ZUMA

NACHRICHTEN

30

ZUMA NACHRICHTEN

Nr. 30

Mai 1992

Herausgeber:
Zentrum für Umfragen, Methoden und Analysen (ZUMA)
ZUMA ist Mitglied der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen e.V. (GESIS)

Vorsitzender: Prof. Dr. Max Kaase
Geschäftsführender Direktor: PD Dr. Peter Ph. Mohler

Anschrift:
B2, 1
Postfach 12 21 55
6800 Mannheim 1

Telefon:
Zentrale 0621/18004-0
Telefax 0621/18004-49
Redaktion 0621/18004-78

EARN/BITNET: OØ5 at DHDURZ2

Redaktion:
Dr. Paul Lüttinger

ISSN 0721-8516 16. Jahrgang

Die ZUMA-Nachrichten erscheinen im Mai und November eines Jahres. Sie werden Interessenten auf Anforderung kostenlos zugesandt.

ZUMA-Tagungen 1992

In eigener Sache

Erhebungen der amtlichen Statistik sind in zunehmendem Maße für viele Forschungsfragen in den Sozialwissenschaften eine wichtige Ressource. Insbesondere die Nutzung von Individualdaten ist jedoch mit dem Bundesstatistikgesetz von 1980 erschwert worden. Die Übermittlung solcher Daten wurde nur zugelassen, wenn gesichert war, daß keinerlei Chance bestand, aus den vorliegenden Daten auf die konkrete Person zu schließen, die "hinter" diesen Daten stand (Prinzip der absoluten Anonymität). Diese Regelung verhinderte in aller Regel eine Datenweitergabe; wenn sie dennoch erfolgte, waren die entsprechenden Daten in ihrer Nutzung erheblich beeinträchtigt. Im Bundesstatistikgesetz von 1987 wurde dem wissenschaftlichen Bedarf nach Einzeldaten allerdings insofern Rechnung getragen, als dort das Konzept der faktischen Anonymität eingeführt worden ist. Heike Wirth berichtet in diesem Zusammenhang über die Ergebnisse eines von ZUMA, dem Statistischen Bundesamt und der Universität Mannheim gemeinsam durchgeführten Forschungsprojektes, in dem konkrete Empfehlungen für die Umsetzung dieses Konzeptes entwickelt wurden. Die Ergebnisse werden mittlerweile auch für die Erstellung faktisch anonymisierter Mikrodatenfiles angewendet. Der Aufsatz von Georg Heer und Bernhard Schimpl-Neimanns resultiert ebenfalls aus einem gemeinsamen Forschungsprojekt des Statistischen Bundesamtes und ZUMA. Die Autoren beschäftigen sich mit einem bislang wenig beachteten Aspekt bei der Nutzung amtlicher Mikrodaten im Rahmen des Statistischen Informationssystems des Bundes (STATIS-BUND), und zwar mit den Auswirkungen des bei der Erstellung von Tabellen benutzten Anonymisierungsverfahrens auf multivariate Analysen.

Im Beitrag von Hartmut Götze wird ein Stichprobendesign für die neuen Bundesländer vorgestellt, das von der Empirisch-Methodischen Arbeitsgruppe am Institut für Soziologie und Sozialpolitik der Akademie der Wissenschaften der DDR entwickelt wurde. Die Forschungsbeiträge werden abgerundet von Heiner Ritter und Cornelia Züll. Sie berichten über ihre Erfahrungen mit dem Programm DBMS/COPY beim Datenaustausch zwischen Statistikprogrammen.

Im personellen Bereich haben sich im Berichtszeitraum erneut Änderungen ergeben. Petra-Victoria Steinhoff schied Ende Mai bei ZUMA aus; die Abteilung Datenorganisation wird nach ihrem Ausscheiden in der bisherigen Form nicht weitergeführt werden. Dr. Peter Schrott, bisher in Mannheim im Rahmen einer international vergleichenden Wahlstudie für den Bereich Inhaltsanalyse des deutschen Projektteils verantwortlich, wird im Juli Leiter der Abteilung Textanalyse, Medienanalyse und Vercodung (TEMEV). Als neue

Mitarbeiterinnen sind bei ZUMA ferner Dr. Angelika Glöckner-Rist, die vorerst im Rahmen eines DFG-Projektes bei ZUMA tätig ist, sowie Dr. Caroline Kramer, die bis zum Juni 1994 im Rahmen des Hochschulsonderprogramms II der Bundesregierung als Postdoktorandin gefördert wird. Zu berichten ist auch, daß Privatdozent Dr. Peter Ph. Mohler einen Ruf auf eine Gründungsprofessur in Leipzig erhalten hat.

Das vorliegende Heft ist die dreißigste Ausgabe der ZUMA-Nachrichten. Wir haben dieses kleine Jubiläum zum Anlaß genommen, für Sie ein Gesamtverzeichnis aller bisher erschienenen Beiträge zu erstellen. Darüber hinaus finden Sie in dieser Ausgabe auch eine Liste unserer seit 1974 erschienenen wichtigsten Publikationen. Ich hoffe, daß Sie auch dieses Mal den ZUMA-Nachrichten interessante Anregungen entnehmen können.

Max Kaase
Vorsitzender des ZUMA e.V.

Die faktische Anonymität von Mikrodaten: Ergebnisse und Konsequenzen eines Forschungsprojektes

Von Heike Wirth

Erhebungen der amtlichen Statistik stellen für die Untersuchung vieler Forschungsfragen seit langem eine außerordentlich wichtige und umfangreiche Datenressource dar. In den letzten Jahrzehnten hat sich das Nutzungsbedürfnis bezüglich dieser Daten jedoch nachhaltig geändert. Die Weiterentwicklung und Verfeinerung statistischer Analyseverfahren mit hohem Erkenntniswert und die verbesserten Möglichkeiten der Datenverarbeitung erlauben nicht nur eine stärkere Nutzung von Massendaten, sondern setzen vielfach auch die Verwendung von Individualdaten voraus. Der hieraus resultierende, zunehmende Bedarf an Individualdaten der amtlichen Statistik konnte allerdings nicht annähernd befriedigt werden, da nach dem Bundesstatistikgesetz von 1980 Individualdaten nur übermittelt werden durften, wenn sie absolut anonym waren. Spezifisch auf wissenschaftliche Nutzungsbedürfnisse ausgerichtet, wurde daher im neuen Bundesstatistikgesetz (1987) das Konzept der faktischen Anonymität eingeführt. Das Anonymisierungsprojekt hatte das Ziel, Empfehlungen für die konkrete Umsetzung der faktische Anonymität zu entwickeln.

1. Hintergrund des Forschungsprojektes

Zumindest seit Beginn der siebziger Jahre besteht ein von seiten der empirischen Sozialforschung vielfach artikulierter Bedarf nach Übermittlung anonymisierter Mikrodaten[1)] der amtlichen Statistik (vgl. hierzu u.a. Brennecke/Schneider 1977; Kaase et al. 1980; Zapf 1985; Müller 1982; Müller/Hauser 1987; Krupp/Preißl 1989). Bislang waren diesem Anliegen der empirischen Sozialwissenschaften durch den Gesetzgeber allerdings enge Grenzen gesetzt. So wurde im Bundesstatistikgesetz von 1980 (Paragraph 11 Abs.5) zwar erstmals geregelt, daß anonymisierte Einzelangaben, sofern diese Angaben dem Befragten nicht mehr zuordenbar sind (absolute Anonymität), von den statistischen Ämtern übermittelt werden dürfen. Diese Vorgabe war in der Praxis jedoch mit erheblichen Problemen behaftet. Denn zum einen wurde von Experten schon zum damaligen Zeitpunkt darauf hingewiesen, daß die Deanonymisierung eines Einzeldatensatzes - selbst bei weitreichenden datenorientierten Schutzvorkehrungen - nie mit absoluter Sicherheit ausgeschlossen werden kann (Brennecke 1980; Schlörer 1980). Zum anderen lagen keine Erkenntnisse über das Ausmaß tatsächlicher Reidentifikationsrisiken vor (Scheuch 1980), die eine theoretisch und empirisch abgesicherte Risikoabschätzung bei Übermittlung von anonymisierten Individualangaben ermöglicht hätten (Hamacher 1980).

Die Beschlußempfehlung des Innenausschusses des Bundestages, nach welcher vor einer Übermittlung anonymisierter Einzelangaben lediglich sichergestellt werden sollte, daß eine potentielle Reidentifikation nach Kenntnissen der statistischen Ämter zweifelsfrei ausgeschlossen werden kann (Südfeld 1987), war angesichts dieser allgemeinen Unsicherheit wenig fruchtbar. Im Sinne einer höchstmöglichen Risikoausschließung verhielten sich die statistischen Ämter bei Anforderungen von Individualdaten mit umfangreichem Merkmalskatalog äußerst zurückhaltend, bzw. die Datenweitergabe war mit solch weitreichenden datenmodifizierenden Anonymisierungsmaßnahmen verbunden, daß die Nutzungsbedürfnisse der Wissenschaft nicht befriedigt und das wissenschaftliche Potential der Daten auch nicht annähernd ausgeschöpft werden konnte.

Die im Konzept der absoluten bzw. zweifelsfreien Anonymität implizit enthaltene Befürchtung einer mißbräuchlichen Verwendung von anonymisierten Individualdaten durch die Wissenschaft beruht nicht auf empirischen Erfahrungswerten (Scheuch 1980, 1987), sondern vielmehr auf einem allgemeinen Spannungsverhältnis zwischen Datenschutz und Forschungsfreiheit. Nach Kaase et al. (1980:283) ist diese Konfliktlinie dadurch gekennzeichnet, daß mit dem umfassenden Informationsbedarf der Forschung unter Umständen ein Eingriff in die persönliche Integrität von Individuen einhergehen kann, umgekehrt jedoch auch auf Datenschutzargumente zurückgegriffen wird, um unbequeme Forschungsvorhaben vom Datenzugang auszuschließen.

Diese im internationalen Vergleich restriktive Handhabung der Datenweitergabe zuungunsten der Wissenschaft wurde vom Gesetzgeber erst relativiert, als das Bundesverfassungsgericht im sogenannten Volkszählungsurteil den Nutzungsbedarf von amtlichen (anonymisierten) Mikrodaten durch die Wissenschaft nicht nur ausdrücklich anerkannte, sondern zugleich betonte, daß das Spannungsverhältnis zwischen Datenschutz und Forschungsinteressen nicht einseitig zuungunsten der Wissenschaft gelöst werden dürfe.

In der Novellierung des Bundesstatistikgesetzes (BStatG) von 1987 wurde deshalb eine - in Anlehnung an eine von seiten der Forschung schon sehr früh vorgetragene Forderung (Mohler/Kaase 1980:108) - spezifische Wissenschaftsklausel eingeführt (Paragraph 16 Abs.6, BStatG), die an den Begriff der faktischen Anonymität anknüpft, wie er bereits durch die European Science Foundation definiert wurde. Danach 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 Forschung übermittelt werden, wenn die Einzelangaben nur mit einem

unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft zugeordnet werden können (...)" (Dorer/Mainusch/Tubies 1988:87).

Das Konzept der faktischen Anonymität ist ein wesentlicher Fortschritt gegenüber der alten Regelung. Denn zum einen ermöglicht die Beschränkung auf den Wissenschaftskontext eine Präzisierung der für einen Reidentifikationserfolg maßgeblichen Randbedingungen, was insbesondere bei der Berücksichtigung von Deanonymisierungsmotiven und Zusatzwissen von großer Bedeutung ist. Zum anderen wird das Mißtrauen gegenüber einer mißbräuchlichen Datenverwendung insofern relativiert, als man von einem rational kalkulierenden Datenangreifer ausgeht, der zwischen dem durch einen Reidentifikationsversuch erzielten Nutzen und den hierfür anfallenden Kosten abwägt.

Eine unmittelbare Umsetzung dieser neuen Regelung in die Weitergabepraxis war allerdings nicht möglich, weil bislang keine Kenntnisse zu den zentralen Problemen der faktischen Anonymität vorlagen. Die in der Vergangenheit durchgeführten Untersuchungen zu Reidentifikationsrisiken orientieren sich an der damaligen Gesetzeslage, bei welcher der Empfängerkreis von (absolut) anonymisierten Mikrodaten nicht näher definiert war. Da hierbei hypothetisch von einem allumfassenden Angriffsszenario und unbegrenzten Mitteln ausgegangen werden mußte, fehlen Untersuchungen, die Reidentifikationsrisiken unter dem Kosten-Nutzen-Aspekt analysieren ebenso wie Untersuchungen, die sich spezifisch auf die im Wissenschaftskontext vorhandenen Randbedingungen konzentrieren.

In Anknüpfung an eine von der amtlichen Statistik schon 1986 aufgegriffene Initiative[2] wurde daher in Kooperation zwischen dem Statistischen Bundesamt, der Universität Mannheim und ZUMA ein Forschungsprojekt[3] durchgeführt, um das Konzept der faktischen Anonymität zu operationalisieren. Ziel war, eine konsensfähige Lösung für die zukünftige Weitergabepraxis zu entwickeln.

Die zentralen Arbeiten des Anonymisierungsprojektes bezogen sich auf die Analyse des potentiellen Reidentifikationsrisikos von amtlichen Mikrodaten und auf die hierbei anfallenden Kosten. Hierfür wurden zunächst die Randbedingungen des Wissenschaftsszenarios untersucht, die für einen Reidentifikationsversuch von Bedeutung sein könnten. Auf dieser Grundlage wurden fünf Angriffsszenarien entwickelt, für die das Reidentifikationsrisiko sowie der damit verbundene Aufwand eingehend überprüft wurde.

Die allgemeinen Fragen zum Problembereich "faktische Anonymität" stellen sich natürlich für alle Arten von Mikrodaten. Aufgrund der vielfältigen

Formen von Mikrodaten ist es allerdings nicht möglich, allgemeine Antworten zu erwarten, die für alle Datenarten gleichermaßen zutreffen. Die Analysen beschränkten sich daher ausschließlich auf amtliche Mikrodaten und ihre Nutzung durch die Wissenschaft. Auch für Daten der amtlichen Statistik stellt sich das Problem unterschiedlich für verschiedene Datentypen. Betriebs- und Wirtschaftsdaten werfen andere Probleme auf als Personen- und Haushaltsdaten (Südfeld 1987; Krupp/Preißl 1989). Vor diesem Hintergrund wurde eine weitere Eingrenzung auf Personen- und Haushaltsdaten und hier wiederum ausschließlich auf den Mikrozensus und die Einkommens- und Verbrauchsstichprobe (EVS) vorgenommen.

In diesem Beitrag werden einige der zentralen Projektergebnisse aufgegriffen und die hieraus abgeleiteten Weitergabeempfehlungen dargestellt.

2. Voraussetzungen und Methoden einer Reidentifikation

2.1 Voraussetzungen einer Reidentifikation

Anonyme Daten unterscheiden sich von personenbezogenen Daten insofern, als sie keine direkten Identifikatoren, wie beispielsweise Namen und Anschriften, enthalten. Das Fehlen direkter Identifikatoren schließt allerdings nicht aus, daß für anonyme Daten nicht nachträglich ein Personenbezug rekonstruiert, d.h. eine Reidentifikation vorgenommen werden kann.[4]

Eine notwendige Voraussetzung hierbei ist, daß ein Angreifer über sogenanntes Zusatzwissen (auch als Identifikationsfile bezeichnet), d.h. personenbezogene Informationen verfügt, das zum einen gemeinsame Merkmale (Überschneidungsmerkmale) mit dem anonymen Mikrodatenfile aufweist und das sich zum anderen - zumindest in Teilen - auf die gleichen Personen wie das Mikrodatenfile bezieht. Ein Angreifer könnte dann versuchen, durch einen Abgleich der Überschneidungsmerkmale auf Identität oder sehr große Ähnlichkeit jene Datensätze im Mikrodaten- und Identifikationsfile zu ermitteln, die von ein und derselben Person stammen. Eine Reidentifikation wäre dann gegeben, wenn es anhand der Überschneidungsmerkmale möglich ist, für einen Datensatz des Mikrodatenfiles eine eins-zu-eins Relation zu einem Datensatz des Zusatzwissens herzustellen und wenn sichergestellt ist, daß sich diese Datensätze auf ein und dieselbe Person beziehen.

Das reale Reidentifikationsrisiko ist im wesentlichen durch drei Eigenschaften der Datenbasis beeinflußt (Paaß/Wauschkuhn 1985):

(1) Informationsgehalt der Überschneidungsmerkmale:
Eine Reidentifikation setzt die Einzigartigkeit von Ausprägungskombinationen voraus. Je höher die Anzahl der Überschneidungsmerkmale, der Differenzierungsgrad der Ausprägungen sowie ihrer Verteilungen ist, desto deutlicher sind die Datensätze im Merkmalsraum voneinander abgegrenzt und desto eher werden eins-zu-eins Zuordnungen möglich sein (für formale Abschätzungen des Informationsgehalts von Überschneidungsmerkmalen vgl. Müller et al. 1991:101f.)

(2) Stichprobeneigenschaften der Daten:
Eine Reidentifikation ist nur dann möglich, wenn eine Person in Mikrodatenfile und Zusatzwissen erfaßt ist. Die Stichprobeneigenschaft stellt daher eine prinzipielle Schranke für die Erfolgswahrscheinlichkeiten von Deanonymisierungsversuchen dar. Wird eine beliebige in der Grundgesamtheit enthaltene Person in einer amtlichen Stichprobe gesucht, entspricht die maximale Erfolgswahrscheinlichkeit dem Auswahlsatz dieser Stichprobe. Ist das Zusatzwissen ebenfalls nur eine Stichprobe - und beide Stichproben sind voneinander unabhängig - ergibt sich die Wahrscheinlichkeit, daß eine beliebige Person in beiden Datenfiles enthalten ist, durch die Multiplikation der Auswahlsätze.
Ein wesentlicher Unsicherheitsfaktor bei Reidentifikationsversuchen ist hierbei durch mögliche statistische Doppelgänger in der Grundgesamtheit gegeben, d.h. durch Personen, die identische Ausprägungskombinationen aufweisen. Ist eine Ausprägungskombination in der Grundgesamtheit mehrfach besetzt und stehen für einen Reidentifikationsversuch nur Stichproben zur Verfügung, ist eine Reidentifikation auch bei eins-zu-eins Zuordnungen nicht mehr mit Sicherheit, sondern nur noch mit einer gewissen Wahrscheinlichkeit möglich, da Verwechslungen mit statistischen Doppelgängern nicht ausgeschlossen werden können (für formale Modelle zur Abschätzung der Populationseinzigartigkeit von Ausprägungskombinationen siehe u.a. Marsh et al. 1991; Bethlehem et al. 1990).
Die von Stichproben ausgehende Schutzwirkung besteht dann nicht mehr, wenn ein Angreifer weiß, welche in seinem Zusatzwissen enthaltenen Personen an der Mikrodatenerhebung teilgenommen haben (Teilnahmekenntnis). Wird eine einzigartige Ausprägungskombination im Mikrodatenfile gefunden, die mit der Kombination des gesuchten Falls im Identifikationsfile übereinstimmt, dann kann - unter der Annahme, daß die Daten kompatibel abgebildet sind - ein Angreifer davon ausgehen, daß es sich um den gesuchten Fall handelt, unabhängig davon, ob in der Grundgesamtheit statistische Doppelgänger existieren oder nicht (vgl. Müller et al. 1991:95). Für eine Angriffssituation mit "Teilnahmekenntnis" ist daher ein wesentlich erhöhtes Reidentifikationsrisiko anzunehmen.

(3) Dateninkompatibilitäten zwischen Mikrodatenfile und Zusatzwissen:
Der oben skizzierte Reidentifikationsprozeß als einfacher Abgleich von Ausprägungskombinationen setzt voraus, daß die gesuchten Datensätze in Mikrodatenfile und Zusatzwissen identisch abgebildet sind (vgl. Block/Olsson 1976; Marsh et al. 1991). Diese Voraussetzung ist unter empirischen Bedingungen nicht immer erfüllt. So ist aus sozialwissenschaftlichen Untersuchungen bekannt, daß es bei jeder Datenerhebung in einem gewissen Grad zu Abweichungen von den "wahren" Werten kommt (Schnell/Hill/Esser 1988).[5] Als mögliche Ursachen sollen hier nur Antwortverzerrungen und einfache Aufbereitungsfehler genannt werden (für eine detaillierte Analyse siehe Müller et al. 1991:114ff.). Werden - wie bei einem Reidentifikationsversuch - Datenbestände aus unterschiedlichen Generierungsprozessen verglichen, können diese originären Datenfehler mit anderen Quellen möglicher Abweichungen kumulieren, welche beispielsweise bedingt sein können durch unterschiedliche Erhebungszeitpunkte, -ziele oder -kontexte oder unterschiedliche Meßinstrumente bei der Datenerhebung.

Auf individueller Ebene können solche Inkompatibilitäten dazu führen, daß Informationen, die sich auf ein und dieselbe Person beziehen, in unterschiedlichen Datenbeständen in abweichender Weise abgebildet sind. Eine Zuordnung auf Basis von Ausprägungsidentität wäre daher nicht möglich. Ebenso ist es vorstellbar, daß Datensätze, die ursprünglich nur sehr ähnliche Merkmalsausprägungen aufweisen, aufgrund von Inkompatibilitäten nun identisch abgebildet sind, wodurch eine weitere Quelle möglicher Verwechslungen bei Reidentifikationsversuchen entsteht. Unter der Annahme, daß bei einem realistischen Angriffsszenario sowohl die Stichprobeneigenschaften von Daten wie auch das Auftreten von Dateninkompatibilitäten eine sichere Reidentifikation wesentlich behindern können, stellt sich die Frage, inwieweit die einem Angreifer potentiell zur Verfügung stehenden Reidentifikationstechniken diese Unsicherheitsfaktoren berücksichtigen.

In der Literatur werden hauptsächlich zwei Reidentifikationstechniken diskutiert, die in bezug auf ihre Leistungsfähigkeit und den mit ihnen verbundenen Aufwand jeweils Extrempunkte einer Skala denkbarer Angriffstechniken darstellen. Da wir für die Operationalisierung der faktischen Anonymität auf diese Techniken zurückgegriffen haben, sollen sie im folgenden kurz skizziert werden.[6]

2.2 Reidentifikationstechniken

Am unteren Ende der Skala sind einfache Abgleichtechniken anzusiedeln, die in der Literatur als Hintertreppenidentifikation, Sortier- oder Selektionstechniken (Schlörer 1980; Block/Olsson 1976; Dittrich/Schlörer 1985) bezeichnet werden. Diesen Verfahren liegt ein Prinzip zugrunde, das in der empirischen Sozialforschung als sogenanntes Matching zur Anwendung kommt, wenn beispielsweise mehrere Dateien, die sich auf die gleiche Population beziehen, miteinander verknüpft werden sollen. Hierbei werden die Einzeldatensätze auf der Basis von Schlüsselvariablen in einer eins-zu-eins Relation zusammengefügt. Ein analoges Vorgehen ist auch bei einem Reidentifikationsversuch vorstellbar, wobei die Überschneidungsmerkmale als Schlüsselvariablen eingesetzt werden. Über Suchroutinen können dann jene Datensätze des Zusatzwissens ermittelt werden, die eine identische, einzigartige Ausprägungskombination zu Datensätzen des Mikrodatenfiles aufweisen. Der Vorteil einer solch einfachen Abgleichtechnik ist in der allgemeinen Verfügbarkeit (nahezu jedes Statistikprogramm bietet die Möglichkeit des Matching) und der wenig aufwendigen Umsetzung zu sehen. In einigen Untersuchungen wird diesen einfachen Abgleichtechniken daher direkt (Fischer-Hübner 1986; Brunnstein 1987) oder indirekt (Dalenius 1977, 1986, 1988; Bethlehem/Keller/Pannekoek 1990) ein hohes Gefährdungspotential zugeschrieben.

Eine Reidentifikationstechnik, die als Zuordnungskriterium nur die Ausprägungsidentität in Verbindung mit der Einzigartigkeit von Datensätzen voraussetzt, weist bezüglich der oben aufgezeigten Unsicherheitsfaktoren allerdings beträchtliche Defizite auf. Einerseits können Datensätze, die inkompatibel abgebildet sind, beim einfachen Ausprägungsabgleich nicht zugeordnet werden. Treten Dateninkompatibilitäten auf und ist ein gesuchter Fall hiervon betroffen, wird der Datensatz im Verlauf des Suchprozesses als nicht passend aussortiert. Andererseits ist es möglich, daß Datensätze, die bei korrekter Abbildung nur sehr ähnliche Werte aufweisen würden, aufgrund von Inkompatibilitäten identisch abgebildet sind. Hier würden mit einer einfachen Abgleichtechnik Falschzuordnungen erfolgen. Abweichungen vom "wahren" Wert sind nur dann unproblematisch, wenn sie übereinstimmen und damit die Abbildung wiederum kompatibel wäre. Schließlich kann es, wenn sowohl das Mikrodatenfile wie auch das Identifikationsfile nur als Stichproben zur Verfügung stehen, aufgrund von statistischen Doppelgängern zu Verwechslungen und damit zu Falschzuordnungen kommen. Das Ausmaß von Nicht- und Falschzuordnungen ist durch die einfache Abgleichtechnik nicht kontrollierbar.

Empirische Befunde, die konkrete Aussagen über das von einem einfachen Abgleich ausgehende Gefährdungspotential ermöglicht hätten, lagen bislang

allerdings nicht vor. Insofern das einfache Matching die untere Aufwandsgrenze bei einem Reidentifikationsversuch markiert und daher bei einem Angriffsszenario sehr wahrscheinlich an erster Stelle stehen würde, wurde diese Technik ungeachtet der offensichtlichen Defizite für die Operationalisierung der faktischen Anonymität berücksichtigt.

Komplexere Techniken versuchen, die aufgezeigten Probleme durch Fehlerabschätzungen und Wahrscheinlichkeitsberechnungen über das Auftreten von statistischen Doppelgängern zu lösen. Als das - auch international - leistungsfähigste Verfahren gilt das von der Gesellschaft für Mathematik und Datenverarbeitung im Rahmen des AIMIPH-Projektes entwickelte Verfahren zur Abschätzung von Reidentifikationsrisiken (Paaß/Wauschkuhn 1985). Dieser auf der Diskriminanzanalyse und Dichteschätzung beruhende Algorithmus ermittelt die Wahrscheinlichkeiten von spezifischen 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 Zusatzwissens derjenige Datensatz des Mikrodatenfiles ermittelt werden, der in seiner Ausprägungskombination die höchste Übereinstimmung aufweist. Zugleich wird die Wahrscheinlichkeit berechnet, mit welcher diese Zuordnung korrekt ist. Diese ist um so höher, je weniger ähnliche Datensätze es im Mikrodatenfile gibt. Liegt die ermittelte Wahrscheinlichkeit über einer festzulegenden Sicherheitsschwelle (z.B. 0.9), so ist davon auszugehen, daß sich die zugeordneten Datensätze auf eine spezifische Person beziehen und damit eine Reidentifikation vorliegt. Im Gegensatz zu der einfachen Abgleichtechnik liefert diese Technik damit konkrete Entscheidungskriterien, wann eine Zuordnung von Datensätzen als korrekt angesehen werden kann und wann nicht.

Im Rahmen des AIMIPH-Projektes wurde die diskriminanzanalytische Methode einer empirischen Überprüfung unterzogen (Paaß/Wauschkuhn 1985). Hierfür standen ein Mikrodatenfile und ein hieraus erzeugtes - mit Fehlern überlagertes - Identifikationsfile zur Verfügung. Das Verfahren hat bei diesem Material unter spezifischen Randbedingungen vergleichsweise hohe Reidentifikationsquoten gezeigt. Aufgrund dieser Ergebnisse war davon auszugehen, daß von diesem Verfahren ein sehr hohes Gefährdungspotential ausgeht, weshalb auch diese Methode im Anonymisierungsprojekt einer empirischen Überprüfung - allerdings unter Verwendung von realen Mikrodaten und realem Zusatzwissen - unterzogen wurde (Bender 1990; Müller 1991; Bender/Blien/Müller 1990a,b).

3. Randbedingungen eines Deanonymisierungsversuchs im wissenschaftlichen Kontext: Deanonymisierungsmotive und Zusatzwissen

3.1 Deanonymisierungsmotive

Ob, abgesehen von der technischen Machbarkeit, überhaupt Deanonymisierungsversuche zu erwarten sind, hängt ab von dem Nutzen, den sich ein Angreifer von einer Reidentifikation verspricht. Um den potentiellen Nutzen deanonymisierter Daten im Wissenschaftskontext zu bestimmen, wurde geprüft, welche Logik der wissenschaftlichen Datennutzung zugrunde liegt und welche Motive sich hieraus für Deanonymisierungsversuche ergeben könnten (vgl. Müller et al. 1991:132ff.).
Diese Analyse führte zu dem Befund, der auch schon vom Bundesverfassungsgericht angeführt wurde: "(...) der Wissenschaftler ist regelmäßig nicht an der einzelnen Person interessiert, sondern an dem Individuum als Träger bestimmter Merkmale." (Neue Juristische Wochenschrift 1984:428, vgl. auch Scheuch 1980; Hamacher 1980; Zapf 1985). Plausible Motive sind aufgrund der beruflichen Interessenlage eines empirisch arbeitenden Sozialwissenschaftlers kaum überzeugend rekonstruierbar. 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. Etwa um eine Auswahlbasis für eine eigene Stichprobenerhebung zu gewinnen oder um eine eigene Datenbasis mit den im Mikrodatenfile enthaltenen Informationen zu ergänzen. Diesen beruflich motivierten Angriffsszenarien ist gemeinsam, daß das Interesse nicht auf die Reidentifikation einiger weniger, sondern einer Vielzahl von Einzeldatensätzen gerichtet ist. Um Reidentifikationen in größerem Maßstab durchführen zu können, müßte - bedingt durch den relativ geringen Auswahlsatz der hier betrachteten Mikrodatenfiles - entweder eine sehr leistungsfähige Reidentifikationstechnik oder sehr umfangreiches Zusatzwissen zur Verfügung stehen.[7]

Es ist sicherlich diskussionswürdig, inwieweit im Zusammenhang einer Datenweitergabe für Forschungsvorhaben auch wissenschaftsfremde Deanonymisierungsmotive zu berücksichtigen sind. Will man jedoch hypothetisch mögliche - wenn auch sehr unwahrscheinliche - Fälle von Datenmißbrauch im Wissenschaftskontext berücksichtigen, so sind berufsfremde Motive ebenfalls in die Analyse einzubeziehen.

Im Unterschied zu beruflichen bietet sich bei außerberuflichen ein weites Spektrum hypothetisch denkbarer Motive an (Knoche 1989; Müller et al. 1991:151ff.). Diese reichen von persönlicher Neugier über ökonomisch

motivierte Deanonymisierungsversuche (beispielsweise der Verkauf von Information an Adressenhändler) bis hin zu eindeutig krimineller Motivation, bei welcher deanonymisierte Daten etwa für Erpressungsversuche herangezogen werden könnten. Da einzelne dieser Motive auf die Reidentifikation einer oder nur weniger Person(en) abzielen, könnte hier nicht nur der Aufwand für die Beschaffung von Zusatzwissen gering sein, sondern bereits eine einfache Abgleichtechnik die gewünschten Erfolge erbringen.

3.2 Zusatzwissen innerhalb des Wissenschaftsbereichs

Die Reidentifikation einer Person ist nur dann möglich, wenn ein Datenangreifer über entsprechendes Zusatzwissen verfügt. Allgemein läßt sich das erlangbare Zusatzwissen kaum abschließend abgrenzen (Burkert 1979, 1980). Die Weitergabebeschränkung faktisch anonymer Daten auf den wissenschaftlichen Kontext beinhaltet jedoch, daß die Risikoabschätzung nicht vor dem Hintergrund eines beliebig zur Verfügung stehenden Zusatzwissens erfolgen muß, sondern spezifisch auf den Datenempfängerkreis ausgerichtet sein kann.

Nach den im Anonymisierungsprojekt durchgeführten Analysen sind es im wesentlichen zwei Arten von Informationsquellen, die als Zusatzwissen genutzt werden könnten (vgl. Beckmann 1988):

Für den Bereich von öffentlich oder für einen beschränkten Personenkreis zugänglichen Registern oder privaten Datenquellen wurden insbesondere berufsgruppenspezifische Handbücher als möglicherweise riskant charakterisiert. In derartigen Handbüchern finden sich zwar in der Regel nur relativ wenige Überschneidungsmerkmale zu amtlichen Daten, es handelt sich jedoch zum Teil um detaillierte berufsbezogene Angaben in Verbindung mit Regionalangaben für teilweise deutlich von der Durchschnittsbevölkerung abgegrenzte Subpopulationen. Da in den meisten Handbüchern jeweils eine Vollerfassung der entsprechenden Berufsgruppen angestrebt wird, könnte die Suchrichtung bei einer Reidentifikation umgekehrt werden: Der Angreifer sucht sich eine dieser spezifischen Subpopulation angehörende Zielperson im Mikrodatenfile aus und versucht dieser den entsprechenden Datensatz im Zusatzwissen zuzuordnen (Strategie des Fischzugs). Je vollständiger die jeweilige Subpopulation im Zusatzwissen erfaßt und je deutlicher sie von durchschnittlichen Merkmalsträgern abgegrenzt ist, desto höher sind die Erfolgschancen einer solchen Fischzugsstrategie einzuschätzen. Es ist daher vorstellbar, daß unter Umständen ein erheblicher Teil der relevanten Mikrodatensätze aufgrund der in einem spezifischen Handbuch enthaltenen Informationen reidentifizierbar ist.

Als zweite wichtige Quelle von Zusatzwissen sind die im Kontext der Sozialwissenschaft professionsgemäß zur Verfügung stehenden Daten zu berücksichtigen, die im wesentlichen aus eigenen, freiwilligen Erhebungen stammen. Im Unterschied zu öffentlich zugänglichen Informationsquellen handelt es sich bei sozialwissenschaftlichen Erhebungen in der Regel um Stichproben. Der Stichprobenumfang ist meist relativ klein, kann aber in Einzelfällen auch einige tausend Personen umfassen. Für diesen Personenkreis wird zum Teil ein sehr umfangreicher Merkmalskatalog erfaßt. Die Stichprobeneigenschaft stellt zwar eine generelle Schranke für Reidentifikationsversuche dar, eine massenhafte Deanonymisierung kann nahezu ausgeschlossen werden. Darüber hinaus verfügen gerade bei umfangreichen Erhebungen die Wissenschaftler in der Regel nicht über die Adressen, da die Erhebungen durch professionelle Umfrageinstitute durchgeführt werden und die Daten ohne Adressen an die Wissenschaftler weitergegeben werden. Dennoch liegt die Annahme nahe, daß beim Vorliegen einer eigenen, personenbezogenen Erhebung der Informationsgehalt der Überschneidungsmerkmale so hoch sein könnte, daß dem Versuch, einzelne Personen zu reidentifizieren, gewisse Erfolgschancen zugebilligt werden könnten.

4. Empirische Überprüfung des Unverhältnismäßigkeitskriteriums

4.1 Untersuchungsanlage

Wie aus den bisherigen Ausführungen deutlich geworden sein sollte, ist die Wahrscheinlichkeit einer korrekten Zuordnung unter realen Bedingungen durch das Zusammenwirken einer Reihe unterschiedlicher Faktoren bestimmt. Hierzu zählen sowohl die Eigenschaften von Zusatzwissen und Mikrodatenfile (beispielsweise Umfang und Repräsentativität), der Informationsgehalt der Überschneidungsmerkmale, der Grad der Kompatibilität beziehungsweise Inkompatibilität unterschiedlicher Datenbestände wie auch die zur Verfügung stehenden Reidentifikationstechniken. Schon der Versuch, die Bedeutung der einzelnen Komponenten für das Reidentifikationsrisiko zu bestimmen, erweist sich als schwierig. Die Auswirkung der kombinierten Effekte eindeutig zu präzisieren, muß daher auch bei einer sehr sorgfältigen Analyse als in hohem Maße spekulativ erscheinen.

Die meisten Untersuchungen zur Bestimmung von Reidentifikationsrisiken unterstellen daher in aller Regel das reine Wirken einzelner Faktoren und grenzen über Modellannahmen störende Randfaktoren, wie beispielsweise Dateninkompatibilitäten, aus. So werden in einem Teil der Arbeiten Inkompatibilitäten überhaupt nicht erwähnt. Diesen Arbeiten liegt die

Annahme zugrunde, daß Datensätze mit übereinstimmenden und eindeutigen Ausprägungskombinationen, vor allem wenn sie sich auf die Grundgesamtheit beziehen, ohne Einschränkung reidentifizierbar sind (Fischer-Hübner 1986; Brunnstein 1987; Dalenius 1977, 1986, 1988). In anderen Studien werden Dateninkompatibilitäten als möglicher Störfaktor bei Reidentifikationsversuchen zwar explizit erwähnt. Bei der Abschätzung von Reidentifikationsrisiken wird dann jedoch vereinfachend unterstellt, daß die Daten vollständig kompatibel abgebildet sind (Spruill 1983; Dittrich/Schlörer 1985; Bethlehem et al. 1990). Nur wenige Arbeiten stellen Inkompatibilitäten für die Abschätzung von Reidentifikationsrisiken direkt in Rechnung (Paaß/Wauschkuhn 1985; Skinner et al. 1990; Marsh et al. 1991). Bei letzteren wird jedoch, wie in den anderen angeführten Untersuchungen auch, das für einen Reidentifikationsversuch notwendige Zusatzwissen als externe Randbedingung ausgeklammert.

Im Anonymisierungsprojekt wurde ein im Vergleich zu diesen Untersuchungen fast konträrer Weg gewählt, indem potentielle Reidentifikationsrisiken für reale Mikrodatenfiles unter Verwendung von realem Zusatzwissen ermittelt wurden. Entsprechend dem Konzept der faktischen Anonymität stand hierbei das erzielbare Endergebnis von realistischen Angriffsversuchen im Mittelpunkt des Interesses. Mögliche Störfaktoren wurden nicht ausgeklammert, sondern konnten in der Weise wirksam werden, wie dies bei einem realen Datenangriff gegeben wäre. Wenn es dabei auch nicht möglich war, einzelne Wirkungsfaktoren eindeutig zu isolieren, so konnte jedoch das Unverhältnismäßigkeitskriterium konkretisiert, das heißt der Aufwand realistischer Angriffsversuche ermittelt und mit dem Ertrag im Sinne erfolgreicher Deanonymisierungen verglichen werden.

Hierfür wurden im Hinblick auf die Gefährdungssituationen, die sich aus der Analyse des potentiellen Nutzens von deanonymisierten Daten und des im Wissenschaftskontext potentiell verfügbaren Zusatzwissens ergaben, fünf Angriffssituation spezifiziert (vgl. Müller et al. 1991:235ff.). Für zwei dieser Szenarien wurde das Reidentifikationsrisiko und der hierbei entstehende Aufwand empirisch überprüft. Drei Szenarien wurden unter Annahme plausibler Randbedingungen und den Ergebnissen der empirischen Überprüfung einer argumentativen Analyse unterzogen, bei welcher die Kosten eines Reidentifikationsversuchs den Kosten einer alternativen Beschaffung gleichwertiger Informationen gegenüber gestellt wurden (vgl. Helmcke 1989; Müller et al. 1991:351ff.).

Die Verwendung von realen Datenbeständen bedeutet nicht, daß gesetzeswidrig Deanonymisierungen versucht oder vorgenommen wurden.

Konkrete Aussagen über die Erfolgswahrscheinlichkeit von Reidentifikationsrisiken erfordern jedoch, daß das Verhältnis von korrekten und falschen Zuordnungen ebenso bekannt ist wie das Gesamtpotential der hypothetisch möglichen korrekten Zuordnungen. Zu diesem Zweck wurde ein aufwendiges Doppelt-Blind-Verfahren entwickelt, bei welchem ein Treuhänder zwischen die Adressenbesitzer einerseits und die mit der Durchführung der Experimente betrauten Forscher andererseits geschaltet wurde. Die Ergebnisse der mit anonymen Daten durchgeführten Experimente wurden an den Treuhänder weitergeleitet, der in datenschutzgerechter Weise überprüfte, welche der Zuordnungen korrekt, welche falsch und wie viele korrekte Zuordnungen maximal möglich gewesen wären (vgl. Müller et al. 1991:243f.).

Die empirische Überprüfung potentieller Reidentifikationsrisiken bei unterschiedlichen Angriffsvarianten und mit unterschiedlichen Reidentifikationstechniken erfolgte jeweils am Beispiel des Mikrozensus. Denn zum einen weist der Mikrozensus in der Regel mehr Überschneidungsmerkmale mit anderen Informationsquellen auf als die EVS (Müller et al. 1991:190-210), was auf ein höheres Gefährdungspotential hinweist. Zum anderen ist der Auswahlsatz des Mikrozensus um den Faktor fünf größer als der der EVS, das heißt die Wahrscheinlichkeit, daß eine beliebige Person sowohl im Zusatzwissen als auch im Mikrodatenfile erfaßt wurde, ist für den Mikrozensus größer als für die EVS.

Konkret wurde jeweils der Mikrozensus 1987 von Nordrhein-Westfalen genutzt. Dieser wurde dankenswerterweise vom Landesamt für Statistik und Datenverarbeitung Nordrhein-Westfalen mit den jeweils benötigten Überschneidungsmerkmalen und dem vollen Auswahlsatz (N=169.368) zur Verfügung gestellt. Die Daten waren anonym, das heißt sie enthielten keine personenbezogenen Angaben. Sonstige Anonymisierungsmaßnahmen wurden nicht vorgenommen (Müller et al. 1991:251f.).

Für ein erstes Angriffsszenario wurde - repräsentativ für öffentlich zugängliche Informationsquellen - Kürschners Deutscher Gelehrtenkalender als Zusatzwissen herangezogen. Alle in diesem Handbuch enthaltenen Informationen, die als Überschneidungsmerkmale zum Mikrozensus in Betracht kamen, wurden nach den Konventionen des Mikrozensus kodiert und auf Datenträger übernommen. Da der Gelehrtenkalender die Hochschullehrer nahezu vollständig erfaßt, konnte mit dieser Datenbasis überprüft werden, inwieweit der Versuch einer massenhaften Deanonymisierung realistisch ist.

Als zweite wichtige Quelle von Zusatzwissen wurde exemplarisch für sozialwissenschaftliche Datenbestände eine der umfangreichsten, in den alten Bundesländern durchgeführten, repräsentativen Erhebungen herangezogen.[8] Am Beispiel dieser Studie sollte zum einen das Gefährdungspotential eingegrenzt werden, daß sich aus dem Zugang zu umfangreichen, personenbezogenen Datenbeständen in den Sozialwissenschaften ergeben könnte. Zum anderen können die überwiegend erwerbs-, haushalts- und familienspezifischen Überschneidungsmerkmale zum Mikrozensus, als typische Beispiele für Alltags- beziehungsweise Anschauungswissen über andere Personen (Nachbarn, Bekannte oder Arbeitskollegen) gelten. Auf diese Weise konnte für etwa zweieinhalb Tausend unterschiedlich realistische Fälle geprüft werden, welches Risiko besteht, eine beliebige Person in einem Mikrodatenfile anhand von Alltagswissen zu deanonymisieren.

4.2 Empirische Überprüfung des Gelehrtenszenarios

4.2.1 Datenbasis

Kürschners Deutscher Gelehrtenkalender ist das umfassendste Verzeichnis von Forschern und Gelehrten im deutschsprachigen Raum ist und damit eine nach bisherigem Wissen sehr riskante Quelle von Zusatzwissen. Die Ausgabe von 1987 umfaßt circa 45.000 Gelehrte, bezogen auf Nordrhein-Westfalen sind etwa 8000 Fälle enthalten.

Der Gelehrtenkalender enthält zehn teilweise sehr differenzierte Überschneidungsmerkmale zum Mikrozensus (vgl. Übersicht 1) mit einem sehr hohen Informationsgehalt. Anhand der Merkmale "Beruf" und "Branche" ist es möglich, die Gelehrtenpopulation im Mikrozensus einzugrenzen. Zugleich ist diese Gelehrtenpopulation durch die Angaben "Fachzugehörigkeit" und "Geburtsjahr" stark differenziert. Mit den verfügbaren Regionalinformationen "Bundesraumordnungsregion" und "Gemeindegrößenklasse" ist es (in Verbindung mit dem Bundesland) möglich, im Mikrozensus Regionaleinheiten mit weniger als 200.000 Einwohnern in der Grundgesamtheit einzugrenzen (vgl. Müller et al. 1991:445f.). Es ist davon auszugehen, daß die im Gelehrtenkalender enthaltenen Angaben wenig fehlerbehaftet sind, da den Befragten die Möglichkeit geboten wird, die Angaben vor der Publizierung nochmals zu überprüfen. Um bei der Datenaufbereitung Quellen möglicher Inkompatibilitäten zu minimieren, wurde die Umsetzung der Klartextangaben des Gelehrtenkalenders in die Mikrozensus-Verkodung sowie die maschinelle Aufbereitung von Mitarbeitern des Statistischen Bundesamtes vorgenommen. Da diese Fachkräfte mit der Mikrozensusverkodung vertraut sind, ist von einer wesentlich höheren Vergleichbarkeit der

Angaben auszugehen, als dies der Fall wäre, wenn ein Angreifer eine solche Verschlüsselung vornimmt.

Hierbei zeigte sich, daß eine Verschlüsselung von Klartextangaben - auch wenn sie professionell erfolgt - durchaus problembehaftet sein kann. Nicht alle der im Gelehrtenkalender enthaltenen Angaben ließen sich eindeutig einer bestimmten Kategorie zuordnen. Ein solcher Fall war z.B. dann gegeben, wenn ein Hochschullehrer zugleich Arzt war, den Angaben aber nicht entnommen werden konnte, inwieweit beide Tätigkeiten gleichzeitig verfolgt werden, bzw. bei welcher Angabe es sich um die Haupttätigkeit handelt. Um das höchstmögliche Reidentifikationspotential auszuschöpfen, wurden in solchen Zweifelsfällen Alternativ-Verschlüsselungen vorgenommen. Insgesamt betrafen diese Alternativen die vier Überschneidungsmerkmale "Beruf", "ausgeübte Tätigkeit", "Wirtschaftszweig" und "Fachrichtung des Hochschulabschlusses" (vgl. Übersicht 1). Um diese Alternativvariablen zu berücksichtigen, wurden entsprechend der Variationsmöglichkeiten 2^4 unterschiedliche Identifikationsfiles erzeugt und bei der empirischen Überprüfung berücksichtigt. Für die konkrete Überprüfung des Szenarios wurde sowohl die einfache Abgleichtechnik wie auch die diskriminanzanalytische Methode von Paaß/Wauschkuhn herangezogen.

4.2.2 Ergebnisse

4.2.2.1 Einfache Abgleichtechnik

Das hohe Reidentifikationsrisiko der hier zur Verfügung stehenden Datenbasis kam insbesondere in der Zahl der einzigartigen Ausprägungskombinationen sowohl im Mikrodatenfile wie auch in den unterschiedlichen Identifikationsfiles zum Ausdruck. In Abhängigkeit der berücksichtigten Alternativvariablen, wiesen zwischen 45 und 65 Prozent der 7983 Datensätze im Gelehrtenkalender eine einzigartige Ausprägungskombination auf (vgl. Müller et al. 1991:274). Für die 169.368 Datensätze des Mikrozensus betrug die Einzelfallquote 38,7 Prozent. Berücksichtigt man nur die bei der Verwendung einer einfachen Abgleichtechnik relevante Zielpopulation, d.h. Fälle, deren Merkmalsausprägungen auch im Gelehrtenkalender enthalten sind, stehen im Mikrozensus noch 3099 Datensätze zur Verfügung. Hiervon weisen 79,6 Prozent eine einzigartige Ausprägungskombination auf.

In einer ersten Angriffssituation wurde unterstellt, daß ein Angreifer über einen Massenfischzug versucht, möglichst viele Fälle des Mikrozensus zu reidentifizieren. Hierbei könnte er den hohen Auswahlsatz des Gelehrtenkalenders nützen, indem er, ausgehend von der relevanten Subpopulation im Mikrozensus (n=3099) anstrebt, die zugehörigen Datensätze im Zusatzwissen zu ermitteln.

Übersicht 1: Überschneidungsmerkmale im Gelehrtenszenario

Gelehrtenkalender[9)] N=7983	Anzahl Merkmals-ausprägungen im Gelehrtenkalender
Gemeindegrößenklasse	9
Geschlecht	2
Geburtsjahr	61
Geburtsmonat	2
Wirtschaftszweig1	15
Wirtschaftszweig2	40
Stellung im Beruf	4
Beruf1	26
Beruf2	47
Ausgeübte Tätigkeit1	5
Ausgeübte Tätigkeit2	7
Fachricht. Hochschulabschluß1	71
Fachricht. Hochschulabschluß2	68
Bundesraumordnungsregion	16

Tabelle 1: Ergebnisse der experimentellen Überprüfung des Gelehrtenszenarios (Zuordnungskriterium: identische Ausprägungskombinationen)

	Teilnahmekenntnis: ja	nein
Anzahl der Fälle: Gelehrtenkalender*	7983	53
Mikrozensus	3099	151
Anteil Einzelfälle (gerundet)	80%	80%
eins-zu-eins Zuordnungen	14	9
hiervon korrekt	4	9
mehrdeutige Zuordnungen	15	2
(potentiell korrekt)	(6)	(1)
Wahrscheinlichkeit zu einem Fall des Gelehrtenkalenders den entsprechenden Datensatz im Mikrozensus zu lokalisieren	(4/7983) 0,0005	(9/53) 0,16

* Die Fälle des Gelehrtenkalenders wurden jeweils für beide Situationen in 16 unterschiedlichen Verkodungsvarianten überprüft.

In einer zweiten Angriffssituation wurde die Risikokonstellation insofern verschärft, als unterstellt wurde, daß der Angreifer über Teilnahmekenntnisse verfügt, also weiß, welche in seinem Zusatzwissen enthaltenen Personen an der Mikrozensuserhebung teilgenommen haben. Unter dieser zusätzlichen Randbedingung wäre es möglich, das Identifikationsfile auf jene 53 Fälle zu beschränken, die nach der Überprüfung des Treuhänders in beiden Datenbeständen erfaßt wurden. Analog wurde die Zielpopulation im Mikrozensus eingegrenzt, indem alle Fälle ausgeschlossen wurden, die von den 53 gesuchten Fällen des Gelehrtenkalenders abweichende Ausprägungen aufwiesen. Von den verbleibenden 151 Mikrozensusdatensätzen weisen 80 Prozent eine einzigartige Ausprägungskombination auf.

Wie Tabelle 1 zu entnehmen ist, variiert das Risiko hinsichtlich dieser Szenarien. Besteht keine Teilnahmekenntnis, können 14 der 7983 Datensätze des Gelehrtenkalenders in einer eins-zu-eins Relation 14 Datensätzen des Mikrozensus zugeordnet werden. Das heißt, diese Datensätze weisen im Mikrodatenfile und im Zusatzwissen eine je einzigartige und je identische Ausprägungskombinationen für die zehn Überschneidungsmerkmale auf. Weiteren 15 Datensätzen des Mikrozensus werden 29 Datensätzen des Gelehrtenkalenders in mehrdeutiger Weise zugeordnet.[10] Bei mehrdeutigen Zuordnungen kann ohne zusätzliche Informationen nicht entschieden werden, welche der zugeordneten Datensätze sich auf eine spezifische Person beziehen. Eine Reidentifikation wäre daher auf der Basis der im Gelehrtenkalender enthaltenen Informationen nicht möglich. Aber auch eine eindeutige Zuordnung ist nicht schon per se gleichbedeutend mit einer Reidentifikation. Nach der Überprüfung der Ergebnisse durch den Treuhänder beziehen sich lediglich vier der 14 eindeutigen Zuordnungen auch in der Realität auf ein und dieselbe Person. Die Wahrscheinlichkeit, für eine beliebige im Gelehrtenkalender enthaltene Person den entsprechenden Datensatz im Mikrozensus zu lokalisieren, ist mit 0,0005 daher äußerst gering. Zugleich liegt bei den wenigen zugeordneten Datensätzen die Wahrscheinlichkeit einer falschen Zuordnung mit 0,71 wesentlich höher als die einer korrekten Zuordnung.

Obwohl sich die absolute Zahl der potentiell korrekt zuordenbaren Datensätze nicht verändert, ist für die Angriffssituation "Teilnahmekenntnis" die Wahrscheinlichkeit einer Reidentifikation mit 0,17 höher als in der Situation "ohne Teilnahmekenntnis". Dies ist darauf zurückzuführen, daß sich ein Angreifer hier - wie oben ausgeführt - auf jene 53 Fälle im Zusatzwissen konzentrieren kann, für welche er sicher weiß, daß sie an der Mikrozensuserhebung teilgenommen haben. Auf diese Weise werden störende statistische Doppelgänger aus dem Zusatzwissen ausgegrenzt und ursprünglich mehrdeutige Zuordnungen liegen nun in eindeutiger Weise vor.

Hierdurch bedingt, ist es nun für neun der 53 gesuchten Fälle möglich, den entsprechenden Datensatz im Mikrodatenfile in eindeutiger Weise zu lokalisieren. Dennoch wird auch unter der Bedingung "Teilnahmekenntnis" die überaus wichtige Schutzfunktion von Dateninkompatibilitäten deutlich: Selbst wenn die mehrdeutigen Zuordnungen ebenfalls berücksichtigt werden, sind von den 53 in beiden Datenfiles erfaßten Fällen 81 Prozent so gelagert, daß sie in mindestens einem Überschneidungsmerkmal inkompatibel abgebildet sind. Diese Fälle sind daher mit einer einfachen Abgleichtechnik auch dann nicht zuordenbar, wenn beliebig viele zusätzliche Überschneidungsmerkmale zur Verfügung stehen würden (vgl. Müller et al. 1991:54).

Folgerungen für die faktische Anonymität:
Das Bundesstatistikgesetz definiert die faktische Anonymität anhand der Unverhältnismäßigkeit des Aufwandes an Zeit, Kosten und Arbeitskraft, der für die Herstellung eines Personenbezugs notwendig ist. Allerdings ist im Gesetzestext über die angegebene Formulierung hinaus nicht näher konkretisiert, im Vergleich zu welchem Gut der Aufwand abzuwägen ist. Dies ist nur implizit zu entnehmen. Im wesentlichen handelt es sich hierbei um das Verhältnis zwischen dem Wert beziehungsweise dem durch einen Reidentifikationsversuch erzielten Nutzen und den hierfür anfallenden Kosten (vgl. Helmcke 1989, Müller et al. 1991:212ff.). Dieser Nutzen läßt sich objektiv am ehesten im Vergleich zu den Kosten einer alternativen Informationsbeschaffung ermessen. Im folgenden wird deshalb das Unverhältnismäßigkeitskriterium durch einen Vergleich mit alternativen Informationsbeschaffungskosten bestimmt.

Legt man einen solchen Vergleich zugrunde, so war die faktische Anonymität unter der Randbedingung "Massenfischzug" zweifellos gegeben. Allein für die Aufbereitung des Gelehrtenkalenders fielen Kosten in Höhe von circa 44.000 Mark an. Für die Vorbereitung und Durchführung der Experimente einschließlich der Rechenzeit (ohne Testläufe) müssen circa 17.000 Mark veranschlagt werden. Der damit für die Ermittlung der 14 eindeutigen Zuordnungen angefallene Betrag von 61.000 Mark stellt eine untere Grenze dar, da weitere Kosten entstanden wären, um die vier korrekten Zuordnungen zu ermitteln. Im Vergleich dazu würde bei einer Alternativbeschaffung dieser Informationen - etwa durch ein professionelles Umfrageinstitut - im Schnitt zwischen 120 und 150 Mark pro Interview anfallen.[11] Selbst unter der Bedingung, daß die obere Grenze bei 150 Mark angesetzt wird, hätte das Reidentifikationsrisiko damit noch um ein Vielfaches höher liegen können, und die faktische Anonymität wäre dennoch gegeben gewesen.

Eine unmittelbare Ermittlung der unter der Annahme "Teilnahmekenntnis" anfallenden Kosten ist nicht möglich, da wir unser "Teilnehmerwissen" über den Treuhänder bezogen haben. Es ist zwar plausibel, daß ein Angreifer eventuell für einige wenige Personen weiß, daß sie an einer amtlichen Erhebung teilgenommen haben, aber sehr unwahrscheinlich, daß dieses Wissen beispielsweise für alle in einem Handbuch enthaltenen Fälle besteht. Wenn wir dennoch von dieser äußerst unrealistischen Annahme ausgehen und stark vereinfachend unterstellen, daß keine zusätzlichen Recherchekosten anfallen, kann der Betrag von 61.000 Mark proportional umgerechnet werden. Ausgehend von den 7983 Fällen des Gelehrtenkalenders hätte der Aufwand für die Überprüfung eines einzelnen Falles etwa 7,60 Mark betragen. Für die Überprüfung der 53 gesuchten Fälle und die Ermittlung der neun korrekten Zuordnungen wären dann Kosten in Höhe von etwas mehr als 400 Mark angefallen. Legt man wiederum die obigen Interviewkosten mit 150 Mark pro Interview zugrunde, hätte die Alternativbeschaffung der gewünschten Informationen etwa 1350 Mark gekostet, so daß das Unverhältnismäßigkeitskriterium nicht erfüllt gewesen wäre. Wie erwähnt, ist bei dieser Kalkulation vorausgesetzt, daß ein Angreifer - bezogen auf ganz Nordrhein-Westfalen - weiß, welche der im Gelehrtenkalender erfaßten Personen auch an der Mikrozensuserhebung teilgenommen haben und sich diese Informationen nicht erst mühsam beschaffen muß. Unabhängig hiervon zeigt sich jedoch, daß man bei einer deutlich von den durchschnittlichen Merkmalsträgern abgrenzbaren Subpopulation, wie beispielsweise den "Gelehrten", unter der Randbedingung "Teilnahmekenntnis" von einem erhöhten Gefährdungspotential ausgehen muß und dies bei entsprechenden Schutzvorkehrungen zu berücksichtigen ist.

4.1.2 Diskriminanzanalytische Methode nach Paaß/Wauschkuhn

Bei gleicher Datenkonstellation wurde im weiteren überprüft, ob und inwieweit die Nutzung einer Reidentifikationstechnik, welche Dateninkompatibilitäten und statistische Doppelgänger in der Grundgesamtheit berücksichtigt, die Erfolgschancen einer Reidentifikation erhöht (Blien/Müller 1991).

Die Verwendung dieser Methode setzt die Spezifizierung eines Fehlerprozesses und damit Informationen über Strukturen und Ausmaß möglicher Dateninkompatibilitäten voraus. Derartige Kenntnisse liegen in den Sozialwissenschaften bislang allerdings nur bruchstückhaft vor (vgl. u.a. Esser 1986; Schnell/Hill/Esser 1988). Es gibt Untersuchungen zu Veränderungsraten im Zeitablauf, der Reliabilität von Erhebungsinstrumenten und der Häufigkeit von Erhebungsfehlern (Koch 1986; Porst/Zeifang 1987; Schwarz/Hippler/Strack 1988). Jüngere Untersuchungen beschäftigen sich auch mit der Wirkung unterschiedlicher Erhebungskontexte auf das

Antwortverhalten (Schwarz/Hippler/Noelle-Neumann 1989). Eine abgeschlossene Theorie liegt bislang nicht vor.

Die sich hierdurch ergebenden Probleme bei der Festlegung von Fehlerprozessen begründen zwar generelle Zweifel an der Einsatzfähigkeit dieses Verfahrens. Ein zentrales Ergebnis der AIMIPH-Studie bestand jedoch darin, daß das Verfahren relativ robust auf Fehlspezifikationen (in bezug auf die Höhe der Fehler) reagiert (Paaß/Wauschkuhn 1985:186; Paaß 1987). Daher wurde analog zum Vorgehen von Paaß/Wauschkuhn ein Fehlerprozeß spezifiziert (Müller 1991). Um den Problemen einer adäquaten Fehlerschätzung gerecht zu werden, wurden (in Höhe und Struktur) verschiedene Fehlerprozesse spezifiziert, wobei insgesamt fünf unterschiedliche Datenkonstellationen geprüft wurden (vgl. Müller et al. 1991:283).

Bei einer Sicherheitsschwelle von 99 Prozent wurden über alle fünf Situationen insgesamt 29 unterschiedliche Zuordnungen ermittelt. Die Überprüfung der Ergebnisse durch den Treuhänder ist allerdings desillusionierend: von den 29 Zuordnungen bezogen sich nur drei auch in der Realität auf ein und dieselbe Person. Alle drei Zuordnungen waren bereits mit der einfachen Abgleichtechnik ermittelt worden, weil die betreffenden Ausprägungskombinationen kompatibel abgebildet waren. Was sich zunächst als Vorteil dieser Methode darstellt, nämlich die höhere Zahl von Zuordnungen, erweist sich bei näherer Betrachtung insofern als Nachteil, als lediglich der Anteil der Falschzuordnungen steigt. Dieser liegt mit etwa 90 Prozent sogar noch höher als bei einem einfachen Abgleich der Ausprägungskombinationen. Die ausgewiesen hohe Sicherheitsschwelle von 99 Prozent trifft sowohl auf falsche wie auf korrekte Zuordnungen zu. Analog zu einer einfachen Abgleichtechnik ist es daher nicht möglich, zwischen korrekten und falschen Zuordnungen zu unterscheiden. [12)]

Folgerungen für die faktische Anonymität:
Setzt man dieses Ergebnis in Relation zum erbrachten Aufwand, war das Unverhältnismäßigkeitskriterium bei Verwendung der diskriminanzanalytischen Methode bereits durch die Vorbereitungsarbeiten erfüllt. Da das Verfahren nicht standardmäßig zur Verfügung steht und aus plausiblen Gründen auch nur spärlich dokumentiert ist, nahm die Rekonstruktion des Algorithmus, die Programmanpassung an die Datenstruktur und die Implementation auf den vorhandenen Rechner etwa ein Jahr Arbeitszeit in Anspruch. Erst dann war die Generierung der Fehlerverteilung und die Durchführung der Experimente möglich. Hinzu kommen hohe Rechenkosten, da der Algorithmus sehr viel CPU-Zeit und Speicherplatz beansprucht. Die Gesamtkosten beliefen sich auf etwa 261.000 Mark (vgl. Müller et al. 1991:311). Der Einsatz einer solch aufwendigen Methode innerhalb des

Wissenschaftskontextes kann damit ebenso wie die eigene Entwicklung eines äquivalenten Algorithmus mit an Sicherheit grenzender Wahrscheinlichkeit für ein Datenangriffsszenario ausgeschlossen werden. Auf eine weitere Überprüfung dieser Reidentifikationstechnik wurde daher verzichtet.

4.3 Empirische Überprüfung des sozialwissenschaftlichen Szenarios

4.3.1 Datenbasis

Die spezifische Risikokonstellation des Gelehrtenszenarios ergab sich aus dem klaren Bezug auf eine deutlich abgrenzbare Subpopulation und deren nahezu vollständiger Erfassung. Im Gegensatz hierzu weist die sozialwissenschaftliche Erhebung einen relativ kleinen Auswahlsatz auf, der einen repräsentativen Querschnitt der Bevölkerung widerspiegelt. Für die hier erfaßten Personen steht mit 35 Überschneidungsmerkmalen zum Mikrozensus allerdings ein sehr umfangreiches Zusatzwissen zur Verfügung. Von besonderem Interesse sind die zahlreichen Merkmale zum Haushaltskontext der Befragten. In verschiedenen Untersuchungen wird darauf hingewiesen, daß schon einige wenige Haushaltsmerkmale zu einzigartigen Ausprägungskombinationen in einem Datenfile führen können (Brunnstein 1987; Fischer-Hübner 1986; Greenberg 1990). Da davon auszugehen ist, daß gerade im Alltagswissen Informationen über Haushalte von Dritten verfügbar beziehungsweise relativ einfach beschaffbar sind, wird daher ein erhöhtes Reidentifikationsrisiko unterstellt, wenn ein Mikrodatenfile detaillierte Angaben über den Haushaltskontext enthält.

Um das von Haushaltskontextmerkmalen ausgehende Gefährdungspotential empirisch zu präzisieren, wurden die Reidentifikationsexperimente in drei Phasen durchgeführt. In einer ersten Phase wurden nur Merkmale berücksichtigt, die sich auf eine spezifische Person im Haushalt beziehen. In einer zweiten Phase wurde zusätzlich der allgemeine Haushaltskontext einbezogen. In einer dritten Phase schließlich wurden detaillierte Haushaltsinformationen auch über die im Haushalt lebenden Partner berücksichtigt. Eine detaillierte Auflistung der jeweils berücksichtigten Überschneidungsmerkmale gibt Übersicht 2.

4.3.2 Ergebnisse

Analog zum Gelehrtenszenario wird zwischen den Angriffssituationen "keine Teilnahmekenntnis" und "Teilnahmekenntnis" unterschieden. Im folgenden werden zunächst die Ergebnisse der Angriffssituation "keine Teilnahmekenntnis" dargestellt.

Übersicht 2: Sozialwissenschaftliches Szenario: Überschneidungsmerkmale in Zuordnungsphase 1 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		3	21	35

* Diese Angaben beziehen sich auf die sozialwissenschaftliche Erhebung, da nur die dort auftretenden Merkmalsausprägungen für die Zuordnungen mit einer einfachen Reidentifikationsmethode relevant sind.

Tabelle 2: Ergebnisse der experimentellen Überprüfung des sozialwissenschaftlichen Szenarios unter der Annahme, daß *keine Teilnahmekenntnis* vorliegt (Zuordnungskriterium: identische Ausprägungskombinationen)

Phase	I	II	III
Zahl der Überschneidungsmerkmale	13	21	35
Anzahl der Fälle:			
Sowi. Stichprobe	2685	2685	2685
Mikrozensus	94.747	94.747	53.441
Anteil Einzelfälle (gerundet)	20%	79%	84%
Zuordnungen (insgesamt)	1107	298	74
Wahrscheinlichkeit zu einem Fall der sowi. Erhebung den entsprechenden Datensatz im Mikrozensus zu lokalisieren	0	0	0

Tabelle 3: Ergebnisse der experimentellen Überprüfung des sozialwissenschaftlichen Szenarios unter der Annahme, daß *Teilnahmekenntnis* vorliegt (Zuordnungskriterium: identische Ausprägungskombinationen)

Phase	I	II	III
Sowi. Erheb. Fall Nr:	Anzahl der berücksichtigten Überschneidungsmerkmale (in Klammern)	13	21
1)	1 (12)	0	0
2)	55 (11)	2	0
3)	22 (10)	0	0
4)	137 (11)	0	0
5)	53 (12)	12	0
6)	235 (12)	159	1
7)	192 (9)	0	0
8)	25 (9)	0	0
9)	588 (9)	0	0
10)	27 (12)	5	0

Gemessen am Anteil der Datensätze mit einzigartigen Ausprägungskombinationen ist den Haushaltsinformationen ein hohes Risikopotential zuzuschreiben. Stehen nur die personenspezifischen Überschneidungsmerkmale zur Verfügung, weist das Mikrodatenfile eine Einzelfallquote von lediglich 19,6 Prozent auf. Werden alle 35 Merkmale berücksichtigt, sind 84 Prozent der Personen in dem Mikrodatenfile durch eine einzigartige Ausprägungskombination gekennzeichnet.

Wie aus Tabelle 2 hervorgeht, spiegelt sich die mit den haushaltsbezogenen Merkmalen einhergehende Trennschärfe auch in den Zuordnungsquoten wider: Von den 1107 aufgrund von identischen Wertekombinationen ermittelten ein- bzw. mehrdeutigen Zuordnungen in Phase I verbleiben 298, wenn die allgemeinen Haushaltsinformationen berücksichtigt werden. Werden alle 35 Überschneidungsmerkmale als Zusatzwissen eingesetzt, reduziert sich die Zahl der Zuordnungen auf 74. Hiervon sind 35 eindeutig.

Wie Tabelle 2 entnommen werden kann, ist die Einzelfallquote in den Daten auch in diesem Szenario kein Indikator für ein real bestehendes Reidentifikationsrisiko. Alle hier auf der Basis identischer Ausprägungskombinationen ermittelten Zuordnungen waren falsch: für keinen der zehn Fälle, die - nach der Überprüfung des Treuhänders - in beiden Erhebungen erfaßt wurden, war es möglich, die entsprechenden Partnerdatensätze aus Mikrozensus und sozialwissenschaftlicher Erhebung korrekt zuzuordnen.

Noch stärker als im Gelehrtenszenario wirken damit Dateninkompatibilitäten und statistische Doppelgänger als Schutz vor erfolgreichen Reidentifikationsversuchen: Durch einen Vergleich der zehn Partnerdatensätze im Mikrozensus und in der sozialwissenschaftlichen Erhebung konnte gezeigt werden, daß schon in Phase I lediglich fünf von dreizehn Merkmalen für alle zehn Datensätze kompatibel abgebildet waren. Bei Berücksichtigung nur dieser fünf Merkmale wies jedoch keiner der zehn gesuchten Datensätze eine eindeutige Ausprägungskombination auf (Müller et al. 1991:335). Es müßten daher weitere Überschneidungsmerkmale berücksichtigt werden. Alle weiteren Merkmale sind jedoch mit einer gewissen Wahrscheinlichkeit von Inkompatibilitäten betroffen, die eine korrekte Zuordnung verhindern.

Das Dilemma zwischen der notwendigen Eindeutigkeit eines Datensatzes einerseits und der mit jedem zusätzlich berücksichtigten Überschneidungsmerkmal steigenden Wahrscheinlichkeit von Inkompatibilitäten andererseits wird unter der Randbedingung "Teilnahmekenntnis" noch deutlicher. Hier wurde - bezogen auf die personenspezifischen Merkmale - als zusätzliche Annahme unterstellt, daß der Angreifer aufgrund von Plausibilitätsüberlegungen weiß, welche dieser Merkmale kompatibel abgebildet sind. Er könnte

sich bei einem Reidentifikationsversuch in einer ersten Phase daher auf jene Datensätze im Mikrozensus konzentrieren, die für diese Merkmale jeweils kompatibel abgebildet sind. Das Ergebnis dieser Suche ist Tabelle 3/Phase I zu entnehmen.

Wie aus Tabelle 3 hervorgeht, ist es selbst unter dieser äußerst riskanten und höchst unwahrscheinlichen Randbedingung nur für einen Datensatz (Phase I/Fall Nr.1) möglich, eine eindeutige (und korrekte) Zuordnung vorzunehmen. In allen anderen Fällen werden weitere Überschneidungsmerkmale benötigt, um statistische Doppelgänger auszuschließen. In der nächsten Phase (II) sind fünf Fälle aufgrund von Inkompatibilitäten nicht mehr zuordenbar. Bei vier Fällen scheint eine weitere Eingrenzung möglich. Die tatsächlich gesuchten Datensätze sind in diesen Merkmalen jedoch inkompatibel abgebildet und deshalb in der eingekreisten Subpopulation nicht mehr enthalten. Bedingt durch die Inkompatibilitäten würde der Suchprozeß in eine völlig falsche Richtung laufen. Noch deutlicher tritt dieser Sachverhalt in Phase III zutage. Hier ist zwar eine eins-zu-eins Zuordnung möglich, diese ist jedoch falsch. Dieser Sachverhalt würde einem Angreifer verborgen bleiben. Da er aufgrund seiner Teilnahmekenntnis sicher ist, daß die von ihm gesuchte Person im Mikrodatenfile enthalten sein muß, besteht für ihn kein Grund, an der Qualität dieser Zuordnung zu zweifeln: Solange sich für den gesuchten Fall zumindest ein identischer Datensatz in einem Mikrodatenfile ermitteln läßt, enthalten die Daten kein Indiz in bezug auf mögliche Verwechslungen aufgrund von Inkompatibilitäten.

Die im sozialwissenschaftlichen Szenario ermittelten Resultate ergänzen die Befunde aus dem Gelehrtenszenario insofern, als sich zeigt, daß es auch unter der Annahme von Teilnahmekenntnis nahezu unmöglich ist, eine beliebige Person anhand einiger weniger Merkmale zu reidentifizieren. Sofern eine gesuchte Person nicht eine in der Grundgesamtheit äußerst selten vertretene Ausprägungskombination in diesen Merkmalen aufweist, werden sich in der Grundgesamtheit und daher auch im Mikrodatenfile eine Vielzahl von statistischen Doppelgängern finden, die einen Reidentifikationsversuch nachhaltig stören. Das notwendige Zusatzwissen wirkt damit in paradoxer Weise auf den Deanonymisierungsprozeß: Je mehr Überschneidungsmerkmale im Zusatzwissen enthalten sind, desto höher ist der Anteil der einzigartigen Ausprägungskombinationen. Mit jedem zusätzlichen Merkmal erhöht sich jedoch zugleich auch die Wahrscheinlichkeit, daß ein gesuchter Fall in diesem Merkmal inkompatibel abgebildet ist, woraus Falsch- bzw. Nichtzuordnungen resultieren können.

Folgerungen für die faktische Anonymität:
Obwohl in dem hier untersuchten Szenario auf bereits maschinenlesbare Datenfiles zurückgegriffen werden konnte, waren umfangreiche Vorarbeiten notwendig, um die in der sozialwissenschaftlichen Erhebung enthaltenen Informationen als Überschneidungsmerkmale zum Mikrozensus nutzen zu können (vgl. Müller et al. 1991:259ff.). Hierbei entstanden für die Entwicklung des Konzepts, die Anpassung (Rekodierung) der Überschneidungsmerkmale und die Angleichung der Datenstruktur von Zusatzwissen und Mikrodatenfile, Arbeitskosten in Höhe von circa 15.000 Mark. Aufgrund der umfangreichen Vorarbeiten einschließlich der Durchführung der Zuordnungsexperimente entstanden zusätzlich Rechenkosten in Höhe von etwa 13.000 Mark. Der gesamte finanzielle Aufwand betrug circa 28.000 Mark. Angesichts des vorliegenden Ergebnisses, nach welchem selbst unter der Randbedingung "Teilnahmekenntnis" und nur unter äußerst extremen Zusatzannahmen, maximal eine korrekte Zuordnung möglich gewesen wäre, wäre eine Alternativbeschaffung der gewünschten Informationen in diesem Szenario in jedem Fall die kostengünstigere Alternative gewesen.

5. Zusammenfassende Bewertung der empirischen Überprüfung von Reidentifikationsrisiken

Als wichtigstes Ergebnis der empirischen Überprüfung ist festzuhalten, daß das Reidentifikationsrisiko bei der Verwendung von realen Daten weitaus niedriger ist, als dies bislang aufgrund wahrscheinlichkeitstheoretischer Berechnungen und mit ganz oder teilweise synthetisch generierten Daten durchgeführter Experimente zu vermuten war.

Eine wesentliche Erkenntnis war, 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 das Identifikationsfile im Gelehrtenszenario einer Vollerhebung nahekam, war die Zahl der mit Sicherheit richtig vorgenommenen Reidentifikationen äußerst gering. Der wichtigste Grund dafür ist die praktische Unvermeidbarkeit von Inkompatibilitäten zwischen unterschiedlichen Datenbeständen, die in ihrer Wirkung auf Reidentifikationsrisiken bislang unterschätzt wurden.

Gegen das Ergebnis einer "natürlichen" Schutzfunktion von Inkompatibilitäten könnte argumentiert werden, daß es im wesentlichen auf einer empirischen Illustration durch einige wenige Fälle beruht. Dem ist

entgegenzuhalten, daß sich auch in anderen Untersuchungen Hinweise auf Dateninkompatibilitäten finden. So berichten Marsh et al. (1991) von Abweichungen, die im Rahmen einer Nacherhebung zum britischen Zensus 1981 festgestellt wurden. Sie betrugen beispielsweise für das Merkmal Haushaltsgröße 2,4 Prozent, für Haus- und Wohnungseigentum 3,2 Prozent, für Erwerbstätigkeit 7,8 Prozent und für die Zugehörigkeit zu grob definierten Berufsklassen (sechs Kategorien) 13 Prozent. Ähnliche Befunde fanden sich auch bei einer Reanalyse der ALLBUS Test-Retest Studie. 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). Obwohl insbesondere die soziodemographischen Merkmale eine relativ hohe Stabilität aufweisen, unterscheiden sich zwischen Welle 1 und Welle 3 (vgl. Müller et al. 1991:124):

- 45 Prozent der Befragten bei den Angaben zur Einkommenshöhe,
- 18 Prozent hinsichtlich der geleisteten Arbeitswochenstunden,
- 16 Prozent bei den angegebenen beruflichen Ausbildungsabschlüssen,
- 13 Prozent hinsichtlich ihrer überwiegenden Einkünfte.

Auf individueller Ebene zeigen sich von Welle 1 zu Welle 2 sowie von Welle 1 zu Welle 3 ähnliche Abweichungen wie im sozialwissenschaftlichen Szenario. Bei elf berücksichtigten Merkmalen machen von Welle 1 zu Welle 2 lediglich 22 Prozent der Befragten identische Angaben. Knapp 39 Prozent weichen in einem Merkmal und weitere 39 Prozent in mindestens zwei Merkmalen ab (vgl. Übersicht 3).

Übersicht 3: ALLBUS Test-Retest-Studie 1984: Fallspezifische Häufigkeiten von aufgetretenen Inkompatibilitäten bei elf berücksichtigten Merkmalen[13] 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 (in %)	Welle 1 zu Welle 3 (in %)
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

Als Maßnahme zum Schutz gegen Deanonymisierung ist deshalb nicht vorrangig auf die Verhinderung einmaliger Ausprägungskombinationen abzustellen. 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).

Während eine massenhafte Deanonymisierung von Datensätzen nahezu ausgeschlossen ist und auch der Versuch, beliebige Personen zu reidentifizieren, in aller Regel scheitern wird, sind der empirischen Analyse jedoch auch Anhaltspunkte für eine Risikokonstellation zu entnehmen, bei welcher eine erfolgreiche Reidentifikation nicht ausgeschlossen werden kann. Dieser - allerdings äußerst unwahrscheinliche - Fall setzt das Zusammentreffen sehr spezifischer Risikofaktoren voraus und kann wie folgt charakterisiert werden: [14)]

1) Eine im Mikrodatenfile gesuchte Person gehört einer sehr kleinen, durch ein spezielles Merkmal identifizierbaren Subpopulation an (sachliche Tiefengliederung);
2) der Mikrodatenfile enthält stark differenzierte Regionalinformationen, so daß in den Regionaleinheiten nur wenige Personen der spezifischen Subpopulation leben (regionale Tiefengliederung);
3) der Datenangreifer weiß, daß die gesuchte Person im Mikrodatenfile enthalten ist (Teilnahmekenntnis);
4) die Merkmale der Person sind genau in der Weise im Mikrodatenfile erfaßt, wie es der Forscher vermutet (Kompatibilität).

Beim Zusammentreffen dieser vier Bedingungen erscheint die Reidentifikation von einzelnen Fällen ohne großen Aufwand als möglich. Bereits wenn eine der Bedingungen nicht gegeben ist, ist die Wahrscheinlichkeit einer Reidentifikation nach den durchgeführten Experimenten als äußerst gering einzustufen. Das gleichzeitige Zusammentreffen aller Bedingungen kann bei Stichprobenerhebungen zwar als außergewöhnlich seltenes Ereignis betrachtet werden. Dennoch sollten bei der Datenübermittlung Vorkehrungen getroffen werden, damit auch eine solche Risikokonstellation ausgeschlossen ist.

Neben obligatorischen vertraglichen Verpflichtungen, die beispielsweise eine Reidentifikation verbieten (vgl. Knoche 1991), sowie technisch-organisatorischen Sicherungsmaßnahmen, die insbesondere der Datenzugriffskontrolle dienen (vgl. Blien 1990), müssen auch datenorientierte Schutzmaßnahmen getroffen werden. In einem weiteren Schritt wurde daher die Wirkung ausgewählter Anonymisierungsmaßnahmen am Beispiel des

Gelehrtenszenarios überprüft (vgl. Müller et al. 1991:386ff.). Die empirische Überprüfung erfolgte im Hinblick auf die oben angeführten Bedingungen eins bis drei).

Zur Verringerung des Reidentifikationsrisikos aufgrund der Zugehörigkeit zu einer kleinen spezifischen Subpopulation oder der Existenz tiefgegliederter Regionalinformationen auf seiten des Angreifers, wurde die von Ausprägungsvergröberungen ausgehende Schutzwirkung geprüft.

Gegen das spezifische Gefährdungspotential der Teilnahmekenntnis wurde die Substichprobenziehung geprüft. Mit fallender Substichprobengröße verringert sich die Wahrscheinlichkeit, daß eine beliebige Person in den übermittelten Mikrodaten enthalten ist. Hierdurch erhöht sich in einem informationstheoretischen Sinn die Unsicherheit eines Angreifers über die Korrektheit möglicher Zuordnungen beträchtlich. Selbst wenn ein Angreifer weiß, daß eine spezifische Person an der Erhebung teilgenommen hat, kann er - bei der Übermittlung von Substichproben - auch bei einer eins-zu-eins Zuordnung nicht mehr sicher sein, ob es sich hierbei um die gesuchte Person oder einen statistischen Doppelgänger handelt.

Im folgenden Abschnitt werden die hieraus für die Wahrung der faktischen Anonymität abgeleiteten datenorientierten Empfehlungen dargestellt (für eine ausführliche Darstellung und Begründung siehe Müller et al. 1991:443ff.). Diese Empfehlungen beziehen sich nur auf den Mikrozensus und die EVS, weil die Untersuchung nur auf diese Datenfiles ausgerichtet war. Inwieweit diese Empfehlungen auf andere Datenbestände übertragbar sind, müßte gesondert untersucht werden.

6. Datenorientierte Empfehlungen für die Übermittlung faktisch anonymer Daten

Datenorientierte Schutzvorkehrungen beruhen letztendlich immer auf einer Reduktion der in den Daten enthaltenen Informationen, womit auch eine Verringerung des Analysepotentials einhergeht. Wenn ein gewisser Informationsverlust aus Datenschutzgründen auch unvermeidbar ist, so sollten die Anonymisierungsmaßnahmen dennoch so gestaltet sein, daß das Analysepotential möglichst geringfügig beeinträchtigt wird.

Das wissenschaftliche Potential von amtlichen 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. Dies gilt auch für regionalspezifische Analysen. Nur aufgrund der umfangreichen Stichproben vom amtlichen Erhebungen, wie z.B. des Mikrozensus, können aussagekräftige Analysen auch regional disaggregiert durchgeführt werden. Es ist davon auszugehen, daß die Wissenschaft amtliche Mikrodaten genau zu diesen Zwecken benötigt.

Die gleichzeitige regionale und sachliche Tiefengliederung ist jedoch ein wesentlicher Faktor der oben dargestellten Risikokonstellation. Bei bestimmten Analyseverfahren (tabellarischen Aufgliederungen) ist die gleichzeitige sachliche und regionale Tiefengliederung nicht sinnvoll, da Zellenbesetzungen sehr klein werden und wegen großer Zufallsschwankungen wenig aussagefähig sind. Hier sind entweder in der regionalen oder sachlichen Analysedimension aus statistischen Gründen Vergröberungen vorzunehmen. Bei multivariaten Analyseverfahren sind solche Aggregierungen nicht erforderlich und beeinträchtigen das Analysepotential. Dennoch erschien es als die sinnvollste Lösung, aus Datenschutzgründen bei der Datenweitergabe für den Regelfall für den Mikrozensus zwei unterschiedliche Datenfiles vorzusehen: ein sogenanntes Grundfile und ein sogenanntes Regionalfile. Mit Hilfe des Regionalfiles soll es möglich sein, Mikrodaten auf einer Ebene von Regionaleinheiten zu analysieren, für welche die Daten entsprechend dem Stichprobenplan noch als repräsentativ gelten. Bei der EVS sind dies die Bundesländer. Eine weitergehende Regionalisierung als im Grundfile des Mikrozensus vorgesehen, erschien daher für die EVS nicht sinnvoll.

Im Grundfile sind die Regionalinformationen nur in undifferenzierter Form enthalten. Sie schließen die Angabe über das Bundesland (außer für die Bundesländer Bremen und Saarland) und eine Klassifikation des siedlungsstrukturellen Typs ein. Empfohlen wurden hierbei Typisierungsmerkmale, wie zum Beispiel die von der Bundesforschungsanstalt für Landeskunde und Raumordnung entwickelte Gemeindetypologie, beziehungsweise eine vergröberte Klassifikation der Gemeindegrößenklasse als Alternative. Alle übrigen Merkmale sollen in möglichst großer Differenzierung enthalten sein, wobei auch der Haushaltszusammenhang der Befragten erhalten bleiben soll.

Das Regionalfile enthält stärker differenzierte Regionalinformationen, grenzt dafür aber die Differenzierungstiefe bei den übrigen Variablen ein. Damit wurden 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 - das haben die Experimente gezeigt - bereits durch die Entfernung der personenbezogenen Angaben faktisch anonym. Als zusätzliche Sicherung wird für das Grundfile empfohlen:

Mikrozensus:

Festlegung eines Minimums in den univariaten Randverteilungen, so daß jede ausgewiesene Merkmalsausprägung für die Grundgesamtheit der Bundesrepublik mindestens 5000 Fälle umfaßt. Dies entspricht circa 50 Fällen im Datensatz des Mikrozensus.

- Es darf keine einzelne Gemeinde eingrenzbar sein, die weniger als 500.000 Einwohner umfaßt.
- Ein Gemeindetyp (z.B. Gemeindegrößenklasse), dem mehrere Gemeinden zugehören, darf in keinem Bundesland weniger als 400.000 Einwohner umfassen.
- Angaben über Staatsangehörigkeit werden nur so weitergegeben, daß eine Nationalität oder eine identifizierbare Gruppe von Nationalitäten in der Bundesrepublik insgesamt wenigstens 50.000 Einwohner umfaßt. Dies entspricht circa 500 Fällen im Mikrozensus.

Einkommens- und Verbrauchsstichprobe:

"Sichtbare" oder über die Zeit vergleichsweise stabile Merkmale - wie Geburtsjahr, Stellung im Beruf oder Besitz auffälliger Konsumgüter - sollen so aggregiert werden, daß nur Merkmalsausprägungen ausgewiesen werden, die für die Grundgesamtheit der Bundesrepublik mindestens 5000 Fälle umfassen. Dies entspricht circa zehn Fällen in der EVS.

Bei öffentlich wenig bekannten oder über die Zeit wenig stabilen, jedoch differenziert erfaßten Merkmalen - im wesentlichen die nichtgruppierten 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.

Im Regionalfile ist die faktische Anonymität durch weitere Ausprägungsvergröberungen bei den sehr differenziert erfaßten Merkmalen Beruf, Wirtschaftszweig, Geburtsjahr und Nationalität zu sichern. Im Detail wurden folgende Einzelmaßnahmen für das Regionalfile des Mikrozensus vorgeschlagen:

Regionalfile:

- Durch die Kombination von Regionalklassifikationen soll keine Regionaleinheit ermittelbar sein, die eine Einwohnerzahl von weniger als 100.000 Personen aufweist.
- Die Überschneidungsmerkmale Beruf, Wirtschaftszweig, Nationalität und Alter sollen so vergröbert werden, daß keine Ausprägungen ausgewiesen werden,
  - die in der Grundgesamtheit nicht wenigstens 50.000 Einwohner umfassen;
  - die 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.
- Alle übrigen Variablen sollen - falls erforderlich - so aggregiert werden, daß jede ausgewiesene Merkmalsausprägung für die Grundgesamtheit der Bundesrepublik mindestens 5000 Fälle umfaßt.

Weiterhin wurde die Ziehung von Substichproben empfohlen. Die Substichprobenziehung verhindert, daß ein Datenangreifer mit Sicherheit weiß, ob eine bestimmte Person im übermittelten Mikrodatenfile enthalten ist. Hierdurch wird der Unsicherheitsfaktor bei einem Reidentifikationsversuch erhöht. Zugleich verringert sie prinzipiell das Reidentifikationspotential. Die Ziehung von Substichproben 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 circa 200.000 Einwohnern. Für solche Einheiten enthält der Mikrozensus circa 2000 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 wurde 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. Insbesondere in Analysen mit multivariaten tabellarischen Aufgliederungen werden auch bei sehr umfangreichen Stichproben wie dem Mikrozensus sehr schnell die Grenzen sichtbar, unterhalb derer eine Substichprobenziehung an der

Substanz der Analysemöglichkeiten des Mikrozensus rührt. Es wurde deshalb empfohlen, beim Grundfile als unterste Grenze eine Substichprobe von 70 Prozent anzusetzen.

Bei der EVS soll das Datenfile in Abhängigkeit der benötigten Erhebungsteile zukünftig mit folgenden Auswahlsätzen weitergegeben werden:

- 98 Prozent:
  Haushalts- und Personenmerkmale aus dem Grundinterview + 1 Erhebungsteil, [15)]
- 90 Prozent:
  Haushalts- und Personenmerkmale aus dem Grundinterview + 2 Erhebungsteile, [15)]
- 80 Prozent:
  Haushalts- und Personenmerkmale aus dem Grundinterview + 3 Erhebungsteile. [15)]

Ergänzend wurde die Weitergabe einer Ein-Prozent-Substichprobe aus dem Mikrozensus empfohlen. Diese sollte, mit Ausnahme der Regionalangaben, sämtliche Merkmale des Mikrozensus ohne weitergehende Anonymisierungsmaßnahmen enthalten. Durch den Wegfall von Regionalinformationen ist die starke Substichprobenziehung eine hinreichende Schutzmaßnahme. Dieses Subfile ist für Analysen gedacht, in welchen eine Massenbasis nicht erforderlich ist oder für Wissenschaftler, die eine Massendatenanalyse nicht selbst vornehmen, jedoch an einem kleineren Datenfile geplante Analysen testen wollen.

7. Ausblick

Die hier dargestellten Empfehlungen sind eine erste Konkretisierung für faktisch anonymisierte Mikrodatenfiles der EVS und des Mikrozensus, die in einem Standardfall angewandt werden können. Nach einiger Zeit der Praxis und Erfahrungssammlung sollten sie nochmals 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. Es ist zwar angestrebt, durch diese Epfehlungen eine gewisse Routinisierung und damit auch Aufwandseinsparung, sowie - falls möglich - eine

Kostensenkung bei der Datenweitergabe zu erreichen. Es wird 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 auf spezifische Forschungsvorhaben abgestimmt sind (Knoche 1991).

Anmerkungen

1) Mikrodaten beziehen sich - im Gegensatz zu Makrodaten oder Aggregatdaten - auf Informationen über einzelne Elementareinheiten, daher werden sie gelegentlich auch als Einzelangaben bezeichnet. Als Elementareinheiten kommen hierbei sowohl Personen bzw. Individuen, Betriebe, Institutionen oder sonstige Objekte in Frage, sofern sie Gegenstand einer Datensammlung sind. Beziehen sich die Mikrodaten auf Individuen, werden sie auch als Individualdaten bezeichnet (vgl. Müller et al. 1991:1).
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.
3) Die Finanzierung der Hauptkosten des Anonymisierungsprojektes (1988-1990) erfolgte durch das Bundesministerium für Forschung und Technologie. Eine ausführliche Darstellung der Projektergebnisse findet sich in Müller/Blien/Knoche/Wirth (1991): Die faktische Anonymität von Mikrodaten.
4) Im allgemeinen spricht man hierbei von Personenbeziehbarkeit. Die Abgrenzungsprobleme zwischen Personenbezug, Personenbeziehbarkeit und Anonymität werden unter anderem bei Brennecke (1980) diskutiert. Für eine kritische Diskussion von Personenbeziehbarkeit als Prinzip siehe Scheuch (1987).
5) Die wichtigsten allgemeinen theoretischen Ansätze zu Fehlerquellen bei der Datenerhebung basieren einerseits auf Grundlagen der Kognitions-, Emotions- und Motivationspsychologie (Schwarz et al. 1988, 1989; Hippler et al. 1987; Schwarz 1990; Sudman/Bradburn 1974; Bradburn/Sudman 1979), andererseits auf einer allgemeinen Theorie des sozialen Handelns, bei welcher das Verhalten in Befragungssituationen als Spezialfall sozialen Handelns angesehen wird (Esser 1984a, b, c, d, 1986).
6) Für eine ausführliche Darstellung siehe Müller et al. (1991:49-85); Bender (1990); Müller, M. (1991).
7) Die angesprochenen Dimensionen können an einem Beispiel verdeutlicht werden. Hypothetisch wird unterstellt, daß ein Angreifer 100 Personen im Mikrozensus deanonymisieren will: Unter der Annahme, mit einer Reidentifikationstechnik könnte jede zehnte Person, deren Daten sowohl im Zusatzwissen als auch im Mikrozensus enthalten sind, reidentifiziert werden, wäre bedingt durch den Auswahlsatz des Mikrozensus von einem Prozent nur jeder tausendste Fall des Identifikationsfiles auch im Mikrozensus auffindbar. Für eine Reidentifikation von 100 Fällen, müßte das Identifikationsfile demnach 100.000 Datensätze enthalten.
8) In dieser Untersuchung wurden über 10.000 Bundesbürger befragt. Bezogen auf Nordrhein-Westfalen enthielt sie 2685 Fälle. Aus Datenschutzgründen wird, in Abweichung von der üblichen Praxis, der Name der Erhebung und der beteiligten Institutionen nicht genannt.
9) 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.

10) Im Falle einer mehrdeutigen Zuordnung liegt keine eins-zu-eins, sondern eine 1:n, n:1 beziehungsweise n:m Zuordnung von Datensätzen vor.
11) Für den ALLBUS 1990 wurde dem Umfrageinstitut pro Interview etwa 130 Mark bezahlt (inklusive Datenfilerstellung und Plausibilitätskontrollen).
12) Für eine detaillierte Erklärung dieser Ergebnisse siehe Müller et al. 1991:302ff.
13) Berücksichtigte Variablen (in Klammern: Anzahl der Merkmalsausprägungen): Geschlecht (2); Alter (58); Familienstand (5); Allgemeiner Schulabschluß (5); Beruf. Ausbildungsabschluß (5); Beruf. Erwerbstätigkeit derz.(22); Arbeitslos (2); Berufliche Stellung derz. (22); Berufliche Stellung früher (20); Arbeitswochenstunden (74); Monatliches Netto-Einkommen (56).
14) Vgl. hierzu auch die Ergebnisse der argumentativen Analyse einzelner Szenarien (Müller et al. 1991:351ff.).
15) Erhebungsteile in diesem Sinn sind das Schlußinterview, der Erhebungsteil über die Nahrungs- und Genußmittel sowie der Erhebungsteil über die Jahresrechnung.

Literatur

Beckmann, P., 1988: Die Bedeutung des Zusatzwissens vor dem Hintergrund einer potentiellen Deanonymisierung von Mikrozensus und Einkommens- und Verbrauchsstichprobe, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.5.

Bender, S., 1990: De-Anonymisierung von Individualdaten bei statistischen Erhebungen. Eine Diskussion des diskriminanzanalytischen Verfahrens von Paaß/Wauschkuhn (Diplomarbeit, Universität Mannheim).

Bender, S./Blien, U./Müller, M., 1990a: Grundidee der diskriminanzanalytischen Methode von PAASS und WAUSCHKUHN zur Zuordnung anonymisierter Datensätze, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.11.

Bender, S./Blien, U./Müller,M., 1990b: Implementation und erste Tests der diskriminanzanalytischen Methode von PAASS und WAUSCHKUHN zur Zuordnung anonymisierter Datensätze.Erfahrungsbericht und Abschätzung des notwendigen Arbeitsaufwandes, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.16.

Bethlehem, J.G./Keller, W.J./Pannekoek, J., 1990: Disclosure Control of Microdata. Journal of the American Statistical Association, Volume 85:38-45.

Blien, U., 1989: Technisch-organisatorische Sicherungsmaßnahmen gegen unbefugte Datenzugriffe bei faktisch anonymen Daten, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.8.

Blien, U./Müller, M., 1991: Empirische Überprüfung der Anonymität des Mikrozensus mit der diskriminanzanalytischen Methode von Paaß und "Kürschners Gelehrtenkalender" als Zusatzwissen, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.18.

Block, H./Olsson, L., 1976: Bakvägsidentifiering, in: Statistiskal Tidskrift 14:133-144. (Engl. Version: Backwardsidentification (unveröffentlicht)).

Bradburn, N./Sudman, S. and Associates, 1979: Improving Interview Method and Questionnaire Design. San Francisco: Jossey Bass.

Brennecke, R., 1980: Kriterien zur Operationalisierung der faktischen Anonymisierung, in: Kaase et al.:158-175.

Brennecke, R./Schneider H., 1977: Zur Problematik des Bundesdatenschutzgesetzes für die Forschung. SPES-Arbeitspapier Nr.63, Sozialpolitische Forschungsgruppe Frankfurt/ Mannheim.

Brunnstein, K., 1987: Über die Möglichkeit der Re-Identifikation von Personen aus Volkszählungsdaten, in: Appel, R. (Hrsg.): Vorsicht Volkszählung! Köln: Volksblattverlag, (2.Auflage).

Burkert, H., 1979: Die Eingrenzung des Zusatzwissens als Rettung der Anonymisierung? Datenverarbeitung im Recht, Bd. 8, Heft 1:63-75.

Burkert, H., 1980: Das Problem des Zusatzwissens, in: Kaase et al.:143-147.

Dalenius, T., 1977: Towards a Methodology for Statistical Disclosure Control. Statistical Review 5/1977:429ff.

Dalenius, T., 1986: Finding a Needle in a Haystack or Identifying Anonymous Census Records. Journal of Official Statistics 2:329-336.

Dalenius, T., 1988: Controlling Invasion of Privacy in Surveys, Continuing Education Program, Statistics Sweden.

Dittrich, K./Schlörer, J., 1985: Anonymisierung von Forschungsdaten. Bericht im Auftrag des Ministeriums für Wissenschaft und Kunst Baden-Württemberg, Mai 1985.

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.

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 37:314-336.

Fischer-Hübner, S., 1986: Zur Anonymität und Reidentifizierbarkeit statistischer Daten, Mitteilungen Nr. 143 des Fachbereichs Informatik der Universität Hamburg.

Greenberg, B., 1990: Disclosure Avoidance Research at the Census Bureau, paper presented at the 1990 Annual Research Conference, Bureau of the Census, Arlington, Virginia.

Hamacher, B., 1980: Resümee zu Datenschutzmaßnahmen, in Kaase et al.:219-224.

Helmcke, T., 1989: Allgemeine Kosten-Nutzen-Überlegungen zu Deanonymisierungsversuchen, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.6.

Hippler, H.-J./Schwarz, N./Sudman, S. (Hrsg.), 1987: Social Information Processing and Survey Methodology, New York/Heidelberg: Springer.

Kaase, M./Krupp, H.-J./Pflanz, M./Scheuch, E.K./Simitis, S. (Hrsg.), 1980: Datenzugang und Datenschutz. Königstein/Ts.:Athenäum.

Knoche, P., 1989: Außerberufliche Motive, Arbeitsbericht aus dem Anonymisierungsprojekt Nr.12.

Knoche, P., 1991: Der neue Leitfaden des Statistischen Bundesamtes für die Weitergabe von Einzeldaten des Mikrozensus und der Einkommens- und Verbrauchsstichprobe. Vortrag auf der Tagung "Faktisch anonyme Einzeldaten der amtlichen Statistik" in Mannheim, Dezember 1991.

Koch, A., 1986: Wie zuverlässig lassen sich Berufs- und Bildungsvariablen messen? Ergebnisse einer Test-Retest-Studie zur Allgemeinen Bevölkerungsumfrage der Sozialwissenschaften 1984, Diplomarbeit, Universität Mannheim.

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

Marsh, C./Skinner, C./Arber, S./Penhale, B./Openshaw, S./Hobcroft, J./Lievesley, D./Walford, N., 1991: The Case for Samples of Anonymized Records from the 1991 Census. Journal of the Royal Statistical Society, Vol.154.

Mohler, P., Ph./Kaase, M., 1980: Formen der Erhebung in der empirischen Sozialforschung, in: Kaase et al.:107-110.

Müller, M., 1991: Reidentifikation von Individualdaten: Experimentelle Überprüfung im Rahmen eines sozialwissenschaftlichen Szenarios mit der Methode von Paaß, Wauschkuhn (Diplomarbeit, Universität 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./Blien, U./Knoche, P./Wirth, H., 1991: Die faktische Anonymität von Mikrodaten (Band 19 der Schriftenreihe Forum der Bundesstatistik). Metzler-Poeschel, Stuttgart.

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):419-428.

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./Wauschkuhn, U., 1985: Datenzugang, Datenschutz und Anonymisierung. Analysepotential und Identifizierbarkeit von anonymisierten Individualdaten, München, Wien: R. Oldenbourg.

Porst, R./Zeifang, K., 1987: Wie stabil sind Umfragedaten? Beschreibung und erste Ergebnisse der Test-Retest-Studie zum ALLBUS 1984. ZUMA Nachrichten 20:8-31.

Scheuch, E.K., 1980: Die Weiterentwicklung des Datenschutzes als Problem der Sozialforschung, in: Kaase et al.:252-275.

Scheuch, E.K., 1987: Risikointerpretation beim Datenschutz, in: Statistisches Bundesamt 1987: 121-145.

Schnell, R./Hill, P.B./Esser, E., 1988: Methoden der empirischen Sozialforschung, Oldenbourg, München.

Schlörer, J., 1980: Anonymisierung von Mikrodaten: Technische Aspekte, in: Kaase et al.:118-142.

Schwarz, N., 1990: Assessing Frequency Reports of Mundane Behaviors: Contributions of Cognitive Psychology to Questionnaire Construction, ZUMA-Arbeitsbericht 1988/10, abgedruckt in: Hendricks, C., Clark, M. (Hrsg.): 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-Arbeitsbericht Nr. 89/18.

Schwarz, N./Hippler, H.-J./Noelle-Neumann, E., 1989: Einflüsse der Reihenfolge von Antwortvorgaben bei geschlossenen Fragen. ZUMA-Nachrichten 25:24-38.

Schwarz, N./Hippler, H.-J./Strack, F., 1988: Kognition und Umfrageforschung: Themen, Ergebnisse und Perspektiven. ZUMA-Nachrichten 22:15-28.

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.

Spruill, N., 1983: Testing Confidentiality of Masked Business Microdata, Working Paper, PRI 83-07.09, The Public Research Institute, Alexandria, Virginia .

Statistisches Bundesamt (Hrsg.), 1985:Datennotstand und Datenschutz, Ergebnisse des 1. Wiesbadener Gesprächs 30./31. Oktober 1984. Stuttgart: Kohlhammer.

Statistisches Bundesamt (Hrsg.), 1987: Nutzung von anonymisierten Einzelangaben aus Daten der amtlichen Statistik. Bedingungen und Möglichkeiten, Stuttgart, Mainz: Kohlhammer.

Südfeld, E., 1987: Anonymisierungsstandards und generelle Abwicklungsregelungen für Anforderungen nach anonymisierten Einzelangaben im Statistischen Bundesamt, in: Statistisches Bundesamt (1987).

Sudman, S./Bradburn, N., 1974: Response Effects in Surveys. Chicago: Aldine Publishing Company.

Zapf, W., 1985: Der Zugang der Wissenschaft zur statistischen Information - Forderung und Realität. In: Statistisches Bundesamt (1985).

Anhang

Die Ergebnisse des Anonymisierungsprojektes wurden auf einer Tagung am 11.12.1991 an der Universität Mannheim vorgestellt. Nachfolgend geben wir die auf der Tagung stattgefundene Podiumsdiskussion in leicht gekürzter Form wieder. Thema der Podiumsdiskussion war vor allem die Übertragbarkeit der Ergebnisse des Anonymisierungsprojektes auf andere Forschungsbereiche.

Podiumsdiskussion: Die neuen Erkenntnisse zur faktischen Anonymität und ihre Übertragbarkeit auf andere Daten- und Forschungsbereiche in der Wissenschaft und der amtlichen Statistik

Prof. Dr. Allerbeck, Universität Frankfurt:
Der Titel unserer Diskussion umfaßt vier Zeilen aus denen ich als operative Worte eigentlich zwei entnehme - "Übertragbarkeit" und "andere". Der Kreis, der hier am Podium sitzt ist so illuster, daß es zuviel Zeit kosten würde ihn vorzustellen. Ich will dies unterlassen und einfach die Teilnehmer in der Reihenfolge der Liste bitten, ihren ersten Beitrag abzugeben.

Dr. Schmidt, Bundesbeauftragter für den Datenschutz, Bonn:
Zunächst möchte ich mich für die Gelegenheit bedanken, daß ich hier so interessante und in jeder Richtung fundierte Beiträge zu einem schwierigen Thema hören konnte, das nicht zu unrecht als Dilemma bezeichnet wurde. Der Versuch ist hiermit gelungen, die in Paragraph 16, Absatz 6 Bundesstatistikgesetz als Voraussetzung einer Datenübermittlung an die Forschung genannte faktische Anonymisierung zu beschreiben und sie nicht nur zu beschreiben, sondern auch erreichbar zu machen. Über diese abstrakte Bemerkung hinaus ist auch noch plausibel dargelegt worden, nicht nur daß, sondern auch mit welchen einfachen Mitteln und aus meiner Sicht für die Forschung auch erträglichen Mitteln, die Daten faktisch hinreichend gut anonymisiert werden können. Wenn nun trotzdem auch für diese faktisch anonymisierten Daten Sicherungsmaßnahmen und weitere Auflagen und Einschränkungen geboten sind, und das bedeutet im Ergebnis, daß der gedankliche Abstand zum sogenannten public file noch ganz erheblich ist, dann sollen diese flankierenden Maßnahmen den aufgezeigten Weg nicht versperren, sondern gewährleisten, daß er auch wirklich mit Erfolg beschritten werden kann. Ich halte es deshalb für angebracht, daß die Überlassung faktisch anonymisierter Daten maßgeschneidert durchgeführt wird und das fallweise geprüft wird. Beides wird durch die hier diskutierten Erkenntnisse aus meiner Sicht erheblich erleichtert und praktikabel. Damit

scheint mir eine Entwicklung in Richtung auf eine wirklich zu Buche schlagende Aufwandsminderung durch wiederholbare Vorgehensweise sehr leicht möglich. Dies dürfte die Durchführbarkeit von Forschungsvorhaben schon in einer sehr frühen Planungsphase weit besser kalkulierbar machen.

Prof. Dr. Dr. Häfner, Zentralinstitut für seelische Gesundheit, Mannheim:
Ich muß zunächst darauf verweisen, daß ich Mediziner bin - mein Arbeitsgebiet ist die psychiatrische Epidemiologie. Vorweg möchte ich das Fazit ziehen, daß dieses Projekt meiner Wissenschaft direkt wenig Nutzen bringt, aber dennoch in einem weiteren Zusammenhang erheblichen Nutzen bringen könnte, so hoffe ich jedenfalls. Relativ wenig Nutzen haben wir insofern, als der größte Teil der medizinischen Forschung, und dazu zählt auch die analytische Epidemiologie, auf Identifikatoren oder mindestens sprechende Codes angewiesen ist, damit Informationen aus unterschiedlichen Quellen und über Zeit einem Individuum zugeordnet werden können. Der zweite Grund liegt darin, daß die Daten der amtlichen Statistik der Bundesrepublik über die Gesundheit der Bevölkerung relativ wenig tief gegliederte, valide Informationen enthalten. Es gibt im Grunde außer der Todesursachenstatistik, deren Daten wenigstens in einigen Bereichen hinreichend valide sind, um als zuverlässige Gesundheitsindikatoren auf der Makroebene zu dienen, eigentlich nur die Daten des Mikrozensus. Sie stellen ein Gemenge aus subjektiven und objektiven Gesundheitsvariablen dar. Ihre geringe Validität läßt sich bereits an einem schlichten Vergleich der Erhebungsergebnisse über verschiedene Querschnitte hinweg feststellen. Sie weisen in einigen Kategorien so erhebliche Unterschiede auf, daß eine Erklärung durch Morbiditätstrends nicht mehr plausibel ist.

Daß die deskriptive Epidemiologie und die Gesundheitssystemforschung unter diesen ungünstigen Bedingungen leiden, läßt sich auch an ihren Entwicklungsdefiziten ablesen. Besonders deutlich wird dies beim Defizit der Public-Health-Forschung in der Bundesrepublik, zu deren Grundlagenwissenschaften das epidemiologische Studium der Beziehung zwischen wirtschaftlichen, sozialen und ökologischen Indikatoren einerseits und Gesundheitsindikatoren andererseits zählt. Der Bundesforschungsminister hat dieser Tage ein hochdotiertes Programm für die Förderung und Institutionalisierung von Public-Health-Forschung aufgelegt, das die Situation im Lande in eindrucksvoller Weise widerspiegelte: einen erschreckenden Mangel an anspruchsvoller epidemiologischer Forschung, aber auch an hinreichend zuverlässigen Gesundheitsdaten sowohl auf Bundesebene als auch auf der Ebene der Länder, Gebietskörperschaften und Kommunen, die als Grundlagen für gute epidemiologische Public-Health-Forschung geeignet wären. Man kann Public-Health-Forschung nur sinnvoll betreiben, wenn ein hinreichender Pool zuverlässiger Daten auf diesem

Gebiet zur Verfügung steht. Hier ist nicht nur eine bessere Kooperation zwischen Datenschutz, Wissenschaft und Gesetzgeber gefordert; hier muß auch mehr Problembewußtsein, vor allem gegenüber dem von unkritischen Datenschützern oft gebrauchten Wort der "Datensammelwut", zum Tragen kommen.

Wenn Public-Health-Forschung mehr an den Veränderungen des Gesundheitszustands der Bevölkerung im Zusammenhang mit Risikofaktoren und im Kontext von ökologischen und Bevölkerungsvariablen interessiert ist, so sind die analytische Epidemiologie und der größte Teil der medizinischen Forschung überhaupt, wie schon eingangs angesprochen, auf Individualdaten angewiesen. Die Quellen, aus denen institutsübergreifende medizinische Forschung, und das gilt für epidemiologische Projekte nahezu in allen Fällen, Informationen über krankheitsrelevante Sachverhalte gewinnen muß, sind in der Regel Ärzte oder ärztliche geleitete Institutionen. Hier kommt mit der ärztlichen Schweigepflicht das Thema der faktischen Anonymisierung ins Spiel. Der Paragraph 203 StGB wird durch Standesrecht ausgefüllt. Die entsprechenden Bestimmungen in der Bundesärzteordnung und in den Landesärzteordnungen erlauben eine Weitergabe ärztlicher Daten zu Zwecken der Forschung nur mit Einwilligung des Patienten oder anonymisiert. Wo die Einwilligung des Kranken nicht erlangt werden kann, beispielsweise weil er nicht einwilligungsfähig oder bereits verstorben ist oder die Einholung einer Einwilligung nur unter unbilligem Aufwand an Zeit und Kosten zu erreichen wäre, konkretisiert sich die Problematik der Anonymisierung. In den vergangenen Jahren ist, nicht zuletzt im Kontext der von einzelnen Datenschutzbeauftragten willfährig geschürten öffentlichen Datenschutzpsychose, der Anspruch an Anonymisierung so hoch geschraubt worden, daß eine Zuordnung der übermittelten Informationen zu einem einzelnen Fall mit hoher Zuverlässigkeit verhindert wurde.

Die Folge davon ist, daß die epidemiologische Forschung in der Bundesrepublik empfindlich beeinträchtigt wurde und eine Reihe von gravierenden medizinischen Problemen, etwa Altersdemenz bzw. Alzheimersche Erkrankung, dort ununtersuchbar wurden, wo Einwilligung und Einwilligungsfähigkeit unabdingbare Voraussetzungen für die Gewinnung der notwendigen Daten sind. Nicht weniger schwerwiegend sind die Risiko- und Therapieforschung betroffen. Wenn über Behandlungserfolge und -risiken ausgesagt werden soll, dann müssen institutionsüberschreitend Informationen über den weiteren Verlauf und muß beispielsweise die Information über einen Todesfall zuverlässig und mit der Möglichkeit der Zuordnung zu dem betreffenden Kranken gewonnen werden. Das gleiche gilt für die analytische Risikoforschung, die die kausale Beziehung zwischen Risikofaktoren und später eintretenden Gesundheitsschäden erfassen will,

ein Thema, das in der modernen Ökologie und Umweltpathologie zentrale Bedeutung einnimmt, aber einer soliden wissenschaftlichen Untersuchung in weiten Bereichen entzogen worden ist.

Für die Untersuchung kleiner Risikopopulationen, etwa solcher, die besonderer Exposition ausgesetzt oder mit spezifischen Vulnerabilitäten belastet sind, ist die Registerforschung ein wichtiges Instrument. Auch hier ist die Bundesrepublik in einer extrem restriktiven Position. Derzeit ist die Forschung mit Krankheitsregistern nur auf der Grundlage einer spezialgesetzlichen Regelung zulässig, wobei der Gesetzgeber alle zu registrierenden Items im Gesetz festgelegt. Diese Lösung ist in zweierlei Hinsicht unsinnig. Einmal führt sie dazu, daß nur eine minimale Anzahl von Registern und diese ausschließlich für solche Krankheiten, die im öffentlichen Bewußtsein stark präsent sind - sprich Krebserkrankungen - legalisiert werden. Zum anderen ist die Festschreibung der zu registrierenden Informationen im Gesetz eine unglückliche Regelung, denn ein Register muß jederzeit für den wissenschaftlichen Fortschritt offen sein, und wissenschaftlicher Fortschritt ist schneller als der Fortschritt der Gesetzgebung. Schließlich soll nicht übersehen werden, daß Risiken, die nur sehr kleine Populationen betreffen, und das gilt auch für die Exposition gegenüber Umweltgiften oder Gefahrensituationen am Arbeitsplatz, in der Tat nur mit nicht anonymisierten Individualdaten untersucht werden können. Hier muß die Abwägung zwischen konfligierenden Grundrechtsgütern doch noch anders gesehen werden als in der politischen und sozialwissenschaftlichen Forschung, denn hier geht es um die Bedrohung der Gesundheit und mitunter um die Gefährdung des Lebens von Menschen.

Damit möchte ich zum Schluß kommen: Wir haben in den letzten Jahren wegen der unserer Meinung nach irrationalen rechtlichen Beurteilung von Fallregistern einen wichtigen Teil unserer psychiatrisch-epidemiologischen Forschung mit den Daten des nationalen Dänischen Fallregisters und mit Daten der Weltgesundheitsorganisation aus zehn verschiedenen Ländern durchführen müssen. Das ist eine Situation, die eigentlich eines zivilisierten, vernünftigen Staatswesens, das die gegenwärtigen und künftigen Interessen seiner Bürger an der Erhaltung und Wiederherstellung der Gesundheit ernstnimmt, nicht würdig. Meine ganze Hoffnung baut darauf, daß allmählich die Vernunft oder eine unvoreingenommene, besonnene Abwägung zwischen bescheidenen Risiken des Mißbrauchs ärztlicher Daten in der Forschung und den Folgen der massiven Einengung medizinischer Forschung für die betroffenen Kranken und für die medizinische Wissenschaft die Oberhand gewinnen. Ich sehe in den Ergebnissen des Projekts zur faktischen Anonymität insofern einen echten Schritt vorwärts, weil es aus

dem Konsens von Sozialwissenschaft, Statistik und Bevölkerungswissenschaft und Politik - wenn auch nicht der Mehrheit aller Politiker - geboren, einen bedeutsamen Schritt auf die Notwendigkeit empirisch-sozialwissenschaftlicher Forschung zum Nutzen von Daten der amtlichen Statistiken getan hat. Ich sehe vor allem in dem Konsens, der sich in der Beurteilung der Ergebnisse dieses Projekts herausgebildet hat, einen hoffnungsvollen Schritt zu einer vernünftigeren Behandlung des Problems Geheimnisschutz und Forschung und hoffe, daß sich dieser Schritt zu einem Trend entwickeln möge, der auch der medizinischen Forschung wieder diejenigen Möglichkeiten zurückgibt, die ohne substantielles Risiko der Verletzung von Geheimnisschutz den Wiedereinstieg in die Bearbeitung einiger großer, ungelöster Forschungsfragen erlaubt.

Dr. Nowak, Statistisches Bundesamt, Wiesbaden:
In den Mittelpunkt dieser Diskussion wurde die Frage gestellt, inwieweit die Ergebnisse des hier vorgestellten Projekts auf andere Bereiche übertragen werden können. Für die Bundesstatistik gilt es dabei, sowohl die rechtlichen als auch die inhaltlichen Aspekte dieser Frage zu sehen. Den rechtlichen Rahmen bildet Paragraph 16 des Bundesstatistikgesetzes. Es hat wenig Zweck, die statistischen Ämter des Bundes und der Länder wegen dieser rechtlichen Grenzen zu schelten. Wir müssen uns einfach daran halten. Innerhalb dieser gesetzlichen Rahmenbedingungen gilt es, die inhaltlichen Kriterien zu konkretisieren. Ich glaube, hier hat uns das Projekt geholfen, eine ganze Reihe weiterer Kriterien zu erkennen. Kriterien die ansetzen an dem Begriff der Unverhältnismäßigkeit und an den Fragen: was kostet es den Angreifer, wie leicht fällt ihm eine Zuordnung und was nutzt sie ihm. Damit ist auch deutlich geworden, daß man hier die Frage stellen kann, ob es alternative Wege gibt, wie er zu diesen Informationen kommen kann. Wir haben anhand der Zahlen des Projekts gesehen, wie wenig rational es ist, den Versuch zu machen, über eine Reidentifikation vier Ergebnisse zu bekommen und dafür 100.000 Mark zu zahlen, wenn das gleiche Ergebnis über einen noch so schlechten Privatdetektiv wahrscheinlich für einen Bruchteil dieser Summe erhältlich ist. Ob man die Ergebnisse des Projekts zu Fragen der Inkompatibilität der Datensätze kurzerhand auf andere Bereiche übertragen kann, wird zu prüfen sein. Ich nehme an, daß es weiter bei der Einzelfallprüfung bleiben wird, da man auch zukünftig die Entwicklungen im Bereich der Inkompatibilität untersuchen und im Auge behalten muß.

Es ist hier mehrfach gesagt worden, daß die Ergebnisse des Projektes sich nur auf personenbezogene Informationen beziehen, nicht jedoch auf den Bereich der Wirtschaftsstatistik. Für einen großen Teil der Wirtschaftsstatistiken dürfte nach meinen Überlegungen das hier gezeigte Verfahren der

faktischen Anonymisierung ausscheiden. Man muß andere Ansätze prüfen und wird dann sehen, wie man weiterkommt.

Prof. Dr. Heinz, Arbeitsgruppe strafrechtliche Rechtstatsachenforschung und empirische Kriminologie, Institut für Rechtstatsachenforschung, Universität Konstanz:
Die Ergebnisse des hier vorgestellten Projektes sind zweifelsohne beeindruckend. In vielen Teilbereichen der sozialwissenschaftlichen Forschung wird damit das Problem der faktischen Anonymisierung lösbar sein. Allerdings gibt es auch, wie soeben schon Herr Häfner kritisch angemerkt hat, wissenschaftliche Bereiche, für die die Ergebnisses dieses Projektes deshalb geringe oder überhaupt keine Relevanz haben, weil diese Bereiche auch weiterhin auf die Erhebung von personenbezogenen Einzeldaten angewiesen sind, die erst im Prozeß der statistischen Analyse aggregiert werden können. Zu diesen Teilbereichen zählt auch die Kriminologie, die entweder die amtliche Statistik selbst als Forschungsgebiet hat, indem sie z.B. die Reliabilität der Strafverfolgungsstatistik durch Abgleich mit Daten aus anderen Quellen zu bestimmen versucht, oder aber mittels der Daten der amtlichen Statistik Untersuchungen durchführt. In beiden Fällen ist kriminologische Forschung regelmäßig auf personenbezogene Einzelangaben angewiesen. Illustrieren will ich dies an zwei konkreten Beispielen aus der empirischen kriminologischen Forschung, wobei im ersten Fall die Daten der amtlichen Statistik Gegenstand der Auswertung sind, im zweiten dagegen Grundlage für die Ziehung einer repräsentativen Stichprobe.

Wie schon aus den amtlichen Rechtspflegestatistiken hervorgeht, ist Kriminalität, jedenfalls in ihren schwereren Erscheinungsformen, ein relativ seltenes Ereignis. Die Strafverfolgungsstatistik der Bundesrepublik Deutschland weist dementsprechend schon bei der Gesamtzahl der Verurteilten gelegentlich nur einen einzigen Verurteilten aus. In Vergangenheit und Gegenwart hat man immer wieder versucht, die Strafzumessungspraxis der Gerichte im zeitlichen Längsschnitt und im regionalen Querschnitt darzustellen, insbesondere die Frage von Gleichmäßigkeit oder Ungleichmäßigkeit zu klären. Früheren Versuchen, die sich auf die veröffentlichten Daten stützten, wurde zu Recht vorgehalten, keine vergleichbaren Gruppen gebildet zu haben. Voraussetzung hierfür ist die Kontrolle der strafzumessungsrelevanten Faktoren, insbesondere der personenbezogenen Informationen der Strafverfolgungsstatistik (der Verurteilung zugrundeliegende Tat, Alter, Geschlecht und Zahl der Vorstrafen der Verurteilten). Die elektronische Datenverarbeitung ermöglicht es der sozialwissenschaftlichen Forschung, die Rohdaten z.B. der Strafverfolgungsstatistik zur Bildung derartiger homogener Gruppen zu nutzen. Bei entsprechender Gruppenbildung ist es aber, da die Strafzumessungspraxis in kleinen regionalen

Einheiten, z.B. Landgerichtsbezirken, miteinander verglichen werden soll, von vornherein nicht auszuschließen, daß man auf Einzelfälle stößt. Dies ist selbst bei der Untersuchung von insgesamt häufiger vorkommenden Delikten erwartbar. Erst bei Bildung derartiger merkmalshomogener Gruppen können Art und Höhe der Strafe in den verschiedenen Regionen auf Unterschiede hin überprüft werden. Und erst nachdem diese homogenen Gruppen gebildet sind, können etwa auftretende Einzelfälle von der weiteren Auswertung ausgeschlossen werden. Würde stattdessen von vornherein auf die Übermittlung und Auswertung von Einzelangaben verzichtet, wäre die Bildung homogener Gruppen prinzipiell verhindert bzw. derart erschwert, daß nicht mehr entschieden werden könnte, ob festgestellte Unterschiede auf Eigenschaften der abhängigen (regionale Einheit) oder der unabhängigen Variablen (Strafzumessungsfaktoren) beruhen.

Das andere Beispiel der Verwendung der Daten der amtlichen Statistik als Grundlage für die Ziehung einer repräsentativen Stichprobe bildet eine Untersuchung, die meine Mitarbeiter und ich in den letzten Jahren durchgeführt haben. Hier ging es unter anderem darum, die Effekte unterschiedlicher Erledigungs- und Sanktionierungsstrategien im Jugendstrafrecht auf die Wiederauftretenswahrscheinlichkeit der betroffenen jugendlichen Straftäter zu ermitteln. Als Grundgesamtheit wurde von uns die Gesamtzahl aller im Jahr 1979 in Baden-Württemberg durch die Staatsanwaltschaften durch Verfahrenseinstellung oder durch Anklage erledigten Jugendstrafverfahren gewählt. Die Daten zu diesen Verfahren wurden durch eine Totalerhebung aus dem Rohdatensatz der Staatsanwaltschaftsstatistik für das Land Baden-Württemberg gewonnen, den das Statistische Landesamt Baden-Württemberg auf Magnetband übermittelt hatte. Nach Ziehung einer quotierten Stichprobe wurden anhand der im Rohdatensatz enthaltenen Aktenzeichen bei den Staatsanwaltschaften die Personalien der Beschuldigten ermittelt, gegen die sich die Verfahren gerichtet hatten. Für diese Personen wurden Auskünfte aus dem Zentral- bzw. Erziehungsregister eingeholt. Erst danach war es möglich, die Daten durch Aggregation zu anonymisieren.

Im Unterschied also zu dem hier vorgestellten Anonymisierungsprojekt, das die faktische Anonymisierung bereits bei Datenerhebung gewährleisten will, ist die Situation in der Kriminologie grundlegend anders. In der Kriminologie ist es regelmäßig so, daß es notwendig ist, entweder Einzeldaten in die Auswertung einzubeziehen oder Einzeldaten aus verschiedenen Datenquellen miteinander zu verknüpfen, etwa bei der Sanktions- und Wirkungsforschung. Dementsprechend stellt sich auch das Datenschutzproblem anders. Für kriminologische Forschung ist in einer ersten Untersuchungsphase der Zugang zu personenbezogenen Einzeldaten der amtlichen Statistik unverzichtbar. Dies gilt zum einen uneingeschränkt für Untersuchungen, bei denen diese Daten als Basis für Sekundärdatenanalysen benötigt werden.

Dies gilt aber auch, jedenfalls solange die Geschäftsstellenautomation noch nicht flächendeckend eingeführt ist, auch für die Stichprobenziehung für Zwecke von Aktenanalysen. Daraus resultiert die Forderung der Kriminologie nach entsprechenden gesetzlichen Forschungsklauseln und nach gesetzlichen Regelungen über die Übermittlung von Daten zu Zwecken der wissenschaftlichen Forschung bei strenger Zweckbindung, die gewährleistet, daß Daten nur im Rahmen des Forschungsvorhabens verwertet und zum frühestmöglichen Zeitpunkt anonymisiert werden.
Bei der automatisierten Datenverarbeitung selbst muß dann freilich der hohen Sensibilität dieser Daten durch entsprechende technische und organisatorischen Maßnahmen (vgl. Paragraph 9 BDSG) Rechnung getragen werden. Von besonderer Bedeutung sind vor allem die Anforderungen bezüglich der Speicher-, Zugriffs-, Übermittlungs- sowie der Eingabekontrolle. In Konstanz haben wir, um auch hier ein konkretes Beispiel zu erwähnen, dieses Problem dadurch gelöst, daß wir zum einen den Zugang nur über virtuelle Maschinen erlauben, auf die nur über dedizierte Terminals zugegriffen werden kann, und wir zum anderen ein maschinenseitiges Protokollierungsprogramm einsetzen, das jeden Datenzugriff mitprotokolliert. Um freilich den Stand der Technik in vollem Umfang zu erreichen, bedarf es professioneller Lösungen, die sich derzeit typischerweise an den Sicherheitsanforderungen des vom US-amerikanischen National Computer Security Center erarbeiteten Kriterienkatalogs zur Bewertung von Informationstechnik-Systemen ("Trusted Computer System Evaluation Criteria"), dem sogenannten Orange Book, orientieren. Dem wird z.B. bei der Neuanschaffung eines Rechners im Hochschulrechenzentrum der Universität Konstanz Rechnung zu tragen versucht. Die Anforderungen des Datenschutzrechts, die die Forschung zu beachten hat, können nur bei entsprechender Ausstattung der Hochschulrechenzentren erfüllt werden. Diese müssen entsprechend ausgestattet werden. Hierauf zu bestehten, ist Aufgabe auch der Forschung.

Dr. Vorschulte, Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, Düsseldorf:
Das statistische Bundesamt und die statistischen Ämter der Länder geben ja bereits heute faktisch anonymisierte Einzeldatensätze an die Wissenschaft weiter. Sie tun das in wenigen Fällen und sie tun das nicht abgestimmt und mit recht unterschiedlichen Verfahren. Diejenigen von Ihnen, die bereits solche Unterlagen bekommen haben wissen, wie lange Gespräche geführt worden sind, bis schließlich ein vereinbarter Datensatz vorlag. Wir geben diese Angaben weiter an Hochschulen und an sonstige Einrichtungen mit der Aufgabe unabhängiger wissenschaftlicher Forschung. Das hier nun vorgestellte Forschungsprojekt zeigt die Anonymisierung von Einzeldatensätzen im Mikrozensus und in der Einkommens- und Verbrauchsstichprobe. Es

hat dazu als Fazit der Untersuchung Empfehlungen gegeben und es gibt, an die Empfehlungen angelehnt, einen Leitfaden, der im Statistischen Bundesamt entwickelt und hier vorgetragen wurde. Diese Empfehlungen sind nach meiner Auffassung und Einschätzung auch auf andere Statistiken in modifizierter Form übertragbar. Und ich meine, nachdem diese Arbeit vorliegt und die Summe der Empfehlungen Ihnen bekannt ist, ist es jetzt an der Wissenschaft, mit Datenwünschen auf uns zuzukommen und wir müssen dann gemeinsam überlegen, wie diese Modifizierungen letztlich aussehen müssen, damit sie den Datenbedarf befriedigt bekommen, der für die Durchführung Ihrer Aufgaben erforderlich ist. Das wollen wir Ihnen liefern. Wir sind schließlich und endlich wissenschaftsfreundlich und nicht wissenschaftsfeindlich und wollen Ihnen die Informationen, die mit viel Mühe und Aufwand erarbeitet worden sind, zugänglich machen, aber wir müssen letztendlich auch darauf achten, daß die Grundforderung der Statistik Bestand behält, daß Einzelangaben nicht einer bestimmten Person zugeordnet werden können.

Prof. Dr. Brennecke, Freie Universität Berlin:
Als Sozialmediziner gehöre ich einem Fachgebiet an, welches sowohl personenbezogene Daten als auch anonymisierte Daten braucht. Ich denke, daß dieses Forschungsprojekt auch für personenbezogene Daten und für Untersuchungen damit ein sehr deutliches Ergebnis liefert. Wer sich an die alte Diskussion erinnert, wird wissen, daß es von Wissenschaftsrichtungen und z. T. auch von politisch oder ideologisch ausgerichteten Gruppen die Argumentation gab, daß Mediziner eigentlich keine personenbezogenen Daten bräuchten, sondern, da anonymisierte Daten auch zur Zuordnung und zur Verknüpfung von Datenbeständen benutzt werden könnten, man auch auf diesem Hintergrund medizinische Forschung betreiben könne. Dieses ist mit diesem Projekt ganz sicher widerlegt. Eine solche Zusammenführung gibt es nicht, sie ist nicht sicher. Registerforschung läßt sich nicht durchführen, indem man auf dieser Ebene arbeitet. Ich halte das für eine Unterstreichung, daß es für bestimmte abgegrenzte Forschungsvorhaben nach wie vor personenbezogene Daten geben muß. Ich denke, daß auch in bezug auf die anonymisierten Daten, die genauso benötigt werden, wesentliche Fortschritte erreicht worden sind. In bezug auf die Übertragbarkeit der Ergebnisse schweben mir zwei Bereiche vor, Einen, den ich sehr global sehe: die amtliche Statistik und alles was damit zusammenhängt. Da gibt es von wissenschaftlicher Seite natürlich sehr viele Anforderungen, die nicht genau so wie in diesem Forschungsprojekt, aber doch ähnlich sind. Ein Beispiel hierfür ist die Mikrozensus-Gesundheitserhebung. Obwohl die genaue Einordnung der Diagnosen sicherlich zweifelhaft ist, kann man doch, wenn man nach Krankheitsgruppen aggregiert, einen Einblick in die Morbidität bekommen, den sonst keine andere Statistik bietet. Auch hier

wäre es notwendig, anonymisierte Individualangaben zu bekommen. Und ich hoffe, daß bei einer Prüfung tatsächlich eine Übertragbarkeit dieser Ergebnisse beispielsweise auf die Mikrozensus-Gesundheitsstatistik möglich ist. Ein zweiter Bereich, und da haben mich die Ausführungen von Herrn Nowak bedenklich gestimmt, ist die Kostenstrukturstatistik, insbesondere die Kostenstrukturstatistik der Ärzte, die, wie ich eben gehört habe, wohl eher "Unternehmensstatistiken" zugerechnet werden. Dennoch hoffe ich, daß - da ja viele Praxisinhaber auch Personen sind - sich zumindestens Teile der Forschungsergebnisse zur faktischen Anonymisierung übertragen lassen. Ein Bereich, der mir ganz wichtig erscheint sind jedoch die prozeßproduzierten Daten. Ich habe den Eindruck, daß dieses Projekt erste Anstöße gegeben hat, in welche Richtung man überlegen muß, um die Anonymisierungsproblematik bei prozeßproduzierten Daten etwas besser in den Griff zu bekommen. Sicher ist heute noch die juristische Problematik ein Hindernis sowie der Umstand, daß mit dem Sozialdatengeheimnis ein grundsätzlich anderer juristischer Tatbestand geschaffen ist als derjenige im Bundesstatistikgesetz. Das Problem, wird sich aber lösen lassen. Wie geht man jedoch bei prozeßproduzierten Daten für die faktische Anonymisierung vor, bei denen ja eine Genauigkeit der Merkmale Voraussetzung für die Verarbeitung ist und damit, wenn ich das mal so bezeichnen darf, das "Geräusch" in den Merkmalsausprägungen, die Unsicherheiten, vermutlich geringer sein werden. Wie geht man dort mit der Anonymisierung vor und worauf hat man zu achten? Ich habe den Eindruck, daß wir solche Daten brauchen werden, auch im Rahmen der Gesundheitsberichterstattung als Querschnitt, vielleicht aber auch als Längsschnitt. Man kann hierfür aus dem Projekt viel Nutzen ziehen und weiterarbeiten. Dafür bin ich Ihnen sehr dankbar. Ich denke, als eine Folgerung ergibt sich, wie häufig bei Forschungsprojekten, ein noch viel größeres Arbeitsgebiet und ich bin zuversichtlich, daß mit dem entsprechenden Elan auch daran gegangen wird.

Prof. Dr. Zimmermann, Universität München:
(...) Nun darf ich vielleicht etwas sagen zu den Interessen meines Fachbereichs, der ja hier nicht unmittelbar beteiligt war. Es ist so, in der Bundesrepublik stehen die Wirtschaftswissenschaften oder auch die Statistik an den Universitäten ganz weit hinter den Sozialwissenschaften, was die empirische Forschung anbetrifft, das auch im Gegensatz dazu wie Ökonomen international forschen und lehren. International ist es üblich, Wissenschaft in einer Verbindung aus Theorie, empirischer Analyse und statistischen Tests zu betreiben. Hier haben die Deutschen - auch in meinem Fachbereich - einen Wettbewerbsnachteil, der begründet ist in der mangelnden Verfügbarkeit von Datenmaterial, das die amtliche Statistik zur Verfügung stellt, und auch in der Tradition der Wirtschaftswissenschaften,

die keine eigenen Daten erhebt. Man mag das bedauern, aber es ist so und insoweit sind wir wirklich in einer Notlage. Und insoweit wird das, was jetzt begonnen worden ist, uns weiterhelfen. Denn zumindest die, die in der Bundesrepublik Haushaltsökonomie betreiben, werden mit Freude hören, daß die Einkommensstichprobe und der Mikrozensus, die zentralen Instrumente, hier nun mittelfristig zur Verfügung stehen werden. Das wird die Forschung mit Sicherheit deutlich befruchten, insbesondere auch vor dem Hintergrund, daß die Mikroökonometrie, die Statistik der Mikrodaten in den Wirtschaftswissenschaften, eine phantastische Blüte in den letzten Jahren erfahren hat, daß wir die methodischen Verfahren haben, um diese Daten auch adäquat zu untersuchen. Die andere Seite ist die Unternehmensstatistik. Der Ökonom möchte natürlich beide Marktseiten, die Haushalte und die Unternehmen, zur Verfügung haben. Es wäre sicherlich aus der Sicht meiner Disziplin wünschenswert, den Weg weiter zu gehen, der begonnen worden ist und ich sehe eigentlich keine prinzipiellen Probleme, warum man hier nicht weitermachen sollte. Lassen sie mich zum Schluß noch einige kurze Kommentare zu den Empfehlungen geben. Überwiegend verstehe ich, daß es diese Auflagen geben muß. Die Stichprobenlösung ist etwa für den Bereich der Statistik und Ökonometrie ein Problem, das uns nicht besonders bedrückt. Die Frage, die schwieriger ist, ist etwa die der Nutzungsbegrenzung, die zu enge Einordnung der Daten, die a priori zur Verfügung gestellt werden. Ich will auch noch einmal darauf hinweisen, daß es Sinn macht, eine Standardisierung dieses Materials vorzunehmen. Aus mehrfacher Hinsicht ist es für beide Seiten nicht nur kostengünstiger und praktikabler, sondern es erleichtert auch den Vergleich der Ergebnisse und die Forschungsergebnisse müssen vergleichbar bleiben. Insofern würde ich auch aus inhaltlicher Sicht dafür plädieren, ein Paket zu machen. Einen Datensatz zu produzieren und den bei hinreichend guten Argumenten zur Verfügung zu stellen.

Prof. Dr. Allerbeck, Universität Frankfurt:
Ich meine, daß natürlich leicht vernachlässigt wird, daß es in den Jahren, die seit den Tagungen verstrichen sind, auf die Herr Kaase eingangs Bezug nahm, natürlich hier und da Fortschritte gegeben hat. Ich denke an eine sicherlich weithin unbekannte Regelung, wie die Ausführungsbestimmungen zum Wissenschaftsparagraphen des hessischen Datenschutzgesetzes, die festlegen, daß die Anonymisierung durch den Wissenschaftler oder die wissenschaftliche Institution vorgenommen werden kann, wenn die Stelle, die die Daten abgibt, selbst dazu nicht in der Lage ist. Also es gibt schon Dinge, die außerhalb unseres Kontextes hier geschehen sind, die Wege weisen, wie man das zur allgemeinen Befriedigung machen kann. Jedenfalls habe ich nicht gehört, daß es in Hessen über diese Bestimmung zu irgendwelchen großartigen Konflikten gekommen ist. Insofern gibt es auch

für diese Identifikationsprobleme hier und da Lösungen, die auch Rechtsnormen darstellen. Auch wenn der Bereich im großen und ganzen unbefriedigend - um es sehr zurückhaltend zu sagen - geregelt ist. Für unsere Diskussion sollten wir vielleicht so vorgehen, daß wir noch eine Runde am Podium haben und dann Fragen oder Stellungnahmen aus dem Rest des Saals willkommen sind.

Prof. Dr. Müller, Universität Mannheim:
Darf ich eine kurze Bemerkung machen. Wenn ich mir die in der Diskussion vertretenen Bereiche ansehe, fällt mir auf, daß wir einen Bereich vernachlässigt haben, nämlich die empirische Sozialforschung. Und ich bedaure, daß wir dazu niemanden eingeladen haben. Wir dachten, daß dieser Bereich ohnehin bekannt ist. Aber wir haben Herrn Mochmann hier, den Geschäftsführer des Zentralarchivs für empirische Sozialforschung, der vielleicht diese Lücke füllen könnte. Entschuldigen Sie Herr Mochmann, daß ich sie nicht vorher explizit auf das Podium gebeten habe.

Ekkehard Mochmann, Zentralarchiv für empirische Sozialforschung, Köln:
Wenn hier eine besondere Legitimation, die Sozialforschung zu vertreten, gefragt ist, so darf ich mich Ihnen auch als Sprecher des Vorstandes der Gesellschaft Sozialwissenschaftlicher Infrastruktureinrichtungen, kurz GESIS, vorstellen. Ich möchte aufgreifen, was Herr Zimmermann schon gesagt hat, nämlich die standardisierte Bereitstellung der Daten. Aus der praktischen Erfahrung des Zentralarchivs spricht im Sinne der intersubjektiven Überprüfbarkeit sehr viel dafür, einen einmal aufbereiteten Datensatz standardisiert und nicht mit individueller Variation verfügbar zu machen. Ganz aus der Praxis gesprochen, sind jetzt Gewichtungsprobleme bei einer in Amerika aufbereiteten Version der Eurobarometer aufgetreten. Wir haben die Absicht, die Nationengewichtung in den Datensatz einzufügen und nicht nur das vorhandene Europagewicht bereitzustellen. Das führt dazu, daß der Datensatz mit anderen Recodierungen bei einzelnen Merkmalen aufbereitet wird und, wenn man nicht aufpaßt, daß in der Literatur in Zukunft zwei Datenbasen für die gleiche Behauptung oder gegenläufige Behauptungen genommen werden. Niemand weiß dann, ist es ein Analyseartefakt oder ist es ein Artefakt, das aus den Daten selbst resultiert? Ich trete hier nachdrücklich dafür ein, daß man für Datensätze einen Standard für Datensatzkennzeichnungen entwickelt, z.B. eine Internationale Datensatz Nummer (ISDN), daß man also wie bei Publikationen weiß, auf welche Ausgabe man sich bezieht. Das zweite Argument in dieser Richtung ist genauso gefallen und ich kann das nochmals aus der praktischen Erfahrung unterstreichen. Wir haben uns im Zentralarchiv Zurückhaltung auferlegt, Spezialsubsets von bestimmten Daten zu erstellen. Für bestimmte Zwecke, z.B. die Lehre, kann es sinnvoll sein, so etwas zu machen, aber in der Regel

stellen die Forscher dann in der Analysephase fest, daß sie doch noch die eine oder andere Variable brauchen, und dann hat man zwei oder drei verschiedene Datenversionen. Zur Frage der Sozialforschung: Die positive Einschätzung des Projektes, die hier wiederholt gegeben worden ist, teile ich auch. Darüber hinaus haben wir zunehmend einen Bedarf an der Verknüpfung nicht nur von Mikrodaten der amtlichen Statistik, sondern insbesondere auch von Meinungs- und Einstellungsfragen mit den regionalen Kontexten. Das ist eine europaweite Entwicklung, die ich als ganz wichtig ansehe. Ich weiß noch nicht, welche Effekte dies nun hat, weil wir ja gerade die Regionalisierungsmerkmale, ganz abgesehen von den Gemeindekennziffern, beschneiden müssen, um die Identifizierbarkeit zu reduzieren. Andererseits erscheint mir aber eine größere Genauigkeit für diesen Bereich und die Verknüpfbarkeit mit Meinungs- und Einstellungsdaten bezogen auf unterschiedliche Regionen zunehmend wichtiger.

Prof. Dr. Müller, Universität Mannheim:
Eine Frage an Herrn Nowak im Hinblick auf den Punkt, den auch Herr Brennecke schon angesprochen hat. Zwar sind nicht alle Punkte, die das Anonymisierungsprojekt untersucht hat, auf andere Datenkonstellationen übertragbar, aber kann man nicht dennoch etwas gewinnen für andere Typen von Daten, die von der amtlichen Statistik aufbereitet werden, insbesondere wenn es sich um sehr ähnliche Datenkonstellationen handelt? Ich denke zum Beispiel an die Zeitbudgetstudie, die das Statistische Bundesamt mittlerweile durchgeführt hat oder an die Beschäftigtenstatistik der Bundesanstalt für Arbeit. Für Sozialdaten gibt es allerdings Spezialregelungen, die zu berücksichtigen sind. Es gibt Testerhebungen oder Sondererhebungen, die das Statistische Bundesamt weiterhin auf der Basis von Paragraph 16 Bundesstatistikgesetz durchführen kann. Auch Daten solcher Erhebungen sind für Sozialwissenschaftler von Interesse. In der Bildungsstatistik gibt es im Grunde eine ganz ähnliche Datenkonstellation: Es gibt Individualdaten einer Reihe von Merkmalen von bestimmten Personen, die man in einer ähnlichen Weise anonymisieren könnte.

Dr. Nowak, Statistisches Bundesamt, Wiesbaden:
Da Sie mich direkt ansprechen, will ich auch direkt antworten. Meine Aussage bezog sich auf den Bereich der Wirtschaftsstatistiken. Herr Vorschulte hat zu recht gesagt, daß man auch hier prüfen müssen wird, ob sich die für die personenbezogenen Statistiken vorliegenden Untersuchungen - Mikrozensus und EVS - im Grundansatz übertragen lassen. Für die Bereiche, die Sie genannt haben und die ich im weitesten Sinne als personenbezogene Statistiken ansehen würde, glaube ich, daß man diese Grundprüfung ähnlich durchsetzen könnte, daß heißt, man wird auch hier prüfen müssen, welche Überschneidungsmerkmale vorliegen, welches Risiko

einer Deanonymisierung sich daraus entwickelt, und ob der Gedanke der Dateninkompatibilität genauso zu sehen ist, wie in den Fällen, die untersucht worden sind, ohne daß man die sehr aufwendigen Szenariotechniken nochmals nachstellen muß. Damit stellt sich auch die Frage, die Herr Brennecke schon angesprochen hat: Sind prozeßproduzierte Daten im Hinblick auf Dateninkompatibilität genauso zu betrachten wie andere, die auf unterschiedlichen Erhebungswegen eingeholt worden sind oder habe ich eine geringere Inkompatibilität, wenn ich aus der gleichen Quelle schöpfe? Ich glaube, das sind Fragen, die man noch untersuchen muß. Wir müssen dabei allerdings den Rahmen, den der Gesetzgeber mit dem Paragraph 16 des Bundesstatistikgesetztes gezogen hat, mit allen Bedingungen sehen, da können wir als Statistiker nicht einfach darüber hinwegspringen.

Bertram Raum, Bundesbeauftragter für den Datenschutz, Bonn:
Ergänzend will ich etwas zu der Möglichkeit sagen, vielleicht einmal ein anonymisiertes File zu erstellen. Ich stehe diesem Wunsch sehr skeptisch gegenüber. Grundsätzlich ist es so, daß Daten an die Forschung wie auch an andere Stellen nur weitergegeben werden können, soweit sie für diesen speziellen Zweck erforderlich sind. Bei einem public use file würde sie jedoch auch Daten bekommen, die für das spezielle Forschungsvorhaben nicht erforderlich sind. Und deshalb müssen diese Stellen sagen, diese Daten brauche ich für dieses Forschungsvorhaben. Dann besteht die Möglichkeit, daß sie diese Daten auch bekommen. Und dann zu dem Problem, das von Herrn Häfner, Herrn Brennecke und auch Herrn Allerbeck angesprochen wurde. Dieses Forschungsprojekt hier bezog sich ja nur auf statistische Daten, die in Paragraph 16 Bundesstatistikgesetz, dem Statistikgeheimnis geregelt sind. Auch Herr Heinz hat Probleme in seiner Forschung, die er auf personenbezogene Daten zurückführt. Das sind aber Sonderprobleme. Sie brauchen da keine statistischen Daten, keine Daten die dem Statistikgeheimnis unterliegen. Für die Daten, die Sie brauchen wird ja schon lange an einer Wissenschaftsklausel in der Strafprozeßordnung gearbeitet. Jüngsten Gerüchten zufolge hat man die Arbeit darüber wieder aufgegriffen. Hier gelten zunächst einmal auch nicht die Wissenschaftsklauseln in den Datenschutzgesetzen, sowohl weder das Bundesdatenschutzgesetz (innerhalb ihrer Forschungsstelle: Paragraph 40 BDSG; hinsichtlich der öffentlichen Stellen, die Daten übermitteln Paragraph 14 BDGS), noch die Datenschutzgesetze der Länder. Beispielhaft ist ja schon das sehr fortschrittliche Datenschutzgesetz von Hessen genannt worden.

Prof. Dr. Heinz, Arbeitsgruppe strafrechtliche Rechtstatsachenforschung und empirische Kriminologie, Institut für Rechtstatsachenforschung, Universität Konstanz:

Auf die von Herrn Raum angesprochene Frage der bereichsspezifischen Regelung im Referentenentwurf eines Gesetzes zur Änderung und Ergänzung des Strafverfahrensrechts (StVÄG) vom 3. Nov. 1988, durch das die Einsicht in Strafverfahrensakten für wissenschaftliche Zwecke die notwendige gesetzliche Grundlage erhalten soll, will ich nicht näher eingehen. Nur zur Erläuterung des von Herrn Raum angesprochenen Problems sei darauf hingewiesen, daß neben Befragung und Beobachtung die Dokumentenanalyse die dritte anerkannte Erhebungsmethode der empirischen Sozialforschung ist. Ein Großteil der kriminologischen Forschung beruht auf Aktenanalysen. Derzeit fehlt immer noch die gesetzliche Grundlage für Aktenauskunft und Akteneinsicht. Kriminologische Aktenanalysen sind, da die Akteneinsicht für wissenschaftliche Vorhaben acht Jahre nach dem Volkszählungsurteil des BVerfG vom 15.12.1983 schwerlich noch auf Nr. 185a RiStBV in Verbindung mit einem "Übergangsbonus" gestützt werden kann, derzeit nicht mehr möglich. Deshalb ist die kriminologische Forschung dringend auf die von Herrn Raum erwähnte Regelung im StVÄG angewiesen. Eingehen will ich vielmehr auf einen hier noch nicht angesprochenen Effekt des hier vorgestellten Anonymisierungsprojekts: Ich befürchte, daß dieses Projekt einen unerwünschten Nebeneffekt haben könnte. Diesen unerwünschten Nebeneffekt sehe ich darin, daß nunmehr angenommen wird, Paragraph 16 Absatz 6 BStatG sei eine Regelung, mit der die Forschungsbedürfnisse vieler Wissenschaftszweige hinreichend befriedigt seien, weil die Probleme der Verhältnismäßigkeit usw. jetzt geklärt seien, so daß die sozialwissenschaftliche Forschung mit faktisch anonymisierten Daten arbeiten könne. Dies ist jedoch nicht für alle sozialwissenschaftliche Disziplinen der Fall. Für die Kriminologie jedenfalls gilt, daß sie auf die personenbezogenen Daten der Strafverfolgungsstatistik angewiesen ist, wie ich bereits am Beispiel der Stichprobenziehung und der Forschungen zur Strafzumessungspraxis zu zeigen versucht habe. Alternativen zu den Daten der Strafverfolgungsstatistik gibt es derzeit - und auf absehbare Sicht - nicht. Die Eintragungen im Bundeszentralregister enthalten z.B. keine Freisprüche, auch die Mehrzahl der Einstellungen werden nicht eingetragen. Ob das geplante "länderübergreifende staatsanwaltschaftliche Informationssystem" die erforderlichen Angaben enthalten wird, steht noch nicht fest. Würde z.B. der Arbeitsentwurf "Strafverfolgungsstatistikgesetz" vom 30.5.1989 in Kraft treten, dann würde Paragraph 16 Abs. 6 BStatG auch für die Kriminologie gelten und damit das Ende für einen nicht unerheblichen Teil kriminologischer Forschung bedeuten, insbesondere für den rechtspolitisch besonders wichtigen Bereich der Sanktions- und Wirkungsforschung. Möglich wären dann vielleicht noch Untersuchungen zur Sanktionspraxis von Massendelikten, wie z.B. "Diebstahl geringwertiger Sachen" in ausgewählten, sehr großen Landgerichtsbezirken. Nicht mehr möglich wären dagegen Untersuchungen in den rechtspolitisch besonders

wichtigen Fallgruppen, etwa der Gewaltkriminalität, weil in diesen relativ seltenen Ereignissen die Einzelangaben, die strafzumessungsrelevant sind, nicht übermittelt werden dürfen. Paragraph 16 Abs. 6 BStatG bedarf deshalb der Ergänzung durch den Gesetzgeber, entweder durch eine weitere Wissenschaftsklausel, in der die Belange der von mir angesprochenen Wissenschaftszweige im Sinne einer "praktischen Konkordanz" von Wissenschaftsfreiheit und dem Recht auf informationelle Selbstbestimmung geregelt werden, oder durch eine bereichsspezifische Regelung in den speziellen Statistikgesetzen, also z.B. im Strafverfolgungsstatistikgesetz. Insoweit verweise ich z.B. auf Paragraph 40 Abs. 2 Bundeszentralregistergesetz. Nur durch eine derartige Regelung wird gewährleistet, daß auch künftig kriminologische Forschung die Daten der amtlichen Statistik für die Durchführung von Forschung, sei es als unmittelbarer Gegenstand, sei es für die Planung und Durchführung von Aktenanalysen (Stichprobenziehung) wird nutzen können. Für viele sozialwissenschaftliche Disziplinen wird der hier vorgestellte Weg der faktischen Anonymisierung gangbar und ausreichend sein. Ich plädiere aber nochmals dringend dafür, daß wissenschaftlichen Forschungsvorhaben, die auf personenbezogene Daten angewiesen sind, auch künftig diese Daten zur Verfügung gestellt werden. Aufgabe und selbstverständliche Verpflichtung der in diesen Bereichen tätigen Forscher ist es, einerseits durch die erforderlichen technisch-organisatorischen Maßnahmen und andererseits durch Anonymisierung der Daten zum frühestmöglichen Zeitpunkt sicherzustellen, daß der einzelne durch den Umgang mit seinen Daten in seinem Persönlichkeitsrecht nicht beeinträchtigt wird.

Prof. Dr. Brennecke, Freie Universität Berlin:
Vielen Dank, daß Sie das nochmal gesagt haben. Das ist auch in meinem Sinn. Ich möchte in bezug auf das, was Herr Mochmann und Sie über das public use file gesagt haben, noch etwas hinzufügen. Ich denke, der Begriff public use file ist sehr unglücklich. Ganz sicher wird es nicht so aussehen - und so stelle ich mir das auch nicht vor -, daß das Statistische Bundesamt ein wie auch immer geartetes File an X gibt und von X an Y und Z usw. Aber die Erfahrung der Auswertung von Mikrodaten hat doch eigentlich gezeigt, und das ist implizit heute auch angeklungen, daß man im Prinzip alle Möglichkeiten eines solchen Datensatzes nutzen möchte. Im Gegenteil, meistens ist die Situation so, daß man sich an vielen Stellen fragt, ob es nicht gut wäre, wenn dieses oder jenes Merkmal noch im Datensatz wäre. Ich vermute deshalb, daß die Anforderungen, die von verschiedenen Seiten bezüglich des Mikrozensus kommen werden, in bezug auf die Merkmale und Variablen sehr ähnlich sein werden. Und ich vermute deshalb, daß es sich von seiten des Statistischen Bundesamtes lohnen wird, diesen Gedanken vorzugreifen und für alle jene Fälle, die mit ähnlicher und begründeter

Einzelanforderung kommen, aber in der Struktur gleich sind, ein einheitliches File zu konstruieren. Das würde unter Umständen den Arbeitsaufwand minimieren und vielleicht auch den Forderungen oder dem Wunsch der Forscher relativ nahe kommen.

Prof. Dr. Dr. Häfner, Zentralinstitut für seelische Gesundheit, Mannheim:
Wenn wir Themen angeschnitten haben, die weit über den Paragraph 16 des Statistikgesetzes hinausgehen und, was mich angeht, nicht nur das Datenschutzgesetz, sondern auch den Paragraph 203 des Standesrechts zum Inhalt hatten, dann hat das damit zu tun, daß man am Ende einer solchen Tagung ein wenig das Recht fühlt, sich zu fragen, ob es über den begrenzten direkten Effekt für mein Fach auch eventuell einen indirekten Effekt gibt. Auch da ist natürlich das Problem - das versuchte ich deutlich zu machen -, daß die Bedürfnisse der Forschung auf meinem Gebiet so sind, daß ich nicht nur de lege lata sondern auch de lege ferenda denken oder diskutieren muß. Und ein zweiter Punkt, der noch ein bißchen weiter weg geht vom heutigen Thema, den ich aber auch mit angeschnitten habe, ist die Fragmentierung der Gesundheitsstatistiken in der Bundesrepublik. Wir haben mit diesem Problem solche negative Folgen für die Forschung, daß wir im Grunde genommen Public-Health Forschung und Epidemiologie nur in Grenzbereichen führen können. Es ist nun leider nicht so, daß uns die Regelungen, wie etwa Anonymisierung, die den Zugang zu Daten einzelner Statistiken erleichtern würden, viel helfen würden. Denn die unterschiedlichen Institutionen die Daten sammeln und zur Verfügung stellen können, haben außerdem noch unterschiedliche Variablendefinitionen, unterschiedliche Erhebungsmethoden, aber auch unterschiedliche Grundgesamtheiten. Wenn ich beispielsweise die Arbeitslosenstatistik vergleichen will mit der Mortalitätsstatistik des Statistischen Bundesamtes, um zu prüfen, ob Arbeitsplatzverlust oder längerfristige Arbeitslosigkeit Suizidraten erhöht, entsteht das Problem, daß ich unterschiedliche Grundgesamtheiten habe, die ich im Grunde nicht vergleichen kann. Das Problem bleibt also, daß man die Ausgangssitutation für Gesundheitsdaten in diesem Land verbessern muß.

Dr. Schmidt, Bundesbeauftragter für den Datenschutz, Bonn:
Ich möchte etwas sagen zu der Problematik, die aus meiner Sicht mit diesem Projekt kaum etwas zu tun hat. Bei den Verlaufsstatistiken, die wir natürlich vor allen Dingen da brauchen, wo es auf die Folgen von Maßnahmen ankommt, wie zum Beispiel bei Verurteilungen, ist sicher eine Anonymisierung, wie heute in diesem Projekt vorgestellt, überhaupt nicht denkbar. Die Richtung, die zur Zeit wohl sehr konsequent gedacht ist, ist die, daß man Verlaufsregister mit einem Anonymisierungstreuhänder schafft, das scheint mir ein aussichtsreicher Weg. Und ich glaube, das Hindernis liegt hier nicht

im Bereich der Datenschützer oder der Datenschutzbeauftragten. Ganz konkret: es gab zum Beispiel in der DDR Register, die unter diesen Umständen hier nie geführt worden wären, die aber gleichwohl für die Forschung auch jetzt noch von erheblicher Bedeutung sein können. Hier geht die Bereitschaft der Datenschutzbeauftragten, über die Wahrung der Interessen der Betroffenen entsprechende Forschung möglich zu machen, sicher sehr weit und ich würde auch mal sagen, weiter als die Bereitschaft mancher, die jetzt das Dach über dem Kopf und den Pförtner bezahlen müßten, damit die Daten auch aufbewahrt werden können. Also da sieht sich zumindest die Datenschutzseite von dem Vorwurf frei, hier ein Hindernis zu bieten. Es sind andere, die da bremsen. Zum public use file möchte ich sagen, dies ist eine Geschichte, die sollte man nach Möglichkeit nicht mit diesem Projekt verknüpfen. Man könnte die Ergebnisse dieses Projekts überstrapazieren, wenn man damit allgemein benutzbare - und sei es nur in der Forschung allgemein benutzbare - Standarddatensätze machen würde. Dann hätte man nämlich vermutlich ein ganz neues Angriffsszenario. Die Interessen, auch die publizistischen Interessen, eine definierte Datenbasis als verletzbar hinzustellen, könnten zu Zusammenarbeiten führen, die ein ganz anderes Szenario aufzeigt. Deshalb meine ich, daß man zumindest mit den jetzt aufgestellten Ergebnissen gut daran tut, weiterhin Forschungsvorhaben bezogenene Weitergaben zu erlauben, die vernünftigerweise aus einer konkreten, dafür geeigneten Basis mit einfachen Mitteln gezogen werden können. Zu denen man dann vielleicht auch leichter als das bisher möglich war, auch mal noch ein Datenfeld nachliefern kann. Aber zu sagen, jeder vernünftig scheinende Antrag bekommt die selben Daten, dafür scheinen mit die Ergebnisse dieses Projektes nicht ausreichend tragfähig. Wir tun deshalb gut daran, zunächst die Erfahrungen mit der individuellen Prüfung abzuwarten und zu sehen, wie das ausgeht.

Dr. Nowak, Statistisches Bundesamt, Wiesbaden:
Ich möchte mich der Meinung anschließen, daß das Wort "public use" für unsere Diskussion irreführend ist. Faktisch anonymisiertes Einzelmaterial ist nach den Bestimmungen des Gesetzgebers nur für einen sozusagen privilegierten Empfängerkreis und nicht für die Öffentlichkeit zugänglich. Aber es ist hier auch - und ich erinnere an die Aussage von Frau Marsh heute Mittag - gesagt worden, Kernpunkt der Akzeptanz ist das Vertrauen. Dem Auskunftgebenden für die Statistik reicht sozusagen das subjektive Mißtrauen, um der Statistik die Mitarbeit aufzukündigen. Das ist etwas sehr Sensibles, woran wir denken müssen. Was Herr Schmidt hier aufführt ist ein Angriffsszenario, das man unter diesem Blickwinkel sehr ernst nehmen muß. Erlauben Sie mir noch eine Anmerkung zu dem, was Herr Heinz gesagt hat. Wir sollten nicht nur gebannt auf den Paragraph 16 Absatz 6 des Bundesstatistikgesetzes sehen. Der Paragraph 16 sagt ja ausdrücklich, daß

der Gesetzgeber im begründeten Fall auch andere Formen der Weiterleitung von Einzelangaben im jeweiligen Gesetz zulassen kann. Er muß es dort aber regeln. Und daher müssen diejenigen, die solche Angaben benötigen an den Gesetzgeber appellieren und nicht nachträglich den Statistikern sagen, nun gebt mir endlich die Daten, auch wenn es im Gesetz so nicht erlaubt ist. Dasselbe gilt auch für die Frage, darf denn ein Wissenschaftler Daten an einen anderen weitergeben. Nachdem wie ich den Paragraph 16 des Bundesstatistikgesetztes lese, steht im Abschnitt 8 genau das Gegenteil. Da steht, daß er es nicht darf. Und letztlich sollten wir auch deshalb hier nicht so gebannt nur auf den Paragraph 16 Absatz 6 des Bundesstatistikgesetztes sehen, weil die Leistung der amtlichen Statistik nicht nur die Übermittlung von Einzelangaben ist - anonymisiert oder wie auch immer - sondern die Übermittlung von statistischen Ergebnissen in Aggregatform. Wir sollten nicht vergessen, daß es dafür auch ganz gute Zwecke gibt und man damit auch ganz gut arbeiten kann. Ein allerletztes Wort zu dem, was Herr Häfner sagte: Eine der Grundlagen für eine fachliche Konzentration der Arbeit der amtlichen Statistik ist die Koordinierungsfunktion, damit Sie eben nicht immer wieder vor unterschiedlichen Statistiken stehen, deren Ergebnisse vielleicht als Insellösung isoliert ganz brauchbar sind, die aber nicht mit anderen kombiniert verwendet werden können.

Prof. Dr. Allerbeck, Universität Frankfurt:
Da das Stichwort public use drei- oder viermal gefallen ist, können sie vielleicht verstehen, daß ich mich selbst auf die Rednerliste setze. Zunächst ist einmal klar, daß unter dem Absatz des Paragraphen, mit dem wir uns die ganze Zeit beschäftigt haben, ein public use file nicht möglich ist. Das ist vollständig unstrittig. Der Aspekt, den Herr Mochmann angesprochen hat ist, daß die Leute, die Daten haben oder mit Daten arbeiten, diese ja fortlaufend immer ein klein wenig modifizieren, weil sie etwas merken, was nicht in Ordnung ist, was besser sein könnte. Die Befürchtung, die da angesprochen worden ist, und die ich teile ist, daß das Statistische Bundesamt, so lange die Daten im Hause sind an den Daten schneidert und sicherlich die eine oder andere Korrektur vornehmen wird. Dieses maßgeschneiderte Verfahren der Variablenauswahl wäre für mich als Wissenschaftler nur dann akzeptabel - und jetzt ist das Stichwort nicht Publizistik sondern einfach Wissenschaft - wenn lediglich Variablen eliminiert würden und sonst nichts mehr mit dem Datensatz geschieht sobald er für fertig erklärt ist. Es darf dann nichts mehr geändert werden. Sonst ist dieses Verfahren der Maßschneiderung unbrauchbar und für die Wissenschaft einfach nicht akzeptabel. Was nun die Szenarien angeht, so wäre manche Diskussion vor der letzten Volkszählung bei uns sehr viel einfacher gewesen, wenn nicht die Informatikdiplomanden sondern unsere Studenten Zugang zu einem public use file gehabt hätten oder einem subset

eines public use file. Weil sie dann nämlich gewußt hätten, was in dieser Volkszählung typischerweise vorkommt. Und das hätte der Statistik nicht geschadet, sondern definitiv geholfen. Wenn man in Ihrem Amt ist, Herr Schmidt, dann bekommt man natürlich immer Alarmmeldungen aus aller Welt. Und hat von diesen einen reichen Vorrat und stellt sich dann vor, was könnte noch dazu kommen. Man könnte sich natürlich auch Strategien der amtlichen Statistik vorstellen, ihr Veröffentlichungsprogramm über das Erscheinen des statistischen Jahrbuchs zu verbreitern. Und eine Form dieser Verbreiterung wäre die Publikation von Daten in einer zeitgenössischen Form.

Dr. Vorschulte, Landesamt für Datenverarbeitung und Statistik Nordrhein-Westfalen, Düsseldorf:
Dazu kann ich gleich sagen, daß wir das jedes Jahr versuchen. Wir haben schon im letzten Jahr eine CD-Rom mit den Ergebnissen der Volkszählung auf kleinster räumlicher Ebene erstellt. Natürlich nur für Nordrhein-Westfalen. Und wir werden in diesem Jahr eine Diskette vorlegen mit Ergebnissen nach dem Bundesraumordnungsprogramm. Für die Herrichtung wird sicherlich vieles getan und da bemühen wir uns sehr, und im übrigen bleibt ja noch festzustellen, daß jeder Forscher die Möglichkeit hat, sich an das jeweilige Statistische Landesamt oder das Statistische Bundesamt zu wenden. Dort sind alle Datensätze vorhanden und bestimmte Probleme kann man dort abarbeiten lassen. Das hat ja Herr Mochmann heute Morgen als Normalfall für England, Dänemark oder Schweden vorgestellt, daß zentral gerechnet wird. Dort werden die Dinge nicht aus dem Haus gegeben, sondern die Probleme werden besprochen und dann wird zentral gerechnet. Das ist natürlich ein Angebot, das die amtliche Statistik hier in der Bundesrepublik auch immer gemacht hat. Wir jedenfalls in Nordrhein-Westfalen wären dazu jederzeit bereit.

Prof. Dr. Allerbeck, Universität Frankfurt:
Das Angebot ist natürlich in der Praxis die Aufforderung, für das Statistische Bundesamt einen Blankoscheck auszustellen. Der heißt natürlich nicht so, sondern der heißt Kostenübernahmeerklärung. Wir sind weit fortgeschritten in der Zeit, vielleicht sollte Herr Müller noch die Gelegenheit bekommen, ein abschließendes Wort in dieser Diskussion anzufügen.

Prof. Dr. Müller, Universität Mannheim:
Ja, gerne. Als erstes möchte ich sagen, daß ich diese Diskussion nicht angeregt habe, damit wir soviel Lob bekommen - worüber ich mich natürlich gefreut habe -, es hätte ja auch anders ausgehen können. Sondern es ging mir darum, Vertreter aus den verschiedenen Bereichen zu Wort kommen zu lassen, genau deshalb, weil die Wissenschaft ein sehr differenziertes

Unternehmen ist und weil klar ist, daß nicht zu allen Zwecken die gleichen Mittel taugen. Wenn man faktische Anonymität diskutiert und sieht, daß es damit für bestimmte Bereiche nun einen Schritt weitergeht, muß man gleichzeitig sagen, daß dieses nicht alle anderen Probleme löst, die wir haben. Deshalb waren die Ergänzungen, caveats und die Hinweise darauf, wie zu den verschiedenen Bereichen Problemlösungen aussehen müßten, ganz in meinem Sinn. Das heißt auch, man möge dieses Projekt und seine Ergebnisse nicht überfrachten und sie nicht auch für Bereiche nutzen wollen, für die sie nicht primär gedacht waren. Deshalb ist wichtig, was zum public use file gesagt worden ist. Man muß da sehr vorsichtig sein. Man soll die Ergebnisse des Anonymisierungsprojektes zunächst nur dazu nehmen, wozu das Projekt gedacht worden ist. Zum Abschluß möchte ich noch sagen, daß ich den heutigen Tag so wahrgenommen habe, wie das ganze Projekt. Ich fand die Zusammenarbeit zwischen den verschiedenen Stellen - zwischen Statistik, Wissenschaft und Datenschutz - überaus fruchtbar. Und wenn wir heute eine vergleichsweise unverkrampfte Diskussion gehabt haben und eine Diskussion, aus der man auch etwas lernen kann, war das zu Beginn des Projektes nicht so. Ich weiß noch, wie wir uns zunächst sehr vorsichtig begegnet sind. Hinter jedem Argument haben wir gewissermaßen eine verdeckte Attacke vermutet. Das hat sich im Laufe der Zeit jedoch in großem Maße verändert. Es liegt mir sehr daran, wenn dieses Projekt jetzt gelobt worden ist, dies wirklich an alle Beteiligten weiterzugeben: Neben den beteiligten Wissenschaftlern an den Datenschutz und die amtliche Statistik, die in vielfältiger Weise mitgewirkt haben. Sie haben nicht nur die Daten zur Verfügung gestellt, sondern auch unkonventiolle Lösungen ermöglicht. Ich habe zum Beispiel noch die lange Diskussion in Erinnerung, bei der wir überlegt haben, wie denn die faktische Anonymität mit Daten geprüft werden kann, die eigentlich nur als faktisch anonym weitergegeben werden dürfen. Daß hier Lösungen gefunden worden sind, ist das Verdienst aller, die mitgewirkt haben. Ich kann heute Abend nur einen großen Dank aussprechen und dabei auch alle einschließen, die sich heute als Referenten und Teilnehmer am Podiumsgespräch zur Verfügung gestellt haben und die dazu beigetragen haben, daß wir einen sehr interessanten Tag verbringen konnten.

Multivariate Analysen mit zufallsüberlagerten Tabellen aus dem Statistischen Informationssystem des Bundes (STATIS-BUND)

Von Georg Heer und Bernhard Schimpl-Neimanns

Das Statistische Informationssystem des Bundes (STATIS-BUND) bietet Nutzern außerhalb der amtlichen Statistik unter bestimmten Voraussetzungen die Möglichkeit, per Online-Anschluß amtliche Mikrodaten nach eigenen Wünschen auszuwerten. Die Nutzer erhalten Fallzahltabellen, die aus Geheimhaltungsgründen mit Zufallsvariablen überlagert sind. Ein Vergleich der Analysen von überlagerten Tabellen mit Analysen der Originaltabellen am Beispiel von Mikrozensus-Daten zeigt keine wesentlichen Verzerrungen in den Ergebnissen. Lediglich sehr schwach besetzte Tabellenfelder verursachen Unterschiede in Teilergebnissen. Des weiteren werden Möglichkeiten diskutiert, den Überlagerungsfehler bei multivariaten Analysen zu berücksichtigen. Die Untersuchungen über die Auswirkungen dieser Überlagerung auf die Ergebnisse multivariater Analysen wurden in Zusammenarbeit zwischen dem Statistischen Bundesamt und ZUMA durchgeführt. Georg Heer ist Referent in der Gruppe *Statistisches Informationssystem des Bundes* im Statistischen Bundesamt.

1. Einleitung

Mit STATIS-BUND bietet das Statistische Bundesamt auch Nutzern außerhalb der amtlichen Statistik Zugang zu Ergebnissen aus der Bundesstatistik. Das EDV-gestützte Informationssystem enthält eine Vielzahl von Zeitreihen und Strukturdaten in Tabellenform sowie komplexe Auswertungs- und Analysemöglichkeiten.

Daneben besteht unter gewissen Voraussetzungen die Möglichkeit, Auswertungen aus Mikrodaten nach eigenen Vorgaben zu erhalten, die der gesetzlich vorgeschriebenen statistischen Geheimhaltungspflicht genügen. Dies erfolgt, indem im Online-Verfahren anonymisierte Fallzahltabellen aus dem Einzelmaterial erstellt werden. Dabei spezifizieren die Nutzer die Tabellen nach eigenen Anforderungen und sind somit nicht an die amtlichen Klassifikationen gebunden. Bei dieser Form des indirekten Zugriffs auf amtliche Mikrodaten wird kein Einzelmaterial weitergegeben. Der vorliegende Aufsatz betrifft damit nicht die Weitergabe eines sogenannten "faktisch anonymisierten Mikrodatenfiles" an die Wissenschaft nach Paragraph 16 Abs. 6 BStatG (vgl. hierzu den Aufsatz von Heike Wirth in diesem Heft).

Auch bei der Weitergabe von Tabellen dürfen keine Einzelangaben offenbar werden. Hier wurden bisher hauptsächlich manuelle Anonymisierungsverfahren angewendet (Streichung von Zellenbesetzungen kleiner als drei,

Aggregation, Klassenbildung etc.), die jedoch aufwendig, fehleranfällig und mit einem Informationsverlust verbunden sind. Diese Nachteile sollen in STATIS-BUND durch ein Verfahren der automatischen Überlagerung von Tabellen mit ganzen Zufallszahlen so weit als möglich vermieden werden.

Bei der Abschätzung des durch die Überlagerung entstehenden Fehlers wurde bereits festgestellt, daß dieser bei der freien Hochrechnung einzelner Tabellenfelder nicht wesentlich über dem beim Mikrozensus üblichen Stichprobenfehler liegt (Kühn/Pfrommer/Schrey 1984). Auf multivariate Analysen ist dieser Befund jedoch nicht ohne weiteres übertragbar, da die Struktur der Tabelle in die bisherigen Fehlerrechnungen nicht einging. Um die Auswirkungen der Überlagerung in STATIS-BUND auf multivariate Analysen zu untersuchen, vergleichen wir in diesem Aufsatz die Ergebnisse von Logit-Modellen für Originaltabellen mit den Ergebnissen für überlagerte Tabellen. Um eine möglichst praxisrelevante Auswertungssituation zu haben, wurden Fragestellungen der schichtspezifischen Bildungsungleichheit untersucht. Methodisch konzentrieren wir uns dabei auf eine spärlich besetzte Tabelle, bei der zu vermuten ist, daß sich die Überlagerung besonders kritisch auswirkt. Darüber hinaus geben wir Hinweise, wie die Auswirkungen des Überlagerungsfehlers möglichst klein gehalten werden können.[1)]

2. Das Statistische Informationssystem des Bundes

2.1 Leistungsumfang

STATIS-BUND enthält statistische Ergebnisse in Form von Zeitreihen und Tabellen mit Strukturdaten aus allen Bereichen der amtlichen Statistik. Das Spektrum des Angebots in den circa 900000 Zeitreihen und in den Strukturtabellen reicht von Daten zu Bevölkerung, Erwerbstätigkeit und Wahlen über Betriebs- und Unternehmensstatistiken bis hin zu Außenhandelszahlen und Volkswirtschaftlichen Gesamtrechnungen. Einen Überblick über das aktuelle Angebot gibt das jährlich vom Statistischen Bundesamt herausgegebene *Datenbestandsverzeichnis*. Die Statistiken sind innerhalb des Systems fachlich und technisch ausführlich beschrieben. Die Dokumentation enthält alle Informationen, die bei der Auswahl und Verwendung der Daten sowie bei der Beurteilung ihrer Qualität erforderlich sind.

Neben dem Angebot an allgemein zugänglichen aggregierten Daten besteht unter bestimmten Voraussetzungen die Möglichkeit, amtliche Mikrodaten für die Weiterverarbeitung in einer anonymisierten Form zu nutzen. Diese Möglichkeit ist unter Punkt 2.2 näher beschrieben.

Ein Nutzer kann Online auf Daten, Analyse- und Auswertungsverfahren in STATIS-BUND zugreifen, sofern er über einen Datex-P-Hauptanschluß verfügt und einen entsprechenden Vertrag mit dem Statistischen Bundesamt abgeschlossen hat. Diese Nutzungsart ist vor allem dann sinnvoll, wenn die vielfältigen Möglichkeiten des Systems intensiv genutzt werden. Eine komfortable Benutzersprache erlaubt dem Anwender die Weiterverarbeitung von Systemdaten sowie eigener Daten mit den folgenden Mitteln:

Auswertungssystem: Verfahren zur Tabellenerstellung einschließlich hierarchischer Auswertung, Mischen, Sortieren und Bearbeiten von Materialien mittels Rechenoperationen sowie Druckaufbereitung.

Analysesystem: Mathematisch-Statistische Analysemethoden, z.B. Lösen von Gleichungssystemen, Regression, Interpolation, Prognose, Zeitreihenanalyse, Bevölkerungsanalyse und -prognose, Varianzanalyse, Faktorenanalyse, Lag-Untersuchungen.

Grafiken: grafische Darstellung von Ergebnissen, Kurven-, Balken- Kreis- oder Tortendiagramme, Deutschlandkarte, Europakarte, Ausgabe in vordefinierter oder in beliebiger Form.

Neben dem Online-Anschluß gibt es noch eine Reihe weiterer Möglichkeiten, Daten aus STATIS-BUND zu erhalten. Für den Bezug kleinerer Datenmengen oder die gezielte Auswahl spezieller Zeitreihen bietet sich der *Diskettenservice* an. Die Disketten können einmalig oder im Rahmen eines Jahresabonnements monatlich, viertel- bzw. halbjährlich bezogen werden. Umfangreichere Datenbestände können per *Magnetband* geliefert werden.[2)]

2.2 Erstellung zufallsüberlagerter Fallzahltabellen aus Einzeldaten

Wie eingangs erwähnt, besteht darüber hinaus die Möglichkeit, Auswertungen aus Mikrodaten zu erhalten. Der Benutzer beschreibt die gewünschte Tabelle in der Benutzersprache des Systems und erteilt durch ein Kommando den Auftrag zur Tabellenerstellung. Aufgrund dieses Auftrags wird die Berechtigung des Benutzers automatisch geprüft und das Originalmaterial zu einer Tabelle ausgezählt, auf die der Benutzer jedoch keinen Zugriff hat. Diese *Originaltabelle* verbleibt im Statistischen Bundesamt für Kontrollzwecke. Aus ihr wird in einem zweiten Schritt eine weitere Tabelle erzeugt, indem zu jeder Fallzahl der Originaltabelle eine ganzzahlige

Zufallszahl addiert wird. Die verwendeten Zufallszahlen sind über mehrere verschiedene Tabellen hinweg annähernd normalverteilt mit Mittelwert Null und Varianz Drei.[3] Die mit Zufallszahlen *überlagerte Tabelle* wird dem Benutzer zur Verfügung gestellt, der damit eine Tabelle nach seinen Vorgaben erhält, die sich von der Originaltabelle in jedem Feld zufallsabhängig geringfügig unterscheidet. Hat die Tabelle viele schwach besetzte Felder, so wirkt sich die Überlagerung naturgemäß relativ stark aus. Der Benutzer kann dann durch gezielte Änderungen der Spezifikationen, zum Beispiel durch das Zusammenfassen von Gliederungspositionen, und einen erneuten Auftrag versuchen, eine aussagefähigere Tabelle zu erhalten.

Das Verfahren hat den Vorteil, daß es für den Anwender keinerlei Einschränkungen hinsichtlich Anzahl und Spezifikation der zu erstellenden Tabellen gibt. Dafür müssen Genauigkeitsverluste in den Einzelfeldern hingenommen werden. Die Überlagerung ist so konstruiert, daß durch Differenzbildung die wahre Größe eines geheimzuhaltenden Wertes nicht ermittelt werden kann. Für eine konkrete Tabelle sind demzufolge die Zufallszahlen nicht normalverteilt mit Mittelwert Null und Varianz Drei. Diese Aussage gilt nur näherungsweise für viele verschieden aufgebaute Tabellen.

Die Überlagerung hat zur Folge, daß in einer Tabelle die Summe von überlagerten Werten in der Regel nicht mit der überlagerten Summe übereinstimmt. Zur Bestimmung einer möglichst exakten Summe sollte diese mit ausgezählt werden und nicht nachträglich aus überlagerten Werten errechnet werden, da sonst die Varianz des Zufallsfehlers steigt (vgl. dazu Abschnitt 3.1).

Die Güte der überlagerten Einzelwerte wurde bereits durch eine Fehlerbetrachtung anhand von Auswertungen des Mikrozensus untersucht (Kühn/Pfrommer/Schrey 1984). Dazu wurde der einfache absolute und der relative Standardfehler für verschiedene Fallzahlen in einer Mikrozensustabelle mit und ohne Überlagerung geschätzt.[4] Für kleine Fallzahlen ist der Unterschied im relativen Standardfehler noch relativ groß (210 Prozent mit Überlagerung gegenüber 160 Prozent ohne Überlagerung bei einer Besetzungszahl von eins, hochgerechnet 100), er sinkt jedoch schnell. Bereits bei einer Besetzungszahl von zehn beträgt der relative Standardfehler 54 Prozent gegenüber 51 Prozent in der Originaltabelle. Für eine Fallzahl von 50 liegen die Werte mit 22,8 Prozent bzw. 22,6 Prozent praktisch zusammen. Bei der Überlagerung entsteht also nur bei denjenigen Fallzahlen ein bedeutender zusätzlicher Fehler, die schon wegen ihres großen Stichprobenfehlers nicht gesichert sind. Diese Überlegungen gelten jedoch nur für ein einzelnes Tabellenfeld. Wie sich die Überlagerung mit Zufallszahlen auf die

Gesamtstruktur einer Tabelle auswirkt, wird in den Abschnitten drei und vier untersucht.

3. Einführung in das Problem

In diesem eher didaktisch orientierten Abschnitt werden die Verzerrungen einfacher statistischer Kennzahlen von Tabellen, deren Zellenbesetzungen mit Zufallsvariablen überlagert sind, aufgezeigt. Die dabei verwendeten synthetischen Daten sind in der Tabelle 3.1 beschrieben.

Tabelle 3.1: Anzahl der Schüler nach Schulart und sozialer Herkunft (fiktive Werte)

Schulart	Mittel-schicht	Unter-schicht	Zusammen
weiterführend	30	25	55
Hauptschule	20	40	60
Zusammen	50	65	115

Zu jedem Tabellenfeld dieser Ausgangstabelle wird, abweichend von der Überlagerung in STATIS-BUND, eine normalverteilte Zufallszahl mit Mittelwert Null und Varianz Eins addiert. Die Zufallszahlen sind stochastisch unabhängig.

3.1 Randsummen und Anteilswerte

Für den Mittelwert und die Varianz einer Randsumme, die aus den Besetzungszahlen von K Zellen ($X_1, \ldots, X_K$) gebildet werden, gilt:

(3.1) $E(\Sigma X_k) = \Sigma E(X_k)$

(3.2) $Var(\Sigma X_k) = \Sigma var(X_k) = K\sigma^2$

da die Zufallsfehler voneinander unabhängig sind. Zählt man jedoch die Randverteilungen zusätzlich zu den Zellen aus, weisen diese nur jeweils eine Varianz von σ^2 auf, sind also weniger fehlerbehaftet als die durch Summierung der Zellen gebildeten Randsummen.

Um die Verzerrung eines Anteilswertes festzustellen, gehen wir von der tatsächlichen Zellenbesetzung $\mu(x)$ und der tatsächlichen Randsumme $\mu(n)$ aus. Die überlagerte Zellenbesetzung X besitzt den Mittelwert $\mu(x)$ und die Varianz $\sigma^2(n)$. Die Randsumme N mit Mittelwert $\mu(n)$ und Varianz $\sigma^2(n)$ kann entweder die einfach überlagerte Randsumme sein oder sich als Summe überlagerter Einzelzellen ergeben. Für den *erwarteten Anteilswert*, der besonders bei kleinen Zellenbesetzungen überschätzt wird, erhalten wir näherungsweise (vgl. Punkt 1 im Anhang):

$$(3.3) \quad E\left(\frac{X}{N}\right) \approx \frac{\mu_x}{\mu_n}\left(1 + \frac{\sigma_n^2}{\mu_n^2} - \frac{\sigma_{xn}^2}{\mu_x \mu_n}\right)$$

Da in der Näherungsformel nur die ersten beiden Momente berücksichtigt werden, wurden Monte-Carlo Simulationen zum Stichprobenumfang 1000 durchgeführt. Eine Verzerrung durch die Überlagerung konnte für die Beispieltabelle praktisch nicht festgestellt werden: Der Anteil an Schülern auf weiterführenden Schulen lag im Mittel bei 0,6 bzw. 0,3845 bei einer empirischen Varianz von 0,0141 bzw. 0,0113.

Die Varianz eines Anteilswertes $p=\mu_x/\mu_n$ oder eines Tabellenfeldes μ_x beträgt nach dem Binomialansatz

$$\sigma^2(p)=p(1-p)/\mu_n = \mu_x/\mu_n^2(1-\mu_x/\mu) \text{ bzw. } \sigma^2(\mu_x) = p(1-p)\,\mu_n = \mu_x - \mu_n^2/\mu_n .$$

Beide Maße sind infolge des Anonymisierungsverfahrens verzerrt. Die *Erwartungswerte der geschätzten Varianzen* lassen sich auch hier näherungsweise ermitteln:

$$(3.4) \quad E(\hat{\sigma}_p^2) \approx \frac{\mu_x}{\mu_n^2}\left(1 - \frac{\mu_x}{\mu_n} - \frac{\sigma_x^2}{\mu_x \mu_n} + \frac{3\sigma_n^2}{\mu_n^2} - \frac{6\mu_x\sigma_n^2}{\mu_n^3} - \frac{2\sigma_{xn}^2}{\mu_x \mu_n} + \frac{6\sigma_{xn}^2}{\mu_n^2}\right)$$

$$(3.5) \quad E(\hat{\sigma}_x^2) \approx \mu_x - \frac{\mu_x^2}{\mu_n}\left(1 + \frac{\sigma_n^2}{\mu_n^2} + \frac{\sigma_x^2}{\mu_x^2} - \frac{2\,\sigma_{xn}^2}{\mu_x\,\mu_n}\right)$$

Im Fall der obigen Beispieltabelle zeigt sich kein nennenswerter Unterschied zwischen den Binomialvarianzen der Original- und der überlagerten Tabelle. Der Effekt der Überlagerung wirkt sich jedoch bei sehr kleinen Zellenbesetzungen in einer Unterschätzung der Varianz und einer Überschätzung der Anteilswerte aus, wie die Simulationsergebnisse (siehe Tabelle 3.2) belegen.

Tabelle 3.2: Einfluß der Zufallsüberlagerung auf Anteilswert, geschätzte Standardabweichung und geschätzte relative Standardabweichung (Simulationsergebnisse bei tatsächlichem Anteil von 50%)

	Original - Tabelle			Überlagerte Tabelle		
Fall-zahl	Anteil an der Rand-summe in %	Standard-abweichung	relative Std.abw. in %	Anteil an der Rand-summe in %	Standard-abweichung	relative Std.abw. in %
5	50.00	1.5811	31.6228	50.41	1.5372	32.1843
10	50.00	2.2361	22.3607	50.16	2.2232	22.3995
25	50.00	3.5355	14.1421	50.04	3.5328	14.1406
50	50.00	5.0000	10.0000	49.99	4.9986	10.0056
100	50.00	7.0711	7.0711	49.99	7.0707	7.0725
250	50.00	11.1803	4.4721	50.00	11.1800	4.4721
500	50.00	15.8114	3.1623	50.00	15.8113	3.1622

3.2 Statistische Tests

Das Prozentsatzverhältnis (Odds-Ratio) der Beispieltabelle berechnet sich aus: (30*40)/(20*25). Es drückt die relativen Chancenverhältnisse zwischen den beiden Schichten aus, eine weiterführende Schule zu besuchen oder nicht. Da die Überlagerungen der einzelnen Zellen voneinander unabhängig sind, kann für die Berechnung des Erwartungswerts die folgende Näherung verwendet werden (vgl. Anhang):

$$(3.6) \quad E\left(\frac{X_1 X_4}{X_2 X_3}\right) \approx \frac{\mu_{x_1} \mu_{x_4}}{\mu_{x_2} \mu_{x_3}} \left(1 + \frac{\sigma_{x_2^2}}{\mu_{x_2^2}} + \frac{\sigma_{x_3^2}}{\mu_{x_3^2}}\right)$$

Man erhält mit (3.6) für das Odds-Ratio den Erwartungswert von 2,41, der nur geringfügig von dem Wert der Originaltabelle (OR=2,40) abweicht. Die Teststatistik für das Odds-Ratio ist: $t=\ln(OR)/\sigma$, mit: $\sigma=(\Sigma 1/n_{ij})^{1/2}$. Der t-Test für die Originaldaten ergibt t=2,27. In den Simulationen zeigte sich, daß in 5,7 Prozent aller überlagerten Tabellen der t-Test fälschlicherweise keinen signifikanten Zusammenhang ermittelte. Ähnlich reagierte auch der Chi-Quadrat-Unabhängigkeitstest auf die Überlagerung; in 5,2 Prozent der durchgeführten Simulationen wurde der "falsche" Schluß der statistischen Unabhängigkeit gezogen. Diese Ergebnisse hängen natürlich von den kleinen Zellenbesetzungen der Beispieltabelle ab und gelten nur für diese Art der Überlagerung.

Zusammenfassend läßt sich festhalten, daß die Abschätzung der Verzerrung statistischer Kenngrößen um so schwieriger wird, je komplexer diese Größen berechnet werden; dies gilt insbesondere für multivariate Analysen. Zudem sind ohne genaue Kenntnis der statistischen Verteilung des Überlagerungsfehlers in STATIS-BUND kaum exakte Abschätzungen möglich. Diese

Verteilung kann jedoch aus Geheimhaltungsgründen nicht veröffentlicht werden und ist ohnehin mathematisch wesentlich schwieriger zu handhaben als die Normalverteilung. Aus diesen Gründen wenden wir uns im nächsten Abschnitt einem Praxistest zu, bei dem nicht überlagerte und überlagerte Fallzahltabellen aus STATIS-BUND mit den gleichen statistischen Verfahren analysiert werden. Der Vergleich der Ergebnisse kann zeigen, wie sich die Überlagerung mit Zufallsfehlern für die Nutzer von STATIS-BUND auswirkt.

4. Vergleich von multivariaten Analysen mit Original- und überlagerten Fallzahltabellen

Für die Wahl der Testtabellen, mit denen der Vergleich durchgeführt wird, wurden zwei Kriterien zugrunde gelegt. Wie die obige Einführung zeigt, wiegt das Problem der Analyse zufallsüberlagerter Tabellen besonders schwer bei schwach besetzten Zellen. Unter methodischen Gesichtspunkten stellt eine tief gegliederte und spärlich besetzte Tabelle einen 'harten Test' für die Güte der Ergebnisse multivariater Analysen dar. Aus der Nutzerperspektive sollte es sich um eine praxisnahe Anwendung handeln. Einer nach inhaltlichen Kriterien erstellten Tabelle ist gegenüber einer Tabelle mit synthetischen Daten der Vorzug zu geben. Wir haben uns hier an einer Fragestellung aus der schichtspezifischen Bildungsforschung orientiert. Der Schwerpunkt dieser Arbeit ist jedoch methodisch-statistisch. Die inhaltlichen Kriterien für die Wahl der Testtabelle werden im folgenden kurz vorgestellt.

4.1 Bildungschancen in Abhängigkeit von Merkmalen der sozialen Herkunft

4.1.1 Inhaltliche Fragestellung und Beschreibung der Daten

Die Verteilung der Schüler nach sozialer Herkunft auf weiterführende Schulen ist auch nach den Reformversuchen der sechziger Jahre ungleich. Nach den klassischen Schichtindikatoren berufliche Stellung, Bildungsabschluß und Einkommen in der Familie gegliedert, zeigt sich ein Zusammenhang zwischen Schulbesuch/-erfolg und sozialer Schicht: je höher die berufliche Stellung, je höher die Bildungsqualifikation und je höher das Einkommen der Eltern ist, desto größer sind die Chancen, eine weiterführende Schule zu besuchen.

Mit der Variablen berufliche Stellung werden zumeist die Auswirkungen der Arbeits- und Berufserfahrungen der Eltern auf die Erziehungswerte und -handlungen operationalisiert. Es wird angenommen, daß Kinder in Mittelschichtfamilien größere Handlungsspielräume besitzen, die ihre kognitiven Fähigkeiten fördern. Der elterliche Bildungsabschluß ist ein

Indikator für die Vertrautheit der Familie mit den Ausbildungsanforderungen der Schule. Ein hoher Bildungsstatus der Eltern fördert die kindliche Intelligenzentwicklung. Die Einkommenssituation beeinflußt direkt (z.B. Verfügbarkeit von Büchern, Wohnsituation) und indirekt (z.B. das Angewiesensein auf ein frühes Einkommen der Kinder) die Bildungschancen. Seit die Länder keine Daten mehr zu Schulbesuchsquoten nach sozialer Herkunft erheben, ist der Mikrozensus die einzige Quelle der amtlichen Statistik, mit der diese Fragen untersucht werden können. Bildungsforscher kritisieren jedoch die eingeschränkten Analysemöglichkeiten, "... weil die Einteilung der Bevölkerung durch die amtliche Statistik in 4 oder 5 Berufsgruppen sehr undifferenziert und soziologisch nur sehr bedingt brauchbar ist" (Geißler 1987: 88). Daneben fehlt es an aussagekräftigen multivariaten Analysen zu den partiellen Effekten der einzelnen Schichtindikatoren auf die Bildungschancen. Man weiß zwar, daß der Effekt der beruflichen Stellung des Familienvorstandes stärker ist als der Einkommenseffekt (Böttcher 1991: 155, 157), aber gilt dies auch nach Kontrolle des Bildungsniveaus? Wir vermuten, daß kognitive Ressourcen des Elternhauses, vereinfacht gemessen mit dem Bildungsniveau des Familienvorstandes, die Bildungschancen am stärksten beeinflussen.

Wir versuchten die Testtabelle auch nach inhaltlichen Kriterien so zu gestalten, daß ansatzweise die genannten Kritikpunkte überwunden werden können, bzw. Wege aufgezeigt werden, die dem Datenbedarf der Bildungsforschung entgegenkommen. Für die Testtabelle wurden die im Mikrozensus 1987 erfaßten 13 bis 14jährigen Schüler in den Schularten des dreigliedrigen Schulsystems ausgewählt. Diese Altersgruppe eignet sich besonders dafür, die Wirkung der ersten Selektionshürde des deutschen Schulsystems zu untersuchen. [5] Die Testtabelle ist neben dem dichotomen Merkmal Schulart (Hauptschule/weiterführende Schule) nach den Merkmalen Bildungsabschluß des Familienvorstandes, Stellung im Beruf des Familienvorstandes und Einkommen des Familienvorstandes gegliedert. Spalte 1 in Tabelle 4.1 gibt die tatsächliche Anzahl der Schüler je Schulart an. Spalte 2 wurde durch Aggregation der überlagerten Testtabelle gewonnen, während Spalte 3 durch einfache Überlagerung von Spalte 1 entstanden ist.

Aus den Angaben zum allgemeinen und beruflichen Bildungsabschluß des Familienvorstandes wurde eine einfache Bildungsskala erstellt. In Tabelle 4.2 kann man einen fast linearen Zusammenhang zwischen dem Bildungsniveau des Familienvorstands und den Bildungschancen der Kinder erkennen (vgl. Spalte 2). Die insgesamt schwach besetzte Kategorie "Abitur" ist von der Überlagerung stark betroffen; der über die Summe überlagerter Zellen ermittelte Anteilswert (vgl. Spalte 3) liegt um 5,5 Prozentpunkte über dem tatsächlichen Anteilswert. Auf diese Kategorie entfallen bei 42 Zellen

Tabelle 4.1: Anzahl der Schüler nach Schulart, ermittelt als Randsumme der Original-Tabelle, Randsumme der überlagerten Tabelle und überlagerte Randsumme der Original-Tabelle

Schulart	Randsummen Original-Tabelle	Randsummen überlagerte Tabelle	überlagerte Original-Randsummen
(1) Hauptschule	5285	5328	5286
(2) Realschule, Gymnasium	6443	6539	6443
Zusammen	11728	11867	11729

Tabelle 4.2: Schüler in weiterführenden Schulen nach Bildungsabschluß des Familienvorstandes, absolut und als prozentualer Anteil an Schülern aller Schularten

	absolut	in Prozent, ermittelt aus:		
Bildungsabschluß des Familienvorstandes	Randsummen Original-Tabelle	Randsummen Original-Tabelle	Randsummen überlagerte Tabelle	überlagerte Original-Randsummen
(1) Volks-/Hauptschule ohne Lehre, o. Ang.	769	28.1	28.2	28.1
(2) Volks-/Hauptschule mit Lehre	2958	51.7	51.2	51.7
(3) Realschule	1124	77.3	77.9	77.5
(4) Abitur	272	76.0	81.5	74.9
(5) Fachhochschule, Hochschule	1320	90.9	89.8	90.9

Tabelle 4.3: Schüler in weiterführenden Schulen nach beruflicher Stellung des Familienvorstandes, absolut und als prozentualer Anteil an Schülern aller Schularten

	absolut	in Prozent, ermittelt aus:		
Berufliche Stellung des Familienvorstandes	Randsummen Original-Tabelle	Randsummen Original-Tabelle	Randsummen überlagerte Tabelle	überlagerte Original-Randsummen
(1) Un- und angelernte Arbeiter	691	29.0	29.7	29.1
(2) Vor- und Facharbeiter, Meister	1075	46.4	47.3	46.4
(3) Einfache Angestellte und Beamte	1284	68.3	68.5	68.3
(4) Höhere Angestellte und Beamte	1892	85.3	84.7	85.3
(5) Selbständige	985	65.5	64.6	65.4
(6) Nichterwerbstätige, keine Angabe	516	36.1	36.8	35.9

nur 272 Schüler in weiterführenden Schulen. Die große Differenz der Anteile deutet auf Probleme bei der multivariaten Analyse hin. Um den Fehler zu verringern, könnte man rekodieren und neu auszählen. Dies haben wir jedoch nicht getan, da einerseits die Kategorie "Abitur" in der Forschung häufig verwendet wird und andererseits gerade die Folgen der Überlagerung bei geringen Besetzungszahlen untersucht werden sollen.[6]

Zur Differenzierung des Merkmals Stellung im Beruf wurde das Mikrozensusmerkmal Stellung im Betrieb verwendet. Mit den darin enthaltenen Informationen lassen sich die Gruppen der Arbeiter, Angestellten und Beamten entsprechend der Position in der betrieblichen Hierarchie gliedern.[7] Tabelle 4.3 zeigt, daß abgesehen von der heterogenen Gruppe der Selbständigen eine höhere berufliche Stellung mit größeren Anteilen der Kinder in weiterführenden Schulen einhergeht.

In Tabelle 4.4 sind die Anteile der Besucher weiterführender Schulen nach dem Nettoeinkommen des Familienvorstands gegliedert.[8] Ähnlich wie bei der schwach besetzten Kategorie "Abitur" in der Bildungsvariablen kann man für die Einkommenskategorie "unter 1000 DM" eine durch die Überlagerung erzeugte Verzerrung im Vergleich der Anteilswerte feststellen. Die Differenz beträgt hier aber nur drei Prozentpunkte.

Tabelle 4.4: Schüler in weiterführenden Schulen nach monatlichem Nettoeinkommen des Familienvorstandes, absolut und als prozentualer Anteil an Schülern aller Schularten

	absolut	in Prozent, ermittelt aus:		
monatl. Nettoeinkommen des Familienvorstandes	Randsummen Original-Tabelle	Randsummen Original-Tabelle	Randsummen überlagerte Tabelle	überlagerte Original-Randsummen
(1) unter 1000 DM	192	32.4	35.3	32.2
(2) 1000 - 2000 DM	1210	38.8	38.7	38.8
(3) 2000 - 3000 DM	2128	50.9	51.1	51.0
(4) 3000 - 4000 DM	1132	76.4	76.5	76.5
(5) 4000 - 5000 DM	597	86.9	86.3	87.0
(6) 5000 DM und mehr	663	90.6	90.4	90.6
(7) Landwirte, o.Angabe	521	55.4	54.7	55.4

4.1.2 Zum Überlagerungsfehler

In der Testtabelle gibt es K=5*6*7=210 Merkmalskombinationen der 3 unabhängigen Merkmale. n_{k2} sei die Anzahl der Schüler auf weiterführenden Schulen mit Merkmalskombination k und n_k sei die entsprechende Gesamtzahl der Schüler. Die analysierte *Originaltabelle* hat dann die Gestalt:

(4.1) (n_{k2}, n_k), k=1, ... ,K

Die analysierte *überlagerte Tabelle* läßt sich formal als

$$(4.2) \quad (n_{k2} + z_{k2}, n_k + z_k),\ k=1, \ldots, K$$

darstellen, wobei z_{k2} und z_k Realisationen ganzzahliger Zufallszahlen sind. Die analysierte überlagerte Tabelle ist *nicht* identisch mit der von STATIS-BUND gelieferten Tabelle. Die in der Tabelle vorhandenen negativen Zellenbesetzungen müssen aus Plausibilitätsgründen noch nachträglich korrigiert werden. Die für die Logit-Analyse verwendete Tabelle, entstand somit in zwei Schritten:

(i) Bereitstellung der zufallsüberlagerten Tabelle durch STATIS-BUND
(ii) Korrektur unplausibler Fallzahlen durch den Benutzer:
  (a) Negative Fallzahlen werden auf Null gesetzt. [9]
  (b) Die Gesamtzahl der Schüler für eine Merkmalskombination k wird gegebenenfalls soweit erhöht, daß $n_{k2} + z_{k2} < n_k + z_k$ gilt. [10]

Bei einer teilweise so schwach besetzten Tabelle wirkt sich die notwendige Korrektur unplausibler Fallzahlen zusätzlich zu dem STATIS-BUND Überlagerungsfehler aus. Die Kenngrößen für den Gesamtfehler vor und nach der Korrektur sind in Tabelle 4.5 zusammengefaßt.

Tabelle 4.5: Kenngrößen des Gesamtfehlers in der überlagerten Tabelle

Kenngrößen des Gesamtfehlers	vor Korrektur	nach Korrektur	Freiheits-grade
Mittelwert	0.40	0.56	-
Varianz	3.27	3.36	-
Chi-Quadrat-Statistik:			
mit Nullzellen	734	823	419
ohne Nullzellen	185	207	354

Die Chi^2-Statistik wurde zweifach berechnet; einmal mit Berücksichtigung der Nullzellen in der Originaltabelle, die dazu mit dem Wert 0,5 belegt wurden und ein zweitesmal ohne Berücksichtigung der 65 Nullzellen in der Originaltabelle. Beide Ergebnisse weisen auf eine insgesamt gute Anpassung der überlagerten Tabelle an die Originaltabelle hin.

Während die Fallzahlen unserer Testtabelle durch die nachträgliche Korrektur zusätzlich verzerrt werden, ist für die eigentlich interessierenden Anteilswerte eher das Gegenteil festzustellen. Der Anteil an Schülern auf weiterführenden Schulen beträgt in der Originaltabelle 54,94 Prozent, in der überlagerten Tabelle 55,39 Prozent und in der korrigierten Tabelle 55,10 Prozent. Die Werte aus der überlagerten und plausibel gemachten Tabelle

liegen im allgemeinen sehr nahe bei den Originalwerten (siehe Tabellen 4.2-4.4). Sowohl der STATIS-BUND Überlagerungsfehler als auch die Korrektur unplausibler Fallzahlen wirkt sich bei kleinen Fallzahlen stärker aus.

4.2 Vergleichende Analyse

4.2.1 Binäre Logit-Modelle

Bevor wir die Ergebnisse für die Original- und überlagerte Tabelle miteinander vergleichen, stellen wir die Grundzüge der Analyse mit binären Logit-Modellen dar (vgl. Arminger/Küsters 1986). Dabei gehen wir zunächst von der Originaltabelle aus und verwenden folgende Bezeichnungen:

N	:	Stichprobenumfang (hier: N = 11728)
K	:	Anzahl der Merkmalskombinationen der unabhängigen Merkmale (hier: K = 210)
Y	:	Y = 1: Besuch einer Hauptschule Y = 2: Besuch einer weiterführenden Schule
n_k, k=1, ... ,K	:	Anzahl der Fälle mit Merkmalskombination k
n_{k2}, k=1, ... ,K	:	Anzahl der Fälle mit Merkmalskombination k und Y=2
$\pi_k \equiv P(Y=2 \mid K)$	:	Wahrscheinlichkeit für Y=2 bei Vorliegen der k-ten Merkmalskombination

In der Gesamtstichprobe vom Umfang N trete die k-te Merkmalskombination genau n_k-mal auf. Die Zufallsvariable N_{k2}, die die Anzahl der Fälle mit Merkmalskombination k und Y=2 beschreibt, ist dann binomialverteilt:

$$(4.3) \quad P(N_{k2}=n_{k2}) = \binom{n_k}{n_{k2}} \pi_k^{n_{k2}} (1-\pi_k)^{n_k-n_{k2}}$$

Weiterhin wird angenommen, daß der Logarithmus der relativen Chancen $\pi_k/(1-\pi_k)$ eine weiterführende Schule statt eine Hauptschule zu besuchen, linear von den Ausprägungen der unabhängigen Merkmale abhängt:

$$(4.4) \quad \gamma_k = \ln\left(\frac{\pi_k}{(1-\pi_k)}\right) = \beta_0 + \beta_1 X_{k1} + \beta_2 X_{k2} + \ldots + \beta_p X_{kp} = X\beta$$
$$x_{kj} \in \{0,1\}, \ \beta_j \in \mathbb{R}.$$

Das *Nullmodell* M_0 mit nur einem Parameter β_0 ist das einfachste Logit-Modell. Es enthält die Hypothese, daß die Chancen nicht von den erklärenden Merkmalen abhängen. Demgegenüber ist das *saturierte Modell* M_s, das soviele β-Parameter wie Merkmalskombinationen enthält (p+1=K), das komplizierteste, aber am wenigsten informative Modell. Unter der

jeweiligen Modellannahme werden die Maximum-Likelihood-Schätzer (ML-Schätzer) für die ß-Koeffizienten bestimmt. Die zugehörige *Log-Likelihood-Funktion* lautet:

$$(4.5) \quad l(\beta) = \sum_{k=1}^{K} \gamma_k n_{k2} - \log(1+\exp(\gamma_k))\, n_k + \sum_{k=1}^{K} \log\binom{n_k}{n_{k2}}$$

Bei der folgenden Logit-Analyse bleibt der Überlagerungsfehler unberücksichtigt. Die Log-Likelihood-Funktion für die überlagerte Tabelle hat daher die Gestalt (4.5) mit $n_{k2} + z_{k2}$ statt n_{k2} und $n_k + z_k$ statt n_k. Der zugehörige Schätzer für ß wird als *Quasi-Maximum-Likelihood-Schätzer* bezeichnet (vgl. Küchenhoff 1990). In Abschnitt fünf werden Möglichkeiten aufgezeigt, den Zufallsfehler zu berücksichtigen.

Zur Beurteilung der Modellanpassung werden zwei häufig verwendete Kriterien herangezogen, die sich beide aus dem Devianzmaß ableiten. Die *Devianz* ist ein Maß für die Diskrepanz zwischen der unter der Modellannahme geschätzten und der beobachteten Tabelle.

$$(4.6) \quad D(M) = -2\,[\,l(\hat{\beta}_M) - l(\hat{\beta}_{M_S})\,]$$

Dabei bezeichnet $\hat{\beta}_M$ die ML-Schätzung für ß unter dem Modell M. Sollen zwei geschachtelte Modelle M_1 und M_2 miteinander verglichen werden, so bietet sich die Differenz der Devianzen als Teststatistik an, da $D(M_1)-D(M_2)$ unter der Nullhypothese M_1 asymptotisch Chi^2 verteilt ist. $D(M_1)$ bezeichne die Devianz der Individualdaten, das heißt die maximal mögliche Devianz für alle denkbaren Tabellen, und $D(M_0)$ die Devianz des Nullmodells, die maximal mögliche Devianz für die vorliegende Tabelle. Das Bestimmtheitsmaß *Pseudo-R^2*

$$(4.7) \quad R^2(M) = \frac{D(M_0) - D(M)}{D(M_I)}$$

gibt den Anteil an Devianz in den Daten an, der durch die in einem Modell M enthaltenen Effekte erklärt werden kann.

4.2.2 Modellwahl und Modellanpassung

Erhielte man mit überlagerten Fallzahltabellen ein anderes passendes Modell als mit der Originaltabelle, wäre das ein fataler Fehler. Deshalb wird

zunächst geprüft, ob bei einer induktiven Modellsuche gleiche Ergebnisse gefunden werden.

In Tabelle 4.6 sind die unter verschiedenen Modellannahmen ermittelten Kenngrößen sowohl für die Originaltabelle als auch für die überlagerte Tabelle wiedergegeben. Die absoluten Devianzen für die überlagerte Tabelle werden ebenso wie die zugehörigen Freiheitsgrade systematisch überschätzt. Auf den Vergleich von zwei Modellen hat dies jedoch keine Auswirkung, da die Differenz der Devianzen und Freiheitsgrade dicht beieinander liegen und die Chi^2-Tests für beide Tabellen stets zum gleichen Ergebnis, dem Verwerfen der jeweiligen Nullhypothese, führen.

Tabelle 4.6: Kenngrößen verschiedener Logit-Modelle

Modell	Original-Tabelle	Differenz zum Vergleichsmodell	Überlagerte Tabelle	Differenz zum Vergleichsmodell
Modell 0: 1				
Devianz	3107.63		3234.38	
Freiheitsgrade	181		207	
R^2	0		0	
Modell 1: 1+Bild		Modell 0		Modell 0
Devianz	982.06	2125.57	1039.67	2194.71
Freiheitsgrade	177	4	203	4
R^2	0.1317	0.1317	0.1344	0.1344
Modell 2: 1+StiB		Modell 0		Modell 0
Devianz	1037.84	2069.79	1250.69	1983.69
Freiheitsgrade	176	5	202	5
R^2	0.1282	0.1282	0.1215	0.1215
Modell 3: 1+Eink		Modell 0		Modell 0
Devianz	1567.41	1540.22	1700.24	1534.14
Freiheitsgrade	175	6	201	6
R^2	0.0954	0.0954	0.0940	0.0940
Modell 4: 1+Eink+StiB		Modell 7		Modell 7
Devianz	746.81	437.59	905.31	523.56
Freiheitsgrade	170	4	196	4
R^2	0.1462	0.0271	0.1426	0.0321
Modell 5: 1+Eink+Bild		Modell 7		Modell 7
Devianz	643.29	334.07	692.65	310.90
Freiheitsgrade	171	5	197	5
R^2	0.1526	0.0207	0.1557	0.0190
Modell 6: 1+Bild+StiB		Modell 7		Modell 7
Devianz	404.29	95.07	498.83	117.08
Freiheitsgrade	172	6	198	6
R^2	0.1675	0.0058	0.1675	0.0072
Modell 7: 1+Bild+StiB+Eink		Modell 0		Modell 0
Devianz	309.22	2798.41	381.75	2852.63
Freiheitsgrade	166	15	192	15
R^2	0.1733	0.1733	0.1747	0.1747
Maximales R^2	0.1925		0.1981	

Bei großen Fallzahlen sind Chi^2-Tests oft wegen ihrer asymptotischen Eigenschaften wenig informativ für die Beantwortung der Frage, welches Modell die Daten ausreichend gut und sparsam beschreibt. Deshalb war für den Vergleich und die Auswahl eines geeigneten Modells das Bestimmtheitsmaß R^2 eine heuristische Entscheidungsgrundlage. Die daraus abgeleiteten Schlüsse erweisen sich als unabhängig von der Überlagerung. Das maximale R^2 beträgt für die Originaltabelle (überlagerte Tabelle) 19,25 Prozent (19,81 Prozent). Damit enthält die untersuchte Tabelle wesentliche Bestimmungsgründe für die Zielgröße. Das Modell 7 "unabhängiger Einfluß aller drei Einzelmerkmale" mit R^2 = 17,33 Prozent (17,47 Prozent) kann als geeignet akzeptiert werden. Es weicht zwar signifikant von den Daten ab, "erklärt" aber bereits 90,0 Prozent (88,2 Prozent) der maximal erklärbaren Devianz. Diese Entscheidung legt auch der Vergleich weiterer, hier nicht dargestellter Modelle mit Interaktionen der unabhängigen Variablen nahe. Von den drei Einzelmerkmalen hat das Einkommen des Familienvorstandes den geringsten partiellen Einfluß auf die Bildungschance des Kindes. Das partielle R^2 ergibt sich durch einen Vergleich der R^2-Werte für die Modelle 4-6 mit dem Modell 7 (siehe Spalten "Differenz zum Vergleichsmodell" in Tabelle 4.6). Die jeweils in den Modellen 4-6 nicht enthaltene Variable verursacht ein geringeres R^2; der in Spalte "Differenz..." ausgewiesene Wert entspricht dem Nettoeffekt der Variablen. Die überlagerte Tabelle überschätzt im Vergleich zur Originaltabelle die partiellen Effekte Bildungsabschluß (3,21 Prozent vs. 2,71 Prozent) und Einkommen (0,72 Prozent vs. 0,58 Prozent), wohingegen der partielle Einfluß der beruflichen Stellung (1,90 Prozent vs. 2,07 Prozent) unterschätzt wird. Die absoluten Verzerrungen sind jedoch gering und die für die inhaltliche Interpretation wichtige relative Anordnung der drei Merkmale wird durch die Zufallsüberlagerung nicht verändert. Insgesamt kann festgehalten werden, daß die Zufallsüberlagerung die Modellanpassung und die Wahl eines geeigneten Modells nicht wesentlich beeinflußt.

4.2.3 Regressionskoeffizienten

Wir wenden uns nun einem Vergleich der Analyseergebnisse für das ausgewählte Modell 7 zu. Tabelle 4.7 enthält die geschätzten ß-Koeffizienten und ihre geschätzten Standardabweichungen. Richtung und Größenordnung stimmen für überlagerte und Originaltabelle weitgehend überein. Da die Schätzwerte asymptotisch normalverteilt sind, lassen sich näherungsweise 95 Prozent-Konfidenzintervalle für die ß-Koeffizienten der Originaltabelle ableiten. Es zeigt sich, daß nur einer der sechzehn ß-Werte der überlagerten Tabelle außerhalb des Konfidenzintervalls liegt und somit signifikant von dem ß-Koeffizienten der Originaltabelle abweicht. Der Koeffizient zu "Bildungsabschluß = Abitur" wird durch die Überlagerung überschätzt (siehe "Zum Überlagerungsfehler" Abschnitt 4.1.2).

Tabelle 4.7: Vergleich der ß-Koeffizienten und ihrer Standardabweichungen für Modell 7

	Original - Tabelle		Überlagerte Tabelle	
Merkmals-ausprägung	Koeffizient	Standard-abweichung	Koeffizient	Standard-abweichung
1	-1.1842	0.0611	-1.1772	0.0608
Bild(2)	0.6140	0.0559	0.6055	0.0555
Bild(3)	1.3736	0.0832	1.4405	0.0831
Bild(4)	1.3422	0.1387	1.7409	0.1383 *)
Bild(5)	1.8623	0.1180	1.8093	0.1102
StiB(2)	0.4475	0.0664	0.4612	0.0657
StiB(3)	1.0495	0.0737	1.0275	0.0732
StiB(4)	1.3338	0.0911	1.2565	0.0889
StiB(5)	0.8694	0.0852	0.7936	0.0843
StiB(6)	0.1958	0.0803	0.1894	0.0795
Eink(1)	-0.2639	0.1094	-0.2204	0.1057
Eink(2)	-0.1138	0.0533	-0.1321	0.0533
Eink(4)	0.4225	0.0766	0.4495	0.0763
Eink(5)	0.6497	0.1305	0.6732	0.1259
Eink(6)	0.9040	0.1457	0.9994	0.1410
Eink(7)	0.0721	0.0881	0.0476	0.0868

*) ß der überlagerten Tabelle außerhalb des 95% Konfidenzintervalls von ß der Original-Tabelle

Tabelle 4.8: Gegenüberstellung der t-Werte, die zu verschiedenen Ergebnissen bei einem t-Test auf Differenz von zwei ß-Koeffizienten führen

	Original - Tabelle		Überlagerte Tabelle	
Merkmalsausprägung	ß-Differenz	t-Wert	ß-Differenz	t-Wert
Bild(4)-Bild(3)	-0.0314	-0.22	0.3005	2.07
Bild(5)-Bild(4)	0.5201	3.19	0.0684	0.43
StiB(3)-Bild(4)	-0.2928	-1.77	-0.7134	-4.37
StiB(4)-Bild(4)	-0.0084	-0.05	-0.4845	-2.83
StiB(5)-Bild(2)	0.2553	2.31	0.1881	1.72
Eink(4)-Bild(2)	-0.1916	-2.00	-0.1560	-1.64
Eink(6)-Bild(2)	0.2900	1.86	0.3939	2.61
Eink(6)-StiB(4)	-0.4298	-2.36	-0.2571	-1.46

Tabelle 4.8 vergleicht die Ergebnisse bei einem t-Test auf signifikanten Unterschied zwischen je zwei ß-Koeffizienten. Bei einem Signifikanzniveau von fünf Prozent und einem kritischen Wert von ±2 führt die Überlagerung in 8 von 120 Fällen zu einem anderen Ergebnis. Besonders deutlich ist die Diskrepanz in den vier Fällen, in denen der Koeffizient zu "Bildungsabschluß = Abitur" betroffen ist.

Wie schon im Abschnitt 3.1 angesprochen, bewirkt die Fehlerüberlagerung von Zellenbesetzungen tendenziell eine Überschätzung des Anteilswertes von überlagerten Zellen im Vergleich zur Originaltabelle, was sich besonders bei kleinen Fallzahlen deutlich zeigt (siehe Formel 3.3). Dies gilt auch für die durch das Logit-Modell 7 geschätzten Zellenbesetzungen (m_u) der überlagerten Tabelle im Vergleich zu den geschätzten Zellenbesetzungen (m_o) der Originaltabelle.

In Abbildung 4.1 sind die relativen Differenzen geschätzter Zellenbesetzungen $rd_k=(m_{ku}-m_{ko})/m_{ko}$ dargestellt.[11] Eine relative Differenz von Eins zum Beispiel bedeutet, daß bei der überlagerten Tabelle die geschätzte Fallzahl doppelt so groß ist, wie bei der Originaltabelle. Die Verzerrungen werden etwa ab einer geschätzten Fallzahl von zehn zunehmend kleiner und können bei Werten größer 20 praktisch vernachlässigt werden.

Mit Bezug auf Abschnitt 3.1, Formel 3.5, war tendenziell eine Unterschätzung der Varianz zu erwarten. Die in Abbildung 4.2 dargestellte Überschätzung der geschätzten Varianz scheint zunächst damit in Widerspruch zu stehen. Das klärt sich jedoch auf, wenn man beachtet, daß auch die zugrundeliegende geschätzte Fallzahl der überlagerten Tabelle im Vergleich zur Originaltabelle größer ist (siehe Abbildung 4.1).

4.2.4 Residuenanalyse

Weiteren Aufschluß über die Modellanpassung liefert die exploratorische Residuenanalyse, in der die beobachteten Werte mit den geschätzten Werten verglichen werden. Wir verwenden die Pearson-Residuen:

$$(4.8) \qquad S_k = \frac{n_{2k} - \hat{\pi}_k n_k}{\sqrt{\hat{\pi}_k (1 - \hat{\pi}_k)\, n_k}}$$

Ist das Modell gut angepaßt und gilt für jede Merkmalskombination k, daß $n_k \pi_k (1-\pi_k) > 9$ so sind die Pearson-Residuen annähernd standardnormalverteilt. $|S_k| \geq 2$ ist somit ein Indiz für eine durch das Modell schlecht angepaßte Merkmalskombination.

Abbildung 4.1: Relative Differenzen geschätzter Zellenbesetzungen

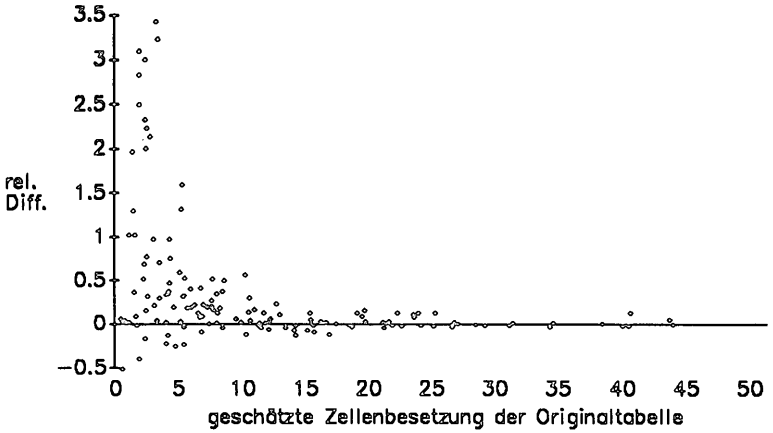
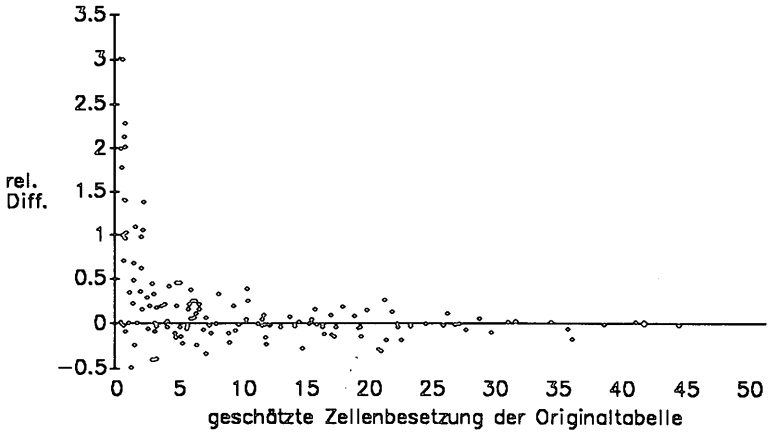


Abbildung 4.2: Relative Differenzen geschätzter Varianzen



In Tabelle 4.9 sind diejenigen Merkmalskombinationen ausgewiesen, die zu unterschiedlichen Ergebnissen bei der Residualanalyse führen. Im Hinblick auf die konservative Bedingung $n_k \pi_k (1-\pi_k) > 9$ wurden in Anlehnung an eine häufig verwendete Faustregel nur Residuen mit $n_k \pi_k \geq 5$ betrachtet. Bei der Originaltabelle lagen elf Residuen im kritischen Bereich; bei der überlagerten Tabelle waren es 16. Ein Vergleich der Residuen in den kritischen Bereichen zwischen Original- und überlagerter Tabelle zeigt jedoch, daß sowohl bei der Originaltabelle Merkmalskombinationen schlecht angepaßt sind, die bei der überlagerten Tabelle gut angepaßt sind, als auch umgekehrt. Insofern kann aus der Residuenanalyse nicht auf einen Überlagerungseffekt geschlossen werden.

Tabelle 4.9: Gegenüberstellung der Pearson-Residuen, die zu verschiedenen Ergebnissen in der Residuenanalyse führen

Merkmalskombination			Original - Tabelle		Überlagerte Tabelle	
Bild	StiB	Eink	geschätzte Fallzahl	Residuum	geschätzte Fallzahl	Residuum
2	2	1	5.66	-0.91	5.43	-2.49
2	3	6	7.20	-0.16	8.11	-2.51
2	4	6	26.92	1.49	26.99	2.44
2	5	6	44.61	-1.75	44.02	-2.22
2	5	7	156.81	-2.23	148.55	-1.44
3	4	2	6.43	-3.05	4.80	1.23
3	5	2	16.56	1.13	14.32	2.32
3	6	7	9.80	1.13	9.96	2.08
4	5	1	2.05	1.18	5.30	-2.03
4	6	2	9.51	1.21	11.06	2.51
5	3	6	7.46	-2.07	7.48	-0.68

4.2.5 Direkter Test auf Unterschiede zwischen den Tabellen

Um den Einfluß der Überlagerung auf die geschätzten Effektkoeffizienten im Modell 7 direkt zu testen, wurde eine zusammengesetzte Tabelle gebildet. Diese neue Tabelle besitzt als viertes Merkmal die Tabellenart mit zwei Ausprägungen (TAB = 1: Originaltabelle, TAB = 2: überlagerte Tabelle).

Tabelle 4.10: Direkter Test auf Überlagerungseffekt

Modell	Devianz	Freiheits-grad	R^2 in %
1) 1+Tab+Bild+Stib+Eink	697.83	373	17.38
2) 1+Tab+Bild+Stib+Eink +Tab*(Bild+Stib+Eink)	690.97	358	17.40

Modell 1 enthält die Hypothese, daß der (geschätzte) Einfluß der drei unabhängigen Merkmale auf die Chance eines Schülers, eine weiterführende Schule zu besuchen, nicht von der Tabellenart abhängt. Im Gegensatz dazu wird im zweiten Modell unterstellt, daß dieser Einfluß sehr wohl von der Tabellenart und damit von der Überlagerung abhängt. Es zeigt sich, daß die 15 zusätzlichen Koeffizienten eine nicht signifikante Devianzreduktion von 6,86 erzielen. Ein t-Test für die einzelnen Koeffizienten ergibt ferner, daß lediglich der Koeffizient zu "Bild = Abitur" * "Tab = 2" signifikant von Null abweicht. Die geringfügige Verringerung der Devianz im zweiten Modell ist also im wesentlichen auf eine Verbesserung der Schätzungen für diese kritische Kategorie zurückzuführen.

5. Möglichkeiten der Einbeziehung eines Zufallsfehlers

Im letzten Kapitel wurde eine Logit-Analyse einer STATIS-BUND Tabelle durchgeführt, die beim Schätzalgorithmus die Zufallsüberlagerung völlig außer acht ließ (Quasi-ML-Schätzung). In diesem Kapitel werden drei Möglichkeiten aufgezeigt, wie der Überlagerungsfehler berücksichtigt werden könnte. Die Vorschläge behandeln *exakt normalverteilte Fehler*. Da für eine konkrete Tabelle aus STATIS-BUND die Normalverteilungsannahme verletzt ist, lassen sich daraus keine allgemeingültigen Empfehlungen für die praktische Arbeit ableiten.

5.1 Fehlerverringerung mittels linearer Regression

Da die Summen fehlerüberlagerter Zellen stärker um die wahren Randsummen streuen als die ausgezählten und überlagerten Randsummen, liegt es nahe, eine Tabelle zu schätzen, die in sich möglichst konsistent ist.[12)] Diese geschätzte Tabelle ließe sich in einem zweiten Schritt für die eigentlichen Analysen verwenden. Zur Fehlerverringerung könnte die folgende einfache Methode dienen, mit der berücksichtigt wird, daß zwischen den Zellen und den Randverteilungen eine lineare Beziehung besteht. Anhand der Beispieltabelle 3.1 aus Kapitel 3 soll diese Methode beschrieben werden. Die abhängige Variable Y besteht aus den vier überlagerten Zellhäufigkeiten, je zwei überlagerten Spalten- bzw. Zeilensummen sowie der überlagerten Gesamtsumme. Die vier zu schätzenden Originalfallzahlen bilden die Regressionskoeffizienten $\beta = (\beta_1, \ldots, \beta_4)$. Unter der Annahme, daß Y normalverteilt ist mit $E(Y) = X*\beta$, wobei die (9x4) - Designmatrix X nur aus Nullen oder Einsen besteht, kann eine lineare Regression durchgeführt werden. Bei einem exakt normalverteilten Fehler konnte der Gesamtfehler in den vier überlagerten Fallzahlen verringert werden (vgl. Tabelle A1 im Anhang).

Für STATIS-BUND Tabellen ist der Erfolg dieses Verfahrens gering, da die Normalverteilungsannahme nicht erfüllt ist. Dennoch wurde das Verfahren auf die in Kapitel 4 beschriebene Testtabelle angewendet. Da bei der Testtabelle sehr viele schwach besetzte Zellen vorliegen, wird man mit dem Problem geschätzter negativer Zellenbesetzungen konfrontiert sein. Um dies zu umgehen, wurde die überlagerte Testtabelle nachträglich über das Einkommensmerkmal aggregiert. Dieses Vorgehen bietet zugleich einen empirischen Test, wie sich eine Aggregation überlagerter Tabellenfelder auswirkt. Zum Vergleich wurde eine weitere Tabelle ausgezählt, in der, wie in der nachträglich aggregierten Testtabelle, das Einkommensmerkmal unberücksichtigt blieb, jedoch sind hier die Zellen nur einfach mit Zufallsvariablen überlagert. Die nachträglich aggregierte Testtabelle (Schulbesuch * Bildungsabschluß * berufliche Stellung) weicht mit einem Chi-Quadrat Wert von 393 bei 59 Freiheitsgraden signifikant von der aggregierten Originaltabelle ab. Folglich weichen auch die Analysen dieser Tabelle gravierend von den Ergebnissen der Originaltabelle ab. Wendet man das beschriebene Verfahren an, reduziert sich der Chi-Quadrat Wert auf 43, d.h. die geschätzte Tabelle weicht nicht signifikant von der Originaltabelle ab. In einem Fall wurde eine negative Zellenbesetzung geschätzt. Die vor der Überlagerung über das Einkommensmerkmal aggregierte Tabelle weicht von der Originaltabelle nur um einen nicht signifikanten Chi-Quadrat Wert von 8 ab. Nach Anwendung des Verfahrens auf diese Tabelle zeigt sich jedoch der eingeschränkte Nutzen des Verfahrens, denn die geschätzte Tabelle ist mit einen Chi-Quadrat Wert von 11 schlechter als die überlagerte Tabelle.

5.2 Logit-Modell mit Fehlern in den Fallzahlen

Im folgenden wird ein binomiales Logit-Modell für Tabellen mit normalverteilten Fehlern formuliert und dem binomialen Logit-Modell aus Abschnitt 4.2.1 gegenübergestellt. Es wird angenommen, daß die Randsumme W_k normalverteilt ist ($W_k \sim N(n_k, \sigma^2)$) und die Fallzahl W_{k2} die unabhängige Summe aus einer binomialverteilten Größe ($B(n_k, \pi_k)$) und einer normalverteilten Größe ($N(0,\sigma^2)$) ist. Die Aspekte der Ganzzahligkeit und Positivität von Fallzahlen werden in diesem Modell außer acht gelassen:

(a) Logit-Modell mit normalverteilten Überlagerungsfehlern unbekannte Parameter: $\beta = (\beta_0, \ldots \beta_p)^t, \pi_k, n_k, \sigma^2$

(5.1) $W_{12}, \ldots, W_{K2}, W_1, \ldots, W_K$ *stochastisch unabhängig mit*

(i) $P\{W_{k2} \le t\} = \sum_{i=0}^{n_k} N_{(i,\sigma^2)}(-\infty, t]\, B_{(n_k,\pi_k)}(i), \quad t \in \mathbb{R}$

(ii) $P\{W_k \le t\} = N_{(n_k,\sigma^2)}(-\infty, t), \quad t \in \mathbb{R}$

(iii) $\pi_k = \exp(X_k^t \beta) \;/\; (1 + \exp(X_k^t \beta))$

(b) Logit-Modell ohne Überlagerungsfehler unbekannte Parameter:
$\beta = (\beta_0, \ldots \beta_p)^t$, π_k

(5.2) $N_{12}, \ldots, N_{K2}$ *stochastisch unabhängig mit*

(i) $P\{N_{k2} \leq t\} = B_{(n_k, \pi_k)}(-\infty, t]$, $t \in \mathbb{R}$

(ii) $\pi_k = \exp(X_k^t \beta) \; / \; (1 + \exp(X_k^t \beta))$

Unter der einfachsten Modellannahme $\pi = \pi_k = \exp(\beta_0)/(1+\exp(\beta_0))$ vergleichen wir die beiden Likelihood-Gleichungen, die sich aus der Log-likelihood-Funktion durch Differentiation nach β_0 ergeben. Gleichungen für die Parameter n_k und σ^2 im Fall (a) sind nicht aufgeführt. Phi(i, σ^2) bezeichnet die Dichte der Normalverteilung mit Mittelwert i und Varianz σ^2.

(5.3)

$$a) \quad \sum_{k=1}^{K} \sum_{i=0}^{n_k} \frac{f_{ki}(\pi)}{f_k(\pi)} [i - n_k \pi] = 0 \quad mit$$

$$f_k(\pi) = \sum_{i=0}^{n_k} f_{ki}(\pi) = \sum_{i=0}^{n_k} \varphi_{(i, \sigma^2)}(w_{k2}) \binom{n_k}{i} \pi^i (1-\pi)^{n_k - i}$$

$$b) \quad \sum_{k=1}^{K} [n_{k2} - n_k \pi] = 0$$

Hieraus ist bereits ersichtlich, welche rechentechnischen Schwierigkeiten beim Lösen der Likelihood-Gleichungen einer zufallsüberlagerten Tabelle mit normalverteilten Fehlern im Vergleich zum Standardmodell auftreten können. Lösungsansätze für Maximum-Likelihood-Schätzungen von gemischten Verteilungen, wie sie die Verteilung von W_{k2} darstellt, finden sich zum Beispiel in Everitt/Hand (1981).

Die in der Literatur diskutierten 'Fehler-in-den-Variablen-Modelle' behandeln Regressionsprobleme, bei denen die erklärenden Variablen fehlerbehaftet sind (vgl. Bickel/Ritov 1987, Küchenhoff 1990). Küchenhoff (1990) befaßt sich unter anderem mit Logit-Modellen der Form $P(Y=1 \mid X=x)=(1+\exp(-\alpha-\beta x))^{-1}$, wobei die normalverteilte Einflußgröße X nicht beobachtbar ist, da sie von einer ebenfalls normalverteilten Störgröße überlagert wird. Ein verwandtes Problem ist unter dem Stichwort "overdispersion" bekannt (vgl. McCullagh/Nelder 1983). Es tritt insbesondere dann auf, wenn wichtige erklärende Merkmale nicht verfügbar sind oder eine zu hohe Aggregationsstufe gewählt wurde. Die Erfolgswahrscheinlichkeit π_k in Kategorie k wird dann als zufällig angesehen. Pierce/Sands (1975) untersuchten zum Beispiel ein Logit-Modell der Form $\pi_k = P(Y=1 \mid x_k) = (1+\exp(x_k^t \beta + \sigma z))^{-1}$, wobei hier die Störgröße z nicht beobachtbar ist. Ist z normalverteilt, so besitzt π_k eine

logistisch-normale Verteilung (vgl. Williams 1982). Aus der Literatur sind uns Logit-Modelle mit Fehlern gemäß (5.1) nicht bekannt.

5.3 Modifizierte Maximum-Likelihood-Schätzung

Offensichtlich ist es nicht einfach, ein statistisches Verfahren zu finden, das den Zusatzfehler in multivariaten Analysen überlagerter Fallzahltabellen geeignet berücksichtigt. Im folgenden werden erste, pragmatisch orientierte Überlegungen in diese Richtung vorgestellt.

In Abschnitt 3.1 wurde gezeigt, daß der Anteilswert bei überlagerten Fallzahlen tendenziell um den Faktor σ_n^2/μ_n^2 überschätzt wird (siehe Formel 3.3). Um diese Verzerrung zu korrigieren, verringern wir die interessierenden Fallzahlen n_{k2} um diesen Faktor. Mithilfe einer weiteren Näherung (Taylor-Reihe) läßt sich zeigen, daß die Korrektur unter der Normalverteilungsannahme Verbesserungen bringt, wenn die Randsumme größer gleich fünf ist.[13] Man erhält mit (5.4) die neuen Fallzahlen n^*_{k2}:

$$(5.4) \quad n^*_{k2} = n_{k2} \left(1 - \frac{\sigma_n^2}{n_k^2}\right)$$

Im Gegensatz zur Überschätzung des Anteilswertes wurde im Abschnitt 3.1 eine Unterschätzung der Varianzen nach dem Binomialansatz festgestellt (siehe Formeln 3.4 und 3.5). Da die ML-Schätzung für Logit-Modelle mittels einer iterativen gewichteten linearen Regression erfolgt, bei der die geschätzten Varianzen als Gewichtungsfaktoren eingehen, ist eine Korrektur dieser Varianzen naheliegend. Mit m_{k2} seien die geschätzten Fallzahlen bezeichnet; die Gesamtschülerzahl sei n_k. Die Varianzfunktion für normale Logit-Modelle lautet, der Einfachheit halber ohne Indizes: $var(m) = m - m^2/n$. In Anlehnung an Formel 3.5 korrigieren wir mit (5.5) die iterativ berechnete Varianz durch Vertauschen der Vorzeichen in der Klammer:[14]

$$(5.5) \quad \hat{\sigma}^2_{m_{k2}} = m_{k2} - \frac{m_{k2}^2}{n_k} \left(1 - \frac{\sigma^2_{n_k}}{n_k^2} - \frac{\sigma^2_{m_{k2}}}{m_{k2}^2}\right)$$

Beobachtungen mit hoher Varianz werden bei der Parameterschätzung weniger berücksichtigt als Beobachtungen mit geringer Varianz. Die modifizierte Varianzfunktion (5.5) ist plausibel, da die Varianz besonders bei kleinen Zellenbesetzungen künstlich größer wird. Die Varianz der Überlagerung wird mit Drei angenommen. Der Einfachheit halber wird diese Annahme auch für die Varianz der geschätzten Fallzahlen m getroffen. Eventuell durch die Modellierung entstehende Kovarianzen zwischen den

geschätzten Zellenbesetzungen und der Randsumme bleiben dabei ebenfalls außer Acht.

Nach diesen beiden Korrekturen zeigen die Ergebnisse in Tabelle 5.1 eine deutliche Verbesserung im Vergleich zu den vorherigen Ergebnissen (siehe Tabellen 4.6 und 4.7).

Tabelle 5.1: Modifizierte ML-Schätzung für Modell 7

Devianz: 342.04
Freiheitsgrade: 192

Merkmals-ausprägung	Koeffizient	Standard-abweichung
1	-1.1931	0.0614
Bild(2)	0.6050	0.0562
Bild(3)	1.3991	0.0845
Bild(4)	1.5384	0.1476
Bild(5)	1.7595	0.1153
StiB(2)	0.4709	0.0666
StiB(3)	1.0632	0.0742
StiB(4)	1.3326	0.0914
StiB(5)	0.8424	0.0861
StiB(6)	0.2154	0.0812
Eink(1)	0.4293	0.0773
Eink(2)	0.5609	0.1298
Eink(4)	0.8763	0.1466
Eink(5)	-0.2970	0.1113
Eink(6)	-0.1280	0.0537
Eink(7)	0.0139	0.0887

Gegenüber einer Devianz für Modell 7 bei der überlagerten Tabelle (Originaltabelle) von 381,75 (309,22) erhält man nach der Modifikation eine Devianz von 342,04. Die ß-Werte sind entsprechend näher an den Ergebnissen der Originaltabelle. Insbesondere weicht nun der Parameter für Bildungsabschluß = "Abitur" nicht mehr signifikant vom Schätzwert der Originaltabelle ab (vgl. Tabellen 4.7 und 5.1). Bei den t-Tests der Differenzen von je zwei ß-Werten sind nur noch drei Differenzen kritisch; zuvor waren es acht Differenzen. Ähnliche Ergebnisse erhält man beim Residuenvergleich.

Berücksichtigt man, daß die Modifikationen nur eine erste, grobe Annäherung an die schwierige Frage der Behandlung von Fehlern in den beobachteten Fallzahlen bei multivariaten Analysen darstellen, sind die Verbesserungen überraschend gut. Dies könnte ein Anlaß sein, in dem Bereich der Modifikation von Anteilswerten und Binomialvarianzen weiter nach Lösungen zu suchen.

6. Schluß

STATIS-BUND bietet neben einer Vielzahl von Zeitreihen und Strukturdaten unter bestimmten Voraussetzungen die Möglichkeit, per Online-Anschluß schnell und flexibel Auswertungen aus Einzeldaten der amtlichen Statistik zu erhalten. Dieser Zugriff erfolgt in Form von frei spezifizierbaren Fallzahltabellen, die aus Anonymisierungsgründen mit Zufallszahlen überlagert werden. Mit dem vorliegenden Aufsatz wurden erste Untersuchungen über die Auswirkungen dieser Zufallsüberlagerung auf multivariate Analysen vorgestellt. Zusammenfassend läßt sich festhalten, daß selbst bei einer sehr spärlich besetzten Testtabelle keine gravierenden Verzerrungen multivariater Analyseergebnisse auftreten. Die Wahl eines statistischen Modells, das die Struktur der Tabelle hinreichend gut beschreibt, wird durch die Überlagerung nicht wesentlich beeinflußt. Das gleiche gilt für die relative Bedeutung der Einzelmerkmale, wenngleich ihr partieller Einfluß auf die Zielgröße unter- bzw. überschätzt wird. Wie zu erwarten, sind Einzelergebnisse, die sich aus sehr schwach besetzten Kategorien ableiten, stark verzerrt. In unserer Beispieltabelle ist dies die Kategorie "Abitur" des Merkmals Bildungsabschluß des Familienvorstands. Als Präventivmaßnahme schlagen wir vor, die Randsummen der Einzelmerkmale stets auszuzählen. Weicht die überlagerte Randsumme stark von der Summe der überlagerten Einzelfelder ab, so ist eine erneute Tabellenerstellung unter Zusammenfassung kritischer Kategorien zu empfehlen; auch wenn dies unter inhaltlichen Gesichtspunkten nicht immer leicht fallen mag.

Bei der (bewußt) so tief gegliederten Testtabelle ist wegen der Größe des Stichprobenfehlers die statistisch gesicherte Interpretation der Analyseergebnisse teilweise nicht mehr gegeben. Weitere Logit-Analysen mit stärker besetzten STATIS-BUND-Tabellen zeigten keine nennenswerten Verzerrungen.[15] Dennoch lassen sich diese Befunde nicht ohne weiteres verallgemeinern, da hierzu weitere Analysen mit anderen Verfahren (lineare Regression, Log-lineare Modelle etc.) und anderen Tabellen erforderlich sind.

Neben der Darstellung eines praxisorientierten Vergleichs von Logit-Analysen wurden in Abschnitt 3.1 Näherungsformeln entwickelt, die Hinweise auf Art und Größenordnung der Verzerrungen wichtiger Schätzwerte liefern. Darüber hinaus wurden in den Abschnitten 5.2-5.3 erste Ansätze vorgestellt, wie man Fehler in den Fallzahlen bei multivariaten Analysen angemessen berücksichtigen könnte. In diesem Bereich ist weitere Forschung nötig.

Anmerkungen

1) Wir danken Siegfried Gabler für hilfreiche Hinweise zu mathematischen Fragen.
2) In Zukunft wird auch ein direkter Datentransfer von einem *Liefer-PC* zu dem PC eines Interessenten über das Telefonnetz möglich sein. Der Kunde sendet dazu über seinen mit einem Telefonmodem ausgestatteten PC die Datenanforderung an den PC des Statistischen Bundesamtes, der dann seinerseits automatisch die Bereitstellung der gewünschten Daten veranlaßt.
3) Aus Geheimhaltungsgründen kann die genaue Verteilung der Zufallszahlen nicht veröffentlicht werden.
4) Dabei wurde von einem durchschnittlichen Zuschlagsfaktor von k=1,6 des Standardfehlers für den Design-Effekt ausgegangen. Mit diesem Faktor wird berücksichtigt, daß sich der Standardfehler durch die Klumpenauswahl des Mikrozensus gegenüber einer reinen Zufallsstichprobe im allgemeinen erhöht.
5) Besucher von Sonderschulen werden im Mikrozensus wie Hauptschüler erfaßt. Gesamtschüler blieben außer Acht. Die Daten des Mikrozensus 1987 beziehen sich auf die Bevölkerung am Familienwohnsitz.
6) Bei den folgenden Analysen berücksichtigen wir die Vergrößerung des Stichprobenfehlers durch den Design-Effekt nicht (siehe Anmerkung 4). Es muß außerdem darauf hingewiesen werden, daß die statistische Aussagefähigkeit dieser Testtabelle aufgrund der geringen Besetzungszahlen zum Teil gering ist. Streng genommen können die Analysen nur heuristischen Charakter haben.
7) Auf eine Trennung zwischen Angestellten und Beamten wurde verzichtet.
8) Da für Landwirte im Mikrozensus keine Einkommensangabe erhoben wird, aber eine vollständige Gliederung angestrebt wurde, sind die Landwirte mit der Gruppe ohne Angabe bzw. ohne Einkommen zusammengefaßt.
9) Eine nachträgliche Korrektur negativer Fallzahlen ist nicht nötig, wenn man in STATIS-BUND die Option "Fallzahlen = positiv" wählt. Die Option bewirkt quasi, daß die Korrektur bereits intern bei der Erstellung der Tabelle durchgeführt wird. In jedem Falle ist aber der gesamte Überlagerungsfehler für kleine Fallzahlen tendenziell positiv.
10) Diese Korrektur ist nicht erforderlich, wenn man (z.B. in GLIM; Payne 1985) eine logistische Regression schätzt. Statt der Vektoren Besucher weiterführender Schulen (n_{k2}) und Gesamtschülerzahl ($n_{\cdot}$) im Logit-Modell wird in der logistischen Regression ein Fallzahlvektor Schulbesuch (Hauptschüler | Besucher weiterführender Schulen) je Merkmalskombination spezifiziert. Da jedoch die Varianz des Zufallsfehlers für die daraus zu berechnenden Gesamtschülerzahlen vergrößert wird, wurde das Logit-Modell gewählt.
11) Berücksichtigt wurden nur Zellen, die sowohl in der überlagerten als auch in der Originaltabelle mit Werten größer als Null besetzt waren. Der Übersichtlichkeit halber sind in den Abbildungen nur relative Differenzen für geschätzte Fallzahlen bis 50 ausgewählt worden.
12) Es handelt sich dabei um ein Problem der Anpassung von Zellen an vorgegebene Randverteilungen. Das Verfahren des Iterative Proportional Fitting (IPF) kann dafür nicht verwendet werden, da es von Fehlern ausgeht, die proportional zur Zellenbesetzung sind. Diese Voraussetzung wird jedoch von der Überlagerung im STATIS-BUND nicht erfüllt.
13) Tatsächlich wurden jedoch für die Testtabelle auch bei schwächer besetzten Randsummen Verbesserungen festgestellt.
14) Entstehen bei der Modifikation der Zellenbesetzungen negative Werte, werden sie auf Null gesetzt. Im Schätzalgorithmus muß aus diesem Grund ein Wert größer Null vergeben werden; hier 0,5.

15) Einer der Vergleiche wurde mit der über das Einkommensmerkmal aggregierten Tabelle durchgeführt (siehe Abschnitt 5.1). Bei einem anderen Test verwendeten wir eine Tabelle zur Fragestellung des Erwerbsstatus von Frauen in Abhängigkeit von der Kinderzahl. Diese Tabelle wies mit einer Dimension von 3*5 Zellen in der am schwächsten besetzten Zelle eine Fallzahl von 91 auf.

Literatur

Arminger, G./Küsters, U., 1986: Statistische Verfahren zur Analyse qualitativer Variablen (=Forschungsbericht der Bundesanstalt für Straßenwesen, Bereich Unfallforschung, 147). Bergisch-Gladbach.

Bickel, B.J./Ritov, Y., 1987: Efficient Estimation in the Errors in Variables Model. The Annals of Statistics, 15, 2:513-540.

Böttcher, W., 1991: Soziale Auslese im Bildungswesen. Die Deutsche Schule, 83, 2: 151-161.

Christensen, R., 1990, Log-linear Models, New York: Springer.

Everitt, B.S./Hand, D.J., 1981: Finite Mixture Distributions. London: Chapman and Hall.

Geißler, R., 1987: Soziale Schichtung und Bildungschancen, S. 79-110, in: ders. (Hrsg.): Soziale Schichtung und Lebenschancen in der Bundesrepublik Deutschland, Stuttgart: Enke.

Küchenhoff, H., 1990: Logit- und Probitregression mit Fehlern in den Variablen (=Mathematical systems in Economics, Vol. 177). Frankfurt: Hain.

Kühn, J./Pfrommer, F./Schrey, E., 1984: Zur technischen Weiterentwicklung des Statistischen Informationssystems. Wirtschaft und Statistik, 12:981-987.

McCullagh, P./Nelder, J.A., 1983: Generalized Linear Models. London: Chapman and Hall.

Payne, C.D., 1985: The GLIM Systems. Release 3.77. Oxford: NAG.

Pierce, D.A./Sands, B.R., 1975: Extra-Bernoulli variation in binary data. Technical Report 46, Dept. of Statistics, Oregon State University.

Statistisches Bundesamt, 1992: Statistisches Informationssystem des Bundes. Datenbestand 1992/1993. Wiesbaden.

Wauschkuhn, U., 1982: Anpassung von Stichproben und n-dimensionalen Tabellen an Randbedingungen. München/Wien: Oldenbourg.

Williams, D.A., 1982: Extra-Binomial Variation in Logistic Linear Models. Applied Statistics, 31: 144-148.

Anhang

1. Bestimmung der Erwartungswerte einer Funktion durch Taylor-Reihen

X und Y seien Zufallsvariablen mit Mittelwert $\mu(x)$ bzw. $\mu(y)$ und Varianz $\sigma^2(x)$ bzw. $\sigma^2(y)$. $\mu(x)$ sei dabei die Häufigkeit in der interessierenden Zelle der Originaltabelle und $\mu(y)$ die entsprechende Randsumme. Der Erwartungswert einer Funktion f(X,Y) läßt sich näherungsweise mit Hilfe einer Taylor-Reihen-Entwicklung (unter entsprechenden Voraussetzungen an f) bis zur zweiten Ordnung berechnen. Es gilt:

$$(A.1) \quad E(f(X,Y)) \approx f(\mu_x,\mu_y) + \frac{1}{2}\sigma_x^2 f_{xx}(\mu_x,\mu_y) + \frac{1}{2}\sigma_y^2 f_{yy}(\mu_x,\mu_y) + \sigma_{xy}^2 f_{xy}(\mu_x,\mu_y)$$

Hieraus lassen sich unmittelbar die Formeln (3.3)-(3.5) ableiten. Die Näherungsformel (3.6) für das Odds-Ratio ergibt sich mit $X=X_1X_4$, $Y=X_2X_3$ und $f(X,Y)=X/Y$, wenn die überlagerten Werte X_1, X_2, X_3, X_4 als unabhängig angenommen werden. In diesem Fall gilt nämlich:

(A.2) $\mu_{(x_1 \cdot x_2)} = \mu_{x_1} \cdot \mu_{x_2}$

(A.3) $\sigma^2_{(x_1 \cdot x_2)} = \mu^2_{x_1}\sigma^2_{x_2} + \mu^2_{x_2}\sigma^2_{x_1} + \sigma^2_{x_1}\sigma^2_{x_2}$

2. Fehlerverringerung mittels linearer Regression

Wir verwenden für die Darstellung der Methode eine Realisation der Überlagerung der Beispieltabelle 3.1 mit Zufallsvariablen. Tabelle A1 zeigt die abhängige Variable, den Aufbau der Designmatrix und die Schätzergebnisse. Wie man im Vergleich der Chi^2-Werte sieht, kann dadurch insgesamt eine gute Anpassung der überlagerten Tabellenfelder an die Originalwerte erzielt werden. Im Einzelfall (siehe Zeile 1) kann die Schätzung aber auch schlechtere Ergebnisse bringen.

Tabelle A1: Verringerung des Gesamtfehlers mittels linearer Regression (Tabelle 3.1 als Original-Tabelle)

Tabellen-feld	Schätzung	Fallzahlen Original	Überlagert	Design-Matrix X			
n_{11}	30.28	30.00	30.22	1	0	0	0
n_{12}	24.97	25.00	25.49	0	1	0	0
n_{21}	19.85	20.00	18.93	0	0	1	0
n_{22}	39.48	40.00	39.14	0	0	0	1
$n_{1.}$	55.25	55.00	54.43	1	1	0	0
$n_{2.}$	59.33	60.00	59.37	0	0	1	1
$n_{.1}$	50.13	50.00	51.61	1	0	1	0
$n_{.2}$	64.45	65.00	65.35	0	1	0	1
$n_{..}$	114.58	115.00	113.99	1	1	1	1
Chi-Quadrat							
Zeilen 1-4:	0.0105	-	0.0870				
Zeilen 1-9:	0.0256	-	0.1621				

Das Stichprobendesign der Empirisch-Methodischen Arbeitsgruppe (EMMAG): Darstellung und Bewertung

Von Hartmut Götze

Die Empirisch-Methodische Arbeitsgruppe (EMMAG) am Institut für Soziologie und Sozialpolitik in der DDR begann im Jahr 1990 mit dem Aufbau eines eigenen Interviewernetzes zur Durchführung sozialwissenschaftlicher Untersuchungen. Im ersten Teil des folgenden Beitrags wird kurz auf die Situation der Umfrageforschung in der DDR in den Jahren 1989/90 eingegangen. Im zweiten Teil werden die Grundgedanken des entwickelten und eingesetzten Stichprobendesigns dargestellt. Zum Abschluß wird auf einige Untersuchungen eingegangen, um an ausgewählten Ergebnissen den entwickelten Ansatz zu evaluieren.

Mit der Wende im Jahr 1989 in der damaligen DDR ging auch ein grundlegender Wandel in der sozialwissenschaftlichen Forschung einher. Der Beitrag zeigt die Nutzung der damals neu entstandenen Möglichkeiten empirischer Sozialforschung und die damit verbundenen Problemlösungen. Somit versteht sich der Text auch als ein Beitrag zur Geschichte der empirischen Sozialforschung der DDR.

1. Zur Situation der Umfrageforschung in der DDR bis 1989

Eine mit der in der Bundesrepublik qualitativ und quantitativ vergleichbare Sozialforschung existierte in der DDR faktisch nicht. Es gab aber Institute und Forschungsgruppen, die relativ häufig empirische Untersuchungen zu eng begrenzten Themenbereichen durchführten, genannt seien hier die soziologischen Arbeitsgruppen beim Rundfunk und Fernsehen, die sich in erster Linie mit dem Rezeptionsverhalten beschäftigten. Es gab auch Institute, die zu unterschiedlichen Themen vereinzelte Untersuchungen mit nur bedingtem Repräsentanzanspruch [1)] durchführen konnten, wie z.B. das Institut für Soziologie und Sozialpolitik zu Fragen der Entwicklung und Bedingtheit des Kinderwunsches, zu Fragestellungen aus der Arbeitswelt und der Freizeit oder die soziologische Forschungsgruppe an der Akademie der Pädagogischen Wissenschaften zu Problemstellungen im schulischen Bereich. Eine gewisse Sonderstellung nahm das Zentralinstitut für Jugendforschung in Leipzig ein. Der von den dort tätigen Wissenschaftlern bearbeitete Themenbereich, zu dem auch eine große Anzahl empirischer Untersuchungen durchgeführt wurde, war relativ umfangreich. So wurden z.B. Forschungsprojekte bearbeitet, die sich mit der Erfassung und dem Vergleich der Lebensweise, den Einstellungen und Sozialisationsproblemen verschiedener Subpopulationen der Jugend beschäftigten. Für alle hier

genannten Wissenschaftseinrichtungen galten, wenn auch in unterschiedlichem Maße, eine Reihe von Restriktionen, auf die hier zur Situationsbeschreibung kurz eingegangen werden soll.

- Die Entscheidung über Forschungsthemen, die Art der Bearbeitung und auch der Umfang und die Art der Datengewinnung durch Umfrageforschung wurde nicht in den Wissenschaftseinrichtungen entsprechend den Sacherfordernissen getroffen, sondern von politischen Stellen und/oder Institutionen der SED. Allgemeine Umfragen oder Bevölkerungsumfragen waren prinzipiell durch den Ministerrat der DDR zu genehmigen.
- Ein großer Teil der konkreten empirischen Sozialforschung war eigentlich Auftragsforschung des Staates oder der Partei. Das führte dazu, daß die Ergebnisse in Forschungsberichte eingingen, die ausschließlich diesen Stellen zur Verfügung standen. Eine wissenschaftliche Diskussion konnte so nicht stattfinden.
- Was für die Soziologie im allgemeinen galt, traf im besonderen auf die Ergebnisse empirischer Untersuchungen zu: Die Zahl der Publikationen war außerordentlich gering. Auch gab es keine soziologische Zeitschrift. Kam es zu Publikationen, so wurden die Ergebnisse nur als stark verallgemeinerte Aussagen formuliert und aus dem Kontext der Untersuchung herausgelöst, so daß eine Sachdiskussion, ein wissenschaftlicher Meinungsstreit, praktisch unmöglich gemacht wurde.

Die Zeit vom Herbst 1989 bis zum Frühjahr 1990 war die Zeit des schnellen Aufbaus der Umfrageforschung auf dem Gebiet der DDR. Aufgrund der beschriebenen Defizite wurde dieser Aufbau im wesentlichen durch Marktforschungsinstitute aus den alten Bundesländern bestritten. Dazu wurden die wenigen vorhandenen Interviewernetze entsprechend dem Standard des jeweiligen Instituts reorganisiert bzw.- was die Regel war - völlig neue Interviewerstäbe aufgebaut. Das soll zur knappen Beschreibung der Situation in dieser Zeit genügen.

Wissenschaftler des Institutes für Soziologie und Sozialpolitik entschieden sich im Frühjahr 1990 vor dem Hintergrund dieser Entwicklungen für den Aufbau eines eigenen Interviewernetzes. Jede andere Entscheidung hätte den Verzicht auf eigene sozialwissenschaftliche Untersuchungen bedeutet. Diese Aufgabe wurde durch die am Institut tätige Empirisch-Methodische Arbeitsgruppe (EMMAG) realisiert. Grundidee dieser Entscheidung war es, Untersuchungen (eigene und die anderer Forschungseinrichtungen) mit einem Interviewerstab durchzuführen, der durch ein sozialwissenschaftliches Institut aufgebaut, betreut und primär für die Realisierung wissenschaftlicher Aufträge genutzt wird. Berücksichtigt wurde damit auch die Situation in der damaligen DDR, daß die Staatliche Zentralverwaltung

für Statistik bereits im Oktober 1989 in den Verdacht geraten war, Daten gefälscht zu haben.

Nicht ohne Einfluß auf diese Entscheidung war der Fakt, daß sich durch diese enge Verbindung zwischen Forschung und Feldarbeit auch günstige Möglichkeiten für methodische Untersuchungen zu einer Reihe von Sachverhalten ergeben, wie z.B. Aspekte der Fragebogengestaltung, des Interviewerverhaltens, dem Zusammenhang von Themenakzeptanz und Teilnahmebereitschaft. Und nicht zuletzt konnte durch diese Entscheidung ein eigenes Stichprobendesign entwickelt und getestet werden.

2. Darstellung des gewählten Stichprobendesigns

2.1. Allgemeine Probleme

Die Entwicklung eines Designs für repräsentative Bevölkerungsstichproben war zu dieser Zeit mit einer Reihe von Problemen verbunden, die so für die Umfrageforscher auf dem Gebiet der Bundesrepublik nicht bestanden. In erster Linie resultierten diese Schwierigkeiten aus der generellen Anlage der amtlichen Statistik der DDR, die eine Stichprobenentwicklung analog der in der Bundesrepublik weit verbreiteten Anwendung der ADM-Muster-Stichproben-Pläne unmöglich machte. Das betraf vor allem das Fehlen eines statistischen Materials, das dem der Auflistung der Bundestagswahl-Stimmbezirke und den damit verbundenen statistischen Angaben entsprach.

Das zweite Problem, das das Ziehen einer Stichprobe und vor allem die Beurteilung der tatsächlichen Repräsentativität erschwerte, ergab sich aus dem großen Zeitraum, der seit der letzten Volkszählung (1981) vergangen war und den Entwicklungen seit dem Sommer 1989, die die Strukturdaten der Bevölkerung nachhaltig veränderten (Bevölkerungsrückgang durch Migration und damit verbunden z.B. Veränderungen in der Alters-, Geschlechter-, Beschäftigtenstruktur).

2.2. Grundgesamtheit und Basis der Stichprobenauswahl

2.2.1 Abgrenzung der Grundgesamtheit

Als Gesamtheit wird die Personengruppe bezeichnet, aus der die Stichprobe gezogen wird und über die Aussagen getroffen werden sollen. In unserem Fall ist die Gesamtheit die in Privathaushalten lebende Bevölkerung der DDR. Das bedeutet, daß Personengruppen wie in der DDR lebende Ausländer (damals weniger als ein Prozent der Wohnbevölkerung) und Heim- oder Anstaltsbewohner nicht zur Grundgesamtheit gerechnet wurden und damit auch keinen Eingang in die Stichprobe fanden. Private Haushalte wurden als zusammen lebende und zusammen wirtschaftende Personen

definiert, unabhängig von ihrer verwandtschaftlichen Stellung zueinander. Allein Wohnende und allein Wirtschaftende bilden die Ein-Personen-Haushalte.

2.2.2 Zielgruppen der Untersuchungen

Je nach Thematik differierten die Zielgruppen der Untersuchungen, die EMMAG durchführte, nur nach dem Mindestalter der Teilnahme. Die Realisierung dieser Restriktion erfolgt auf der noch zu besprechenden Stufe "Auswahl der Befragungsperson".

2.3. Das Auswahlverfahren[2)]

Unsere Entscheidung für das konkrete Stichprobendesign (s.u.) wurde zum einen durch die schon erwähnten Sachzwänge diktiert und war und ist zum anderen ein Versuch, ein von den ADM-Stichproben abweichendes und für sozialwissenschaftliche Forschungen geeignetes Design zu entwickeln. Obwohl zu dieser Zeit in der DDR noch eine zentrale Einwohnermeldedatei existierte und auch die Möglichkeit bestand, daraus auf ausgewählte Daten für sozialwissenschaftliche Untersuchungen zuzugreifen, entschieden wir uns gegen eine (theoretisch mögliche) reine Zufallsauswahl aus der oben definierten Grundgesamtheit. Die Gründe dafür liegen auf verschiedenen Ebenen. Zum einen wollten wir die zu den Verwaltungseinheiten der DDR vorliegenden Strukturdaten für eine Plausibilitätskontrolle unseres Ansatzes nutzen und zum anderen hätte die Entscheidung für eine reine Zufallsauswahl nicht aufzubringende Kosten in der Phase der Datenerhebung verursacht, d.h. wir hätten ein extrem großes Interviewernetz aufbauen müssen. Das wäre mit folgenden Nachteilen verbunden gewesen:

- überdurchschnittlich hohe Kosten für die Gewinnung und Schulung der Interviewer,
- geringe Auslastung der Interviewer, da die Anzahl der Interviews pro Interviewer und Untersuchung in diesem Fall sehr klein ist,
- sehr hoher Aufwand für die Feldorganisation.

Die andere (theoretische) Möglichkeit hätte darin bestanden, diese reine Zufallsauswahl mit einem "normalen" Interviewernetz zu realisieren. Auch hier sprechen die Nachteile gegen diese Vorgehensweise:

- extrem lange Feldzeiten, die die Zielstellung, mittels Umfrageforschung Zeitpunktdaten zu erheben, ad absurdum führen,
- hohe Gesamtkosten durch hohe Reisekosten.

Der Vorteil des Vorhandenseins eines zentralen Einwohnermeldespeichers wurde deshalb von EMMAG mit einer Vorgehensweise kombiniert, die die oben genannten Nachteile einer reinen Zufallsauswahl minimiert. Die Entscheidung fiel auf ein mehrstufiges Zufallsverfahren, wobei bei der

Auswahl keine Schichtung struktureller Merkmale berücksichtigt wurde. Die verfügbaren Strukturmerkmale wurden nach erfolgter Zufallsauswahl zur Überprüfung der Plausibilität der ersten Auswahlstufe herangezogen.

2.3.1 Die erste Auswahlstufe[3)]

Die Grundlage stellte die räumliche Gliederung der DDR dar. Das kleinste Element, für dessen Beschreibung durch die amtliche Statistik Aussagen über sozio-demografische Merkmale zur Verfügung standen, waren die Kreise, wobei nach Land- und Stadtkreisen (kreisfreie Städte) unterschieden wurde. Auf dieser Stufe ging es darum, mittels einer Zufallsauswahl von Kreisen eine Untergesamtheit der oben beschriebenen Grundgesamtheit zu erzeugen, die hinsichtlich zu bestimmender Merkmale nicht wesentlich von dieser abweicht. Ziel dieser Stufe war es, ein verkleinertes Abbild der damaligen DDR zu schaffen, das mit dieser hinsichtlich verschiedener überprüfbarer struktureller Merkmale übereinstimmte. Abbildung 1 vermittelt einen Eindruck über die territoriale Verteilung der 34 Kreise, die aus den 227 existierenden uneingeschränkt zufällig gezogen wurden.

Die Überprüfung der (auf dieser Ausbaustufe des Netzes) zufällig ausgewählten Kreise mit der Grundgesamtheit wurde anhand folgender Kriterien vorgenommen: Altersstruktur, Geschlechtsstruktur, Qualifikationsstruktur, Urbanisierungsgrad.

Zur Erfassung dieser Kriterien benutzten wir die nachstehenden Merkmale, die in der amtlichen Statistik dokumentiert wurden:

- Altersstruktur:
  - Bevölkerung im Kindesalter (bis 15 Jahre),
  - Bevölkerung im arbeitsfähigem Alter (16-60 bzw.16-65 Jahre),
  - Bevölkerung im Rentenalter (über 60 bzw. 65 Jahre).
- Geschlechtsstruktur:
  - weibliche Bevölkerung,
  - männliche Bevölkerung.
- Berufstätigenstruktur:
  - Berufstätige in Industrie und Bauwesen,
  - Berufstätige in Land- und Forstwirtschaft,
  - Berufstätige in den Bereichen Wissenschaft, Bildung, Kultur, Gesundheits- und Sozialwesen.
- Qualifikationsstruktur:
  - Berufstätige mit Hochschulabschluß,
  - Berufstätige mit Fachschulabschluß,
  - Facharbeiter und Meister.
- Urbanisierungsgrad:
  - Bevölkerung in Gemeinden mit 10000 und mehr Einwohnern.

Abbildung

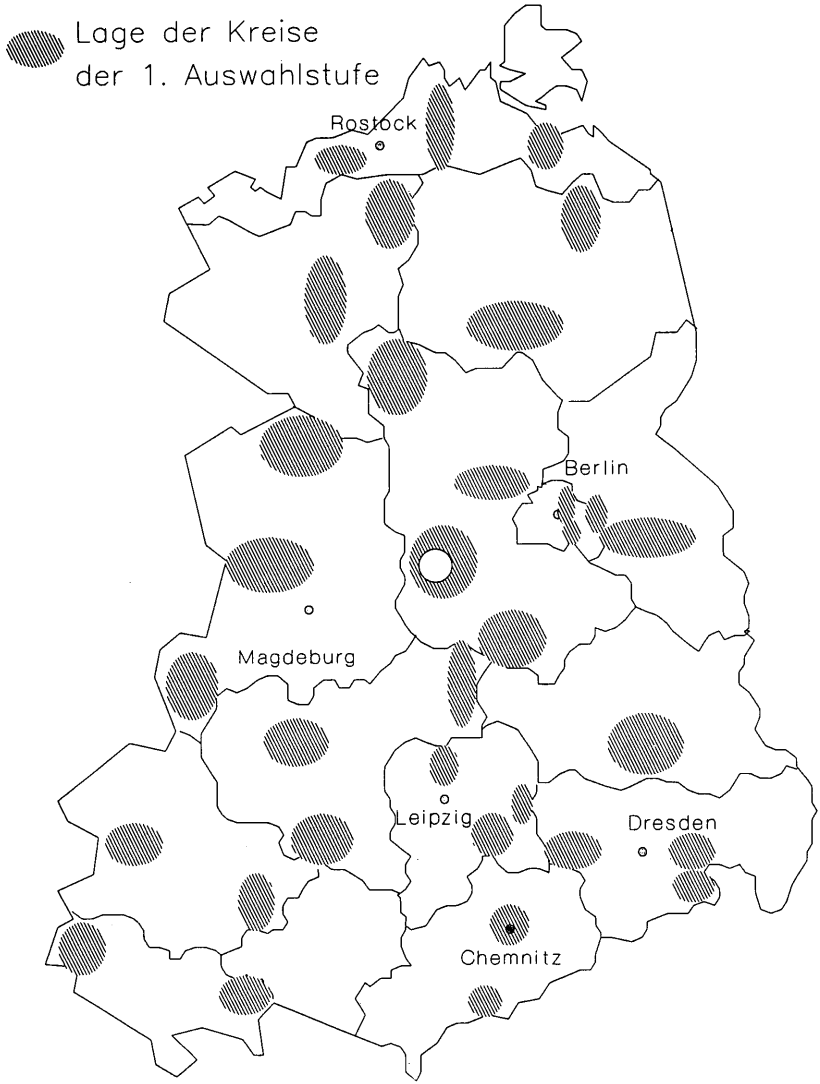

Lage der Kreise
der 1. Auswahlstufe
Rostock
Berlin
Magdeburg
Leipzig
Dresden
Chemnitz

Tabelle 1: Struktur der sozio-demographischen Kriterien

	ausgewählte Kreise	DDR insgesamt
- ALTERSSTRUKTUR		
Anteil an der Bevölkerung insgesamt		
* Kinder	22.2	21.6
* im arbeitsfähigem Alter	62.1	62.3
* Rentner	15.7	16.1
- GESCHLECHTSSTRUKTUR		
* Anteil der weiblichen Bevölkerung an der Gesamtbevölkerung	52.1	52.2
- BERUFSTÄTIGENSTRUKTUR		
Anteil der Wirtschaftsbereiche an den Berufstätigen		
* Industrie/Bau	46.4	46.8
* Land- und Forstwirtschaft	13.2	11.4
* Wissenschaft, Bildung, Kultur, Gesundheits- und Sozialwesen	16.5	16.0
- QUALIFIKATIONSSTRUKTUR		
Anteil der Berufstätigen in volkseigenen Betrieben und Produktionsgenossenschaften insg.		
* mit Hochschulabschluß	7.8	8.2
* mit Fachschulabschluß	14.2	14.1
* Facharbeiter/Meister	64.8	64.8
* Teilfacharbeiter	3.3	3.3
* ohne Berufsabschluß	9.9	9.7
- URBANISIERUNGSGRAD	57.7	66.7

Angaben in Prozent. Quelle: Statistisches Jahrbuch der DDR, 1988 und 1990 und andere Daten der Staatlichen Zentralverwaltung für Statistik.

Von diesen Strukturmerkmalen wurde jeweils die Struktur für die Summe der ausgewählten Kreise berechnet und mit der der DDR insgesamt verglichen. Als zweites Vergleichskriterium benutzten wir den Anteil der oben genannten einzelnen Kennziffern an den jeweiligen Werten für die DDR insgesamt. Die Ergebnisse können den Tabellen 1 und 2 entnommen werden.

Tabelle 2: Anteile verschiedener sozio-demographischer Merkmale der auf der ersten Stufe ausgewählten Kreise an der DDR insgesamt

	ausgewählte Kreise (in 1000)	Anteil an DDR insgesamt (in %)
- BEVÖLKERUNG insgesamt	2588.1	15.5
° im Kindesalter insg.	575.1	15.9
- davon weiblich	280.6	16.0
- davon männlich	294.5	15.9
° im arbeitsfähigen Alter insgesamt	1607.7	15.5
- davon weiblich	767.9	15.4
- davon männlich	839.8	15.5
° im Rentenalter insgesamt	405.3	15.1
- davon weiblich	298.9	15.2
- davon männlich	106.4	15.1
° weiblich insgesamt	1347.3	15.5
° männlich insgesamt	1240.8	15.6
- BERUFSTÄTIGE in		
° Industrie/Bau	746.0	15.3
° Land- und Forstwirtsch.	211.5	17.8
° Wissenschaft, Bildung, Kultur, Gesundheits- und Sozialwesen	264.7	16.0
- BERUFSTÄTIGE		
° mit Hochschulabschluß	126.4	20.6
° mit Fachschulabschluß	241.2	22.3
° Facharbeiter/Meister	1067.3	21.1
- URBANISIERUNGSGRAD		
° Bevölkerung in Gemeinden mit 10000 und mehr Einwohnern	1492.9	13.4

Die Grundaussage der in den Tabellen vorgelegten Daten lautet: Die Summe der ausgewählten Kreise ergibt ein verkleinertes Abbild der damaligen DDR. In den wesentlichen Strukturen (Alter, Geschlecht, Berufstätigkeit) stimmen die Werte der zufällig ausgewählten Kreise genau genug mit denen der Zielgesamtheit überein. Ein Problem gibt es bezüglich des Verhältnisses von Stadt- und Landbevölkerung. Deutlich wird es in den Angaben zum Urbanisierungsgrad und zu den Berufstätigen in der Land- und Forstwirtschaft. Es zeigt sich, daß die Landbevölkerung leicht überrepräsentiert ist. Dieses leichte Manko wurde auf dieser Ausbaustufe der Stichprobe in Kauf genommen und bei der Feldsteuerung im Rahmen konkreter Untersuchungen berücksichtigt. Soviel zur Vorgehensweise auf der ersten Auswahlstufe.

Ergänzend sei hier noch bemerkt, daß es im Verlauf der Arbeit mit diesem Stichprobendesign einige Modifikationen gab, mit denen auf Veränderungen in den Strukturmerkmalen reagiert wurde. Ein zweiter Grund für Veränderungen ergab sich aus der Wiedereinführung des (Bundes-) Länderprinzips. Ziel dieser Veränderungen war und ist es, die Relationen zwischen den Bundesländern (hinsichtlich der genannten Strukturmerkmale und der darüber verfügbaren Daten) auch in der Stichprobe adäquat zu berücksichtigen.[4)]

2.3.2 Die zweite Auswahlstufe

Charakteristisches Merkmal dieser Stufe ist im EMMAG-Design das Arbeiten mit zufällig gezogenen Startadressen. Hierzu wurde, wie oben schon erwähnt, das zentrale Einwohnermelderegister genutzt. Entsprechend der Größe der jeweiligen Stichprobe, die zwischen 1000 und 5000 Personen lag, wurde eine entsprechende Anzahl von Adressen zufällig und proportional zur definierten Bevölkerung aus dem Register gezogen. Ausgehend von diesen Startadressen erfolgt nach konkreter Begehungsvorschrift die Auflistung der Haushalte und die Durchführung der Interviews. Im Verlauf der Arbeit gingen wir von der Vorgehensweise wieder ab, bei der die Startadresse gleichzeitig die erste Befragungsadresse darstellte. Um auf unvorhersehbare Schwierigkeiten im Feld reagieren zu können, liegt die Zahl der pro Kreis gezogenen Startadressen immer etwas über der theoretisch benötigten. Ein Beispiel: Der Kreis Wernigerode ist auf der ersten Auswahlstufe zufällig ausgewählt worden. Aufgrund seiner Einwohnerzahl werden im Rahmen einer Stichprobe mit n=1500 in diesem Kreis 60 Interviews durchgeführt. Davon ausgehend, daß von einer Startadresse aus maximal fünf Befragungen durchgeführt werden, würden also zwölf Startadressen benötigt. Erfahrungen besagen aber, daß dieser Idealfall sehr selten auftritt. (Die dieser realen Situation zugrundeliegenden Ausfall- und Verweigerungsgründe sind temporär und territorial in qualitativer und quantitativer Hinsicht sehr verschieden und können deshalb im Rahmen dieses Beitrages nicht behandelt werden.) Dementsprechend wird für jeden Kreis die Anzahl von Adressen gezogen, die die Realisierung der jeweiligen Stichprobengröße bei der zu erwartenden durchschnittlichen Ausfallquote garantiert. Um auf sich abzeichnende überdurchschnittlich hohe Ausfallquoten in einzelnen Kreisen reagieren zu können, werden für jede Untersuchung einige Reserveadressen gezogen.

2.3.3 Die dritte Auswahlstufe

Diese Stufe ist die, wie auch in anderen Vorgehensweisen übliche, zufällige Auswahl der Zielperson, also der Person in einem Haushalt, die befragt werden soll. Zwischen den dafür zur Verfügung stehenden Möglichkeiten haben wir uns im EMMAG-Design vorerst für das Prinzip "nächster

Geburtstag" entschieden. Es wird also immer die Person im Haushalt befragt, die zur Zielgruppe gehört und als nächste nach dem Befragungstermin Geburtstag hat. Die Entscheidung für dieses System stellt keine absolute Festlegung dar. Bei Bedarf oder Notwendigkeit kann auch ein anderes System eingesetzt werden. Durch den auf dieser Stufe notwendigen Umstieg von Personen auf Haushalte ergibt sich eine unterschiedliche Wahrscheinlichkeit für die Einbeziehung einzelner Personen in die Stichprobe. Diese ist, je nach Definition der Zielgruppe der Untersuchung, abhängig von der Haushaltsgröße oder der Personenanzahl in einem Haushalt, die zur Zielgruppe gehören.

2.4 Strukturmerkmale in Untersuchungen - einige Zahlen

Seit 1990 wurden mit dem vorgestellten Stichprobendesign verschiedene sozialwissenschaftliche Untersuchungen durchgeführt.[5] Im folgenden soll an einigen ausgewählten Beispielen die Leistungsfähigkeit des Stichprobenansatzes und der Feldorganisation dargestellt werden. Die Werte, die sich auf die verschiedenen Umfragen beziehen, sind nicht gewichtet. Ebenso fand keine gezielte Nachrekrutierung von Befragten statt.

1. Ergebnisse aus der Pretest-Erhebung "Zu Erwerb und Verwertung beruflicher Qualifikation in den neuen Bundesländern..." (siehe Anm. 5).

Tabelle 3: Strukturdaten der Pretest-Erhebung "Zu Erwerb..." (in Prozent)

	Stichprobe Berufstätige	Stat. Jahrbuch '89 Gesamtbevölkerung
- GESCHLECHT		
° männlich	50	48
° weiblich	50	52
- GEMEINDEGRÖSSENGRUPPEN		
° bis unter 2.000	15	23
° 2.000 " 5.000	8	11
° 5.000 " 20.000	19	16
° 20.000 " 50.000	26	15
° 50.000 " 100.000	6	8
° 100.000 und mehr	24	27
- BERUFSTÄTIGE		
° Arbeiter und Angestellte	93	89
° Mitglieder von Produktionsgenossenschaften	4	9
° übrige Berufstätige (privat, selbst.)	2	2

Stichprobenumfang: n=1000

Diese Zahlen, speziell die zu den Gemeindegrößengruppen, lassen sich nur bedingt miteinander vergleichen, da sich die Stichprobe aus Berufstätigen zusammensetzt und sich die Angaben der amtlichen Statistik auf die Gesamtbevölkerung beziehen. Es kann jedoch davon ausgegangen werden, daß die Verteilung in der Stichprobe tendenziell der Realität entspricht. In der ehemaligen DDR gab es eine ständige Land-Stadt-Wanderung speziell der Bevölkerungsgruppen im arbeitsfähigen Alter. Damit stellt diese Gruppe in den kleineren Gemeinden einen geringeren Anteil an der Bevölkerung als in den größeren. Die in der Tabelle enthaltenen Aussagen zum Geschlecht und zu den Berufstätigen deuten auf ein hohes Maß an Übereinstimmung zwischen der Stichprobe und den Berufstätigen in der Gesamtbevölkerung.

2. Unmittelbar nach den gesamtdeutschen Bundestagswahlen begann die Feldzeit der Untersuchung "ISSP plus" (siehe Anmerkung 5). Da in dieser Untersuchung auch die Frage nach der gewählten Partei gestellt wurde, bot sich der Vergleich zum realen Wahlergebnis in den neuen Bundesländern an (Tabelle 4).

Tabelle 4: Strukturdaten - Bundestagswahl vom 2.12.90

	Stichprobe (in %)	Bevölkerung der neuen Bundesländer (in %)
• CDU	37.1	41.8
• SPD	25.0	24.3
• FDP	13.7	12.9
• PDS	12.1	11.1
• Bündnis 90/Grüne	10.1	6.0
• DSU	0.8	1.0
• Republikaner	0.3	1.3
• andere	0.9	1.6

Stichprobenumfang: n=1000

Das Wahlergebnis in der Stichprobe stimmt im hohen Maß mit dem Wahlverhalten der Gesamtbevölkerung der neuen Bundesländer überein. Diese Übereinstimmung ist ein besonders überzeugendes Indiz für die Repräsentativität des gewählten Stichprobenansatzes, da diese Wahlergebnisse aktuelle und gesicherte Vergleichsdaten darstellen. Die Stichprobe wies einen Frauenanteil von 49.6% auf, der verglichen mit Angaben der Bevölkerungsstatistik vom 31.12.89 um 3.5% zu gering war.

3. Untersuchung "Wahrnehmung von AIDS im Kontext anderer Gesundheitsrisiken in den neuen Bundesländern" (siehe Anmerkung 5).

Tabelle 5: Strukturdaten - Alter und Familienstand

	Stichprobe (in %)	Bevölkerung der neuen Bundesländer (Daten vom 31.12.89) (in %)
- GESCHLECHT:		
° männlich	40.7	46.9
° weiblich	59.3	53.1
- ALTERSGRUPPEN:		
° 14-17 Jahre	1.8	k.A.
° 18-24 Jahre	10.4	13.2
° 25-34 Jahre	25.9	20.9
° 35-44 Jahre	23.2	16.1
° 45-59 Jahre	23.8	26.1
° über 60 Jahre	14.8	23.7
- FAMILIENSTAND:		
° verheiratet	67.3	63.6
° ledig	17.1	18.6
° geschieden	10.1	7.8
° verwitwet	5.6	10.0

Stichprobenumfang: n=2000

Die Abweichungen lassen sich zum Teil mit der Thematik der Befragung erklären. Die Themenstellung führte zu einer überdurchschnittlich hohen Verweigerung bei den älteren Bürgern. Damit korreliert auch der geringe Anteil der Verwitweten in der Stichprobe. Die Altersgruppe der 18-24jährigen ist in der Stichprobe etwas unterrepräsentiert, da diese Gruppe aus verschiedenen Gründen (Studium, Dienstreise, bestimmte Freizeitaktivitäten) am Wohnort überdurchschnittlich stark nicht erreichbar ist. Die Relation zwischen den Frauen und Männern läßt sich nur zum Teil durch die höhere Verweigerungsrate bei den Männern erklären.

3. Zusammenfassung und Ausblick

Unseres Erachtens kann der Schluß gezogen werden, daß das von EMMAG entwickelte und in der Praxis eingesetzte Stichprobendesign seine Leistungsfähigkeit bewiesen hat. Abweichungen verschiedener Strukturdaten in den durchgeführten Untersuchungen von denen in der Grundgesamtheit lassen sich erklären und wurden in Abhängigkeit vom Untersuchungsziel

durch Gewichtungsprozesse minimiert. Für die Zukunft gilt, daß dieser Ansatz weiterentwickelt und eingesetzt wird. Dafür sprechen meines Erachtens mehrere Gründe. Der erste ergibt sich aus der Organisation der Auswahlstufen im vorgestellten Design. Mit der Verbesserung der Strukturdatenbasis für die neuen Bundesländer ist bei Verwendung des vorgestellten Designs eine mehrfache Überprüfung der Repräsentanz möglich.[6] Auch für die Feldorganisation bietet der gewählte Ansatz Vorteile. So kann die Identität von Befragungsgebieten mit territorialen Verwaltungseinheiten für eine gezielte Information über geplante Studien genutzt werden, um so die Ausschöpfung zu erhöhen. Dazu gibt es erste positive Erfahrungen. Perspektivisch soll die Anzahl der Kreise noch erhöht werden, um so eine noch größere Streuung innerhalb der Bundesländer zu erreichen. Tiefergehende Analysen der Repräsentativität und die Erhöhung der Ausschöpfung bei Untersuchungen - auf diese beiden Aufgaben wird sich die Arbeit am EMMAG-Stichprobendesign in der Zukunft konzentrieren. Eine Reihe von Problemen, die in der Arbeit mit dem ADM-Design auftreten, werden auch durch diesen anderen Ansatz nicht gelöst. Deshalb sollte meines Erachtens von allen, die sich aus theoretischen oder praktischen Gründen mit der Stichprobenproblematik beschäftigen, nach Wegen (auch unkonventionellen) gesucht werden, um das Abbild der Realität durch die Stichprobe zu verbessern. Nach wie vor werden in der Umfrageforschung Bevölkerungsgruppen ausgeschlossen, per Definition und aufgrund einer erschwerten Erreichbarkeit oder Ansprechbarkeit. Die Umfrageforschung hat die Aufgabe, das zu ändern.

Anmerkungen

1) Die nur bedingte Repräsentanz zahlreicher Untersuchungen hat verschiedene Gründe, vor allem politische und finanzielle. Da auch der überwiegenden Mehrzahl der Sozialwissenschaftler keine professionellen Interviewer zur Verfügung standen, wurde die Datenerhebung in vielen Fällen durch die Wissenschaftler selbst durchgeführt. Das führte dazu, daß die Fallzahlen oftmals zu klein für repräsentative Aussagen waren. In anderen Untersuchungen mußten Abstriche am Prinzip der Zufallsauswahl der Probanden gemacht werden, um überhaupt Untersuchungen durchführen zu können. Das bedeutet, daß von Stichprobenplänen im eigentlichen Sinn nicht nicht gesprochen werden kann.

2) Eine Einführung in die Stichprobentheorie und die Möglichkeiten der ADM-Muster-Stichproben bietet das 1979 vom Arbeitskreis Deutscher Marktforschungsinstitute herausgegebene und von F. Schaefer bearbeitete Buch: Muster-Stichproben-Pläne für Bevölkerungs-Stichproben in der Bundesrepublik Deutschland und West-Berlin.

3) Eine detailliertere Beschreibung der Vorgehensweise findet sich in einem unveröffentlichten Forschungsbericht von Sabine Nowossadeck und Enno Nowossadeck, der eine Grundlage für das EMMAG-Stichprobendesign und den Aufbau des Interviewernetzes darstellte. Dieser Bericht "Repräsentativitätsuntersuchung als Grundlage für den Aufbau eines Interviewernetzes" ist bei der Abteilung Methodenentwicklung von ZUMA einsehbar. Diese Abteilung setzt sich aus Mitarbeitern zusammen, die der Empirisch-Methodischen Arbeitsgruppe angehörten: Dr.sc. Michael Häder (Leiter der Gruppe), Dipl.-Soz. Hartmut Götze, Dr. Bernhardt Krüger, Dr. Sabine Nowossadeck.

4) Im Rahmen dieses Beitrages kann nicht auf alle Veränderungen in der Arbeit mit dem vorgestellten Design eingegangen werden. Die Empirisch-Methodische Arbeitsgruppe arbeitet am Sozialwissenschaftlichen Forschungszentrum Berlin-Brandenburg (SFZ) weiter und steht für Auskünfte zur Verfügung. Ansprechpartner sind: Dipl.-Psych. Rainer Schubert, Dr. Jochen Brandt, Dipl.- Phil. Dagmar Schreiber. Anschrift: EMMAG, O-1086 Berlin, Jägerstr. 10/11.
5) Einige sollen hier genannt werden:
   - Pretest-Erhebung zum Thema "Zu Erwerb und Verwertung beruflicher Qualifikation in den neuen Bundesländern", Feldzeit: September/Oktober 1990, Stichprobenumfang: n=1000;
   - "Leben '91"- sozialwissenschaftliche Untersuchung in den neuen Bundesländern, Februar 1991, n=1500;
   - "ISSP plus"- als Nacherhebung in den neuen Bundesländern, Dezember 1990, n=1000;
   - Untersuchung zum Thema "Wahrnehmung von AIDS im Kontext anderer Gesundheitsrisiken in den neuen Bundesländern", Juni 1991, n=2000.
6) So werden z.B. von infas verschiedene Studien mit einer Vielzahl von Kreisstrukturdaten und detaillierten Aussagen zu Bevölkerungsprognosen angeboten. Darin enthalten sind Kennziffern die in dieser Art für die Grundbausteine des ADM-Stichprobendesigns, die Stimmbezirke oder die Gemeinden, nicht zur Verfügung stehen. Dadurch ergibt sich die Möglichkeit, sowohl die Repräsentativität der ausgewählten Kreise für das gesamte Territorium der neuen Bundesländer als auch die Repräsentativität von Untersuchungsergebnissen mittels sozialwissenschaftlich relevanter Daten zu überprüfen.

Erfahrungen und Problemlösungen beim Datenaustausch zwischen Statistikprogrammsystemen

Von Heiner Ritter und Cornelia Züll

Der Datenaustausch zwischen verschiedenen Softwaresystemen ist ein immer wieder auftretendes Problem. Wir wollen uns im nachfolgenden Artikel mit dem Austausch von Daten zwischen Statistikprogrammen befassen. Wichtig ist dabei für uns, wie - neben den reinen Datenwerten - Informationen wie Labels für Variablen und Werte, Angaben über fehlende Werte und die fehlenden Werte selbst ausgetauscht werden können, ohne daß eine Neueingabe im anderen System gemacht werden muß. Seit einiger Zeit ist mit DBMS/COPY (und DBMS/COPY Plus) ein Produkt auf dem Markt, das alle diese Probleme lösen bzw. vereinfachen soll. Unsere Erfahrungen mit diesem Programm werden aufgezeigt und Empfehlungen für den Datenaustausch gegeben.

1. Einleitung

Datenaustausch ist ein sehr weit gestecktes Gebiet. Mit dem Begriff stellt sich sofort die Frage, zwischen wem oder was Daten ausgetauscht werden sollen und was unter Daten zu verstehen ist. Datenaustausch heißt erst einmal nicht mehr, als daß Zahlen, Buchstaben (manchmal nur die Großbuchstaben, vielfach aber auch die Kleinbuchstaben) und einige Sonderzeichen, die in Dateien gespeichert sind, ausgetauscht werden können. Damit ist explizit nicht der Austausch von Maschinencode, also von Programmen oder ähnlichem gemeint.

Datenaustauschprobleme gibt es z.B. mit internen Formaten von Textverarbeitungssystemen und Grafikprogrammen, zwischen Statistikprogrammen und Datenbanksystemen. Wir wollen uns im folgenden Beitrag auf den Datenaustausch zwischen Statistikprogrammen untereinander und zwischen Statistikprogrammen und Datenbanksystemen (wie z.B. dBASE IV) unter MS DOS beschränken.

Ohne weiteres ist ein Austausch von ASCII-Dateien möglich. Das sind solche Dateien, die Zahlen, Buchstaben und einige definierte Sonderzeichen umfassen. Somit ist ein rudimentärer Datenaustausch möglich. Allerdings sind in jedem Programm von neuem alle Variablen- und Fallzuordnungen zu definieren, was eine Zahl/Zahlengruppe oder ein Buchstabe/Wert usw. zu

bedeuten hat. Weitere programminterne Definitionen (z.B. Missing-Data-Angaben) werden nicht übernommen und müssen somit neu definiert werden.

Ein Austausch der Dateien mit internen Formaten (Systemdateien) ist in der Regel nur sehr beschränkt möglich (z.B. können dBASE-Dateien in SPSS eingelesen werden). Unter Systemdateien werden in diesem Zusammenhang Dateien verstanden, die vom jeweiligen Programm erstellt wurden und alle vorher vom Benutzer definierten Datenbeschreibungen und Daten enthalten. Diese Informationen werden von jedem Programm in einer anderen Form, häufig auch binär gespeichert, so daß sie nicht direkt von Programm zu Programm übergeben werden können.

2. Austausch von statistischen Daten

Im folgenden wollen wir ausführlicher den Datenaustausch zwischen Statistikprogrammen und Datenbanken betrachten.

Daß es überhaupt solche Umsetzprogramme gibt ist von zweierlei Interesse: zum einen haben die Hersteller z.B. von Statistikprogrammen erkannt, daß Datenanalyse nicht nur in einem einzigen Programm durchgeführt wird, sondern daß es notwendig ist, Daten aus einer Datenbank, wie z.B. dBASE, in ein Statistikprogramm einlesen zu können; auch die Abspeicherung von Daten in einer Art Transportformat für den Austausch zwischen verschiedenen Rechnertypen aber innerhalb eines Programmsystems, z.B. das SAS-Transportprotokoll oder die SPSS-Exportdatei, ist von Bedeutung.

Somit ist zunächst bei jedem Statistikprogramm - und auf diese wollen wir uns im folgenden konzentrieren - zu prüfen, aus welchem Fremdsystem Daten gelesen und auf welches Daten geschrieben werden können. In der Regel beschränkt sich dies auf Tabellenkalkulationsprogramme (z.B. Lotus 1-2-3) oder Datenbanken (z.B. dBASE).

3. Was ist beim Austausch von Statistikdaten zu beachten?

Um die Probleme zu verdeutlichen, soll der Datenaustausch zwischen zwei Statistikprogrammen auf die bisher konventionelle Art, vermittelt über eine ASCII-Datei, beschrieben werden.

Im Statistikprogramm A sind die Daten in Form von Variablen und Beobachtungen/Fälle abgelegt. Die Variablen sind in einem bestimmten

Datentyp gespeichert, wie z.B. Integer, Floating Point, Double Precision. Diese Form der Speicherung hat in der Regel Einfluß auf deren weitere Behandlungsmethode bei Berechungen durch das Programm. Sodann ist auch das genaue Format einer Variablen gespeichert, z.B. I4, I1 oder F10.3, F3.0. Für die Variablennamen, z.B. V10, ist häufig ein selbsterklärendes Variablenlabel wie "BILDUNG" oder "GESCHLECHT" definiert. Den einzelnen Variablenkategorien bzw. Variablenwerten, wie z.B. "1" oder "2" bei "GESCHLECHT", ist oft ebenfalls ein selbsterklärender Name, das Wertelabel, beigegeben, z.B. 1 = "weiblich", 2 = "maennlich". Des weiteren ist bei einem Fall auch kenntlich zu machen, ob z.B. eine Person bei einer Befragung die Antwort verweigert hat und daher diese in der entsprechenden Variablen als fehlender Wert (Missing Data) nicht in die Analyse eingehen soll.

Falls nun eine ASCII-Ausgabe aus dem Statistikprogramm A erfolgt, werden Zahlen in einem bestimmten Format in einer bestimmten Reihenfolge (im festen oder freien Format) in eine Ausgabedatei geschrieben. Hierdurch gehen diverse Informationen aus der Ursprungsdatei verloren. Diese sind:

- Position und Breite der Variablen (Format),
- der Variablentyp,
- der Variablenname und das Variablenlabel,
- das Wertelabel,
- die Definition der Werte, die als fehlend zu behandeln sind.

Um diese Zahlen in das Statistiksystem B wieder einzulesen, müssen alle diese Angaben für das Statistiksystem B neu definiert werden. Ein Programm, das in irgendeiner Form ohne den Verlust all dieser Informationen bei einem solchen Datenaustausch helfen kann ist somit von großem Nutzen, da es Zeit und Arbeit spart.

Wie wichtig und wirkungsvoll ein solches Programm ist, hängt allerdings davon ab, welche Art von "Problemen" bei der Umsetzung bekannt sein müssen bzw. welche Art von Umsetzungsfolgewirkungen man akzeptieren muß.

4. Der Praxistest mit dem Programm DBMS/COPY Plus

Im folgenden Praxistest haben wir das Programm DBMS/COPY Plus herangezogen. Zuerst wird der theoretische Leistungsumfang des Programms beschrieben und dann an Umsetzungsbeispielen mit einer kleinen Testdatei sowie mit einer großen Datendatei der Leistungsumfang beurteilt.

Tabelle 1: Statistikprogramme, deren Dateien DBMS/COPY umsetzen kann

ABstat	Glim	SPSS/PC+
Across	MicroStat-II	Stata
Bass	Minitab	STATGRAPHICS
BMDP	NCSS	StatPac Gold
CSS	PRODAS	StatPlan III
EGRET	SAS/PC	Stats+
EpiInfo	SCA	SYSTAT
Gauss	SOLO	YStat

Das Programm DBMS/COPY und seine Erweiterung DBMS/COPY Plus ("The Tool for Software Connectivity") ist von der US-Firma Conceptual Software entwickelt worden und wird mittlerweile von verschiedensten Herstellern von Statistiksoftware als Zusatztool erwähnt und meist auch vertrieben. Das Programm kann zwischen 72 Programmen bzw. Programm-Versionen Daten und die dazugehörenden Datenbeschreibungen umsetzen. Tabelle 1 zeigt alle auswählbaren Statistikprogramme, deren Dateien umgesetzt werden können. Daneben können auch alle gängigen Formate der häufig eingesetzten Datenbanksysteme und der Tabellenkalkulations-Programme umgesetzt werden. DBMS/COPY erlaubt zudem eine ASCII-Ausgabe der Dateien in festem oder freiem Format.

In der "Plus"-Version besteht zusätzlich zur reinen Umsetzung die Möglichkeit, Datenmanipulationen vorzunehmen, wie z.B.:

- eine Variablenauswahl,
- eine Fallauswahl,
- die Kreierung neuer Variablen (auch unter Einsatz einer Funktionen-Bibliothek),
- eine Variablen-Formatdefinition.

Hierzu werden Kommandos zur Verfügung gestellt, mit deren Hilfe verschachtelte und differenzierende Abfragen programmiert werden können. Das Programm ist im Stapelbetrieb (Batch) und im interaktiven Modus (Fenstersystem) einzusetzen.

Der Umsetzvorgang geht bei DBMS/COPY Fall für Fall vor sich. D.h. aus der Ausgangsdatei, z.B. einer SPSS/PC+ Systemdatei, wird ein Fall eingelesen und alle Variablen in die entsprechenden Felder im Speicher zwischengespeichert. Von dort wird der Fall dann in das Format des Zielstatistikprogramms umgesetzt. Soweit bei der Umsetzung Informationen

mitgeteilt werden müssen, was z.B. mit programminternen Variablen geschieht, wird eine "LOG"-Datei erstellt, die nach der Umsetzung gedruckt werden kann.

Das Programm handhabt die folgenden Problemfelder:
- Variablentypen,
- Variablenformate,
- fehlende Werte,
- Variablennamen.

Es werden also keine Wertelabels umgesetzt.

4.1 Die Testdaten

Im folgenden werden mit Hilfe der Testdaten die Umsetzerfolge und -probleme beschrieben, um gut handhabbare Umsetztips vermitteln zu können.

Die erste Testdatei ist eine kleine Datei (54 Fälle) mit Variablen verschiedensten Typs:

- eine Integer-Variable, 5-stellig (I5), mit einem Variablenlabel (Länge: 49 Zeichen);
- eine Integer-Variable, 1-stellig (I1) mit Wertelabels;
- eine Integer-Variable, 1-stellig (I1) mit dem Variablenlabel; die Variable enthält 2 System-Missing-Data Fälle ("."), sowie einen Fall (Wert 8) als User-Missing-Data;
- eine Character-Variable, 8-stellig (A8) mit 27 Fällen, die nur Leerzeichen enthalten und als User-Missing-Data (Blank) definiert sind;
- eine floating point Variable, double precision, mit einem Variablenlabel (Länge: 42 Zeichen).

Die zweite Testdatei besteht aus circa 500 Variablen und circa 3000 Fällen (Allgemeine Bevölkerungsumfrage der Sozialwissenschaften, ALLBUS 1990).

4.2 Informationen zu den Statistiksystemen

Wir haben uns bei unseren Tests auf die folgenden Statistiksysteme beschränkt: SPSS/PC+ (Version 4.0), SAS/PC (Version 6.03), SYSTAT (Version 5.0), STATGRAPHICS (Version 5.0), Gauss (Version 386 VM 2.1), CSS (Version 3.1) und das Datenbanksystem dBASE IV. In Tabelle 2 werden die Möglichkeiten dieser Systeme in bezug auf Variablen- und Fallzahlen, fehlende Werte, Labels etc., die jeweils maximal verarbeitet werden können,

gegenübergestellt. Diese Übersicht legt schon unterschiedliche Fragestellungen nahe: was passiert, wenn die Labels zu lang sind oder was passiert mit fehlenden Werten, wenn die Systeme unterschiedliche Arten unterstützen? Im folgenden versuchen wir, solche Fragen zu beantworten.

Tabelle 2: Datenbeschreibungsmöglichkeiten bei verschiedenen Systemen

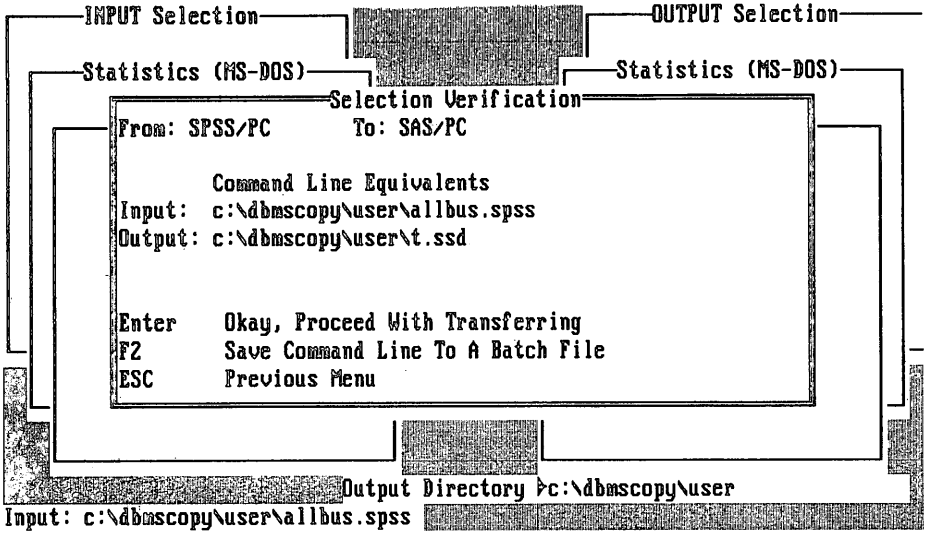
	SPSS/PC+	SAS/PC	SYSTAT	Gauss	CSS	STATGRAPHICS
Variablenzahl	500	°	256	8190	300	°
Fallzahl	°	°	°	°	°	°
Länge Variablenname	8	8	8	8	8	10
Variablen Labels	ja	ja	nein	nein	ja	nein
Länge der Var.Labels	60	40	-	-	40	-
Wertelabels	ja	ja	ja	nein	ja	ja
Länge der Wertelab.	60	40	-	-	40	-
Character-Variablen	ja	ja	ja	ja	ja	ja
Länge Character Vars	255	200	12	8	40	70
System-Missing-Data	ja	ja	ja	ja	ja	ja
User-Missing-Data	ja	ja	nein	nein	nein	nein
Zahl der User-Miss.	1	<= 27	-	-	-	-

° hardwareabhängig und/oder Prozedurabhängig

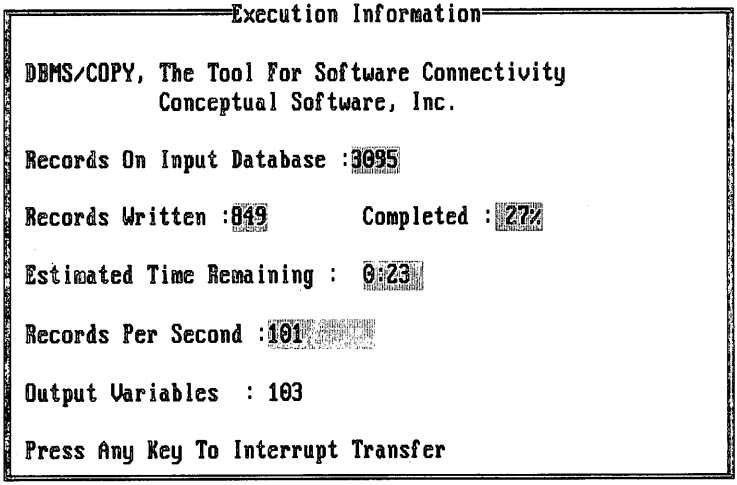
4.3 Beschreibung eines Umsetzvorgangs

Dem Programm DBMS/COPY ist die Ausgangs- und Zieldatei bekannt zu machen. Mit Hilfe des Menüsystems kann das gewünschte Programmsystem zusammen mit der Angabe der Pfade für die Speicherorte der Ausgangs- und Zieldatei gewählt werden. Für die Umsetzung selbst sind die Statistikprogrammsysteme nicht erforderlich (Bild 1). Während des laufenden Umsetzvorgangs werden einige statistische Daten angezeigt (Bild 2).

Abbild. 1: Informationsübersicht vor dem Starten des Umsetzvorgangs mit dem Programm DBMS/COPY



Abbild. 2: Statistische Informationen zum Umsetzungsvorgang



4.4 Ergebnisse der Umsetzungen zwischen den Statistikprogrammen

Mit DBMS/COPY wurden die Testdatensätze für die genannten Systeme umgesetzt und die Ergebnisse jeweils im Zielsystem überprüft. Das Hauptergebnis dieser Tests läßt sich unabhängig vom Ausgangssystem und dem Zielsystem zusammenfassen. Um es vorwegzunehmen, alle Datenwerte werden immer richtig übernommen. Bei keinem der untersuchten Systeme traten in diesem Bereich Probleme auf. Die Umsetzung geht in der Regel je nach eingesetzter Hardware sehr schnell: selbst für die große Testdatei werden nur circa zweieinhalb Minuten für eine Umsetzung benötigt, eine Ausnahme bildete die Umsetzung in STATGRAPHICS mit etwa acht Minuten. Als PC stand uns ein IBM kompatibler PC mit einem 80486 Prozessor/33 MHz zur Verfügung.

DBMS/COPY setzt Variablennamen, die nicht den Konventionen des Zielsystems entsprechen, automatisch um. Wenn also eine Datei von SPSS in Gauss umgesetzt wird, werden alle Variablennamen für numerische Variablen in Großbuchstaben umgestellt, die von Character-Variablen in Kleinbuchstaben. Genauso werden z.B. die Namen der Systemvariablen $DATE, $CASENUM und $WEIGHT in SPSS/PC+ geändert, da die meisten anderen Systeme Variablennamen mit einem $ im Variablennamen nicht erlauben (z.B. in _date für SAS). Diese Änderungen werden am Bildschirm angezeigt und in einer LOG-Datei zur Dokumentation gespeichert.

Probleme kann es bei der Umsetzung mit den Dateigrößen der neu erstellten Datei geben. DBMS/COPY liest die Variablen aus der Ausgangsdatei und speichert sie in einer internen Datei. Dabei werden z.B. alle numerischen Variablen aus einer SPSS/PC+ Datei automatisch in einem double precision Feld gespeichert, d.h. jede Variable benötigt acht Bytes (Positionen), auch wenn sie nur die Werte 0 oder 1 annehmen kann. Bei unserer großen Testdatei führt dies dazu, daß aus der circa zwei MB umfassenden SPSS/PC+ Systemdatei eine circa 12 MB große SAS/PC "Datenbank"-Datei wird. Im DBMS/COPY-Handbuch ist jeweils dokumentiert, wie die Variablen der Ausgangsdatei intern gespeichert werden. Falls nicht ausreichend Speicherplatz auf der Festplatte des PCs vorhanden ist, bricht während der Umsetzung das Programm zwangsweise ab. Leider prüft das Programm vor der Umsetzung nicht, ob der erforderliche Speicherplatz für die Zieldatei auf der Festplatte verfügbar ist.

Variablenlabels werden in der Regel übernommen. Sind die Labels des Ausgangssystems zu lang für das Zielsystem, werden sie von rechts her abgeschnitten. Wertelabels werden, wie schon oben erwähnt, grundsätzlich nie übernommen, auch wenn sowohl das Ausgangs- wie auch das Zielsystem Wertelabels erlauben.

Tabelle 3: Häufigkeitsauszählung einer Beispielvariablen ("Bildung") mit SPSS/PC+ vor der Umsetzung in SAS/PC mit DBMS/COPY

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	1	34	63.0	66.7	66.7
	2	12	22.2	23.6	90.2
	3	5	9.3	9.8	100.0
		2	3.7	Missing	
	8	1	1.9	Missing	
				-------	------
	Total	54	100.0	100.0	

Valid cases 51 Missing cases 3

Tabelle 4: Häufigkeitsauszählung der Beispielvariablen ("Bildung") mit SAS/PC nach der Umsetzung mit DBMS/COPY aus SPSS/PC+

BILDUNG	Frequency	Percent	Cumulative Frequency	Cumulative Percent
.	3	.	.	.
1	34	66.7	34	66.7
2	12	23.5	46	90.2
3	5	9.8	51	100.0

Frequency Missing = 3

DBMS/COPY kennt nur eine Art von fehlendem Wert, den System-Missing-Data Wert. Sind in einer Datei Werte als fehlend deklariert, werden sie zwangsläufig in System-Missing-Data umgesetzt und sind von den anderen System-Missing-Data nicht mehr zu unterscheiden. In unserer Beispieldatei waren bei einer Variablen zwei Arten von fehlenden Werten vorhanden: zwei Fälle hatten System-Missing-Data, ein Fall hatte den Wert 8, der als fehlender Wert deklariert war. Nach einer Umsetzung mit DBMS/COPY in welches System auch immer, wird der Fall mit dem Wert 8 in den System-Missing-Data umgesetzt und der Wert 8 verschwindet. Die drei Fälle sind nun nicht mehr zu unterscheiden. Die Tabellen 3 und 4 verdeutlichen für SPSS/PC+ bzw. SAS/PC das Ergebnis einer Häufigkeitsauszählung dieser Beispielvariable.

Leider kennt DBMS/COPY die Begrenzungen der Zielsystem bezüglich der Variablenzahl nicht. So wird z.B. ohne weiteres und vor allem ohne Warnung eine SPSS/PC+ Datei oder eine CSS-Datei mit mehr als 500 Variablen erstellt, die dann nicht mehr ohne Zusatzangaben in SPSS/PC+ bzw. CSS eingelesen werden kann.

4.5 Spezielle Probleme bei einzelnen Systemen

Zusätzlich zu den oben beschriebenen Problemen, die praktisch alle untersuchten Systeme betreffen, treten einige spezielle Verfahrensweisen und/oder Probleme bei der Umsetzung zwischen den einzelnen Systemen auf, die vom Ausgangs- oder Zielstatistiksystem abhängig sind.

Eine SPSS/PC+ Datei enthält immer die drei Systemvariablen - Anzahl der Fälle ($CASENUM), das Datum ($DATE) sowie eine Gewichtungsvariable ($WEIGHT). Diese Variablen werden immer mit in die Zieldatei übernommen.

Wie schon dem DBMS/COPY Manual zu entnehmen ist, kann die SAS/PC-Zieldatei nur mit einem einzigen Buchstaben benannt werden. Dies soll an einem "kryptischen" Feld in der SAS-Systemdatei liegen. Die Definition des fehlenden Wertes bei Charactervariablen geht bei der Umwandlung von SAS/PC in SPSS/PC+ verloren; es wird ein normaler, gültiger Character-Wert daraus. Im Handbuch von DBMS/COPY steht, daß CSS keine Variablenlabels kennen würde, was nicht den Tatsachen entspricht. Dadurch gehen bei der Umsetzung in eine CSS-Systemdatei alle Variablenlabels verloren.

Character-Variablen werden ebenfalls von DBMS/COPY nicht in eine CSS-Datei übernommen. Dies wird während des Umsetzprozesses als Warnung angezeigt. Die Character-Variable führt auch bei der Umsetzung von einer SPSS/PC+ Datei oder einer SAS/PC-Datei in Gauss zu Problemen: die Variable wird zwar übernommen, erscheint auch in der Variablenliste bei Gauss, aber der Versuch, für diese Variable eine Häufigkeitsverteilung zu erstellen, führt zu einem Programmfehler. Von der großen Testdatei war keine Umsetzung in eine STATGRAPHICS-Datei möglich. Zwar arbeitete DBMS/COPY bis zur Angabe, daß 100% der Daten umgesetzt seien, das Programm blieb jedoch zu diesem Zeitpunkt hängen und ließ sich nur durch einen Warmstart des PCs beenden, was zu einer Zerstörung der umgesetzten STATGRAPHICS-Datei führte.

4.6 Erzeugen von ASCII-Dateien

DBMS/COPY erlaubt das Erstellen einer ASCII-Datei. In den meisten Statistikprogrammen oder Datenbanken steht diese Option direkt zur

Verfügung. Eine sinnvolle Anwendung der Umsetzung in ASCII könnte sich nur dann ergeben, wenn ein Programm die Werte nur tabellarisch, d.h. in einem festen Format ausgeben kann, das Zielstatistiksystem die Daten aber in freiem Format einlesen möchte (d.h. die einzelnen Werte für alle Variablen fortlaufend, durch Komma getrennt geschrieben). In DBMS/COPY kann zwischen freiem und festem Format gewählt werden.

Für fehlende Werte kann ein Zeichen definiert werden, das für jede fehlende Angabe ausgegeben werden soll. System-Missing-Data-Werte und User-Missing-Data-Werte werden dabei gleich behandelt und alle auf dieses eine Zeichen gesetzt. Bei freiem Format können die Werte mit Komma oder einem anderen Sonderzeichen getrennt ausgegeben werden. Zusätzlich werden die Variablennamen in der ersten Datenzeile gespeichert.

Bei der Ausgabe in einem festen Format werden alle Werte, wenn Sie aus einem double precision Feld in DBMS/COPY kommen (z.B. bei SPSS/PC+) 12-stellig mit zwei Dezimalstellen gespeichert. Zusätzlich zur ASCII-Datei im festen Format wird eine zweite Datei angelegt, die eine Beschreibung zur ASCII-Datei enthält (z.B. Variablennamen, Spaltenpositionen). Ein Beispiel für solch eine Datei zeigt Tabelle 5. Sie kann als Dokumentation zur Datendatei verwendet werden, bietet aber keine direkte, automatische Hilfe beim Wiedereinlesen der Daten.

Tabelle 5: Dateibeschreibung von DBMS/COPY für ASCII-Dateien

```
dictionary
extension=dat
missing=.
numeric=n
fixed=y
dictionary=dct
date=mm/dd/yyyy

     13     8     c       $DATE
     21    12     R       $WEIGHT
     33    12     R       ID1
     45    12     R       V127
     57    12     R       GESCHL
     69    12     R       ALTER
     81    12     R       BILDUNG
     93     7     c       CHARV
    100    12     R       DVAR
endvars
```

Will man ASCII-Dateien in festen Format ausgegeben, sollte man auf die von den Statistikprogrammen selbst zur Verfügung gestellten Prozeduren zugreifen, um alle Variablen in der tatsächlich nötigen Breite gespeichert zu bekommen und um alle fehlenden Werte zu erhalten. Benötigt man jedoch Daten im freien Format, ist DBMS/COPY eine hilfreiche Alternative.

4.7 Empfehlungen zum Arbeiten mit DBMS/COPY

Der nach unseren Erfahrungen beste Weg, Dateien umzusetzen, ist zunächst die Option, die die Programme selbst bieten. So kann SPSS/PC+ verschiedene Datenbank- und Spreadsheetformate sowohl einlesen als auch erstellen. CSS kann neben dBASE-Dateien auch SPSS-Export-Dateien lesen und erstellen. In beiden Fällen entfallen einige der oben beschriebenen Probleme, da die echten Breiten der Variablen berücksichtigt und die Ausgabedateien nicht unnötig aufgebläht werden. An CSS werden hierbei sowohl die Variablen- als auch die Wertelabels von SPSS/PC+ korrekt übergeben. Da aber nicht alle Systeme die Möglichkeit bieten, Dateien anderer Systeme einzulesen, ist ein Programm wie DBMS/COPY eine gute Alternative.

Um sinnvoll mit DBMS/COPY zu arbeiten, sollte man sich vorher unbedingt über die Restriktionen des Zielsystems im Klaren sein.

Mit dem Zusatzmodul DBMS/COPY Plus kann eine Variablenauswahl getroffen werden, wenn zuviele Variablen in der Ausgangsdatei vorhanden sind. Wir hatten allerdings mit DBMS/COPY Plus einige Probleme, da das System zwar laut Dokumentation Bereichsangaben bei Variablenlisten erlaubt (z.B. V1-V70), aber in der Praxis doch alle Variablen, die übernommen werden sollten, einzeln aufgeführt werden mußten.

Bei großen Dateien sollte vorher an Hand des DBMS/COPY-Handbuches überprüft werden, ob die Variablen in einem "double precision" Feld gespeichert werden und bei ein- oder zwei-stelligen Variablen Vorkehrungen getroffen werden, damit die Zieldatei nicht zu groß wird. Mit DBMS/COPY Plus lassen sich für die Umsetzung pro Variable ein gegenüber dem Standard reduzierter Byte-Bereich angeben; hierzu ist allerdings ein größerer Arbeitsaufwand erforderlich.

Eventuell zu lange Variablenlabels sollten im Ursprungssystem sinnvoll gekürzt werden, damit sie nach der Umsetzung brauchbar sind.

Soll die Trennung zwischen User-Missing-Data und System-Missing-Data beibehalten werden, muß vorher im Ausgangssystem die Deklaration des

Datenwerts als fehlend aufgehoben werden und später im Zielsystem neu gesetzt werden.

Wertelabels müssen auf jeden Fall im Zielsystem - soweit vorhanden - neu eingegeben werden.

Am Beispiel der Umsetzung der großen Testdatei ALLBUS 1990 aus dem Statistikprogramm SPSS/PC+ in das Statistikprogramm SAS/PC sollen diese Vorüberlegungen demonstriert werden. Bezüglich der einzelnen Programmvorgaben verweisen wir auf die Tabelle 2. Vor der Umsetzung mit DBMS/COPY sind die im folgenden aufgeführten Überlegungen/Entscheidungen anzustellen bzw. zu treffen:

- Da SAS/PC keine programminterne Begrenzungen für die Variablen und Fallanzahl hat, ist hier nicht mit Problemen zu rechnen.
- Da beide Programmsysteme Charaktervariablen kennen, ist hierbei zu beachten, daß die Länge einer Charaktervariable in SPSS/PC+ länger ist als in SAS/PC, d.h., daß eventuell Werte abgeschnitten werden.
- Da die numerischen Variablentypen, wie sie in SPSS/PC+ gespeichert sind bei der Umsetzung in double precision Typen umgewandelt werden, entsteht eine um ein vielfaches größere Ausgabedatei, die auch auf keine Diskette mehr zu speichern ist. Hierbei kann man sich mit einem Komprimierprogramm behelfen (die circa 12 MB große SAS/PC-Datei konnte auf gut 1 MB komprimiert werden) oder man reduziert die zu übertragende Variablenzahl auf die unabdingbar notwendige, entweder schon mit Hilfe des Ausgangsstatistiksystems oder über die Variablenauswahl in DBMS/COPY Plus. Ebenfalls über die Variablen-Formatdefinition in DBMS/COPY Plus läßt sich das "Aufblähen" der Zieldatei steuern. Dies scheint uns jedoch nur dann ein sinnvolles Verfahren zu sein, wenn für eine große Variablenanzahl die gleiche Variablen-Formatänderung durchgeführt werden kann bzw. wenn die Variablenanzahl begrenzt ist.
- Sowohl SPSS/PC+ wie auch SAS/PC kennen auf ihre spezifische Weise benutzerdefinierte fehlende Werte. Da dies jedoch bei der Umwandlung durch DBMS/COPY im System-Missing-Werte umgesetzt werden, ist zu entscheiden, ob dies für die Datenanalyse im Zielsystem von Bedeutung ist, d.h. müssen auch im Zielsystem Variablenwerte als benutzerdefinierte fehlende Werte behandelbar sein, so sind die Deklarationen als fehlender Wert in SPSS/PC+ zu entfernen und in SAS/PC neu zu definieren.
- Bei der Umsetzung der Variablennamen ist kein Problem zu erwarten, da beide Statistiksysteme maximal acht Zeichen zulassen. Lediglich bei der Länge der Variablenlabels könnten bei der Umsetzung Verstümmelungen vorkommen, da SPSS/PC+ 60 Zeichen, SAS/PC jedoch nur 40 Zeichen zuläßt.

5. Zusammenfassung

Ein Programm wie DBMS/COPY ist ein sehr hilfreiches Instrument bei der Umsetzung von Dateien aus einem Statistik-Programmsystem in ein anderes. Allerdings sollte man bei einer solchen Umsetzung möglichst viel über beide Systeme, das Ausgangssystem und das Zielsystem wissen, um nicht böse Überraschungen zu erleben. Verglichen mit dem Aufwand, alle Angaben von Spaltenpositionen der Variablen in der Datei über Variablenlabels bis hin zu Wertelabels neu eingeben zu müssen, sind der Aufwand und die Informationsverluste vergleichsweise gering, die beim Umsetzen mit DBMS/COPY zu bewältigen sind.

Zu hoffen bleibt, daß auch der Hersteller des Systems noch einige Arbeit in DBMS/COPY investiert, um Probleme mit dem Zusatzmodul DBMS/COPY Plus zu bereinigen und vor allem einige der oben beschriebenen Mängel zu beheben. Zu wünschen wäre auch, daß die Dokumentation der internen Struktur der von DBMS/COPY erzeugten Umsetzdatei verbessert wird.

Literatur

Conceptual Software, Inc., 1989: dbms/copy. The Tool For Software Connectivity. Houston: Conceptual Software.

Conceptual Software, Inc., 1990: dbms/copy Plus. The Tool For Software Connectivity. Houston: Conceptual Software.

Mitteilungen

Neue PC-Version CLUSTAN 3.3

Ab sofort steht eine neue PC-Version des Clusteranalyse-Programms CLUSTAN zur Verfügung. CLUSTAN ist ein von D. Wishart, Edinburgh, entwickeltes Programm zur Klassifikation von Daten nach unterschiedlichsten Kriterien. Es bietet sowohl für kontinuierliche als auch für binäre Daten eine große Auswahl an Ähnlichkeitsmaßen, wie z.B. eine Reihe von verschiedenen Algorithmen zur eigentlichen Typenbildung. Graphische Darstellungen der Ergebnisse sind sowohl auf dem Bildschirm und auch auf einem Postscript-Drucker möglich.

CLUSTAN 3.3 enthält folgende Hauptfunktionen:

- Datenklassifikation:
  CLUSTAN 3.3 bietet mehr als 15 unterschiedliche Hauptverfahren der Datenklassifikation (Cluster) und der Überprüfung gewonnener Klassifikationen.

- Dateneingabe, Datenaufbereitung:
  Es können sowohl Rohdaten als auch Ähnlichkeitsmatrizen eingelesen werden. Die Behandlung von fehlenden Werten und Labels ist möglich. Rohdaten müssen für die Klassifikationsprozeduren in Ähnlichkeitsmatrizen überführt werden. CLUSTAN führt dazu entweder eine Faktorenanalyse (Hauptachsenlösung) oder eine Berechnung unterschiedlichster Assoziations- bzw. Dissoziationskoeffizienten durch (42 unterschiedliche Koeffizienten stehen zur Wahl). Alle Variablen können standardisiert werden.

- Ausgabe der Analyseergebnisse:
  Die Ergebnisse der einzelnen Klassifikationen können sowohl tabellarisch als auch graphisch (auf Bildschirm/Postscript-Druckern) dargestellt werden. Daneben können die Ergebnisse in eine Datei zur Weiterverarbeitung mit anderen Programmen oder zur Wiedereingabe in CLUSTAN gespeichert werden.

Neu gegenüber Release 3.2 ist vor allem die verbesserte Bildschirmdarstellung, die Möglichkeit, seitenweise zu blättern und die Grafikausgabe, die sowohl am Bildschirm als auch auf Postscript-Druckern erfolgen kann.

Um CLUSTAN PC einsetzen zu können, wird ein IBM PC oder ein kompatibler PC mit mindestens 512k Kernspeicher und mit einer Festplatte benötigt. Extended oder Expanded Memory wird nicht genutzt. Ein mathematischer Co-Processor ist nicht erforderlich; ist er vorhanden reduziert sich die Laufzeit der Prozeduren deutlich. Grafiken können an PCs mit CGA, EGA, VGA und Herkules-Grafikkarten dargestellt werden. Die Ausgabe der Grafiken kann auf Postscript-Druckern erfolgen, eine Unterstützung von HP Laserjet-Druckern ist in Vorbereitung, ist aber derzeit noch nicht möglich.

Die PC Version kostet als Einzelkopie US $1000.-. Preise für Mehrfach-Lizenzen sind zu erfragen. Neben der PC-Version sind eine Vielzahl von Mainframe-Versionen des Programmsystems erhältlich.

Anfragen richten Sie bitte an Cornelia Züll.

Literatur

D. Wishart, 1987: CLUSTAN User Manual (Fourth Edition). Computer Laboratory, University of St. Andrews.

°

Der SOZIALWISSENSCHAFTEN-BUS mit neuen Preisen

In gemeinsamer Verantwortung von ZUMA und GFM-GETAS (Hamburg) wird dreimal jährlich der Sozialwissenschaften-Bus durchgeführt. Dieser Bus ist eine Service-Einrichtung für die deutschen Sozialwissenschaften. Hiermit wird es Sozialwissenschaftlern an Hochschulen sowie an Forschungsanstalten und Einrichtungen des Bundes und der Länder ermöglicht, sich mit Einzelfragen bzw. kleinen Fragebatterien an einer sozialwissenschaftlichen Mehrthemenumfrage zu beteiligen. Die Besonderheiten dieser in der Tradition des ZUMA-Bus stehenden Mehrthemenumfrage gegenüber herkömmlichen Mehrthemenumfragen der Sozialforschungsinstitute sind:

- Durch die Beschränkung auf sozialwissenschaftliche Fragestellungen wird eine zu große Fragenheterogenität vermieden;
- durch die Verknüpfung von inhaltlichen Frageneinschaltungen mit der ZUMA-Standarddemographie ist eine den höheren Anforderungen der Sozialwissenschaften angemessene Auswertung möglich;
- anders als bei den üblichen Buseinschaltungen findet (entsprechend dem bisherigen ZUMA-Bus) vor Beginn des Hauptfeldes ein Pretest statt;

- durch Feldkontrollen und Datenbereinigung mit einem bei Mehrthemenumfragen nicht branchenüblichen Aufwand wird eine außergewöhnlich hohe Datenqualität gewährleistet.

Preise Sozialwissenschaften-Bus 1992

Die Abrechnung für die einzelnen Einschaltungen erfolgt nicht nach der Anzahl der Fragen, sondern nach dem jeweils notwendigen Befragungs-Zeitaufwand. Der Preis schließt bereits eine dem Gesamtumfang der Einschaltung normalerweise entsprechende Anzahl von Befragungshilfen (Listen, Kärtchenspiele, Skalenvorlagen) ein.

Nachdem Befragungen in den östlichen Bundesländern sowohl hinsichtlich der Stichprobe als auch hinsichtlich der Datenerhebung durch einen gutausgebauten großen Interviewerstab heute nicht mehr aufwendiger sind als in den westlichen Bundesländern, können wir für die Einschaltungen "West" und "Ost" vom selben Preis pro Minute als Kalkulationsbasis ausgehen. Hierbei bezeichnet der Preis in der ersten Spalte den Preis pro Minute für einen einzuschaltenden Fragenblock bis unter einer Viertel Stunde Dauer und der in der zweiten Spalte den Preis pro Minute bei Fragenblöcken, die 15 und mehr Minuten Zeit im Interview einnehmen:

Erhebungsgebiet	Anzahl Interviews	Preis Befragungszeit pro Minute *	
		bis 15 Min.	ab 15 Min.
West oder Ost	2.000	DM 3.400,-	DM 3.150,-
West oder Ost	1.000	DM 2.600,-	DM 2.450,-
nur Ost	500	DM 1.950,-	DM 1.850,-

* Den Kosten ist die gesetzliche Mehrwertsteuer hinzuzurechnen

Bei 2.000 Interviews entspricht dieser Preis in etwa einem mittleren Preis von DM 1.600,- pro Frage. Für offene Fragen wird der zusätzliche Vercodungsaufwand gesondert berechnet. Einzelfragen werden nach Aufwand kalkuliert. Sonderrabatt bei Mehrfacheinschaltungen.

Im Preis enthalten sind:

- Beratung bei der Fragenerstellung und bei der Fragenabfolge,
- Pretest,
- Schreiben und Drucken der Befragungsunterlagen,
- Durchführung der Feldarbeit,
- Datenkontrolle und Datenbereinigung sowie Datenqualitätskontrollen,
- Vercoden geschlossener Fragen,

- eine normale Anzahl von Listen oder Karten,
- das Grundmodul der ZUMA-Standarddemographie.

Termine

Terminplan für den Sozialwissenschaften-Bus III/92

Pretestbeginn		03.09.
Hauptfeld	Beginn:	14.10.
	Ende:	20.11.
Datenbandauslieferung		22.12.

Der Abgabetermin für die einzuschaltenden Fragen ist jeweils vier Wochen vor Pretest-Termin, bei Verzicht auf eine Pretest-Überprüfung 14 Tage vor Beginn des Hauptfeldes.

Buchbesprechungen

Walter Müller/Uwe Blien/Peter Knoche/Heike Wirth u.a.: Die faktische Anonymität von Mikrodaten, Band 19 der Schriftenreihe FORUM der Bundesstatistik, herausgegeben vom Statistischen Bundesamt, Metzler-Poeschel Verlag, 1991. 482 Seiten, 23,20 Mark. ISBN 3-8246-0231-8.

Viele Forschungsthemen in den Sozialwissenschaften, der Medizin und anderen Gebieten sind auf Mikrodaten, d.h. anonymisierte Daten über einzelne Personen angewiesen. Die amtliche Statistik erhebt solche Daten, die für Forschungszwecke schon immer hochinteressant waren. Mit dem Bundesstatistikgesetz in der Fassung von 1987 wurde eingeführt, daß Mikrodaten an die wissenschaftliche Forschung in einer "faktisch anonymisierten Form" weitergegeben werden dürfen. Unklar war bisher, wann Daten als faktisch anonymisiert gelten können. Das Buch von Müller et al. ist das Ergebnis eines groß angelegten Forschungsprojektes, in dem theoretisch und empirisch überprüft wurde, unter welchen Bedingungen zwei Erhebungen der amtlichen Statistik, der Mikrozensus und die Einkommens- und Verbrauchsstichprobe, als faktisch anonymisiert angesehen werden können.

In dem Buch werden sehr ausführlich, zum Teil etwas breit, die zu diesem Themenkomplex gehörenden Probleme sachkundig dargestellt. Nach einführenden Teilen, in denen die Begriffe Mikrodaten, Anonymität, faktische Anonymität, die sogenannte Unverhältnismäßigkeit der Reidentifikation sowie die gesetzlichen Grundlagen verdeutlicht werden, enthält das zweite Kapitel den Stand der Forschung und die gegenwärtige Praxis der Datenweitergabe im nationalen und internationalen Kontext. Das dritte Kapitel erschließt methodische Grundlagen von Reidentifikationsversuchen, das vierte Kapitel beschreibt Probleme, die bei der Reidentifikation auftreten können. Im fünften Kapitel werden Motive behandelt, die für eine Deanonymisierung ausschlaggebend sein können. Das sechste Kapitel analysiert das Zusatzwissen zur Deanonymisierung, d.h. personenbezogene Daten oder Nachschlagwerke, welche normalerweise im wissenschaftlichen Bereich zugänglich sind. Im siebten Kapitel werden Kosten-Nutzen-Überlegungen zu Reidentifikationsversuchen angestellt.

Die neuen Forschungsergebnisse des Projektes sind im achten bis zehnten Kapitel enthalten. Dort werden die Entwicklung konkreter Angriffsszenarien zur Reidentifikation sowie die Datenbasen, mit denen eine empirische Überprüfung erfolgte, beschrieben. Das neunte Kapitel enthält die von einem Treuhänder überprüften Ergebnisse von Reidentifikationsversuchen mit

verschiedenen Verfahren. Wichtig für die Forschung ist hierbei, daß bisher das Reidentifikationsrisiko viel zu hoch eingeschätzt wurde. Bedingt durch Dateninkompatibilitäten und andere Faktoren lassen sich tatsächlich selbst aus so großen Stichproben wie dem Mikrozensus nur wenige Daten zu bestimmten Personen zuordnen. Die Überprüfung der Zuordnung ergab jedoch, daß der größte Teil dieser vermeintlich eindeutigen Zuordnungen falsch war. Im zehnten Kapitel werden dann für die Daten, die nicht empirisch überprüft werden konnten, Kosten-Nutzen-Überlegungen für Deanonymisierungsstrategien diskutiert.

Das Buch schließt mit einer Darstellung von Anonymisierungsmaßnahmen sowie technisch organisatorischen Sicherheitsmaßnahmen gegen unbefugten Dateizugriff, einer zusammenfassenden Bewertung der untersuchten Szenarien sowie Empfehlungen zur Weitergabe von Einzelangaben aus dem Mikrozensus und der EVS an die Wissenschaft.

Der Band ist in einem flüssigen, gut lesbaren Stil geschrieben, verläßt jedoch nie die saubere wissenschaftliche Argumentation. Bedingt durch die unterschiedliche Autorenschaft der einzelnen Kapitel und vermutlich auch das relativ schnelle Erscheinen des Bandes sind jedoch einige Wiederholungen vorhanden.

Ich halte das Buch für lesens- und beachtenswert für alle, die empirisch mit Mikrodaten arbeiten. Aus dem Werk kann man entnehmen, welche möglichen Angriffsstrategien denkbar sind und wie man sich davor schützen kann. Zusätzlich ist der Band eine ausgezeichnete Hilfe, wenn man selbst Daten vom Statistischen Bundesamt erhalten möchte. Es lohnt sich, zwar nicht empirisch, jedoch theoretisch, anhand der Vorgaben dieses Buches zu prüfen, ob die Daten, die man selbst benötigt, als faktisch anonym eingeordnet werden können. Insoweit weist das Buch weit über den Mikrozensus und die Einkommens- und Verbrauchsstichprobe hinaus.

Ralph Brennecke, Freie Universität Berlin

ZUMA-Arbeitsberichte

Nachfolgend sind die ZUMA-Arbeitsberichte, die seit November 1991 publiziert worden sind, in Form von Abstracts kurz dargestellt. ZUMA-Arbeitsberichte werden Interessenten auf Anfrage zugesandt. Bestellungen sind zu richten an:

Zentrum für Umfragen, Methoden und Analysen
ZUMA-Publikationen
Postfach 12 21 55
6800 Mannheim 1

* * *

Schneid, Michael: Einsatz computergestützter Befragungssysteme in der Bundesrepublik Deutschland. ZUMA-Arbeitsbericht 91/20.
Im Sommer 1991 führte ZUMA eine Befragung bei bundesdeutschen Marktforschungsinstituten durch, um detaillierte Angaben darüber zu erhalten, welcher Stellenwert Computern bei der Datenerhebung in der kommerziellen Markt- und Meinungsforschung zugewiesen wird. Von den 79 auskunftsbereiten Instituten setzt derzeit etwa ein Drittel ein oder mehrere computergestützte Befragungssysteme ein. Etwa jedes vierte Institut trägt sich noch mit dem Gedanken, ein computergestütztes Befragungssystem einzusetzen, wobei sich die Erwartungen dieser Noch-Nicht-Nutzer weitestgehend mit den Erfahrungen der Nutzer decken: die Vorteile eines Befragungssystems sind vor allem darin zu sehen, daß die Befragungsergebnisse schnell vorliegen und komplexere Befragungsabläufe durchgeführt werden können. Als nachteilig werden bei einem computergestützten Befragungssystem die hohen Kosten sowie die Beschränkung auf standardisierte Fragen angesehen. Die Institute können in verschiedene Nutzer-Typen unterteilt werden, die die Befragungssysteme unterschiedlich bewerten. So beurteilen jene Institute, die im vergangenen Jahr vergleichsweise viele CATI-Studien durchgeführt haben, die Fähigkeiten und Möglichkeiten insgesamt positiver als jene Institute, die nur wenig (oder keine) Studien mit computergestützten Befragungssystemen durchgeführt haben. Die Hälfte der befragten Institute setzt kein Befragungssystem ein und wird auch zukünftig kein System einsetzen. Begründet wird der Nicht-Einsatz vor allem mit der Arbeitsweise und den methodischen Schwerpunkten des Instituts, sowie mit den Kosten eines computergestützten Befragungssystems.

Porst, Rolf/Schneid, Michael: Software-Anforderungen an computergestützte Befragungssysteme. ZUMA-Arbeitsbericht 91/21.
In dem Arbeitsbericht wird beschrieben, welche Anforderungen aus der Sicht empirischer Methodenforschung an die Software für computergestützte Befragungen gestellt werden und wie ein bei ZUMA eingesetztes Programm auf diese Anforderungen reagiert. Abgehandelt werden Erwartungen und Realisierung in den Bereichen Vorbereitung der Befragung, Durchführung der Befragung und Datenorganisation nach Abschluß der Befragung; breiter Raum ist dem Befragungsinstrument gewidmet. Die Abhandlung schließt mit Perspektiven der computergestützten Datenerhebung.

° ° °

Mueller, Ulrich: The Reproductive Success of the Elites in Germany, Great Britain, Japan and the USA during the 19th and 20th Century. ZUMA-Arbeitsbericht 91/22.

° ° °

Hartmann, Peter H./Schimpl-Neimanns, Bernhard: Zur Repräsentativität sozialdemographischer Merkmale des ALLBUS - Multivariate Analysen zum Mittelschichtbias der Umfrageforschung. ZUMA-Arbeitsbericht 92/01.
(Erscheint in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, Jg. 44, Heft 2, 1992: Sind Sozialstrukturanalysen mit Umfragedaten möglich? Analysen zur Repräsentativität einer Sozialforschungsumfrage.)
Umfragedaten und amtliche Statistik sollen die Verteilung demographischer und sozialstruktureller Merkmale der Bevölkerung korrekt wiedergeben. Aufgrund des selektiven Ausfalls bestimmter Bevölkerungsgruppen kommt es bei Umfragedaten jedoch zu Verzerrungen, die für verschiedene Umfragen schon oft durch univariate Vergleiche mit amtlichen Daten nachgewiesen werden konnten. Nachgewiesen wurden insbesondere selektive Ausfälle bei kleinen Haushalten und bei Personen mit niedriger sozialer Schichtzugehörigkeit (sog. Mittelschichtbias). Für die Sozialforschungsumfrage ALLBUS konnten im Vergleich zum amtlichen Mikrozensus diese Abweichungen reproduziert werden. Die Verzerrung bei der Haushaltsgröße erweist sich als besonders folgenreich, weil sie eine personenrepräsentative Gewichtung der Haushaltsstichprobe verhindert. Im Kontext der sozialen Schicht konnte durch multivariate Analysen nachgewiesen werden, daß die Ausfälle vor allem als Effekte von Bildungsabschluß und Beteiligung am Erwerbsleben erklärbar sind. Der sogenannte Mittelschichtbias erweist sich dabei im wesentlichen als Bildungsbias. Bei Versuchen, mit Hilfe von Umfragedaten Merkmale der Haushaltsgröße, der sozialen Schichtzugehörigkeit oder der

Beteiligung am Erwerbsleben zu beschreiben oder zu erklären, ist mit Verzerrungen zu rechnen.

° ° °

Bohner, Gerd/Crow, Kimberly/Erb, Hans-Peter/Schwarz, Norbert: Affect and persuasion: Mood effects on the processing of message content and context cues, and subsequent behavior. ZUMA-Arbeitsbericht 92/02.
(Erscheint in: European Journal of Social Psychology.)
The impact of recipients' mood on the processing of simple persuasive communications and subsequent behavior is explored. As expected on theoretical grounds, elated moods reduced subjects' motivation to process message content and contextual cues in field and laboratory experiments. Implications for persuasion theories and the interplay of affect and cognition are discussed.

° ° °

Bless, Herbert/Bohner, Gerd/Hild, Traudel/Schwarz, Norbert: Asking difficult questions: Task complexity increases the impact of response alternatives. ZUMA-Arbeitsbericht 92/03.
(Erscheint in: European Journal of Social Psychology.)
In providing behavioral frequency reports, respondents use the range of the response alternatives as a frame of reference. This results in higher frequency estimates on scales that offer high rather than low values. The present study demonstrates that the size of this scale effect increases with increasing question difficulty.

° ° °

Bandilla, Wolfgang/Gabler, Siegfried/Wiedenbeck, Michael: Methodenbericht zum DFG-Projekt ALLBUS Baseline-Studie 1991. ZUMA-Arbeitsbericht 92/04.
Innerhalb des Forschungsprogramms ALLBUS wurde im Sommer 1991 die Basisumfrage Gesamtdeutschland durchgeführt. Hierbei handelt es sich um eine Sonderstudie, die von der Deutschen Forschungsgemeinschaft (DFG) finanziert wurde und deren besonderer Stellenwert sich aus der deutschen Vereinigung erklärt. Mit der Studie wurden vornehmlich zwei Ziele verfolgt. Zum einen sollte ein möglichst umfassendes Bild der unterschiedlichen aktuellen Realitäten in beiden deutschen Staaten gewonnen werden, und zwar noch bevor sich neue Strukturen in größerem Umfang gebildet und verfestigt haben. Zum zweiten ging es darum, für die künftige Erfassung des zu erwartenden gesellschaftlichen Wandels die hierzu notwendigen Zeitreihen zu begründen. Im Methodenbericht werden Planung, Vorbereitung

und Durchführung der Studie ausführlich dokumentiert. Breiten Raum nimmt die Darstellung und Begründung des Fragenprogramms ein. Darüber hinaus werden das Stichprobendesign (einschließlich der unterschiedlichen Auswahlverfahren in Ost- und Westdeutschland) sowie die Feldphase der Studie dokumentiert.

° ° °

Faulbaum, Frank: Von der Variablenanalyse zur Evaluation von Handlungs- und Prozeßzusammenhängen. ZUMA-Arbeitsbericht 92/05.
Der Arbeitsbericht beschäftigt sich mit den theoretischen Restriktionen der in den empirischen Verhaltens- und Sozialwissenschaften vorherrschenden funktionalen Modellparadigmen bei der Erklärung von Zusammenhängen zwischen beobachteten und/oder unbeobachteten Variablen. Es wird vorgeschlagen, sogenannte operative Modellparadigmen in Betracht zu ziehen, mit denen es möglich ist, Annahmen über unbeobachtete Handlungs- und Prozeßzusammenhänge beliebiger Komplexität in erklärende Modelle einzubeziehen.

° ° °

Borg, Ingwer: Überlegungen und Untersuchungen zur Messung der subjektiven Unsicherheit der Arbeitsstelle. ZUMA-Arbeitsbericht 92/06.
Eine kurze Skala zur Messung der subjektiven Unsicherheit der Arbeitsstelle (SUSA) wird entwickelt. Theoretisch werden zunächst zwei Dimensionen der SUSA unterschieden, eine affektive (Befürchtungen, Bangen) und eine kognitive (Bedenken). Die Antworten von 200 Arbeitnehmern verschiedener Berufe auf entsprechende Items zeigen klar die zwei-dimensionale Struktur der SUSA. Bedenken erweisen sich als notwendige, aber nicht hinreichende Bedingung für Befürchtungen. Die differentielle Validität beider Dimensionen zeigt sich darin, daß bei stärkeren internalen Kontrollüberzeugungen SUSA-Befürchtungen und vor allem SUSA-Bedenken abnehmen; bei stärkeren externalen Überzeugungen nehmen beide SUSA-Komponenten zu, vor allem die Befürchtungen. Zudem korreliert höhere soziale Unterstützung mit geringeren SUSA-Befürchtungen, ist aber unkorreliert mit SUSA-Bedenken. In einer weiteren Erhebung werden die Items der leicht revidierten SUSA-Skala von 125 Arbeitnehmern einer Textilfirma beantwortet. Die Faktorenstruktur der SUSA repliziert die der ersten Stichprobe. Die Dimensionen haben wieder unterschiedliche Vorhersagevalidität. SUSA-Befürchtungen korrelieren positiv mit ökonomischer Abhängigkeit, Verbundenheit mit der Arbeit (Involvement) und affektiver bzw. kalkulativer Bindung (Commitment) an die Firma. SUSA-Bedenken korrelieren negativ

mit affektiver Bindung, Arbeits- und Job-Bindung und positiv mit Entfremdung.

* * *

Borg, Ingwer/Braun, Michael: Arbeitsethik und Arbeitsinvolvement als Moderatoren der psychologischen Auswirkungen von Arbeitsplatzunsicherheit. ZUMA-Arbeitsbericht 92/07.
Frühere Befunde von Borg (1989) über die Korrelate der subjektiven Unsicherheit der Arbeitsstelle (SUSA) werden hier an einer deutschen Repräsentativ-Stichprobe überprüft. Dabei steht vor allem die Frage im Vordergrund, ob sich SUSA bei Personen mit hoher Arbeitsethik anders auf ihre Einstellungen und Meinungen auswirkt als bei Personen mit niedriger Arbeitsethik. Für Personen mit hoher Arbeitsethik ist er flacher z.T. sogar U-förmig, d.h. sehr hohe SUSA ist hier verbunden mit teilweise positiveren Einstellungen (z.B. Arbeitszufriedenheit) und Meinungen (z.B. über die Beziehungen von Mitarbeitern und Vorgesetzten). Eine Reanalyse der Borg-Daten zeigt dort ähnliche Trends. Die Details der Trendverläufe sind aber abhängig von ihren Ausgangswerten in den jeweils untersuchten Populationen, d.h. davon, wie positiv die Meinungen und Einstellungen bei vollständiger Arbeitsplatzsicherheit sind.

* * *

Singer, Eleanor/Hippler, Hans-J./Schwarz, Norbert: Confidentiality assurances in surveys: Reassurance or threat? ZUMA-Arbeitsbericht 92/08.
(Erscheint in: International Journal of Public Opinion Research.)
Over the last three decades, the public's willingness to take part in surveys has gradually declined, and the decline has been attributed in part to increasing concern about the confidentiality of the data requested. This paper reviews the early literature bearing on confidentiality assurances and willingness to respond, and then reports on three experiments designed to investigate the effects of confidentiality on the expectations of respondents and on their willingness to take part in a survey. The results of all three experiments confirm our expectation that confidentiality assurances are not always perceived as reassuring, and do not necessarily increase the public's willingness to respond.

Bless Herbert/Mackie, Diane/Schwarz, Norbert: Mood Effects on Attitude Judgments: The Independent Effects of Mood Before and After Message Elaboration. ZUMA-Arbeitsbericht 92/09.
(Erscheint in: Journal of Personality and Social Psychology.)
The independent effects of induced mood on the encoding of persuasive messages and on the assessment of attitude judgments is investigated. In Experiment 1, positive or negative mood was induced either prior to the encoding of a counterattitudinal message or prior to the assessment of attitude judgments. When mood was induced prior to message presentation, subjects in a bad mood were more persuaded by strong than by weak arguments, whereas subjects in a good mood were equally persuaded by strong and by weak arguments. When subjects encoded the message in a neutral mood, however, the advantage of strong over weak arguments was more pronounced when subjects were in a good rather than in a bad mood at the time of attitude assessment. In Experiment 2, subjects exposed to a weak or strong counterattitudinal were required to form a global evaluation or a detailed representation of the message. Positive, negative, or neutral mood was then induced. Subjects in a good mood were most likely and subjects in a negative mood least likely to base their reported attitudes on global evaluations. Implications for the impact of moods on processing strategies are discussed.

° ° °

ZUMA-Publikationen

ZUMA veröffentlicht die Ergebnisse seiner Arbeit regelmäßig in Sammelbänden, Monographien und Arbeitsberichten. Die nachfolgende Liste gibt einen Überblick über die wichtigsten Publikationen seit der Gründung des Instituts. Die Mitarbeiter von ZUMA veröffentlichen darüber hinaus regelmäßig Buchbeiträge und Aufsätze in Fachzeitschriften, die in den Jahresberichten von ZUMA aufgeführt werden. Zudem berichten wir in zwei Zeitschriften und einem Newsletter über neueste Forschungsergebnisse, Konferenzen, Termine und vieles mehr. Diese Zeitschriften senden wir Ihnen auf Anfrage kostenlos zu. Neben den ZUMA-Nachrichten ist dies der "Informationsdienst Soziale Indikatoren" (ISI), der seit 1989 jeweils im Januar und Juli erscheint und Beiträge zur Sozialberichterstattung und Sozialindikatorenforschung enthält. Der "Newsletter on Cognition and Survey Research" dient der Kommunikation zwischen Forschern, die an kognitiven Aspekten der Umfragemethodologie, insbesondere der Fragebogenkonstruktion interessiert sind. Er enthält Abstracts neuerer Arbeiten aus dem Bereich der Kognitionspsychologie und Umfragemethodik sowie Tagungsankündigungen. Der Newsletter erscheint in unregelmäßigen Abständen.

I. Methoden und Verfahren der Empirischen Sozialforschung

Hypothesen, Gleichungen und Daten. Spezifikations- und Meßprobleme bei Kausalmodellen für Daten aus einer und mehreren Beobachtungsperioden.
Von E. Weede.
Königstein/Ts.: Athenäum, 1977. 28 Mark. ISBN 3-7610-8201-0.
ZUMA-Monographien Sozialwissenschaftliche Methoden, Band 1 (vergriffen).
Inhalt: Möglichkeiten und Probleme der kausalen Abhängigkeitsanalyse von nicht-experimentellen Daten. Untersucht wird die Umsetzung inhaltlicher Hypothesen in Gleichungen und die Frage nach Annahmenbelastungen bei Überprüfungen.

Sozialstrukturanalysen mit Umfragedaten. Probleme der standardisierten Erfassung von Hintergrundsmerkmalen in allgemeinen Bevölkerungsumfragen.
Hrsg. von F.U. Pappi.
Königstein/Ts.: Athenäum, 1979. 46 Mark. ISBN 3-7610-8200-2.
ZUMA Monographien Sozialwissenschaftliche Methoden, Band 2 (vergriffen).
Inhalt: Vorschlag für ein umfassendes Instrumentarium zur Erhebung zentraler demographischer und anderer Hintergrundsvariablen in der Umfrageforsachung. Experten aus verschiedenen Disziplinen nehmen zu diesem Vorschlag Stellung und berichten über exemplarische Auswertungen.

Datenzugang und Datenschutz. Konsequenzen für die Forschung.
Von M. Kaase/H.J. Krupp/M. Pflanz/E.K. Scheuch/S. Simitis.
Königstein/Ts.: Athenäum, 1980. 57 Mark. ISBN 3-7610-8228-2.
ZUMA-Monographien Sozialwissenschaftliche Methoden, Band 3 (vergriffen).
Inhalt: Ergebnisse einer Kolloquienreihe zur Datenschutzproblematik in den Sozialwissenschaften. 25 Einzelbeiträge zu den Bereichen Datenbedarf, Datenzugangsproblematik und Datenschutzmaßnahmen.

Computerunterstützte Inhaltsanalyse in der empirischen Sozialforschung.
Hrsg. von H.-D. Klingemann.
Frankfurt: Campus, 1984. 46 Mark. ISBN 3-593-33254-X.
ZUMA-Monographien Sozialwissenschaftliche Methoden, Band 4.
Inhalt: Berichte über Forschungen aus der Pädagogik, den Kommunikationswissenschaften, der Politologie, den Wirtschaftswissenschaften und der Soziologie, die sich der computerunterstützten Inhaltsanalyse bedient haben.

Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. Beiträge zu methodischen Problemen des ALLBUS 1980.
Hrsg. von K.U. Mayer/P. Schmidt.
Frankfurt: Campus, 1980. 49 Mark. ISBN 3-593-33262-0.
ZUMA-Monographien Sozialwissenschaftliche Methoden, Band 5.
Inhalt: Beiträge zu methodischen Problemen der Datenerhebung des ALLBUS 1980 (Stichproben- und Gewichtungsprobleme, Interviewereffekte), exemplarische Analysen (Einstellungen zu Gastarbeitern, Postmaterialismus, Schichtidentifikation und Klassenkonflikte, Einkommensdiskriminierung von Frauen).

Soziale Empfindungen.
Von D. Krebs/K. Schuessler.
Frankfurt: Campus, 1987. 258 Seiten, 38 Mark. ISBN 3-593-33875-0.
ZUMA-Monographien Sozialwissenschaftliche Methoden, Band 6.
Inhalt: 9 Skalen zur Messung sozialer Empfindungen: Entfremdung, Vertrauen zu anderen Menschen, Selbstbestimmung, Zukunftsorientierung, Desillusion über Funktionsweisen des politischen Apparates,Glaube an die Demokratie, Arbeitszufriedenheit, subjektive Moral. Ihre interkulturelle Anwendung in der Survey-Forschung wird am Beispiel einer deutschen und einer amerikanischen Umfrage vorgestellt und diskutiert. Behandelt werden sowohl methodische Themen wie soziologisch bedeutsame Unterschiede innerhalb und zwischen den verglichenen Ländern.

Multivariate Analyse von Verlaufsdaten. Statistische Grundlagen und Anwendungsbeispiele für die dynamische Analyse nichtmetrischer Merkmale.
Von H.-J. Andreß.
Mannheim: ZUMA, 1985. 317 Seiten, 25 Mark. ISBN 3-924220-02-6.
Inhalt: Möglichkeiten und Probleme der kausalen Analyse von Daten, die Art und Zeitpunkt von Veränderungen auf individueller Ebene (z.B. geographische und berufliche Mobilität, Bildungskarriere, Erwerbsbiographien, Wechsel des Familienstandes) für eine größere Stichprobe möglichst genau erfassen.

An Empirical Study of the Reliability and Stability of Survey Research Items.
Hrsg. von G.W. Bohrnstedt/P.Ph. Mohler/W. Müller.
Sociological Methods and Research 15, 1987.
Inhalt: Am Beispiel der Daten des ALLBUS Dreiwellenpanels aus dem Jahr 1986, werden sowohl klassische als auch moderne Verfahren für die Messung von Item-Reliabilität und Stabilität vorgestellt und diskutiert. Zu den Autoren gehören u.a. G. Arminger, W. Jagodzinski, F. Faulbaum, R. Porst und P. Schmidt.

Lineare Modelle zur Analyse von Paneldaten.
Von G. Arminger/F. Müller.
Opladen: Westdeutscher Verlag, 1990. 228 Seiten, 38 Mark.
ISBN 3-531-12176-6.
Inhalt: Einführung in lineare Modelle zur Analyse von Paneldaten. Der Schwerpunkt liegt auf Modellen mit latenten Variablen, wobei auch Modelle mit latenten Konstruktvariablen und ordinalen Indikatoren beschrieben und analysiert werden. Alle Modelle werden anhand eines Beispieldatensatzes mit beigefügten LISREL -und LISCOMP-Programmen dargestellt und analysiert.

Modellbildung und Simulation in den Sozialwissenschaften.
Von K.G. Troitzsch.
Opladen: Westdeutscher Verlag, 1990. 206 Seiten, 36 Mark.
ISBN 3-531-12150-2.
Inhalt: Einführung in die wichtigsten Verfahren der mathematischen und computergestützten Modell- und Theoriebildung in den Sozialwissenschaften. Behandelt werden deterministische und stochastische Modelle von Prozessen der individuellen Meinungsbildung und der Veränderung von Einstellungen in größeren Kollektiven.

Merkmale einer allgemeinen Standarddemographie. Gegenüberstellung sozlodemographischer Variablen aus dem Mikrozensus, der Einkommens- und Verbrauchsstichprobe, der Volkszählung und der Standarddemographie des Zentrums für Umfragen, Methoden und Analysen.
Bearbeitet von Manfred Ehling und Jürgen H.P. Hoffmeyer-Zlotnik.
Heft 4 der Schriftenreihe Ausgewählte Arbeitsunterlagen der Bundesstatistik. Hrsg. vom Statistischen Bundesamt, Wiesbaden, 1988.
Inhalt: Synoptische Übersicht und methodische Hinweise über zentrale demographische Merkmale der Volkszählung 1987, des Mikrozensus, der Einkommens- und Verbrauchsstichprobe und der ZUMA Standarddemographie.

Social Information Processing and Survey Methodology.
Hrsg. von H-J. Hippler/N. Schwarz/S. Sudman.
New York: Springer, 1987. 223 Seiten, 65 Mark.
ISBN 3-540-96570-X.
Inhalt: Umfrageforschern ist seit langem bekannt, daß die Art und Weise, in der eine Frage gestellt wird, die erhaltenen Befunde beeinflussen kann. Ziel der Beiträge dieses Buches ist es, die kognitiven und kommunikativen Prozesse zu identifizieren, die solchen Kontexteffekten zugrunde liegen. Das Buch bietet eine Einführung in Theorien der Informationsverarbeitung und berichtet über exemplarische Untersuchungen zum Einfluß des Fragenkontextes und der Fragenformulierung aus kognitionspsychologischer Sicht.

Context effects in social and psychological research.
Hrsg. von N. Schwarz/S. Sudman.
New York: Springer, 1992, 84 Mark.
ISBN 3-540-97705-8.
Inhalt: Kognitionspsychologen und Umfrageforscher berichten über Untersuchungen zum Auftreten von Kontexteffekten bei Befragungen, in Laborexperimenten und in psychologischen Tests. Von besonderem Interesse sind Einflüsse der Reihenfolge von Fragen und von Antwortvorgaben. Mehrere Beiträge stellen theoretische Modelle zur Erklärung und Vorhersage von Kontexteffekten zur Verfügung.

Werte und Wandel: Ergebnisse und Methoden einer Forschungstradition.
Hrsg. von H. Klages/H-J. Hippler/W. Herbert.
Frankfurt/New York: Campus, 1992, 694 Seiten, 98 Mark.
ISBN 3-593-34469-6.
Inhalt: Im September 1989, also noch vor Beginn der zur Wiedervereinigung Deutschlands führenden Ereignisse im Spätherbst, fand am Forschungsinstitut für öffentliche Verwaltung bei der Hochschule für Verwaltungswissenschaften in Speyer eine gemeinsam mit dem Zentrum für

Umfragen, Methoden und Analysen in Mannheim (ZUMA) veranstaltete internationale Konferenz zu Standort und Zukunft der Werteforschung statt. Der Band enthält 31 Vorträge von Referenten aus dem universitären, Marktforschungs- und dem Verwaltungsbereich. Der gemeinsame Tenor der Beiträge, welche sich mit den aktuellen Trends des Wertewandels in den achtziger Jahren beschäftigten, kann dabei zur Formel "Selbstverwirklichung durch (hedonistischen) Konsum und nicht durch (politisches) Engagement" verkürzt werden.

Analyse verbaler Daten: über den Umgang mit qualitativen Daten.
Hrsg. von Jürgen H.P. Hoffmeyer-Zlotnik.
Opladen: Westdeutscher Verlag, 1992, 424 Seiten.
ISBN 3-531-12360-2.
Inhalt: Das Buch behandelt die unterschiedlichen Möglichkeiten und Ansätze der Analyse qualitativer Daten. Besonderer Wert wurde darauf gelegt, daß die unterschiedlichen Analyseansätze nicht nur methodologisch diskutiert, sondern auch hinsichtlich methodischer Anwendung an jeweils konkreten Projektfragestellungen und -daten demonstriert werden. Damit ist dieser Band eine praxisorientierte Einführung in die Analyse verbaler Daten.

II. "Allgemeine Bevölkerungsumfrage der Sozialwissenschaften" (ALLBUS) und "International Social Survey Program" (ISSP)

Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) 1980.
Hrsg. vom Zentralarchiv für empirische Sozialforschung der Universität zu Köln und vom Zentrum für Umfragen, Methoden und Analysen, Mannheim.
Codebuch mit Methodenbericht und Vergleichsdaten. ZA-Nr. 1000.
Projektleitung: M. Rainer Lepsius, Erwin K. Scheuch, Rolf Ziegler, Köln: Zentralarchiv für emirische Sozialforschung.
1. Auflage 1982, 353 Seiten, 60 Mark. ISBN 3-88387-020-X.
Inhalt: Randverteilungen der jeweiligen Fragen aus allen verfügbaren deutschen und ausländischen Vorbildstudien, chronologisches und alphabetisches Verzeichnis der Vorbildstudien, ausführliche Übersicht über die spezifischen Fragenmodifikationen, umfassender Methodenbericht über die Vorbereitung und Durchführung des ALLBUS 1980 (Stichprobendesign, Ausschöpfung und Durchführung der Interviews). Im gleichen Format erschienen ist der "ALLBUS 1982".

Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) 1982.
Hrsg. vom Zentralarchiv für empirische Sozialforschung der Universität zu Köln und vom Zentrum für Umfragen, Methoden und Analysen, Mannheim. Codebuch mit Methodenbericht und Vergleichsdaten. ZA-Nr. 1160. Projektleitung: M. Rainer Lepsius, Erwin K. Scheuch, Rolf Ziegler. Köln: Zentralarchiv für empirische Sozialforschung, 1. Auflage 1984, 498 Seiten, 25 Mark.
ISBN 3-88387-028-5.

Allgemeine Bevölkerungsumfrage der Sozialwissenschaften. Beiträge zu methodischen Problemen des ALLBUS 1980.
Hrsg. von K.U. Mayer/P. Schmidt.
Frankfurt: Campus, 1980. 49 Mark. ISBN 3-593-33262-0.
ZUMA-Monographien Sozialwissenschaftliche Methoden, Band 5.
Inhalt: Beiträge zu methodischen Problemen der Datenerhebung des ALLBUS 1980 (Stichproben- und Gewichtungsprobleme, Interviewereffekte), Exemplarische Analysen (Einstellungen zu Gastarbeitern, Postmaterialismus, Schichtidentifikation und Klassenkonflikte, Einkommensdiskriminierung von Frauen).

Blickpunkt Gesellschaft. Einstellungen und Verhalten der Bundesbürger.
Hrsg. von W. Müller/P.Ph. Mohler/B. Erbslöh/M. Wasmer.
Westdeutscher Verlag, 1990. 224 Seiten, 38 Mark.
ISBN 3-531-12170-7.
Inhalt: In neun Beiträgen werden die Daten des ALLBUS und des ISSP genutzt, um durch Informationen zu neueren Entwicklungen in Einstellungen und Verhaltensweisen der Bundesbürger einen Beitrag zur allgemeinen Sozialberichterstattung zu leisten.

Attitudes to Inequality and the Role of Government.
Hrsg. von J.W. Becker/J.A. Dawis/P. Ester/P.Ph. Mohler.
Rijswijk, Netherlands: Social and Cultural Planning Office, 1990.
ISBN 90-377-0041-1.
Inhalt: Das Buch gibt eine erste zusammenfassende Darstellung von Ergebnissen des International Social Survey Program (ISSP), u.a. mit Beiträgen zu Sozialer Ungleichheit und zu politischen Einstellungen.

III. Handbuch Sozialwissenschaftlicher Skalen

Hrsg. vom Zentrum für Umfragen, Methoden und Analysen, Mannheim und vom Informationszentrum Sozialwissenschaften, Bonn.
Wissenschaftliche Bearbeitung: Jutta Allmendinger, Peter Schmidt und Bernd Wegener. Dokumentarisch Bearbeitung: Theodor Eikelmann und Peter Ohly. Bonn 1983 (Teil 1 und 2). ISBN 3-8206-0019-1.
Derzeitiger Stand: Gesamtausgabe 110 Mark inclusive aller bisher erschienenen Nachlieferungen (1985-1988). Bearbeitung fortgeführt von Dagmar Krebs (wissenschaftlich) und H. Peter Ohly (dokumentarisch).

Das ZUMA-Handbuch Sozialwissenschaftlicher Skalen ist eine umfassende Arbeitsunterlage für Lehre und Forschung in der empirischen Sozialwissenschaft. Es dokumentiert in mittlerweile drei Bänden (Loseblatt) 136 deutschsprachige Einstellungsskalen aus Soziologie, Politscher Wissenschaft und Sozialpsychologie zu Themen wie Anomie, Autoritarismus, Entfremdung, Kontrollerwartung, Wertwandel, Politische Partizipation, Arbeitszufriedenheit und Arbeitsorientierungen, Religiosität, Bürokratie und Organisation sowie soziale Ungleichheit. Die Dokumentation erfolgte nach einem Kriterienkatalog, der u.a. eine genaue Testkennzeichnung, den Forschungskontext und theoretischen Hintergrund, die Testdarstellung (Skalenvorgabe, Skalierungsmodell), die methodische Qualität von Indikatoren und Konstrukten (Korrelationen, Dimensionalität der Skala, Reliabilität und Validität) und die genaue Formulierung der Items umfaßt. Neben den Skalenbeschreibungen enthält das Handbuch eine Einführung in das dokumentarische Vorgehen und die Theorie der Skalenbildung sowie ein Autoren und Sachregister.

IV. Datenverarbeitung und Statistik

Wissenschaftliche Anwendung von Statistik-Software SoftStat Tagungsbände

Die SoftStat-Konferenzen befassen sich mit dem Einsatz und der Untersuchung von Methoden und Werkzeugen für die Statistik, insbesondere mit der Anwendung und dem Vergleich von statistischen Auswertungssystemen. 1982 zum erstenmal durchgeführt, fand im April 1991 bereits die sechste Konferenz statt.

Statistik-Software in der Sozialforschung.
Hrsg. von Wilke H. u.a. Berlin: Quorum, 1983.
10 Mark. ISBN 3-88726-006-6.
Berichtsband der zweiten Konferenz über die wissenschaftliche Anwendung von Statistik-Software. Beiträge über spezielle statistische Verfahren und Programmvergleiche (z.B. Kontingenztafelanalyse, Clusteranalyse, Netzwerkanalyse, LISREL), Abhandlungen über Lösungsalternativen für Aufgaben der Datenverwaltung, Übersichtsbeiträge über das Programmangebot zur Erstellung tabellarischer und graphischer Darstellungen sowie thematische Karten.

Statistik-Software. 3. Konferenz über die wissenschaftliche Anwendung von Statistik-Software 1985.
Hrsg. von W. Lehmacher/A. Hörmann.
Stuttgart: Fischer, 1986. 399 Seiten, 58 Mark.
Berichtsband über die dritte SoftStat Koferenz. Beiträge über den Vergleich von Programmsystemen, GLIM, Pfadmodelle mit latenten Variablen, Software für Analyse von Verweildauern, Software für Biologisch-Medizinische Anwendungen in APL, Software für Textanalyse.

Fortschritte der Statistik-Software 1. 4. Konferenz über die wissenschaftliche Anwendung von Statistik-Software.
Hrsg. von F. Faulbaum/H.-M. Uehlinger.
Stuttgart: Fischer, 1988. 595 Seiten, 76 Mark.
ISBN 3-437-50320-0.
Berichtsband über die vierte SoftStat Konferenz. Beiträge über Statistik-Programme zur Datenanalyse (Datenerhebung und Datenmanagement, Vergleich und Bewertungen), Statistik-Pakete im PC-Bereich, Expertensysteme, Individuelle Modellierung mit Statistik-Software, Exploratorische Datenanalyse, Skalierung und Klassifikation, Simulation, Statistische und DV-Aspekte linguistischer Datenverarbeitung.

SoftStat '89, Fortschritte der Statistik-Software 2. 5. Konferenz über die wissenschaftliche Anwendung von Statistik-Software.
Hrsg. von F. Faulbaum/R. Haux/K.-H. Jöckel.
Stuttgart: Fischer, 1990. 98 Mark. 644 Seiten.
ISBN 3-437-50335-9.
Berichtsband über die fünfte SoftStat Konferenz. Berichte über Statistische Auswertungssysteme für Datenverwaltung und Datenanalyse, Individuelle Modellierung mit Statistik-Software, Graphik und explorative Datenanalyse, Expertensysteme in der Statistik, Kartographie und geographische Informationssysteme, Statistik und Datenerhebung, Simulation, Skalierung und Klassifikation, Statistik und linguistische Datenverarbeitung,

Verarbeitung großer Datenbestände mit statistischen Auswertungssystemen, Statistik-Ausbildung und Statistik-Software, Rechnernetze in der Statistik.

SoftStat '91, Advances in Statistical Software 3.
Hrsg. von F. Faulbaum.
Stuttgart - Jena - New York: Gustav Fischer 1992. 536 Seiten, 94 Mark.
ISBN 3-437-40280-3.
Berichtsband über die 6. Konferenz mit Originalbeiträgen zu den folgenden Themenbereichen:
Statistische Auswertungssysteme in Datenverwaltung und Datenanalyse, Individuelle Modellierung, Wissensbasierte Systeme in der Statistik, Simulation, Systeme und Ansätze zur interaktiven graphischen Datenanalyse, Kartographie und Geographische Informationssysteme, Statistik und Linguistische Datenverarbeitung, Behandlung anonymisierter Daten, Computerunterstützte Datenerhebung, Rechnernetze und Paralleles Rechnen in der Statistischen Datenverarbeitung, Statistikausbildung und Statistik-Software, Statistische Arbeitsplatzsysteme, weitere Bereiche der Anwendung von Statistik-Software wie Statistische Qualitätskontrolle und Versuchsplanung.

TEXTPACK PC.
Von P.Ph. Mohler/C. Züll.
Mannheim: ZUMA, 1990. 50 Mark.
ISBN 3-924220-05-0.
TEXTPACK PC ist ein für die computerunterstützte Inhaltsanalyse (cui) entwickeltes Programmsystem. Es ermöglicht Dateien so auszugeben, daß sie anschließend mit Statistikprogrammen wie SAS, SPSS oder SIR eingelesen und weiter analysiert werden können. Die englischsprachige Dokumentation beschreibt die Handhabung des Systems ausführlich und anhand zahlreicher Beispiele.

Computer-Aided Text Classification. The General Inquirer III.
Von C. Züll/R.P. Weber/P.Ph. Mohler.
Mannheim: ZUMA, 1989. 258 Seiten, 75 Mark.
ISBN 3-924220-04-2.
Der General Inquirer und die zu ihm gehörenden inhaltsanalytischen Diktionäre (Harvard III, Harvard IV und Lasswell Value Dictionary) sind ein Analysesystem für die computerunterstützte Inhaltsanalyse. In der Dokumentation werden die Programme und die Diktionäre im Detail beschrieben und die theoretischen und methodologischen Hintergründe für die computerunterstütze Analyse behandelt.

Computerunterstütze Inhaltsanalyse mit TEXTPACK PC. Release 4.0 für IBM XT/AT und Kompatible unter MS/DOS ab Version 3.0.
Von C. Züll/P. Ph. Mohler/A. Geis.
Stuttgart: Fischer 1991. 157 Seiten, 44 Mark.
ISBN 3-437-40243-0.
TEXTPACK PC ist ein für die computerunterstützte Inhaltsanalyse (cui) entwickeltes Programmsystem. Es ermöglicht Dateien so auszugeben, daß sie anschließend mit Statistikprogrammen wie SAS, SPSS oder SIR eingelesen und weiter analysiert werden können. Diese und weitere Möglichkeiten der Textdeskription und Textanalyse werden ausführlich und anhand zahlreicher Beispiele beschrieben. Ergänzt wird das Buch durch einen Einblick in die Inhaltsanalyse allgemein und cui im besonderen.

PC-Graphik-Programme in der Statistik. Vergleichende Gegenüberstellung von PC-Graphik-Programmen mit Anwendungsbeispielen.
Von H. Ritter.
Stuttgart: Fischer, 1991. 230 Seiten, 39 Mark.
ISBN 3-437-40242-0.
Darstellung und vergleichende Bewertung von statistischen Graphik-Softwareprodukten (SPSS/PC+Graphics, HARVARD Graphics, CHART, SAS/GRAPH, STATGRAPHICS und STATA für den DOS-Bereich; DATA DESK Professional und JMP für Apple Macintosh); Hardware-Anforderungen.

V. Verschiedenes

Herausforderungen der Empirischen Sozialforschung. Beiträge aus Anlaß des zehnjährigen Bestehens des Zentrums für Umfragen, Methoden und Analysen.
Hrsg. von M. Kaase/M. Küchler.
Mannheim: ZUMA, 1985. 23 Mark. ISBN 3-924220-03-4 (vergriffen).
Reden und Referate zur Leistungsfähigkeit und Leistungsmöglichkeiten der Empirischen Sozialforschung anläßlich des zehnjährigen Bestehens von ZUMA.

Evolution und Spieltheorie.
Hrsg. von U. Mueller.
München: R. Oldenbourg Verlag, 1990. 215 Seiten, 88 Mark.
ISBN 3-486-55839-0.
Zehn originäre Aufsätze aus dem Forschungsbereich der evolutionären Spieltheorie, die seit 1973 erschienen sind, liegen hiermit erstmals in deutscher Übersetzung vor (unter anderem von J.M.Smith, G.R.Price und D.T. Bishop).

ZUMA-Tagungen 1992

Die Teilnahmegebühren für ZUMA-Workshops reduzieren sich für Studenten und arbeitslose Wissenschaftler auf die Hälfte des jeweils angegebenen Beitrags für Teilnehmer aus den neuen Bundesländern reduzieren sich die Teilnahmegebühren auf den Anteil Ihres Gehalts, gemessen an entsprechenden Positionen in den alten Bundesländern.

*

Workshop: "GAUSS", 15. bis 16. September 1992

Der Workshop soll einen Überblick über die Programmiersprache GAUSS geben. Neben den Grafikmöglichkeiten werden vor allem die Applikationsmodule behandelt. Die Berechnung grundlegender statistischer Kenngrößen wird ebenso besprochen, wie Probit-, Logit- und Loglineare Modelle. Wir widmen uns der Optimierung von Zielfunktionen mittels OPTMUM und MAXLIK sowie der Lösung nichtlinearer Gleichungssysteme. Interessenten werden gebeten, sich bis zum 30.Juni 1992 bei ZUMA, Tagungssekretariat anzumelden. Die Teilnehmerzahl ist auf 20 Personen begrenzt. Für die Teilnahme wird ein Beitrag von 40 Mark erhoben. Der Workshop wird von *Siegfried Gabler* geleitet.

*

Workshop: "Praktische Anwendungen der theoretischen Panelforschung", 13. bis 15. Oktober 1992

Im Mittelpunkt des Workshops stehen diesmal Heuristiken zur Entdeckung von Strukturen in zeitabhängigen Daten und ihre Probleme. Dabei wird einerseits gezeigt, wie man konfirmatorische Verfahren der kausalen Modellierung unter Einsatz neuer Techniken der Modellmodifikation zur Entdeckung von Modellen für zeitabhängige Daten einsetzen kann. Andererseits soll vermittelt werden, wie komplexe exploratorische Techniken (z. B. die multiple Korrespondenzanalyse) und interaktive graphische Verfahren zur Identifikation zeitlicher Strukturen herangezogen werden können. Zur Anwendung kommen u.a. die Computerprogramme EQS, TETRAD und CATEGORIES. Die Teilnehmerzahl ist auf 20 Personen begrenzt. Grundkenntnisse in multivariaten Verfahren werden vorausgesetzt. Die Teilnahmegebühr beträgt 60 Mark. Die Anmeldung wird bis zum 11. September 1992 beim ZUMA Tagungssekretariat erbeten. Referenten sind

Frank Faulbaum und *Michael Wiedenbeck*, die den Workshop auch organisatorisch betreuen.

°

Workshop: "Einführung in die Korrespondenzanalyse" 3. bis 6. November 1992

Gemeinsam bieten das Zentrum für Umfragen, Methoden und Analysen und das Zentralarchiv für empirische Sozialforschung vom 3. bis 6. November 1992 in Mannheim eine Einführung in die Korrespondenzanalyse an. Die Veranstaltung soll Anfängern über Demonstrationen und praktischen Übungen am PC einen Einstieg in dieses Analyseverfahren ermöglichen. An drei der insgesamt vier Veranstaltungstagen wird das von M.J. Greenacre entwickelte Programm SimCA verwendet. Am vierten Tag kommt das an der Universität Leiden entwickelte SPSS Modul "Categories" zum Einsatz. Nach einer kurzen Einführung in den Algorithmus wird anhand inhaltlicher Beispiele die Anwendung der Korrespondenzanalyse diskutiert. Organisiert und betreut wird der Workshop von *Jürgen H.P. Hoffmeyer-Zlotnik* (ZUMA) und *Jörg Blasius* (ZA) unter Mitarbeit von *Herbert Matschinger* (Zentralinstitut für seelische Gesundheit, Mannheim). Der Veranstaltungsort ist Mannheim. Für die Teilnahme wird ein Beitrag von 80 Mark erhoben. Die Teilnehmerzahl ist auf 20 Personen begrenzt. Interessenten werden gebeten, sich bis zum 25. August 1992 bei ZUMA, Tagungssekretariat anzumelden.

°

Workshop: "Einführung in die computerunterstützte Inhaltsanalyse (cui) mit TEXTPACK PC", 10. bis 11. November 1992

Der Workshop hat zum Ziel, Anfängern die Grundsätze der cui zu vermitteln und in das Arbeiten mit TEXTPACK PC einzuführen. Es werden sowohl Vorträge zu Grundproblemen der cui und spezifischen Anwendungen als auch intensive Übungen am PC angeboten. Der Workshop wendet sich ausschließlich an Personen, die noch keine Erfahrung mit der cui haben. Referenten sind *Alfons J. Geis*, *Peter Ph. Mohler* und *Cornelia Züll* (ZUMA). Interessenten werden gebeten, sich bis zum 30. September 1992 bei ZUMA, Tagungssekretariat, anzumelden. Für die Teilnahme wird ein Beitrag von 60 Mark erhoben. Die Teilnehmerzahl ist auf 20 Personen begrenzt.

Workshop: "Amtliche Daten der DDR-Statistik", 26. November 1992

In der DDR wurden von der Zentralverwaltung für Statistik regelmäßige Bevölkerungsbefragungen, wie beispielsweise die Einkommensstichprobe in Arbeiter- und Angestelltenhaushalten oder die Statistik des Haushaltsbudgets, durchgeführt. Da diese Daten wichtige Informationen über die Lebensverhältnisse der DDR-Gesellschaft vor der Umwandlung in ein marktwirtschaftliches System enthalten, sind sie für die Rekonstruktion der damaligen Lebensverhältnisse ebenso wie für die Untersuchung der Transformationsprozesse noch immer von Aktualität. Hinzu kommt, daß für eine Übergangsphase die wichtigsten dieser regelmäßigen Erhebungen in den neuen Bundesländern weiter erhoben wurden, so daß hier eine kontinuierliche Datenbasis zur Verfügung steht. Die in der DDR erhobenen Daten unterlagen bislang jedoch unterschiedlichen Geheimhaltungsstufen, so daß die darin enthaltenen Informationen von der Forschung nur sehr eingeschränkt genutzt werden konnten. Der Workshop wird über die Inhalte und die Validität der wichtigsten amtlichen (und noch verfügbaren) Bevölkerungserhebungen der DDR informieren. Weiterhin wird ein Überblick darüber vermittelt, welche dieser Statistiken fortgeführt beziehungsweise welche in den neuen Bundesländern eingeführt wurden. Darüber hinaus werden Datenzugangsmöglichkeiten dargestellt und erste Ergebnisse präsentiert. Referenten sind Experten der amtlichen Statistik sowie Forscher, die bereits mit diesen Daten arbeiten. Der Workshop wird von *Heike Wirth* betreut. Interessenten werden gebeten, sich bis zum 14. September bei ZUMA, Tagungssekretariat, anzumelden. Für die Teilnahme wird ein Beitrag von 20 Mark erhoben.

*

SOFTSTAT '93
7. Konferenz über die wissenschaftliche Anwendung von Statistik-Software
14. bis 18. März 1993 in Heidelberg

Die SoftStat-Konferenzen wurden 1981 vom Zentrum für Umfragen, Methoden und Analysen (ZUMA) ins Leben gerufen und finden seitdem regelmäßig alle zwei Jahre statt. Sie befassen sich mit dem Einsatz und der Untersuchung von Methoden und Werkzeugen der Informatik für die Statistik, insbesondere mit Neu- und Weiterentwicklungen, Anwendungen, Vergleichen und Bewertungen statistischer Auswertungssysteme einschließlich der mathematischen Verfahren, welche den Software-Realisierungen zugrundeliegen. Darüber hinaus thematisieren die Konferenzen die Rolle von Auswertungssystemen im wissenschaftlichen

Forschungsprozeß, in Ausbildung und Lehre sowie bei der Lösung konkreter Probleme in einzelnen Anwendungsbereichen. Die Konferenzen begnügen sich aber nicht nur mit der Deskription von Softwarelösungen. Vielmehr wollen sie auch über die technologischen Bedingungen (z. B. bestimmte Hardwarebedingungen, Betriebssystem-Umgebungen, Rechnernetze etc.) für die Realisierung konkreter Systeme informieren.

Die SoftStat-Konferenzen sind traditionell interdisziplinär orientiert. Damit soll nicht nur der Erfahrungsaustausch zwischen unterschiedlichen Disziplinen gefördert werden, sondern auch der Beobachtung Rechnung getragen werden, daß originäre Entwicklungen im Bereich der Statistik-Software häufig durch konkrete Probleme in spezifischen Anwendungsbereichen angestoßen werden. Die auf den Konferenzen vertretenen Anwendungsgebiete umfassen Disziplinen wie die Biometrie oder die Biomedizin ebenso wie die Ökonometrie, die empirische Sozialforschung oder die Linguistik.

Die Organisation der SoftStat '93 liegt wieder in den Händen von ZUMA. Unterstützt wird ZUMA auch dieses Mal von der Arbeitsgruppe "Statistische Auswertungssysteme" der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie und der Arbeitsgruppe "Computational Statistics" der Deutschen Region der Internationalen Biometrischen Gesellschaft.

Neben den wissenschaftlichen Vortragsveranstaltungen bietet auch diese Konferenz wieder die Gelegenheit zu praktischen Software-Demonstrationen. Kommerziellen und nichtkommerziellen Software-Anbietern wird darüber hinaus die Möglichkeit geboten, im Rahmen einer begleitenden Ausstellung mit einem fachkundigen Publikum in Kontakt zu treten.

Das wissenschaftliche Programm gliedert sich in folgende Themenbereiche: *Statistische Auswertungssysteme, Wissensbasierte Systeme in der Statistik, Simulation, Versuchsplanung, Interaktive graphische Datenanalyse, Behandlung von Massendaten, Kartographie und Geographische Informationssysteme, Statistik und Linguistische Datenverarbeitung, Computerunterstützte Datenerhebung, Skalierung und Klassifikation, Datenmodelle und Datenbankentwurf, Statistische Arbeitsplatzsysteme und Rechnerkommunikation, Verteiltes Rechnen, Paralleles Rechnen, Statistikausbildung und Statistik-Software, Statistische Algorithmen, weitere Bereiche der Anwendung von Statistik-Software, wie z.B.: Nutzung externer Informationssysteme, Probleme der Statistikberatung, Symbolisches Rechnen, Konnektionistische Anwendungen.*

Die Tagung sieht die folgenden Veranstaltungsformen vor:

(A) *Integrierende Referate* (45 Minuten), (B) *Überblicksdarstellungen* (45-60 Minuten), (C) *Positionsreferate* (40 Minuten), (D) *Einzelreferate zu Themenbereichen des wissenschaftlichen Programms* (20 Minuten), (E) *Poster*, (F) *Demonstrationen von Software-Produkten.*

Arbeitsgruppen
Arbeitsgruppen sehen zu einem Thema aufeinander bezogene Beiträge mehrerer Beteiligter vor. In ihnen sollen relativ kleine Gruppen die Möglichkeit haben, bestimmte Themen- und Problemstellungen nach ihren Wünschen zu diskutieren. Die Art der Diskussion soll der jeweiligen Arbeitsgruppe vorbehalten bleiben. Der Leiter der jeweiligen Arbeitsgruppe koordiniert vor der Tagung die einzelnen Beiträge und legt die Form und den Ablauf der Veranstaltung fest. Für jede Arbeitsgruppe steht ein halber Tag (bis 4 Stunden) zur Verfügung.

Tutorien
Bei Tutorien handelt es sich um Lehrveranstaltungen, bei denen interessierten Teilnehmern bestimmte Fähigkeiten vermittelt werden. Für die Durchführung von Tutorien steht auch noch Freitag, der 18. März 1993, zur Verfügung.

Symposien
Symposien sind Diskussionsveranstaltungen im Rahmen eines größeren Auditoriums zu bestimmten aktuellen Problemstellungen, die eine besondere Organisation und wechselseitige Abstimmung der einzelnen Beiträge erfordern. Zur Verfügung stehen 3 Zeitstunden. Im Regelfall folgt der Ablauf eines Symposiums dem folgenden Schema:
- Eröffnung und Problemaufbereitung durch den Leiter,
- Präsentation der Beiträge durch die angemeldeten Teilnehmer,
- Diskussion unter den Symposiumsteilnehmern,
- Einbeziehung des Auditoriums.

Ausstellungen
Für Ausstellungen kommerzieller und nicht-kommerzieller Software-Anbieter stehen im Veranstaltungsbereich Flächen für Ausstellungsstände zur Verfügung. Ausstellungen sind grundsätzlich kostenpflichtig. Die Kosten richten sich nach Ausstellungsfläche und Ausstattung. Die Konditionen für Aussteller sind in getrenntem Informationsmaterial zusammengestellt, das beim Konferenzsekretariat angefordert werden kann.

Anmeldefristen
Referate: 1. August 1992

Poster und Software-Demonstrationen: 15. Oktober 1992

Eine spätere Anmeldung von Software-Demonstrationen ist nur möglich, wenn noch Raum- und Zeitkapazität vorhanden ist. Da zu diesem Zeitpunkt die Ausstattungsentscheidungen getroffen werden, müssen verspätet angemeldete Demonstrationen mit den gebotenen Möglichkeiten Vorlieb nehmen. Sie können außerdem nicht mehr in das vorläufige Programm aufgenommen werden.

Arbeitsgruppen und Symposien: 15. Oktober 1992

Später angemeldete Arbeitsgruppen und Symposien können im vorläufigen Programm nicht mehr angekündigt werden und können nur noch nach Maßgabe der dann noch vorhandenen Zeit- und Raumkapazität akzeptiert werden.

Tutorien: 15. Oktober 1992

Bei späterer Anmeldung ist ein Versand der Anmeldeunterlagen zusammen mit dem vorläufigen Programm nicht mehr möglich.

Ausstellungen:
Ausstellungen sollten wegen der begrenzten Ausstellungsfläche so früh wie möglich angemeldet werden. Nur bis zum 15. Oktober angemeldete Ausstellungen können im vorläufigen Programm berücksichtigt werden. Letzter möglicher Anmeldetermin ist der 11. Dezember 1992.

Tagungsgebühren

(1)	Nicht-Referenten	DM 200,-- (Überweisung bis zum 31.12.1992)
		DM 260,-- (Überweisung nach dem 31.12.1992)
(2)	für Referenten	DM 170,-- (Überweisung bis zum 31.12.1992)
		DM 260,-- (Überweisung nach dem 31.12.1992)
(3)	für Studenten	DM 50,--

Für Mitglieder der Deutschen Region der Internationalen Biometrischen Gesellschaft und für Mitglieder der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie reduziert sich der Tagungsbeitrag jeweils um DM 30,-.

Berichte über Veranstaltungen

Anwender-Konferenz: "Analyse mit TEXTPACK PC" 18. bis 20. März 1992

Im März fand eine Konferenz statt, auf der TEXTPACK-Anwender - aber auch Nutzer anderer Systeme - über ihre Erfahrungen mit der computerunterstützten Inhaltsanalyse (cui) und dem Programmpaket TEXTPACK berichteten. Die Vorträge und Diskussionen behandelten grundsätzliche Fragen zur Anwendung der computerunterstützten Inhaltsanalyse, aber auch konkrete Anwendungen, und zwar aus den Bereichen der Soziologie, Pädagogik, Psychologie und Psychoanalyse; das Themenspektrum reichte von der Verwendung des Systems in der Lehre über Verschriftungsprobleme und Diktionärskonstruktion bis hin zur Präsentation von Projektergebnissen. Neben kleineren Berichten aus aktuellen Projekten standen sieben Hauptvorträge auf dem Programm. Die großzügige Zeitplanung bot den über 30 Teilnehmern ausreichend Gelegenheit zu Gesprächen und Informationsaustausch. Die Tagungsbeiträge werden in den nächsten Monaten als Buch im Westdeutschen Verlag veröffentlicht.

Die Konferenz wurde vorbereitet und betreut von *Alfons J. Geis, Peter Ph. Mohler* und *Cornelia Züll.*

*

Workshop "Kognitionspsychologie und Fragebogenkonstruktion" 28. bis 30. April 1992

Gegenstand des von 26 Teilnehmern besuchten Workshops waren die kognitiven und kommunikativen Prozesse, die Antworten von Befragten zugrunde liegen. Einer Einführung in die Aufgaben der Befragten aus psychologischer Sicht folgte eine Diskussion der Probleme retrospektiver Verhaltensberichte. Besondere Aufmerksamkeit fanden Möglichkeiten zur Verbesserung der Validität retrospektiver Berichte durch geeignete Fragenführung sowie der Einfluß von Antwortvorgaben. Daran anschließend wurden Probleme der Einstellungsmessung behandelt, wobei insbesondere die Kontextabhängigkeit von Einstellungsurteilen Beachtung fand. Der Workshop schloß mit einem Vergleich unterschiedlicher Modi der Datenerhebung (persönliche, telefonische und schriftliche Befragungen) und ihren Implikationen für die Datenqualität.

Der Workshop wurde von *Hans-J. Hippler* und *Norbert Schwarz* geleitet.

Symposium: "Gewichtung in sozialwissenschaftlichen Umfragen" 5. bis 6. Mai 1992

Im Mai fand bei ZUMA eine Expertenrunde mit Vertretern aus Forschungs- und Erhebungsinstituten, bestehend aus Mathematikern und Sozialwissenschaftlern, zum Thema "Gewichtung in sozialwissenschaftlichen Umfragen" statt. Dieses Symposium ist als ein Einstieg in einen entsprechenden Dialog zwischen universitärer Sozialforschung, kommerzieller Sozial- und Marktforschung und amtlicher Statistik zu sehen. Es wurden unterschiedliche Sichtweisen von Schichtung (bei der Stichprobenziehung) und Gewichtung angesprochen, Stichprobendesigns erläutert und die Notwendigkeit bzw. Fragwürdigkeit einer Gewichtung überhaupt diskutiert und für den Umgang mit sozialwissenschaftlichen Umfragedaten Extrempositionen relativiert. Ein Band mit den Referaten und der aufgearbeiteten Diskussion soll im Laufe dieses Jahres erstellt werden. Für einen Workshop zum Thema "Gewichtung" ist es allerdings, wie dieses Symposium deutlich gezeigt hat, noch zu früh: denn nicht die Technik der Berechnung, sondern Konventionen der Vergleichbarkeit würden solch einen Workshop rechtfertigen. Konventionen der Vergleichbarkeit müssen jedoch noch entwickelt werden. Allerdings ist als ein Ergebnis der Tagung deutlich geworden, daß die gängigen Stichprobendesigns in zentralen Punkten der Schichtung und Gewichtung (-sanweisung) für den Nutzer der Erhebungsdaten nicht ausreichend dokumentiert sind. Aus diesem Grund werden wir versuchen, für eine der nächsten Ausgaben der ZUMA-Nachrichten eine Dokumentation über die Schichtung bei der Ermittlung der sample-points und die Gewichtungsanweisung für die Erhebungsdaten zusammenzustellen.

Organisiert und geleitet wurde das Symposium von *Siegfried Gabler, Jürgen H.P. Hoffmeyer-Zlotnik* und *Dagmar Krebs.*

Gesamtverzeichnis ZUMA-Nachrichten 1 - 29 1977-1992

Nummer 11 (November 1982)

Nummer 12 (Mai 1983)

Nummer 13 (November 1983)

Nummer 14 (Mai 1984)

Nummer 15 (November 1984) *vergriffen*

Nummer 19 (November 1986)

Nummer 20 (Mai 1987)

Nummer 25 (November 1989)

Nummer 26 (Mai 1990) *vergriffen*

Nummer 27 (November 1990)

Nummer 28 (Mai 1991)

Nummer 29 (November 1991)

Durchwahl-Rufnummern (Stand: Juni 1992)

Sie erreichen ZUMA unter der Sammelnummer: (0621) 18004-0. Die Mitarbeiter der Abteilungen ALLBUS und Methodenentwicklung sind über die Sammelnummer: 102360 bzw. 102304 direkt anzuwählen. Die Telefonzentrale ist von Montag bis Donnerstag von 8.30 bis 17.00 und freitags von 8.30 bis 15.30 besetzt. Die mit (S) bezeichneten Mitarbeiterinnen nehmen Sekretariatsaufgaben wahr.

GESCHÄFTSFÜHRUNG

Geschäftsführender Direktor

PD Dr. Peter Ph. Mohler	42
Lisbeth Koch (S)	41
Jolantha Müllner (S)	43
Jutta Flachs (S)	43

Stellv. Geschäftsf. Direktorin

Carol Cassidy	24

Verwaltung

Dipl.-Kfm. Volker Neureither	19
Brigitte Müller	17
Dipl.-Soz. Angelika Stiegler	17
Maria Groh	91

WISSENSCHAFTLICHE LEITUNG

Wissenschaftliche Leiter

Prof. Dr. Ingwer Borg	20
PD Dr. Peter Ph. Mohler	42
PD Dr. Dr. Ulrich Mueller	30
PD Dr. Norbert Schwarz	34

Projektleiter

PD Dr. Frank Faulbaum	32
Dr. Hans-Jürgen Hippler	40
Dr. Jürgen H.P. Hoffmeyer-Zlotnik	44
PD Dr. Dagmar Krebs	45
Dagmar Haas (S)	31

ABTEILUNGEN

Computerabteilung

Carol Cassidy	24
Heiner Ritter	26
Cornelia Züll	26

Feldabteilung

Dipl.-Soz. Rolf Porst	62
Dipl.-Psych. Peter Prüfer	61
Margrit Rexroth, M.A.	64
Dipl.-Soz. Michael Schneid	60
Christa Muhr (S)	65

Abteilung Textanalyse, Medienanalyse, Vercodung

Dr. Peter Schrott	51
Alfons J. Geis, M.A.	54
Ingrid Weickel	50

Statistikabteilung

PD Dr. Siegfried Gabler	14
Dipl.-Math. Michael Wiedenbeck	16

ALLBUS 1023-60/04

Dr. Michael Braun	"
Dr. Wolfgang Bandilla	"
Dr. Janet Harkness	"
Dipl.-Soz. Achim Koch	"
Dipl.-Soz. Reiner Trometer	"
Dipl.-Soz. Martina Wasmer (beurl.)	"
Maria Kreppe-Aygün (S)	"

Mikrodaten

Dr. Georgios Papastefanou	76
Dr. Paul Lüttinger	78
Dipl.-Soz. Bernhard Schimpl-Neimanns	72
Joachim Wackerow	73
Dipl.-Soz. Heike Wirth	79
Rita Haaf (S)	75

Soziale Indikatoren

Dr. Heinz-Herbert Noll	48
Dipl.-Soz. Friedrich Schuster	25
Dipl.-Soz. Stefan Weick (beurl.)	25
Dipl.-Soz. Erich Wiegand	29
Ursula Palm (S)	47

Methodenentwicklung 1023-60/04

Dr. Michael Häder	"
Dipl. Ing. Hartmut Götze	"
Dr. Bernhard Krüger	"
Dr. Sabine Nowossadeck	"

Spezielle Projekte

Dr. Angelika Glöckner-Rist	1023-60/04
Dr. Caroline Kramer	38

Adressenpflege

Ich bin umgezogen. Senden Sie mir die ZUMA-Nachrichten ab sofort bitte an:

Name .....
Vorname .....
Titel .....
Institut .....
.....
Straße .....
PLZ/Ort .....

Falls der Versand an eine andere als die Institutsadresse erfolgen soll:

Straße .....
PLZ/Ort .....

Ich habe noch einen Interessenten für Sie. Senden Sie die ZUMA-Nachrichten bitte an:

Name .....
Vorname .....
Titel .....
Institut .....
.....
Straße .....
PLZ/Ort .....

Noch eine Bitte: Teilen Sie uns bitte mit, wenn Sie die ZUMA-Nachrichten **nicht** mehr zugesandt bekommen wollen.

ZUMA

ISSN 0721-8516