Krause, Peter; Pischner, Rainer; Wagner, Gert G.

Optimale Verarbeitung von Längsschnittdaten: Das Beispiel des Sozio-oekonomischen Panels (SOEP)

DIW Discussion Papers, No. 75

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

DIW Diskussionspapiere
Discussion Papers

Diskussionspapier Nr. 75

Optimale Verarbeitung von Längsschnittdaten - Das Beispiel des Sozio-oekonomischen Panels (SOEP)

von
Peter Krause, Rainer Pischner und Gert Wagner*

Deutsches Institut für Wirtschaftsforschung, Berlin
German Institute for Economic Research, Berlin

Deutsches Institut für Wirtschaftsforschung

Diskussionspapier Nr. 75

Optimale Verarbeitung von Längsschnittdaten - Das Beispiel des Sozio-oekonomischen Panels (SOEP)

von
Peter Krause, Rainer Pischner und Gert Wagner*

*) Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

Berlin, im August 1993

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 1000 Berlin 33
Telefon: 49-30 - 82 991-0
Telefax: 49-30 - 82 991-200

Optimale Verarbeitung von Längsschnittdaten - Das Beispiel des Sozio-oekonomischen Panels (SOEP)

von Peter Krause, Rainer Pischner und Gert Wagner

Für die Analyse von Erhebungsdaten, die ständig in einem bestimmten Rhythmus anfallen oder gar als echte Längsschnittsdaten erhoben werden, also als wiederholte Befragung immer derselben Einheiten, stellt sich die Frage, ob diese Daten in eine spezielle Datenbank eingespeichert werden sollen. Dadurch ergeben sich bestimmte technische Vorteile, ist aber mit dem Nachteil verbunden, daß die Analyse der Daten zusätzliche Kenntnis über die Benutzung der Datenbank erfordert.

Am Beispiel eines sehr komplexen Datensatzes, den, des Sozio-oekonomischen Panels (SOEP), geben wir Hinweise für den praxisgerechten Aufbau und Umgang mit einer Datenbank im Bereich wirtschafts- und sozialwissenschaftlicher Grundlagenforschung. Weiterhin wird auf das wichtige Ziel der "Datenweitergabe" eingegangen, die Re-Analysen ermöglichen.

Die SOEP-Daten

Das Sozio-oekonomische Panel (SOEP) ist eine großangelegte Wiederholungs-holungsbefragung, die seit 1984 in der Bundesrepublik Deutschland läuft. Die Daten dienen der sozial- und wirtschaftswissenschaftlichen Grundlagenforschung[1] sowie der laufenden Sozialberichterstattung[2]. Im Juni 1990 wurde diese Befragung auf Ostdeutschland ausgeweitet[3].

Pro Erhebungswelle werden gegenwärtig etwa 13 000 Personen im Alter von über 16 Jahren in etwa 6 500 Privathaushalten einmal jährlich befragt. Zusätzliche Informationen liegen darüber hinaus für die nicht befragten Kinder bis 16 Jahre sowie die gegenüber dem Vorjahr aus der Befragung temporär oder ganz ausgeschiedenen Personen und Haushalte vor.

---

1 Vgl. z.B. U. Rendtel und G. Wagner (Hg.), Lebenslagen im Wandel - Zur Einkommensdynamik in Deutschland seit 1984, Frankfurt, New York 1991.

2 Vgl. jüngst Statistisches Bundesamt (Hg.), Datenreport 1992, Bonn 1992.

3 Vgl. Projektgruppe Panel, Das Sozio-oekonomische Panel (SOEP) im Jahre 1990/91, in: Vierteljahrshefte zur Wirtschaftsforschung, Heft 3-4, 1991, S. 146-155.

Die Befragungsinhalte erstrecken sich über ein breites sozio-ökonomisches Themenspektrum wie Demographie, Arbeitsmarkt, Bildung, Gesundheit, subjektive Indikatoren, Einkommen, Wohnen, Haushaltszusammensetzung etc. Die 30 haushaltsspezifischen Fragen und knapp 100 personenspezifischen Fragen werden für jedes Erhebungsjahr in etwa 1000 Variablen abgespeichert.

Die Daten stehen allen Universitäten und Forschungsinstituten in der Bundesrepublik Deutschland für wissenschaftliche Zwecke grundsätzlich kostenlos zur Verfügung, wenn die datenschutzrechtlichen Bedingungen erfüllt sind. Auch in das Ausland werden diese Daten unter Berücksichtigung zusätzlicher Datenschutzauflagen für Forschungszwecke weitergegeben. Die Daten werden gegenwärtig von über 100 Nutzergruppen im Bereich Ökonometrie, Wirtschafts- und Sozialwissenschaften sowie der aktuellen Politikberatung ausgewertet. Die technische Durchführung der Datenanalyse erfolgt überwiegend durch hochqualifizierte Wissenschaftler. Der Einsatz von Programmierern für laufende Auswertungen bildet eher die Ausnahme.

Komplexe Datenstrukturen im Bereich der Sozial- und Wirtschaftswissenschaften

Der beschriebene Datenbestand des SOEP ist in mehrfacher Hinsicht komplex. So sind Daten inhaltlich und hinsichtlich ihres Zeitbezuges sehr heterogen. Das Gros der Daten wird zu den breitgefächerten Themenschwerpunkten als Stichtagsinformation zum Erhebungszeitpunkt erfaßt. Darüber hinaus gibt es auch retrospektiv zeitdiskret erfaßte Informationen zum Erwerbs- und Einkommensstatus auf Monatsbasis ("Kalendarien"), Einkommensdaten, die für den Vormonat oder das Vorjahr erhoben werden, ereignisorientiert erfaßte Veränderungen der Beziehungen im Haushalt des vergangenen Jahres sowie Informationen zur Biographie, die sich auf das gesamte bisherige Lebensalter beziehen.

Die Daten werden in vielfältigen Forschungszusammenhängen ausgewertet. Entsprechend heterogen ist bei den Datennutzern die Erfahrung zum Verarbeiten von Mikrodaten. Die Daten müssen demzufolge in einer möglichst übersichtlichen und einfachen Struktur organisiert sein, die zudem aber auch komplexe Anwendungen unterstützen muß.

Die Hardwareausstattung sowie die in Einsatz befindlichen Softwareprodukte sind bei den verschiedenen Nutzergruppen sehr unterschiedlich. Hinzu kommt in diesem Bereich ein in den letzten Jahren forciert zu beobachtende Abkehr von reinen Mainframelösungen hin zu PC-, Workstation-, Netzwerk- oder mit der Mainframe kombinierte Konfigurationen. Diese Entwicklung wird in der wissenschaftlichen Anwendung durch neue spezialisierte Statistiksoftware (LIMDEP) und Programmiersprachen (GAUSS) , die nur noch für PC's entwickelt wird, noch verstärkt. Diese Uneinheitlichkeit der Hard- und Softwareausstattung erfordert zur Vereinfachung von Reanalysen logischen Datenstrukturen, die möglichst reibungslos zwischen unterschiedlichen Datenbanksystemen übertragen werden können und die auch problemlos in datenbankunabhängige standardisierte Softwareprodukte implementiert sind.

Aufgrund ihrer permanenten Aktualität sind die Daten sowohl für Querschnittauswertungen vor allem im Bereich der laufenden Sozialberichterstattung als auch als Verlaufsinformationen im Längsschnitt von wissenschaftlichem Interesse. Diese divergierenden Analysekonzepte werden jedoch nicht optimal durch eine einheitliche Datenorganisation unterstützt. Mit zunehmender Zahl von Erhebungswellen werden überwiegend querschnittbezogene Datenstrukturen, die insbesondere für neue Nutzer einfacher nachvollziehbar sind, für Längsschnittauswertungen ineffizient. Eine gewisse zusätzliche Redundanz der Datenhaltung für bestimmte inhaltliche Schwerpunkte ist unter diesen Gesichtspunkten langfristig unumgänglich.

Notwendigkeiten für eine Datenbank

Um große Datenmengen möglichst sparsam abzuspeichern und - jeweils dokumentierte - Manipulationen an den Daten vorzunehmen, sind Datenbanken besonders geeignet. Derjenige, der ständig große Datensätze erhebt und verwaltet, sollte die Vorteile eines modernen Datenbanksystems (DBMS) auf jeden Fall nutzen[4]. Die DBMS-Software bietet nicht zuletzt gute datenschutzrechtliche

4 Wer freilich nur Querschnittsdaten erhebt und verwaltet, die nicht in einem schnellen Rhythmus wiederholt anfallen, für den bietet ein Datenbanksystem gegenüber einem Statistik-Programmpaket wie SPSS oder SAS kaum praktischen Vorteile. Bei Querschnittsdaten ist nach einer ersten Phase der Datenbereinigung der Veränderungsprozeß an dem Datensatz rasch abgeschlossen (bzw. die Veränderungen werden von den jeweiligen Nutzern selbst vorgenommen, d.h. nicht in den Originaldatensatz zurückgespielt). Die in der Sozialforschung übli-

Möglichkeiten der Zugriffskontrolle und -dokumentation[5]. Nur Datenbanken bieten zuverlässige Protokolle der Updates von Längsschnittdaten.

Die Speicherung in einem DBMS ist übersichtlich und zugleich sind effiziente Zugriffe möglich, da alle Daten physisch zusammengelegt sind. Die Gliederung des Datenbestandes erfolgt ausschließlich virtuell über vorher logisch definierte Fenster (Views, Tables, Rectypes). Die Statistik-Programmpakete SPSS und SAS verarbeiten - ebenso wie Programmiersprachen wie GAUSS - demgegenüber physisch voneinander unabhängige Rechteckfiles. Jede Datenverknüpfung zwischen diesen Rechteckfiles setzt hier gleiche Sortierfolgen voraus, was insbesondere bei varianten Verknüpfungskriterien (z.B. Haushaltsnummer) erhebliche physische Umsortierungsvorgänge erfordern kann.
Eine Zusammenführung über einzelne oder eine Kombination eineindeutiger Schlüsselvariablen verlangt natürlich, daß bei der Erhebung mit Sorgfalt darauf geachtet wird, daß die Schlüsselvariablen fehlerfrei vergeben werden. Dies ist nach unserer Erfahrung der Erhebungspraxis keine Selbstverständlichkeit. Für das SOEP wurde deswegen in Zusammenarbeit mit dem Erhebungsinstitutn Infratest ein Personen- und Haushaltsnummernsystem entwickelt, das diese Eineindeutigkeit garantiert, aber auch im Alltag des Erhebungsgeschäftes funktioniert[6].

Verarbeitung von wissenschaftlichen Längsschnittdaten

Ein wichtiges Prinzip der "Informatik großer Datensätze" ist allerdings nicht auf die Realisierung mit Hilfe von Datenbank-Software angewiesen: Die "dynamische Effizienz"[7]. Gemeint ist damit, daß nicht der gesamte Datensatz neu strukturiert werden muß, wenn eine weitere Erhebungswelle hinzukommt. Dies wäre dann der Fall, wenn für jede Erhebungseinheit die Informationen aller Erhebungswellen in einem Record abgelegt werden. Dieser Record müßte mit je-

---

che Stichprobengröße bereitet heutzutage auf Mainframes keinerlei Probleme und bei inzwischen üblichen Festplattengrößen von 100 MB und mehr auf PCs auch nicht mehr, so daß die "Datenkompression", die Datenbanken bieten können, für die meisten Nutzer nicht entscheidend ins Gewicht fallen.

5 Vgl. auch eine praxisorientierte Darstellung bei B. Engel, Effiziente relationale Speicherung komplexer Datenstrukturen, in: B. Engel et al., Datenbankorganisatorische Probleme und Grundlagen des NIFA-Panels. Ergebnisse eines Workshops. Arbeitspapier des Sonderforschungsbereichs 187, Ruhr-Universität Bochum 1991.

6 Vgl. dazu Infratest Sozialforschung, Das Sozio-oekonomische Panel - Paneldatei, Dokumentation, München 1988

7 Vgl. M. David, The Science of Data Sharing, in: J. Sieber (Hg.), Sharing Social Science Data, Newbury Park, CA 1991.

der neuen Welle verlängert werden. Eine derartige recordorientierte Speicherung kommt zwar der späteren Analyse entgegen, die - zumindest mit den üblichen Statistik-Paketen - auf einer recordweisen Verarbeitung basiert. Allerdings werden die Records auf diese Art und Weise sehr lang. Die amerikanische PSID-Studie ("Panel Study of Income Dynamics"), die inzwischen 25 Befragungswellen enthält, stößt mit einer Recordlänge von 35000 Spalten inzwischen an die logisch zulässige Höchstmenge auf Mainframe-Maschinen.

Nicht zuletzt, um die Daten auch unschwer wellenweise auswerten zu können (ohne daß der ganze Datensatz benutzt werden muß), bietet es sich an, die Daten für jede Erhebungswelle in einem getrennten File abzuspeichern. Die dynamische Effizienz ist damit offensichtlich gewährleistet, da mit jeder Welle nur ein weiteres File verarbeitet werden muß, jedoch die alten Files nicht verändert werden müssen (sieht man von der Generierung neuartiger Variablen ab).

Das Zusammenführen von Informationen, die in mehreren Files enthalten sind, für jeweils eine Erhebungseinheit, ist inzwischen auch mit den üblichen Statistik-Paketen kein Problem. Bei SPSS, um ein Beispiel zu nennen, müssen Match-Prozeduren benutzt werden. Einzige Voraussetzung für das Funktionieren dieser Prozeduren ist, daß alle Datensätze nach dem gleichen Merkmal durchsortiert sind. Also beispielsweise aufsteigend nach einer Personennummer. Ein Umsortieren, beispielsweise nach der Haushaltsnummer ist mit den Statistik-Programm-Paketen selbst jederzeit - und meist auch rechenzeiteffizient - möglich[8]. Lediglich bei spezifischen Anwendungen wie z.B. der Disaggregation von Haushaltsinformationen auf Personendaten bei von Welle zu Welle potentiell wechselnden Haushaltszusammensetzungen sind Datenbankabfragen zeiteffizienter und übersichtlicher.

Für eine benutzerfreundliche Verarbeitung von sozialwissenschaftlichen Längsschnittdaten sind nicht alle technischen Möglichkeiten, die ein DBMS bietet, auch effizient im Hinblick auf einen sparsamen Zeiteinsatz des Nutzers von Daten.

---

8 Für - zum Teil sehr komplizierte Beispiele -, die mehrere Sortier- und Matchstufen enthalten - vgl. J. Witte, Data Management and Analysis of the German Socio-economic Panel Using SPSS, Dokumentation des DIW, Berlin 1992. Alle DIW-Dokumentationen können gegen eine Schutzgebühr vom DIW, SOEP-Sekretariat, Königin-Luise-Str. 5, 14195 Berlin 33, bezogen werden.

In kommerziellen Datenbanken werden die Daten im allgemeinen stark gegliedert abgelegt. Dies fördert die Übersichtlichkeit des Datenbestandes und spart Platz. Bei wissenschaftlichen Datenabfragen werden aber selbst bei gleicher inhaltlicher Fragestellung natürlicherweise immer eine Vielzahl unterschiedlicher Zusatzinformationen zur Beschreibung oder Erlärung herangezogen. Eine zu stark gegliederte Datenstruktur hätte hier Effizienzverluste bei der Abfrage innerhalb von Datenbanksystemen und erst recht außerhalb bei der separaten Verwaltung vieler vieler Einzeldateien zur Folge. In diesem Zusammenhang führt eine in gewissem Umfang redundante Datenspeicherung wie z.B. das "Mitschleppen" von Erwerbsvariablen, die alle den Code "trifft nicht zu" enthalten, für Nichterwerbstätige letzendlich zu Effizienzgewinnen bei der Verarbeitung.

Mit dem letzten Beispiel wird ein weiteres Postulat der Datenbankimplementation - die Minimierung von Redundanz (im Idealfall werden nur noch die Verknüpfungsindikatoren mehrfach geführt) - angesprochen. Dieses erscheint ebenfalls in erster Linie für den kommerziellen Markt von Bedeutung. Die Widersprüchlichkeit oder Inkonsistenz von Informationen ist meist nur durch mehr oder weniger restriktive Annahmen eliminierbar, die ihrerseits wiederum Gegenstand wissenschaftlicher Fragestellungen sein können. Dies gilt selbst für so "eindeutige" Variablen wie Geburtsjahr: So lassen sich gewisse Schwankungen in den Altersangaben als Funktion des Lebensalters selbst oder - wie im Falle bestimmter Ausländerpopulation - als Datum der amtlichen Registrierung interpretieren. Auftretende Fehler können in wissenschaftlichen Datensätzen in erster Linie durch Optimierung von Feldarbeit und Datenediting (Nachfragen; Doppelerfassugn) und erst in zweiter Linie nachträglich im Datenbanksystem aufgrund der Vielzahl der verfügbaren Zusatzinformationen eliminiert werden.

Im Prinzip ist bei Längsschnittdaten ohne (redundante) Zusatzinformationen erkennbar, wie lange eine Befragungseinheit im Datensatz enthalten ist (vorher und nachher existieren keine validen Informationen bzw. sind sie auf spezielle Missing-Codes gesetzt). Für eine Analyse der Daten ist es allerdings sehr umständlich, wenn bei jeder neuen Analyse immer wieder geprüft werden muß, ob eine Befragungseinheit in einer bestimmten Welle vorhanden ist oder nicht. Die Praxis der Auswertung des SOEP hat gezeigt, daß es sehr nützlich, ja unumgänglich ist, "Metadaten" zu generieren, die angeben, für welche Zeitpunkte für jede

Befragungseinheit Informationen vorliegen. Diese Metadaten, die pro Befragungseinheit aus einem Record bestehen, in dem für jede Erhebungswelle die Information über die Verfügbarkeit von Daten in einer Variablen enthalten ist, eignen sich für rasche Filterprozesse, wie sie typischerweise bei Analysen vorgenommen werden[9].

Bei Längsschnittbefragungen wird bei der Fragebogengestaltung typischerweise viel Mühe darauf verwandt, den Befragten so wenig wie möglich mit überflüssigen Fragen zu belasten, da er schließlich bereit sein soll, auch in Zukunft an der Befragung teilzunehmen. Nicht zuletzt spart diese Vorgehensweise Befragungszeit und damit Kosten. Diese Vorgehensweise bereitet aber bei der Auswertung der Daten Schwierigkeiten. Bei Personenbefragungen wird z.B. nur bei der jeweils ersten Befragung einer Person der Schulabschluß erhoben (im Falle des SOEP sind dies mehrere Fragen für die allgemeinbildenden und berufsbezogenen Abschlüsse). In den Folgewellen wird lediglich mit einer Filterfrage ermittelt, ob sich eine Veränderung ergeben hat und wenn nicht - was bei den meisten Erwachsenen der Fall ist - werden die Fragen nach dem Schulabschluß übersprungen. Wenn eine Wiederholungsbefragung mehrere Wellen lang gelaufen ist, ist es recht mühsam, für die jeweils aktuellste Befragungswelle den Bildungsabschluß eines jeden Befragten zu ermitteln, denn dazu muß bei den meisten Befragten der gesamte Datensatz "rückwärts" bis ggf. zur ersten Welle durchsucht werden, um den Bildungsabschluß ermitteln zu können. Es ist deswegen sinnvoll, eine "Statusvariable" zu generieren, die den jeweils gültigen Bildungsabschluß pro Befragungswelle im Datensatz angibt. Ähnliches gilt für andere Variablen, wie z.B. Merkmale des Arbeitgebers, berufliche Tätigkeit oder die Wohnungsausstattung.

Sowohl durch Metadaten als auch durch Statusvariablen wird das Gebot der Datenintegrität verletzt, denn dieselbe Information wird zwei- oder mehrfach abgespeichert, wodurch es - in der Praxis - zu widersprüchlichen Informationen im Datensatz kommen kann. Allerdings macht diese Redundanz das Leben bei der Auswertung sehr viel leichter.

Die Befürworter von Datenbanksystemen führen an, daß eine redundante Speicherung nicht notwendig sei, da man die Programme (Retrievals), mit denen Da-

9 Für eine nähere Beschreibung vgl. J. Frick, Die SIR-Datenbank des Sozio-ökonomischen Panels - ein Tutorial zu Aufbau, Syntax und problemorientierten Anwendungen (Version 90.1), Dokumentation des DIW, Berlin 1990.

ten generiert und anschließend im Datensatz abgelegt werden, auch in der Datenbank ablegen und bei jeder Analyse, die eine bestimmte Variable benötigt, unschwer wieder laufen lassen könne. Dieses Argument ist im Grundsatz richtig, geht jedoch an der Praxis bislang vorbei. "Virtuelle Variablen", die nicht tatsächlich im Datensatz abgelegt sind, können nur von denjenigen Nutzern der Daten angesprochen werden, die über das Datenbanksystem verfügen, mit dem Daten ursprünglich generiert wurden. Große Datensätze wie das SOEP werden jedoch typischerweise mit Hilfe einer Vielzahl von Software bearbeitet. Hinzu kommt, daß nur derjenige das Retrieval nutzen kann, der über alle Teilmengen (Wellen) des Datensatzes verfügt. Dies ist bei vielen Nutzern von Längsschnittdaten nicht der Fall, die nur an ausgewählten Wellen Interesse haben.

Weitergabe von Daten

Nur durch vielfache Re-Analysen von komplexen Daten, kann deren voller Informationsgehalt erschlossen werden. Nicht zuletzt sind Re-Analysen notwendig, um Fehler aufzudecken.

Soll ein Datensatz von vielen Nutzern analysiert werden, müssen die Hürden, die vor der Auswertung stehen, möglichst niedrig sein. Beim Ziel einer "nutzerfreundlichen Datenweitergabe" ergeben sich eine Fülle von Detailfragen, die im folgenden Beispiel des SOEP dargestellt werden sollen. Die Lösung dieser Aufgabe ist umso schwieriger, je größer der Datensatz ist. So umfaßt der Datensatz des SOEP nach der 10. Welle über eine viertel Milliarde Zeichen, die in über 100 großen Files (Tabellen bzw. Records) gespeichert sind. Unter diesen Vorausetzungen sind im Rahmen des Weitergabeprozesses die zumutbaren Hardware-Kapazitäten der Datenproduzenten - hier das DIW - und natürlich die der Empfänger zu berücksichtigen. Das gleiche gilt für die zu verwendende Software, also für die Auswahl der Datenbanksysteme bzw. der Analyse-Programme. Weiterhin ist zu berücksichtigen, daß Manpower meist sehr eng limitiert ist. Schließlich - dies gilt in hohem Maße für eine Wiederholungsbefragung des SOEP - muß die Datenweitergabe im Einklang mit den datenschutzrechtlichen Bestimmungen erfolgen.

Um die Erfüllung der datenschutzrechtlichen Schutzbestimmungen voll zu gewährleisten, müssen künftige Nutzer des SOEP - ob aus dem Inland oder Aus-

land - einen schriftlichen Antrag auf Datenweitergabe stellen und ihren Analysezweck in groben Umrissen benennen. Diese Nennung ist notwendig, da Daten nur zielbezogen gespeichert und ausgewertet werden dürfen. Liegt der schriftliche Antrag vor, erfolgt die Prüfung der Datenschutzmaßnahmen bei den externen Nutzern durch die Datenschutzbeauftragte des DIW. Hierbei unterstützt das DIW die potentiellen Datennutzer im Hinblick der für den Datenschutz notwendigen Maßnahmen, so daß bisher aus Datenschutzgründen noch kein Antrag abgelehnt zu werden brauchte.

Bei Längschnittsdatensätze bietet es sich an, die neuen Daten nach jeder neuen Welle weiterzugeben. Dies kann aber erst geschehen, wenn die Daten aufbereitet worden sind. Dies erfordert einige Zeit, zumal der Datensatz des SOEP sehr kompliziert ist. Es empfiehlt sich ein Pflichtenheft anzulegen, in dem zu vermerken ist, welche Arbeiten in welcher Reihenfolge zu erledigen sind. Für das SOEP sind es in grober Gliederung:

- Erweiterung des Datenbankschemas (d.h. der Formate);
- Aufbereitung und Fehlerbereinigung der gelieferten Daten und Aufnahme in die Datenbank
- Verknüpfungschecks und Plausibilitätsprüfungen im Quer- und im Längssschnitt,
- Aktualisierung bzw. Erweiterung wellenübergreifender Merkmale in der Datenbank.
- Generierung von Hochrechnungsfaktoren und von sog. Statusvariablen und
- Dokumentation des erweiterten Datenbestandes.

Erst nach Erledigung dieser Arbeiten können die Daten weitergeben werden. Im allgemeinen geschieht dies beim SOEP 12 bis 18 Monate, nachdem die Feldarbeit der Erhebung abgeschlossen ist.

Bei aller Sorgfalt unterlaufen bei der Aufbereitung Fehler. Dennoch ist es nicht zweckmäßig, berichtigte Daten sofort neu zu verteilen. Effizienter ist es, die Nutzer von den Fehlern zu informieren, kurzfristig Lösungsvorschläge zu unterbreiten und nur in schwerwiegenden Fällen einzelne Records auf Anfrage neu zu verteilen, wenn sie für die jeweiligen Analysen unbedingt erforderlich sind. E-mail bietet ideale Möglichkeiten der Distribution von Nutzerinformationen.

Aus Datenschutzgründen dürfen Mikrodaten selbst nicht über e-mail versandt werden.

Was wird weitergeben?

Am technisch einfachsten zu realisieren wäre, würde immer der gesamte Datenbestand weitergeben werden. Dies ist indes nicht nur ineffizient, sondern auch mit dem Datenschutzgesetz unverträglich.

Ineffizient wäre eine solche Vorgehensweise, weil "alte" Datennutzer den größten Teil der Daten schon besitzen und unnötigerweise diese noch einmal implementieren müßten. Sie bekommen natürlich nur die neuen und die geänderten Records.

Unverträglichkeit mit dem Datenschutzgesetz ergäbe sich bei der Weitergabe aller Variablen, da dadurch die Identifizierung einzelner Personen erleichtern würden; z.B. Klartextangabe zur Person oder tief disaggregierte Regionalinformationen, die für Erhebung und Weiterverfolgung im Datensatz einer Wiederholungsbefragung enthalten sein müssen.

Die Weitergabe der Daten in das Ausland führt zu weiteren Problemen und bringt zusätzlichen Arbeiten mit sich. Denn das DIW nimmt den Datenschutz sehr ernst, - schließlich kann sich eine Wiederholungsbefragung keinen Skandal leisten - und erstellte in Zusammenarbeit mit dem Berliner Datenschutzbeauftragten eine besonders stark anonymisierte Version des SOEP-Datensatzes, den sog. "Public-Use-File". Dieser Datensatz kann einerseits bedenkenlos an unabhängige Forscher in aller Welt weitergeben werden, andererseits werden seine Analysemöglichkeiten nur marginal eingeschränkt. Im Public-Use-File wurden dazu einige Variablen gelöscht und andere durch Recodierungen aggregiert. Haupteinschränkung ist die mit dem Datenschutzbeauftragten vereinbarte Reduzierung des Datensatzes auf 95% - sowie die Ausblendung von Regional- und Klartextangaben - Deanonymisierung extrem erschwert wird. Bei einer solchen Kürzung eines samples, ist natürlich darauf zu achten, daß er seine Längschnitteigenschaften nicht beinträchtigt werden. Dies wurde erreicht durch Herausnahme von 5% der Ursprungshaushalte (die in der ersten Welle befragt wurden) und systematisches Löschen alle daraus abgeleiteten Datensätze.

Wie sollen die Daten weitergeben werden?

Nach Abschluß der vertraglichen Regelungen müssen den Nutzern verschiedene Optionen angeboten werden, in welcher Form sie die Daten erhalten können. Die erste Auswahl betrift das Medium, auf dem die Daten gespeichert sind. Früher kamen nur Magnetbänder für Großrechner in Frage. Heute werden die Daten aufgrund der enorm gestiegenen Leistungsfähigkeit der "Personal Computer" (PC) hinsichtlich Rechengeschwindigkeit und Speicherkapazität, nicht zuletzt auch wegen des gebotenen Komforts, zunehmend auch auf PC's ausgewertet.

Um eine breite Nutzung eines Datensatzes zu ermöglichen, darf dieser Trend nicht unberücksichtigt bleiben. Die Entwicklung einer Weitergabeform für PC auf Disketten erwies sich somit als unumgänglich. Doch sind Disketten im Gegensatz zu Magnetbändern leicht zugänglich und auch leichter lesbar. Dies ist datenschutzrechtlich bedenklich und erfordert die Erfüllung zusätzlicher Auflagen. So können die Daten grundsätzlich nur komprimiert und durch ein Passswort zusätzlich verschlüsselt weitergegeben werden. Desweiteren werden sie auch nicht mit normaler Post, sondern als Wertbrief verschickt. Das Passwort - ohne dieses sind die Disketten völlig unbrauchbar - wird gemeinsam mit der Dokumentation in einem gesondertem Umschlag mit getrennter Post und mit mehreren Tagen Verzögerung den Datennutzern getrennt mitgeteilt.

Weitere mögliche Medien zur Speicherung der Daten wären Streamer, dies sind Bandgeräte für PC's, und CD-ROM's (Compact-Disks). Streamer gibt es bislang nur in zu vielen und nicht genormten Varianten und kommen deshalb nicht in Frage. Anders sieht es bei den CD-ROM's aus. In den USA werden die Panel-Daten zweier Erhebungen[10] bereits auf diesem Medium weitergeben. Probleme bereiten gegenwärtig nicht mehr das Einlesen der Daten von CD's durch Sekundärnutzer, schließlich sind die Laufwerke hierfür inzwischen sehr preiswert geworden, sondern das Beschreiben der Datenträger . Dieses ist noch sehr aufwendig und so teuer, daß das DIW die Daten zur Zeit nur auf Magnetband für Großrechner oder auf Disketten für PC vertreiben kann.

Eine weitere Möglichkeit zur breiten Nutzung der Daten eröffnet die Datenfernverarbeitung. Technische Probleme gibt es kaum noch, doch stehen dem bislang

---

10 NLS und PSID

grundsätzliche Bedenken der Datenschützer entgegen, die berechtigt sind, das eine Kontrolle des Zuganges zu Fernleistungen sehr schwierig ist.
Neben dem Medium können die SOEP-Nutzer wählen, für welches Software-System sie die Daten erhalten wollen. Dies wirft die grundsätzliche Frage auf, in wievielen verschiedenen Formen Daten weitergegeben werden sollten. Zwei Extrema sind denkbar: Einmal die Weitergabe der Daten allein als Exportfile in dem vom Vertreiber benutzten Datenbank- System, (im Falle des SOEP also SIR). Oder die Weitergabe in allen Formaten, die von den Anwendern der Daten benutzt werden. Beide Optionen sind nicht hilfreich, wenn eine breite Nutzung der Daten angestrebt wird. Denn die erste schränkt den Anwenderkreis ein, die zweite würde zu derart vielen Weitergabeformen führen, daß dieser Service die eigentliche Datenaufarbeitung behindern würde. Dies ist schlicht zu aufwendig.

Eine gute Lösung besteht in der Kombination beider Extrema: Selbstverständlich steht das SOEP als Exportfile des Systems zur Verfügung, mit dem es selbst erstellt wird. Damit ist dem Anwender durch die Wahl von SIR als Datenbanksystem ein Weg gezeigt, schnell an die Daten und zu Ergebnissen zu kommen. Auch der Know-how-Transfer vom DIW zum Nutzer ist dann maximal. Allerdings erfordert SIR eine Einarbeitung. Für viele Nutzer, die nicht auf Support (z.B. von ihren Universitätsrechenzentren) hoffen können, sind andere Softwaresysteme effizienter.
Die SOEP-Daten können mit beliebiger Software analysiert werden, wenn die Rohdatenfiles vertrieben werden, d.h. von Datensätzen im Zeichenformat, (ASCII-Format). Bei Verzicht auf Struktur der Datenbank, sowie auf die automatische Labelung der Variablen gibt es keine grundsätzlichen Schwierigkeiten die Daten in andere System zu implementieren. Die Arbeit, eine geeignete Datenstruktur zu entwerfen und aufzubauen, kann den Anwendern ohnehin nicht abgenommen werden. Zudem ist es möglich, die Variablen - und Wertebezeichnungen getrennt als ASCII-File mitzugeben.

Deshalb hat sich das DIW für die Weitergabe der SOEP-Daten zu folgender Kompromißlösung entschieden: Die Rohdatenweitergabe für die am häufigsten angewandten Auswertungssystem - SAS und SPSS - wird unterstützt. D.h. für jedes einzelne File (Record) der SIR -Datenbank wird je ein Rohdaten- und ein Dictionary-File erzeugt. Letzterer enthält die Formatangaben für die Variablen, sowie ihre Labels. Die Unterstützung für SAS und SPSS erfolgt durch Zu-

satzsoftware für PC, mit denen die Dictionary-Files in lauffähige SPSS- und SAS-Programme umgesetzt werden können. Die Zusatzprogramme analysieren die Dictionary Files, generieren die systemspezifischen Einlese-Formate für Daten und für die Variablenbeschreibungen. Lediglich die Filedefinition muß der jeweiligen Hardware angepaßt werden. Die erzeugten Programme können dann die Daten einschließlich ihrer Labels einlesen.

Daten im SAS oder SPSS-Format können wiederum von anderen Datenbanksystemen über Schnittstellen, die auch Labels weitergeben, eingelesen werden. Würden allerdings nur SPSS-oder SAS-Exportfiles weitergegeben, könnten Anwender, die z.B. GAUSS benutzen, die Daten nicht verarbeiten. Deswegen ist die Weitergabe von Rohdaten und Dictionary-Files eine optimale Form der Distribution.