

Rainer Lenz und Markus Zwick*

Integrierte Mikrodatenfiles – Methoden zur Verknüpfung von Einzeldaten

1 Einleitung

Die breite Basis sozio-ökonomischer Erhebungen in der amtlichen Statistik ermöglicht Sozialwissenschaftlern und Ökonomen eine differenzierte Beschreibung und Analyse der besonderen Lebensverhältnisse gesellschaftlicher Gruppen. Jede der vorliegenden Statistiken – insbesondere der Mikrozensus (MZ), die Einkommens- und Verbrauchsstichprobe (EVS) und die Lohn- und Einkommensteuerstatistik – bildet dabei einen eigenen Ausschnitt der sozio-ökonomischen Verhältnisse in Deutschland ab. Der Blickwinkel dieser Statistiken wird durch ihren Erhebungszweck festgelegt und richtet sich insofern auf speziell zugeschnittene Grundgesamtheiten und Variablenauswahlen. Der hohe Grad an Spezialisierung hat allerdings seinen Preis: Vergleichende Analysen können sich nur auf Variablen beziehen, die in allen einbezogenen Statistiken vorliegen, und allgemeine Analysen etwa der Einkommensverteilung sind nur innerhalb einer möglichst umfassenden Schnittmenge ihrer Grundgesamtheiten sinnvoll. Mit dem vorliegenden Bestand amtlicher Statistiken sind beide Arten von Analysen nur für wenige und zumeist stark eingeschränkte Fragestellungen durchführbar.

Zur Überwindung dieser Problematik wird in der empirischen Sozialforschung die Integration verschiedener Mikrodatenbestände herangezogen. D. h., die in unterschiedlichen Statistiken vorliegenden Mikrodaten werden mit einem geeigneten mathematischen Verfahren zusammengeführt und stehen danach als eigenständiges Datenmaterial mit breiterem Merkmalskanon und einer erweiterten Anzahl von Merkmalsträgern für Analysen zur Verfügung.

Die Integration der Datenbestände basiert auf einem Mix mathematisch-statistischer Methoden (u. a. Statistical Matching Verfahren und verteilungsbasierte multiple Imputation), die in einem einheitlichen theoretischen Rahmen kombiniert werden. Auf beiden Gebieten existieren in der Literatur eine Reihe konkurrierender Ansätze, die in Bezug auf ihre Anwendbarkeit bei dem vorliegenden Datenmaterial zu erproben sind. Statistisches Kriterium für den Einsatz solcher Verfahren sind die Eigenschaften von Schätzfunktionen über dem integrierten Material (insb. Erwartungstreue und Varianz von Schätzfunktionen der ersten Verteilungsmomente). Diese Eigenschaften sind in der Regel nicht analytisch herleitbar, sondern nur in umfangreichen Monte-Carlo-Simulationen zu ermitteln.

2 Die Konstruktionsidee von „Integrierten Mikrodatenfiles“

Eine Grundgesamtheit, die sich zweidimensional definiert über ihre Merkmale und Merkmalsträger, kann durch eine einzelne Erhebung niemals vollständig beschrieben werden. Darüber hinaus ist jede Erhebung mit einem Zweck verbunden, der dazu führt, dass nur zweckdienliche Angaben erfasst werden und alle anderen möglichen Angaben negiert werden.

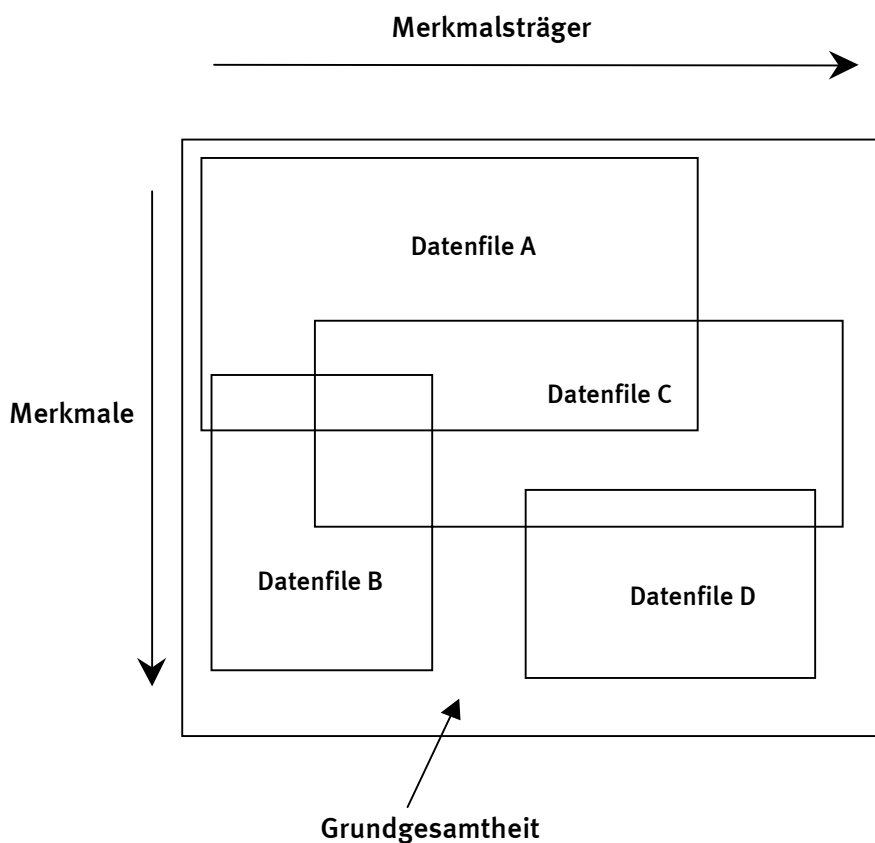
Aus diesem Grund ist leicht ersichtlich, dass es nicht möglich ist, einen allumfassenden Datensatz mittels Erhebung zu generieren. Auch sekundärstatistische Angaben sind hier zunächst keine Hilfe. In der Regel erfassen diese Datenbestände zwar eine Vielzahl von Merkmalsträgern, so z. B. die Steuerdaten oder die Daten der Einwohnermeldeämter, aber die Merkmale sind primär auf das Erfassungsziel, Steuererhebung bzw. Einwohnernachweis, ausgelegt, so dass weitergehende statistische Analysen oftmals nur beschränkt möglich sind.

Die hier geschilderte Problematik lässt sich mittels „Integrierter Mikrodatenfiles“ entschärfen. Unter der Annahme, dass verschiedene primäre und sekundäre Statistiken mit ihren Einzeldaten dieselbe Grundgesamtheit beschreiben, könnte ein Datensatz, der alle Erhebungen umfasst, ein möglichst breites bis

* Dr. Rainer Lenz, Forschungsdatenzentrum des Statistischen Bundesamtes
Markus Zwick, Leiter des Forschungsdatenzentrums des Statistischen Bundesamtes

vollständiges Bild einer Grundgesamtheit vermitteln. Die einzelnen statistischen Quellen erweitern den Blick auf die Grundgesamtheit im oben beschriebenen Maße in zwei Dimensionen. Die Dimension der Merkmale erweitert sich, wenn zwei Erhebungen mit unterschiedlichen Merkmalen eine identische Grundgesamtheit beschreiben. So z. B. die EVS und der MZ, die beide die Gruppe der Haushalte mit unterschiedlichen und gemeinsamen Merkmalen erfassen. Zur Erweiterung der Dimension der Merkmalsträger werden Datenbestände zusammengeführt, die jeweils nur einen Teilbereich der Gesamtpopulation beschreiben und im Idealfall sogar keine gemeinsamen Merkmalsträger besitzen (und damit im mathematischen Sinne disjunkt sind). So beschreibt die Sozialhilfestatistik in der Regel Personen bzw. Haushalte ohne steuerliches Einkommen, die Einkommensteuerstatistik hingegen nur Haushalte mit steuerpflichtigen Einkünften. Im allgemeinen allerdings überlappen sich sowohl Merkmalsträger als auch Merkmale verschiedener Erhebungen. Eine Überlappung der Merkmale ist eine notwendige Bedingung zur Integration verschiedener Datenbestände; ob sie auch hinreichend ist, hängt von der Qualität der gemeinsamen Merkmale ab.

Übersicht 1
Beschreibung einer Grundgesamtheit mittels verschiedener Erhebungen



Selbst bei der Integration aller zur Verfügung stehender Erhebungen einer Gesamtpopulation wird es nicht gelingen, alle für einen beliebigen Forschungszweck relevanten Angaben zu erfassen. Sind die Erfassungslücken, die in der Regel an den Rändern einer Grundgesamtheit entstehen, nicht allzu groß, so können diese Lücken mittels mathematischer Verfahren, hier insbesondere Imputationsverfahren, geschlossen werden.

3 Methoden der Datenverknüpfung

Die folgenden Methoden beschäftigen sich mit der Verknüpfung von möglichst identischen Merkmalsträgern und damit mit der Erweiterung des Merkmalskanons für eine gegebene Grundgesamtheit. Die Er-

weiterung eines Datensatzes in Form von Merkmalsträgern gestaltet sich in der Praxis als nicht aufwendig, wenn die Merkmalsträger eindeutig identifizierbar und damit Doppelerfassungen ausgeschlossen sind. In manchen Fällen, wie dem in Kapitel 2 aufgeführten Beispiel der Gegenüberstellung von Sozialhilfestatistik und Einkommensteuerstatistik, können Doppelerfassungen wegen sich nicht überlappender Berichtskreise a priori ausgeschlossen werden.

Größere Probleme entstehen bei der Erweiterung von Datensätzen um Merkmale. Hierzu sollen zwei Formen der Datenverknüpfung zur Anwendung kommen. Diese sind zum einen die exakte und zum anderen die multivariate Verknüpfung.

Exakte Verknüpfung

Bei der exakten Verknüpfung werden zwei Datenbestände, die für die gleiche Menge an Merkmalsträgern mit gleichen und unterschiedlichen Merkmalen vorliegen, mittels einer oder mehrerer Identifikationsvariablen zu einem neuen Datenbestand miteinander verbunden. Das heißt, gleiche Merkmalsträger werden in beiden Datenbeständen mittels der Identifikationsvariablen eindeutig identifiziert. Diese Merkmalsträger werden in einen neuen Datensatz mit ihren jeweiligen Merkmalen der einzelnen Datenbestände übernommen.

Als Identifikationsvariablen stehen bei der exakten Verknüpfung grundsätzlich eindeutige Kennziffern (Personenkennziffern, Unternehmenskennziffern) bzw. ein Bündel mit in der Summe eindeutigen Angaben (Name, Anschrift, Geburtsdatum) zur Verfügung. Das deutsche Datenschutzrecht lässt die exakte Verknüpfung von Datenbeständen nur in sehr wenigen Ausnahmen, wie z. B. beim Unternehmensregister zu. Da in Deutschland keine direkten Identifikationsvariablen wie etwa Unternehmenskennziffern zur Verfügung stehen, ist die oftmals sehr aufwendige und fehleranfällige Identifikation über ein Bündel von Merkmalen nötig. Sogar wenn dieses Bündel Merkmale mit Angaben zu Name oder Anschrift der Befragten enthält, ist nach bisherigen Erfahrungen mit sehr viel Aufwand bei der Verknüpfung zu rechnen. Hier können Verfahren wie z. B. das so genannte *pattern matching* eingesetzt werden. Dabei wird versucht, ein Ähnlichkeitsmaß bzw. Distanzmaß zwischen zwei Zeichenketten zu definieren. Erfolg versprechend scheinen die Ansätze der n-Gramme (es werden die in beiden Zeichenketten übereinstimmenden Zeichenfolgen der Länge n gezählt) und die so genannte Levenstein-Metrik (es wird die Anzahl der nötigen Vertauschungen und Ersetzungen von Buchstaben bei der Überführung der einen Zeichenkette in die andere als Abstandsmaß zugrunde gelegt). Beide Ansätze sind sehr verwandt mit dem im folgenden Abschnitt beschriebenen Vorgehen des *Statistical Matchings*.

Multivariate Verknüpfung (Statistical Matching)

Innerhalb des Verfahrens der multivariaten Verknüpfung werden Merkmalsträger unterschiedlicher Datenbestände aufgrund fehlender direkter Identifikatoren über ihre ‚Ähnlichkeit‘ zusammengespielt. Ähnlichkeit definiert sich in diesem Zusammenhang als Ähnlichkeit innerhalb eines Bündels von Merkmalen (im Folgenden Überschneidungsmerkmale genannt). Da man bei dieser Art von Verknüpfung vereinzelt auch fehlerhafte Zuordnungen in Kauf nehmen muss, entsteht ein synthetischer Datensatz.

Die Merkmalsträger des so entstandenen Datensatzes enthalten neben den Überschneidungsmerkmalen vor allem die Merkmale, die vor der Verknüpfung nur in jeweils einem der beiden Datenbestände vorhanden waren.

Ein Grundproblem dieses Verfahrens ist die Beurteilung darüber, welche Datensätze als ‚ähnlich‘ einzustufen sind. Ein häufig verwendetes Verfahren zur Quantifizierung der Ähnlichkeit besteht in dem Einsatz von Distanzfunktionen, die z. B. auf der Grundlage von Euklidischen Distanzen gebildet werden. Bei dieser Verfahrensweise wird die Distanz einer Merkmalsausprägung zu jeder anderen Merkmalsausprägung des gleichen Merkmals für alle Datensätze bestimmt. Innerhalb eines simultanen Optimierungsverfahrens werden in einem nächsten Schritt diejenigen Merkmalsträger zusammengeführt, deren Gesamtdis-

tanz über alle getroffenen Zuordnungen minimal ist. Ein vergleichbares Verfahren wurde bereits in einem anderen Kontext in der amtlichen Statistik erfolgreich eingesetzt (siehe Lenz).

Bei der multivariaten Verknüpfung von Stichprobenerhebungen ist darüber hinaus zu entscheiden, wie mit den Hochrechnungsfaktoren umzugehen ist. Hier bieten die Verfahren des ‚constrained‘ bzw. des ‚unconstrained matching‘ unterschiedliche Herangehensweisen. Beim ‚unconstrained matching‘ werden die Datensätze eins zu eins zusammengeführt und nach der Verknüpfung wird für den entstandenen synthetischen Datenbestand ein neuer an Zusatzinformationen gebundener Hochrechnungsrahmen entwickelt. Das ‚constrained matching‘ übernimmt den Hochrechnungsrahmen aus beiden Erhebungen, dies führt aber zu einem erheblichen mehr an Rechenaufwand. Die Merkmalsträger werden hierbei nicht einzeln, sondern jeweils gemäß ihrem Repräsentationsgrad zusammengeführt. Ein Merkmalsträger aus dem Datenbestand A mit z. B. dem Hochrechnungsfaktor zehn wird mit dem ähnlichsten Merkmalsträger aus dem Datensatz B verbunden, verbleibt im Datenbestand A nunmehr aber mit dem Hochrechnungsfaktor 9 und steht für weitere Zusammenführungen zur Verfügung. Dieses Verfahren erhält exakt die Randverteilungen und die Korrelationen innerhalb der Datenbestände, bläht aber den Datensatz auf die Anzahl der Einheiten in der Grundgesamtheit auf.

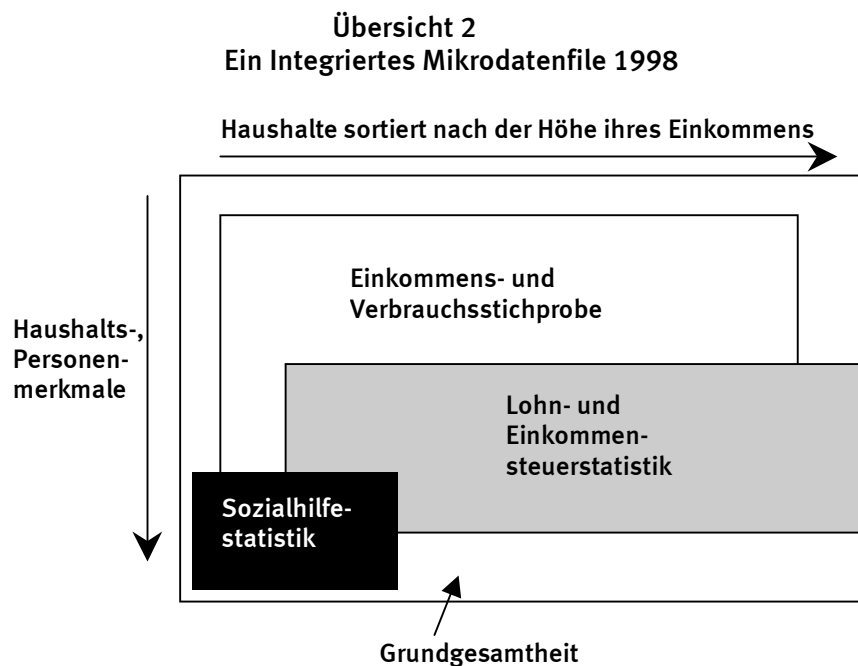
4 Integrierte Mikrodatenfiles (IMDAF) in Deutschland

Unter dem Namen IMDAF69 entstand während des Projekts „Sozialpolitisches Entscheidungs- und Indikatorensystem für die Bundesrepublik Deutschland (SPES)“ in den siebziger Jahren ein erstes integriertes Mikrodatenfile auf der Grundlage von amtlichen Einzeldaten. Hierzu standen u. a. die Einzeldaten der Einkommens- und Verbrauchsstichprobe für das Jahr 1969, des Mikrozensus sowie die Mikrozensuszusatzenerhebung aus dem Jahre 1971 zur Verfügung. Des Weiteren wurden Angaben aus der Volkswirtschaftlichen Gesamtrechnung (VGR) und der Einkommensteuerstatistik genutzt. Mit diesem Datenbestand konnten vielfältige Fragestellungen beantwortet werden, die mit den jeweils einzelnen Datenbeständen nicht zu beantworten gewesen wären. Die Verfahren, die zum IMDAF69 führten, wurden später rückwirkend auch für die Entwicklung eines IMDAF61/62 genutzt.

Diese Idee, Einzeldatenbestände der amtlichen Statistik zu einem integrierten Mikrodatenfile zusammenzuführen, ist im Nachgang zum Ersten Armuts- und Reichtumsbericht der Bundesregierung von den Statistischen Ämtern des Bundes und der Länder wieder aufgegriffen und mit verschiedenen Experten erörtert worden¹. Innerhalb dieses Gesprächs wurde insbesondere die Möglichkeit zur Generierung eines „Integrierten Mikrodatenfiles“ mit den Daten der EVS 1998 und den Einzeldaten der Lohn- und Einkommensteuerstatistik 1998 diskutiert. Als wünschenswert wurde es angesehen, dass ein solcher Datenbestand unter dem Namen IMDAF98 zum kommenden Zweiten Armuts- und Reichtumsbericht der Bundesregierung in den Jahren 2003/2004 den Forschern zur Verfügung steht. Dieses Ziel konnte aufgrund einer mangelnden Finanzierung zum damaligen Zeitpunkt nicht umgesetzt werden. Die simultane Betrachtung von Einkommen, Vermögen und Altersvorsorge scheint aber im Hinblick auf zukünftige Armuts- und Reichtumsberichterstattungen weiterhin sinnvoll. Die Datenintegration könnte über Haushalte oder Steuerpflichtige erfolgen; welche Verfahrensweise sinnvoller ist, müssen die methodischen Arbeiten zeigen. Als mögliche Fusionsmerkmale stehen in der Einkommens- und Verbrauchsstichprobe und der Lohn- und Einkommensteuerstatistik das klassierte Einkommen, die Quellen des Einkommens, Alter, Familientyp sowie auch die Region zur Verfügung.

Übersicht 2 zeigt die Konstruktionsidee des IMDAF98. Die Einkommensverteilung ließe sich über die nahezu gesamte Spreizung, insbesondere wenn es gelänge, auch die Einzeldaten der Sozialhilfestatistik zu integrieren, mit diesen Daten deutlich besser analysieren.

1 am 16. November 2000 mit Vertretern des BMF, BMA, DIW, Infratest Burke, ZUMA, GMD, Sachverständigenrat zur Begutachtung der gesamtwirtschaftlichen Entwicklung und den Universitäten Frankfurt am Main und Lüneburg



Die jeweiligen Datensätze enthalten keine direkten Identifikatormerkmale, so dass nur eine Integration mittels multivariater Verknüpfung sinnvoll erscheint. Eine erste rechtliche Überprüfung ergab, dass die Erstellung eines synthetischen Mikrodatenfiles als IMDAF98 mit dem Bundesstatistikgesetz vereinbar ist. Parallel zu den methodischen Arbeiten müsste aber auch dieser Aspekt noch einmal intensiver betrachtet werden.

Die Vorstellung der zeitnahen Erstellung eines „Integrierten Mikrodatenfiles“ erscheint aus heutiger Sicht sehr optimistisch. Die Qualität der Zusammenführung hängt sehr stark von der Wahl der hierzu verwendeten Verfahren ab. Z. B. kann eine geringfügige Modifikation der in Kapitel 3 erwähnten Distanz- und Ähnlichkeitsmaße zu deutlich anderen Ergebnissen bei der Verknüpfung führen. Es gibt hierzu Positionen, die davon ausgehen, dass die Fehler bei der Datenintegration größer werden können als der zusätzliche Nutzen einer integrierten Datei ausgleichen könnte. Diese pessimistische Position wird seitens der amtlichen Statistik nicht geteilt, es zeigt sich aber auch durch solche Positionen, dass im Bereich der Integration von amtlichen und ggf. auch nichtamtlichen Mikrodatenbeständen noch vielfältige Grundlagenforschung zu betreiben ist.

5 Ausblick

In diesem Aufsatz wurde ein starker Fokus auf die Verknüpfung der EVS 1998 und den Einzeldaten der Lohn- und Einkommensteuerstatistik 1998 gerichtet. In einem nächsten Schritt ist geplant, die auf diese Weise integrierten Daten mit dem Mikrozensus zu verknüpfen. Dabei wird jedem Einkommensteuerfall im Mikrozensus eindeutig ein Merkmalsträger der integrierten Datei zugeordnet. Da zwar einerseits bekannt ist, dass die Merkmalsträger des Mikrozensus in der Lohn- und Einkommensteuerstatistik und damit auch mit großer Wahrscheinlichkeit in der integrierten Datei enthalten sind, aber andererseits keine direkten Identifikatormerkmale zwischen den beiden Datenbeständen existieren, ist geplant, auf die Methode des in Kapitel 3 beschriebenen Statistical Matchings zurückzugreifen. Weil eine solche Verknüpfung nur auf Haushaltsebene möglich ist, müssen zunächst die aus der Lohn- und Einkommensteuerstatistik gewonnenen Personenangaben zu Haushalten aggregiert werden. Im Weiteren muss durch den Einsatz von Methoden der multiplen Imputation sichergestellt werden, dass die zusammen gespielten Merkmale beider Erhebungen konsistent sind.

Ein hauptrangiges Ziel der Arbeiten wird darin bestehen, das integrierte Mikrodatenfile über die verschiedenen Datenzugangswege der Forschungsdatenzentren des Bundes und der Länder, darunter insbesondere die Bereitstellung der Daten als so genanntes Scientific-Use-File, der Wissenschaft zugänglich zu machen. In diesem Zusammenhang muss später ein Anonymisierungskonzept entwickelt werden, welches gleichermaßen auf die Erfüllung der Regeln der statistischen Geheimhaltung und den bestmöglichen Erhalt an Potenzial für wissenschaftliche Analysen abstellt.

Literatur:

Bork, C. (2000); Steuern, Transfers und private Haushalte, Finanzwissenschaftliche Schriften, Band 99, Verlag Peter Lang; Frankfurt am Main, Berlin, Bern

Hauser, R. (Hrsg.) (1994); Mikroanalytische Grundlagen der Gesellschaftspolitik, Band 1 und 2, Ergebnisse aus dem gleichnamigen Sonderforschungsbereich 3 der DFG, Akademie Verlag, Berlin

Kortmann, K. (1982); Verknüpfung und Ableitung personen- und haushaltsbezogener Mikrodaten, Campus Forschung, Band 272, Campus Verlag, Frankfurt/ New York

Lenz, R. (2005); Measuring the disclosure protection of micro aggregated business microdata – An analysis taking the example of German Structure of Costs Survey, erscheint in: Journal of Official Statistics, Schweden

Zwick, M. (2004); Integrierte Mikrodatenfiles, Statistik und Wissenschaft, Band 1, S. 287-294, 2004