center, Final Project Report, Bureau of the Census.

Kirschner, H.-P. (1987): Stichprobenverfahren und Auswahlsätze als Mittel der Anonymisierung, in: Statistisches Bundesamt (1987), S. 113-120.

Koch, A. (1986): Wie zuverlässig lassen sich Berufs- und Bildungsvariablen messen? Ergebnisse einer Test-Retest-Studie zur Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften 1984, unveröffentlichte Diplom-Arbeit, Universität Mannheim.

König, R. (1976) (Hrsg.): Das Interview. Formen-Technik-Auswertung, Köln: Kiepenheuer & Witsch.

Kooiman, P., Willenborg, L.C.R.J. (1990): Assessing Disclosure Risks for Microdata, Netherlands Central Bureau of Statistics, Pros M 1-79-589, BPA no.: 15832-90-M 1.

Kortmann, K. (1982): Verknüpfung und Ableitung personen- und haushaltsbezogener Mikrodaten, Frankfurt a. M., New York: Campus.

Kortmann, K., Schmaus, G. (1976): Generierung des Mikrodatenfiles 1969 für die Bundesrepublik Deutschland (IMDAF 69), in: Nachrichtendienst für die öffentliche und private Fürsorge 5, S. 144ff.

Krebs, D., Schmidt, P. (Eds.) (in Druck): Relevance of Attitude Measurement in Social Science. New York/Heidelberg: Springer

Kriz, J. (1981): Methodenkritik empirischer Sozialforschung: Stuttgart.

Krug, W. (1976): Quantifizierung des systematischen Fehlers in wirtschafts- und sozialstatistischen Daten. Dargestellt an der Statistik der Erwerbstätigkeit, Berlin: Duncker und Humblot.

Krug, W., Nourney, M. (1982): Wirtschafts- und Sozialstatistik: Gewinnung von Daten, München, Wien: Oldenbourg.

Krupp, H.-J. (1980): Der Datenbedarf in der quantitativen Ökonomie, in: Kaase et al., S. 1-9.

Krupp, H.-J. (1982): Empirische Sozialwissenschaft und amtliche Statistik aus der Sicht der sozialwissenschaftlichen Politikberatung. Sonderdruck der Referate zur 29. Tagung des Statistischen Beirates. Beilage zu Wirtschaft und Statistik.

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

Kruse, H., Mangold, R., Mechler, B., Penger, O. (1991): Programmierung Neuronaler Netze, Bonn usw.: Addison-Wesley .

Küchler, M. (1983): Wissenschaftliche Arbeit und publizistische Vermarktung. Überlegungen zu einem Spannungsverhältnis oder: Brauchen wir einen "Code of Ethics"?, in: Soziologie 1, S. 49ff.

Kühn, J. (1987): Automatisierte Anonymisierungsverfahren für Kurzbandsätze, in: Statistisches Bundesamt (Hrsg.), Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik, Band 5 der Schriftenreihe: Forum der Bundesstatistik, Stuttgart und Mainz: Kohlhammer.

Kühn, J., Pfrommer, F., Schrey, E. (1984): Zur technischen Weiterentwicklung des Statistischen Informationssystems, in: Wirtschaft und Statistik, 12/1984, S. 981 ff.

Kühn, J., Radtke, M. (1987): Automated Anonymization Procedures for Short Tape Records, International Statististical Institute (ISI), 46th Session, Contributed Papers, Tokio 1987, S. 245 ff.

Kühn, J., Radtke, M. (1988): Automated Anonymization Procedures for Short Tape Records, Federal Statistical Office, Wiesbaden.

Kürschners Deutscher Gelehrtenkalender (1976, 1987): DeGruyter. Berlin.

Lamnek, S. (1988): Qualitative Sozialforschung, Band 1: Methodologie, München, Weinheim.

Lamnek, S. (1989): Qualitative Sozialforschung, Band 2: Methoden und Techniken, München, Weinheim.

Lang, M. (1989): Sicherung von PC und vernetzter Datenverarbeitung vornehmlich gegen Mißbrauch, in: Datenschutz und Datensicherung 10, S. 505-511.

Lessler, J.T. (1984): Measurement Error in Surveys, in Turner et al, S. 405-440.

Leue, J. (1986): Einsichtsrechte in öffentliche Register, In: Vollkommer, M., Datenverarbeitung und Persönlichkeitsschutz, S. 83-105. Universitätsbund Erlangen-Nürnberg e.V. Erlangen.

Levine, R.D., Tribus, M. (1979) (Eds.): The Maximum Entropy Formalism, Cambridge (Mass.), London: MIT-Press.

Lietmeyer, V. (1986): Microanalytic Tax Simulation Models in Europe: Development and Experience in the German Federal Ministry of Finance, in Orcutt, Merz, Quinke (1986).

Little, R.J.A., Rubin, D.B. (1987): Statistical Analysis With Missing Data, New York (Wiley).

Litz, H.-P., Litowatz, T. (1986): "Amtliche Statistik in marktwirtschaftlich organisierten Industriegesellschaften", Frankfurt a. M., New York: Campus

Lockemann, P.C., Schmidt, J.W. (1987) (Hrsg.): Datenbank-Handbuch, New York/Heidelberg: Springer.

Lord, F.M., Novick, M.R. (1974): Statistical Theories of Mental Test Scores (2. Auflage), London: Addison-Wesley.

Madow, W. G., Olkin, I., Rubin, D.B. (1983) (Eds.): "Incomplete Data in Sample Surveys, Vol. 2: Theory and Bibliographics", New York usw.: Academic Press

Mangold, W. (1980): Selbstregulierung der Forschung als Teil des Datenschutzes und Voraussetzung für Datenzugang, in: Kaase et al. (1980).

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

Mayer, K.U., Blossfeld P. (1990): Die gesellschaftliche Konstruktion sozialer Ungleichheiten im Lebensverlauf, in: Berger, P.A., Hradil, S. (Hrsg.), Sonderband der Sozialen Welt "Das Ende der sozialen Schichtung?", Göttingen, S. 297-318.

McClelland, J.E., Rumelhart, D.E. et al. (1986) (Eds.): Parallel distributed Processing. Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models, Cambridge (Mass.), London: MIT Press.

McClelland, J.L., Rumelhard, D.E. (1988): Explorations in Parallel Distributed Processing. A Handbook of Models, Programs, and Exercises: MIT Press/ Bradford Books .

Mertens, D. (1988) (Hrsg.): Konzepte der Arbeitsmarkt- und Berufsforschung. Eine Forschungsinventur des IAB (Beiträge zur Arbeitsmarkt- und Berufsforschung Band 70, Neuauflage), Nürnberg.

Miehle, N., Tubies, H. (1980): Bedeutung von Sicherheitssoftware, in: Datenschutz und Datensicherung 4, S. 197ff.

Mohler, P., Ph. (1987): Bewährt, aber noch nicht ausgereift - Bemerkungen zum Datenschutz in der Empirischen Sozialforschung, in: Sozialwissenschaften und Berufspraxis, Jg. 10/2, Juni, S. 128ff.

Mohler, P., Ph., Kaase, M. (1980): Formen der Erhebung in der empirischen Sozialforschung, in: Kaase et al. (1980).

Mokken, R.J., Pannekoek, J., Willenborg, L. (1990): Microdata and Disclosure Risks, paper presented at the U. S. Census Bureau annual conference, Washington D.C.

Müller, G.F. (1983): Umsetzung der Datenschutzforderung in Datensicherungsmaßnahmnen, in: Datenschutz und Datensicherung, Heft 4, S. 289ff.

Müller, M. (1991): Reidentifikation von Individualdaten: Experimentelle Überprüfung im Rahmen eines sozialwissenschaftlichen Szenarios mit der Methode von Paaß, Wauschkuhn (Diplomarbeit, Universität Mannheim).

Müller, W. (1978): Klassenlage und Lebenslauf. Untersuchungen zu Prozessen sozialstrukturellen Wandels in der Bundesrepublik Deutschland (unveröffentlichte Habilitationsschrift), Mannheim.

Müller, W. (1982): Empirische Sozialwissenschaft und amtliche Statistik aus der Sicht der empirisch orientierten Forschung. Sonderdruck der Referate zur 29. Tagung des Statistischen Beirates. Beilage zu Wirtschaft und Statistik.

Müller, W., Hauser, R. (1987): Der Bedarf der Wissenschaft an anonymisierten Einzelangaben, in: Statistisches Bundesamt (1987).

Neue juristische Wochenschrift (1984): Urteil des Bundesverfassungsgerichts zum Volkszählungsgesetz (8), S. 419-428.

Niemann, H. (1981): Pattern Analysis, New York/Heidelberg: Springer.

Niemann, H. (1983): Klassifikation von Mustern, New York/Heidelberg: Springer.

Okunda, T., Tanaka, H., Asai, K. (1984): Discrimination Problem with Fuzzy States and Fuzzy Information, in: Zimmermann, Zadeh, Gaines (1984).

Opp, K.-D. (1976): "Methodologie der Sozialwissenschaften. Einführung in die Probleme ihrer Theoriebildung", Reinbek b. Hamburg: Rowohlt

Orcutt, G., Merz, J., Quinke, H. (1986) (Eds.): Microanalytic Simulation Models to Support Social and Financial Policy, Amsterdam: North-Holland .

Paaß, G (1987): Re-Identifikationsrisiko von Einzelangaben, in: Statistisches Bundesamt (Hrsg.): Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik. Bedingungen und Möglichkeiten, Stuttgart, Mainz: Kohlhammer.

Paaß, G. (1985a): Anonymität statistischer Einzelangaben, in: Datenschutz und Datensicherung Nr. 2, (1985), S. 97ff.

Paaß, G. (1985b): Anonymität von Individualdaten in statistischen Datenbanken, in: Spies (1985) .

Paaß, G. (1986): Statistical match: evaluation of existing procedures and improvements by using additional information, in: Orcutt, Merz, Quinke (1986).

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

Paaß, G. (1988b): Stochastic Generation of a Synthetic Sample from Marginal Information, Gesellschaft für Mathematik und Datenverarbeitung, Arbeitspapiere der GMD, Nr. 308, St. Augustin .

Paaß, G. (1989): Structured probabilistic neural networks, in: Proceedings of the Neuro-Nimes 1989, S. 345-359 .

Paaß, G., Wauschkuhn, U. (1985): Datenzugang, Datenschutz und Anonymisierung. Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten, München, Wien: R. Oldenbourg.

Pfeifer, G., Strickert, H.-G. (1961): Personenstandsgesetz, Schweitzer Verlag. Berlin.

Pflanz, M., Greiser, E. (1980): Datenbedarf für die Epidemiologie, in: Kaase et al. (1980).

Phillips, D.L., Clancy, K.J. (1972): Some Effects of "Social Desirability" in Survey Studies. American Journal of Sociology 77, S. 921-940.

Piller, E., Weißenbrunner, A. (1986): Software-Schutz. Rechtliche, organisatorische und technische Maßnahmen, New York/Heidelberg: Springer.

Popper, K.R. (1966): Logik der Forschung, Tübingen.

Popper, K.R. (1972): Die Logik der Sozialwissenschaften, in: Adorno, T.W. u.a. (Hrsg.): Der Positivismusstreit in der deutschen Soziologie, Neuwied/Berlin.

Porst, R. (1985): Praxis der Umfrageforschung, Teubner. Stuttgart.

Porst, R., Schmidt, P., Zeifang, K. (1987): Comparisons of Subgroups by Models with Multiple Indicators. In: Sociological Methods & Research, Vol.15, No.3, S. 303-316.

Porst, R., Zeifang, K. (1987a): Wie stabil sind Umfragedaten? Beschreibung und erste Ergebnisse der Test-Retest-Studie zum ALLBUS 1984, in: ZUMA Nachrichten 20, S. 8-31.

Porst, R., Zeifang, K. (1987b): A Description of the German General Social Survey Test-Retest Study and a Report on the Stabilities of the Sociodemographic Variables. In: Sociological Methods & Research, Vol.15, No. 3, S. 177-218.

Pütter, P.S. (1981): Datensicherung - ihre Realisierung und Kontrolle, in: ÖVD 5, S. 3ff.

Radtke, M. (1988): Anonymisierung eines Mikrodatenfiles durch Generierung synthetischer Sätze, in: Allgemeines Statistisches Archiv, Band 72, 1988, S. 399 ff.

Rainwater, L. (1990): Changing in Equalities Structures in Europe: The Challenge to Social Science. Conference on Social Science Perspectives in Western Europe. Kronberg Germany.

Rao, J.N.K., Wu, C.F.J. (1988): Resampling Inference with Complex Survey Data, in: Journal of the American Statistical Association, Volume 83, 1988, S. 231 ff.

Rauterberg, M. (1984): Wie anonym sind anonymisierte Daten wirklich?, in: Datenschutz und Datensicherheit 3, S. 174ff.

Redfern, P. (1986): Studie über die Zukunft der Volkszählung: Alternative Ansätze.

Rendtel, U. (1990): Teilnahmebereitschaft in Panelstudien: Zwischen Beeinflussung, Vertrauensbildung und sozialer Selektion, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie 42/2, S. 280-299

Reuter, A. (1987): Maßnahmen zur Wahrung von Sicherheits- und Integritätsbedingungen, in: Lockemann, Schmidt (1987).

Rihaczek, K. (1980): Die Verwendung kryptographischer Verfahren, in: Datenschutz und Datensicherung 2, S. 99ff.

Rihaczek, K. (1985): Datenmißbrauch: Verhindern besser als verbieten, in: Spies (1985).

Rihaczek, K. (1987a): Anonym, Anonymisieren, in: Datenschutz und Datensicherung 3, S. 137f.

Rihaczek, K. (1987b): Ein Kompromißvorschlag zur Datenverschlüsselung, in: Datenschutz und Datensicherung 6, S. 299ff.

Roethlisberger, D. (1939): Management and the Worker.

Roth, E. (Hrsg.) (1987): Sozialwissenschaftliche Methoden. Lehr- und Handbuch für die Forschung und Praxis, München, Wien: Oldenbourg.

Rothe, G., Wiedenbeck, M. (1987): Stichprobengewichtung: Ist Repräsentativität machbar?, in: ZUMA Nachrichten 21, S. 43ff.

Rubin D.B. (1987): Multiple Imputations for Nonresponse in Surveys, New York (Wiley).

Rudolph, H. (1986): Die Fluktuation in sozialversicherungspflichtiger Beschäftigung. Erste Ergebnisse aus der Beschäftigtenstichprobe des IAB, in: Mitteilungen aus der Arbeitsmarkt- und Berufsforschung, 19. Jg./2, S. 257ff.

Rumelhart, D.E., McClelland, J.L. et al. (1986) (Eds.): Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Vol. 1: Foundations, Cambridge (Mass.), London: MIT Press .

Scherer, J. (1980): Datenzugang des Forschers unter dem amerikanischen Freedom of Information Act, in: Kaase et al. (1980).

Scherer, J. (1980): Datenzugang des Forschers zwischen Informationsanspruch und Geheimhaltungsgrundsatz - Das Beispiel des Bildungsforschers, in: Kaase et al. (1980)

Scheuch, E.K. (1980): Die Weiterentwicklung des Datenschutzes als Problem der Sozialforschung, in: Kaase et al., S. 252-275.

Scheuch, E.K. (1987): Risikointerpretation beim Datenschutz, in Statistisches Bundesamt 1987, S. 121-145.

Scheuch, E.K., Graef, L., Kühnel S. (1989): Volkszählung, Volkszählungsprotest und Bürgerverhalten - Ergebnisse der Begleituntersuchung zur Volkszählung 1987 - . Band 12 der Schriftenreihe Forum der Bundesstatistik, Metzler-Poeschel Stuttgart.

Schimpl-Neimanns, B. (1991): Zur Repräsentativität soziodemographischer Merkmale des ALLBUS im Vergleich mit dem Mikrozensus. Vortrag bei der Sektion Methoden der DGS Frühjahrsitzung.

Schlörer, J. (1977): Confidentiality and Security in Statistical Data Banks, in: Gaus, W., Henzler, R. (Eds.): Data Documentation: Some Principles and Applications in Science and Industry. München, Verl. Dok., S. 103-123.

Schlörer, J. (1978a): Disclosure from Statistical Databases: Quantitative Aspects of trackers. ACM Trans. Database System.

Schlörer, J. (1978b): Security of Statistical Databases: Multidimensional Transformation - I.Transformable Matrices, TB-IMSD 2, Inst. med. Statist. Doc., Universität Gießen, August.

Schlörer, J. (1980): Anonymisierung von Mikrodaten: Technische Aspekte, in: Kaase et al. (1980).

Schnell, R. (1991): Wer ist das Volk? Zur faktischen Grundgesamtheit bei "allgemeinen Bevölkerungsumfragen": Undercoverage, Schwerereichbare und Nichtbefragbare. Kölner Zeitschrift für Soziologie und Sozialpsychologie 1/91, S. 106-137.

Schnell, R., Hill, P.B., Esser, E. (1988): Methoden der empirischen Sozialforschung, München, Wien: Oldenbourg.

Schönberg, V. (1984): Organisatorische und softwaregestützte EDV-Sicherheit, in: Zimmerli, Liebl (1984).

Schöneburg, E. et al. (1990): "Neuronale Netzwerke", Haar bei München, Markt und Technik.

Schuessler, K., Hittle D., Cardaseia, J. (1978): Measuring responding desirably with attitude-opinion items, in: Social Psychology 41, S. 224-235.

Schwartz, J.E. (195): The neglected Problem of measurement Error in Categorical Data. Sociological Methods and Research 13, S. 435-466.

Schwarz, N. (1990): Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Costruction, ZUMA-Arbeitsbericht 1988/10, abgedruckt in: Hendricks, C., Clark, M. (Eds.): Research Methodology (Review of Personality and Social Psychology, Vol. 11), Beverly Hills: Sage.

Schwarz, N., Hippler, H.-J., Noelle-Neumann, E. (1989): Response Order Effects in Long Lists: Primacy, Recency, and Asymmetric Contrasts Effects, Mannheim, ZUMA-Arbeitsberichte Nr. 89/18.

Schwarz, N., Hippler, H.-J., Noelle-Neumann, E. (1989): Einflüsse der Reihenfolge von Antwortvorgaben bei geschlossenen Fragen, in: ZUMA Nachrichten 25, S. 24-38.

Schwarz, N., Hippler, H.-J., Strack, F. (1988): Kognition und Umfrageforschung: Themen, Ergebnisse und Perspektiven, in: ZUMA Nachrichten 22, S. 15-28.

Shannon, C.E. (1976): Die mathematische Theorie der Kommunikation, in: Shannon, C.E., Weaver, W.: Mathematische Grundlagen der Informationstheorie, München, Wien: R. Oldenbourg.

Shannon, C.E., Weaver, W. (1976): Mathematische Grundlagen der Informationstheorie, München, Wien: R. Oldenbourg.

Siebter Tätigkeitsbericht der Landesbeauftragten für den Datenschutz in Baden-Württemberg 1986: Datenschutz für unsere Bürger, Stuttgart.

Simitis, S. (1980): Datenschutz und Wissenschaftsfreiheit, in: Kaase et al., S. 83-91.

Simitis, S. (1981): Einleitung, in: Simitis et al. (1981).

Simitis, S. (1987): Datenschutz und kriminologische Forschung, in: Jehle (Hrsg.), S. 47-63.

Simitis, S. et al. (1981): Kommentar zum Bundesdatenschutzgesetz (3. Auflage), Baden-Baden: Nomos Verlagsgesellschaft.

Singer, E. (1978): Informed Consent: Consequences for Response Rate and Response Quality in Social Surveys, in: American Sociological Review, 43, S. 144-162.

Skinner, C.J. (1990): Statistical Disclosure Issues for Census Microdata, Paper presented at the International Symposium on Statistical Disclosure Avoidance, Voorburg, Niederlande 1990.

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

Smith, T.W. (1984): Nonattitudes: A Review and Evaluation, in: Turner, C.F., Martin, E. (eds.), Surveying Subjective Phenomena, Vol.1, S. 215-255.

Smithson, M. (1987): Fuzzy Set Analysis for Behavioral and Social Sciences, New York/Heidelberg: Springer.

Spies, P.P. (1985) (Hrsg.): Datenschutz und Datensicherung im Wandel der Informationstechnologien, New York/Heidelberg: Springer.

Spruill, N. (1983): Testing Confidentiality of Masked Business Microdata, Working Paper, PRI 83-07.09, The Public Research Institute, Alexandria, Virginia .

Spruill, N.L. (1983): The Confidentiality and Analytic Usefulness of Masked Business Microdata, in: Proceedings of the Section on Survey Research Methods, American Statistical Association, S. 602 ff.

Stadler, N. (1980): Datensicherung durch Organisation. Voraussetzungen des Datenschutzes (2. Auflage), Freiburg i. Br.

Statistisches Bundesamt (1960) (Hrsg.): Stichproben in der amtlichen Statistik, Wiesbaden.

Statistisches Bundesamt (1978): Volkszählung vom 27. Mai 1970, Untersuchungen zur Methode und Genauigkeit der Volkszählung 1970, Fachserie 1, Heft 26: Wiesbaden.

Statistisches Bundesamt (1987) (Hrsg.): Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik. Bedingungen und Möglichkeiten, Stuttgart, Mainz: Kohlhammer.

Statistisches Bundesamt (1989): Datenreport 1989. Zahlen und Fakten über die Bundesrepublik Deutschland.

Statistisches Bundesamt (1990): Arbeitsunterlagen zum Mikrozensus. Das Erhebungsprogramm des Mikrozensus seit 1957.

Steinmüller, W. (1980): Ein organisationsunterstütztes Verfahren zur Anonymisierung von Forschungsdaten, in: Kaase et al. (1980).

Steinmüller, W. (1987): Betroffenenschutz bei offenen Netzen, in: Hohmann (1987).

Steinmüller, W. (1988a) (Hrsg.): Verdatet und vernetzt. Sozialökologische Handlungsspielräume in der Informationsgesellschaft, Frankfurt a. M.: Fischer.

Steinmüller, W. (1988b): Demokratische und soziale Informationstechnologiepolitik, in: Steinmüller (1988a).

Struif, B. (1987): Mit der Chipkarte ins BS2000, in: Datenschutz und Datensicherung 3, S. 122ff.

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

Sudman, S., Bradburn, N. (1974): Response Effects in Surveys. Chicago: Aldine Publishing Company.

Szameitat, K., Deininger, R. (1969): Some Remarks on the Problem of Errors in Statistical Results, in: Bulletin of the International Statistical Institute, Vol.42, Book 1.

Szyperski, N., Grochla, E., Homberger, H.-J. (1982) (Hrsg.): Datensicherung für die betriebliche Praxis, Ein Maßnahmenkatalog von der Zugangs- bis zur Organisationskontrolle.

Taschenbuch des öffentlichen Lebens (1983/84): Festland. Bonn.

Teske, W. (1986): Der Zugang zu den Datenbanken der Wirtschaft, In: Vollkommer, M., Datenverarbeitung und Persönlichkeitsschutz, S. 105-131. Universitätsbund Erlangen-Nürnberg e.V. Erlangen.

Tessaring, M. (1988): Arbeitslosigkeit, Beschäftigung und Qualifikation: Ein Rück- und Ausblick, in: Mitteilungen aus der Arbeitsmarkt- und Berufsforschung 21/2, S. 177ff.

Tessaring, M., Blien, U., Fischer, G., Hofmann, I., Reinberg, A. (1990): Bildung und Beschäftigung im Wandel. Die Bildungsgesamtrechnung des IAB (Beiträge zur Arbeitsmarkt- und Berufsforschung 126), Nürnberg: Bundesanstalt für Arbeit.

Theil, H. (1967): Economics and Information Theory, Amsterdam.

Thome, R. (1979): Datenschutz, München: Franz Vahlen.

Titterington, D.M., Murray, G.D., Murray, L.S.; Spiegelhalter, D.J., Skene, A.M., Habbema, J.D.F., Gelpke, G.J. (1981): Comparison of Discrimination Techniques Apllied to a Complex Data Set of Head Injured Patients. J.R.Statist.Soc. A144, Part 2, S. 142-175.

U.S. Department of Commerce (1978): Report on Statistical Disclosure and Disclosure-Avoidance Techniques, Statistical Policy Working Paper 2, Office of Federal Statistical Policy and Standards, Washington, D.C.: Government Printing Office.

Uhlig, S. (1987): Justizdaten aus dem Bundeszentralregister und dem Gewerbezentralregister für die Forschung, In: Jehle, J.-M. (Hrsg.): Datenzugang und Datenschutz in der kriminologischen Forschung. Kriminologische Zentralstelle e.V. Wiesbaden.

Waidner, M., Pfitzmann, B., Pfitzmann, A. (1987): Über die Notwendigkeit genormter kryptographischer Verfahren, in: Datenschutz und Datensicherung 6, S. 293ff.

Weber, M. (1980): Wirtschaft und Gesellschaft. Grundriss der verstehenden Soziologie, Fünfte Auflage, Tübingen: J.C.B. Mohr (Paul Siebeck) .

Weck, G. (1984): Datensicherheit, Stuttgart: Teubner.

Wedel, E. (1989): Haushalten 1987 - Methode und Ergebnis der Volkszählung, in Wirtschaft und Statistik (5), S. 273-276.

Wegener, B. (1983): Theoretische Einleitung: Wer skaliert? Die Meßfehler-Testtheorie und die Frage nach dem Akteur, in: ZUMA-Handbuch Sozialwissenschaftlicher Skalen (Wissenschaftliche Bearbeitung: Allmendinger, J., Schmidt, P., Wegener, B.,. Dokumentarische Bearbeitung: Eickelmann, T., Ohly, P.), Bonn.

Weigert, K. (1980): Auskünfte aus dem Melderegister nach dem Meldegesetz, In: Datenschutz und Datensicherung, Heft 4, S. 191-196.

Weise, K.-Th. (1981): Zur Organisation der Datensicherung, in: Datenschutz und Datensicherung, S. 34ff.

Werhahn, P. (1911): Die privatrechtliche Bedeutung der öffentlichen Bücher und Register, Dissertation. Rößler & Herbert. Heidelberg.

Weyer, H., Pütter, P.S. (1983): Organisation und Technik der Datensicherung. Empfehlungen aus der Kontrollpraxis, Köln: Datakontext.

White, H. (1989): Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models, in: Journal of the American Statistical Association, S. 1003 ff.

Who is who in Germany (1974): Who is who Publishing GmbH. München.

Willenborg, L.C.R.J. (1990): Disclosure Risks for Microdata Sets: Stratified Populations and Multiple Investigators, Netherlands Central Bureau of Statistics, Pros. M 1-79-589, BPA no.: 10088-90 M 1.

Willenborg, L.C.R.J. (1990): Remarks on Disclosure Control of Microdata, Netherlands Central Bureau of Statistics.

Willenborg, L.C.R.J., Mokken, R.J., Pannekoek, J. (1990): Microdata and Disclosure Risks, Annual Research Conference, Proceedings 1990, U.S. Department of Commerce, Bureau of the Census, S. 167 ff.

Wilson, A. (1970): Entropy in Urban and Regional Modelling, London: Pion Ltd.

Young, T.Y., Calvert, T.W. (1974): Classification, Estimation and Pattern Recognition, New York: Elsevier

Zadeh, L.A. (1965): Fuzzy Sets, in: Information and Control 8, S. 338-353.

Zapf, W. (1985): Der Zugang der Wissenschaft zur statistischen Information - Forderung und Realität. In: Statistisches Bundesamt (Hrsg.), Datennotstand und Datenschutz, Ergebnisse des 1. Wiesbadener Gesprächs 30./31. Oktober 1984. Stuttgart: Kohlhammer.

Zeifang, K. (1987): Die Test-Retest-Studie zum ALLBUS 1984 - Tabellenband - (ZUMA - Arbeitsbericht Nr. 87/01), Mannheim.

Zeifang, K. (1987a): Die Test-Retest-Studie. Methodenstudie des ALLBUS. Abschlußbericht, Mannheim, ZUMA-Arbeitsbericht Nr. 87/02.

Ziegler-Jung, B. (1980): Codes of Ethics und forschungspezifisches Datenschutzgesetz, in: Kaase et al. (1980).

Zimmerli, E., Liebl, K. (1984) (Hrsg.): Computermißbrauch Computersicherheit. Fälle - Abwehr - Aufdeckung, Ingelheim: Peter Hohl.

Zimmermann, H.-J., Zadeh, L. A., Gaines, B. R. (1984) (Eds.): Fuzzy sets and decision analysis, Amsterdam: North-Holland.