

### Verknüpfung und Generierung von Mikrodaten: dargestellt am Beispiel des integrierten Mikrodatenfiles 1969 für die Bundesrepublik Deutschland

Kortmann, Klaus; Krupp, Hans-Jürgen

Veröffentlichungsversion / Published Version

Sammelwerksbeitrag / collection article

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:

GESIS - Leibniz-Institut für Sozialwissenschaften

#### Empfohlene Zitierung / Suggested Citation:

Kortmann, K., & Krupp, H.-J. (1977). Verknüpfung und Generierung von Mikrodaten: dargestellt am Beispiel des integrierten Mikrodatenfiles 1969 für die Bundesrepublik Deutschland. In P. J. Müller (Hrsg.), *Die Analyse prozeß-  
produzierter Daten* (S. 109-140). Stuttgart: Klett-Cotta. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-325075>

#### Nutzungsbedingungen:

*Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen.*

*Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.*

#### Terms of use:

*This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public.*

*By using this particular document, you accept the above-stated conditions of use.*

## Verknüpfung und Generierung von Mikrodaten

Dargestellt am Beispiel des Integrierten Mikrodatenfiles 1969 für die Bundesrepublik Deutschland

Klaus Kortmann / Hans-Jürgen Krupp

### 1. Perspektiven des Einsatzes integrierter Mikrodatenfile

In den letzten Jahren hat sich die Verwendung von Mikrodaten als eine der wesentlichen Innovationen in den Sozialwissenschaften erwiesen. Diese Entwicklung, die erst durch den Bau leistungsstarker Computer mit großen Speicherkapazitäten ermöglicht wurde, eröffnet insbesondere den empirischen Sozialwissenschaften neue Perspektiven.

Generell erlaubt die Verwendung von Mikrodaten eine Reduzierung des Aggregationsgrades sozialwissenschaftlicher Analysen. Dieses gilt gleichermaßen für einfache statistische Analysen zur Bildung und Überprüfung empirisch gehaltvoller Theorien wie auch für die gesellschaftspolitische Anwendung.

Im Bereich der statistischen Analysen kann nun auf die durch Tabellenprogramme vorgegebenen festen Bevölkerungsgruppen verzichtet werden. Es ist möglich, die Situation von Individuen oder Randgruppen zu untersuchen. Je nach Art der Fragestellung lassen sich diese Gruppen beliebig und wechselnd abgrenzen. So können auch die Besonderheiten von Randgruppen aufgezeigt werden. Ihr Einbezug in einen repräsentativen Mikrodatensatz erlaubt zugleich, ihr quantitatives Gewicht in der Gesellschaft abzuschätzen.

Ein breit gefächertes Mikrodatensatz liefert darüber hinaus detailliertere Informationen bezüglich der untersuchungsrelevanten Merkmale von Personen, Familien, Haushalten oder anderen sozioökonomischen Gruppen, die üblicherweise in höher aggregierten Gruppendaten nicht vorhanden sind. Besonders hervorzuheben ist, daß für beliebige Gruppen nicht nur die Durchschnittswerte,

sondern auch Verteilungs- und Streuungskennzahlen angegeben werden können.

Generell kann man sagen, daß ohnehin viele Verfahren der statistischen Methodenlehre die Existenz von Mikrodaten voraussetzen. Viele der seit Jahren gehüteten Lehrbuchweisheiten können auf diese Art und Weise zum ersten Mal in breitem Umfang angewendet werden.

Damit wächst zugleich die Chance zur Bildung und Überprüfung empirisch gehaltvoller Theorien. Es ist möglich, Verhaltenshypothesen auf einem noch sinnvoll interpretierbaren Aggregationsniveau zu formulieren und zu testen. Die üblichen Verfahren der statistischen Hypothesenüberprüfung können herangezogen werden. Im Prozeß der Hypothesenentstehung kann man die Korrelation zwischen zahlreichen Variablen vorher untersuchen.

Verläßt man schließlich die Bereiche der statistischen Analyse und der Theorienbildung und wendet sich der Anwendung auf politische Fragestellungen zu, sind Mikrodaten erneut die Grundvoraussetzung für eine operable Bewältigung von Entscheidungs- und Politiksystemen. Gerade in den Sozialwissenschaften werden derartige Systeme bis auf weiteres als Simulationssysteme entwickelt werden müssen. Die Qualität der Ergebnisse von Simulationssystemen hängt jedoch im wesentlichen Umfang von der Qualität der Eingabedaten ab.

Der Siegeszug der Mikrodaten wäre freilich nicht möglich gewesen, wenn es bei den inhaltlichen Beschränkungen geblieben wäre, welche in der Erhebung einzelner Stichproben gegeben sind. In allen genannten Bereichen ergaben sich erhebliche Schwierigkeiten solange das theoretisch abdeckbare Spektrum durch den Merkmalsumfang der jeweiligen Stichprobe begrenzt wurde. Da aber Erhebungsgesichtspunkte den Merkmalsumfang von Einzelstichproben begrenzen, ergaben sich hieraus zugleich nennenswerte Begrenzungen der statistischen Auswertungsmöglichkeiten, des theoretisch greifbaren Spektrums wie auch der möglichen Politikmodelle.

Aus diesem Grunde sind in den letzten Jahren zahlreiche Techniken zur Verknüpfung von Mikrodaten entwickelt worden. Auch hierzu war die Existenz leistungsfähiger Computer eine unabdingbare Voraussetzung.

Erst die Erforschung der Integrations- und Verknüpfungstechniken erlaubte die aus der Begrenzung des Merkmalsumfangs sowie aus der oft unvollständigen Erfassung aller Bevölkerungsgruppen resultierenden Restriktionen zu überwinden. Es erwies sich als eine sinnvolle Strategie, verschiedene Stichproben so zusammenzuführen, daß nun auch Aussagen möglich wurden, die auf der Basis der jeweiligen isolierten Stichprobe nicht gemacht werden konnten. Damit ergeben sich zahlreiche Vorteile.

Es ist nun zum Beispiel möglich, Interdependenzen des menschlichen Verhaltens in viel höherem Maße als bisher zu berücksichtigen. Diese Breite erlaubt es dann auch, unterschiedliche Fragestellungen miteinander zu verknüpfen. Theorie- wie Politiksysteme können breit angelegt werden, so daß auch die jeweiligen Folgen, Nebenwirkungen oder Maßnahmen berücksichtigt werden können. Erst die jetzt mögliche inhaltliche Breite läßt die Entwicklung leistungsfähiger Politiksysteme erhoffen.

Schon die wenigen eben genannten Gesichtspunkte machen deutlich, daß sich aus der Verwendung integrierter Mikrodatenfiles neue Perspektiven für die Sozialwissenschaften ergeben. Viele Anwendungen, insbesondere auf politischem Gebiet, haben die Entwicklung integrierter Mikrodatenfiles zur unabdingbaren Voraussetzung. Es lohnt sich daher, der Frage nach der Erzeugung derartiger Datensätze Beachtung zu schenken.

## 2. Arten der Verknüpfung von Mikrodaten

Bei der Entstehung neuer Wissenschaftsgebiete entwickelt sich ein einheitlicher Sprachgebrauch nur zögernd. Aus diesem Grunde kann hier kein allseits akzeptiertes Begriffssystem vorgestellt

werden. Erschwerend kommt hinzu, daß die Probleme der Mikrodatenverknüpfung im deutschen Sprachraum bisher kaum beachtet worden sind. Aber auch im angelsächsischen Sprachraum werden die Begriffe unterschiedlich verwandt und unklar voneinander abgegrenzt.

Die Verknüpfung von Mikrodaten kann sich grundsätzlich auf die logische Zusammenführung von Beobachtungseinheiten oder die logische Zusammenführung von Merkmalen beziehen. Bei der logischen Zusammenführung von Beobachtungseinheiten werden die Ergebnisse unterschiedlicher Stichproben zusammengeführt, weil bestimmte Teile der Bevölkerung, über die man Aussagen machen will, in den isolierten Stichproben jeweils nicht vorhanden sind. Ist man zum Beispiel an Aussagen über die Wohnbevölkerung in Privathaushalten interessiert, hat aber nur eine Stichprobe deutscher Privathaushalte zur Verfügung, ist es notwendig, diese mit Angaben aus einer Ausländerstichprobe zusammenzuführen. Die Verknüpfung mehrerer Stichproben bezieht sich darauf, daß hier die Einzelstichproben jeweils nur einen Teil der interessierenden Gesamtbevölkerung enthalten.

Relativ unproblematisch gestaltet sich die logische Zusammenführung von Beobachtungseinheiten dann, wenn zwar keine der Einzelstichproben für sich die Grundgesamtheit repräsentiert, dies aber beide gemeinsam tun und alle relevanten Merkmale in beiden Stichproben in übereinstimmenden Ausprägungen erhoben werden. In diesem Fall beschränkt sich die Integrationsarbeit im wesentlichen auf eine exakte Hochrechnung der Angaben auf makroökonomische Daten.

Nur selten aber sind diese Bedingungen erfüllt, da unterschiedliche Stichproben in der Regel mit unterschiedlichen Zielsetzungen durchgeführt werden. Ihren Niederschlag findet dies zumeist in der Erhebung unterschiedlicher Merkmale beziehungsweise in ihrer unterschiedlichen Abgrenzung. So werden Einkommensangaben beispielsweise als absolute Brutto- oder Nettowerte, als unterschiedlich stark aggregierte Bruttobeträge sowohl bezüglich der Komponenten (Einkommen aus unselbständiger Arbeit, Einkommen aus Unternehmertätigkeit, etc.) als auch bezüglich der jeweili-

gen Einheit (Haushalt, Personen) erhoben, Nettoeinkommen oft nur nach unterschiedlichen Größenklassen. In diesem Fall entsteht die Notwendigkeit, durch hypothetische Merkmalskorrektur die unterschiedlichen Einkommensangaben auf ein einheitliches Niveau zu bringen, das den aus den Untersuchungszielen resultierenden Anforderungen Genüge trägt.

In nicht seltenen Fällen liegen für bestimmte Bevölkerungsgruppen - in der Regel Randgruppen - keine direkt verwertbaren Mikrodaten vor, d.h., die insgesamt zur Verfügung stehenden Einzelstichproben repräsentieren nicht die Grundgesamtheit. In diesen Fällen erweist es sich als erforderlich, über die skizzierte hypothetische Merkmalskorrektur hinaus Beobachtungseinheiten synthetisch zu generieren. Für eine derartige Generierung werden Hypothesen benötigt, die empirisch fundiert sein sollten. Mindestvoraussetzung zur Erfüllung dieser Forderung sind Informationen aus Gruppendaten, die zumindest die Bestimmung der Randverteilungen erlauben. Diese Informationen können freilich durchaus aus verschiedenen Quellen stammen.

Die verschiedenen Arten der Verknüpfung bei der logischen Zusammenführung von Beobachtungseinheiten, die hier als integrative Verknüpfung bezeichnet werden soll, sind in Übersicht 1 nochmals gegenübergestellt.

Hiervon zu unterscheiden ist die logische Zusammenführung von Merkmalen. Ausgangssituation ist in diesem Fall entweder eine für die Grundgesamtheit repräsentative Stichprobe oder ein Datenfile, das nach der logischen Zusammenführung von Beobachtungseinheiten nunmehr als repräsentativ anzusehen ist.

Notwendig wird eine logische Zusammenführung der Merkmale dieser Datenbasis mit einer (oder mehreren) weiteren Stichprobe(n), wenn ein begrenzter Satz von Merkmalen ein Erreichen der angestrebten Forschungsziele unmöglich macht. Voraussetzung ist, daß die herangezogenen Datenbasen Merkmale für im Prinzip die gleiche Grundgesamtheit enthalten.

Übersicht 1

Die logische Zusammenführung von Beobachtungseinheiten

Tatbestand	Art der Verknüpfung
<p>Die Grundgesamtheit ist vollständig, wenn auch unter Umständen mit unterschiedlichen Auswahlätzen in der Stichprobe repräsentiert.</p>	<p>Keine Verknüpfung notwendig</p>
<p>Keine der Einzelstichproben ist repräsentativ für die Grundgesamtheit, zusammengeführt erlauben sie aber eine Repräsentation der Grundgesamtheit. Die für den Forschungszweck relevanten Merkmale sind in allen Stichproben in übereinstimmenden Abgrenzungen erhoben.</p>	<p>Integrative Verknüpfung im engeren Sinne; Stichprobenkumulation</p>
<p>Keine der Einzelstichproben ist repräsentativ für die Grundgesamtheit, zusammengeführt erlauben sie aber eine Repräsentation der Grundgesamtheit. Die in den einzelnen Stichproben für den Forschungszweck relevanten Merkmale sind unterschiedlich abgegrenzt.</p>	<p>Stichprobenkumulation mit hypothetischer Merkmalskorrektur</p>
<p>Es stehen nicht genügend Stichproben zur Verfügung, um die Gruppengesamtheit vollständig zu repräsentieren.</p>	<p>Synthetisch-hypothetische Generierung von Beobachtungseinheiten</p>

In der Praxis wird in beiden Fällen in der Regel so vorgegangen, daß eine der Stichproben als Ausgangsstichprobe benutzt wird, die dann durch Angaben aus anderen Stichproben ergänzt wird. Bei der logischen Zusammenführung von Beobachtungseinheiten werden zusätzliche Beobachtungseinheiten in die Basisstichprobe eingespielt. Bei der logischen Zusammenführung von Merkmalen werden für die Beobachtungseinheiten der Basisstichprobe zusätzliche Merkmale aus anderen Stichproben übernommen.

Je nach der Art und Qualität der zur Verfügung stehenden Stichproben können zur Merkmalszusammenführung exakte oder statistische Verknüpfungstechniken herangezogen werden. Während die exakte Verknüpfung eine eindeutige Zuordnung von Einheiten aus verschiedenen Stichproben voraussetzt, werden bei der statistischen Verknüpfung Merkmale von nicht identischen Beobachtungseinheiten zusammengeführt, deren Merkmalsausprägungen aber in einer Reihe von wichtigen Größen übereinstimmen.

Sind derartige Merkmale in einer oder der anderen Stichprobe unsicher, fehlerbehaftet oder gar nicht vorhanden, so ist vorab eine hypothetische Korrektur beziehungsweise eine synthetische Generierung von Merkmalen erforderlich.

Übersicht 2 stellt die verschiedenen Verknüpfungsarten für die logische Zusammenführung von Merkmalen nochmals gegenüber.

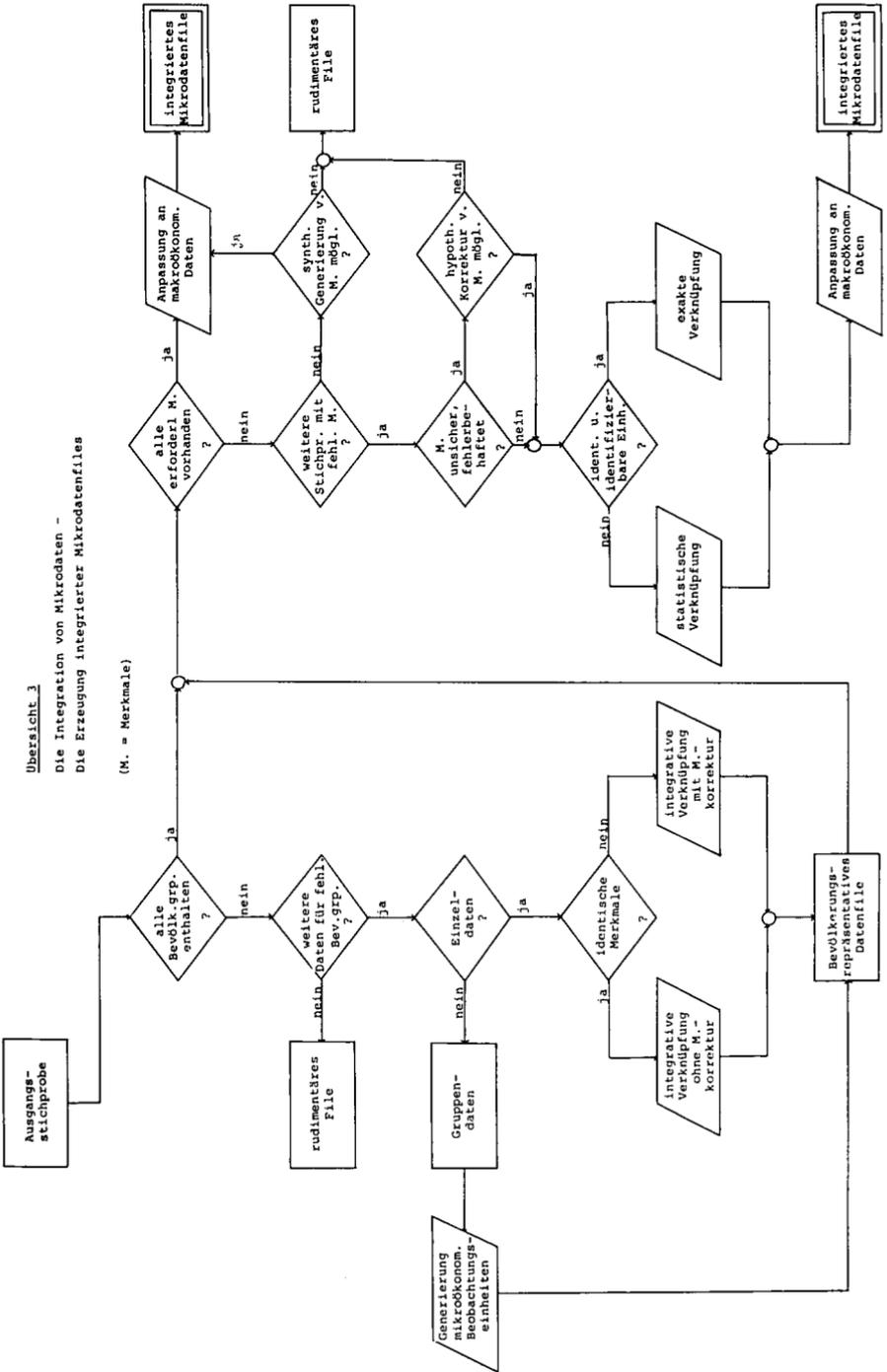
Das Ergebnis sowohl der logischen Zusammenführung von Beobachtungseinheiten wie auch der anschließenden logischen Zusammenführung von Merkmalen wird hier als integriertes Mikrodatenfile bezeichnet.

Die Übersicht 3 zeigt in Form eines Flußdiagrammes nochmals die einzelnen Generierungsschritte auf. Der linke Teil beschreibt die Zusammenführung der Beobachtungseinheiten, der rechte die Zusammenführung der Merkmale.

## Übersicht 2

### Die logische Zusammenführung von Merkmalen

Tatbestand	Art der Verknüpfung
Die für den Forschungszweck relevanten Merkmale sind in einer Stichprobe vollständig erhoben.	Keine Verknüpfung notwendig
Die für den Forschungszweck relevanten Merkmale sind nicht in einer einzelnen Stichprobe vorhanden. Es existieren aber Stichproben mit identischen und identifizierbaren Beobachtungseinheiten, die insgesamt alle relevanten Merkmale enthalten.	Exakte Verknüpfung
Die für den Forschungszweck relevanten Merkmale sind nicht in einer einzelnen Stichprobe enthalten. Es existieren jedoch Stichproben, die alle relevanten Merkmale enthalten. Die Beobachtungseinheiten der verschiedenen Stichproben sind freilich entweder verschieden oder zumindest nicht identifizierbar.	Statistische Verknüpfung
Einige Merkmale sind in einigen Stichproben unsicher, fehlerbehaftet oder gar nicht vorhanden.	Hypothetische Korrektur bzw. synthetisch-hypothetische Generierung von Merkmalen.



### 3. Ein Beispiel: Das Integrierte Mikrodatenfile für die Bundesrepublik Deutschland 1969 (IMDAF 1969)

In diesem Abschnitt soll nun versucht werden, die bisher weitgehend abstrakten Darlegungen zu veranschaulichen. Hierzu werden am Beispiel der Generierung des Integrierten Mikrodatenfiles für Bundesrepublik Deutschland 1969 (IMDAF 1969) die angewandten Verknüpfungstechniken erläutert.

#### 3.1 Die Ausgangssituation

##### 3.1.1 Forschungsziele und hieraus resultierende Anforderungen

Die Erstellung des IMDAF 1969 erfolgte im Rahmen der Arbeiten am Sozialpolitischen Entscheidungs- und Indikatorensystem für die Bundesrepublik Deutschland (SPES-Projekt). Im Entscheidungssystem werden Wirkungen, Nebenfolgen und Interdependenzen spezifischer gesellschaftspolitischer Maßnahmen untersucht<sup>1)</sup>. Im Mittelpunkt steht die Entwicklung eines für die Bundesrepublik repräsentativen Simulationssystems für kurzfristig abrufbare Alternativrechnungen in neun Teilbereichen der Sozial- und Gesellschaftspolitik. Das Entscheidungssystem verknüpft diese Bereiche untereinander und mit makroökonomischen Simulationsprozessen. Im Indikatorensystem werden die Ziele, auf die sich politische Maßnahmen ausrichten, und das Ausmaß der Zielrealisierung untersucht. Definiert und berechnet werden soziale Indikatoren, die den gesellschaftlichen Istzustand beschreiben sollen<sup>2)</sup>. Die Realisierung beider Teile, sowohl des Entscheidungs- wie auch des Indikatorensystems, setzen das Vorhandensein einer geschlossenen Mikrodatenbasis voraus, die alle Bevölkerungsgruppen umfaßt.

Im Rahmen dieses Berichtes ist es weder möglich noch sinnvoll, die ganze Breite des geplanten Merkmalsumfangs darzustellen. Wir wollen uns stattdessen auf den Sektor der Einkommensverteilung beschränken. Dieses entspricht auch dem bisherigen Arbeitsablauf, für den Fragen der Einkommensverteilung von zentraler Bedeutung waren.

Allein die Analyse des Indikatorensystems auf dem Gebiet der

Einkommensverteilung zeigt den erheblichen Datenbedarf, wenn man die unterschiedlichen Dimensionen der Einkommensverteilung berücksichtigen will. Dieses sind:

- a. Die Ungleichmäßigkeit der Bedarfsdeckungsmöglichkeiten
- b. Die Leistungsbezogenheit der Einkommen
- c. Die Armut in der Einkommensdimension
- d. Die Stetigkeit und Sicherheit des Einkommensstromes<sup>3)</sup>

Die Anforderungen, die das Entscheidungssystem an die Mikrodatenbasis stellt, haben zwei Aspekte. Zum einen dient sie zur Gewinnung und zum Testen von Hypothesen zum Beispiel bezüglich der Einkommensverteilung und somit zur Entwicklung eines Moduls zur Erklärung und Fortschreibung der personellen Einkommensverteilung, zum anderen bildet sie die Ausgangsdatenbasis für das Mikrosimulationssystem, mit dessen Hilfe die Konsequenzen gesellschaftspolitischen Handelns für einzelne Bevölkerungsgruppen untersucht werden sollen.

Grundvoraussetzung zur Erfüllung dieser Forschungsziele ist die Organisation von Daten auf der Mikroebene, d.h. auf der Ebene von Personen und Haushalten, die dann beliebig zu größeren Gruppen zusammengefaßt werden können. Als nahezu ideal zur Erreichung der skizzierten Ziele ist ein Datensatz anzusehen, der

1. für alle Bevölkerungsgruppen (insbesondere der Randgruppen) detaillierte demographische Angaben liefert, die es erlauben, sowohl haushalts- wie auch personenspezifische Auswertungen vorzunehmen. Ferner sind
2. differenzierte Angaben der funktionellen Einkommenskomponenten auf Personenebene sowie
3. Brutto- und Nettoeinkommensangaben unabdingbar notwendig.
4. Anzustreben ist fernerhin eine Kompatibilität der Ergebnisse mit der Volkswirtschaftlichen Gesamtrechnung sowie der Einkommensteuerstatistik, Schließlich sind
5. die Bevölkerungsgruppen auf die Ergebnisse der Volkszählungen hochzurechnen<sup>4)</sup>.

### 3.1.2 Zur Qualität vorhandener Primärstatistiken

Eine Primärstatistik, die allen diesen Anforderungen entspricht, liegt für die Bundesrepublik nicht vor.

Zwar stehen im Rahmen des SPES-Projektes mit den Einkommens- und Verbrauchsstichproben (EVS) relativ leistungsstarke Primärstatistiken für die Jahre 1962, 1969 (die folgenden Aussagen beziehen sich auf die EVS 1969) zur Verfügung, doch umfassen sie lediglich die deutschen Privathaushalte außerhalb des Anstaltsbereiches. Auch sind die Angaben für Bezieher hoher Einkommen (über 100.000 DM p.a.) wenig zuverlässig. Bezieher von Einkommen über 250.000 DM p.a. enthält die Stichprobe überhaupt nicht. Weitere wichtige Bevölkerungsgruppen wie die Ausländer und die Wohnbevölkerung im Anstaltsbereich wurden von vornherein aus der Erhebung ausgeschlossen<sup>5)</sup>.

Zudem waren seitens des Statistischen Bundesamtes die drei Teile der EVS 1969, nämlich das Grund-, Haupt- und Schlußinterview, die jeweils unterschiedliche Themenschwerpunkte beinhalten, nicht zusammengeführt worden.

Während die EVS den Anforderungen sowohl bezüglich der demographischen wie auch der Einkommensangaben der durch sie erfaßten Bevölkerungsgruppen in vieler, wenn auch nicht in jeder Hinsicht genügt, sind die statistischen Informationen bezüglich der nicht in der EVS eingeschlossenen Bevölkerungsgruppen wesentlich lückenhafter.

Für Ausländer und die Wohnbevölkerung im Anstaltsbereich (Anstaltsinsassen und Personen in Privathaushalten im Anstaltsbereich) stehen zwar in den Mikrozensen hinreichende demographische Angaben zur Verfügung, die in etwa denen in der EVS entsprechen. Das Einkommen wird demgegenüber nur als Größenklasse des Nettoeinkommens erfaßt. Weitere Informationen über Ausländer liefert zwar die Ausländererhebung der Bundesanstalt für Arbeit, die uns allerdings aus Gründen des Datenschutzes bisher nicht zugänglich gemacht wurde. Über die Wohnbevölkerung im Anstalts-

bereich liegen über die Angaben im Mikrozensus hinaus keinerlei Informationen vor.

Für die Bezieher hoher und höchster Einkommen existieren lediglich die stark aggregierten Angaben der Einkommensteuerstatistik.

Die Erstellung einer geschlossenen Datenbasis für die Bundesrepublik Deutschland verlangt demzufolge einerseits eine logische Zusammenführung von Merkmalen, im konkreten Fall der drei Interviewteile der EVS mit Hilfe exakter und statistischer Verknüpfungsverfahren. Im Anschluß daran ist eine logische Zusammenführung von Beobachtungseinheiten erforderlich, also die Zusammenführung der Angaben über deutsche Privathaushalte außerhalb des Anstaltsbereiches aus der EVS mit den Angaben über Ausländer und die Wohnbevölkerung innerhalb des Anstaltsbereiches aus dem Mikrozensus. Diese Zusammenführung macht zudem eine hypothetische Merkmalskorrektur notwendig, d.h. aus den mit Sicherheit fehlerbehafteten Angaben der Nettoeinkommensgrößenklasse muß das Bruttoeinkommen in absoluten Beträgen ermittelt werden.

Ein synthetische Generierung von Einheiten muß schließlich im Bereich hoher Einkommen durchgeführt werden, da über diese Gruppen keine primärstatistischen Informationen vorliegen.

Der Prozeß der Generierung des Integrierten Mikrodatenfiles soll im folgenden kurz geschildert werden, wobei es besonders darauf ankommt, die typischen Charakteristika der einzelnen Verknüpfungsarten näher aufzuzeigen.

### 3.2 Die logische Zusammenführung von Merkmalen

#### 3.2.1 Exakte Verknüpfung

Unter exakter, direkter oder personeller Verknüpfung versteht man die Verknüpfung von Angaben über identische Personen aus zwei oder mehr statistischen Quellen<sup>6)</sup>.

Dieses Verfahren, das u.a. bereits in amerikanischen Projekten<sup>7)</sup> angewendet wurde, setzt voraus, daß in allen zu verknüpfenden Stichproben ein eindeutiges Merkmal - zum Beispiel eine Personen- oder Haushaltsnummer - vorhanden ist, das eine zweifelsfreie Zuordnung erlaubt.

Eine derartige Größe stand in Form der Haushaltsnummer prinzipiell zur Verknüpfung aller drei EVS-Teile (Grund-, Haupt- und Schlußinterview) zur Verfügung<sup>8)</sup>.

Diese grundsätzlich simple und gleichzeitig zuverlässige Methode stößt dort auf ihre Grenzen, wo einfache Verkodungsfehler die eindeutige Zuordnung von Erhebungen unmöglich machen. Sofern es nicht möglich ist, diese Fehler manuell zu beheben (wenn man etwa beim Vergleich einzelner Datenteile offensichtliche Verdrehungsfehler entdeckt), ist ein exaktes Verknüpfen nicht mehr möglich. Das gilt auch dann, wenn wie im Falle der EVS etwa im Verlauf des Erhebungsjahres Haushalte aus einer Untersuchung ausscheiden, so daß nicht alle Interviewteile zur Verfügung stehen.

In derartigen Fällen steht man vor der Alternative, entweder nur diejenigen Haushalte zu berücksichtigen, für die vollständige Informationen vorliegen oder aber synthetische Verknüpfungsverfahren anzuwenden. Da einige Gründe dafür sprechen, daß insbesondere ein Ausscheiden aus der Erhebung nicht ausschließlich zufallsbedingt, sondern vielmehr korreliert mit demographischen und finanziellen Merkmalen ist, würde ein Verzicht auf Haushalte mit unvollständiger Information einen Bias in das Datenfile einbringen, dessen Ausmaß allerdings nur schwer abzuschätzen ist.

Deshalb ist eine statistische Verknüpfung - die zweifellos mit Fehlern behaftet ist - einem Verzicht auf Information vorzuziehen.

Im Falle der Verknüpfung der verschiedenen Teile der Einkommens- und Verbrauchsstichprobe war es aber immerhin in 96.5 % aller Fälle möglich, dem zentralen Hauptinterview der EVS, das alle wesentlichen Einkommensangaben enthält, ein Grundinterview und sogar in 98.5 % aller Fälle ein Schlußinterview exakt zuzuordnen. Alle übrigen Fälle wurden statistisch verknüpft.

### 3.2.2 Statistische Verknüpfung

Unter statistischer Verknüpfung versteht man eine Reihe von Verfahren, die dazu entwickelt wurden, aus verschiedenen Quellen stammende Daten vergleichbarer Einheiten, die aber der gleichen Grundgesamtheit entstammen, miteinander zu verknüpfen. Während bei der exakten Verknüpfung nur Daten identischer Einheiten (Personen, Haushalte) zusammengeführt werden, setzt die statistische Verknüpfung lediglich eine Übereinstimmung in wichtigen, untersuchungsrelevanten Merkmalen voraus. Bedingung hierzu ist, daß zumindest eine Teilmenge von Merkmalen in beiden herangezogenen Erhebungen gemeinsam vorhanden ist.

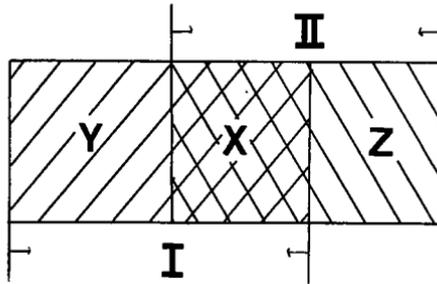
D.h. (vgl. Schaubild I), die Teilmenge  $x$  der Merkmale wird sowohl in Stichprobe I wie II erfaßt, wohingegen die Merkmalsgruppen  $Y$  und  $Z$  in nur jeweils einer Erhebung enthalten sind. Notwendige Bedingung für ein Zusammenführen zweier Einheiten ist nun eine möglichst gute Übereinstimmung der Merkmalsausprägungen der Merkmale der Gruppe  $X$ . Zur Feststellung der Güte der Übereinstimmung wurden verschiedene Verfahren erarbeitet, die im folgenden vorgestellt werden.

Angewandte statistische Verknüpfungstechniken<sup>9)</sup> werden weitgehend bestimmt von der Art der zur Verfügung stehenden Daten sowie von den Zielen, die man bei der Erstellung des Datenfiles verfolgt. Da zwei Projekte selten in ihren Zielen kongruent sind, und sie oft auch auf unterschiedliche Ausgangsdaten zurückgreifen, hat sich bis heute noch kein Verfahren durchgesetzt, wenn auch gewisse gemeinsame Grundzüge vorhanden sind.

Der allen Verfahren gemeinsame erste Schritt des statistischen Verknüpfungsvorganges besteht in der Gruppierung der Einheiten (Haushalte, Familien, Personen, Steuerfälle etc.) anhand von Ausprägungen von den beiden Stichproben gemeinsam vorhandenen Merkmalen, d.h., die Einheiten beider Stichproben werden in bestimmte Zellen sortiert.

Schaubild I

Merkmalsgruppen bei der statistischen Verknüpfung von Stichproben



Von den aus der Menge der X-Merkmale des obigen Schaubilds hierzu ausgewählten Gruppierungsmerkmalen wird verlangt, daß sie einen möglichst starken Bezug zumindest zu den Variablen der Gruppe Y und Z haben, denen das primäre Forschungsziel gilt. Die Enge des Zusammenhanges und damit die konkrete Auswahl erfolgt entweder wie im Falle Okners<sup>10)</sup> und Budds<sup>11)</sup> aufgrund von Plausibilitätsüberlegungen ohne nähere statistische Untersuchung. Allerdings ist in beiden Vorhaben die Anzahl gemeinsamer Variablen ohnehin stark eingeschränkt, so daß sich eine Vorauswahl weitgehend erübrigt.

Demgegenüber bestimmt Alter<sup>12)</sup> die Erklärungskraft der einzelnen X-Merkmale, die er als relativen Anteil der Variablen am Bestimmtheitsmaß von Testgleichungen definiert, und verwendet diejenigen Variablen mit der höchsten Erklärungskraft als Gruppierungsgrößen. Einen ähnlichen Weg beschreiten Nancy und Richard Ruggles<sup>13)</sup> zur Auswahl der Verknüpfungsmerkmale und ihrer Merkmalsausprägungen. Mit Hilfe des CHI-Quadrat-Tests stellen sie zunächst fest, ob eine bestimmte Differenzierung einer X-Variablen zu einer identischen Verteilung der Einheiten über Ausprägungen von Y- und Z-Variablen führt. Ist dies der Fall (hoher Wert für CHI-Quadrat), so ist die untersuchte Differenzierung der X-Variablen insignifikant. Ergeben sich signifikante Differenzen, so wird ihre Stärke mit einem Regres-

sionsansatz überprüft. Erst wenn ein niedriger Korrelationskoeffizient die mittels des CHI-Quadrat-Test festgestellte unterschiedliche Verteilung der Einheiten statistisch untermauert, wird die X-Variable in der getesteten Differenzierung als Gruppierungsmerkmal herangezogen.

Verdeutlicht werden soll dieses Vorgehen anhand eines von Ruggles beschriebenen Anwendungsfalles (vgl. Tabelle 1). Untersucht wird, ob eine Differenzierung zwischen Arbeitnehmern im Privaten und Öffentlichen Sektor (X-Variable) relevant ist. Der Test erfolgt anhand der Verteilung der Arbeitnehmer über die Familiengröße (Y-Variable). Der niedrige CHI-Quadrat-Wert deutet auf eine gleichförmige Verteilung hin. Da das hohe Bestimmtheitsmaß dies bestätigt, wird auf die ins Auge gefaßte Disaggregation verzichtet.

Tabelle 1  
Auswahltest für Gruppierungs-  
merkmale nach Ruggles

DISTRIBUTION OF FAMILY SIZE FOR PRIVATE AND GOVERNMENT  
EMPLOYEES

y variable Size of Family (Number of Persons)	x variable: Class of Worker			
	Private Company Employee		Government Employee	
	Number of Observations	Percent	Number of Observations	Percent
1	869	12.4	186	13.6
2	1394	19.9	279	20.4
3	2075	29.6	439	32.1
4	1445	20.6	288	21.1
5	728	10.4	115	8.4
6	289	4.1	38	2.8
7	124	1.8	13	1.0
8	50	0.7	6	0.4
9 or more	17	0.2	3	0.2
Total cases	8537	100.0	1707	100.0

Comparison between distributions:

Chi Square Probability 0.9536 (based on distributions of number of observations)

Correlation Coefficient 0.9966 (based on percentage distributions)

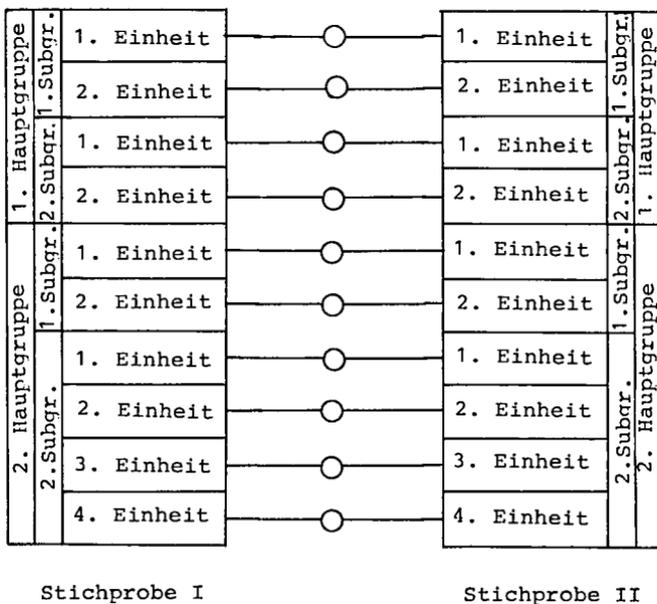
Quelle: Ruggles, Nancy und Richard, A Strategy for Merging and Matching Microdata Sets, in: Annals of Economic and Social Measurement 3/2 1974, S. 353-371, S.S. 363

Im zweiten und unter Umständen weiteren Schritten werden die Hauptgruppen zunächst weiter untergliedert, sofern es aufgrund der Anzahl gemeinsamer Variablen und Stichprobengröße möglich ist.

Innerhalb dieser Subgruppen kann entweder eine Reihung der Einheiten (etwa nach der Höhe der absoluten oder relativen Einkommen) erfolgen<sup>14)</sup>, woran anschließend diejenigen Einheiten beider Stichproben miteinander verknüpft werden, die den gleichen Rang innerhalb einer Gruppe haben (vgl. Schaubild II).

Schaubild II

Zuordnung von Einheiten nach Rangnummern innerhalb von Subklassen



Eine weitere Möglichkeit der Zuordnung innerhalb der Subgruppen besteht darin, solche Einheiten einander zuzuordnen, die hinsichtlich weiterer Merkmale eine möglichst hohe Übereinstimmung haben.

Hierzu ist es möglich, die Merkmale unterschiedlich mit Konsistenzpunkten zu gewichten, wobei letztlich die Gewichte nach subjektiven Gesichtspunkten gesetzt werden. Die Erreichung einer Mindestpunktzahl wird dann als Matchbedingung verlangt. Diese Mindestpunktzahl kann sukzessive gesenkt werden, sofern unter Zugrundelegung strengerer Maßstäbe keine entsprechenden Haushalte gefunden werden können. Verknüpft werden können dann entweder diejenigen Einheiten, die die höchste Punktzahl erzielen oder diejenigen, die im Suchprozeß als erste die Mindestpunktzahl erreichen<sup>15)</sup> oder aber aus der Menge von Einheiten, die über der Mindestpunktzahl liegen, wird eine durch einen Zufallsprozeß ausgewählt<sup>16)</sup>.

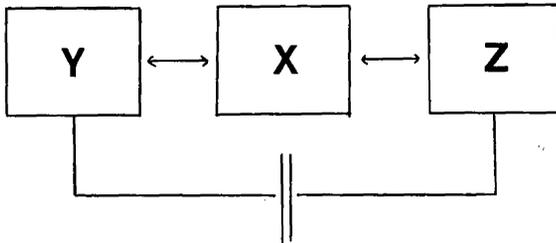
Eine dritte Möglichkeit besteht schließlich darin, innerhalb der Subgruppen nach einem reinen Zufallsverfahren zu verknüpfen. Dieses Verfahren kann dann zu akzeptablen Ergebnissen führen, wenn die Subgruppen stark disaggregiert und sehr dünn besetzt sind.

Allen oben beschriebenen Verfahren gemeinsam ist, daß für den Verknüpfungsvorgang zwar die Beziehungen zwischen den Variablen Y und X einerseits sowie X und Z andererseits berücksichtigt werden, demgegenüber aber die Beziehungen zwischen Y und Z außer acht gelassen sind, d.h., es wird für alle in einer Gruppierungszelle zusammengefaßten Einheiten eine konstante Beziehung zwischen den Variablen der Gruppe Y und Z unterstellt (vgl. Schaubild III).

Eine derartige konstante Beziehung ist aber ernsthaft in Zweifel zu ziehen. Insbesondere Sims<sup>17)</sup> erhob daher in mehreren Kommentaren zu den bereits erwähnten Arbeiten die Forderung, die Beziehungen zwischen Y und Z zu berücksichtigen. Ansetzen

Schaubild III

Berücksichtigung von Interdependenzen zwischen Merkmalsgruppen bei der statistischen Verknüpfung



kann eine solche Forderung auf der Ebene der Subgruppen, d.h. bei der endgültigen Zusammenführung von Einheiten, die bislang mit Hilfe von Rangnummern oder Konsistenzpunkten durchgeführt wurde.

Scheitern dürfte ihre Realisierung allerdings bereits daran, daß es in aller Regel keine externen Daten gibt, mit deren Hilfe sich die Beziehungen zwischen Y und Z schätzen lassen, denn nicht zuletzt das Fehlen derartiger Informationen macht die Verknüpfung ja erst erforderlich<sup>18)</sup>. Wäre eine solche Schätzung dennoch möglich, so würden andererseits bereits in der Entstehungsphase des Datenfiles Beziehungen festgeschrieben, die spätere Benutzer, die die Entstehungsgeschichte des Files nicht kennen, unter Umständen als Forschungsergebnisse erneut aus dem Datenbestand ableiten würden.

Statistische Verknüpfungsverfahren sind die am weitesten verbreiteten Verfahren zur logischen Zusammenführung von Merkmalen, da es aus erhebungstechnischen Gründen nur selten möglich ist, identische Einheiten zu befragen. Geschieht dies doch, so werden die Ergebnisse aus Gründen des Datenschutzes in aller Regel so anonymisiert, daß ein nachträgliches exaktes Zusammenführen von Merkmalen letzten Endes doch nicht mehr möglich ist.

Im Rahmen der Arbeiten am IMDAF 1969 spielten diese Verfahren zur logischen Zusammenführung von Merkmalen allerdings eine weniger bedeutende Rolle, da die Zusammenführung der verschiedenen EVS-Interviews in der überwiegenden Anzahl der Fälle aufgrund der Haushaltsnummer exakt erfolgen konnte.

Dies galt nicht für 3.5 % aller Grund- und für 1.5 % aller Schlußinterviews. Erleichtert wurde die Aufgabe der statistischen Verknüpfung dieser Fälle durch die Tatsache, daß diese Teilinterviews von identischen Einheiten gegeben wurden, die lediglich im Nachhinein nicht mehr unmittelbar zu identifizieren waren. Demgegenüber bestand in allen oben beschriebenen Projekten das Problem, daß die Interviews nicht von identischen Einheiten gegeben wurden. Aus diesem Grunde konnte man im Fall der EVS bei einer Übereinstimmung einer Reihe von demographischen Merkmalen davon ausgehen, daß man tatsächlich Merkmale identischer Einheiten miteinander verknüpft hatte.

Zu unterscheiden waren zwei Fälle<sup>19)</sup>.

In der ersten Gruppe standen für jeweils ein Hauptinterview mehrere Grund- beziehungsweise Schlußinterviews mit der gleichen Haushaltsnummer zur Verfügung. Diese Identität war ausschließlich auf Vercodungsfehler zurückzuführen. Durch einen Vergleich der demographischen Merkmale zwischen den Haushalten mit gleicher Nummer war es dann unproblematisch, jeweils identische Einheiten zu finden und deren Merkmale zu verknüpfen.

In lediglich 2.1 % bzw. 0.5 % war es auch mit diesem Verfahren nicht möglich, dem Hauptinterview ein Grund- bzw. ein Schlußinterview zuzuordnen. Gemäß der sozialen Stellung des Haushaltsvorstandes (6 Klassen), der Größenklasse des Haushaltsnettoeinkommens (11 Klassen) und der Haushaltsgröße (5 Klassen) wurden die verbleibenden Einheiten in insgesamt 330 Zellen gruppiert. Innerhalb dieser Zellen wurden dann diejenigen Einheiten miteinander verknüpft, die hinsichtlich

aller weiteren gemeinsamen Merkmale möglichst gut übereinstimmen, wobei jede Übereinstimmung gleich gewichtet wurde. Erleichtert und begünstigt wurde der Auswahlprozeß dadurch, daß die Anzahl der insgesamt für die Zuordnung zur Verfügung stehenden Grundinterviews sechsmal höher war als die benötigte und die des Schlußinterviews etwa viermal über der Anzahl der zu vervollständigenden Hauptinterviews lag. In Tabelle 2 sind diese Angaben noch einmal gegenübergestellt.

Tabelle 2<sup>20)</sup>

Zuordnung von Grund- und Schlußinterview zum Hauptinterview der EVS 1969

Kriterium der Zuordnung zum Hauptinterview	Grundinterview	Schlußinterview
1. Eindeutige Zuordnung nach Nummern	45.748	46.787
2. Doppelte Nummern, Zuordnung nach sozialen Merkmalen	628	355
3. Zuordnung nach der besten Übereinstimmung der soz. Angaben, Korrektur von Drehfehlern etc.	1.007	241
4. Zwischensumme	47.383	47.383
5. Überschuß	5.368	725
6. Summe	52.751	48.108

Auf dem Hintergrund, daß alle Interviews von identischen Einheiten gegeben wurden, garantiert dieses Verfahren eine annähernd optimale statistische Verknüpfung von Merkmalen.

#### 4. Die logische Zusammenführung von Beobachtungseinheiten

##### 4.1 Integrative Verknüpfung von Mikrodaten

Während die bisher beschriebenen Verfahren der exakten und statistischen Verknüpfung von Mikrodaten zur logischen Zusammenführung von Merkmalen dienen, bezeichnet die integrative Verknüpfung die Möglichkeit der logischen Zusammenführung von Beobachtungseinheiten, d.h. der Kumulation von Stichproben. Die Notwendigkeit hierzu trat bei der Konstruktion des IMDAF 1969 auf, da weder die Wohnbevölkerung im Anstaltsbereich noch die Ausländer in der EVS erfaßt sind. Uns zugängliche Informationen bezüglich dieser Gruppen standen lediglich im Mikrozensus zur Verfügung. Außerdem fehlten die Bezieher sehr hoher Einkommen.

Die Probleme bei der integrativen Verknüpfung von Informationen liegen deshalb weniger darin, in zwei Stichproben entsprechende Einheiten zu finden, die man hätte verknüpfen können, sondern insbesondere vielmehr darin, die unterschiedlichen Variablenabgrenzungen und -ausprägungen in zwei Stichproben zu beseitigen. Die schwierigste Aufgabe im Rahmen der Arbeiten am IMDAF 1969 erwuchs daraus, daß die EVS zwar ein breites Spektrum an Bruttoeinkommenskomponenten sowie das Nettoeinkommen in absoluten Beträgen liefert, der Mikrozensus demgegenüber aber lediglich des Nettoeinkommens in sieben Größenklassen erfragt. Da wesentliche Dimensionen der Einkommensverteilung (etwa die Leistungsbezogenheit der Einkommen) nur dann untersucht werden können, sofern Bruttoeinkommenskomponenten vorliegen, bestand die unabdingbare Notwendigkeit, die Angaben der Nettoeinkommensgrößenklasse mittels hypothetischer Merkmalskorrektur in Bruttoeinkommenskomponenten umzurechnen. Ähnliche Schwierigkeiten, wenn auch in ihren Auswirkungen von wesentlich untergeordneter Bedeutung, traten auch bei anderen Variablen auf.

Die integrative Verknüpfung kennzeichnet demzufolge keine Verknüpfungsmethoden im engen Sinn, sie ermöglicht vielmehr

unter Umständen in ihrem Anschluß die Durchführung weiterer exakter bzw. statistischer Verfahren. In aller Regel bedingt sie aber umfangreiche Arbeiten zur Abstimmung von Variablen und zwar dann, wenn mangels existierender oder aber wissenschaftlichen Zwecken zur Verfügung stehender Daten es nicht möglich ist, entsprechende Einheiten exakt oder statistisch zu verknüpfen.

#### 4.2 Hypothetische Korrektur von Merkmalen<sup>21)</sup>

Mittels der hypothetischen Merkmalskorrektur war es im Anschluß an die logische Zusammenführung von Daten der deutschen Wohnbevölkerung außerhalb des Anstaltsbereiches aus der EVS sowie von Ausländern und der Wohnbevölkerung innerhalb des Anstaltsbereiches (Privathaushalte im Anstaltsbereich und Anstaltsinsassen) aus dem Mikrozensus notwendig, aus den Angaben der Nettoeinkommensgrößenklassen des Mikrozensus nach Einkommensarten differenzierte Bruttoeinkommen zu berechnen, entsprechend den Angaben in der EVS auf Personenebene.

Hierzu war es zunächst erforderlich, die Nettogrößenklassenangaben in absolute Werte umzurechnen. Nachdem verschiedene Versuche mit SPLINE-Funktionen keine zufriedenstellenden Ergebnisse erbracht hatten, wurde hierzu differenziert nach 60 Bevölkerungsgruppen<sup>22)</sup> und 7 Nettoeinkommensklassen eine Identität der Einkommensklassendurchschnitte in EVS und Mikrozensus unterstellt, es wurden also aus der EVS Werte errechnet und diese auf die aus dem Mikrozensus stammenden Angaben übertragen. Dieses Verfahren ist zulässig, da die theoretische Abgrenzung des Nettoeinkommensbegriffes in beiden Erhebungen die gleiche ist.

Die 1. Hypothese der Einkommensangabenkorrektur lautet demnach: Für gleiche Bevölkerungsgruppen aus der gleichen Grundgesamtheit gelten bei der gleichen Abgrenzung eines Einkommensbegriffes innerhalb identischer Einkommensgrößenklassen die gleichen Einkommensklassendurchschnitte.

Ein mit Hilfe der aus der EVS abgeleiteten Klassendurchschnitte durchgeführter Nettoeinkommensvergleich zwischen Personen in deutschen Privathaushalten außerhalb des Anstaltsbereiches in EVS und Mikrozensus zeigt allerdings recht beträchtliche Diskrepanzen, die darauf zurückzuführen sind, daß entsprechende Einheiten bei der Erhebung von Einkommensangaben in Größenklassen zu Unterangaben neigen<sup>23)</sup>. Diese Diskrepanzen wurden im folgenden dadurch beseitigt, daß wiederum differenziert nach 60 Bevölkerungsgruppen die zunächst aus der EVS ermittelten Klassendurchschnitte mit dem Quotienten aus durchschnittlichem Nettoeinkommen laut EVS und Mikrozensus multipliziert wurden.

Dieser Vergleich konnte nur für Personen in deutschen Privathaushalten außerhalb des Anstaltsbereiches durchgeführt werden, da nur diese Bevölkerungsgruppe in beiden Erhebungen erfaßt wird. Zu bestimmen waren jedoch die absoluten Nettoeinkommensangaben von Personen in Ausländerhaushalten sowie im Anstaltsbereich.

Die 2. Hypothese zur Korrektur der Einkommensangaben lautet deshalb: Personen in Ausländerhaushalten und im Anstaltsbereich haben bei gleichen demographischen Merkmalen und bei gleicher Nettoeinkommensklasse das gleiche durchschnittliche Nettoeinkommen wie deutsche Personen in Privathaushalten außerhalb des Anstaltsbereiches.

Die 3. Hypothese lautet: Es gibt keine statistisch signifikanten Unterschiede in der Zuverlässigkeit von Einkommensangaben zwischen Deutschen und Ausländern einerseits sowie zwischen Personen innerhalb und außerhalb des Anstaltsbereiches andererseits.

Alle drei genannten Hypothesen entziehen sich bislang einer empirischen Überprüfung, da keine entsprechende Daten vorliegen. Ihre Gültigkeit muß deshalb trotz einiger Bedenken bis zu einer eventuellen Falsifikation unterstellt werden.

Sie erlauben nunmehr, die aus der EVS für deutsche Personen

in Privathaushalten außerhalb des Anstaltsbereiches ermittelten korregierten Nettoeinkommensklassendurchschnitte auf Ausländer und deutsche Personen in Privathaushalten innerhalb des Anstaltsbereiches zu übertragen.

Im zweiten Schritt mußten die Nettoeinkommensangaben in Bruttoeinkommen umgerechnet werden. Dieses Problem wurde gelöst mittels eines von Gernold Frank entwickelten Programmmoduls, das mit Hilfe der reziproken Steuerformeln sowie unter Beachtung der Vorschriften des Steuer-, Renten- und Krankenversicherungsrechtes aus den Nettoeinkommen die Bruttoeinkommenssumme berechnet. Dieses Programm wurde anhand der EVS-Angaben getestet und lieferte sehr gute Ergebnisse auf der Personenebene. Da weder Steuer- noch Renten- oder Krankenversicherungsrecht zwischen Deutschen und Ausländern auf der Beitragsseite unterscheidet, bestand kein Anlaß zu der Vermutung, daß die Anwendung dieses Moduls auf Ausländer sowie die Wohnbevölkerung im Anstaltsbereich zu verzerrten Ergebnissen führen könnte.

Dem abschließenden Schritt der hypothetischen Merkmalskorrektur zur Splitterung der Bruttoeinkommen in seine Komponenten<sup>24)</sup> lag die folgende 4. Hypothese zugrunde: Für alle Personen innerhalb einer Zelle, die durch die soziale Stellung, das Geschlecht, die Stellung zum Haushaltsvorstand sowie das Lebensalter definiert ist, ist die Relation zwischen den Bruttoeinkommenskomponenten konstant.

Gemäß dieser Hypothese wurde die sich aus der EVS ergebende Struktur der Bruttoeinkommen erwerbstätiger deutscher Personen außerhalb des Anstaltsbereiches auf die entsprechenden deutschen Personen innerhalb des Anstaltsbereiches übertragen.

Ausdrücklich verzichtet wurde bisher auf die Aufsplitterung der Bruttoeinkommen nicht erwerbstätiger Deutscher im Anstaltsbereich sowie von Ausländern. Während es sich bei der ersten Gruppe um untypische Personen handelt, für die es außerhalb der Anstalten keine zuverlässige Vergleichsgruppe gibt, wurde eine hypothetische Aufspaltung der Ausländerbruttoeinkom-

men in der Hoffnung auf einen späteren Zugang zu besserem Datenmaterial einstweilen zurückgestellt.

#### 4.3 Synthetisch-hypothetische Generierung von Beobachtungseinheiten<sup>25)</sup>

Von der statistischen Verknüpfung von Stichproben muß die synthetische Generierung von Beobachtungseinheiten abgegrenzt werden. Dieses Verfahren wurde im Rahmen der Arbeit am Integrierten Mikrodatenfile 1969 angewandt zur Generierung von Haushalten mit höchsten Einkommen.

Zu verstehen sind hierunter alle diejenigen Schritte, die dazu erforderlich sind, um aus stark aggregierten Informationen über Bevölkerungsgruppen Individualdaten zu bestimmen.

Ausgangspunkt hierzu sind die aus den Gruppendaten ableitbaren Randverteilungen. Diese Randverteilungen werden sukzessive disaggregiert mit Hilfe von nach weiteren Merkmalen gegliederten Informationen über die gleiche Grundgesamtheit aus der gleichen statistischen Quelle. Diese Aufspaltung erfolgt dabei stets so, daß als Rahmenbedingung die vorgegebenen Randverteilungen auf jeder Disaggregationsstufe eingehalten sind. Zeigt sich nach der Berücksichtigung aller Informationen aus der ursprünglichen Statistik die Notwendigkeit einer tieferen Disaggregation der Gruppen, so können weitere Differenzierungen erfolgen anhand externer Informationen über die jeweilig aufzusplittenden Bevölkerungsgruppen.

Im Falle der Bezieher von Einkommen über 250.000 DM pro Jahr erfolgte zunächst eine Differenzierung nach der Größenklasse des Gesamtbetrages der Einkünfte und der überwiegenden Einkunftsart. Die Angaben standen in der Einkommensteuerstatistik 1968 zur Verfügung. Gemäß einer weiteren Tabellierung der Steuerpflichtigen nach der Größenklasse des Gesamtbetrages der Einkünfte und der Anzahl der Kinderfreibeträge konnten diese Zellen weiter aufgespalten werden.

Da aus der Steuerstatistik keine weiteren demographischen Merkmale abgeleitet werden konnten, wurde die weitere Differenzierung anhand von Verteilungen durchgeführt, die sich getrennt nach überwiegender Einkunftsart und Anzahl der Kinder aus der Einkommens- und Verbrauchsstichprobe errechnen ließen. Als Kriterium hierzu wurde zunächst das Alter des Haushaltsvorstandes und im Anschluß daran das Alter der Kinder herangezogen. Dieses Verfahren stellte sicher, daß die auf der nächst höheren Aggregationsstufe vorgegebenen Randverteilungen auch im Zuge der einzelnen Differenzierungsschritte erhalten blieb.

Differenziert nach drei Größenklassen des Gesamtbetrages der Einkünfte, sieben überwiegende Einkunftsarten, fünf Klassen des Kinderfreibetrages, fünf Altersklassen des Haushaltsvorstandes und vier Altersklassen der Kinder standen somit für 2100 Zellen die Besetzungshäufigkeiten fest. Eine größere Anzahl dieser Zellen war natürlich bereits ex definitione unbesetzt. Die zu generierenden Einheiten waren mithin so zu wählen, daß ihre demographische Struktur identisch war mit derjenigen, die sich aus der oben skizzierten Zellenverteilung ergab.

Eine weitere Differenzierung der genannten Gruppen zur Bestimmung weiterer demographischer Merkmale hätte zur Ableitung statistisch gesicherter Ergebnisse eine Stichprobengröße erfordert, die weit über der Anzahl der über 47.000 in der EVS erfaßten Haushalte gelegen hätte. Aus diesem Grunde wurde darauf verzichtet und nunmehr mittels eines Zufallsprozesses solche Haushalte aus der EVS gezogen, die die erforderlichen Merkmalsausprägungen besaßen. Diese Haushalte spiegeln sowohl die sich aus der Einkommensteuerstatistik 1968 ergebende Verteilung der Steuerpflichtigen wider wie auch innerhalb dieser Gruppen die Verteilung der deutschen Haushalte gemäß der EVS. Da sich die aus der EVS hervorgegangenen Informationen nicht auf Haushalte mit hohem Einkommen beziehen - da sie in der EVS nicht eingeschlossen sind - mußte die Annahme einer Unabhängigkeit dieser Angaben von der Einkommenshöhe gemacht werden. Diese durchaus zu diskutierende Hypothese wird durch das Prinzip des unzureichenden Grundes solange gerechtfertigt,

wie eine Falsifikation anhand empirischer Daten nicht möglich ist.

## 5. Resumee

Die Darstellung der verschiedenen Verfahren zur Verknüpfung, Korrektur und Generierung von Mikrodaten offenbaren die vielfältigen Schwierigkeiten bei der Generierung eines Mikrodatafiles. Diese beruhen sowohl auf der unvollständigen Erfassung aller Bevölkerungsgruppen wie auch der fehlerhaften Erhebung von Merkmalen. Selbst dort, wo Daten im Prinzip vollständig erhoben werden, unterliegen die Informationen bewußten oder unbewußten Fehlangaben oder einfachen Fehlvercodungen richtiger Angaben. Allen diesen Fehlern muß im Einzelfall nachgegangen werden und entsprechende Korrekturen müssen erfolgen.

Wie in diesem Beitrag gezeigt wurde, erfordert sowohl die Korrektur wie auch die Verknüpfung von Mikrodaten eine Reihe von Hypothesen, deren Gültigkeit unterstellt werden muß. Die Formulierung dieser Hypothesen ist weitgehend in das Ermessen der einzelnen Forscher gestellt. Alter schreibt denn auch "that the creation of synthetic datafiles by way of matching is just as an art as it is a science"<sup>26</sup>). Ein Verzicht auf derartige Hypothesen zöge als einzige Alternative den Verzicht auf jegliche Verknüpfung und Generierung von Mikrodaten nach sich. Ein solcher Preis erscheint aber in Anbetracht der unzureichenden Informationen über soziale Lagen zu hoch. Nur muß man sich bei der Interpretation der Ergebnisse, die aus einem solchen Datenfile abgeleitet werden, der Annahmen, die zu ihrer Generierung gemacht wurden, bewußt sein. Hierzu bedarf es einer Offenlegung aller Hypothesen.

## Anmerkungen

- 1) Vgl. Krupp, Hans-Jürgen, Sozialpolitisches Entscheidungs- und Indikatorensystem für die Bundesrepublik Deutschland (SPES), in: Allgemeines Statistisches Archiv, 1973, Heft 3/4, S. 380-387, s.S. 381
- 2) Vgl. Zapf, Wolfgang u.a., Das SPES-Indikatorensystem 1975, SPES-Arbeitspapier Nr. 46
- 3) Vgl. Glatzer, Wolfgang, Krupp, Hans-Jürgen, Soziale Indikatoren des Einkommens und seiner Verteilung für die Bundesrepublik Deutschland, in: Zapf, Wolfgang (Hrsg.), Soziale Indikatoren - Konzepte und Forschungsansätze, Band III, Frankfurt/Main 1975, S. 193-238
- 4) Vgl. Krupp, Hans-Jürgen, Stand der Statistik der personellen Einkommensverteilung, in: Wirtschaftsdienst, 1975, Heft 1, S. 36-41, s.S. 37
- 5) Vgl. Krupp, Hans-Jürgen, Möglichkeiten der Verbesserung der Einkommens- und Vermögensstatistik, Göttingen 1975, S. 80-90
- 6) Vgl. Okner, Benjamin A., Data Matching and Merging: An Overview, in: Annals of Economic and Social Measurement, 1974, 3/2, S. 347-352, s.S. 347
- 7) Vgl. David, M.H., Gates, W.A. and Miller, R.F., Linkage and Retrieval of Microeconomic Data. A Strategy for Data Development and Use, Lexington 1974
- 8) Vgl. Krupp, Hans Jürgen, Ergänzung der Volkswirtschaftlichen Gesamtrechnungen durch Vermögensrechnungen; Anforderungen an die Einkommens- und Verbrauchsstichprobe im Hinblick auf die Bereitstellung von Ausgangsdaten für Gesamtwirtschaftliche Vermögensrechnungen (Ergänzungsbericht). Forschungsbericht im Auftrag des Bundesministers für Arbeit und Sozialordnung unter Mitarbeit von Dr. Peter Hecheltjen und Arno Weigend, Frankfurt, August 1973, S. 8-13
- 9) Vgl. Alter, Horst, Creation of a Synthetic Data Set by Linking Records of the Canadian Survey of Consumer Finances with the Family Expenditure Survey 1970, in: Annals of Economic and Social Measurement, 1974, 3/2, S. 373-394.  
Budd, E.C., Radner, D.B., Hinrichs, J.C., Size Distribution of Family Personal Income: Methodology and Estimates for 1964, U.S. Bureau of Economic Analysis Staff Paper No. 21 (1973).  
Okner, Benjamin A., Constructing a New Data Base from Existing Microdata Sets: The 1966 Merge File, in: Annals of Economic and Social Measurement, 1972, 1/3, S. 325-342.  
Ruggles, Nancy and Richard, A Strategy for Merging and Matching Microdata Sets, in: Annals of Economic and Social Measurement, 1974, 3/2, S. 353-371

- 10) Okner, Benjamin A., Constructing .
- 11) Budd, E.C, Radner, D.B., Hinrichs, J., Size Distribution.
- 12) Alter, Horst, Creation, S. 377 ff.
- 13) Ruggles, Nancy and Richard, A Strategy, S. 359 ff.
- 14) Dieses Verfahren wenden Budd und Ruggles an.
- 15) Vgl. Alter, Horst, Creation, S. 384.
- 16) Vgl. Okner, Benjamin, A., Constructing, S. 332.
- 17) Sims, Christopher A., Comments, in: Annals of Economic and Social Measurement, 1972, 1/3, S. 343-345  
Sims, Christopher A., Rejoinder, in: Annals of Economic and Social Measurement, 1972, 1/3, S. 355-357  
Sims, Christopher A., Comment, in: Annals of Economic and Social Measurement, 1974, 3/2, S. 395-397.
- 18) Vgl. Okner, Benjamin A., Reply and Comments, in: Annals of Economic and Social Measurement 1972, 1/3, S. 359-362  
Budd, E.C., Comments, in: Annals of Economic and Social Measurement, 1972, 1/3, S. 349-354.
- 19) Bearbeitet wurden sie von Udo Kröber und Günther Schmaus.
- 20) Krupp, H.-J., Ergänzung, S. 11.
- 21) Vgl. Kortmann, Klaus, Krupp, Hans-Jürgen und Schmaus, Günther, Strukturen der Einkommensverteilung 1969 - Erste Ergebnisse und Erfahrungen mit einem integrierten Mikrodatenfile für die Bundesrepublik Deutschland 1969 (IMDAF 1969), S. 359-552 s.S. 542 f .  
Kortmann, Klaus und Schmaus, Günther, Generierung des Mikrodatenfiles 1969 für die Bundesrepublik Deutschland (IMDAF 1969), SPES-Arbeitspapier Nr. 39, S. 9-25.
- 22) Die Differenzierung erfolgte nach der sozialen Stellung (5 Klassen), der Stellung zum Haushaltsvorstand (3 Klassen) sowie der Haushaltsgröße (4 Klassen). Entsprechend der hier erläuterten Korrektur der Einkommen von Personen in deutschen und ausländischen Anstaltshaushalten, verlief die Korrektur der Einkommen der Anstaltsinsassen, lediglich die Gruppenbildung erfolgte nach etwas anderen Kriterien.
- 23) Vgl.: Zur Genauigkeit von Einkommensangaben in Interviews. Dargestellt am Beispiel der Einkommens- und Verbrauchsstichprobe 1969, in: Wirtschaft und Statistik, 1973, Heft 3, S. 193-196.
- 24) Jeweils Bruttoeinkommen aus unselbständiger Tätigkeit, aus Unternehmertätigkeit und Vermögen, aus Renten, aus öffentlichen Pensionen sowie aus einmaligen Übertragungen unter 1000 DM .

- 25) Vgl. Kortmann, Klaus und Schmaus, Günther, Die Generierung,  
S. 25-37.
- 26) Alter, Horst, Creation, S. 378.