

Wagner, Gert G.; Schupp, Jürgen; Rendtel, Ulrich

Working Paper — Digitized Version

Das Sozio-ökonomische Panel: Methoden der Datenproduktion und Aufarbeitung im Längsschnitt

DIW Discussion Papers, No. 31

Provided in Cooperation with:

German Institute for Economic Research (DIW Berlin)

Suggested Citation: Wagner, Gert G.; Schupp, Jürgen; Rendtel, Ulrich (1991) : Das Sozio-ökonomische Panel: Methoden der Datenproduktion und Aufarbeitung im Längsschnitt, DIW Discussion Papers, No. 31, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:

<https://hdl.handle.net/10419/95777>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.



Diskussionspapiere
Discussion Papers

Diskussionspapier Nr. 31

**Das Sozio-ökonomische Panel
Methoden der Datenproduktion und
Aufarbeitung im Längsschnitt**

von
G. Wagner, J. Schupp und U. Rendtel

Die in diesem Papier vertretenen Auffassungen liegen ausschließlich in der Verantwortung des Verfassers und nicht in der des Instituts.

Opinions expressed in this paper are those of the author and do not necessarily reflect views of the Institute.

Deutsches Institut für Wirtschaftsforschung

Diskussionspapier Nr. 31

**Das Sozio-ökonomische Panel
Methoden der Datenproduktion und
Aufarbeitung im Längsschnitt**

von
G. Wagner, J. Schupp und U. Rendtel

Berlin, August 1991

Deutsches Institut für Wirtschaftsforschung, Berlin
Königin-Luise-Str. 5, 1000 Berlin 33
Telefon: (030) 82 991-0
Telefax: (030) 82 991-200

DAS SOZIO-ÖKONOMISCHE PANEL - METHODEN DER DATENPRODUKTION UND -AUFBEREITUNG IM LÄNGSSCHNITT

von Gert Wagner, Jürgen Schupp und Ulrich Rendtel

Im vorliegenden Aufsatz wird der erste umfassende methodische Überblick über die Realisierung der Wiederholungsbefragung des Sozio-ökonomischen Panels für die Bundesrepublik Deutschland gegeben^{1,2}. Darüber hinaus wird die Entwicklung der Stichprobe ausführlich über einen längeren Zeitraum nachgezeichnet und erstmals wird das Konzept der Datenaufbereitung analytisch dargestellt. An geeigneten Stellen wird auch auf Fortschritte bei der Erhebung und Datenaufbereitung gegenüber der Vorbildstudie "Panel Study of Income Dynamics" hingewiesen, die in den USA seit 1968 läuft³.

1 Ziel

Das Sozio-ökonomische Panel (SOEP) ist eine seit 1984 laufende Stichprobenerhebung, die in den Rahmen der Bevölkerungs- und Erwerbsstatistik einzuordnen ist. Die Daten des SOEP ergänzen amtliche Statistiken, sowohl im Hinblick auf zusätzliche Variablen, die erhoben werden, als auch im Hinblick auf den Längsschnittcharakter der Daten, d.h. durch die wiederholte Befragung ein und derselben Stichprobenmitglieder. Diese Methode eignet sich naturgemäß hervorragend zur Beschreibung und Analyse von Veränderungen, wie sie durch die Entwicklung der DDR induziert wurden. Deswegen wurde das SOEP bereits im Juni 1990 auf das Gebiet der damaligen DDR ausgeweitet.

1.1 Lebenslagen und Mikroökonomischer Ansatz

Zur Untersuchung des strukturellen Wandels und sich daraus ergebender Veränderungen auf der Makro-Ebene von Wirtschaft und Gesellschaft ist in der Ökonomie und Soziologie seit den 70er Jahren die empirische Analyse von Mikroeinheiten in den Vordergrund getreten. Nachdem zunächst große Querschnittsdatensätze von Privathaushalten und Personen analy-

siert wurden (vgl. dazu z.B. Engel in diesem Buch), reifte rasch die Erkenntnis, daß - ebenso wie in der traditionellen Zeitreihenanalyse von Aggregatdaten - die zeitliche Verfolgung von Mikroeinheiten notwendig ist, um Hypothesen über Wandlungsprozesse in einer Gesellschaft empirisch testen zu können. Erst retrospektiv erhobene Lebensverlaufsdaten (vgl. Brückner in diesem Buch) und kontinuierlich durch Wiederholungsbefragungen erhobene Paneldaten, über die im folgenden berichtet wird, ermöglichen die Durchführung derartiger Kausalanalysen.

Vor dem Hintergrund des Paradigmas einer "rationalen Politikbetreuung", welches das "Sozialpolitische Entscheidungs- und Indikatorensystem" (SPES) bestimmte (vgl. Krupp 1975, Hauser 1991), sollen mit dem SOEP für zwei theoretische Ansätze Panel-Daten zur Verfügung gestellt werden: Ökonomische Variablen mikroökonomischer und mikroökonomischer Ansätze, die zum einen den Humankapitaltheorien im weitesten Sinne zuzuordnen sind und zum anderen theoretischen Ansätzen zur Arbeitsmarktsegmentation und Armutsforschung (vgl. Krupp und Glatzer 1978). Die soziologischen (und wenige politologische) Variablen werden vom Konzept der Sozialberichterstattung bestimmt (vgl. Habich und Zapf in diesem Band), das insbesondere vom Forschungsprogramm der "Lebensverlaufsanalyse" theoretisch untermauert wird (vgl. dazu Mayer 1990).

Neben dem Längsschnittaspekt und der theoriegeleiteten Variablenauswahl nimmt das Sozioökonomische Panel mit seinem haushaltsbezogenen Erhebungsdesign einen weiteren zentralen soziologischen wie mikroökonomischen Erkenntnisfortschritt auf. Mit der Befragung jeder erwachsenen Person im Haushalt stehen zentrale Informationen des Primär-Netzes 'Haushalt' zur Verfügung. Insbesondere neuere spieltheoretische Ansätze sowie die gesamte Literatur zur Netzwerkanalyse betonen die Notwendigkeit derartiger haushaltsbezogener Datenbasen (Ott 1991).

1.2 Fragenprogramm

Das Erhebungs- wie Analysedesign des SOEP ist in mehrfacher Weise komplex: Dies betrifft die Ausrichtung der Variablenauswahl in einen verzahnten ökonomischen und soziologischen Theorieansatz; die Erweiterung der Individualperspektive um den Haushaltskontext, die repräsentative Querschnitts- und Längsschnittanalyse sowie die überproportionale Einbeziehung einer eigenen Migrantenstichprobe (vgl. Abschnitt 2.3.2.1.2) in den Untersuchungsansatz.

Rein kohortenbezogene Panel-Befragungen (wie z.B. der US-amerikanische National Longitudinal Survey), gehen von weniger breiten theoretischen Grundlagen aus.

Bei der thematischen Auswahl des Fragenprogramms wurde an die umfangreichen empirischen Vorarbeiten innerhalb des Sonderforschungsbereiches angeknüpft (vgl. Becker 1989).

Das Fragenprogramm und die Analysemöglichkeiten erstrecken sich schließlich auf folgende acht inhaltlichen Themenfelder⁴: Demographie und Bevölkerung; Arbeitsmarkt und Beschäftigung; Einkommen, Steuern und soziale Sicherung; Wohnen; Gesundheit; Bildung und Qualifikation; Leistungen privater Haushalte; Grundorientierungen, Partizipation und Integration.

Neben diesem Standardteil der Längsschnittindikatoren gibt es wellenspezifische Zusatzfragen, die bestimmte Themenbereiche vertiefen: Erwerbsbiographie seit dem 15. Lebensjahr (1984); Ehe- und Familienbiographie (1985); Soziale Herkunft, Berufseinstieg, Wohnumfeld (1986); Soziale Sicherung, Vorrubestand, Pflegebedürftigkeit und Kinderbetreuung (1987); Vermögensbilanz (1986); Weiterbildung und Qualifizierung (1989); Zeitverwendung und -präferenzen (1990) sowie Familie und Unterstützungsleistungen (1991).

Die Fragen des SOEPs sind standardisiert und "geschlossen", d.h. daß die Antwortmöglichkeiten genau eingegrenzt werden. Es gibt allerdings eine Reihe von "offenen Fragen", deren

Antworten im Fragebogen im Klartext notiert werden. Offene Klartextangaben werden entweder anhand feststehender Codierungsregeln in Klassifikationsschemata umgesetzt oder von der Editing-Gruppe des Umfrageinstituts auf einer gesonderten Datei mit den notwendigen Identifikatoren gesondert übergeben (dies betrifft z.B. "Sonstige Sorgen" oder "sonstige Bildungsabschlüsse")⁵.

Insgesamt bietet das SOEP damit auch die Möglichkeit, ein gewisses Maß "qualitativer Sozialforschung" zu betreiben, da im Laufe der Zeit ggf. für alle Haushaltsmitglieder eine Vielzahl von Klarschriftangaben im Datensatz enthalten sind. Aus Datenschutzgründen werden diese Angaben freilich nicht an externe Datennutzer weitergegeben (vgl. auch Abschnitt 2.3.6.4 unten).

Standardmäßig werden ebenfalls die berufliche Tätigkeit und der Wirtschaftszweig in Klarschrift erfaßt und erst anschließend von ZUMA vercodet. Dieses Verfahren erwies sich auch als außerordentlich nützlich, um Klassifizierungsprobleme bei der Basisbefragung in der DDR zu umgehen (vgl. dazu Geiss und Hoffmeyer-Zlotnik 1991). Da für die alte DDR-Wirtschaft und die Übergangsprozesse keine international vergleichbaren und bewährten Klassifikationsschemata für Beruf und Branche vorlagen, ist die Erfassung der Klarschriftangaben der Befragten am sinnvollsten, da sie beliebige Re-Codierungen nach alternativen Klassifikationsschemata zulassen.

1.3 Repräsentativität im Längs- und Querschnitt

Die Repräsentativität einer Stichprobe bedeutet im allgemeinen nicht, daß es sich um einen in jeder denkbaren Hinsicht perfekten Mikrokosmos der Grundgesamtheit handelt. Vielmehr sind je nach Fragestellung unterschiedliche Repräsentativitätsbegriffe und -anforderungen sinnvoll zu formulieren, da in der Praxis keine perfekten Stichproben im Sinne eines "Mikrokosmos" gezogen werden können⁶.

Die Ziele von Stichprobenerhebungen können grob in zwei Teilbereiche unterschieden werden: Zum einen sollen Stichproben zur Deskription genutzt werden, d.h., daß die Grundgesamtheit in ihren Aggregaten und ihren Verteilungen möglichst gut⁷ dargestellt werden soll. Dies ist für Zwecke der Politikberatung wichtig. Zum anderen - und dies steht in den Wissenschaften im Vordergrund - können mit Hilfe von Stichprobendaten Theorien durch multivariate Verfahren getestet werden, deren Hypothesen kausal formuliert sind. Die Anforderungen an die Repräsentativität sind für die Überprüfung von Theorien - was auf den ersten Blick vielleicht überraschen mag - niedriger als für rein deskriptive Zwecke.

Ein erstes Problem der Repräsentativität ist die (definitorische) Abgrenzung der relevanten Grundgesamtheit (im Englischen wird von der "target population" gesprochen; vgl. auch Schnell 1991). Die meisten Stichproben in der Bundesrepublik Deutschland repräsentieren von ihrem Ansatz her nicht die gesamte Wohnbevölkerung, weil Bewohner von Heimen (Anstaltsbevölkerung) und Ausländer nicht erfaßt werden sollen. Beides wurde mit dem SOEP versucht, wobei es sich freilich erwies, daß auch im SOEP die Anstaltsbevölkerung aus praktischen Gründen der Stichprobenziehung sowie des Feldzugangs nicht hinreichend repräsentiert werden kann. Für die wichtigsten Ausländergruppen der südeuropäischen ehemaligen Anwerbeländer wurde eine eigene Stichprobe (genannt Stichprobe B) gezogen.

Die "Zielpopulation" des SOEP ist also umfassender als die üblicher Umfragen, allerdings gelingt es auch dem SOEP nicht, die gesamte Wohnbevölkerung zu erfassen. Auf spezielle Probleme durch Zuwanderungen wird noch einzugehen sein. Rein stichprobentechnisch war die Ausweitung der relevanten Grundgesamtheit auf das Gebiet der DDR kein Problem. Auf diesbezügliche Aktivitäten wird im anschließenden Abschnitt eingegangen werden.

Nach der Abgrenzung der "Zielpopulation" wird üblicherweise unter Repräsentativität bzw. Repräsentativitätsproblemen nur die Frage verstanden, ob die beabsichtigte Grundgesamtheit in ihren Strukturen in der Stichprobe richtig wiedergegeben wird bzw. bei disproportional angelegten Stichproben ein geeignetes Verfahren zur "Entzerrung" der Stichprobe existiert.

Bei Panel-Erhebungen ist schließlich als dritte Stufe der Repräsentativität darauf zu achten, ob aufgrund der Wiederholung der Befragung bei den Stichprobenmitgliedern Verhaltensänderungen, bedingt durch die Teilnahme an der Untersuchung selbst, eintreten ("Paneleffekt").

Aufgrund des Gegenstandsbereiches der SOEP-Befragung und der nur einmal jährlich stattfindenden Befragung ist mit ernsthaften Panel-Effekten nicht zu rechnen. Freilich können sich Probleme durch die Verweigerung der Teilnahme von "Zielhaushalten" ergeben. Darauf wird für das SOEP in den Abschnitten 2.3.4.1 und 2.3.5 konkreter eingegangen. Hier sei nur eine allgemeine Anmerkung gemacht.

Für den Test von Theorien ist es nicht notwendig, daß Stichproben völlig unverzerrt sind (vgl. dazu Rendtel 1991a). Vielmehr kommt es darauf an, daß für alle relevanten Beziehungen des Modells (und der Realität) genügend Fälle in der Stichprobe enthalten sind, sodaß statistisch gesicherte Aussagen möglich werden. Insbesondere bei Panels können nicht-zufällige Ausfallprozesse ggf. mit modelliert werden, so daß die Beziehungen unverzerrt geschätzt werden.

1.4 Gang der Darstellung

Im folgenden werden in Abschnitt 2.3.2 die Stichprobenziehung und die Feldarbeit der ersten Wellen des SOEPs beschrieben. Hierbei handelt es sich zwar nicht um panelspezifische Probleme, aber die Ziehung der 1. Welle ist für die spätere Repräsentativität des Längsschnittes gleichwohl extrem bedeutsam. Mit Abschnitt 2.3.3 beginnt die panelspezifische Darstellung. Zuerst wird das sogenannte Weiterverfolgungskonzept vorgestellt, welches für eine Längsschnitterhebung naturgemäß von zentraler Bedeutung ist. Außerdem wird die "panel-spezifische" Konstruktion der Erhebungsinstrumente diskutiert und dokumentiert. In Abschnitt 2.3.4 wird auf die bisherige Entwicklung der Stichproben des SOEPs und in Abschnitt 2.3.5 auf notwendige Gewichtungen und Hochrechnungen der Stichprobenergebnisse eingegangen. In Abschnitt 2.3.6 wird die Aufbereitung der Daten diskutiert und dargestellt, die weit komplizierter ist als bei einfachen Querschnitterhebungen.

In Abschnitt 2.3.7 werden schließlich die wichtigsten Ergebnisse zusammengefaßt und Perspektiven vorgestellt.

2 Stichprobenziehung und Feldarbeit der 1. Wellen

Im folgenden werden die genaue Definition der Stichproben, die Stichprobenziehung und die Feldarbeit der ersten Wellen des SOEPs dargestellt. In der alten Bundesrepublik Deutschland begann das Panel mit den Stichproben A und B (Deutsche und Ausländer) im Jahre 1984. Die Ausweitung auf das Gebiet der damaligen DDR erfolgte mit der Stichprobe C (Deutsche in der DDR) im Juni 1990.

Mit der Durchführung der Erhebung ist "Infratest Sozialforschung" (München) betraut. Neben der Kooperation bei der endgültigen Gestaltung der Erhebungsinstrumente, erfolgt durch Infratest die Feldarbeit, die Datenerfassung und -aufbereitung und die Dokumentation der Erhebung, bevor die anonymisierten Mikrodaten an die Panel-Projektgruppe übergeben werden.

2.1 Grundgesamtheit und Befragungspersonen

Die Grundgesamtheit ("Zielpopulation"), die durch das SOEP repräsentiert werden soll, ist zum einen definiert durch die deutsche Wohnbevölkerung der Bundesrepublik Deutschland einschließlich Berlin (West) im Jahre 1984 und die deutsche Wohnbevölkerung in der DDR im Juni 1990. In der Bundesrepublik wurden neben deutschen Personen in Privathaushalten auch ausgewählte ausländische Personengruppen in die Untersuchung einbezogen.

Befragungspersonen sind alle Haushaltsmitglieder, die mindestens 16 Jahre alt sind. Für die jüngeren Haushaltsmitglieder werden von der Hauptbefragungsperson einige Angaben gemacht. Durch diese "Proxy-Interviews" sind auch für Kinder Verlaufsanalysen möglich, z.B. zu ihrem Bildungsweg. Diese Informationen können mit den späteren Informationen verbunden werden, wenn die Kinder das Befragungsalter erreichen. Da alle erwachsenen Personen

eines Haushaltes interviewt werden sollen, scheiden Telefoninterviews als Erhebungsmethode aus, wenngleich dem Telefon als Kontaktmedium zwischen Befragungshaushalt und Interviewer bzw. Umfrageinstitut bei Interviewerwechsel eine wichtige Rolle zukommt.

Die ursprüngliche Stichprobenziehung für die BRD im Jahre 1984 erfolgte aus stichproben- und erhebungstechnischen Gründen getrennt für zwei Teilpopulationen:

Stichprobe A "Deutsche" umfaßt Personen in Privathaushalten, deren Haushaltsvorstand nicht die türkische, griechische, jugoslawische, spanische oder italienische Staatsangehörigkeit besitzt.

Stichprobe B "Ausländer" umfaßt Personen in Privathaushalten, deren Haushaltsvorstand die türkische, griechische, jugoslawische, spanische oder italienische Staatsangehörigkeit besitzt.

Bewohner von Wohnheimen (Schüler-, Lehrlings-, Studenten-, Berufstätigen- und Altenwohnheime) wurden im Rahmen des ADM-Samples⁸ in die Stichprobenziehung einbezogen. Anstaltsinsassen im engeren Sinne (Krankenhäuser/Sanatorien; Altenpflegeheime; Kasernen) wurden in der Ausgangsstichprobe des Panels nicht erfaßt, im weiteren Panelverlauf jedoch dann, wenn Personen aus den Ausgangsprivathaushalten vorübergehend oder auf Dauer in derartige Institutionen wechseln (vgl. dazu Abschnitt 3.1 unten) und in der Lage sind, weiterhin an der Befragung teilzunehmen.

Stichprobe C "Deutsche in der DDR" umfaßt Personen in Privathaushalten, deren Haushaltsvorstand (Hauptverdiener des Haushalts) die Staatsbürgerschaft der DDR besitzt. Damit wurden ca. 1,7 % der Wohnbevölkerung der DDR im Juni 1990 als Ausländer aus der Stichprobe ausgeschlossen. Da sie überwiegend Anstaltsbewohner sind, wären sie aufgrund des Ausschlusses von Anstaltshaushalten ohnehin bereits nicht in der Stichprobenabgrenzung enthalten.

Die Stichprobengrößen wurden pragmatisch festgelegt. Einerseits sind möglichst große Stichproben aus methodischen Gründen gewünscht, andererseits sind die zur Verfügung stehenden Finanzmittel notorisch knapp.

2.1.1 Stichprobe A "Deutsche in der BRD"

Stichprobe A sollte einen Umfang von 4500 Haushalten netto ausweisen; die realisierte Nettostichprobe umfaßte schließlich 4554 Haushalte.

Für die Stichprobenziehung der Teilstichprobe A diente das ADM-Ziehungsband von 1982 als Grundlage⁹. Aus ihm wurden 584 Sample-Points mittels eines mehrstufigen, geschichteten Stichprobenverfahrens zufällig ausgewählt. Innerhalb der ausgewählten Stimmbezirke erfolgte die Haushaltsauswahl durch die Interviewer nach dem Random-Route-Verfahren: Ausgehend von einer vorgegebenen Startadresse mußten die Interviewer fortlaufend 84 Adressen notieren. Jeder 7. Haushalt war "Zielhaushalt" und somit für die Befragung anzuwerben.

2.1.2 Stichprobe B "Ausländer in der BRD"

Stichprobe B besteht genaugenommen wiederum aus fünf eigenständigen Teilstichproben der fünf zahlenmäßig größten ausländischen Nationalitäten, die 1984 in der Bundesrepublik Deutschland wohnten. Um detailliertere Auswertungen durchführen zu können, wurde die angestrebte Nettofallzahl für die fünf Nationalitäten (die südeuropäischen Anwerbeländer für Arbeitsmigranten) höher angesetzt als es ihrem Bevölkerungsanteil entspricht (disproportionaler Ansatz). Die Festlegung der Fallzahl der Nationalität erfolgte - bei insgesamt 1400 zu befragenden "Ausländerhaushalten" - unter Berücksichtigung der erwarteten Mobilitäts- bzw. Rückkehrerquoten.

Für die Stichprobenziehung der Teilpopulation B wurde zunächst getrennt für jede Nationalität eine Zufallsauswahl von Kreisen und kreisfreien Städten vorgenommen. Anhand der

Ausländerregister der Kreise bzw. kreisfreien Städte wurden sodann die Befragungspersonen nach dem Zufallsverfahren ausgewählt. Der Haushalt der auf diese Weise ausgewählten Befragungsperson gehört zur Stichprobe, sofern der Haushaltsvorstand die gleiche Nationalität besitzt. In einer Reihe von Kreisen bzw. kreisfreien Städten - insbesondere in Baden-Württemberg - war eine Ziehung aus den Ausländerregistern nicht möglich; hier wurde - nach Zufallsauswahl von Gemeinden - ersatzweise das Einwohnermelderegister der Gemeinden bzw. Städte zugrundegelegt.

Die Zahl der Sample-Points wurde mit 80 für die (stark überrepräsentierten) Türken und je 40 für die übrigen Nationalitäten festgelegt. Pro Point wurden aus den Melderegistern 20 Adressen gezogen, von denen ein Teil als Reserveadressen diente.

Die Zahl der eingesetzten Adressen pro Sample-Point bei den Ausländer-Stichproben weist eine stärkere Streuung auf als in der Stichprobe A. Grund ist der erheblich höhere Anteil qualitätsneutraler Ausfälle, für die - meist wegen falscher bzw. nicht mehr aktueller Adresse - eine Reserveadresse einzusetzen war.

2.1.3 Stichprobe C "Deutsche in der DDR"

Obwohl im Frühjahr 1990 die deutsche Vereinigung bereits absehbar war, war es sinnvoll, die Stichprobengröße für die Ost-Stichprobe C des SOEPs so zu wählen, daß eigenständige Analysen für die DDR bzw. die Neuen Bundesländer möglich sind. Im Vergleich zur Stichprobe A wurde deswegen ein größerer Auswahlatz zur Grundgesamtheit gewählt und mindestens 2000 Haushalte als Stichprobengröße angestrebt; erreicht wurden 2179 Haushalte.

Aufgrund des Zugriffes auf Adressen aus der zentralen Einwohnermeldekartei der DDR war ein anderes - und besseres - Verfahren zur Stichprobenziehung möglich als für die Stichproben A und B.

Als Auswahlgrundlage diente das von Infratest für die DDR erstellte Mastersample (Stichtag der Ziehungsanweisung: 14. März 1990). Dieses ist im Unterschied zum ADM-Mastersample¹⁰ in der alten Bundesrepublik eine Zufallsauswahl von Personenadressen, die aus dem zentralen "Einwohnerdatenspeicher" gezogen wurden (vgl. zu diesem Pietzke 1991).

Zur Stichprobenbildung für die Stichprobe C wurde das Infratest-Master-Sample in folgenderweise genutzt:

- Zunächst wurde eine haushaltsproportionale Allokation für 360 Sample-Points berechnet, die der Schichtung des Master-Samples nach Kreisen und Gemeindegrößenklassen folgte.
- Je Schicht wurden anschließend in systematischer Zufallsauswahl die Sample-Points aus dem Mastersample ausgewählt, die dieser haushaltsproportionalen Allokation entsprachen.
- Schließlich wurde aus den für diese Sample-Points vorhandenen Adressen jeweils eine Person ab 16 Jahren als "Startadresse" zufällig ausgewählt. Um zu einer repräsentativen Haushaltsstichprobe zu gelangen (und um durch die Klumpung der Befragungsadressen Wegezeiten und -kosten zu sparen), wurde das Random-Route-Verfahren gewählt. Ausgehend von dieser Startadresse hatte jeder Interviewer auf einem formal beschriebenen und eindeutigen Auswahlweg (Random-Route) die Haushalte aufzulisten, wobei die Startadresse selbst nicht befragt werden durfte.
- Zehn Privathaushalte sollten aufgelistet und zur Panelteilnahme angeworben werden¹¹, soweit sich nicht bei der Kontaktaufnahme herausstellte, daß eine der aufgelisteten Adressen nicht zur Grundgesamtheit gehört (bzw. die Wohnung verlassen ist). In solchen Fällen konnten bis zu zwei zusätzliche Adressen als Ersatz aufgelistet und kontaktiert werden. Jeder dritte Haushalt war "Zielhaushalt" und somit für die Befragung anzuwerben¹².

Von Insidern der empirischen Sozialforschung der DDR wird gegen Adressen aus dem zentralen Melderegister der Einwand vorgebracht, daß in den entsprechenden Wohnungen häufig andere Personen lebten, als dort gemeldet waren. Diese Behauptung ist zwar schwer zu überprüfen, aber aufgrund der Wohnungsnot in der DDR durchaus plausibel und wahrscheinlich. Die Qualität der SOEP-Stichprobe wird aber hierdurch auf keinen Fall beeinträchtigt, da die Adressenstichprobe nur benutzt wurde, um eine zufällige und repräsentative regionale Verteilung der Befragungshaushalte zu gewährleisten. Es ist für die Random-Route unerheblich, ob und von wem der Start-Haushalt tatsächlich bewohnt wird.

2.2 Adressenbrutto und qualitätsneutrale Ausfälle der ersten Wellen

2.2.1 Stichproben A und B

Nach dem Stichprobenplan umfaßte der ursprüngliche Bruttoansatz der Stichprobe A 7008 Adressen. Von den darin enthaltenen 1168 Reserveadressen wurden 158 nicht eingesetzt, weil in dem jeweiligen Sample-Point bereits eine Vollausschöpfung (9 oder 10 Haushalte mit vollständigen Interviews) erreicht war bzw. erreichbar schien. Dagegen wurde die Stichprobe in den Sample-Points mit schlechter Ausschöpfung um insgesamt 1129 Adressen aufgestockt. Insgesamt kamen demnach 7979 Adressen zum Einsatz.

Um das rechnerische Stichprobenbrutto zu erhalten, ist die Gesamtzahl der angesetzten Adressen zu bereinigen um diejenigen, die nicht zur Grundgesamtheit "Privathaushalte ohne die in Stichprobe B gesondert befragten Haushalte" gehören. Die entsprechenden Adressen werden als "qualitätsneutrale Ausfälle" bezeichnet und das Ergebnis als "bereinigtes Brutto".

Wie Tabelle 1 zeigt, gab es in Stichprobe A 5,8% qualitätsneutrale Ausfälle. Das bereinigte Brutto umfaßte 7519 Adressen. Die qualitätsneutralen Ausfälle ergeben sich aus der Art der Adressenentwicklung, nämlich der Auflistung von Haustüradressen durch den Interviewer auf dem vorgegebenen Begehungsweg. Bei einem Teil der Adressen wird erst durch den späteren Haushaltskontakt erkennbar, daß der Haushalt nicht zur Grundgesamtheit gehört (da es sich

In Stichprobe A besteht der allergrößte Teil der Ausfälle aus Verweigerungen. Der Anteil der nicht erreichten Haushalte konnte aufgrund der langen Feldzeit auf 3,2% gesenkt werden (was in normalen Querschnittsbefragungen nicht erreicht wird). Die sonstigen Ausfallgründe setzen sich zusammen aus Haushalten (0,2%), in denen die Befragung wegen Sprachschwierigkeiten nicht durchgeführt werden konnte; es handelt sich dabei um Ausländer, die nicht zu den fünf Nationen der Stichprobe B gehören und daher in Stichprobe A "mitliefen". Schließlich gibt es noch 0,8% der Adressen, über deren Bearbeitung keinerlei Information vorliegt. In der Regel handelt es sich um Reserveadressen, die der Interviewer nicht als zu bearbeitende Adressen erkannt hat.

Bei den Ausländerhaushalten der Stichprobe B sind die Verweigerungsquoten deutlich niedriger als bei den deutschen Haushalten. Dafür ist der Anteil nichterreichter Haushalte höher.

Die Qualität (Querschnittsrepräsentativität) der Stichprobe bemißt sich nach der Übereinstimmung von Merkmalsverteilungen in der realisierten Stichprobe (Nettostichprobe) mit den Merkmalsverteilungen in der Grundgesamtheit. Die Verteilung in der Grundgesamtheit wird durch externe Statistiken geschätzt, die natürlich das Problem aufweisen, daß sie ihrerseits wieder verzerrt sein können (insbesondere die Einkommens- und Verbrauchsstichprobe). Darüberhinaus wurde für das SOEP die Möglichkeit genutzt, eine "interne Validierung" vorzunehmen, die sich durch den Vergleich von Basis-Informationen über Haushalte, für die keine vollständigen Interviews realisiert werden konnten, mit den Informationen für teilnahmebereite Haushalte ergibt.

Bei der internen Validierung geht es um die Frage, ob sich diejenigen Haushalte und Personen, die zur Teilnahme an der Befragung bereit waren, systematisch von denjenigen unterscheiden, die zu einer Teilnahme nicht bereit waren. Die praktische Schwierigkeit bei einer derartigen Validierung besteht darin, daß Strukturmerkmale der Haushalte, die zu einer Teilnahme an einer Befragung nicht bereit sind, normalerweise nicht bekannt sind. Wegen der Bedeutsamkeit der Stichprobenvalidierung für das SOEP wurde hier jedoch versucht,

möglichst viele Informationen auch über die Ausfälle zu erhalten. Dies ist ohnehin schwierig, und wird - zu Recht - von Datenschutzregelungen, die streng einzuhalten sind, erschwert.

Diese Information liegt in Stichprobe A für 31% der Ausfälle vor. In Stichprobe B muß wegen zu geringer Informationen auf diesen Punkt der Analyse verzichtet werden.

Tabelle 2: Interne Validierung der SOEP-Stichproben A und B

	Stichprobe A		Stichprobe B	
	SOEP	Ausfälle	SOEP	Ausfälle
Fallzahl gesamt	4 554	2 906	1 415	652
<u>Haushaltsgröße</u>				
Basis (Fallzahl)	4 554	1 728	1 415	261
%	%	%	%	%
1 Person	26	30	15	15
2 Personen	32	36	17	18
3 Personen	19	16	20	26
4 und mehr Personen	24	18	48	42
<u>Geschlecht des HHV¹⁾</u>				
Basis (Fallzahl)	4 554	1 799	1 415	269
%	%	%	%	%
männlich	76	67	92	91
weiblich	24	33	8	9
<u>Alter des HHV¹⁾</u>	4 554	1 523	1 415	213
%	%	%	%	%
unter 29 Jahre	14	9	12	9
30 - 39 Jahre	18	12	29	28
40 - 49 Jahre	22	18	34	34
50 - 59 Jahre	17	17	22	23
60 - 69 Jahre	13	17	3	6
70 - 79 Jahre	12	20	0	-
80 J. und älter	4	7	0	-
<u>Stellung des HHV¹⁾ im Beruf²⁾</u>				
Basis (Fallzahl)	4 494	894	-	-
Arbeiter	39	33	-	-
Angestellte	36	38	-	-
Beamte	11	10	-	-
Selbständige	11	13	-	-
Auszubildende/Praktikanten	1	1	-	-
Nie erwerbstätig	4	6	-	-
<u>nach Statusgruppen</u>				
niedrigste ³⁾	25	23	-	-
höchste ⁴⁾	15	11	-	-

1) Haushaltsvorstand bzw. Befragungsperson.

2) Bei Nicht-Erwerbstätigen: letzte Stellung.

3) Un- und angelernte Arbeiter, einfache Angestellte.

4) Hochqualifizierte leitende Angestellte, Beamte im höheren Dienst, Freie Berufe, Selbständige mit 10 und mehr Mitarbeitern.

Tabelle 3: Externe Validierung der SOEP-Stichproben A und B sowie C

	A		B		C	
	Stich- probe	Abwei- chung ¹⁾	Stich- probe	Abwei- chung ¹⁾	Stich- probe	Abwei- chung ¹⁾
<u>Haushalte</u>					-	-
<u>Personen</u>						
<u>Bundesland</u>						
Berlin (West)	3,9	-0,6	3,2	-0,6	-	-
Hamburg	2,8	-0,6	2,8	-0,3	-	-
Bremen	1,3	-0,1	1,1	0,0	-	-
Schleswig-Holstein	4,0	-0,2	1,2	-0,7	-	-
Niedersachsen	11,2	0,0	5,7	-0,6	-	-
Nordrhein-Westfalen	26,8	-0,5	28,5	-2,7	-	-
Hessen	8,8	-0,3	13,6	2,6	-	-
Rheinland-Pfalz/Saar	7,5	+0,3	4,2	+0,1	-	-
Baden-Württemberg	15,5	+0,9	26,9	+4,2	-	-
Bayern	18,3	+1,2	12,8	-1,9	-	-
Mecklenburg-Vorpommern	-	-	-	-	11,7	+0,1
Brandenburg	-	-	-	-	16,2	+0,3
Sachsen-Anhalt	-	-	-	-	18,5	+0,3
Thüringen	-	-	-	-	16,7	+0,4
Sachsen	-	-	-	-	30,3	+0,1
Berlin (Ost)	-	-	-	-	6,7	-1,1
<u>Gemeindegrößenklasse</u>						
unter 5.000 Einwohner	12,0	+1,5	2,8	-0,2	-	-
5.000 b.u. 20.000	14,0	+0,2	13,7	-0,3	-	-
20.000 b.u. 100.000	10,1	-0,2	7,7	+0,4	-	-
100.000 b.u. 500.000	16,1	+0,3	17,5	+2,4	-	-
500.000 und mehr						
- Kernzonen	32,4	-1,6	38,1	-3,0	-	-
- Randzonen	15,4	-0,2	20,1	+0,6	-	-
unter 2.000	-	-	-	-	26,6	+2,2
2.000 b.u. 10.000	-	-	-	-	17,2	-2,3
10.000 b.u. 50.000	-	-	-	-	23,9	+0,4
50.000 b.u. 100.000	-	-	-	-	5,8	-1,2
100.000 und mehr	63,9	-1,5	75,7	0,0	26,5	+0,8
<u>Geschlecht</u>						
Männer	47,8	-1,5	-	-	47,5	+0,5
Frauen	52,2	+1,5	-	-	52,5	-0,5
<u>Altersgruppen</u>						
16-19 Jahre	9,0	+0,5	-	-	7,2	+1,0
20-29 Jahre	19,6	+1,3	-	-	20,8	+0,7
30-39 Jahre	16,6	+1,6	-	-	23,8	+4,7
40-49 Jahre	19,6	+1,6	-	-	17,4	+2,2
50-59 Jahre	14,9	+0,2	-	-	16,0	-0,4
60-69 Jahre	10,4	1,1	-	-	9,0	-2,8
70 Jahre und älter	10,0	-4,1	-	-	5,7	-5,5

1) Abweichung der ungewichteten Stichprobe von den amtlichen Bevölkerungszahlen in Prozentpunkten.

In bezug auf die regionalen Verteilungsmerkmale und die Haushaltsstruktur spiegelt die Stichprobe A die bekannten Schwierigkeiten der Umfrageforschung wider. Die Bevölkerung der Ballungsgebiete - und hier insbesondere der Kernzonen - ist für die Teilnahme an Umfragen schwieriger zu gewinnen als die Bevölkerung in den mittleren und kleineren Städten bzw. Gemeinden. Ältere Menschen sind aufgrund von Befragungsproblemen unterrepräsentiert. Beim Vergleich mit den Daten der Bevölkerungsstatistik muß allerdings beachtet werden, daß darin auch Bewohner von Heimen enthalten sind, die mit der SOEP-Stichprobe nicht erfaßt werden. Dadurch sind die Vergleichszahlen etwas verzerrt. Auch diese Verzerrungen sind allerdings relativ leicht durch eine Hochrechnung zu beseitigen.

Die Fallzahl der Ausfälle, für die eine Information zur Stellung im Beruf des Haushaltsvorstandes vorliegt, ist mit $N=894$ recht klein. Als erstes Ergebnis fällt auf, daß der Anteil der Arbeiter unter den Ausfällen niedriger ist als in der realisierten Stichprobe. Angestellte und Selbständige sind demgegenüber unter den Ausfällen leicht überproportional vertreten; bei den Beamten ist insgesamt keine Abweichung erkennbar. Dieses Ergebnis steht in einem bemerkenswerten Widerspruch zu der gängigen These vom "Mittelstandsbias" der Umfrageforschung und belegt die Qualität von Umfragen, die sorgfältig angelegt, durchgeführt und dokumentiert werden.

Die Validierung der Stichprobe B ist weit schwieriger. Während über die deutsche Wohnbevölkerung relativ verlässliche und differenzierte Verteilungsinformationen aus der amtlichen Statistik vorliegen, sind die verfügbaren Strukturdaten über die ausländische Wohnbevölkerung dürftig. Deswegen gibt es kaum Möglichkeiten einer externen Stichprobenvalidierung anhand amtlicher Daten. Umso bedeutsamer ist die interne Validierung (vgl. Tabelle 2).

In bezug auf die soziodemographische Struktur der Stichprobe - Haushaltsgröße sowie Alter und Geschlecht des Haushaltsvorstands - ist das Ergebnis außerordentlich zufriedenstellend. In bezug auf die regionale Verteilung zeigt die Struktur der Ausfälle das gleiche Ergebnis wie für Stichprobe A, nämlich verstärkte Ausfälle in den Kernzonen der Ballungsräume.

Auf die Gewichtung und Hochrechnung der Stichproben wird in Abschnitt 2.3.5 eingegangen.

2.2.2 Stichprobe C

Wie Tabelle 1 (rechter Block) zeigt, wurde für die Stichprobe C in der DDR eine Ausschöpfung von 70 % des bereinigten Ausgangsbruttos erreicht. Das ist ein Feldergebnis, das bei vergleichbaren Studien in der Bundesrepublik auch mit erheblichem Mitteleinsatz kaum noch zu realisieren ist. Auch bei der ersten Welle des SOEP in der Bundesrepublik im Jahre 1984 konnte ein solches Ergebnis trotz großer Anstrengung und mehrmonatiger Feldzeit nicht erreicht werden. Die damalige Ausschöpfung lag mit 60,6 % bei deutschen und 68,0 % bei ausländischen Haushalten deutlich niedriger (linker Block von Tabelle 1).

Von einer bereinigten Bruttostichprobe von 3 114 Haushaltsadressen in der DDR konnten innerhalb einer Feldzeit von etwas mehr als einem Monat insgesamt 2 179 Haushalte zur erfolgreichen Mitarbeit gewonnen werden¹³. In diesen Haushalten wurden insgesamt 4 453 Personen ab 16 Jahren befragt.

Ein spezielles Problem, über das im Jahr 1990 auch andere in Ostdeutschland tätige Umfrageinstitute ähnlich berichten, waren die Totalausfälle von Sample-Points. Totalausfall heißt, daß die Adressen eines Sample-Points völlig unbearbeitet blieben, weil der dafür vorgesehene Interviewer aus irgendeinem Grund diese Aufgabe in der vorgesehenen Zeit nicht übernehmen konnte oder wollte und ein Ersatz nicht oder nicht rechtzeitig verfügbar war (oder die zentrale Feldorganisation gar nicht rechtzeitig Kenntnis davon erhielt, daß die Adressen unbearbeitet blieben).

In der SOEP-Basiserhebung waren 29 von 360 Sample-Points Totalausfälle. Diese verteilen sich jedoch zufällig über die gesamte DDR und führen daher zu keiner erkennbaren Beeinträchtigung der Stichprobe. Die Neutralität dieser Ausfälle wird auch durch die regionale Validierung der Stichproben belegt.

Die Totalausfälle sind einerseits Folge der mangelnden technischen Kommunikationsmöglichkeiten in Ostdeutschland, andererseits auch Ausdruck der Probleme, die mit dem Aufbau bzw. der Neustrukturierung eines Interviewerstabes verbunden sind. Infratest hatte sich dafür entschieden, ein bestehendes Interviewernetz der alten DDR zu übernehmen und auf die neuen Anforderungen hin neu zu organisieren. Ausgewählt wurde das Netz der "Soziologischen Abteilung des Rundfunks der DDR". Eine Analyse eventueller Intervieweffekte in den Befragungsdaten wurde vorgenommen. Sie haben sich - wie auch im Westen - als unbedeutend erwiesen, insbesondere gibt es keine Effekte des Eintrittsdatums (vgl. Riebschläger und Wagner 1991).

Die realisierte Stichprobe wurde sofort nach der Erhebung, soweit Vergleichszahlen vorlagen, mit Daten der amtlichen Statistik der ehemaligen DDR überprüft (vgl. Tabelle 3).

- Regional gibt es praktisch keine Abweichung der Stichprobe in der Verteilung nach Ländern.
- Nach Gemeindegrößenklassen sind zwar gewisse Abweichungen der realisierten Stichprobe zur Grundgesamtheit vorhanden, sie zeigen jedoch nicht die im Westen normalerweise zu beobachtende generelle Unterrepräsentanz der Bevölkerung in Großstädten. Allerdings fällt auf, daß die Personen/Haushalte in Berlin leicht unterrepräsentiert sind.
- In der Verteilung der Geschlechter sind keine nennenswerten Abweichungen erkennbar.
- Ebenso wie in der Stichprobe A sind auch in der Stichprobe C die älteren Altersgruppen - insbesondere die Gruppe der über 70jährigen - in der Nettostichprobe deutlich unterrepräsentiert. Allerdings sind die Vergleichszahlen der Bevölkerungsstatistik etwas verzerrt, da sie nicht um Bewohner von Heimen - die mit dem SOEP nicht erfaßt werden - bereinigt werden können. Gleichwohl: Die Verzerrungen der Altersstruktur im SOEP finden ihren Niederschlag in einer Unterrepräsentation von Einpersonenhaushalten. Deren genaue Repräsentation ist aber - mangels geeigneter amtlicher Rahmendaten - schwer prüfbar. Eine Validierung der sozialen Stellung ist noch schwieriger, da die

Datenbasis der amtlichen DDR-Statistik in diesem Feld noch dürftiger ist (vgl. Wagner und Schupp 1991); zur Hochrechnung vgl. Pischner (1991).

2.3 Interviewer-Stichprobe

Eine wissenschaftliche Panel-Befragung, wie sie das SOEP darstellt, bietet auch die Möglichkeit, als zusätzliche Stichprobe ein "Interviewer-Panel" zu bilden. Um mögliche Interviewer-Effekte überprüfen zu können (vgl. dazu Hoag 1981) werden einige Interviewer-Merkmale vom Umfrage-Institut zusammen mit den eigentlichen Befragungsdaten zur Verfügung gestellt. Diese Angaben der Interviewer sind wie beim SOEP selbst über die Zeit hinweg verknüpfbar und zudem den befragten Haushalten zuordenbar. Damit steht ein zusätzlicher Datensatz zur Verfügung, der wichtige methodische Analysen erlaubt (vgl. dazu z.B. Rendtel 1990; Riebschläger und Wagner 1991).

Die Interviewerstichprobe für die Stichproben A und B umfaßt im Jahr 1984 631 Personen. Die Stichprobe C wurde im Jahre 1990 von 308 Interviewern erhoben.

3 Panelspezifische Konstruktion der Erhebungsinstrumente und Sicherung der Teilnahmebereitschaft

Für eine Panel-Erhebung sind das Konzept der Weiterverfolgung der Stichprobenmitglieder und die Sicherung deren Teilnahmebereitschaft die zentralen Parameter der Studie. Im folgenden wird die für das SOEP entwickelte Vorgehensweise dargestellt, wobei auch deutlich werden wird, daß es aufgrund von Lerneffekten an mehreren Stellen im Laufe der Zeit zu Veränderungen der Vorgehensweise gekommen ist.

3.1 Weiterverfolgungskonzept

Für eine Wiederholungsbefragung, die alle Haushaltsmitglieder erfaßt, ist festzulegen, nach welchen Kriterien Personen weiterverfolgt werden, um die Repräsentation der ausgewählten Grundgesamtheit zu gewährleisten. Wenn man von Zuwanderungen mit eigener Haushalts-

gründung absieht, ist es bei einem entsprechenden Weiterverfolgungskonzept möglich, die natürliche Bevölkerungsdynamik der Grundgesamtheit in der Panel-Stichprobe nachzubilden. Notwendig ist dazu zum einen, daß junge Haushaltsmitglieder, die in das Befragungsalter hineinwachsen (also das 16. Lebensjahr vollenden) auch befragt werden; zum zweiten müssen Personen, die einen "Stammhaushalt" verlassen nach einem bestimmten Muster weiterverfolgt werden. Dies bedeutet, daß Personen, die während der Panel-Laufzeit zugezogen sind und den Haushalt später wieder verlassen, zum Zwecke der Querschnittsrepräsentativität nicht unbedingt weiterverfolgt werden müssen. Weiterverfolgt werden müssen allerdings Kinder, die direkt aus dem Ausland in einen Haushalt gezogen sind (da diese gewissermaßen virtuell zur Grundgesamtheit der ersten Welle gehören), sowie Kinder, die im Verlauf der Paneluntersuchung in Stammhaushalte ziehen und erst später das Befragungsalter erreichen (vgl. dazu Galler 1986). Dieses Konzept der Weiterverfolgung von Stammpersonen wurde ursprünglich für das SOEP gewählt.

Für Längsschnittanalysen, insbesondere von demographischen Ereignissen, ist es aber sinnvoller, nicht nur Stammpersonen, sondern alle Personen weiterzuverfolgen, die jemals mit dem SOEP in Kontakt gekommen sind. Dies führt potentiell dazu, daß sich die Stichprobe schneeballartig vermehrt, da hypothetisch bei genügend großer Mobilität und einer entsprechend langen Laufzeit des Panels am Ende die gesamte Grundgesamtheit in die Stichprobe hineinwächst. Dies ist freilich nur theoretisch der Fall, da in der Praxis durch die sukzessive Verweigerung der Befragten weiter am Panel teilzunehmen, die Stichprobe eher kleiner wird. Wie zu erwarten war ist die Teilnahmebereitschaft bei Personen, die in einen bestehenden Panel-Haushalt hineinziehen, gering. Deswegen ergeben sich keine Größenprobleme für die Stichprobe, wenn Nicht-Stammpersonen weiterverfolgt werden. Im Gegenteil: die Zahl der auswertbaren Fälle sinkt weniger schnell als beim reinen Stammpersonenkonzept. Da die Weiterverfolgung von neu hinzuziehenden Personen bei ereignisorientierten Längsschnittanalysen von besonderem Reiz ist, da es sich bei solchen Personen um überdurchschnittlich

mobile Gruppen handelt, werden seit 1990 (West-Welle 7) alle Nicht-Stammpersonen weiterverfolgt¹⁴.

Bei Querschnitts-Auswertungen auf hochgerechneter Basis werden Haushalte mit Nicht-Stammpersonen etwas niedriger gewichtet als c.p. andere Haushalte, in die keine derartigen neuen Personen hinzugekommen sind (vgl. auch Abschnitt 2.3.5.1 unten). Die vereinfachten Weiterverfolgungsregeln tangieren die Gültigkeit des Stammpersonenkonzepts für Längsschnittuntersuchungen überhaupt nicht. Im Gegensatz zu Querschnitten haben hier nur die Stammpersonen einen positiven Längsschnitt-Hochrechnungsfaktor.

Auch bei dem neuen Weiterverfolgungskonzept stellen allerdings Zuwanderer, die neue Haushalte in der Bundesrepublik Deutschland gründen, ein großes Problem dar. Aufgrund der Tatsache, daß zu Beginn der 80er Jahre nur Zuwanderungen von Arbeitsimmigranten bzw. deren Familienmitglieder eine bedeutsame Rolle spielten, wurde bei der Konzeption des SOEP auf eine sukzessive Ziehung von Zuwandererhaushalten verzichtet. Auch bei der amerikanischen Vorbildstudie PSID wurden trotz der hohen Immigrationsraten in die USA 25 Wellen lang keine Zuwanderer erfaßt. Dies kann damit begründet werden, daß es hauptsächlich auf Längsschnittanalysen ankommt, für die Zuwanderer - definitionsgemäß - keine so große Rolle spielen¹⁵.

Die seit 1988 dramatisch gestiegene Zahl von Zuwanderern in die alte Bundesrepublik Deutschland (Aussiedler, Übersiedler und Asylbewerber) erfordert inzwischen allerdings eine Ergänzung der Panel-Stichprobe, wenn weiterhin Aussagen über die Wohnbevölkerung gemacht werden sollen, da diese Zuwandererpopulation auch quantitativ durchaus bedeutsam wurde (wir schätzen, daß diese Population im Jahre 1991 bereits ca. 3 % der Privathaushalte in der Bundesrepublik Deutschland ausmacht). Auf verschiedene Möglichkeiten einer solchen Ergänzung wird in Abschnitt 2.3.7 eingegangen werden.

3.2 Verlaufsbeobachtung und Nacherhebung

3.2.2 Nacherhebung von Längsschnittmerkmalen bei lückenhaften Verläufen

Beim Sozio-ökonomischen Panel werden 'endgültige' und 'vorläufige' (oder 'temporäre') Ausfälle unterschieden. Als endgültig werden Ausfälle bewertet, bei denen ein Haushalt/eine Person trotz intensiver Bemühungen nicht mehr aufzufinden ist oder wenn eindeutig die weitere Teilnahme an der Befragung verweigert wird. In allen anderen Fällen wird der Ausfall zunächst als 'vorläufig' bewertet. Das heißt, daß in der Folgewelle ein erneuter Versuch unternommen wird, den Haushalt/die Person aufzufinden (falls dies das Problem war) und ihn bzw. sie zur weiteren Teilnahme an der Befragung zu gewinnen. Gelingt dies wiederum nicht, wird der 'vorläufige' zu einem 'endgültigen' Ausfall - dieser Haushalt bzw. diese Person wird in den Folgewellen des Panels nicht mehr bearbeitet. Kann der Haushalt/die Person jedoch zu einer erneuten Teilnahme motiviert werden, so ist durch den Ausfall in der vorhergehenden Welle eine Lücke in den Längsschnittdaten entstanden (vgl. Tabelle 4).

Tabelle 4: Temporäre Ausfälle

Auswertbare Längsschnittdatensätze der Stichproben A und B mit "Lücken"

	Personen	Haushalte
Welle 2	202	91
Welle 3	176	79
Welle 4	83	33
Welle 5	126	55
Welle 6	147	66

Zusätzlich zu diesen haushaltsbezogenen Lücken treten Lücken bei einzelnen Personen auf, wenn ein Haushalt in einer Welle nicht vollständig befragt werden konnte, in der folgenden Welle aber wieder alle Haushaltsmitglieder ein Interview gewähren. So gab es beispielsweise in der 2. West-Welle 123 Haushalte mit unvollständig realisierten Interviews, 53 dieser Haushalte beteiligten sich an der nachfolgenden 3. Welle vollständig. Pro Welle ist demnach mit 250 Personen zu rechnen, die lückenhafte Längsschnittverläufe aufweisen. Diese Fallzahl ist zwar im Verhältnis zur gesamten Stichprobe gering; es ist aber zu beachten, daß die Lücken jede Welle erneut und immer wieder bei anderen Haushalten bzw. Personen auftreten. Es wurde deshalb ein Verfahren entwickelt, nach dem zumindest die wichtigsten Längsschnittmerkmale für das Lückenjahr nacherhoben werden können.

Im Sommer 1987 - etwa parallel zum Dankschreiben für die Teilnahme an der 4. West-Welle - wurde zunächst damit begonnen, die Lücken aus der 3. Welle nachzuerheben. Es wurde ein unerwartet guter Erfolg erzielt. Von den 192 zu erfragenden Personen aus 134 Haushalten konnten für 176 Personen die Längsschnittinformationen nacherhoben werden. In der überwiegenden Zahl der Fälle geschah dies telefonisch. Aufgrund dieses guten Ergebnisses wurde die Entscheidung getroffen, die Lücken von Welle 2 ebenfalls nachzuerheben. Die lange Zeitspanne zwischen dem Feststellen der Lücken von Welle 2 (Sommer 1986) und der Nacherhebung (Herbst 1987) erwies sich jedoch als nachteilig; nur rund 75 % der betroffenen Personen konnten in die Feldarbeit einbezogen werden, da ein relativ hoher Anteil der Zielpersonen zwischenzeitlich aus dem Panel ausgeschieden war.

Seit 1987 (West-Welle 4) erfolgte eine regelmäßige Nacherhebung.

3.2.3 Nacherhebung von biographischen Angaben

Die ersten drei Wellen der Stichproben A und B (West) enthielten als Themenschwerpunkte biographische Angaben, die im Verlauf des Panels nur einmal erhoben werden müssen: Erwerbsbiographie in Jahresschritten, beginnend mit dem Alter von 15 Jahren (Welle 1);

Familienstandsbiographie, Fragen zur Kindheit und zum Auszug aus dem Elternhaus, für Frauen Angaben über die von ihnen geborenen Kinder (Welle 2) sowie Soziale Herkunft, Einstieg in den Beruf (Welle 3).

Diese Angaben fehlen ganz oder teilweise für Personen, die erst in späteren Wellen in das Panel einbezogen wurden und für solche Personen, die in der 2. oder 3. Welle ein temporärer Ausfall waren. Die entsprechenden Fragen wurden deshalb in einem Zusatzfragebogen 'Biographie' zusammengefaßt. Den neuen Personen, die in der 2., 3. oder 4. Welle ins SOEP gekommen sind, wurde der Zusatzfragebogen im Herbst 1987 zusammen mit einem kleinen Geschenk, einem Taschenkalender, zugeschickt. Dabei wurden allerdings die neuen Personen, die gerade 16 Jahre alt geworden waren und deshalb in die Befragung einbezogen worden waren, nicht berücksichtigt, weil die meisten Fragen für sie noch nicht zutreffen bzw. die Angaben zum Elternhaus bereits vorhanden sind. Die Nacherhebung der biographischen Angaben bei den temporären Ausfällen der Welle 2 erfolgte zumeist - zusammen mit der Nacherhebung der Längsschnittmerkmale bei lückenhaften Verläufen (vgl. die Ergebnisse im vorhergehenden Abschnitt) - telefonisch.

Seit 1988 (West-Welle 5) erfolgt eine regelmäßige Lückenerhebung durch einen speziellen Biographie-Fragebogen. Für die Ost-Stichprobe C erfolgt die biographische Nacherhebung für alle Befragungspersonen ab Welle 3 im Jahre 1992.

3.3 Methodenmix der Erhebung

Die Erhebungsinstrumente des SOEP bestehen aus

- Adressenprotokoll,
- Haushaltsfragebogen
- Personenfragebogen.

Die Fragebögen sind bezüglich des Layouts und der Filterführung relativ aufwendig gestaltet, da diese von den Interviewern flexibel eingesetzt werden müssen:

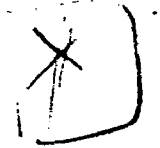
- in den Stichproben A und C ist neben der Grundform des mündlichen Interviews auch ein Selbstausfüllen des Fragebogens durch die Befragten möglich.
- in Stichprobe B kann das Interview wahlweise in deutscher Sprache oder in der jeweiligen Fremdsprache (oder auch in einer Mischform) durchgeführt werden.

Mit dem Adressenprotokoll werden auch eine Vielzahl von Informationen über die Zusammensetzung und Veränderung des Haushaltes festgehalten, die notwendig sind, um komplizierte Längsschnittanalysen durchführen zu können.

Die Erhebungsverfahren des SOEP stellen einen Methoden-Mix dar; für Methodenanalysen werden die Informationen über die benutzte Methode in jedem Falle im auswertbaren Datensatz gespeichert:

- Grundform der Erhebung ist das mündliche, persönlich durchgeführte Interview
- die Befragungsperson darf jedoch den vom Interviewer überreichten und erläuterten Fragebogen auch selbst ausfüllen.
- im Falle von Verweigerung oder nicht angetroffenen Zielpersonen wird ein nochmaliger Interviewerbesuch schriftlich oder durch ein telefonisches Kontaktgespräch vereinbart.
- Wenn die Befragungspersonen dies wünschen, wird auf den (nochmaligen) Interviewerbesuch verzichtet und das Interview ausnahmsweise schriftlich (d.h. postalisch) oder telefonisch durchgeführt.

Eine Vorgabe wird freilich ausgesprochen strikt gehandhabt: Auskünfte über eine Befragungsperson dürfen nur von dieser selbst eingeholt werden. Proxy-Interviews, wie sie z.B. bei der US-amerikanischen SIPP-Studie gängig und beim PSID für alle Haushaltsmitglieder notwendig sind, die nicht Haushaltsvorstand sind, sind grundsätzlich nicht erlaubt. Es gibt nur wenige Ausnahmefälle, in denen z.B. wegen Krankheit eines Haushaltsmitglieds - mit dessen Zustimmung - der Personenfragebogen von einem anderen Haushaltsmitglied beantwortet wird.



3.4 Panelpflege

3.4.1 Instrumente

Die Panelpflege beim SOEP zur Aufrechterhaltung des Stichprobenbestandes gliedert sich in zwei wichtige Bereiche:

- die Kontaktaufnahme und -pflege, um die befragten Personen zur erstmaligen und wiederholten Mitarbeit zu gewinnen (Motivation)
- das Wiederauffinden von Haushalten, die zwischen zwei Befragungen innerhalb der Bundesrepublik Deutschland umziehen (Aktualisierung des Adressenbestandes).

Folgende Maßnahmen der Panelpflege lassen sich beim SOEP unterscheiden:

- Die Kennzeichnung der Studie durch eine populäre Form. Die Befragten in allen Teilstichproben kennen das SOEP unter dem Namen "LEBEN IN DEUTSCHLAND".
- Eine illustrierte Informationsbroschüre über Ziele der Untersuchung (einschließlich der jeweiligen Übersetzung in die jeweilige Heimatsprache bei Stichprobe B)
- Informationsblatt zum Datenschutz¹⁶
- Dankschreiben nach Abschluß der Feldarbeiten jeder Welle
- Jede Befragungsperson erhält ein Monatslos der ZDF-Lotterie "Der Große Preis"
- Seit 1987 (4. West-Welle) erhalten alle Panelhaushalte ein kleines Geschenk ("Treueprämie") im Wert von 5.- bis 10,- DM.

Neben diesen direkten Maßnahmen hat sicherlich die Motivation der Interviewer ebenfalls erhebliche Bedeutung und Einfluß auf die Teilnahmebereitschaft der Befragten. Hierbei zählen eine gute Schulung, ausreichende Informationen über das Projekt, eine übersichtliche Gestaltung der Erhebungsinstrumente und den Informationen über Forschungsergebnisse zu den Grundlagen für eine erfolgreiche Interviewertätigkeit. So wird seit Jahren allen an der Untersuchung beteiligten Interviewern zum Jahresende vom Auftraggeber (dem DIW) eine Dankeskarte zugesandt, um die Relevanz ihres Engagements ausdrücklich zu unterstreichen.

Bei der ersten Welle in der damaligen DDR war es von Vorteil, daß das SOEP gegenüber den Befragten den Projekttitel "Leben in Deutschland" führt. Im Anschreiben und der Broschüre für die Basisbefragung in der DDR wurde darauf hingewiesen, daß die Erhebung in der alten Bundesrepublik Deutschland seit 1984 läuft und ihr hoher Anspruch, das Leben in Deutschland zu begleiten, nunmehr realisiert werden kann.

Im Gegensatz zur ersten Welle des SOEPs in der BRD im Jahre 1984 konnte in der DDR das geplante Interview nicht durch ein postalisch versandtes Ankündigungsschreiben den betroffenen Haushalten frühzeitig mitgeteilt werden. Dies hat jedoch zu keinen nennenswerten Problemen geführt. In der DDR gehörten unangemeldete Besuche zum Alltag, da aufgrund des unzureichenden Telefonnetzes Verabredungen kaum telefonisch angekündigt wurden, wie dies z. B. in der BRD seit Jahren zur Interviewerroutine beim SOEP gehört.

3.4.2 Kontinuität Haushalt - Interviewer

Eine Panel-Untersuchung stellt nicht nur für den Befragten ein Panel dar, sondern auch für die Interviewer selbst.

Beim Interviewereinsatz für das SOEP war zwischen zwei Strategien zu wählen:

- Der Verteilung der Befragungsarbeit auf möglichst viele Interviewer; d.h. im Grenzfall, daß die Zahl der eingesetzten Interviewer der Zahl der Sample-Points entsprechen würde.
- Der Konzentration der Befragungsarbeit auf möglichst wenige, hoch qualifizierte Interviewer.

Ausgangspunkt war im Jahre 1984 die erste Strategie, d.h. ein maximaler Einsatz von Interviewern, um möglichst wenige Interviewer-Effekte zu provozieren. Freilich hat sich im Laufe der Feldarbeit in mehreren Wellen herausgestellt, daß für die hohe Ausschöpfung, die beim SOEP verlangt wird, einige Interviewer besser und andere weniger gut geeignet sind. Hinzu kommt, daß ein Interviewerwechsel eine entscheidende Determinante für die Verwei-

gerung der weiteren Mitarbeit am Panel durch die Befragten darstellt (vgl. Rendtel 1990)¹⁷. Es war also notwendig, ein Optimum zwischen einer möglichst großen Zahl von Interviewern und der Konzentration auf möglichst wenige Interviewer zu suchen, wobei noch zu berücksichtigen ist, daß der Ausfall eines einzigen Interviewers dann die Gefahr vergrößert, daß sehr viele Panel-Haushalte verweigern.

3.5 Adressenverfolgung und Panel-Datei

Die Feldarbeit beim SOEP erstreckt sich über mehrere Monate, da es notwendig ist, eine über 90 %ige Ausschöpfung bei der Weiterverfolgung zu gewährleisten. Um stichtagsbezogene Auswertungen durchführen zu können, ist das Datum des Interviews im Datensatz festgehalten.

Seit Welle 3 wurde der Ablauf der Feldarbeit umgestellt. Jetzt werden zuerst die als schwierig eingestuften Fälle kontaktiert, um hier eine lange Bearbeitungszeit zu ermöglichen, ohne die Dauer der Feldzeit zu belasten.

Im SOEP werden für jeden Befragungshaushalt und jede Befragungsperson, die nicht mehr an derselben Adresse wie im Vorjahr angetroffen werden, Recherchen über den weiteren Verbleib angestellt. Informationen zu diesen Adressverfolgungsarbeiten werden in der Paneldatensatz gespeichert.

Jährlich sind aus der Stichprobe A etwa 7 % aller Haushalte an einer anderen Adresse anzutreffen, wobei dieser Anteil nach einem anfänglichen Ansteigen seit 1988 wieder zurückgeht. Hinzu kommen - wenn man so will - 'Teilumzüge' von Personen, die Haushalte verlassen und an einer neuen Adresse für das SOEP-Weiterverfolgungskonzept auch einen 'neuen' Haushalt bilden. Dies trifft auf jährlich rund 4 % aller Haushalte der Stichprobe A zu.

Der Erfolg der Adressrecherchen des Umfrageinstituts ist als sehr gut zu bewerten, da er seit Beginn des Projektes bei über 90 % liegt. Quellen für die Adressenvermittlung sind in weit

mehr als der Hälfte die Interviewer selbst, die der Einsatzzentrale den neuen Wohnort mitteilen. Die verbliebenen Fälle können zu etwa gleichen Teilen durch Anfragen beim zuständigen Postamt sowie dem Einwohnermeldeamt ermittelt werden.

4 Entwicklung der Stichproben A und B sowie C

Im folgenden wird auf die empirische Entwicklung der Stichprobengröße und -struktur eingegangen.

4.1 Stichproben A und B

4.1.1 Determinanten der Entwicklung

Die Stichprobenentwicklung im SOEP resultiert zum einen aus demographischem Wandel (Abgänge durch Tod oder Umzug ins Ausland, Zugänge durch Erreichen des Befragungsalters oder Entstehen neuer Haushalte). Zum anderen wird der Stichprobenbestand ganz wesentlich durch den Erfolg der Feldarbeit bestimmt. Der Erfolg der Feldarbeit wird im folgenden unter zwei Aspekten beurteilt:

- Gelingt es, mit den Haushalten aus der vorhergehenden Befragungswelle wieder Kontakt aufzunehmen?
- Erhält man von den Haushalten mit Kontaktaufnahme ein Interview?

Tabelle 6 zeigt, daß die Ausfälle aufgrund von Nicht-Erreichbarkeit insgesamt eine untergeordnete Rolle spielen. Auch konnte die Ausfallquote von 1,9 % in Welle 2 auf 0,9 % in Welle 7 gesenkt werden. Allerdings ist die Kontaktaufnahme mit den Haushalten der vorhergehenden Befragungswelle unterschiedlich schwer. Dies belegen in Tabelle 6 die deutlich verschiedenen Ausfallraten in den jeweiligen Haushaltsgruppen. In der Gruppe der verzogenen Ein-Personen-Haushalte ist daher die Ausfallrate wesentlich höher als bei Haushalten ohne Adressenänderung. In beiden Haushaltsgruppen konnte die Feldarbeit erheblich verbessert

werden; noch schwieriger sind "abgespaltene Haushalte" wiederzufinden. Hier zeigt sich auch kein Trend zur Verbesserung der Feldarbeit.

Tabelle 6: Ausfallraten in % aufgrund von Nichterreichbarkeit

	Welle 2		Welle 3		Welle 4		Welle 5		Welle 6		Welle 7	
	N	%	N	%	N	%	N	%	N	%	N	%
Insg.:	6051	1,9	5814	1,4	5465	1,0	5342	0,9	5156	0,9	5044	0,9
davon:												
HH ohne Adressen-änderung:	5413	0,8	5039	0,4	4808	0,1	4683	0,1	4545	0,2	4472	0,0
verzogene Einpersonenhaushalte	119	21,0	180	14,4	142	7,7	143	5,6	126	4,7	122	5,7
abgespaltene Haushalte	221	11,7	295	8,4	242	10,4	242	7,4	246	11,8	263	12,9

N = Anzahl der Fälle insgesamt (alte Haushalte + neue Haushalte - durch Tod aufgelöste Einpersonenhaushalte - ins Ausland verzogene Haushalte)

% = Prozentsatz der Haushalte ohne Kontaktaufnahme (Basis N)

Datenbasis: Das Sozio-ökonomische Panel, Welle 1-7

Der Erfolg der nächsten Stufe der Feldarbeit wird daran gemessen, ob es gelingt, von den wieder erreichten Haushalten ein Interview zu erhalten. Hierbei werden die Haushalte in drei Untergruppen eingeteilt: Teilnehmer des Vorjahres, vorläufige Ausfälle des Vorjahres und die Gruppe der neu entstandenen Haushalte. Die in Tabelle 7 ausgewiesenen Ausfallquoten zeigen den unterschiedlichen Erfolg der Feldarbeit bei den einzelnen Teilgruppen. Erwartungsgemäß wird das beste Ergebnis bei den Haushalten erzielt, die sich auch im Vorjahr am Panel beteiligten. Erfreulich ist hierbei, daß der Erfolg der Feldarbeit wesentlich gesteigert werden konnte. Von Welle 2 bis Welle 4 halbierte sich die Ausfallrate von 9,7 % auf 4,7 %. Allerdings konnte diese positive Entwicklung in Welle 5 nicht fortgesetzt werden. Hier vergrößerte sich die Ausfallrate um 1,7 Prozentpunkte auf 6,4 %. Hier ist jedoch zu beachten,

daß in Welle 5 mit dem Befragungsschwerpunkt 'Vermögen' ein ausgesprochen sensibles Themenfeld erfragt wurde. In der 6. und insbesondere der 7. Welle zeigt sich ein Abklingen dieses negativen Effektes.

Schwieriger zu befragen sind dagegen neue Haushalte (siehe mittlere Zeile von Tabelle 7), die häufig von jungen Personen nach der Trennung von den Eltern gegründet werden. Doch auch hier konnte die Ausfallrate von 29,2 % (Welle 2) auf 17,6 % (Welle 4) beträchtlich gesenkt werden. In Welle 5 steigt diese Ausfallrate gegenüber diesem Niveau um 6,1 Prozentpunkte auf 23,7 %. Der negative Effekt des Befragungsschwerpunktes Vermögen ist in der Gruppe der neuen Haushalte damit wesentlich größer gewesen als bei den Mitgliedern wiederholt befragter Haushalte.

Tabelle 7: Ausfallraten in % der erreichten Haushalte nach Teilnahme im Vorjahr

	Welle 2		Welle 3		Welle 4		Welle 5		Welle 6		Welle 7	
	N	%	N	%	N	%	N	%	N	%	N	%
Teilnehmer des Vorjahres	5 742	9,7	5 235	8,3	5 019	4,7	4 937	6,4	4743	5,8	4630	4,8
Neue Haushalte	195	29,2	238	21,4	182	17,6	194	23,7	183	16,9	198	23,2
Vorläufige Ausfälle des Vorjahres	-	-	259	59,5	197	52,8	154	71,4	169	57,4	154	49,4

N = Anzahl der Fälle insgesamt (Haushalte mit Kontaktaufnahme)

% = Prozentsatz der Haushalte ohne realisiertes Interview (Basis N)

Datenbasis: Das Sozio-ökonomische Panel, Welle 1-7

Allerdings konnte der positive Trend bei der Entwicklung der Antwortraten der neuen Haushalte in der 7. Welle nicht fortgesetzt werden. Hier stieg die Ausfallrate wieder auf 23,2 % an. Neu gegründete Haushalte - wie übrigens auch umgezogene "alte" Haushalte - stellen ein

Haushalts-Panel wie das SOEP vor gewisse Schwierigkeiten bei der Motivation zur Teilnahme (vgl. Rendtel 1990). Hier wird von den Befragten nicht immer verstanden, weshalb sie in einem neuen Haushaltskontext bzw. an einem neuen Wohnort weiter an der Befragung teilnehmen sollen.

Obwohl bei Start des SOEP der Anstaltsbereich nur für die Ausländerstichprobe - aufgrund der Adressenauswahl aus Personenregistern - erhoben werden konnte, bestand anfänglich die Hoffnung, daß die Panelstichprobe - via Weiterverfolgung - in den Folgewellen gleichsam in den Anstaltsbereich hineinwächst. Es zeigte sich jedoch, daß der Übergang in den Anstaltsbereich (z.B. Altenwohnheime) in der Panel-Stichprobe praktisch nicht abgebildet wird. In der deutschen Teilstichprobe stagniert die Zahl der Bewohner im Anstaltsbereich bei ca. 30 Personen, während in der Ausländerstichprobe diese Zahl kontinuierlich von 47 Personen (Welle 1) auf 19 Personen (Welle 7) abnimmt. Dies ist weitgehend dem zu erwartenden Rückgang der Ausländer in Berufstätigenwohnheimen zuzuschreiben. Obwohl also kaum Personen mit Interviews im Anstaltsbereich in der Stichprobe sind, kann das Ereignis des Wechsels doch sehr sicher festgestellt und daher auch analysiert werden.

Eine erneute Befragung von (vorläufigen) Ausfällen ist an sich nicht üblich. Durch diese Anstrengung können inzwischen immerhin ca. 50 % der vorläufigen Ausfälle für die Befragung wieder gewonnen werden (letzte Zeile in Tabelle 7). Diese Haushalte wären andernfalls endgültig aus der Untersuchung ausgeschieden. Pro Welle wird also eine Reduzierung der Stichprobe um ca. 80 Haushalte vermieden.

4.1.2 Ergebnisse zur Stichprobenentwicklung

Aus Abbildung 1 geht hervor, wie sich die Fallzahlen der realisierten Personen- und Haushaltsinterviews bis zur Welle 7 entwickelt haben. Insgesamt zeigt sich eine deutliche Stabilisierung der Fallzahlen des Panels. In Welle 7 wurde mit 4 640 Haushalten und 9 522 Befra-

gungspersonen ein Interview realisiert. Dies entspricht 78,3 % bzw. 77,8 % der Ausgangsgröße von Welle 1.

Hier etwa Abb. 1

Bei der Betrachtung der Personen in den SOEP-Haushalten ist zwischen allen Haushaltsmitgliedern und den persönlich interviewten Personen (Befragungspersonen) zu unterscheiden. Zwar liegen nur für Befragungspersonen eine große Zahl von Informationen vor, insbesondere können natürlich nur für direkt Befragte subjektive Indikatoren erhoben werden. Gleichwohl werden mit Hilfe des Haushaltsfragebogens auch zentrale (Längsschnitt-) Indikatoren für nicht-befragte Kinder erhoben. Insgesamt liegen für die SOEP-Haushalte Informationen für 16 205 Personen in Welle 1 vor. In Welle 7 sind es mit 12 253 Personen noch 75,6 % der Ausgangsgröße.

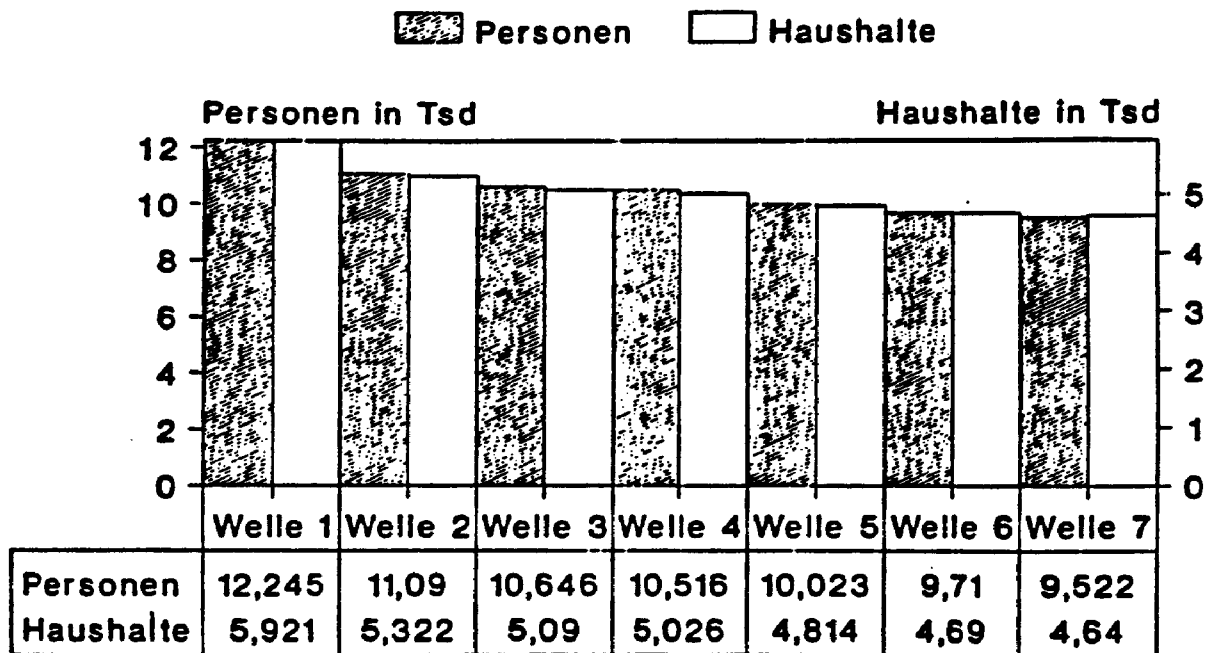
Betrachtet man nur die Befragungspersonen, so zeigt Abbildung 1, daß die Entwicklung des Haushalts- und des persönlich interviewten Personenbestandes annähernd gleichförmig verläuft. Hier wäre zunächst zu erwarten gewesen, daß sich der Personenbestand ungünstiger entwickelt als der Haushaltsbestand. Dieser Effekt tritt ein, wenn in einem nennenswerten Anteil der Haushalte zwar ein Haushaltsinterview realisiert werden kann, nicht jedoch alle Befragungspersonen zu einem Interview bereit sind. Es läßt sich jedoch zeigen, daß die Auskunftsbereitschaft auf Haushalts- und Personenebene weitgehend identisch ist (vgl. Rendtel 1990).

Bei der Entwicklung der SOEP-Stichproben A und B muß beachtet werden, daß mit der Stichprobe B beachtliche Wanderungsbewegungen nachgebildet werden, die in ihrem Netto-Effekt ein Schrumpfen der Ziel-Stichprobe implizieren, da seit 1984 mehr Ausländer, die von der Stichprobe B repräsentiert werden, die Bundesrepublik Deutschland verlassen haben als zugezogen sind. 565 Personen der Panelstichprobe haben die Bundesrepublik Deutschland verlassen, während 210 Personen im Zuge des "Familiennachzuges" in die Stichprobe gelangt

Abbildung 1



Anzahl der realisierten Interviews im Sozio-ökonomischen Panel



Datenbasis: Das Sozio-Ökonomische Panel
Welle 1 bis Welle 7

sind. Allerdings werden bislang nur 122 dieser Personen selbst befragt, da es sich in der Hälfte der Fälle um Kinder unter 16 Jahren handelt.

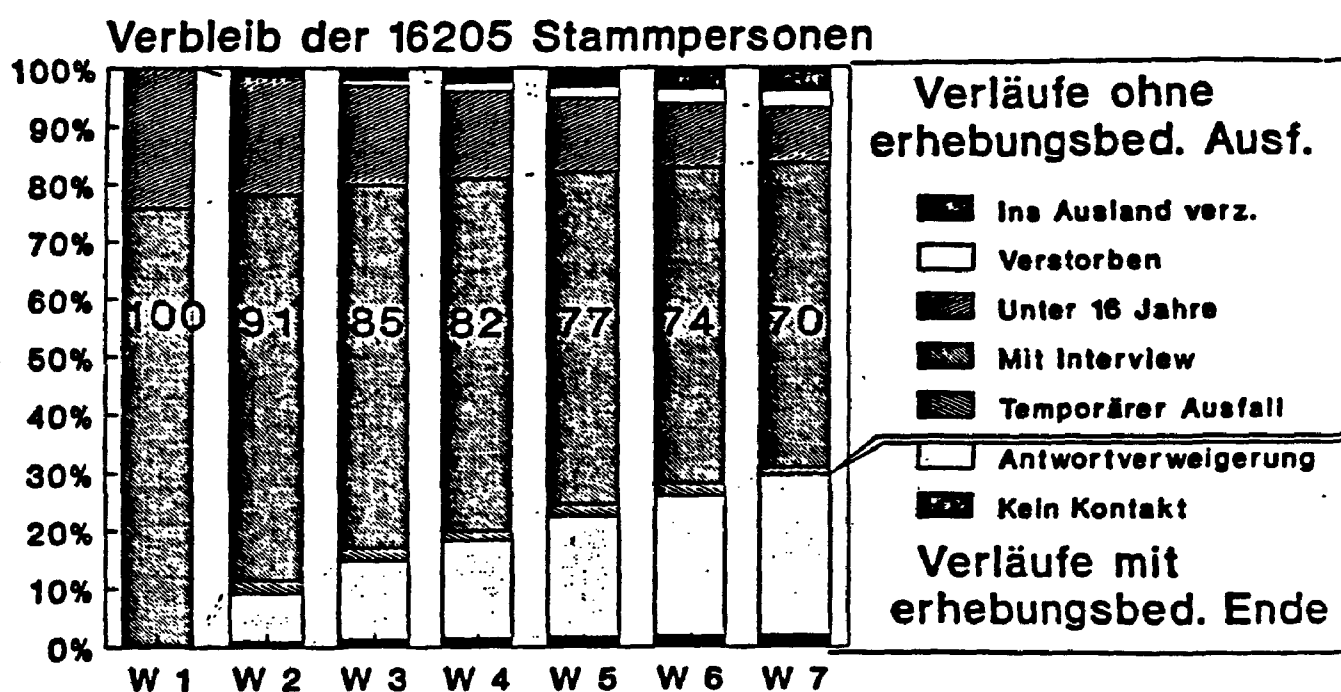
Für Längsschnittauswertungen ist weniger die in der Abbildung 1 ausgewiesene Entwicklung des Querschnittbestandes des SOEPs von Belang als vielmehr der Bestand an Personen, die während der Laufzeit des Panels im Stichprobenbestand sind. Naturgemäß nimmt insbesondere der Bestand derjenigen, die in allen Wellen befragt wurden, durch demographische Bewegungen (Tod, Fortzug ins Ausland) ab. Hinzu kommen erhebungsbedingte Ausfälle (kein Kontakt, Antwortverweigerung). In Abbildung 2 wird die Entwicklung des Bestands der 16 025 Stammpersonen aus Welle 1 dargestellt. Für 70 Prozent des Ausgangsbestandes stehen vollständig Längsschnittinformationen über die ersten sieben Befragungswellen zur Verfügung. Bis Welle 7 belaufen sich die Verluste durch demographische Bewegungen auf 6,9 % des Ausgangsbestandes. Die Verluste durch erhebungsbedingte Ausfälle sind vier mal so groß (29,8 % des Ausgangsbestandes) wie die Verluste durch demographische Veränderungen. Diese Entwicklung ist im Hinblick auf die Analysierbarkeit des SOEPs in keiner Weise beunruhigend: weder im Hinblick auf die Repräsentativität noch im Hinblick auf die analysierbaren absoluten Fallzahlen. Die Repräsentativität der Ergebnisse, die mit dem SOEP erzielt werden, wird durch die Gewichtung und Hochrechnung der Daten gewährleistet (vgl. Abschnitt 2.3.5 unten). Die Fallzahlen selbst sind sowohl im Querschnitt, für den sich die Fallzahlen durch das dynamische Weiterverfolgungskonzept stabilisieren, wie für den Längsschnitt nach wie vor ausreichend. Wenn es für bestimmte Untergruppen Fallzahlprobleme gibt, die sich in mangelnder Signifikanz von Ergebnissen ausdrücken, so sind dies Probleme, die aufgrund der Gesamtfallzahl bereits in der ersten Welle gegeben waren.

Hier etwa Abb. 2

Im Hinblick auf die Schätzung von kausalanalytischen Längsschnittmodellen ist zu bedenken, daß zunehmend Methoden entwickelt werden, die nicht nur diejenigen Fälle im Längsschnitt auswerten, für die über alle Wellen hinweg vollständige Informationen vorliegen (complete

Abbildung 2

Entwicklung Stichprobenbestands Basis : Alle Stammpersonen der 1. Welle



Datenbasis : Das Sozio-ökonomische Panel
Welle 1 bis Welle 7

case analysis), sondern daß auch "unvollständige Verläufe" einbezogen werden können (complete information analysis). Man spricht auch von der Analyse von "unbalanced panels". Für eine ökonometrische Anwendung am Beispiel von Einkommensverläufen vgl. Löwenbein und Rendtel (1991) sowie Licht und Steiner (1991). Die Schätzung von Längsschnittmodellen über "Unbalanced-Panel"-Methoden wird mit zunehmender Laufzeit des Panels immer dringlicher, da der Datenbestand immer weniger die Gestalt eines Rechteckfiles hat. Die Abbildung 3 verdeutlicht diese Entwicklung. Im oberen Teil wird das allmähliche Abschmelzen des Bestands der 12 245 Befragungspersonen der 1. Welle dargestellt. Hierbei sind alle Ausfälle zusammengefaßt worden ("ausgeschieden" = demographische plus erhebungsbedingte Ausfälle). Allerdings wurden temporäre Ausfälle (vgl. zu deren Umfang Abbildung 2) nicht berücksichtigt.

Insgesamt sind von den 12 245 befragten Personen der 1. Welle noch 61,2 % in der Welle 7 in der Stichprobe. Im Schaubild entspricht diese Gruppe gerade 52,6 % der Bezugsgröße von 14 256 Personen. Gleichzeitig wurden in der 7. Welle 2 013 Personen befragt, die in der ersten Welle noch nicht im Befragungsalter waren oder erst später in die Panelstichprobe aufgenommen wurden. Im unteren Teil von Abbildung 3 wird dargestellt, seit wann diese Personen den Fragebogen beantwortet haben. Der Datensatz des Panels hat damit annähernd die Gestalt eines Rhombus. Die Beschränkung auf Personen, die in allen Wellen an der Befragung teilgenommen haben, wertet also nur die Daten von 7 506 oder 14 256 ($= 12\,245 + 2\,013$) Personen aus. Dem entspricht die in Abbildung 3 ausgewiesene Quote von 52,6 %. Noch ungünstiger wird dieses Verhältnis, wenn man einerseits in die Bezugsgröße die 611 Personen einbezieht, die zwischen Welle 2 und Welle 6 in die Panel-Stichprobe eintraten und diese vor der 7. Welle wieder verließen, und andererseits nur Personen ohne temporäre Ausfälle im Rahmen der "Complete-Case"-Analyse berücksichtigt. Dies senkt die Personenzahl nochmals um 415 Personen. Der Anteil der solchermaßen analysierten Personen beträgt dann nur noch 47,7 %.

Hier etwa Abb. 3

Für die Analyse von unvollständigen Verläufen ist die Methode der Ereignisanalyse am bekanntesten (vgl. dazu den Beitrag von Hujer und Schneider sowie Ott und Pötter in diesem Band sowie zahlreiche Beiträge bei Hujer et al. 1991). Für ereignisbezogene Analysen bietet das SOEP trotz sinkender Fallzahlen eine mit jeder Welle zunehmende Zahl von analysierbaren Ereignissen. Abbildung 4 verdeutlicht diese Entwicklung. Interessant ist insbesondere der im Trend insgesamt lineare Anstieg aller Ereignisse. Diese Entwicklung macht auch Prognosen für Fallzahlen künftiger Wellen möglich.

Auch wenn es sich bei einer Vielzahl von Episoden um keine klassischen "unabhängigen Ereignisse" handelt, weil sie bei ein und derselben Person auftreten, so vergrößert die zunehmende Zahl von Spells doch die "Mächtigkeit" von Panel-Daten.

Hier etwa Abb. 4

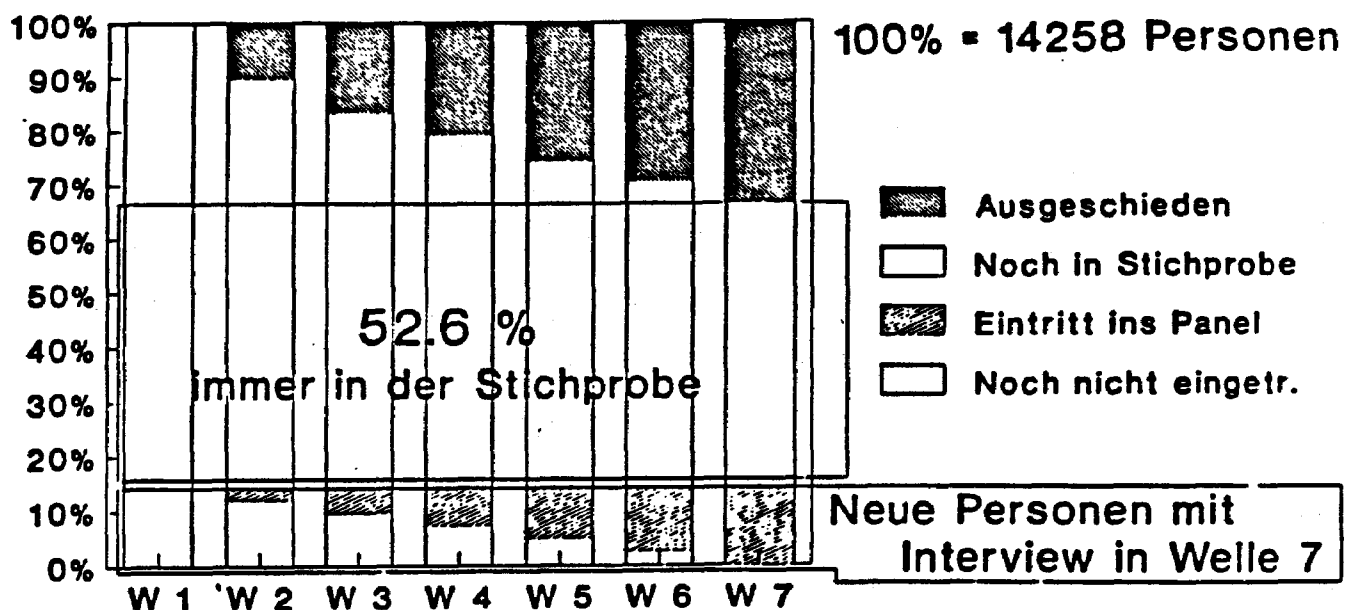
4.2 Stichprobe C

Für die Entwicklung der Stichprobe C liegen bislang nur vorläufige Informationen für Welle 2 vor, die im ersten Halbjahr 1991 erhoben wurde. Es zeigt sich folgendes im ganzen erfreuliches Bild:

Nach der sehr guten Stichprobenausschöpfung der ersten Welle (Ost) mit 70 % hatte sich die Ausgangssituation im März in den neuen Bundesländern zur Durchführung empirischer Erhebungen gewandelt. Der von mehreren Instituten berichtete Stimmungsumschwung in der Bevölkerung dämpfte den Optimismus hinsichtlich der Teilnahmebereitschaft am SOEP im Jahre 1991. Die Feldarbeit ist zum jetzigen Zeitpunkt (Stand 15. Juni 1991) noch nicht abgeschlossen; ein abschließendes Ergebnis kann zum gegenwärtigen Zeitpunkt noch nicht berichtet werden. Die Hauptlieferung der realisierten Haushalte ist jedoch mittlerweile im Prozeß der Datenaufnahme und des Editings.

Abbildung 3

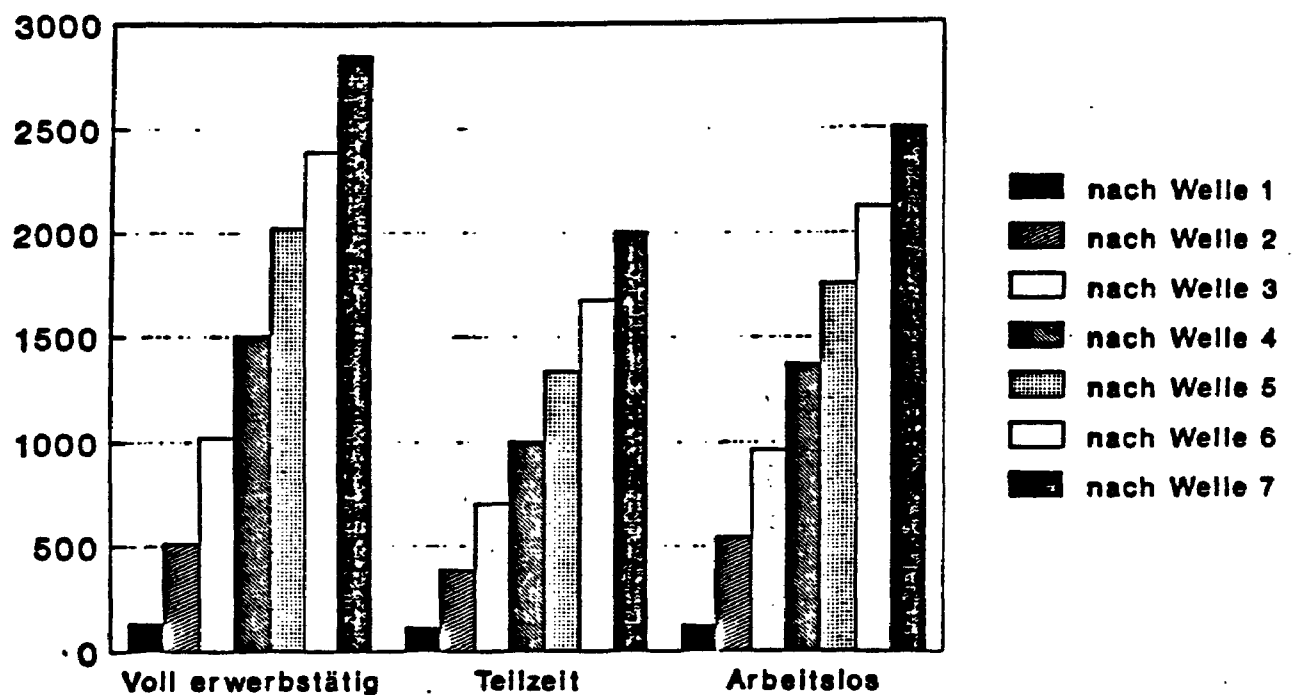
Entwicklung des Stichprobenbestands Vergleich :Befragungspersonen 1. Welle und Zugewinne bis zur 7. Welle



Datenbasis : Das Sozio-ökonomische Panel
Welle 1 bis Welle 7

Abbildung 4

Anzahl unzensierter Spells im Panel Entwicklung bis zur 7. Welle



Quelle: Das Sozio-ökonomische Panel
Welle 1 bis Welle 7

Die Interviewereindrücke sowie schriftlichen Reaktionen der Befragungspersonen bestätigen die gegenwärtigen Schwierigkeiten, empirische Studien zu sozio-ökonomischen Themen durchzuführen. Insgesamt scheint sich der stattgefundenen Stimmungsumschwung jedoch nicht grundlegend negativ auf die Stichprobenausschöpfung ausgewirkt zu haben. Für die 'alten' Haushalte an der alten Adresse gehen die gegenwärtigen Hochrechnungen Infratests von einer Mortalitätsquote von maximal 10 % aus. Damit wäre das Ergebnis des Jahres 1985 für die 2. Welle-West sogar leicht verbessert.

Schwieriger gestaltet sich in den neuen Bundesländern die Bearbeitung neuer Haushalte, bei denen in der Regel auch eine Adressenrecherche notwendig ist. Hier sind die Bearbeitungszeiten wesentlich länger als in den alten Bundesländern. Dabei kann vermutlich das West-Ergebnis von 63 % Ausschöpfung 'neuer' Haushalte - zumindest in der vorgesehenen Feldzeit bis Anfang Juli - nicht erreicht werden. Gegenwärtig wird mit besonderen Anstrengungen versucht, dieses Problem der schwierigen Zusammenarbeit mit Meldestellen verstärkt anzugehen und notfalls die Bearbeitung auch nach dem offiziellen Termin des Feldendes fortzusetzen.

Insgesamt muß freilich das Ergebnis der Feldarbeit für Stichprobe C als außerordentlich positiv bewertet werden, da mit hoher Wahrscheinlichkeit insgesamt die Stichprobenausschöpfung der zweiten Welle im Osten mindestens genauso gut, wenn nicht sogar noch leicht besser sein wird als in Stichprobe A. Nach zwei Wellen ist die kumulative Teilnahmebereitschaft in der Stichprobe C damit um 10 %-Punkte höher als dies in Stichprobe A der Fall war.

4.3 Interviewer-Stichproben

Zur Entwicklung der Interviewer-Stichproben kann nur etwas über die der Interviewer für die Stichproben A und B ausgesagt werden; für den Interviewerstab der Stichprobe C liegen noch keine Längsschnitt-Ergebnisse vor.

Aufgrund des in Abschnitt 2.3.3.4.2 geschilderten Verfahrens des Interviewer-Einsatzes hat sich die Zahl der eingesetzten Interviewer im Laufe der SOEP-Laufzeit drastisch vermindert.

Eine kleine Zahl von Interviewern wird pro Welle neu in das Projekt eingearbeitet. Die Erfolgsquoten dieser neuen Interviewer sind in der Regel etwas niedriger als bei den schon länger im Projekt mitarbeitenden Interviewern.

Ursprünglich wurden 631 Interviewer eingesetzt. In Welle 7 sind es nur noch 280 Interviewer. Die durchschnittliche Zahl der Haushalts-Interviews pro Interviewer stieg von 9 auf 17 an.

Nach wie vor wird der ganz überwiegende Teil der Haushalte über Interviewer vor Ort betreut und nach wie vor ist das mündliche Interview die dominierende Erhebungsmethode. Der Anteil der alternativen Interviewmethoden steigt jedoch kontinuierlich, wenn auch geringfügig, von Welle zu Welle an (vgl. Tabelle 5 oben).

Die Kontrolle von Interviewern durch die erneute Befragung eines Haushalts hat auch eine Schätzung für die Untergrenze von Betrugsfällen bei professionellen Interviewerstäben ergeben. Mit den Ergebnissen der zweiten Welle stellte sich heraus, daß 49 Haushalts-Interviews in der ersten Welle offensichtlich gefälscht sein mußten. Daran waren 4 Interviewer beteiligt, d.h. 0,6 % aller Interviewer. Man kann vermuten, daß bei üblichen Querschnittsbefragungen der Anteil gefälschter Interviews größer ist, da es sich um weniger qualifizierte Interviewer handelt, die zudem keine Kontrolle durch eine 100%ige Wiederbefragung befürchten müssen.

5 Gewichtung und Hochrechnung

Ziel einer jeden Stichprobe ist der Rückschluß von der Erhebung auf die dadurch "abgebildete" Grundgesamtheit. Um auf die Fallzahlen der Grundgesamtheit schließen zu können, ist eine Hochrechnung der Stichprobenfälle notwendig. Wenn es sich um keine einfache (ungeschichtete) Zufallsstichprobe handelt, ist auch eine Gewichtung (bzw. Umgewichtung) der Stichprobendaten notwendig, um die Strukturen der Grundgesamtheit nachzubilden. Auch diese Gewichtung ist beim SOEP notwendig.

Für die Randanpassung der ersten Welle der Stichprobe A und B wurden die im linken Block von Übersicht 1 dargestellten Merkmalskombinationen gewählt, die insgesamt 316 Restriktionen an die Hochrechnungsergebnisse festlegen. Für die Randanpassung der Stichprobe C konnten nur personenbezogene und regionale Restriktionen vorgegeben werden, da für das Jahr 1990 für die DDR keine vertrauenswürdigen Haushaltsstrukturdaten vorlagen. Durch die Hochrechnung der mit Hilfe der im rechten Block in Übersicht 1 dargestellten 115 Merkmalskombinationen ergab sich auch eine plausible Haushaltsstruktur (vgl. Pischner 1991).

Übersicht 1

Hochrechnungsrahmen des SOEP

Stichproben A und B (1984)		Stichprobe C (1990)	
Privathaushalte			
Geschlecht des HV			-
Alter des HV			-
Haushaltsgröße			-
Nationalität			-
Wohnbevölkerung in Privathaushalten			
Geschlecht		Geschlecht	
Alter		Alter	
Familienstand		Familienstand	
Nationalität			
Schüler			
Geschlecht			-
Schultyp			-
Nationalität des HV			
Erwerbstätige in Privathaushalten Stichprobe A:			
Selbständige und Mithelfende nach:			
Geschlecht			-
Alter			
Landwirtschaftlichem Beruf (ISCO = 6)/Andere			
Abhängig Beschäftigte:			
Geschlecht			-
Alter			
Beruf (Hauptgruppe ISCO)			
Erwerbstätige in Privathaushalten Stichprobe B:			
Nationalität			-
Geschlecht			-
Alter			-
Regionale Verteilung			
-		Bezirke der DDR	

Eine Panelerhebung unterscheidet sich von einer Folge von unverbundenen Querschnittsbefragungen dadurch, daß die vorhergehende Panelwelle Ausgangsbestand für die nachfolgende Befragungswelle ist. Verluste treten in der nachfolgenden Welle auf, wenn im Falle eines Haushalts-Panels die betreffenden Personen verstorben sind bzw. aus dem Erhebungsgebiet verzogen sind (Demographische Verluste) oder wenn die Personen in der nachfolgenden Befragungswelle nicht mehr zu einer weiteren Teilnahme bereit sind (Panelmortalität). Gewinne treten für die Stichprobe auf, wenn neue Personen in Panel-Haushalte ziehen bzw. wenn - wie das beim SOEP der Fall ist - Kinder aus der Panelpopulation das Befragungsalter erreichen.

Es ist unmittelbar einleuchtend, daß derartige Besonderheiten der Erhebung bei der Gewichtung und Hochrechnung berücksichtigt werden müssen (vgl. auch Ernst 1989). D.h., die einzelnen Wellen einer Panel-Erhebung sollten nicht wie eine Folge unverbundener Querschnitte hochgerechnet werden. Auch liefert eine solche Herangehensweise keine Antwort auf die Frage, wie Längsschnitte hochgerechnet werden können.

Einzelheiten dieser Vorgehensweise und empirische Erfahrungen mit diesem Ansatz findet man bei Rendtel (1991b).

Allgemein läßt sich das Ziel einer Hochrechnung darin beschreiben, auf Basis der Stichprobenerhebung auf das Vorkommen interessierender Merkmalskombinationen in der Grundgesamtheit zu schließen. Die Hochrechnung über die inversen Auswahlwahrscheinlichkeiten basiert auf dem Randomisierungs-Ansatz. Bei dem Randomisierungs-Ansatz werden die Merkmalsausprägungen Y_i der einzelnen Einheiten der Population als feste Größen betrachtet. Zufällig ist lediglich die Auswahl der einzelnen Einheiten i . Die Zufallsgröße C_i gibt hierbei an, ob die Einheit i der Stichprobe angehört ($C_i = 1$) oder nicht ($C_i = 0$). Will man nun Y durch einen linearen Schätzer $\hat{Y}$ der Form

$$\hat{Y} = \sum_{i=1}^N \alpha_i C_i Y_i$$

bestimmen, so liefert die Forderung nach Erwartungstreue von $\hat{Y}$:

$$\alpha_i = 1/P \quad (C_i = 1).$$

Man erhält damit:

$$\hat{Y} = \sum_{i=1}^N \frac{1}{P(C_i=1)} C_i Y_i = \sum_{i=1}^n \frac{1}{P(C_i=1)} Y_i$$

Dieses Hochrechnungsverfahren wurde von Galler (1986) für die Hochrechnung von Paneldaten vorgeschlagen. Der Auswahlvorgang durch eine Panelerhebung läßt sich als ein mehrstufiger Prozeß beschreiben. Die Auswahl eines Längsschnitts über T Panelwellen wird als Auswahlprozeß mit 2T Stufen betrachtet, die folgendermaßen charakterisiert sind (um die Notation zu vereinfachen, wurde auf den Index i für die Stichprobenelemente verzichtet):

1. Stufe: Design-Auswahl (Anlage der Stichprobe) $P(D=1)$
2. Stufe: Antwort in der ersten Welle $P(R_1=1/D)$
3. Stufe: Erfolgreiche Kontaktaufnahme in der zweiten Welle
 $P(K_2=1/D, R_1)$
4. Stufe: Antwortgewährung in der zweiten Welle
 $P(R_2=1/D, R_1, K_2)$
- .
- .
- .
- 2T.Stufe: Antwortgewährung in der T-ten Welle
 $P(R_T=1/D, R_1, K_2, \dots K_T)$

Die Auswahlwahrscheinlichkeit $P(C=1)$ für den gesamten Auswahlprozeß über alle Einzelstufen ist dann durch das Produkt der Einzelwahrscheinlichkeiten gegeben:

$$\begin{aligned}
P(C=1) &= P(D=1, R_1=1, K_2=1, \dots, R_T=1) \\
&= P(D=1) \cdot P(R_1=1/D) \\
&\quad \cdot P(K_2=1/D, R_1) \\
&\quad \cdot P(R_1=1/D, R_1, K_2) \\
&\quad \cdot \\
&\quad \cdot \\
&\quad \cdot P(R_T=1/D, R_1, K_2, \dots, K_T)
\end{aligned}$$

Das Problem, Längsschnitte hochzurechnen, reduziert sich damit auf die Berechnung der Startwahrscheinlichkeit, an der ersten Welle des Panels teilzunehmen und auf die Bestimmung der Bleibewahrscheinlichkeiten, d.h. der bedingten Wahrscheinlichkeiten, auf der jeweiligen Auswahlstufen im Panel zu verbleiben.

Für Querschnittsauswertungen auf hochgerechneter Basis bedeutet die Berücksichtigung der Nicht-Stammpersonen kein Problem, solange die Ziehungswahrscheinlichkeiten für Haushalte in dem betreffenden Jahr bekannt bzw. abschätzbar sind. Dies stellt auch bei dem seit 1990 vereinfachten Weiterverfolgungskonzept, das auch Nicht-Stammpersonen weiterhin in das SOEP einbezieht, kein wesentliches Problem dar (vgl. Huang 1984). Im Gegenteil: das PSID-Konzept, nur Stammpersonen weiterzuverfolgen und Nicht-Stammpersonen bei der Hochrechnung unberücksichtigt zu lassen (d.h. diese Personen erhalten den Hochrechnungsfaktor 0) führt auf das Problem, daß die Hochrechnungsergebnisse zwischen Haushalts- und Personenebene nicht mehr konsistent miteinander sind¹⁸. An dieser Stelle zeigt sich, daß das demographisch orientierte Konzept der Stammpersonen über die Zeit nicht identisch ist mit dem Konzept der Bevölkerung in Privathaushalten. Diese Identität gilt streng genommen nur für die Startwelle des Panels¹⁹, danach verändern sich die zur Herstellung der Querschnittsrepräsentativität notwendigen Gewichtungsfaktoren in Abhängigkeit von der Entwicklung des Haushalts.

Haushalte mit Zuzügen haben eine erhöhte Auswahlchance; dies gilt auch für das ursprüngliche Weiterverfolgungskonzept, das nur Stammpersonen einbezog. Da sich die Auswahlchan-

cen von Haushalten allein über die Ziehungschancen seiner Mitglieder bei Start des Panels sowie die Weiterverfolgungsregeln ergeben, haben Haushalte mit Zuzügen eine höhere Auswahlchance als Haushalte ohne Zuzüge (da sie mindestens über 2 Pfade zu erreichen waren). Dies führt in der Konsequenz dazu, daß Haushalte mit Zuzügen ein geringeres Gewicht erhalten müssen. Ordnet man allen Personen (d.h. auch den Nicht-Stammpersonen) des Haushalts das Haushaltsgewicht zu, so kompensiert dieses geringere Gewicht den Zuwachs der Fallzahlen durch die Zuzüge. Im Unterschied zu den vereinfachten Weiterverfolgungsregeln impliziert die Weiterverfolgung nach dem Stammpersonen-Konzept eine Fülle von sehr unübersichtlichen Exklusionsbedingungen für einzelne "Pfade" der Personen zwischen den Haushalten, die für die Hochrechnung sehr schwer zu handhaben sind. Letztendlich selektiert eine am Stammpersonenkonzept orientierte Weiterverfolgung Personen mit einer hohen Mobilität aus der Panelstichprobe vollständig heraus, die eigentlich nur ein geringeres Gewicht als immobile Personen bekommen müssen.

Faßt man alle Ausführungen zur Hochrechnung aus der praktischen Sicht dessen zusammen, der die SOEP-Daten auswerten will, so heißt dies, daß für jeden Haushalt im Datensatz pro Welle je ein Querschnitts- und Längsschnitthochrechnungsfaktor zur Verfügung gestellt wird. Der Querschnittsfaktor liefert für die Gesamtstichprobe die für das Erhebungsjahr gültigen Werte in der Grundgesamtheit. Der Längsschnittfaktor gibt an, wie sich aufgrund der Paneldynamik die Auswahlwahrscheinlichkeit der einzelnen Haushalte ändert. Will man für Längsschnittanalysen wissen, wie sich über mehrere Wellen hinweg die richtige Gewichtung berechnet, sind die Längsschnittfaktoren miteinander zu multiplizieren.

6 Datenaufbereitung

Die konsistente, d.h. über Jahre hinweg nach gleichen Regeln erfolgende Datenaufbereitung stellt eine zentrale Aufgabe für ein Panel-Projekt dar. Der mehrstufige Prozeß der Arbeiten an den SOEP-Daten wird im folgenden beschrieben.

6.1 Ablauf der Datenerfassung, -prüfung und -bereinigung

Die Datenerfassung, -prüfung und -bereinigung ist ein mehrstufiger Prozeß. Folgende Arbeitsschritte sind zu nennen, die vom Umfrageinstitut durchgeführt werden:

- **Manuelle Datenprüfung ("editing"):** Bei einer sog. Basisprüfung geht es um die Vollständigkeit und die richtige Zuordnung aller Erhebungsunterlagen. Das Adressenprotokoll wird auf formale Richtigkeit und Übereinstimmung mit der Haushaltsstruktur laut Haushaltsfragebogen geprüft; die auf dem Adressenprotokoll eingetragenen Interviewerkontakte, die auf Datenträger erfaßt werden, werden markiert. Auf der zweiten Stufe erfolgt die Fragebogenprüfung. Dieses sog. editing hat zum Ziel, vor allem solche Fehlerquellen zu beseitigen, die nach der Datenerfassung auf EDV von dem maschinellen Prüfprogramm nicht oder nur schwer als Fehler erkennbar sind.
- **Automatische Datenprüfung:** hier ist in eine Basisprüfung und die Fragebogenprüfung zu unterscheiden. In der Basisprüfung werden die verschiedenen Teile des Datenbestands pro Haushalt zusammengeführt und durch ein maschinelles Prüfprogramm auf Vollständigkeit und Konsistenz geprüft. Anschließend erfolgt die in Abschnitt 2.3.6.2 beschriebene Fehlerbereinigung.
- **Erfassung und Prüfung der Adressenprotokolle von Ausfällen.**
- **Schlußprüfung nach Integration der Datensätze.**

6.2 Fehlerbereinigung und Dokumentation

Ein speziell entwickeltes Prüfprogramm nimmt keine maschinellen Korrekturen vor, sondern sucht unzulässige Werte und druckt das Ergebnis der Fehlersuche aus. Speziell geschulte Prüfer in der Editing-Gruppe des Umfrageinstituts rufen die aufgelisteten Fälle am Bildschirm auf und korrigieren, nach den jeweils mit der Projektgruppe festgelegten Bearbeitungsregeln direkt im Datensatz.

Ist die Angabe eindeutig unplausibel, wird sie seit 1990 gelöscht und durch ein spezielles Missing-Value-Zeichen an die für K.A. vorgesehene Stelle gesetzt. Angaben, die formal fehlerhaft sind, aber im Kontext des Einzelfalles dann doch plausibel erscheinen, bleiben unverändert. Diese Form der Fehler-Korrektur wurde auch rückwirkend für alle Wellen durchgeführt. In einem speziellen Datensatz (DOKU) sind die Originalwerte enthalten. Jeder SOEP-Auswerter kann auch auf diese Werte zugreifen.

6.3 Datenstrukturen

6.3.1 Konzept der logischen Datenaufbereitung

Bei der Erhebung des SOEPs fallen eine große Zahl von Daten an, die unterschiedlichen logischen Ebenen zuzuordnen sind (z.B. die bereinigten Daten und die Original-Werte). Eine Besonderheit des SOEPs besteht darin, daß alle erhobenen und erzeugten Daten, soweit es keine datenschutzrechtlichen Hemmnisse gibt, auch an externe Nutzer in anonymisierter Form weitergegeben werden.

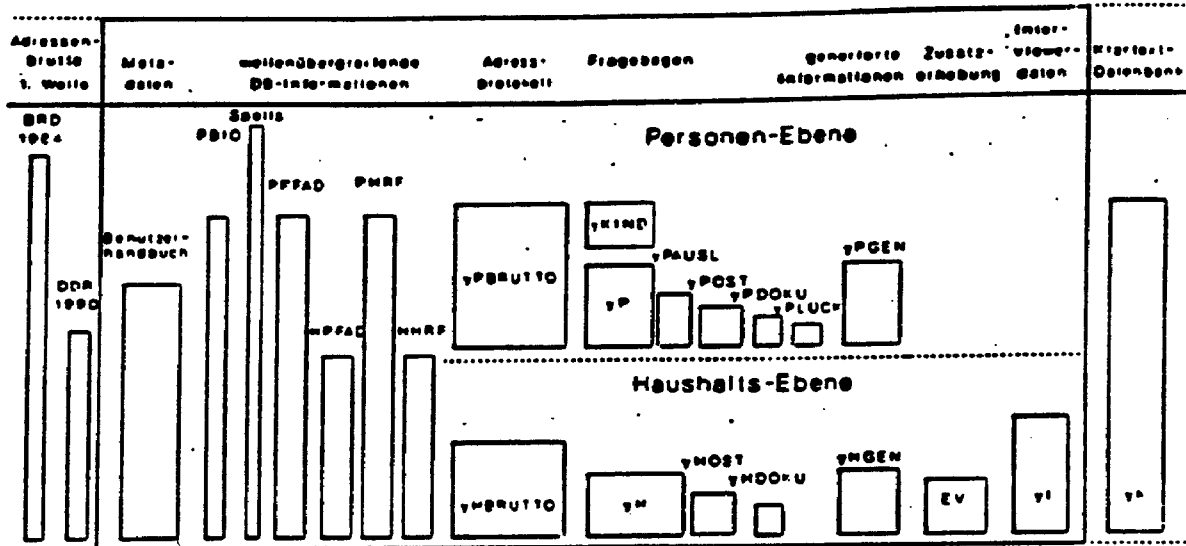
Abbildung 5 zeigt alle maschinenlesbaren Datentypen, die für das SOEP existieren. Die datenschutzrechtlich weitergabefähigen Daten sind im inneren Rechteck der Abbildung enthalten. Die rechts angeordneten Datentypen für Klartextangaben und kleinräumige Regionalinformationen werden für die meisten Auswertungen nicht benötigt. Sie sind im wahrsten Sinne des Wortes von peripherer Bedeutung. Es handelt sich um datenschutzrechtlich äußerst sensitive Daten, die nur innerhalb des DIWs in Zusammenarbeit mit der Panel-Gruppe ausgewertet werden können.

Hier etwa Abb. 5

Die Daten werden mit einem Datenbanksystem verwaltet, welches die Verknüpfung von Daten, die zu unterschiedlichen hierarchischen Ebenen gehören, ebenso erleichtert wie die Verknüpfung über der Zeit²⁰ (vgl. auch den Beitrag von Müller und Worpitz in diesem

Abbildung 5

Die Tabellen der SOEP-Datenbank



Die Stichprobe C (DDR bzw. Ostdeutschland) ist völlig in die Datenbank integriert. Bis 1989 werden die Fälle dieser Stichprobe in den wellenübergreifenden DB-Informationen (PPFAD und HRFAD) als "noch nicht befragt" gekennzeichnet. Die West-Stichprobe ist aus Datenschutzgründen nicht in die eigentliche SOEP-Datenbank integriert. Auswertungen der West-Stichprobe sind nur im Rahmen besonderer Vereinbarungen im DiW möglich. Weitergegeben werden zum Stand 1.1.1991 die stark umrandeten Teile der Datenbank.

Legende:

y:	Laufindex für Welle A (=1984) bis z.Z. G (=1990)
PBIO:	Personenbezogene Biographieinformation
Spells:	Informationen aus den Aktivitäts- und Erwerbskalendarien (monatlich)
PPFAD, HPPAD:	Personenbezogene bzw. haushaltsbezogene Teilnahmeinformationen
PHRF, HHRF:	Personenbezogene bzw. haushaltsbezogene Hochrechnungsfaktoren
yPLÜCKE:	Nacherhobene Personeninformationen bei TeilnahmeLücken
yPBRUTTO, yHBRUTTO:	Personen- und Haushaltsbezogene Informationen aus dem Adressprotokoll
yH:	Haushaltsbezogene Standard-Informationen aus dem Haushaltsfragebogen
yP, yKIND:	Personenbezogene Standardinformationen aus den Personenfragebogen (yP) bzw. aus dem Haushaltsfragebogen (yKIND)
yPAUSL:	Personenbezogene Zusatzinformationen für Ausländer der Stichprobe B
yPOST:	Personenbezogene Zusatzinformationen für die Ost-Stichprobe C
yPDOKU, yHDOKU:	Personen- bzw. haushaltsbezogene Dokumentationsinformationen über bereinigte Variablen
yGEN, yHGEN:	Personen- und haushaltsbezogene generierte Variable
GHOST:	Haushaltsinformationen für die Basisbefragung in der DDR (y=G, für das Jahr 1990)
EV:	Vermögensbilanz der Haushalte in Welle 5 (y=E, für das Jahr 1988)
yI:	Interviewerinformationen
yK:	Personenbezogene Klerschriftangaben aus dem Fragebogen

Band). In den Sozialwissenschaften ist die Haltung von Daten unter Datenbanksystemen umstritten, da in diesem Falle für die Auswertungen der Daten eine zusätzliche Humankapital-Investition notwendig ist, um die Handhabung des entsprechenden Datenbanksystems zu erlernen. Aber nur unter einem Datenbanksystem sind flexible Auswertungen, die den vollen Informationsgehalt von Haushalts-Panels nutzen, möglich (vgl. dazu auch Engel et al. 1991).

6.3.2 Logische Struktur der Variablen

Bei der Aufbereitung der Befragungsdaten selbst entsteht das Entscheidungsproblem, ob der Fragebogen in einer Eins-zu-Eins-Nachbildung in der Datenbank abgebildet werden soll, oder ob eine logische Struktur gewählt wird, die den typischen Benutzerwünschen folgt, anstatt der Struktur, die den Befragten entgegenkam. Ursprünglich wurde die Datenbank als Eins-zu-Eins-Abbildung der Befragungsinstrumente konzipiert, jedoch wurde dieses Konzept aufgrund von vielfältigen Analyseerfahrungen modifiziert:

Ein grundsätzliches Problem bei der Aufbereitung von Panel-Daten besteht darin, daß die Erhebungsinstrumente psychologischen Überlegungen folgen, die das Interview möglichst leicht machen. Dadurch wird die Reihenfolge der Fragen bestimmt und bestimmte Fragen müssen bei Wiederholungsbefragungen nicht in jedem Jahr gestellt werden. Ist z.B. einmal der Schulabschluß ermittelt, muß er nur noch erfragt werden, wenn eine Filterfrage mit Ja beantwortet wird, ob eine Änderung des Schulabschlusses im letzten Jahr erfolgt ist. Eine Eins-zu-Eins-Abbildung des Fragebogens bedeutet dann, daß für viele Personen der Schulabschluß in den Daten der laufenden Welle nicht enthalten ist. Er muß erst aus Daten vergangener Wellen rekonstruiert werden. Dies ist für die meisten Nutzer der Daten eindeutig sub-optimal. Es gilt deswegen, generierte Variablen mit aufbereiteten Informationen zur Verfügung zu stellen, insbesondere sogenannte Statusvariablen, die z.B. den Schulabschluß in jeder Welle enthalten, obwohl er nicht für jeden Befragten erfragt wurde.

Die Logik des Datenbankaufbaus folgt also eher querschnittsbezogenen Regeln. Für längsschnittbezogene Analysen werden dem Nutzer freilich eine Fülle an Variablen zur Verfügung gestellt, um personen- bzw. haushaltsidentische Verknüpfungen über mehrere Wellen zu vereinfachen. Hierzu wurden 'Längsschnittwegweiser-Relationen' gebildet (PPFAD und HPFAD), welche für jede jemals im Verlauf der Studie aufgetretene Person bzw. Haushalt einen Datensatz vorsieht, dessen Attribute (Variablen) Auskunft über zugehörige logische Satzschlüssel in den Querschnittstabellen der Datenbank liefern.

Die Erfahrungen mit dem SOEP haben gezeigt, daß die Generierung derartiger Variablen für die Nutzer wichtiger ist als immer feinere "Fehlerbereinigungen", die ohnehin nie zu einem Ende kommen können. In komplexen Datensätzen wird man immer wieder Inkonsistenzen feststellen, wenn eine neue Fragestellung an den Datensatz herangetragen wird. Letztlich müssen die Nutzer die notwendigen Bereinigungsarbeiten selbst vornehmen²¹.

6.4 Datenweitergabe und Datenschutz

Die Daten des Sozio-ökonomischen Panels werden für wissenschaftliche Zwecke kostenlos weitergegeben. Das Interesse an den Daten hat in der Vergangenheit ständig zugenommen. Bis zum Juli 1991 wurden bereits über 100 Datenweitergabeverträge geschlossen, wobei hinter einem Vertrag häufig mehrere Wissenschaftler stehen, die mit den Daten arbeiten. Hervorgehoben sei, daß auch im Ausland reges Interesse an den Paneldaten besteht.

Da in vielen Staaten weniger strenge Datenschutzbestimmungen gelten als in der Bundesrepublik Deutschland, wurde eine besonders stark anonymisierte Version des Datensatzes entwickelt, die an Forscher in jedem Staat dieser Welt weitergegeben werden kann²².

Eine Wiederholungsbefragung, die auf die permanente Teilnahmebereitschaft der Befragten angewiesen ist, muß zur Sicherung des Vertrauensverhältnisses noch mehr tun, als hierzu gesetzlich vom Datenschutz vorgeschrieben wird. Der Aufwand für die Anonymisierung der

Daten ist beträchtlich, so dürfen als sensibel angesehene Variablen nicht den Nutzern zugänglich gemacht werden.

Im "Public-Use-File" des SOEP wurden einige wenige Variablen gelöscht, die aber auch für die meisten Analysen ohnehin keine große Bedeutung haben. Es handelt sich um die genaue Nationalität von Ausländern in der Stichprobe (man weiß also nur, daß es sich um Nicht-Deutsche handelt), die regionale Zugehörigkeit des Stichprobenhaushalts und um detaillierte Angaben zum Vermögen (grobe Angaben stehen freilich zur Verfügung).

Außerdem wird nur eine 95 %-Stichprobe des SOEP-Samples für public use weitergegeben²³. Dies beruht auf der Überlegung, daß es normalerweise einfach ist, einen Befragten zu de-anonymisieren, wenn man weiß, daß er in einer Stichprobe enthalten ist. Werden aber einige Fälle der Stichprobe entfernt²⁴, kann man nicht mehr sicher sein, daß die betreffende Person in diesem Sub-Sample ist und der Grad der Anonymisierung wird dadurch größer.

7 Perspektiven

Die Daten des SOEPs werden inzwischen von über 100 Nutzergruppen ausgewertet (Stand Juli 1991). Für einen Überblick über inhaltliche Längsschnittergebnisse vgl. neben der Bibliographie des vorliegenden Buches insbesondere Krupp und Schupp (1988), Wagner (1990), Rendtel und Wagner (1991) sowie Hujer et al. (1991). Für erste Ergebnisse der Ost-Stichprobe C vgl. Projektgruppe Panel (1991a).

Die Öffnung der innerdeutschen Grenze und der Beitritt der DDR zur BRD waren natürlich für eine großangelegte Haushaltsstichprobe eine Herausforderung. Zumal dann, wenn diese Untersuchung bei den Befragten unter dem Titel "Leben in Deutschland" eingeführt wurde. Eine Wiederholungsbefragung ist natürlich zur empirischen Erfassung von starken Veränderungen in der DDR und deren Rückwirkungen auf die BRD ideal geeignet.

Die Ausweitung des SOEPs auf das Gebiet der DDR gelang bereits im Juni 1990. Die Instrumente des SOEPs bewährten sich auch bei dieser schwierigen Aufgabe. Das Sozio-ökonomische Panel liefert somit innerhalb des bevölkerungs- und erwerbsstatistischen Berichtssystems der gesamten Bundesrepublik Deutschland seit 1984 einen wichtigen Baustein.

Wie die über 25jährige Erfahrung aus den Vereinigten Staaten zeigt, schließt die mikroanalytische Dauerbeobachtung gesellschaftlichen Wandels zentrale Lücken der politikberatenden Sozialberichterstattung und erweist sich als notwendige Infrastruktureinrichtung für die empirisch fundierte Theoriebildung der Sozial- und Wirtschaftswissenschaft.

Im europäischen Kontext stellt das Sozio-ökonomische Panel eine für international vergleichende Panel-Analyse zentrale nationale Längsschnittdatenbasis dar, deren methodisches wie inhaltliches Konzept für neu entstandene nationale Panel-Studien teilweise als Vorbild diente (z.B. für die British Household Panel Study).

Eine besondere Herausforderung für jede Panel-Studie stellt die Integration von Zuwanderern dar. Da nichts dafür spricht, daß die Migration nach Mitteleuropa in den nächsten Jahren und Jahrzehnten geringer werden wird, wird für das SOEP an einem Konzept für "Zuwanderer-Ergänzungen" gearbeitet.

Literatur

- Becker, Irene 1989: Die Datenbestände des Sfb 3, Sfb 3 Arbeitspapier Nr. 317, Frankfurt, Mannheim
- Beckett, Sean et al. 1988: The Panel Study of Income Dynamics after Fourteen Years - An Evaluation, in: *Journal of Labor Economics*, 6(4), S. 472-492
- Brecht, Beatrix 1990: Aufbau, Struktur und Anwendung des Sozio-ökonomischen Panels in INGRES, Diskussionsbeiträge des Sfb 178, Serie II-Nr. 120, Konstanz
- David, Martin 1989: Managing Panel Data for Scientific Analysis - The Role of Relational Data Base Management Systems, in: D. Kasprzyk, S. 226-241
- Ernst, Lawrence R. 1989: Weighting Issues for Longitudinal Household and Family Estimates, in: D. Kasprzyk, S. 139-159
- Esser, Hartmut u.a. 1989: Mikrozensus im Wandel. Band 11 der Schriftenreihe Forum der Bundesstatistik, herausgegeben vom Statistischen Bundesamt, Stuttgart
- Frick, Joachim 1990: Die SIR-Datenbank des sozio-ökonomischen Panels - Ein Tutorial zu Aufbau, Syntax und problemorientierten Anwendungen, Berlin
- Galler, Heinz P. 1986: Ein Gewichtungskonzept für das Sozio-ökonomische Panel, Sfb 3 Arbeitspapier Nr. 204, Frankfurt, Mannheim
- Geiss, Alfons J. und Jürgen H.P. Hoffmeyer-Zlotnik 1991: Zur Vercodung von Beruf, Branche und Prestige in der DDR, in: Projektgruppe Panel (Hg.): *Lebenslagen im Wandel - Basisdaten und -analysen zur Entwicklung in Ostdeutschland*, Frankfurt, New York, S. 139-147
- Grohmann, Heinz 1985: Vom theoretischen Konstrukt zum statistischen Begriff - Das Adäquationsproblem, in: *Allgemeines Statistisches Archiv*, 69 (1), S. 1-15.
- Hanefeld, Ute 1987: *Das Sozio-ökonomische Panel - Grundlagen und Konzeption*, Frankfurt, New York
- Hauser, Richard 1991: Einleitung, in: R. Hauser et al. (Hg.), *Mikroanalytische Grundlagen der Gesellschaftspolitik - Ausgewählte Probleme und Lösungsansätze*, Weinheim
- Hippler, Hans-J. et al. 1990: Der Einfluß von Datenschutzzusagen auf die Teilnahmebereitschaft an Umfragen, in: *ZUMA-Nachrichten*, 27, S. 54-67
- Huang, H. 1984: Obtaining Cross-Sectional Estimates from a Longitudinal Survey. Experiences of the Income Survey Development Program, *Proceedings of the Section of Survey Research Methods, American Statistical Association*, p. 676-681
- Hujer, Reinhard et al. (Hg.) 1991: *Herausforderungen an den Wohlfahrtsstaat im strukturellen Wandel*, Frankfurt, New York
- Kalton, G. 1989: Modelling Considerations: Discussion from a Survey Sampling Perspective, in: D. Kasprzyk et al., S. 575-585
- Kasprzyk, Daniel et al. (Hg.) 1989: *Panel Surveys*, New York u.a.
- Kirschner, H.P. 1984: Allbus 1980 - Stichprobenplan und Gewichtung, in: K.U. Mayer und P. Schmidt (Hg.), *Allgemeine Bevölkerungsumfragen der Sozialwissenschaften - Beiträge zu methodischen Problemen des Allbus 1980*, Frankfurt, S. 114-182
- Krause, Peter und Gert Wagner 1991: Datenhaltung bei sozialwissenschaftlichen Panel-Studien, in: B. Engel et al., *Datenbankorganisatorische Probleme und Grundlagen des NIFA-Panels - Ergebnisse eines Workshops, Arbeitspapier des Sfb 187 "Neue Informationstechnologie und flexible Arbeitssysteme"*, Bochum
- Krupp, Hans-Jürgen und Wolfgang Glatzer 1978: *Umverteilung im Sozialstaat - Empirische Analysen für die Bundesrepublik*, Frankfurt, New York
- Krupp, Hans-Jürgen und Ute Hanefeld 1987: *Lebenslagen im Wandel - Analysen 1987*, Frankfurt, New York
- Krupp, Hans-Jürgen und Jürgen Schupp 1988: *Lebenslagen im Wandel - Daten 1988*, Frankfurt, New York
- Kruskal, William und Frederik Mosteller 1979: Representative Sampling, I-III, in: *International Statistical Review*, 47, S. 13-24, S. 111-127 und S. 245-265
- Licht, Georg und Viktor Steiner 1991: Stichprobenselektion, unbeobachtete Heterogenität und Humankapitaleffekte bei der Schätzung von Lohnfunktionen mit Paneldaten, in: U. Rendtel und G. Wagner, S. 100-134
- Löwenbein, Oded und Ulrich Rendtel 1991: Selektivität und Panelanalyse, in: U. Rendtel und G. Wagner, S. 156-187
- Mayer, Karl Ulrich 1990: Lebensverläufe und sozialer Wandel. Anmerkungen zu einem Forschungsprogramm, in: Mayer, K.U. (Hrsg.), *Lebensverläufe und sozialer Wandel, Sonderheft 31, Kölner Zeitschrift für Soziologie und Sozialpsychologie*, S. 7-21

- Ott, Notburga 1991: Intrafamily Bargaining and Household Decisions, Berlin u.a.
- Pietzke, Iwan-Rainer 1991: Anwendung der Informations- und Kommunikationstechnik in der Kommunalverwaltung - heute und in der Zukunft, in: Verwaltungsorganisation, 25 (1), S. 15-20.
- Pischner, Rainer 1991: Die Gewichtung der Ost-Stichproben des Sozio-ökonomischen Panels und der Pilotstichprobe des Wohlfahrtssurveys, in: Projektgruppe Panel (Hg.): Lebenslagen im Wandel - Basisdaten und -analysen zur Entwicklung in Ostdeutschland, Frankfurt, New York, S. 97-112
- Pol, Frank van de 1989: Issues of Design and Analysis of Panels, Amsterdam
- Projektgruppe Panel (Hg.) 1991a: Lebenslagen im Wandel - Basisdaten und -analysen zur Entwicklung in den Neuen Bundesländern, Frankfurt, New York
- Projektgruppe Panel (Hg.) 1991b: Das Sozio-ökonomische Panel - Benutzerhandbuch, Version 5/91, Berlin
- Rendtel, Ulrich 1990a: Teilnahmebereitschaft in Panelstudien - Zwischen Beeinflussung, Vertrauen und sozialer Selektion, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 42(2), S. 280-299
- Rendtel, Ulrich 1991a: Über die Behandlung des Selektivitätsproblems bei der Auswertung von Paneldaten - dargestellt an zwei Fallbeispielen aus dem Sozio-ökonomischen Panel, in: Beiträge zur Arbeitsmarkt und Berufsforschung (144), S. 89
- Rendtel, Ulrich 1991b: Die Schätzung von Populations-Werten in Panel-Erhebungen, in: Allgemeines Statistisches Archiv, 75 (4).
- Rendtel, Ulrich und Gert Wagner (Hg.) 1991: Lebenslagen im Wandel - Zur Einkommensdynamik in Deutschland seit 1984, Frankfurt, New York
- Riebschläger, Marlis und Gert Wagner 1991: Interviewerstab und Interviewereffekte in der DDR-Basisbefragung des Sozio-ökonomischen Panels, in: Projektgruppe Panel (Hg.): Lebenslagen im Wandel - Basisdaten und -analysen zur Entwicklung in Ostdeutschland, Frankfurt, New York, S. 127-138
- Schmaus, Günther 1990: Organisation der Daten des Luxemburger Haushaltspanels, Document PSELL No. 17, Walferdange/Luxemburg
- Schnell, Rainer 1991: Wer ist das Volk?, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 43(1), S. xx-xx
- Schnell, Rainer, Paul B. Hill und Elke Esser 1988: Methoden der empirischen Sozialforschung, München, Wien
- Schupp, Jürgen 1985: Erfahrungen mit dem Aufbau einer SIR/DBMS-Datenbank beim Sozio-ökonomischen Panel, in: Status GmbH (Hg.), Datenbank SIR/DBMS Benutzerkonferenz, Berlin, S. 117-142.
- Schupp, Jürgen 1991: Zur Durchführung des Sozio-ökonomischen Panels, in: Herget, Hermann (Hrsg.), Chancen von Panelerhebungen und zeitbezogener Analyse für die Berufsbildungsforschung, Berichte zur beruflichen Bildung, Heft 124, Berlin und Bonn, S. 111-127
- Schupp, Jürgen und Gert Wagner 1991: Die Ost-Stichprobe des Sozio-ökonomischen Panels - Konzept und Durchführung der "SOEP-Basiserhebung 1990" in der DDR, in: Projektgruppe Panel (Hg.): Lebenslagen im Wandel - Basisdaten und -analysen zur Entwicklung in Ostdeutschland, Frankfurt, New York, S. 25-41.
- Solenberger, Peter et al. 1989: Data Base Management Approaches to Household Panel Studies, in: D. Kasprzyk, S. 190-225
- Wagner, Gert 1990: Das Sozio-ökonomische Panel - Ein Instrument zur Dauerbeobachtung privater Haushalte, in: H. Rapin (Hg.) Der private Haushalt im Spiegel sozialempirischer Erhebungen, Frankfurt, New York, S. 93-113
- Wagner, Gert 1991: Die Erhebung von Einkommensdaten im Sozio-ökonomischen Panel (SOEP), in: U. Rendtel und G. Wagner (Hg.): Lebenslagen im Wandel - Zur Einkommensdynamik in Deutschland seit 1984, Frankfurt, New York, S. 26-33
- Wagner, Gert und Jürgen Schupp 1990: Das Sozio-ökonomische Panel im sich einenden Deutschland, Sfb 3 Arbeitspapier Nr. 326, Frankfurt, Mannheim

- 1 Die Autoren bedanken sich bei den weiteren Mitgliedern der Projektgruppe 'Das Sozio-ökonomische Panel' am Deutschen Institut für Wirtschaftsforschung für kollegiale Hilfe bei der Erstellung dieses Aufsatzes: Joachim Frick, Elke Holst, Peter Krause, Rainer Pischner, Marlis Riebschläger und Johannes Schwarze. Bernhard von Rosenblatt (Infratest Sozialforschung) ist für hilfreiche Kommentare und die Überlassung von umfangreichem Material aus den Infratest-Methodenberichten über das Sozio-ökonomische Panel zu danken.
- 2 Dieser Aufsatz beantwortet insbesondere die von Hanefeld (1987) offen gelassenen Fragen der Konzeption.
- 3 Für einen Überblick über die Entwicklung von Panel-Surveys vgl. Kasprzyk et al. (1989) sowie den - allerdings weniger breiten - Überblick von Pol (1989).
- 4 Die Dokumentation aller Fragebögen umfaßt inzwischen drei DIN-A4-Ordner. Selbst ein kondensierter Überblick ist im Rahmen eines Buches nicht möglich. Vgl. deswegen das "Benutzerhandbuch" des SOEPs (Projektgruppe Panel 1991b).
- 5 Weiterhin bietet auch die als "warming down" gedachte abschließende Aufforderung im Personenfragebogen "Vielen Dank für Ihre Mitarbeit! Wenn Sie uns sonst noch etwas mitteilen wollen:" Möglichkeiten qualitativer Analysen. Das "warming down" soll an sich nur die Bereitschaft zur Wiederteilnahme im nächsten Jahr fördern. Die Vercodung dieser Mitteilung wird geprüft.
- 6 Vgl. hierzu den dreiteiligen Übersichtsartikel von Kuskal und Mosteller (1979). Am Beispiel der Deskription der Einkommensverteilung und der Erklärung von Einkommensstrukturen werden die unterschiedlichen Qualitätsanforderungen an Daten am Beispiel des Mikrozensus von Esser et al. (1989) sehr gut dargestellt.
- 7 "Gut" meint genaugenommen im Hinblick auf spezifische Fragestellungen. Je nach Problem und Frage wird die Bewertung unterschiedlich ausfallen (vgl. die Diskussion des "Adäquationsproblems" durch Grohmann 1985, S. 11ff.). Zur Schätzung von Aggregaten ist die möglichst vollständige systematische Erfassung ("Coverage") der relevanten Grundgesamtheit ("target population") sicherlich wichtiger als z.B. eine ausgewogene regionale Repräsentation der Stichprobe. Diese kann allerdings für Verteilungsanalysen wichtig sein, welche aber u.U. keine hundertprozentige Coverage erfordert.
- 8 "ADM" ist die "Arbeitsgemeinschaft Deutscher Marktforschungsinstitute"; für das Konzept des ADM-Mastersamples vgl. Kirschner (1984) und Schnell (1991).
- 9 Vgl. dazu und zum "ADM-Mastersample" Kirschner (1984) und Schnell (1991).
- 10 Der Arbeitskreis deutscher Marktforschungsinstitute (ADM) erstellte erst im Anschluß an die erste gesamtdeutsche Wahl im Dezember 1990 ein Ziehungsband für die neuen Bundesländer auf Stimmbezirkbasis.
- 11 Aus Zeitgründen mußte auf einen getrennten Adressenvorlauf verzichtet werden.
- 12 Dieser Abstand von drei Adressen ist der übliche Abstand beim Random-Route-Verfahren der ADM-Institute. Bei den West-Stichproben A und B wurde ein größerer Abstand gebildet (7 Haushalte; vgl. Abschnitt 2.3.2.1.1 oben), um eine größere Unabhängigkeit der Stichprobenhaushalte zu erreichen. Dies war in der DDR nicht möglich, da die Interviewer mit dem Random-Route-Verfahren ohnehin noch nicht vertraut waren, so daß sie durch einen längeren Begehungsweg nicht zusätzlich belastet werden sollten.
- 13 Drei Prozent der Haushalte wurden erst in der ersten Juliwoche vollständig befragt. Da das Interviewdatum im Datensatz festgehalten wird, ist bei der Auswertung eine Kontrolle möglich, ob die Daten vor oder nach der Währungsunion am 1. Juli erhoben wurden.

- 14 Beckett et al. (1989) empfehlen dies auch für die PSID-Stichprobe. Bei der Umsetzung des ursprünglichen SOEP-Konzeptes zeigte es sich ohnehin, daß das Weiterverfolgungskonzept von Stammpersonen für die praktische Feldarbeit zu kompliziert war, d.h. es wurden auch Personen interviewt, die nach den Regeln des Stammpersonenkonzepts nicht hätten interviewt werden dürfen. Hinzu kommt, daß sich die Hochrechnung bei dem erweiterten Konzept wesentlich vereinfacht, da die Ziehungswahrscheinlichkeiten einfacher zu berechnen sind (vgl. Kalton 1989, S. 583 ff.).
- 15 Freilich ist die Vernachlässigung von Einwanderern für eine Panel-Studie, die das Armutsproblem als einen wesentlichen Analysebereich ins Auge faßt, argumentativ durchaus nicht leicht zu begründen. Im Jahre 1990 wird erstmals eine - extern finanzierte - Zuwandererstichprobe von sog. Hispanics in die PSID einbezogen werden. Als zweite, ebenso bedeutsame Zuwandererpopulation fehlen dann immer noch asiatische Einwanderer im PSID.
- 16 Die Information zum Datenschutz ist bei Wiederholungsbefragungen aus rechtlichen Gründen besonders wichtig, aber insgesamt wahrscheinlich keine "vertrauensbildende Maßnahme", da ein höheres Maß an Sensibilität der Fragen suggeriert wird, als es tatsächlich gegeben ist, wie Testuntersuchungen zeigen (vgl. Hippler et al. 1990).
- 17 Es konnte ebenfalls gezeigt werden, daß die bisherige Teilnahmedauer am Panel - mit Ausnahme der neu eintretenden Personen - keinen statistisch signifikanten Einfluß auf die Teilnahmebereitschaft hat (vgl. Rendtel 1990).
- 18 Dieses Problem, ggf. unterschiedliche Hochrechnungsfaktoren in einem Haushalt verwenden zu müssen, ist für die PSID seit langem bekannt.
- 19 Ähnliche Schwierigkeiten ergeben sich bei der Abgrenzung zum Anstaltsbereich. Während nach dem demographischen Ansatz die Stammpersonen - zumindest theoretisch - in den Anstaltsbereich hineinwachsen, bleibt die Bevölkerung in Privathaushalten definitionsgemäß außerhalb dieses Bereichs.
- 20 Für alternative Datenbankkonzepte vgl. Engel (1991), Brecht (1990) und David (1989); sowie Schmaus (1990) für eine SPSS-Lösung, die bei großen Panels - wie dem SOEP - freilich suboptimal ist.
- 21 Es wäre auch wissenschaftstheoretisch naiv anzunehmen, daß jeder Datennutzer bei einer gleichartigen Fragestellung aus komplexen Daten die exakt selben Antworten erzielen würde. Es gibt grundsätzlich zu viele Operationalisierungsprobleme im Detail und Bereinigungsmöglichkeiten, als daß mit dem SOEP identische Ergebnisse von verschiedenen Nutzern vorgelegt werden, wenn die Fragestellung nur hinreichend komplex ist.
- 22 Zu diesem Zwecke wurde das "Merkblatt zum Datenschutz", das jedem befragten Haushalt ausgehändigt wird, eigens um den Hinweis auf Datenweitergabe in das Ausland geändert.
- 23 Ein Sub-Sample eines Panels darf nicht eine bloße 95 %-Stichprobe der Datenmenge einer jeden Welle darstellen. Dadurch würden die Längsschnittinformationen zerstört. Das Public-Use Sample wird deswegen über eine Auswahl der Haushalte der ersten Welle bestimmt. Die Informationen dieser Haushalte (und aller daraus entstehender split-offs) werden Welle für Welle als Public-Use-File weitergegeben.
- 24 Es müssen eine nennenswerte Zahl von Fällen sein, die entfernt werden. Ansonsten könnte man aus einem Vergleich der Randverteilungen des gesamten Samples und des Sub-Samples die Merkmale der entfernten Personen rekonstruieren.