

Datenerhebung aus Massenakten

Renn, Heinz

Veröffentlichungsversion / Published Version

Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:

Renn, H. (1984). Datenerhebung aus Massenakten. In W. Bick, R. Mann, & P. J. Müller (Hrsg.), *Sozialforschung und Verwaltungsdaten* (S. 168-191). Stuttgart: Klett-Cotta. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-330739>

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.

Datenerhebung aus Massenakten

0. Problemstellung

Bei der Verwendung von Massenakten als Datenquellen der empirischen Sozialforschung müssen folgende *Besonderheiten* in Rechnung gestellt werden. Massenakten sind „*prozeßproduzierte*“ Aufzeichnungen, die nicht zu Forschungszwecken erhoben werden. Auf das Zustandekommen und den weiteren Fortbestand der Akten bis zum Zeitpunkt einer wissenschaftlichen Auswertung hat der Sozialforscher keinen Einfluß¹. Darüber hinaus sind selten alle interessierenden Sachverhalte in einem einzelnen Aktenbestand festgehalten. In der Regel sind die *Informationen* unvollständig und *auf verschiedene Bestände verstreut*². Schließlich handelt es sich bei Massenakten um Informationssammlungen über eine *große Anzahl von Bestandseinheiten*. Die vollständige Analyse aller Einheiten ist aber oft aus ökonomischen, technischen oder methodischen Gründen nicht möglich oder nicht wünschenswert.

Datenerhebung aus überlieferten Beständen von Massenakten ist somit mit *zwei grundlegenden Problemen* konfrontiert:

- Einmal mit dem Problem der *inhaltlichen Abdeckung* des jeweils interessierenden Sachverhalts im einzelnen Aktenbestand („coverage“) verbunden mit dem nachgelagerten Problem der *Verknüpfung von Inhalten* verschiedener Bestände („record linkage“).
- Zum anderen mit dem Problem der *repräsentativen Auswahl* von Untersuchungseinheiten aus der Gesamtzahl identifizierter Bestandseinheiten („sampling“) und der *Schätzung charakteristischer Kennwerte* hinsichtlich der Hypothesen des Sozialforschers („estimation“).

Beim ersten der beiden Probleme geht es um systematische Verzerrungen des überlieferten Inhalts der Akten, die sich sowohl auf den *Ausfall bestimmter Bestandseinheiten* bzw. Einheitengruppen beziehen können als auch auf das *Fehlen bestimmter Merkmalsbeschreibungen* bei identifizierten Einheiten („missing data“). Beim zweiten Problem handelt es sich um die Frage nach der *Auswahl* von Einheiten aus dem vorgefundenen Datenmaterial *aufgrund angemessener Kriterien* sowie den Möglichkeiten einer *zufallskritischen Einschätzung* der Ergebnisse der Datenanalyse.

Die Unterteilung in diese beiden grundlegenden Probleme ist eine rein analytische Angelegenheit. In konkreten Datenerhebungen aus Massenakten sind Fragen der inhaltlichen Abdeckung, der Verknüpfung verschiedener Quellen, der Repräsentativi-

1. Für den Fall historischer Aufzeichnungen siehe Murphey (1973, 140-153).

2. Vgl. Hammarberg (1971, 542).

tät der Auswahl sowie der zufallskritischen Bewertung von Resultaten eng miteinander verflochten. Je nach spezifischer Datenlage haben einzelne Fragen jedoch einen besonderen Stellungswert.

Im folgenden will ich die angesprochenen Probleme der Datenerhebung genauer darstellen und Verfahrensvorschläge zu ihrer Lösung näher erörtern. Dabei beziehe ich mich jeweils auf bestimmte Datenkonstellationen.

1. Ein klassisches Beispiel

Wir gehen dabei aus von der klassischen Untersuchung der Personalverflechtungen von Banken, die Otto Jeidels im Jahre 1905 vorgelegt hat. Hierin wird für die Zeit um die Jahrhundertwende ein vermehrtes Auftreten personeller Verflechtungen zwischen den Banken und der Groß-Industrie nachgewiesen (Jeidels, 1905).

Die Jeidelsche Untersuchung ist nicht nur die erste empirische Arbeit über Personalverflechtungen von Banken; sie hat darüber hinaus theoriegeschichtlich einen besonderen Stellenwert, insoweit als sie von zeitgenössischen marxistischen Theoretikern als Beleg der Hypothese vom „Finanzkapital“, dem zunehmenden Ausmaß der Verflechtungen zwischen Banken und Großindustrie im Verlaufe der kapitalistischen Entwicklung, herangezogen wird³.

Nun sind bei genauerer Betrachtung die Jeidelschen Ergebnisse keineswegs als Aussagen über *die* Banken und *die* Großindustrie zu werten. Zwar ist auch bei Jeidels der Ausgangspunkt die allgemeine Vermutung, daß „die Selbständigkeit des Industriellen ... beschränkt werden (kann) durch die Vertreter und Leiter des Verwendung suchenden Geldkapitals, durch die Banken“ (3). Aufgrund von Schwierigkeiten bei der Datenerhebung war aber in zweifacher Hinsicht eine Einschränkung der Allgemeinheit der in Angriff genommenen Thematik notwendig. Eine Schwierigkeit ergab sich aus dem Umstand, daß „die Beziehungen der Bankwelt zur Industrie sich abspielen zwischen einer großen Zahl von Banken auf der einen und einer großen Zahl von verschiedenen Gewerbezweigen auf der anderen Seite“. Hier war eine Auswahl zu treffen: Jeidels untersucht auf der Bankenseite nur die damaligen sechs Großbanken, auf der Industrieseite nur die Montan- und Eisenindustrie sowie Unternehmungen, „die in den Produktionsprozeß der Eisenindustrie gehören“ (4). Darüber hinaus hatte Jeidels „mit einem bedenklichen Hindernis zu kämpfen, nämlich mit dem Zustand des Quellenmaterials“ (10). Als Datenquellen nutzt Jeidels prozeßproduzierte Aufzeichnungen, die jeweils spezifische Informationslücken aufweisen, die bestenfalls untereinander kompensiert werden können. Im einzelnen verwendet er als Datenquellen die Geschäftsberichte der Großbanken, allgemeine Presseartikel sowie solche in Fachblättern der Industrie und des Bankenbereichs, weiter sogenannte Börsenhandbücher, wie das „Handbuch der Aktiengesellschaften“ und das „Adreßbuch der Direktoren und Aufsichtsräte“.

Es liegt auf der Hand, daß angesichts dieser Gegebenheiten die Jeidelsche Untersuchung keineswegs als repräsentativ für *die* Banken und ihre Beziehungen zu *der* Industrie angesehen werden kann. Sie ist somit nur von eingeschränktem Wert als empirischer Beleg der These vom „Finanzkapital“.

3. So von Lenin, 1917, in seiner Imperialismustheorie.

Angesichts dieser Diskrepanz zwischen schmaler empirischer Basis der Jeidelschen Untersuchung und der großen Bedeutung, die ihr von den marxistischen Interpreten zugeschrieben wird, wäre eine Wiederholung der Arbeit — und zwar für den Zeitraum der Jahrhundertwende — unter den heutigen besseren methodischen und technischen Bedingungen ausgesprochen reizvoll: Die Fragestellung Jeidels müßte erneut aufgegriffen werden, jedoch sei gleichzeitig zu versuchen, die Schwierigkeiten, mit denen Jeidels bei der Datenerhebung konfrontiert war und die die oben beschriebenen Beschränkungen der Aussagekraft zeitigten, mit Hilfe des heutigen methodischen und technischen Instrumentariums zu überwinden.

Damit ist im Hinblick auf diese klassische Untersuchung die eingangs umrissene allgemeine Problematik der Datenerhebung aus Massenakten angesprochen: In der Jeidelschen Arbeit treten pointiert auf

- das Problem der inhaltlichen Abdeckung des theoretisch Gemeinten durch die empirischen Daten, verbunden mit dem Problem der richtigen Verknüpfung von Inhalten verschiedener Quellen,
- das Problem der angemessenen Auswahl von Untersuchungseinheiten, verbunden mit dem Problem der Einschätzung der Repräsentativität der Analyseergebnisse.

Wir wollen im folgenden diese Probleme, ausgehend von der theoretischen Fragestellung Jeidels, darlegen und für die Untersuchung von Personalverflechtungen der Banken Möglichkeiten der Lösung aufzeigen, hierbei zu machende Annahmen illustrieren und nicht überwindbare Schwierigkeiten kennzeichnen.

2. Der Idealfall: Die einzelne vollständige Datenquelle

Die eingangs dargelegte Fragestellung enthält als *Untersuchungsobjekt* die Personalverflechtungen von Banken um die Jahrhundertwende im damaligen Deutschen Reich. Dies ist ein theoretischer Begriff, mit dem der *raumzeitliche Rahmen* der Untersuchung abgesteckt wird und der darüber hinaus die *Untersuchungseinheiten (Analyseeinheiten)* und die interessierenden Merkmale derselben festlegt. Auf der theoretischen Ebene sind gegebenenfalls weitere begriffliche Explikationen notwendig, um auf der Meßebene eindeutige Indikatoren der theoretischen Sachverhalte bestimmen zu können.

In unserem Beispiel sind die Untersuchungseinheiten „Banken“, als Merkmale sollen interessieren Personalverflechtungen und Eigenschaften von Banken, die Personalverflechtungen erklären können. Da wir voraussetzen, daß intuitiv hinreichend klar ist, was man unter einer Bank zu verstehen hat, erübrigen sich für diesen Begriff weitere Präzisierungen. Die Bezeichnung als Bank in einer Datenquelle reicht als Indikator aus. Hingegen wird für das Merkmal „Personalverflechtungen“ festgelegt, daß hierunter die Anzahl von Positionen zu verstehen ist, die Angehörige einer Bank in Kontrollgremien von Unternehmungen im Nichtbanken-Bereich innehaben⁴. Weitere Merkmale seien die Größe der Bank, ausgedrückt etwa in der Höhe der Bilanzsumme, des Eigenkapitals oder der Einlagen, die Rechtsform der Bank u. ä..

Wichtig ist, daß durch den *theoretischen Begriff* des Untersuchungsobjektes auch die *Grundgesamtheit* dessen, was untersucht werden soll, festgelegt wird.

4. Dies ist nur eine von vielen möglichen Definitionen von „Personalverflechtung“.

Die Grundgesamtheit des Untersuchungsobjektes kann formal als eine *Matrix* beschrieben werden. Diese Matrix X_{ij} ist eine — zunächst nur fiktive — Aufstellung der Ausprägungen x_{ij} , die sämtliche Angehörigen der Grundgesamtheit, $0_1, \dots, 0_n$, hinsichtlich der interessierenden Merkmale, x_1, \dots, x_m , besitzen.

Ausführlich lautet sie wie folgt:

$$X_{ij} = \begin{matrix} & x_1 & \dots & x_j & \dots & x_m \\ \begin{matrix} 0_1 \\ \vdots \\ 0_i \\ \vdots \\ 0_n \end{matrix} & \left[\begin{array}{cccc} x_{11} & \dots & x_{1j} & \dots & x_{1m} \\ \vdots & & \vdots & & \vdots \\ x_{i1} & \dots & x_{ij} & \dots & x_{im} \\ \vdots & & \vdots & & \vdots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nm} \end{array} \right. \end{matrix} \quad (1)$$

Die kritische Frage ist nun, ob eine einzelne Datenquelle existiert, in der diese Grundgesamtheit vollständig aufgezeichnet ist. So könnte ein „Handbuch der Banken“ als Datenquelle vorliegen, in dem sämtliche Banken aufgeführt sind, Aussagen über die jeweilige Größe, die Rechtsform u.ä. gemacht und auch die jeweilige Anzahl von Personalverflechtungen der Bank angegeben sind.

In einem solchen Idealfall ist die *Datenquelle* hinsichtlich des theoretischen Begriffs des Untersuchungsobjektes *vollständig*: Für jede *Untersuchungseinheit* auf der theoretischen Ebene existiert eine entsprechende *Bestandseinheit* in der Datenquelle. Die Vollständigkeit eines Datenbestandes ist somit keine isolierte Eigenschaft der Datenquelle; sie ergibt sich aus der Konfrontation von Datenquelle und theoretischem Begriff.

Der Datenquelle kann im Idealfall X_{ij} als *Datenmatrix* entnommen und in der üblichen Weise analysiert werden. Beispielsweise sind die merkmalspezifischen Mittelwerte $\mu \cdot j$ leicht zu berechnen:

$$\mu \cdot j = \frac{\sum_{i=1}^n x_{ij}}{n} \quad (2)$$

Da es sich im Falle von Massenbeständen um eine große Anzahl von Bestandseinheiten handelt, ist es empfehlenswert, aus der Grundgesamtheit der Bestandseinheiten eine *Zufallsstichprobe* zu ziehen und für die Stichprobendaten eine Datenmatrix zu erstellen. Die entsprechenden *Stichprobenkennwerte* $\bar{x} \cdot j$ sind erwartungstreue Schätzwerte der Parameter der Grundgesamtheit.

Bezüglich unseres Beispiels sind wir in der Lage, mit einer bestimmten statistischen Signifikanz z. B. Aussagen über die durchschnittliche Anzahl von Personalverflechtungen der Banken zu machen und Zusammenhänge zwischen der Größe einer Bank und der Anzahl ihrer Personalverflechtungen aufzuzeigen.

Da die Daten einer einzelnen Datenquelle dem theoretischen Begriff gerecht werden, bestehen weder das Problem der inhaltlichen Abdeckung noch das der Verknüpfung von Inhalten verschiedener Datenbestände. Probleme der repräsentativen Auswahl von Bestandseinheiten und der Parameterschätzung aus Stichprobendaten tre-

ten zwar auf, können aber durch die Theorie der Stichprobenziehung und der statistischen Inferenz als zureichend gelöst betrachtet werden⁵.

3. Der Normalfall: Die einzelne unvollständige Datenquelle

Der geschilderte Idealfall der vollständigen Datenquelle ist jedoch nicht der Normalfall der Datenerhebung aus vorgefundenen Beständen. In der Regel ist die zur Verfügung stehende *Datenquelle unvollständig*. Zwei polare Typen der *Unvollständigkeit* sind hier denkbar:

- (1) Einmal fehlen in der Datenquelle bestimmte *Bestandseinheiten*, so daß für die entsprechenden Untersuchungseinheiten keine Informationen vorliegen.
- (2) Zum anderen sind zwar alle Bestandseinheiten identifiziert, es fehlen zu bestimmten *Merkmalen* jedoch die Informationen⁶.

Bezogen auf unser Beispiel bestünde im ersten Fall die Unvollständigkeit der Datenquelle darin, daß in einem Handbuch *nicht alle* Banken aufgeführt wären. Im zweiten Falle enthielte das Verzeichnis zwar alle Banken, bei einigen Merkmalen, z. B. bei der Personalverflechtung, sind jedoch keine Angaben vorhanden.

Wir wollen im folgenden die jeweilige Struktur der Unvollständigkeit einer Datenquelle untersuchen, die Auswirkung auf das Analyseergebnis abschätzen und sodann mögliche Problemlösungen erörtern.

3.1 Die Nichtidentifikation bestimmter Bestandseinheiten

Gehen wir davon aus, daß in einer Datenquelle nur k Bestandseinheiten enthalten sind (identifiziert werden können), so kann die vollständige Datenmatrix X_{ij} nicht mehr aus dem Datenbestand erstellt werden. Sie zerfällt bezüglich ihrer Erhebbarkeit in zwei Teile:

$$X_{ij} = \begin{bmatrix} X_{km} \\ X_{lm} \end{bmatrix} \quad (3)$$

wobei $k+l=n$

$$X_{km} = \begin{matrix} & x_1 & \dots & x_j & \dots & x_m \\ 0_1 & \begin{bmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1m} \\ \vdots & & \vdots & & \vdots \\ x_{k1} & \dots & x_{kj} & \dots & x_{km} \end{bmatrix} \\ & & & & & \end{matrix} \quad (3a)$$

$$X_{lm} = \begin{matrix} & x_1 & \dots & x_j & \dots & x_m \\ 0_{k+1} & \begin{bmatrix} 0 & \dots & 0 & \dots & 0 \\ \vdots & & \vdots & & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix} \\ & & & & & \end{matrix}$$

5. Vgl. für den Fall der historischen Sozialforschung Schofield, 1972.

6. Streng genommen ist (1) der Grenzfall von (2), da bei (1) Informationen zu *allen* Merkmalen fehlen, bei (2) nur die Informationen zu *einigen* Merkmalen. Allerdings setzt (1) nicht die Identifikation der jeweiligen Bestandseinheiten voraus.

Die Matrix X_{km} enthält die gesamte Information der Datenquelle. Die Matrix X_{lm} ist eine Nullmatrix, da für die Untersuchungseinheiten $0_{k+1}, \dots, 0_n$ keine Daten vorhanden sind.

Welche Auswirkungen hat der Ausfall der l nichtidentifizierten Bestandseinheiten auf das Ergebnis der Datenanalyse?

Folgende *Konstellationen* können unterschieden werden:

- (1) Der Ausfall der Bestandseinheiten ist *zufallsbedingt*.
- (2) Der Ausfall der Bestandseinheiten erfolgte aufgrund *systematischer Einflüsse*, wobei die Wirkungsfaktoren des Ausfalls und deren Zusammenspiel mit den interessierenden Eigenschaften des Untersuchungsobjektes (a) *unbekannt* bzw. (b) *bekannt* sein können.

Der *zufällige* Ausfall von Bestandseinheiten ist völlig unproblematisch: In diesem Falle entspricht eine Zufallsstichprobe aus X_{km} einer Zufallsstichprobe aus X_{ij} . Der Ausfall von l Bestandseinheiten hat lediglich einen *größeren Zufallsfehler* zur Folge. Bei relativ großem k kann diese Auswirkung aber praktisch vernachlässigt werden.

Rein formal handelt es sich hier um das gleiche Vorgehen wie bei der Anwendung einer „missing-data“-Routine im Falle unvollständiger Datenmatrizen in der Umfrageforschung⁷. Die dem üblichen „missing-data“-Problem der Umfrageforschung entsprechende Datenlage unterscheidet sich aber grundlegend von der hier betrachteten Situation: Einmal ist X_{lm} eine reine Nullmatrix. Der Einsatz einer „missing-data“-Routine erbrächte somit keine Effizienzsteigerung. Darüber hinaus sind die Bestandseinheiten $0_{k+1}, \dots, 0_n$ nicht identifiziert. Aus diesem Grunde kann die Matrix

7. Vgl. zu den einzelnen Verfahren und ihrer jeweiligen Problematik Hertel (1976); für die deutschsprachige Literatur Lösel und Wüstendörfer (1974); für die historische Sozialforschung werden die entsprechenden Probleme von Bremer (1977) diskutiert. Gemäß einer solchen Vorgehensweise wäre jeder der l Bestandseinheiten, $0_{k+1}, \dots, 0_n$, jeweils der merkmalspezifische Stichprobenmittelwert als erwartungstreuer Schätzwert von $\mu \cdot j$ zuzuweisen. Die derart kompletizierte Datenmatrix X_{ij} sähe dann wie folgt aus:

$$X_{ij} = \begin{matrix} & \begin{matrix} x_1 & \dots & x_j & \dots & x_m \end{matrix} \\ \begin{matrix} 0_1 \\ \vdots \\ 0_k \\ 0_{k+1} \\ \vdots \\ 0_n \end{matrix} & \left[\begin{matrix} x_{11} & \dots & x_{1j} & \dots & x_{1m} \\ \vdots & & \vdots & & \vdots \\ x_{k1} & \dots & x_{kj} & \dots & x_{km} \\ \bar{x} \cdot 1 & \dots & \bar{x} \cdot j & \dots & \bar{x} \cdot m \\ \vdots & & \vdots & & \vdots \\ \bar{x} \cdot 1 & \dots & \bar{x} \cdot j & \dots & \bar{x} \cdot m \end{matrix} \right] \end{matrix}$$

Diese Vorgehensweise, von Hertel (1976, 461–465) „*the sample mean solution*“ genannt (vgl. für die theoretischen Grundlagen Wilks, 1932), bringt im vorliegenden Falle gegenüber dem einfachen Weglassen von Bestandseinheiten (nach Hertel, 1976, 460–461, „*the delete cases solution*“), *keine Effizienzsteigerung*. Durch das Einsetzen der merkmalspezifischen Mittelwerte $\bar{x} \cdot j$ wird die Zahl der Freiheitsgrade *nicht* erhöht. Eine Effizienzsteigerung ist nur dann

X_{ij} gar nicht erstellt werden. Dies ist aber beim zufallsbedingten Ausfall von Bestandseinheiten ohne Bedeutung.

Ist hingegen der Ausfall der Bestandseinheiten *systematisch*, so kann nicht mehr von der Gleichwertigkeit der Information in X_{nm} und X_{km} ausgegangen werden. Dies hat Auswirkungen auf das Analyseergebnis.

Im vorliegenden Falle bilden die k Bestandseinheiten der unvollständigen Datenquelle eine *andere Grundgesamtheit* als die des theoretischen Untersuchungsobjektes. So mögen in unserem fiktiven „Handbuch der Banken“ nicht alle Banken, sondern nur Banken ab einer bestimmten Kapitalausstattung, aus einer bestimmten Region und/oder mit einer bestimmten Publizitätspflicht enthalten sein.

Eine Zufallsstichprobe aus einer derart verzerrten Datenquelle ist aber genauso verzerrt. Die Inferenzstatistik hilft folglich hier nicht weiter. Ein Schluß von entsprechenden Stichprobenkennwerten beträfe in unserem Beispiel nur die Grundgesamtheit *dieser* Banken und keineswegs die Grundgesamtheit *aller* Banken.

Im Falle unvollständiger Datenquellen mit systematischer Verzerrung unbekannter Art und unbekanntem Ausmaßes ist man demzufolge genötigt, vom ursprünglichen theoretischen Begriff des Untersuchungsobjektes abzugehen und eine *der Datenlage entsprechende Umdefinition des Begriffs* vorzunehmen.

Verzerrte Datenbestände sind *kein unlösbares Problem*. Man muß allerdings etwas über das Zustandekommen der Verzerrung und deren Auswirkungen auf Eigenschaften des interessierenden Untersuchungsobjektes und die Anzahl der betroffenen Bestandseinheiten wissen.

Gehen wir davon aus, daß nur *ein einzelner* Faktor den Ausfall der Bestandseinheiten bewirkt. X_{km} und X_{lm} unterscheiden sich dann nach der Wirkung dieses Faktors auf die jeweilige Verteilung der x_j . Die Verteilungen der x_j für die $0_1, \dots, 0_k$ können der Datenquelle entnommen werden. Eine Schätzung der $\mu \cdot j$ aus diesen Verteilungen anhand der entsprechenden merkmalspezifischen Stichprobenmittelwerte $\bar{x} \cdot j$ ist aber verzerrt:

$$\mu \cdot j \neq E(x_j) \quad (4)$$

zu erwarten, wenn X_{ln} keine Nullmatrix ist, sondern für einen Teil der x_{ij} ($i = k + 1, \dots, n$) Angaben in der Datenquelle vorhanden sind, z. B.

$$X_{ln} = \begin{matrix} & x_1 & \dots & x_j & & \dots & x_m \\ \begin{matrix} 0_{k+1} \\ \vdots \\ 0_n \end{matrix} & \left[\begin{array}{cccc} 0 & \dots & x_{(k+1)j} & \dots & x_{(k+1)m} \\ \vdots & & \vdots & & \vdots \\ x_m & \dots & 0 & & \dots & 0 \end{array} \right] \end{matrix}$$

Durch Ersetzen der „missing data“ durch den jeweiligen merkmalspezifischen Stichprobenmittelwert $\bar{x} \cdot j$ und die anschließende Analyse der gesamten Datenmatrix X_{ij} wird auch die in X_{ln} enthaltene Information verwertet. Dies ist das übliche „missing-data“-Problem der Umfrageforschung. Die unserer Diskussion zugrunde liegende Datenlage unterscheidet sich aber von der der Umfrageforschung grundlegend.

Das jeweilige Ausmaß der Verzerrung Δ_j kann formal bestimmt werden als die Differenz zwischen diesen beiden Größen:

$$\Delta_j = E(x_j) - \mu \cdot j. \quad (5)$$

Ausgehend von bestimmten Vorkenntnissen über den Faktor, der den Ausfall von Bestandseinheiten bewirkt hat, sind bestimmte *qualitative Korrekturen* des Ergebnisses der Analyse von X_{km} durchaus plausibel.

So könnte man vermuten, unsere Datenquelle, das „Handbuch der Banken“, enthalte nur Banken ab einer bestimmten Größe. Da bei großen Banken die Einlagenhöhe vermutlich auch größer ist als bei kleinen Banken, wäre die durchschnittliche Einlagenhöhe berechnet für eine Stichprobe aus dem verzerrten Datenbestand im Wert zu hoch. Eine entsprechende Interpretation des Analyseergebnisses wäre somit angebracht. Strebt man darüber hinaus eine *quantitative* systematische Korrektur des Analyseergebnisses an, so sind die Δ_j quantitativ zu bestimmen und $\bar{x} \cdot j$ als Stichprobenschätzung von $E(x_j)$ ist entsprechend zu berichtigen, so daß

$$\mu \cdot j = \bar{x} \cdot j - \Delta_j. \quad (6)$$

Es läßt sich aber zeigen, daß die Δ_j aus einer einzelnen unvollständigen Datenquelle allein *nicht* zu bestimmen sind:

$$\Delta_j = \frac{1}{n} (\bar{x} \cdot j - \xi \cdot j) \quad (7)$$

Das Ausmaß der Verzerrung kann als die mit dem Anteil der l nichtidentifizierten Bestandseinheiten an den n Untersuchungseinheiten gewichteten Differenz zwischen dem merkmalspezifischen Mittelwert der Datenquelle $\bar{x} \cdot j$ und dem entsprechenden Mittelwert der nichtidentifizierten Bestandseinheiten $\xi \cdot j$ geschätzt werden.

Von diesen drei Bestimmungsfaktoren der Verzerrung sind jedoch zwei aus der zur Verfügung stehenden Datenquelle nicht zu ermitteln.

Im Hinblick auf die systematische Korrektur der Analyseergebnisse einer unvollständigen Datenquelle ist man somit auf weitere Informationen angewiesen. Dies können einmal andere Datenquellen sein. Auf dieses Problem kommen wir an anderer Stelle zurück.

Darüber hinaus ist das Problem der systematischen Verzerrung von Analyseergebnissen aus vorgefundenen Datenbeständen nur zu lösen durch eine *gültige Theorie der systematischen Fehlerentstehung und der systematischen Fehlerwirkung*. Nur so kann im Nachhinein eine Ergebniskorrektur hinsichtlich der Generalisierung auf die Grundgesamtheit des theoretischen Untersuchungsobjektes abgesichert werden⁸. Dabei liegt es auf der Hand, daß die Zusammenhänge wesentlich komplexer sind, wenn *mehrere* Faktoren den Ausfall von Bestandseinheiten bewirken.

Eine solche Theorie müßte Angaben enthalten über die Wirkungsfaktoren des Ausfalls von Bestandseinheiten und über das Zusammenspiel dieser Wirkungsfakto-

8. Vgl. für den Fall der historischen Sozialforschung Hammarberg, 1977 a, 459-460.

ren mit den interessierenden Eigenschaften des Untersuchungsobjektes und die Anzahl der betroffenen Bestandseinheiten. Dies bedeutet in unserem Beispiel, daß eine Theorie darüber vorliegen muß, welche Faktoren, z. B. die Einlagenhöhe, von Bedeutung sind, daß Banken in die vorgefundene Datenquelle des „Handbuches der Banken“ aufgenommen werden und wie solche Faktoren mit anderen interessierenden Sachverhalten, z. B. die Personalverflechtung, in Beziehung stehen und wie groß die Anzahl der nicht aufgenommenen Banken ist. Systematische Korrekturen sind um so genauer, je größer der Informationsgehalt der Theorie ist. Dabei sind die Korrekturbemühungen jedoch ohne Wert, ja sie wirken sogar verschlimmernd, wenn die Vermutungen auf denen sie beruhen, falsch sind. In vielen Fällen ist man allerdings auf mehr oder weniger plausible ad-hoc Vermutungen angewiesen⁹. Wo selbst solche sinnvollerweise nicht abgeleitet werden können, bleibt dem Forscher nur die oben angesprochene Umdefinition des Begriffs des ursprünglichen Untersuchungsgegenstandes übrig.

3.2 Fehlende Angaben zu identifizierten Bestandseinheiten

Der zweite Typus der Unvollständigkeit einer einzelnen Datenquelle liegt vor, wenn zwar alle n Bestandseinheiten identifiziert sind, jedoch zu q Merkmalen Informationen fehlen. Auch hier kann aus der Datenquelle die vollständige Datenmatrix X_{ij} nicht erstellt werden. Bezüglich der Erhebbarkeit der Daten zerfällt X_{ij} wiederum in zwei Teile:

$$X_{ij} = [X_{np} | X_{nq}] \quad (9)$$

wobei $p + q = m$

$$\begin{array}{c}
 \begin{array}{c}
 x_1 \quad \dots \quad x_p \\
 0_1 \left[\begin{array}{ccc}
 x_{11} & \dots & x_{1p} \\
 \vdots & & \vdots \\
 x_{i1} & \dots & x_{ip} \\
 \vdots & & \vdots \\
 0_n \left[\begin{array}{ccc}
 x_{n1} & \dots & x_{np}
 \end{array} \right.
 \end{array} \\
 \\
 \begin{array}{c}
 x_{p+1} \quad \dots \quad x_m \\
 0_1 \left[\begin{array}{ccc}
 0 & \dots & 0 \\
 \vdots & & \vdots \\
 0 & \dots & 0 \\
 \vdots & & \vdots \\
 0_n \left[\begin{array}{ccc}
 0 & \dots & 0
 \end{array} \right.
 \end{array}
 \end{array}
 \end{array} \quad (9a)$$

9. Vgl. für den Fall der historischen Sozialforschung Murphy, 1973, 171.

Die Matrix X_{np} enthält hier die gesamte Information der Datenquelle. Die Matrix X_{nq} ist eine Nullmatrix, da für die Merkmale x_{p+1}, \dots, x_m keine Daten vorhanden sind.

Welche Auswirkungen hat das Fehlen der Angaben bei Merkmalen, x_{p+1}, \dots, x_m , auf das Ergebnis der Datenanalyse?

Zunächst muß ein *Spezifikum das Fehlen von Angaben zu einem Merkmal* in einer Datenquelle gegenüber dem Fehlen von Bestandseinheiten herausgestellt werden. Beide Sachverhalte betreffen den Ausfall von Informationen über den Untersuchungsgegenstand. Allerdings besteht ein grundsätzlicher Unterschied zwischen dem Ausfall einer Bestandseinheit und dem Ausfall eines Merkmals.

Durch den *Ausfall einzelner Bestandseinheiten* entstehen in der Regel nur kleine Verzerrungen der in der Datenquelle enthaltenen Informationen. Bei einer relativ großen Anzahl von Bestandseinheiten und einem zufallsbedingten Ausfall gleichen sich diese vielen Einzelverzerrungen aus. Hinsichtlich der nach der theoretischen Fragestellung interessierenden Informationen der Datenquelle ist nur eine inferenzstatistisch kontrollierbare Zufallsverzerrung vorhanden. Demgegenüber liegt eine in dieser Hinsicht *systematische Verzerrung* der Informationen der Datenquelle vor, wenn aufgrund eines systematisch wirkenden Faktors nur ganz bestimmte Bestandseinheiten ausfallen, so daß die entsprechenden Einzelverzerrungen in eine bestimmte Richtung kumulieren.

In 3.1 haben wir diese beiden Fälle betrachtet und die unterschiedlichen Auswirkungen auf das Analyseergebnis untersucht.

Bei *fehlenden Angaben zu Merkmalen* ist eine solche Unterteilung wenig sinnvoll.

Hinsichtlich der fehlenden Information betrifft der Ausfall einer Bestandseinheit *nur eine* Untersuchungseinheit, hingegen betrifft der Ausfall eines Merkmals *alle* Untersuchungseinheiten. Ein solches Merkmal ist aber Indikator eines theoretischen Begriffs. Somit bewirkt sein Ausfall *immer eine systematische Verzerrung* der Informationen der Datenquelle, die für die theoretische Fragestellung der Untersuchung von Belang sind, gleichgültig, ob der Ausfall auf einen systematischen oder einen Zufallsfaktor zurückgeführt werden kann.

Dies trifft auch zu, wenn ein theoretischer Begriff durch mehrere Indikatoren gemessen werden soll und zufallsbedingt alle entsprechenden Merkmale ausfallen. Die Aussage muß allerdings abgeschwächt werden, wenn von mehreren Indikatoren eines theoretischen Konstrukts zufallsbedingt nur ein Teil ausfällt. Hier ist ein Fehlerausgleich, wie beim zufallsbedingten Ausfall von Bestandseinheiten zu erwarten. Angesichts der im Vergleich zu den Bestandseinheiten sehr kleine Zahl der Merkmale ist jedoch immer mit einem sehr großen Zufallsfehler zu rechnen¹⁰.

Aus diesen Erwägungen heraus gehen wir im folgenden davon aus, daß durch fehlende Angaben zu den Merkmalen x_{p+1}, \dots, x_m die in X_{np} enthaltene Information im Hinblick auf den Untersuchungsgegenstand systematisch verzerrt ist. Dies gilt im-

10. Ein zufallsbedingter Ausfall der Angaben zu einem Merkmal könnte aber nach folgender Argumentation plausibel sein: Gehen wir davon aus, daß jedem theoretischen Begriff auf der Meßebene eine Grundgesamtheit von Indikatoren (Operationalisierungen) entspricht. Solche Indikatoren können sich sowohl unter den p in der Datenquelle vorhandenen Merkmalen, x_1, \dots, x_p , befinden als auch unter den q fehlenden Merkmalen, x_{p+1}, \dots, x_m . Unterstellen wir nur einen interessierenden theoretischen Begriff ξ . Gehen wir weiter davon

mer, gleichgültig, ob ein systematischer oder ein Zufallsfaktor den Ausfall bewirkt hat und gleichgültig, ob der systematische Wirkungsfaktor bekannt oder unbekannt ist.

Da somit X_{np} eine andere Grundgesamtheit betrifft als die des Untersuchungsobjektes ist es notwendig, vom ursprünglichen theoretischen Begriff des abzugehen und eine der Datenlage entsprechende Umdefinition des Untersuchungsgegenstandes vorzunehmen.

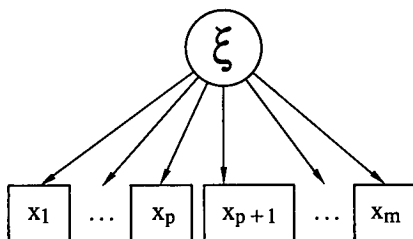
3.3 Mischtypen

Bis hierhin haben wir die beiden polaren Typen der Unvollständigkeit einer einzelnen Datenquelle, die Nichtidentifikation von Bestandseinheiten (3.1) und das Fehlen von Angaben zu identifizierten Bestandseinheiten (3.2) betrachtet. In einer Datenquelle können jedoch *sowohl* Bestandseinheiten, z. B. $0_{k+1}, \dots, 0_n$, *als auch* Angaben zu bestimmten Merkmalen, z. B. zu x_{p+1}, \dots, m , fehlen. Hier zerfällt bezüglich der Erhebbarkeit X_{ij} in vier Teile:

$$X_{ij} = \left[\begin{array}{c|c} X_{kp} & X_{km} \\ \hline X_{np} & X_{nm} \end{array} \right] \quad (10)$$

wobei $k+l=n$ und $p+q=m$

aus, daß alle m Indikatoren eine Stichprobe aus der Grundgesamtheit der möglichen Indikatoren von ξ betrachtet werden können (Abbildung).



In diesem Falle besteht auf der Meßebeine eine Gleichwertigkeit aller Indikatoren als Operationalisierungen des theoretischen Begriffs. Somit machte es nichts für die Gültigkeit der Messung des Konstrukts aus, wenn zufallsbedingt einige dieser Indikatoren, z. B. x_{p+1}, \dots, x_m , im Datenbestand nicht vorhanden sind. Die Zuverlässigkeit der Messung leidet allenfalls.

Wie beim zufallsbedingten Ausfall von Bestandseinheiten (3.1) wäre unter diesen Bedingungen eine Zufallsstichprobe aus X_{np} bis auf einen größeren Zufallsfehler einer Zufallsstichprobe aus X_{ij} gleichwertig. Allerdings ist es höchst unwahrscheinlich, daß in einer Datenquelle für alle ausgefallenen Merkmale unter den vorhandenen Merkmalen i. o. Sinne Entsprechungen vorhanden sind. Die Vermutung, daß ein Ausfall von Merkmalen immer systematische Verzerrungen hervorruft, erscheint somit gerechtfertigt zu sein.

$$\begin{array}{cc}
 \begin{array}{c} \\ \end{array} & \begin{array}{c} x_1 \dots x_p \\ \left[\begin{array}{ccc} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{k1} & \dots & x_{kp} \end{array} \right] \end{array} & \begin{array}{c} \\ \end{array} & \begin{array}{c} x_{p+1} \dots x_m \\ \left[\begin{array}{ccc} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{array} \right] \end{array} \\
 X_{kp} = \begin{array}{c} 0_1 \\ \vdots \\ 0_k \end{array} & & X_{km} = \begin{array}{c} 0_1 \\ \vdots \\ 0_k \end{array} & & & \\
 & & & & & (10a)
 \end{array}$$

$$\begin{array}{cc}
 \begin{array}{c} \phantom{0_{k+1}} \\ \end{array} & \begin{array}{c} x_1 \dots x_p \\ \left[\begin{array}{ccc} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{array} \right] \end{array} & \begin{array}{c} \phantom{0_{k+1}} \\ \end{array} & \begin{array}{c} x_{p+1} \dots x_m \\ \left[\begin{array}{ccc} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{array} \right] \end{array} \\
 X_{np} = \begin{array}{c} 0_{k+1} \\ \vdots \\ 0_n \end{array} & & X_{nm} = \begin{array}{c} 0_{k+1} \\ \vdots \\ 0_n \end{array} & & &
 \end{array}$$

In diesem Fall enthält die Matrix X_{kp} die gesamte Information der Datenquelle. Da sowohl die Untersuchungseinheiten $0_{k+1}, \dots, 0_n$ nicht identifiziert sind als auch Angaben zu den Merkmalen x_{p+1}, \dots, x_m fehlen, sind die übrigen Matrizen Nullmatrizen¹¹.

Unsere in den Abschnitten 3.1 und 3.2 angestellten Überlegungen können auf diese Datenlage übertragen werden: Hinsichtlich fehlender Angaben zu bestimmten Merkmalen ist hiernach *immer* und hinsichtlich des Ausfalls von Bestandseinheiten bei einem systematischen Wirkungsfaktor ein Abgehen vom ursprünglichen theoretischen Untersuchungsobjekt erforderlich. Daher liegt die Schlußfolgerung nahe, daß die beschriebene Datenlage bei einer einzelnen Datenquelle immer die Notwendigkeit einer Umdefinition der Grundgesamtheit erfordert.

Das logische Komplement zum betrachteten Fall ist jedoch nicht vom selben Typ der Unvollständigkeit der Datenquelle. Zwar zerfällt auch hier bezüglich ihrer Erhebbarkeit X_{ij} in vier Teile (siehe Formel 10). Da aber wegen der Komplementarität bezüglich der in der Datenquelle enthaltenen Information genau die Angaben vorhanden sind, die im obigen Fall fehlen, ist nur eine der vier Teilmatrizen eine Nullmatrix, z. B.:

$$\begin{array}{cc}
 \begin{array}{c} \\ \end{array} & \begin{array}{c} x_1 \dots x_p \\ \left[\begin{array}{ccc} x_{11} & \dots & x_{1p} \\ \vdots & & \vdots \\ x_{k1} & \dots & x_{kp} \end{array} \right] \end{array} & \begin{array}{c} \\ \end{array} & \begin{array}{c} x_{p+1} \dots x_m \\ \left[\begin{array}{ccc} x_{1(p+1)} & \dots & x_{1m} \\ \vdots & & \vdots \\ x_{k(p+1)} & \dots & x_{km} \end{array} \right] \end{array} \\
 X_{kp} = \begin{array}{c} 0_1 \\ \vdots \\ 0_k \end{array} & & X_{km} = \begin{array}{c} 0_1 \\ \vdots \\ 0_k \end{array} & & & \\
 & & & & & (10b)
 \end{array}$$

$$\begin{array}{cc}
 \begin{array}{c} \phantom{0_{k+1}} \\ \end{array} & \begin{array}{c} x_1 \dots x_p \\ \left[\begin{array}{ccc} x_{(k+1)1} & \dots & x_{(k+1)p} \\ \vdots & & \vdots \\ x_{n1} & \dots & x_{np} \end{array} \right] \end{array} & \begin{array}{c} \phantom{0_{k+1}} \\ \end{array} & \begin{array}{c} x_{p+1} \dots x_m \\ \left[\begin{array}{ccc} 0 & \dots & 0 \\ \vdots & & \vdots \\ 0 & \dots & 0 \end{array} \right] \end{array} \\
 X_{np} = \begin{array}{c} 0_{k+1} \\ \vdots \\ 0_n \end{array} & & X_{nm} = \begin{array}{c} 0_{k+1} \\ \vdots \\ 0_n \end{array} & & &
 \end{array}$$

11. Für jede andere Kombination von 0_i und x_j gilt Entsprechendes. Wichtig ist, daß von den vier Untermatrizen jeweils drei Nullmatrizen sind.

In diesem Falle sind in der Datenquelle alle Untersuchungseinheiten identifiziert und auch Angaben zu allen theoretisch relevanten Merkmalen vorhanden, *jedoch* fehlen letztere bei *einigen* der identifizierten Bestandseinheiten. Gehen wir davon aus, daß der Ausfall der Angaben zu den Merkmalen x_{p+1}, \dots, x_m bei den Bestandseinheiten $0_{k+1}, \dots, 0_n$ *zufallsbedingt* ist, so könnte X_{ij} durch Einsetzen der merkmalspezifischen Mittelwerte $\bar{x} \cdot j$, $j=p+1, \dots, m$, in einer Stichprobe die Teilmatrix X_{nm} komplettiert werden. Es handelt sich um die Anwendung einer üblichen missing-data-Prozedur.

$$X_{ij} = \begin{matrix} & & x_1 & & x_p & & x_{p+1} & & x_m \\ \begin{matrix} 0_1 \\ \vdots \\ 0_k \\ 0_{k+1} \\ \vdots \\ 0_n \end{matrix} & \left[\begin{array}{cccc} x_{11} & \dots & x_{1p} & x_{1(p+1)} \dots x_{1m} \\ \vdots & & \vdots & \vdots \\ x_{k1} & \dots & x_{kp} & x_{k(p+1)} \dots x_{km} \\ x_{(k+1)1} & \dots & x_{(k+1)p} & \bar{x} \cdot (p+1) \dots \bar{x} \cdot m \\ \vdots & & \vdots & \vdots \\ x_{n1} & \dots & x_{np} & \bar{x} \cdot (p+1) \dots \bar{x} \cdot m \end{array} \right. \end{matrix} \quad (11)$$

Unter der o. a. Annahme des zufallsbedingten Ausfalls der Information erhält man bei diesem Vorgehen ein unverzerrtes Analyseergebnis; wegen der Einschränkung der Freiheitsgrade durch den Einsatz der Mittelwerte allerdings einen im Vergleich zur vollständigen Datenquelle höheren Zufallsfehler.

Sind die Merkmale x_1, \dots, x_p auf der einen Seite mit den Merkmalen x_{p+1}, \dots, x_m auf der anderen Seite korreliert, so schöpft die beschriebene Vorgehensweise die in der Datenquelle enthaltene Information nicht vollständig aus. Hier ist eine Vorgehensweise, die diese Korrelation berücksichtigt, angemessener. Anstelle der merkmalspezifischen Mittelwerte $\bar{x} \cdot j$ können in der Stichprobe Schätzwerte \hat{x}_{ij} eingesetzt werden. Letztere sind zu ermitteln mit einer Schätzfunktion:

$$\hat{x}_{ij} = f(x_{i1}, \dots, x_{ip}) \quad (12)$$

$$i = k+1, \dots, n$$

$$j = p+1, \dots, m$$

Diese Schätzfunktion kann z. B. als Regressionsgleichung

$$\hat{x}_{ij} = b_0 + b_1 x_{i1} + \dots + b_p x_{ip} \quad (13)$$

aus den Daten der beiden Matrizen X_{kp} und X_{km} gewonnen werden:

$$X_{ij} = \begin{matrix} & X_1 & \dots & X_p & X_{p+1} & \dots & X_m \\ 0_1 & \left[\begin{array}{cccc} X_{11} & \dots & X_{1p} & X_{1(p+1)} & \dots & X_{1m} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0_k & X_{k1} & \dots & X_{kp} & X_{k(p+1)} & \dots & X_{km} \\ 0_{k+1} & X_{(k+1)1} & \dots & X_{(k+1)p} & \hat{X}_{(k+1)(p+1)} & \dots & \hat{X}_{(k+1)m} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 0_n & X_{n1} & \dots & X_{np} & \hat{X}_{n(p+1)} & \dots & \hat{X}_{nm} \end{array} \right. & \end{matrix} \quad (14)$$

Voraussetzung ist allerdings, daß die für das Verhältnis X_{kp} zu X_{km} gefundene Funktion auch für das Verhältnis von X_{np} zu X_{nm} gilt. Dies ist so bei einem zufallsbedingtem Ausfall von Informationen, so daß hier ein unverzerrtes Analyseergebnis vorliegt, dessen Zufallsfehler geringer ist als bei der Analyse einer Stichprobe aus der Matrix (11).

Im Falle eines *systematischen* Ausfalls von Einheiten kann in der Regel nicht davon ausgegangen werden, daß die gefundene Funktion auch für das Verhältnis X_{np} zu X_{nm} gilt. Hier besteht die Möglichkeit (12) aufgrund einer Fehlertheorie zu spezifizieren (vgl. allerdings hierzu die Diskussion in 3.1) und \hat{x}_{ij} hiernach zu schätzen. Ist dieser Weg nicht gangbar, so muß die theoretische Fragestellung der Untersuchung entweder hinsichtlich der Untersuchungseinheiten $0_{k+1}, \dots, 0_n$ oder hinsichtlich der Merkmale x_{p+1}, \dots, x_m eingeschränkt werden.

4. Die Verknüpfung unvollständiger Datenquellen

Unsere bisherigen Überlegungen zeigen, daß aus einer *einzelnen unvollständigen Datenquelle* Parameter der Grundgesamtheit nicht ohne eine *systematische Verzerrung* geschätzt werden können. Eine Ausnahme bildet lediglich der zufallsbedingte Ausfall von Untersuchungseinheiten. Allenfalls können aus Vermutungen über die Art der Fehlerentstehung und Fehlerwirkung für bestimmte Datenlagen Korrekturen der jeweiligen systematischen Verzerrung abgeleitet werden.

Hinsichtlich der spezifischen Unvollständigkeit einer Datenquelle liegt es nahe, einer *anderen Datenquelle*, die ihrerseits unvollständig sein kann, die fehlenden Informationen zu entnehmen. Im folgenden untersuchen wir nur die Kombination von zwei Datenquellen. Eine Ausweitung dieser Überlegungen auf die Verknüpfung von mehreren Datenquellen ist möglich. Die hierbei auftretenden speziellen Probleme eines „multiple linkage“ werden jedoch hier nicht behandelt (vgl. die knappe Übersicht zu dieser Problematik von Winchester, 1973, 145–149, und die dort angegebene Literatur).

4.1 Probleme der Verknüpfung komplementärer Datenquellen

Untersuchen wir zunächst die Verknüpfung komplementärer Datenquellen. Zwei Datenquellen sind zueinander komplementär, wenn in der einen Datenquelle genau diejenigen Informationen enthalten sind, die in der anderen Datenquelle fehlen. Für eine vorliegende Datenquelle ist somit eine weitere Datenquelle zu suchen, die die-

sen Anforderungen genügt. Bezogen auf die beiden polaren Typen der Unvollständigkeit von Datenquellen bedeutet dies im Falle *nichtidentifizierter Bestandseinheiten* (3.1), daß man eine weitere Datenquelle sucht, in der die fehlenden Untersuchungseinheiten $0_{k+1}, \dots, 0_n$ identifiziert werden können. Da X_{k_m} der ersten und X_{l_m} der zweiten Datenquelle jeweils entnommen werden kann, liegt ein hinsichtlich des theoretischen Untersuchungsobjektes vollständiger Datenbestand vor. Desgleichen wird im Falle *fehlender Angaben bei identifizierten Bestandseinheiten* eine Datenquelle gesucht, die Angaben zu den Merkmalen x_{p+1}, \dots, x_m enthält. X_{n_p} kann aus der einen und X_{n_q} aus der anderen Datenquelle erhoben werden. Auch hier ist der Datenbestand hinsichtlich der theoretischen Fragestellung komplett¹². Hinsichtlich des Problems der *Parameterschätzung* ist die Verbindung komplementärer Datenquellen der Idealfall. Durch die Kombination unvollständiger Quellen erhält man einen *vollständigen Datenbestand*, der keine besonderen Schätzprobleme aufwirft (vgl. 2).

Wenn wir die eben geschilderte Sachlage als unproblematisch bezeichnen, wird unterstellt, daß die Verknüpfung der Quellen gelungen ist. Das Verknüpfen von Informationen aus verschiedenen Datenquellen ist aber seinerseits äußerst problematisch. Zwar ist nach Verknüpfung komplementärer Quellen das Problem des „source bias“ gelöst; darüber hinaus muß aber auch mit *Verzerrungen* gerechnet werden, die *durch den Verknüpfungsprozeß* selbst zustande kommen („linkage bias“).

Mit dem Problem der Abschätzung der durch die besondere Art des „linkage process“ eingeführten Verzerrung wollen wir uns in diesem Abschnitt beschäftigen. Die Art des Verknüpfungsprozesses hängt dabei vom jeweiligen Typ der Unvollständigkeit der ursprünglichen Datenquelle ab.

4.1.1 Die Identifizierung von Untersuchungseinheiten

Wir untersuchen den Fall identifizierter Bestandseinheiten mit fehlenden Merkmalsangaben (3.2)¹³. In der ursprünglichen Datenquelle liegt eine Liste von n Untersuchungseinheiten vor, je Untersuchungseinheit sind in dieser Quelle Angaben zu den Merkmalen x_1, \dots, x_p enthalten. Nach der theoretischen Fragestellung erforderliche Angaben zu weiteren Merkmalen x_{p+1}, \dots, x_m sind in einer anderen Datenquelle ebenfalls auf die einzelnen Einheiten bezogen enthalten. Die Untersuchungseinheiten $0_1, \dots, 0_n$ der einen Datenquelle sind somit in der anderen Datenquelle zu identifizieren (vgl. Hershberg et al., 1976; Winchester, 1973; 1970).

Beispielsweise könnte ein „Handbuch der Banken“ nur Angaben zum Eigenkapital, zur Einlagenhöhe usw. aller Banken enthalten. In einem Zeitschriftenartikel wären für den Bankenbereich die Anzahl der Personalverflechtungen mitgeteilt. Sollen nun Aussagen über den Zusammenhang zwischen Höhe des Eigenkapitals und der Anzahl der Personalverflechtungen gemacht werden, so sind beide Quellen in einen

12. Für die beiden betrachteten Mischtypen der Unvollständigkeit einer Datenquelle ist entsprechend vorzugehen, so daß die jeweiligen Nullmatrizen aufgefüllt werden können. Die Verknüpfungsprobleme sind hier größer.

13. Die Identifizierungskriterien bei der Hinzugewinnung von Bestandseinheiten, die in der Datenquelle nicht vorhanden sind, sind raum-zeitliche Randbedingungen, die das Untersuchungsobjekt definieren (vgl. Privatbanken versus öffentliche Banken, die beide zusammen den Bereich der Banken abdecken).

Analysebestand zusammenzufassen. Man wird dann für jede der Banken, die in dem „Handbuch“ enthalten sind, im Zeitschriftenartikel die Anzahl der für die einzelne Bank zutreffenden Personalverflechtungen heraussuchen.

4.1.2 Identifizierungskriterien und ihre Eindeutigkeit

Grundlegendes Erfordernis einer jeden Verknüpfung verschiedener Datenquellen ist das Vorliegen einer *Mindestmenge gemeinsamer überlappender Informationen*. Nur solchen überlappenden Informationsmengen können Identifizierungskriterien entnommen werden. Diese Kriterien sind Merkmale, an denen erkannt werden kann, ob eine Untersuchungseinheit der einen Datenquelle identisch ist mit einer Untersuchungseinheit der anderen Datenquelle. Bildlich gesprochen sind Identifizierungskriterien die Berührungspunkte der Datenquellen, ohne die es nicht zu einer Verknüpfung kommen kann. Es liegt auf der Hand, daß es sich um *eindeutige Kriterien* handeln muß. Auf den ersten Blick ist der *Eigennamen einer Untersuchungseinheit* ein solches eindeutiges Identifizierungskriterium. Somit wird „record linkage“ zunächst immer als „nominal record linkage“ (Wrigley, 1973, 1) verstanden. Die Identifikation der Untersuchungsobjekte erfolgt in der Regel über den Eigennamen (vgl. Wrigley und Schofield, 1973). Andere Identifizierungskriterien sind möglich, so bei Personen das Alter, das Geschlecht, der Wohnort u. ä. In vielen Fällen ist jedoch der Name einer Untersuchungseinheit kein eindeutiges Kriterium. Dies gilt insbesondere für die historische Sozialforschung. So berichtet Winchester (1970; 1973) über entsprechende Schwierigkeiten für den englischen Sprachbereich. Diese Schwierigkeiten ergeben sich insbesondere wegen der unterschiedlichen Schreibweise der Zunamen, unterschiedlicher Abkürzungen der Vornamen u. ä. in verschiedenen Datenquellen. In dieser Hinsicht geben für andere Sprachbereiche Blayo (1973) und Herlihy (1973) interessante Belege. Über quasi-alphabetische Codes wird versucht, die Eindeutigkeit zu erhöhen (vgl. Winchester, 1970, 114–117).

Eine weitere Möglichkeit, die Eindeutigkeit zu steigern, besteht in der gleichzeitigen Verwendung mehrerer Identifikationskriterien. So können neben dem Eigennamen, z. B. bei Individuen, Beruf, Wohnort und Alter herangezogen werden. Damit läßt sich zwar die Eindeutigkeit der Zuordnung erhöhen (allerdings mit fallendem Grenzertrag für jedes neu hinzukommende Kriterium), eine völlige Eindeutigkeit der Zuordnung kann jedoch fast nie erreicht werden.

Hier stößt man auf ein Dilemma (Wrigley, 1973, 5–6): Die Verknüpfung von Datenquellen ist nur dann sinnvoll, wenn die hinzukommende Datenquelle neue Informationen zu der bereits vorhandenen hinzufügt und dabei zwischen einer richtigen und einer falschen Verknüpfung unterschieden werden kann. Diese beiden Bedingungen verhalten sich jedoch antagonistisch zueinander:

- Je weniger Informationen die Datenquellen gemeinsam haben, desto vorteilhafter ist ihre Verknüpfung, desto größer ist jedoch die Unsicherheit hinsichtlich der Richtigkeit der Verknüpfung.
- Je mehr Informationen die Datenquellen gemeinsam haben, desto weniger vorteilhaft ist ihre Verknüpfung, desto sicherer kann man jedoch hinsichtlich der Richtigkeit der Verknüpfung sein.

Die Steigerung der Gewißheit der Verknüpfungen („certainty“) führt somit zur Verringerung der Brauchbarkeit der Daten („utility“).

4.1.3 Fehlverknüpfungen und Verzerrungen („linkage bias“)

Mangelnde Eindeutigkeit impliziert die Möglichkeit von Fehlverknüpfungen zwischen den Datenquellen. Eine Fehlverknüpfung ist gegeben, wenn eine der folgenden drei Möglichkeiten vorliegt (Nathan, 1967, 455):

- (a) die Angabe zur Untersuchungseinheit 0_i der Datenquelle Q_1 wird in der Datenquelle Q_2 mit einer Angabe verknüpft, die jedoch eine Einheit betrifft, die nicht zur Datenquelle Q_1 gehört („erroneous matches“).
- (b) die Angabe zur Untersuchungseinheit 0_i der Datenquelle Q_1 wird in der Datenquelle Q_2 *nicht* mit einer Angabe verknüpft, obwohl diese zu 0_i gehört („erroneous non-matches“).
- (c) die Angabe zur Untersuchungseinheit 0_i der Datenquelle Q_1 wird in der Datenquelle Q_2 mit einer Angabe verknüpft, die zur Untersuchungseinheit 0_k gehört. 0_k ist eine Untersuchungseinheit des in Frage stehenden Untersuchungsobjektes („mismatches“).

Welche Verzerrungen des kombinierten Datenbestandes resultieren jeweils aus diesen Formen einer Fehlverknüpfung?

Fehlverknüpfungen des Typs (a) („erroneous matches“) führen dem kombinierten Datenbestand Informationen hinzu, die nicht zur Grundgesamtheit des Untersuchungsobjektes gehören. Demgegenüber werden durch Fehlverknüpfungen des Typs (b) („erroneous non-matches“) in den kombinierten Datenbestand nicht alle Informationen aufgenommen, die zur Grundgesamtheit des Untersuchungsobjektes gehören. Fehlverknüpfungen des Typs (c) („mismatches“), führen nicht zu einer unangemessenen Erweiterung bzw. Einengung des Datenbestandes. Während bei univariaten Schätzungen ein Fehlerausgleich auf der Hand liegt, sind die Schätzungen aller bi- bzw. multivariaten Größen durch „mismatches“ allerdings verzerrt. (Hierauf hat z. B. auch Hammarberg, 1977b, 16, hingewiesen). Wir werden auf diesen Typ (c) im weiteren nicht eingehen (siehe auch die Argumentation von Marks et al., 1974, 101-103).

4.1.4 Die Verzerrung des gesamten Datenbestandes

Im Normalfall der Verknüpfung der Informationen von Datenquellen muß immer mit einer mehr oder weniger großen Zahl von Fehlverknüpfungen gerechnet werden. Entscheidend hinsichtlich der Verzerrung des gesamten kombinierten Datenbestandes ist jedoch nicht die einzelne Fehlverknüpfung, sondern die *Gesamtheit* aller vorkommenden Fehlverknüpfungen. Hierbei ist von Bedeutung nicht die Summe der Fehlverknüpfungen sondern ihre Differenz, der sog. Nettofehler:

$$e_m - e_{nm} . \quad (15)$$

Hierbei bedeuten e_m = Anzahl der „erroneous matches“ und e_{nm} = Anzahl der „erroneous non-matches“.

Treten Fehlverknüpfungen innerhalb eines Verknüpfungsprozesses zufällig und unabhängig voneinander auf, so gleichen sich die Einzelverzerrungen aus. Der Erwartungswert des Nettofehlers ist Null:

$$E(e_m - e_{nm}) = 0 . \quad (16)$$

Hinsichtlich des Gesamtbestandes liegt keine Verzerrung vor.

Allerdings unterscheiden sich Verknüpfungsprozesse danach, ob sie die erste oder die zweite Art von Fehlerverknüpfungen, „erroneous matches“ oder „erroneous non-matches“, begünstigen. So wird eine „strenge“ Verknüpfungsregel die zweite, eine „lockere“ Verknüpfungsregel die erste Art der Fehlerverknüpfung fördern. Es besteht geradezu ein Dilemma insoweit, als das Bestreben, die Anzahl der Fehlerverknüpfungen der einen Art zu vermindern, zwangsläufig zur Erhöhung der Anzahl derjenigen der anderen Art führt („matching dilemma“ nach El-Khorazaty, 1975 und El-Khorazaty et al. 1976).

Ist $e_m < e_{nm}$ („strenge“ Verknüpfungsregel), so liegt eine unzulässige Verkleinerung des Datenbestandes im Sinne der theoretischen Fragestellung vor; ist dagegen $e_m > e_{nm}$ („lockere“ Verknüpfungsregel), so kommen dadurch Einheiten in den Datenbestand, die nicht zum Untersuchungsobjekt gehören.

Somit hängt das Ausmaß der Verzerrung des integrierten Datenbestandes vom spezifischen Vorgehen bei der Kombination der beiden Datenquellen, von der *Verknüpfungsregel*, ab. Konsequenterweise existieren eine Reihe von Versuchen, *optimale* Verknüpfungsregeln abzuleiten. Dabei wird unter einer optimalen Verknüpfungsregel die Vorgehensweise verstanden, die das Ausmaß von Fehlerverknüpfungen minimiert. An dieser Stelle wollen wir uns nicht weiter mit diesem Bereich beschäftigen. Für eine Übersicht siehe Winchester, 1973, 142–145. (Für die Originalarbeiten siehe Fellegi u. Sunter, 1969; Du Bois, 1965; 1969; Nathan 1967; Tepping 1968).

4.1.5 Auswahl der Einheiten und Verknüpfung der Datenquellen

Die Verknüpfung von zwei Datenquellen ist an sich kein Auswahlproblem.

Hinsichtlich der Massenhaftigkeit von Bestandseinheiten in der ursprünglichen Datenquelle sind Überlegungen zur Stichprobenziehung, jedoch eng mit denen der Verknüpfung von Datenquellen verbunden. Es wäre ökonomisch, technisch und methodisch nicht vertretbar, für die in Frage stehende Grundgesamtheit die notwendigen Verknüpfungen durchzuführen und dann erst in der üblichen Weise eine Stichprobe zu ziehen.

Demgegenüber wäre aber die Ziehung von Zufallsstichproben jeweils in einer der beiden zu verknüpfenden Datenquellen ebenfalls nicht sinnvoll, da es höchst unwahrscheinlich ist, daß die in den einzelnen Datenquellen jeweils gezogenen Einheiten übereinstimmen.

Andererseits erscheint es sinnvoll, aus der ursprünglichen Datenquelle eine Zufallsstichprobe zu entnehmen und dann die Einheiten dieser Stichprobe in der zweiten Datenquelle zu identifizieren. Aber auch diese Vorgehensweise ist sehr arbeitsaufwendig, wenn man bedenkt, daß dann alle in der zweiten Datenquelle enthaltenen Einheiten zu sichten sind. Eine bestimmte Ordnung dieser Einheiten etwa in alphabetischer Reihenfolge der Namen könnte sicherlich den Aufwand verringern.

Diesen Ordnungsgesichtspunkt legt Wrigley (1973, 14) zugrunde, wenn er vorschlägt, die Einheiten *beider* Quellen in alphabetischer Reihenfolge der Eigennamen zu ordnen und dann in der ersten Quelle Gruppen mit einem bestimmten Anfangsbuchstaben bzw. Buchstabenkombination der Namen zufällig zu ziehen und diese mit den Einheiten der zweiten Quelle zu verknüpfen. Wrigley weist darauf hin, daß dies z. B. problematisch sei bei weiblichen Personen, die nach der Heirat ihren Namen wechseln. Beachtenswerter erscheint der Einwand, daß der Anfangsbuchstabe eines Namens mit anderen Merkmalen der Untersuchung, z. B. Schichtzugehörigkeit,

korreliert sein kann. Es wäre also zu überprüfen, welches Merkmal in dieser Hinsicht am neutralsten ist und dann dieses als Erhebungsmerkmal zu verwenden.

4.2 Probleme der Verknüpfung nichtkomplementärer Datenquellen

Im Normalfall der Datenerhebung aus Massenakten liegen keine komplementäre sondern nichtkomplementäre Quellen vor. Das Verhältnis zweier Datenquellen soll als *nichtkomplementär* bezeichnet werden, wenn die in der einen Datenquelle fehlenden Informationen nicht voll aus der anderen Datenquelle ergänzt werden können. Das Ergebnis der Verknüpfung der beiden unvollständigen Datenquellen ist ein integrierter Datenbestand, der zwar mehr Informationen enthält als jeweils eine einzelne der beiden Datenquellen aber seinerseits ebenfalls ein unvollständiger Datenbestand ist.

Dies führt zu Schätzproblemen, die je nach dem Typ der Unvollständigkeit des integrierten Datenbestandes zu lösen sind. Somit handelt es sich hier gegenüber der Analyse einer einzelnen Datenquelle keineswegs um ein neues Problem. Man kann den neuen integrierten Datenbestand als einzelne unvollständige Datenquelle nehmen und entsprechend analysieren. Jedoch kulminieren bei einem Datenbestand, der durch Verknüpfung zweier nichtkomplementärer Quellen entstanden ist, die bislang behandelten Einzelprobleme.

Wir greifen im folgenden aus den in der Literatur behandelten Fällen (z. B. Kindahl, 1962; Hammarberg, 1971) einen heraus, an dem wir die besondere Problematik integrierter Datenbestände aus nichtkomplementären Datenquellen beispielhaft illustrieren wollen. Dabei geht es darum, zu zeigen, in welcher Weise versucht wird, verzerrungsfreie Schätzungen der Parameter der Grundgesamtheit zu erreichen. Verzerrungen können jedoch bei dieser Datenkonstellation nicht völlig vermieden werden. Es wird jedoch angestrebt, die Verzerrungen im integrierten Datenbestand auf ein Minimum zu beschränken.

4.2.1 Die Struktur des Datenbestandes

Kindahl (1962) beschreibt folgenden Fall: Geschätzt werden soll die Höhe der Einlagen privater Banken an der Wende der 60er/70er Jahre des 19. Jahrhundert in den USA. Eine Liste aller Privatbanken mit einer entsprechenden Angabe existiert nicht. Aus Veranlagungsunterlagen der Steuerbehörde kann aber eine Liste von Privatbanken mit der jeweiligen Einlagenhöhe erstellt werden. Daneben liegt ein Bankenverzeichnis mit weiteren Angaben — jedoch nicht zur Einlagenhöhe — vor. Ein Teil der Banken ist sowohl in der Veranlagungsliste als auch im Bankenverzeichnis aufgeführt, ein Teil der Banken ist aber entweder nur in der Veranlagungsliste oder im Bankenverzeichnis enthalten.

Allgemein kann das Verhältnis der beiden Datenquellen zueinander wie folgt beschrieben werden: Die Datenquelle Q_1 enthält nur für einen Teil der zum Untersuchungsbereich gehörenden Untersuchungseinheiten. Es existiert eine weitere Datenquelle Q_2 . In ihr fehlen Angaben zu den interessierenden Merkmalen. Die Vereinigungsmenge von Q_1 und Q_2 ist nicht die Gesamtmenge der zum Untersuchungsbebereich gehörenden Untersuchungseinheiten. Es existiert eine Schnittmenge von Q_1 und Q_2 , die diejenigen Untersuchungseinheiten enthält, die sowohl in der einen wie auch in der anderen Datenquelle enthalten sind.

Diese beiden Datenquellen sind zu einem Datenbestand zu verknüpfen. Einerseits sind in diesem Datenbestand nicht alle Einheiten enthalten, die hinsichtlich der theoretischen Fragestellung zu identifizieren sind, andererseits fehlen für einen Teil der identifizierten Einheiten die Angaben zu dem interessierenden Merkmal.

4.2.2 Das Schätzproblem

Wie kann aus diesen Daten die Gesamthöhe der Einlagen privater Banken geschätzt werden? Hätten wir entsprechende Angaben für alle Banken, so wäre die Höhe der Einlagen aus einer Stichprobe als die Summe der einzelnen Einlagenhöhe pro Bank, x_{ij} , leicht zu berechnen:

$$\sum_{i=1}^n x_{ij}$$

Nun sind aber nicht alle Untersuchungseinheiten im Datenbestand identifiziert, so daß diese Möglichkeit ausscheidet. Wir können aber die gesamte Einlagenhöhe auch als das Produkt der Anzahl der Banken und der durchschnittlichen Einlagenhöhe pro Bank bestimmen:

$$\sum_{i=1}^n x_{ij} = \bar{x} \cdot j \cdot n \quad (17)$$

Das Schätzproblem wird so auf die Ermittlung dieser beiden Komponenten verlagert.

Nun haben wir bereits für einen einzelnen unvollständigen Datenbestand mit nichtidentifizierten Einheiten gezeigt, daß die Stichprobenmittelwerte $\bar{x} \cdot j$ verzerrte Schätzungen der entsprechenden Grundgesamtheitparameter $\mu \cdot j$ sind (Formel 6):

$$\mu \cdot j = \bar{x} \cdot j - \Delta \cdot j$$

Andererseits kann argumentiert werden, daß die Veranlagungsliste — wegen der Steuerbemessungsgrenze — nur Privatbanken ab einer gewissen Größe enthält. Wenn wir $\bar{x} \cdot j$ als Schätzwert für $\mu \cdot j$ heranziehen, so ist in Rechnung zu stellen, daß $\bar{x} \cdot j$ zu hoch ist. Da wir aber eine weitere genauere Bestimmung nicht vornehmen können, müssen wir uns mit $\bar{x} \cdot j$ als Schätzwert für $\mu \cdot j$ begnügen und zur Berechnung der Gesamthöhe der Einlagen in Formel (17) einsetzen. Da aber wegen der nichtidentifizierten Einheiten die Größe n unbekannt ist, kann auch eine solche Berechnung nicht ausgeführt werden.

Nun besteht aber die Möglichkeit mit Hilfe von Stichproben aus den Datenquellen Q_1 und Q_2 n zu schätzen. Die gemeinsame Verteilung der gezogenen Einheiten hinsichtlich ihrer Zugehörigkeit zu den beiden Datenquellen ist in *Tabelle 1* dargestellt. Nur die Größen in den schraffierten Feldern der Tabelle können aus den Stichproben entnommen werden.

Wenn nun die Identifikation einer Einheit in der Datenquelle Q_2 unabhängig ist von ihrer Identifikation in der Datenquelle Q_1 , dann ist der Anteil der Gesamtpopu-

Tabelle 1: Gemeinsame Verteilung hinsichtlich der Zugehörigkeit zu den Datenquellen Q_1 und Q_2

		In der Datenquelle Q_2		Σ
		identifiziert	nicht identifiziert	
In der Datenquelle Q_1	identifiziert	n_{11}	n_{12}	n_1
	nicht identifiziert	n_{21}	n_{22}	$n - n_1$
Σ		n_2	$n - n_2$	n

lation der in Q_2 identifiziert wird, gleich dem Anteil der durch Q_2 in der Subpopulation der Quelle Q_1 identifizierten Einheiten:

$$\frac{n_2}{n} = \frac{n_{11}}{n_1} \quad (18)$$

Hieraus ergibt sich für n die Schätzung

$$\hat{n} = \frac{n_1 n_2}{n_{11}} \quad (19)$$

\hat{n} kann so zusammen mit \bar{x}_j in Formel (17) eingesetzt und die Gesamtheit der Einlagen berechnet werden.

Nun ist aber der Umstand, Bestandseinheit in Q_1 zu sein, sicher nicht unabhängig vom Umstand in Q_2 identifiziert zu werden. So waren wahrscheinlich bei unseren Privatbanken gleiche Faktoren ausschlaggebend sowohl für die Aufnahme in die Veranlagungsliste als auch für die Aufnahme in das Bankenverzeichnis.

Nun kann gezeigt werden, daß positive Abhängigkeit der gleichzeitigen Identifikation in verschiedenen Datenquellen dazu führt, daß die Höhe von \hat{n} zu niedrig geschätzt wird. Dies ist zunächst als ein Mangel zu werten. Hinsichtlich des Ausgangsproblems der Schätzung der Gesamthöhe der Einlagen und angesichts der umgekehrten Verzerrung von \bar{x}_j ist dies von Vorteil. Es liegt ein Fehlerausgleich vor, so daß

per Saldo die Gesamthöhe der Einlagen in der Grundgesamtheit richtig getroffen wird.

Wir wollen unsere Darstellung an dieser Stelle abbrechen. Überlegungen hinsichtlich der Wirkung von Fehlverknüpfungen könnten noch angestellt und weitere Möglichkeiten des Verringerns der Verzerrung erörtert werden (Vgl. Hammarberg 1977 b; El-Khorazaty et al. 1976). Es geht hier aber nur um die Darlegung einiger Probleme der Analyse verknüpfter nichtkomplementärer Datenquellen.

5. Zusammenfassung

Ausgehend vom *Idealfall* einer *vollständigen Datenquelle*, in der alle der theoretischen Fragestellung gemäßen Informationen enthalten sind, wird der *Normalfall* historisch-sozialwissenschaftlicher Forschung, die einzelnen *unvollständige Datenquelle*, untersucht. In einer solchen Datenquelle fehlen entweder bestimmte Bestandseinheiten, so daß für entsprechende Untersuchungseinheiten keine Informationen vorliegen, oder es sind zwar alle Bestandseinheiten identifiziert, doch fehlen zu bestimmten Merkmalen die Informationen. Daneben sind u. U. in einer Datenquelle sowohl Bestandseinheiten als auch Angaben zu bestimmten Merkmalen nicht vorhanden.

Die Überlegungen beschränken sich zunächst auf eine einzelne unvollständige Datenquelle. Hinsichtlich der spezifischen Unvollständigkeit liegt es aber nahe, einer anderen Datenquelle, die ihrerseits unvollständig sein kann, die fehlenden Informationen zu entnehmen. Infolgedessen werden im weiteren Probleme der Verknüpfung von Datenquellen erörtert. Dabei geht es zunächst um die Kombination von *komplementären Datenquellen*. Betrachtet werden u. a. Probleme der Identifizierung von Einheiten, Identifizierungskriterien und ihre Eindeutigkeit sowie Fehlverknüpfungen und Verzerrungen der im Datenbestand enthaltenen Information. Schließlich steht die Verknüpfung *nichtkomplementärer Datenquellen* im Mittelpunkt der Betrachtung.

Der vorliegende Beitrag deckt ohne Zweifel nicht die gesamte Problematik der Datenerhebung aus Massenakten ab. Insbesondere fehlen Überlegungen zur Verknüpfung mehrerer Datenquellen sowie die Einbeziehung des Netzwerk-Aspektes beim Problem der Auswahl von Einheiten. Der Leser möge daher den Beitrag als ersten Versuch einer Systematisierung werten.

Literatur

- Blayo, Yves: Name variations in a village in Brie, 1750-1860, in: E. A. Wrigley (Hrsg.), *Identifying People in the Past*, London (1973)
- Bremer, Stuart A.: Statistics sans Samples: Two largely undiscussed problems of statistical analysis, in: QUANTUM-Conference (1977)
- Du Bois, N.S.D.: A solution to the problem of linking multivariate documents, in: *Journal of American Statistical Association*, Nr. 64 (1969), 163-74
- Du Bois, N.S.D.: A document linkage program for digital computer, in: *Behavioural Science*, Nr. 10 (1965), 312-19

- El-Khorazaty, M. N.: Methodological Strategies For the Analysis of Categorical Data From Multiple-Record Systems, University of North Carolina Institute of Statistics, in: Mimeo Series, Nr. 1019 (1975)
- El Khorazaty, M. N., et al.: A Review of Methodological Strategies for Estimating the Total Number of Events With Data From Multiple-Record Systems, University of North Carolina Institute of Statistics, in: Mimeo Series, Nr. 1095 (November 1976)
- Felligi, I. P. und A. B. Sunter: A Theory for Record Linkage, in: Journal of the American Statistical Association, Nr. 64 (1969), 1183-1210
- Hammarberg, Melvyn: Record Linkage and Sampling Strategies, in: QUANTUM-Conference (1977)
- Hammarberg, Melvyn A.: Designing a Sample from Incomplete Historical Lists, in: American Quarterly, Nr. 23 (1971), 542-561
- Herlihy, David: Problems of record linkages in Tuscan fiscal records of the fifteenth century, in: E. A. Wrigley (Hrsg.), Identifying People in the Past, London (1973)
- Hershberg, Th, A. Burstein u. R. Dockhorn: Record Linkage, in: Historical Methods Newsletter, Nr. 9 (1976), 137-163
- Hertel, Bradley R.: Minimizing error variance introduced by missing data routines in survey analysis, in: Sociological Methods and Research, Nr. 4 (1976), 459-474
- Jeidels, Otto: Das Verhältnis der deutschen Großbanken zur Industrie mit besonderer Berücksichtigung der Eisenindustrie, in: Staats- und sozialwissenschaftliche Forschungen, herausg. von Gustav Schmoller und Max Sering, Bd. 24, Leipzig: Dunker & Humblot, (1905)
- Kindahl, James K.: Estimation of means and totals from finite populations of unknown size, in: Journal of the American Statistical Association, Nr. 57 (1962), 61-91
- Lenin, W. I.: Der Imperialismus als höchstes Stadium des Kapitalismus, Petrograd (1917)
- Lösel, Friedrich u. Werner Wüstendörfer: Zum Problem unvollständiger Datenmatrizen in der empirischen Sozialforschung, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, Nr. 26 (1974), 342-357
- Marks, Eli S., William Seltzer und Karol J. Krotki: Population Growth Estimation, in: A Handbook of Vital Statistics Measurement, New York: The Populative Council (1974)
- Murphey, Murray G.: Our Knowledge of the Historical Past, Indianapolis und New York: Bobbs-Merrill (1973)
- Nathan, Gad: Outcome probabilities for a record matching process with complete invariant information, in: Journal of American Statistical Association, Nr. 62 (1967), 454-69
- Schofield, Roger S.: Sampling in historical research, in: Wrigley, E. A. (Hrsg.), Nineteenth-century society. Essays in the use of quantitative methods for the study of social data, Cambridge: University press (1972), 146-190
- Tepping, B. J.: A model for optimum linkage of records, in: Journal of the American Statistical Association, Nr. 63 (1968), 1321-32
- Wilks, S. S.: Moments and Distributions of Estimates of Population Parameter From Fragmentary Samples, in: Annals of Mathematical Statistics, Nr. 3 (1932)

- Winchester, Ian: On referring to ordinary historical persons, in: E. A. Wrigley (Hrsg.), *Identifying People in the Past*, London (1973)
- Winchester, Ian: A brief survey of the algorithmic, mathematical and philosophical literature relevant to historical record linkage, in: E. A. Wrigley (Hrsg.), *Identifying People in the Past*, London (1973)
- Winchester, Ian: The Linkage of Historical Records by Man and Computer: Techniques and Problems, in: *Journal of Interdisciplinary History*, Nr. 1 (1970), 107–124
- Wrigley, E. A.: Introduction, in: E. A. Wrigley (Hrsg.), *Identifying People in the Past*, London (1973)
- Wrigley, E. A. and R. S. Schofield: Nominal record linkage by computer and the logic of family reconstitution, in: E. A. Wrigley (Hrsg.), *Identifying People in the Past*, London (1973)