Vierteljahrshefte
zur Wirtschaftsforschung
77 (2008), 3, S. 77–109

Mikrodaten, Gewichtung und Datenstruktur der Längsschnittstudie Sozio-oekonomisches Panel (SOEP)

von Jan Goebel, Markus M. Grabka, Peter Krause, Martin Kroh, Rainer Pischner, Ingo Sieber und Martin Spieß*

1 Einleitung

Die Daten des Sozio-oekeonomischen Panels (SOEP) werden seit 1984 auf Grundlage eines mit dem Befragungsinstitut Infratest abgestimmten Erhebungskonzepts jährlich erhoben. Erfasst werden die Daten zentral in München bei TNS Infratest, aufbereitet in München *und* beim DIW Berlin, um dann möglichst zeitnah durch die SOEP-Gruppe am DIW Berlin an externe Nutzer weitergegeben zu werden. Das komplexe Zusammenspiel von Erhebung, Aufbereitung, Speicherung und Weitergabe der SOEP-Daten wird laufend aktualisiert und immer wieder an die jeweiligen befragungs- und datentechnischen Erfordernisse angepasst.

Beispielsweise erfolgten Mitte der 80er Jahre noch alle Datenarbeiten im DIW Berlin als Mainframe-Anwendungen (IBM VM/370). Die Daten wurden von Infratest als Rohdaten (EBCDIC) mit fixer Bespaltung auf Magnetbändern übertragen. Der Umstieg auf Unix-Server erfolgte Mitte der 90er Jahre – die Daten wurden dann zum Teil auf Disketten und derzeit komplett auf CD beziehungsweise DVD geliefert. Der inzwischen weit über 20 Jahre umfassende Erhebungszeitraum spiegelt sich jedoch nicht nur in den Veränderungen der technischen Datenaufbereitung, sondern auch in den eigentlichen SOEP-Daten wieder. Neben der durch den Fall der Mauer motivierten Ausdehnung des Erhebungsgebietes auf die neuen Bundesländer zeigen sich in den SOEP-Daten weitere im Laufe der Zeit erfolgte exogene Schocks der Bevölkerungsgröße und -struktur in Deutschland. Beispielsweise stieg im Zuge des Familiennachzugs, u. a. infolge des Bürgerkriegs Anfang der 90er Jahre in Jugoslawien, der Umfang der entsprechenden SOEP-Substichprobe stark an,[1] wohingegen sich z. B. das Subsample der Spanier aufgrund von Re-Migration nach dem Zusammenbruch des Franco-Regimes, den ersten freien Wahlen in Spanien 1977 und der Einbeziehung Spaniens in die EU stark verkleinert hat. Auch die aus den offenen Angaben der Befragten geäußerten Sorgen zeigen die mit weltpolitischen Ereignissen einhergehenden Ängste in der Bevölkerung auf – um AIDS, um Rechtsradikalismus, um die Ost-West-Integration, um Krieg und Terrorismus.

* DIW Berlin, E-Mail: *jgoebel@diw.de*, *mgrabka@diw.de*, *pkrause@diw.de*, *mkroh@diw.de*, *rpischner@diw.de*, *isieber@diw.de*, *mspiess@diw.de*

1 Die maximale Haushaltsgröße im besagten Sample B2 (Ex-Jugoslawien) lag in den Jahren 1990 bis 1994 über zehn Personen.

Das Spannungsverhältnis von Kontinuität und Wandel bestimmt den Charakter einer Längsschnittstudie also auf *allen* Ebenen. Vor diesem Hintergrund stellt der vorliegende Aufsatz die Grundstruktur der Erhebung und der Daten des SOEP für den Zeitpunkt des Jahres 2007 dar, und nimmt gleichzeitig Bezug auf die Veränderungen dieser Strukturen im Laufe der vergangenen Jahr(zehnt)e.[2] Diese Veränderungen und die damit einhergehenden Lerneffekte rund um die Komplexität von haushalts- und personenbezogenen Längsschnitt-Daten werden in Kapitel 2 näher beleuchtet. Gleichzeitig wird versucht, mit der Reihenfolge der beschriebenen Aspekte den Entstehungsprozess einer jeden neuen Welle des SOEP nachzuzeichnen. Die Entwicklung der eingesetzten Erhebungsinstrumente und der Stichproben wird in Kapitel 3 beschrieben. Kapitel 4 stellt dar, wie im SOEP auf Grundlage der erfolgreich interviewten Haushalte und Personen Unterschiede in den Ziehungs- und Ausfallwahrscheinlichkeiten mithilfe der Gewichtung ausgeglichen und fehlende Angaben imputiert werden. Die eigentliche Struktur der Mikrodatenbasis, die nach zusätzlichen Datengenerierungen durch die SOEP-Gruppe letztendlich an externe Nutzer weitergegeben wird, beschreibt Kapitel 5.

Die Nutzung statistischer Daten ist jedoch nicht möglich ohne zusätzliches Wissen über die Daten, in Form von Meta-Daten und Datendokumentation; Kapitel 6 gibt daher einen Überblick über die sehr umfassende Dokumentation des SOEP. Die Möglichkeit zur Anmischung extern erhobener Regionalindikatoren an Mikrodaten wurde in den vergangenen Jahren stark verbessert. Inzwischen sind Verknüpfungsmöglichkeiten bis auf Postleitzahlen- oder Gemeindeebene möglich; Kapitel 7 gibt eine Zusammenstellung der derzeit vorhandenen Möglichkeiten. Das abschließende Kapitel dieses Aufsatzes resümiert die bereits vorgenommenen und gibt einen Ausblick auf mögliche weitere qualitative und quantitative Verbesserungen der Daten des Sozio-oekonomischen Panels.

2 Vom Adressprotokoll zur Daten-DVD, oder: Zur Komplexität von Mikrodaten im Längsschnitt

Die Erhebung der SOEP-Daten erfolgt einmal jährlich, im Regelfall in den ersten Monaten eines Kalenderjahres. Dabei kommt eine Vielzahl von Erhebungsinstrumenten zum Einsatz, um sowohl den bisherigen Lebensverlauf als auch die aktuelle Lebenssituation von Personen im Kontext ihrer Familie und ihres Haushaltes zu erfassen. In zufällig und repräsentativ ausgewählten Haushalten werden alle Erwachsenen (17-Jährige und Ältere) mithilfe von Personenfragebögen direkt befragt. Zusätzlich werden von der „Hauptauskunftsperson" (Haushaltsvorstand) im Haushaltsfragebogen Merkmale über den gesamten Haushalt (z. B. die Wohnung) und über Kinder unter 17 Jahren im Haushalt erhoben. Diese Standardfragebögen beinhalten zum überwiegenden Teil gleich bleibende Fragen zur Messung von Stabilität und Wandel der Lebensbedingungen und einen über die Zeit rotierenden Teil zu einem jeweiligen Schwerpunktthema. Hinzu kommen eine Reihe spezieller Fragebögen, wie z. B. zum Lebenslauf bei neu erfassten Personen oder Fragebögen, mit denen Mütter Angaben über ihre (kleinen) Kinder machen. Im Folgenden soll der Prozess der Erhebung und Aufbereitung der Daten durch die SOEP-Gruppe in Berlin überblicksartig skizziert werden.

2 Der vorliegende Beitrag konzentriert sich dabei auf die datenaufbereitenden Arbeiten der SOEP-Gruppe am DIW. Den Produktionsprozess der SOEP-Daten und die historischen Erfahrungen beschreibt aus Sicht der Münchener SOEP-Gruppe bei Infratest Bernhard von Rosenbladt in diesem Heft.

Das Adressprotokoll

Da bei einer Wiederholungsbefragung eine der wichtigsten Grundvoraussetzungen die korrekte Zuordnung von Personen und Haushalten über die Zeit ist, umfassen die Erhebungsinstrumente nicht nur die diversen direkt an Haushalte und Personen gerichteten Fragebögen, sondern auch das vom Interviewer geführte *Adressprotokoll* mit methodischen Angaben zur Feldsteuerung. Hier sind zunächst alle nach den Vorjahrsangaben zu erwarteten Personen im Haushalt mit demografischen Basismerkmalen (Geschlecht, Geburtsjahr, Stellung zum Haushaltsvorstand) gelistet. Die Interviewer prüfen anhand dieser Angaben die vollständige Erfassung aller Haushaltsmitglieder und ergänzen oder streichen Personeneinträge bei entsprechenden personellen Veränderungen infolge von Geburt und Zuzug beziehungsweise Wegzug und Tod (vgl. dazu auch das anschauliche Beispiel der Familie Söpp bei von Rosenbladt in diesem Heft).

Seit dem Jahr 1994 wird auf der Adressliste auch die jeweilige Nationalität erfasst, sodass diese Angabe seitdem auch für noch nicht befragte Kinder jährlich vorliegt. Auf Haushaltsebene werden seit einigen Jahren auch Telefonanschluss und Email-Adresse aufgeführt. Seit 2007 werden zudem rudimentäre Proxyangaben zum Erwerbsstatus für nicht befragungswillige Mitglieder eines Befragungshaushalts (*Partial Unit Non-Response*, PUNR) erfasst. Bei neuen Stichproben wurden darüber hinaus Quartiersmerkmale durch die Interviewer erhoben, um die Güte der Stichprobenziehung beziehungsweise Ausfallprozesse bei der Erstbefragung zu prüfen. Die im SOEP sehr detaillierten Feldangaben werden laufend erweitert und im sogenannten *Codebook* systematisch dokumentiert, das inzwischen zusammen mit den Brutto-Daten[3] der Adressprotokolle für Personen und Haushalte am Jahresende von Infratest aktualisiert bereitgestellt wird.

Personen- und Haushaltsfragebögen

Die eigentlichen Interviews werden mithilfe von Personen- und Haushaltsfragebögen erhoben, dabei werden seit 1994 für Erst- und Folgebefragte einheitliche Fragebögen mit entsprechenden Filterfragen benutzt. Zuvor wurden jeweils unterschiedliche Fragebögen verwendet, die „grünen" für die älteren und die „blauen" für die erstmals im SOEP Befragten. Seit 1994 werden die bei neu befragten Personen einmalig zu erhebenden Biografieangaben in einem eigenen Lebenslauffragebogen erfasst.[4] Die Biografiemerkmale wurden seither um Merkmale zum Elternhaus, zu Angaben von Kindern (auch bei Vätern) sowie zum Migrationshintergrund erweitert. Seit 2000 erhalten Jugendliche beim Eintritt ins Befragungsalter einen eigenständigen, spezifisch auf diese Alterstufe abgestimmten Jugendbiografie-Fragebogen, der seit dem Erhebungsjahr 2006 für Teile der erstbefragten Jugendlichen den Personenfragebogen ersetzt. Das führt jedoch zu Veränderungen in den Populationsabgrenzungen, da die Befragungspopulation seitdem nicht mehr alleine die Personen mit realisiertem Personeninterview umfasst, sondern auch die Jugendlichen mit ausgefülltem Jugendfragebogen. Insgesamt haben sich in den letzten Jahren die Erhe-

3 Der Begriff der „Brutto"-Daten umschreibt hier die Gesamtheit aller zu befragenden Einheiten; der Begriff „Netto" bezieht sich entsprechend auf die Population der faktisch realisierten Interviews (siehe auch Kapitel 5 und insbesondere Fußnote 16).

4 In den ersten drei Wellen (1984–1986) erfolgte die Biographie-Erhebung peu à peu als jeweiliger Schwerpunkt im Personenfragebogen. D.h. seit Einführung des Biographiefragebogens werden erstmals Befragte deutlich stärker belastet. Bei neuen Teilstichproben erfolgt daher diese einmalige Zusatzbefragung erst in der zweiten bzw. dritten Welle (siehe dazu die Biografie-Dokumentation, Frick und Groh-Samberg 2007).

bungsinstrumente immer weiter diversifiziert (vgl. dazu die Beiträge von Schupp et al., von Rosenbladt sowie Trommsdorff in diesem Heft).

Bis 1994 wurden die Personenfragebögen für die zum Zeitpunkt der Stichprobenziehung 1983 quantitativ bedeutsamsten fünf Zuwanderergruppen (aus der Türkei, Griechenland, Italien, Spanien sowie dem ehemaligen Jugoslawien) in die jeweiligen Landessprachen übersetzt. Diese aufwändige Vorgehensweise hat sich aufgrund der weiteren Felderfahrung nach einigen Jahren als nicht mehr notwendig erwiesen und wurde aufgegeben. Freilich steht nach wie vor eine Übersetzung des Fragebogens als Hilfe für Personen mit sprachlichen Verständigungsproblemen zur Verfügung.

Die Befragungsinhalte werden zum Großteil jährlich, bei manchen Fragen zu Beruf, Gesundheit und Freizeitaktivitäten auch zweijährig wiederholt; in jedem Jahr gibt es zudem spezifische Befragungsschwerpunkte, welche auch in größeren zeitlichen Abständen, in der Regel nach fünf Jahren, wiederholt werden. Ad-hoc-Anpassungen von Frageformulierungen mussten insbesondere bei Fragen mit sozialpolitischem Bezug infolge veränderter Rahmenbedingungen erfolgen (z. B. Einführung des Euro im Jahre 2002 oder ALG II in 2005). Die Fragebögen ändern sich damit faktisch von Jahr zu Jahr, zum einen um mit aktuellen Entwicklungen Schritt zu halten und zum anderen unterliegen sie ständiger Verbesserung. Seit einigen Jahren wird vom Erhebungsinstitut Infratest bereits nach Erstellen des Fragebogens gespeichert, welche Variablen aus welchem Jahr unverändert weitergeführt werden und welche modifiziert wurden. Hierdurch kann, zumindest in beschränktem Maße, die Güte der Vergleichbarkeit korrespondierender Variablen im zeitlichen Verlauf genauer beurteilt werden, wobei nicht nur die exakte Identität der Frageformulierung, sondern die funktionale Äquivalenz der erhobenen Information relevant ist. Die systematische Erfassung dieser Veränderungen im zeitlichen Verlauf ist in den letzten Jahren in Zusammenarbeit mit Infratest immer weiter ausgebaut worden und wird zukünftig zu einer erweiterten Beurteilung der Korrespondenz genutzt.

Prüfprozeduren und „missing values“

Die aus dem Feld eingehenden Daten werden von Infratest zunächst so zusammengeführt, dass alle Untersuchungseinheiten konsistente Personen- und Haushaltsidentifikatoren erhalten. Danach erfolgt eine Vielzahl von inhaltlichen Basisprüfungen; die dazu erforderlichen Prüfprogramme müssen infolge der Veränderungen im Fragebogen jeweils immer wieder neu angepasst werden.[5] Dabei werden – z. B. nach Beantwortung einer Filterfrage – Fragenblöcke für nicht zu Befragende (z. B. Berufsangaben für Nichterwerbstätige) mit „Trifft nicht zu“-Codes („–2“) versehen sowie fehlende Angaben identifiziert und als Verweigerung („–1“, keine Angabe beziehungsweise *item nonresponse*) vercodet.

5 In den ersten Jahren wurden hierzu in Fortran „Plutogramme“ programmiert; derzeit sind diese komplexen Abfragen noch in einer dBase-Anwendung programmiert. Für CAPI-Interviews (bei denen der Interviewer die Angaben in einen Labtop eingibt) wird der Fragebogen samt den entsprechenden Prüfroutinen in jedem Jahr neu erstellt und zeigt bereits im Interview Inkonsistenzen an. Bei bestimmten offenen Fragestellungen zu Einkommen und insbesondere Vermögen werden in CAPI bei Verweigerungen kaskadenartig weitere Fragen gestellt, die gegebenenfalls zumindest grobe Angabe zulassen und wertvolle Hinweise bei der Imputation fehlender Werte liefern können; diese Informationen sind im Rahmen von Paper-and-Pencil (PAPI) Interviews im Allgemeinen nicht erhebbar.

In den ersten Jahren des SOEP wurden inkonsistente Befragungsangaben noch im Datensatz erhalten – seit 1991 werden aber offenkundige Inkonsistenzen bei Vorliegen weiterer Informationen bereinigt oder mit einem zusätzlichen *missing value* Code („–3“) als nachträglich gelöschte Angabe kenntlich gemacht.[6] Alle drei Arten an fehlenden Werten (keine Angabe, trifft nicht zu, nachträglich gelöschter Wert) waren zunächst – um Speicherplatz zu sparen, der Anfang der 80er Jahre noch eine sehr knappe Ressource war – alphanumerisch einspaltig vercodet (&, blank, –) und wurden erst beim Einlesen in die Datenbank am SOEP in die SOEP-spezifische *missing value* Deklaration (–1, –2, –3) überführt. Inzwischen werden auch die Befragungsdaten bei Infratest mit einheitlicher *missing value* Vercodung geführt.

Variablennamen und Itemkorrespondenzliste

Für den Ablauf der Feldarbeit sowie innerhalb der eigenen Prüfprozeduren operiert Infratest mit eigenen „sprechenden“ Variablenbezeichnungen. Diese werden seit dem Jahr 2006 nicht mehr jahresweise variabel sondern im Zeitablauf fix definiert und wurden für die zurückliegenden Jahre bis 2000 in einer konsistenten Strukturtabelle abgelegt, die alle Fragebogeninformationen beinhaltet. In dieser erst seit wenigen Jahren bestehenden jahresübergreifenden Strukturtabelle (im Excelformat) werden auch weitere Merkmale zur Charakterisierung der Variablen (Format, Bespaltung) sowie der Bezug zu den korrespondierenden Vorjahresvariablen, die Kennzeichnung von Veränderungen und die Identifizierung neuer Variablen abgelegt. Diese Strukturtabelle wird inzwischen jährlich mit der Datenlieferung bereitgestellt und bildet zugleich die Grundlage zur Generierung der Variablendefinitionen (Variablen und Werte Labels, Format) zwecks Einlesen der Daten ins SOEP-Format.

Prinzipiell werden inzwischen alle von Infratest geprüften Befragungsdaten (Netto-Daten) und Adressprotokolle (Brutto-Daten) in die SOEP-spezifische Datenstruktur überführt und gespeichert. Die jährlich anfallenden Befragungsdaten werden unter der SOEP-spezifischen Namenskonvention (alphanumerisches Wellenkennzeichen und Verweis auf die Fragebogennummer bei Befragungsdaten beziehungsweise „sprechende“ und wellenübergreifend konstante Namen bei Brutto-Daten) ebenfalls jahresbezogen abgelegt. Ausgegliedert werden lediglich datenschutzrechtlich sensible Informationen (Regional- und Klartextangaben sowie Vornamen[7]).

Die Daten werden dann in die SOEP-spezifische Struktur übertragen und mit dem Datenbankprogramm SIR verwaltet. Die damit korrespondierenden Variableninformationen werden inzwischen aus der Strukturtabelle ausgelesen und zusammen mit weiteren Datenstrukturmerkmalen (Variablen- und Werte-Labels, Häufigkeiten, Filetyp, Fragebogenbezug) in einer weiteren relationalen Datenbank (PostgreSQL) gehalten, die wiederum die Grundlage für das Daten- und Variableninformationssystem des SOEP (SOEPinfo) bildet.

6 Für die zentralen Befragungsdaten (Personen- und Haushaltsfragebögen) werden neben den geprüften Daten auch die ungeprüften Originaldaten von Infratest an das DIW Berlin geliefert, sodass der Prüf- und Bereinigungsprozess nachvollzogen werden kann. Die mit „–3“ vercodeten Fälle werden ggf. wie die aufgrund von item-non-response fehlenden Werte („–1“) behandelt und in entsprechende Imputationsroutinen aufgenommen.

7 Zur Konsistenzprüfung der intertemporalen Zuordnung von Personen werden Vornamen erfragt und von Infratest an die SOEP-Surveygruppe im DIW Berlin geliefert. Davon unabhängig werden Namen sowie die Adresse des Befragten aus Datenschutzgründen nur bei Infratest Sozialforschung in München gehalten (getrennt von den Surveydaten).

Der Vorteil von SIR ist nach wie vor seine fallorientierte hierarchische Struktur, die eine natürliche Entsprechung in längsschnittorientierten Haushalts- und Personendaten besitzt.[8] Auf diese Weise werden die spezifischen Vorteile der jeweiligen Datenbanksysteme im Bereich der Datenhaltung und Generierung einerseits sowie der flexiblen Verknüpfung von Variableninformationen andererseits optimiert.

SOEPinfo

Die Bereitstellung von Itemkorrespondenzlisten, die die korrespondierenden Variablen über die Zeit abbilden, ist für das Arbeiten mit Paneldaten elementar. Die ersten korrespondierenden Variablenlisten wurden in den ersten Jahren noch in Papierform bereitgestellt; im Jahr 1990 wurde dann eine erste in DOS/Clipper kompilierte elektronische Form (SOEPinfo) erstellt, die in den folgenden Jahren immer wieder erweitert wurde und zusammen mit den Daten bis 2004 jährlich ausgeliefert wurde.[9] SOEPinfo wurde im Laufe der Jahre immer weiter zu einem kompakten Informationssystem ausgebaut. Bereits seit Mitte der 90er Jahre läuft dieses Software-Paket als Web-Anwendung und umfasst heute neben der Itemkorrespondenz flexible Suchformen über Variableneingabe und Themenlisten, ungewichtete Häufigkeits-Auszählungen, Abbilder der Fragebögen mit verlinkten Variablenbezeichnungen, grundlegende Fallzahlen auf Personen- und Haushaltsebene im Quer- und Längsschnitt sowie Programmgeneratoren zum Aufbereiten der Analysefiles in SPSS, SAS oder STATA. Insofern werden über die web-basierte Schnittstelle von SOEPinfo keinerlei Mikrodaten zur Verfügung gestellt, sondern Meta-Informationen über die SOEP-Daten sowie aktive Unterstützungsleistungen für externe SOEP-Nutzer.

Datenschutz: Speicherung sensibler Informationen

Datenschutzrechtlich empfindliche Informationen wurden früher komplett außerhalb der Datenbank gehalten. Dieses Vorgehen hat im Laufe der Zeit aber zu Inkonsistenzen und Datenlücken geführt, sodass inzwischen auch sensiblere Daten von vornherein in der SIR-Datenbank abgelegt werden. Der Zugriff auf diese Daten ist hochvertraulich; ein Missbrauch allerdings faktisch ausgeschlossen, da auch innerhalb der SOEP-Gruppe der Zugriff auf die SIR-Datenbank auf wenige Mitarbeiter beschränkt ist (und das DIW Berlin als Ganzes keinerlei Zugriff hat). Die entsprechenden Daten wurden in zum Teil sehr aufwändigen Verfahren rückwirkend aufbereitet und integriert. Hierzu zählen detaillierte Regionalinformationen (siehe Kapitel 7), die in den letzten Jahren bis hin zur Ebene der Postleitzahlen immer weiter verfeinert wurden; seit 2000 liegen zudem über die Adresse zugespielte Wohnumfeldmerkmale (Microm) vor. Für die Nachcodierung von Berufen und Bildungsangaben wurden rückwirkend aus den ursprünglichen Originaldatensätzen Klartextangaben extrahiert und zusammen mit der Neuvercodung in eigenständigen Datensätzen wellenübergreifend aufbereitet und gespeichert. Alle laufenden Klartextangaben werden inzwischen ebenfalls an zentraler Stelle wellenübergreifend geführt.

8 Siehe Krause und Wagner (1991), sowie Frick, Krause und Schupp (1992).
9 Etwa zur gleichen Zeit entstand auch das SOEP-spezifische Literaturverwaltungsprogramms SOEPlit, mit dem bibliographische Angaben zu allen verfügbaren Publikationen auf Basis des SOEP verwaltet werden. Jede/r Datennutzer/in des SOEP verpflichtet sich im Rahmen des Datenweitergabevertrages, eine Kopie aller eigenen SOEP-basierten Publikationen an das DIW Berlin zu übermitteln.

Aufbereitungen für die Datenweitergabe

In den ersten Jahren wurden die Daten für die Nutzer noch „von Hand“ aufbereitet. Im Jahr 1989 erfolgte die erste automatisierte Datenausgabe auf Grundlage einer in REXX für die IBM-Mainframe programmierten interaktiven Datenbankanwendung. Dabei wurde die in der Datenbank angelegte Struktur als Rohdaten mit generierten Einleseschemata ausgegeben. Darauf aufbauend wurden Programmmodule (DIC2SAS, DIC2SPSS) per Hand erstellt, mittels derer die Rohdaten für die Nutzer in die gängigen Statistikpakete (SPSS und SAS) überführt werden konnten. Zu dieser Zeit waren Speicherkapazitäten noch eine knappe (und teure) Ressource und komplexe Längsschnittanalysen wurden von den bestehenden Analysepaketen noch wenig unterstützt. Unter anderem deswegen hat Götz Rohwer auf Basis der SOEP-Daten ein eigenständiges Analysepaket TDA erstellt, das nicht nur die Berechnung komplexer Zeitreihenmodelle bis hin zu Sequenzanalysen erlaubte, sondern auch extrem speicher-effizient war, unter anderem weil es auf in RZOO gepackte Daten zugriff, die für die Analyse nur selektiv entpackt wurden. Weiterentwicklungen von TDA finden bis heute noch unter SOEP-Nutzern Anwendung (siehe auch Krause, Pischner, Wagner 1993).

Inzwischen werden die Daten von Infratest nahezu ausschließlich als SPSS-Datensätze bereitgestellt. Auf Nutzerseite werden neben SPSS- insbesondere STATA-Datensätze zunehmend nachgefragt, während die Zahl der SAS-Nutzer national und international zurückgeht. Vor allem in Statistikerkreisen finden inzwischen zudem Anwendungen in R vermehrt Anklang. Seit einigen Jahren werden die SOEP-Daten intern nach Updates der Basis-Datenbank (in SIR) automatisiert und tagesaktuell in allen gängigen Softwarepaketen erzeugt, da die Datenaufbereitung und -generierung innerhalb der SOEP-Gruppe mit unterschiedlichsten Statistikpaketen erfolgt. Für die Datenweitergabe werden neben der deutschen Version auch eine englischsprachige Version (mit englischsprachigen Labels und Dokumentation) bereitgestellt.

Der Zeitplan

Die Zeitspanne von der Erstellung des Fragebogens bis zur Weitergabe der Daten an die Nutzer umfasst gut zwei Jahre. Im Vorjahr der Erhebung wird zunächst der Fragebogen erstellt. Daran schließt sich die die Ergänzung der Strukturtabelle, die Einarbeitung der Prüfprogramme sowie die Programmierung der CAPI-Befragung an. Die Befragten werden im Rahmen der „Panel-Pflege“ kontaktiert und die Interviewer instruiert. Die Feldarbeit beginnt in der Regel im Januar eines jeden Jahres – bis April des Jahres sind circa 80 % der Interviews durchgeführt. Bis zum Jahresende erfolgt in aufwändiger Weise die Nacherhebung schwer erreichbarer Personen sowie die umfangreiche Datenprüfung.

Die erhobenen und geprüften Daten werden von Infratest jeweils am Ende des Erhebungsjahrs (Mitte bis Ende Dezember) an das SOEP-Team ausgeliefert. Auf Grundlage der Strukturtabellen werden dann Einlese-Schemata für die Speicherung der Daten im SOEP-Format generiert. Die Einleseschemata müssen dabei für die nicht per Algorithmus generierbaren Teile jedes Jahr von Hand angepasst werden. Die eingelesenen Daten werden weiter geprüft und in enger Rücksprache mit Infratest laufend aktualisiert. Ab Anfang März stehen die so erfassten Daten in „erster Lesung“ für die darauf aufbauenden weiteren Datengenerierungen zur Verfügung, die meist bis Mai/Juni abgeschlossen

sind. Mitte des Jahres – bereits ein halbes Jahr nach Erhalt der Felddaten – werden die neuen Daten zunächst zu Testzwecken an Beta-User und dann allgemein an die nationale und internationale Nutzergemeinschaft weitergegeben. Weitergehende Datenarbeiten wie die Nacherhebung von Klartexten, rückwirkende Generierung von Bildungsangaben, Erfassung und Aufbereitung von Berufsvercodungen, Aufbereitung von Fälschungen oder Einbinden neuer Datentypen (Microm), sowie Arbeiten zur Weiterentwicklung der Datenstruktur erfolgen meist in der zweiten Jahreshälfte.

3 Erhebungsweise und Stichproben

Die Zahl und Vielfalt der im SOEP je Welle eingesetzten Erhebungsinstrumente hat sich seit dem Jahr 2000 deutlich gesteigert (siehe Tabelle 1). Neben den langjährig standardisierten Fragebögen zur Erhebung von Personen- und Haushaltsinformationen bei Personen im Befragungsalter werden inzwischen auch regelmäßig weitere Instrumente zur detaillierten Erfassung des Lebensverlaufs eingesetzt. Derzeit werden zudem Informationen zum Lebenslauf bei neu befragten Erwachsenen, bei erstbefragten Jugendlichen sowie bei Müttern/Eltern zur Geburt und für ihre 2- bis 3-jährigen Kinder erhoben. In den nächsten Jahren sollen durch weitere Befragungsinstrumente auch das Vorschul- und Schulalter noch detaillierter abgebildet werden (5- bis 6-jährige Kinder). Bei einem vorzeitigen Ausscheiden aus dem SOEP werden Todesfälle zum Teil durch Nachrecherchen ermittelt und in die vorhandene Datenstruktur integriert. Auf diese Weise werden die verschiedenen Phasen im Lebensverlauf mit jeweils spezifischen Instrumenten detailliert erfasst (siehe Tabelle 2).

Tabelle 1

Erhebungsinstrumente und Datenanreicherung im SOEP

Wiederholt pro Lebenslauf	Einmalig pro Lebenslauf
Adressprotokoll	
Fragebögen	
Personenfragebogen für „alte" Personen (grün bis 1993)	Personenfragebogen für neue Personen (blau bis 1993)
Personenfragebogen für alle Personen	
Nacherhebungsfragebogen für Lückefälle	Lebenslauffragebogen (1984 und seit 1987 für alle neuen Personen)
	Jugendfragebogen (seit 2000)
	Mutterkind-Frabogen I (Kinder im Alter von 0–1 Jahren)
	Mutterkind-Frabogen II (Kinder im Alter von 2–3 Jahren)
	Mutterkind-Frabogen III (Kinder im Alter von 5–6 Jahren)
Andere Erhebungsformen und Tests	
Greifkrafttest	Kognitive Leistungsfähigkeit von 17-Jährigen
Kognitionstests	
Weitere Datenanreicherungen und Recherchen	
Experimente	Verbleibstudie bei Ausfällen
Kleinräumige Regionalinformationen	Wegzug ins Ausland

Quellen: Das Sozio-oekonomische Panel (Wellen A–W), eigene Darstellung.

Tabelle 2

Erhobene Daten mit Bezug zum Lebenslauf im SOEP

Lebensphase	Alter	Auskunftsperson	Analyselevel	Zentrales Datenfile
Fötale Phase	<0	Vater/Mutter	P-Vater/P-Mutter	$P
Geburt und Babyalter	0–1	Mutter	P	BIOAGE01
Kleinkind	2–3	Mutter	P	BIOAGE03
(Vor)Schule	5–6	Mutter	P	BIOAGE06
Kind	0–16	Haushaltsvorstand	P	$KIND
Jugendlicher	17	Befragungsperson	P	BIOAGE17
Erwachsener	18–	Befragungsperson	P	$P
Terminale Phase		Haushaltsvorstand	P	PBR_EXIT
Tod		Haushaltsvorstand/ Nachrecherche	P	PBR_EXIT, PPFAD
Erinnerung und Renten		Hinterbliebener Partner	P-Partner	$P

Anmerkungen: Das $-Zeichen steht für einen Wellenpräfix von A bis derzeit X (1984–2007). Bei den Datensätze BIOAGE01 bis BIOAGE17 beziehen sich die Zahlen auf das Lebensalter der zu befragten Personen oder der Personen über die Informationen erfragt werden. Zur näheren Erläuterung der Datenstruktur und Files siehe Kapitel 5.1.

Quellen: Das Sozio-oekonomische Panel (Wellen A–W), eigene Darstellung.

Bei den Erwachsenen werden neuerdings durch das Einbeziehen von Kognitions- und Greifkrafttests, Recherchen zum Wegzug ins Ausland sowie experimentelle Testsituationen spezifische Verhaltensweisen genauer gemessen. Die Verknüpfung mit kleinräumigen Regionalindikatoren (Umwelt, Zentralität, sozialräumliches Umfeld) sowie mit institutionellen Angaben (Kindergarten, Schule, Arbeitsplatz) sollen die präzise Erfassung des sozialen Umfeldes noch weiter verbessern.

Aufbauend auf der Population der erstmals im Jahr 1984 in Westdeutschland Befragten, werden mit dem Erhebungsjahr 2008 für etwa 2 500 Personen Daten zum fast lückenlosen Lebensverlauf über 25 Jahre vorliegen. In der 1990 in Ostdeutschland gestarteten Stichprobe werden im Jahr 2009 etwa 1 500 Personen zum 20. Mal befragt werden.

Die Stichproben des SOEP

Mit der 2006 gestarteten Ergänzungsstichprobe H umfasst das Sozio-oekonomische Panel nunmehr acht Teil-Stichproben und – für die Stichproben A und B – 24 auswertbare Wellen (1984–2007).[10] Ein Ende der Erhebungen ist nicht geplant. Übersicht 1 zeigt die jeweilige repräsentative Grundgesamtheit der bisherigen acht Stichproben.

Die SOEP-Respondenten werden haushaltsweise für das SOEP rekrutiert. Der Modus der Kontaktaufnahme erfolgt bei allen Teilstichproben in gleicher Weise mithilfe eines Schrei-

10 Eine aktuelle Beschreibung des SOEP findet sich bei Wagner, Frick und Schupp (2007).

Übersicht 1

Die Stichproben des Sozio-oekonomischen Panels

Stichprobe A:	(Deutsche)[1] Haushalte[2] in der Bundesrepublik Deutschland (Hauptstichprobe, Start 1984)
Stichprobe B:	Ausländische Haushalte[3] in der Bundesrepublik Deutschland (Start 1984)
Stichprobe C:	Privathaushalte in der DDR (Start 1990).
Stichprobe D:	Zuwanderer-Privathaushalte in Deutschland (Start 1994/95)
Stichprobe E:	Haushalte in Deutschland (Ergänzungsstichprobe, Start 1998)
Stichprobe F:	Haushalte in Deutschland (Ergänzungsstichprobe, Start 2000)
Stichprobe G:	Hocheinkommens-Privathaushalte in Deutschland (Hocheinkommensstichprobe, Start 2002)
Stichprobe H:	Haushalte in Deutschland, Ergänzungsstichprobe (Start 2006)

1 Genauer: Haushalte, deren Haushaltsvorstand zum Zeitpunkt der Ziehung nicht türkischer, italienischer, jugoslawischer, griechischer oder spanischer Nationalität war. Dies waren ganz überwiegend (99%) deutsche Haushaltsvorstände.

2 Anstaltshaushalte sind bei der Stichprobenziehung nicht eingeschlossen; sie werden zwar auch nicht ausgeschlossen, wenn sie beim Random-Walk gelistet werden und sind insofern im Bruttobestand enthalten, werden aber bei der Durchführung der Befragung in der Regel von den Interviewern bei neuen Samples nicht berücksichtigt. Anstaltshaushalte werden in der Regel erst bei der weiteren Befragung durch Weiterverfolgung per Interviewer erfasst (Umzug ins Altersheim etc); die in den Folgewellen einbezogene Anstaltspopulation ist aber nicht repräsentativ für die Grundgesamtheit.

3 Genauer: Haushalte, deren Haushaltsvorstand zum Zeitpunkt der Ziehung türkischer, italienischer, jugoslawischer, griechischer oder spanischer Nationalität war.

Quellen: Das Sozio-oekonomische Panel (Wellen A–W), eigene Darstellung.

bens der Feldarbeits-Organisation (TNS Infratest München), mit dem das SOEP und ein Interviewer angekündigt werden.[11]

Realisiert wird das Konzept zur Erhebung von Lebensläufen einzelner Personen durch die Ziehung einer Haushaltsstichprobe, bei der alle Personen in diesen Haushalten (genauer: Privathaushalten) selbst Erhebungs-Einheit werden.[12] Alle Personen in diesen gezogenen Haushalten sind sogenannte Stammpersonen oder „*Original Sample Members*“ (OSM). Ziehen Personen aus einem Befragungshaushalt aus, so werden diese innerhalb der Bundesrepublik Deutschland weiter verfolgt, ebenso werden Personen, die mit ihnen zusammenziehen, auf Dauer in das SOEP einbezogen.

Durch den Einbezug von Nicht-Stammpersonen („*Non Original Sample Members*“, NOSM) entsteht im Prinzip eine Schneeballstichprobe (vgl. dazu Tabelle 3). Dies ist beabsichtigt, da so die Lebensläufe der Original-Stichprobenmitglieder besser im Kontext analysierbar sind. Dies gilt auch für Geschiedene, deren Lebenswege sich (scheinbar) völlig trennen (vgl. Spieß et al. 2008). Der Schneeball-Effekt wird durch eine entsprechende Gewichtung der Daten berücksichtigt. Faktisch wird durch dieses „Weiterverfolgungskonzept“ die Stichprobe nicht größer, da den Schneeball-Befragten jene gegenüberstehen,

11 Die Adressen werden unterschiedlich ermittelt: meist per Random-Walk (Stichproben A, E, F und H), einmal per Register (B), einmal per Register und Interviewer (C) und zweimal durch Screenen von Haushalten nach spezifischen Bevölkerungsgruppen (Teilstichprobe D: Zuwanderer, Adressziehung aus Infratest-Bus-Befragung; Realisierung per Standard-Random-Walk; Teilstichprobe G: Hocheinkommenshaushalte, Ziehung aus Standard-Telefon-Interviews).

12 Kinder werden freilich erst mit dem Erreichen des 17. Lebensjahres selbst persönlich befragt. Auch ungeborene Kinder gehören virtuell zu dieser Stichprobe.

Tabelle 3

Zahl der Haushalte im Sozio-oekonomischen Panel im Jahr 2006 nach Stichprobe und Stichprobenstatus der Befragten

Status	Haushalte insgesamt[1]	Nur OSM-Haushalte	Gemischte Haushalte	Nur NOSM-Haushalte[2]	Nur OSM-Haushalte[1]	Gemischte Haushalte	Nur NOSM-Haushalte[2]
Stichprobe	Zahl der Haushalte				Status-Anteile in %		
Insgesamt	12 361	9 490	2 317	554	76,8	18,7	4,5
A	2 821	1 572	950	299	55,7	33,7	10,6
B	655	392	223	40	59,9	34,0	6,1
C	1 717	1 123	461	133	65,4	26,8	7,8
D	222	150	68	4	67,6	30,6	1,8
E	686	567	96	23	82,6	14,0	3,4
F	3 895	3 394	450	51	87,1	11,6	1,3
G	859	786	69	4	91,5	8,0	0,5
H	1 506	1 506	–	–	100,0	–	–

1 OSM-Haushalte: (Original Sample Member): Haushalte, die keine zugezogenen Hausmitglieder enthalten.
2 NOSM-Haushalte: (Non Original Sample Member): Haushalte ohne OSM.

Quelle: Das Sozio-oekonomische Panel (Wellen A–W); eigene Berechnungen.

die nicht mehr bereit sind, am SOEP weiter teilzunehmen („Panel-Ausfälle"). Auch diese Ausfälle werden im Laufe der Panel-Laufzeit durch Gewichtungen berücksichtigt, die unverzerrte Rückschlüsse auf die Grundgesamtheit zulassen (vgl. dazu Abschnitt 4).

Interviewmode und Interviewer-Sample

Die Erhebung wird in der ersten Welle ausschließlich Face-to-Face durchgeführt (mit Paper und Pencil [PAPI]) oder auch – seit 1998 – mit Computer Assisted Personal Interviewing (CAPI). Ab der zweiten Welle ist auch ein Selbstausfüllen mit nur telefonischer Betreuung möglich; diese Möglichkeit wird aber faktisch erst ab Welle 7 (1990) tatsächlich in nennenswertem Umfang genutzt (nach 22 Wellen im Sample A circa 20 %; nach 5 Wellen in Sample F circa 6 % aller realisierten Interviews).

Die verschiedenen Befragungs-Modi sind für jede einzelne Beobachtungseinheit (Person beziehungsweise Haushalt) codiert und im Standard-Datensatz abgelegt und auswertbar. Allein aufgrund der verschiedenen Befragungs-Modi ist das SOEP für survey-methodische Fragen eine wahre Fundgrube (vgl. Schräpler 2007). Hinzu kommt die Identifikation eines jeden einzelnen Interviewers (vgl. Schräpler und Wagner 2000), sodass auch potenzielle Interviewereffekte analysierbar sind (vgl. Schräpler 2004).[13]

13 Da ein nichtexperimentelles Design vorliegt, sind Befragungsartefakte zwar statistisch modellier- und kontrollierbar, aber eine kausale Zuschreibung auf Einzeleffekte bedarf der Setzung von Annahmen.

Ausgewählte Sonderdatenbestände

In Spezial-Datensätzen sind auch Angaben zu Haushalten vorhanden, die sich weigerten, an der Erhebung teilzunehmen („Brutto"-Daten) oder die im Laufe der Zeit ein Interview verweigerten. Da im Längsschnitt auch gefälschte Interviews aufgedeckt werden können, die in einem Querschnitt unentdeckt blieben, wurden im Nachhinein einige (weniger als ein halbes Prozent) Datensätze als vom Interviewer gefälscht identifiziert (vgl. Schräpler und Wagner 2005). Auch diese Daten stehen für Re-Analysen in einem speziellen Datensatz zur Verfügung.

Ab dem Jahr 2008 werden auch die Mikro-Daten einer 2007 durchgeführten Drop-Out-Erhebung allgemein verfügbar sein. Dann werden auch die Daten einer ebenfalls 2007 durchgeführten Interviewer-Befragung vorliegen, die die Daten über die Interviewer, die aus der Buchhaltung des Umfrageinstituts stammen (z. B. Geschlecht und Alter; vgl. Schräpler und Wagner 2000) ergänzen und deutlich vertiefte Analysen von Interviewereffekten zulassen.

Durch die Zuordnung von (kommerziellen) Nachbarschafts-Daten sind auch Analysen von Interaktionseffekte mit „Meso-Variablen" möglich (vgl. Abschnitt 7 unten); aus Datenschutzgründen sind diese kleinräumigen Informationen aber nur innerhalb des DIW Berlin – für jeden registrieren Nutzer – auswertbar. Dies gilt auch für die sozialstrukturelle Analyse der Vornamen der Befragten (vgl. Gerhards und Hans 2006).

4 Gewichtung, effektive Fallzahlen und Imputationen im SOEP

Die Erhebung von Mikrodaten gelingt niemals vollständig – eine Binsenweisheit. Zusätzlich zu den Maßnahmen, die insbesondere auf Seiten des Münchner SOEP-Teams bei Infratest unternommen werden, um die Qualität der Daten bereits im Prozess der Erhebung und Verknüpfung mit dem bestehenden Datenbestand auf hohem Niveau zu halten (vgl. dazu den Beitrag von Rosenbladt in diesem Heft), hat auch das Berliner SOEP Team am DIW verstärkte Bemühungen unternommen, um die Zuverlässigkeit und Vollständigkeit der erhobenen Daten kontinuierlich zu verbessern.

Eine erste und unerlässliche Maßnahme ist die Bereitstellung von Gewichtungsfaktoren. Schon allein aufgrund der unterschiedlichen Samples, insbesondere der überproportionalen Repräsentation von Ostdeutschen, Ausländern und Zuwanderern im SOEP, sowie aufgrund der möglichen selektiven Ausfallwahrscheinlichkeiten ist eine Gewichtung der Daten für Analysen im Quer- wie im Längsschnitt unbedingt notwendig. Da es nur wenige vergleichbar langlaufende Haushaltspanelstudien wie das SOEP auf der Welt gibt, hat das SOEP-Team auch in diesem Bereich Pionierarbeit geleistet.

4.1 Gewichtung und Hochrechnung

Gewichte werden in unterschiedlichsten Analysesituationen benötigt. Besteht etwa Interesse an Kennwerten wie Summen oder Anteilen in einer endlichen konkreten Grundgesamtheit, etwa der in Privathaushalten lebenden Bevölkerung Deutschlands im Jahr 2007, dann werden Gewichte üblicherweise eingesetzt, um bei deren Schätzung die Ziehungs-

und Antwortwahrscheinlichkeiten der betrachteten Stichprobe zu berücksichtigen („design"-basierter Ansatz). Entsprechende Fragstellungen sind typisch etwa im Rahmen der amtlichen Statistik oder Sozialberichterstattung. Sind statistische Modelle, etwa in einem Regressionsmodell Effekte von Variablen wie „Arbeitserfahrung" auf eine abhängige Variable wie „logarithmiertes Einkommen" als Modell für Zusammenhänge in einer abstrakteren Grundgesamtheit von Interesse – dies ist meist in einem akademischen Kontext der Fall –, dann können Gewichte eingesetzt werden, um für mögliche Verzerrungen der Schätzung durch unterschiedliche Antwortwahrscheinlichkeiten zu kompensieren („modell"-basierter Ansatz).

Grundlagen

Gültige Aussagen über Kennwerte in einer Grundgesamtheit auf der Basis einer Stichprobe setzen voraus, dass die dazu verwendeten Schätzer für diese Kennwerte entsprechende, günstige Eigenschaften aufweisen. Kern eines Gewichtungskonzeptes muss es also sein, Analysen zu unterstützen, die angesichts unterschiedlicher Ziehungs- und Antwortwahrscheinlichkeiten gültige Schlüsse ermöglichen.

Ein der Konstruktion der SOEP-Gewichte zugrunde liegender und ursprünglich im designbasierten Rahmen vorgeschlagener Ansatz besteht darin, jede in eine Stichprobe gezogene Einheit, z. B. Haushalte oder Personen, mit dem Kehrwert ihrer Ziehungswahrscheinlichkeit zu multiplizieren, d. h. zu gewichten (Horvitz und Thompson 1952). Voraussetzung ist, dass einerseits die Ziehungswahrscheinlichkeiten für die in die Zufallsstichprobe gezogenen Einheiten bekannt sind und alle gezogenen Stichprobeneinheiten auch beobachtet werden. Dann führt dieses Vorgehen zu gültigen Schlüssen bezüglich des interessierenden Kennwertes in endlicher Grundgesamtheit.

Allerdings werden in allen Datenerhebungen, die auf Freiwilligkeit beruhen, einige der in die Stichprobe gezogenen Einheiten nicht beobachtet werden, etwa weil sie sich weigern, an der Untersuchung teilzunehmen. Die Wahrscheinlichkeit, für eine in der Stichprobe beobachtete Einheit tatsächlich ein Interview zu realisieren, ist also nicht allein aus der Ziehungswahrscheinlichkeit ableitbar, sondern muss aufgrund dieser zweiten Selektionsstufe der Teilnahmebereitschaft eigens geschätzt werden (vgl. hierzu den Abschnitt Startwellengewichte). Bei einem Längsschnittdatensatz kommt hinzu, dass Einheiten über die Zeit hinweg ausfallen („Panel attrition"), also etwa Haushalte, die zu einem Zeitpunkt erfolgreich interviewt wurden, aber zu einem späteren Zeitpunkt nicht mehr Teil der Stichprobe sind (vgl. hierzu den Abschnitt Längsschnittgewichte).

Um die Vielseitigkeit möglicher Analysen mit dem SOEP ausschöpfen zu können, wurde das auf Horvitz und Thompson (1952) zurückgehende Gewichtungskonzept, insbesondere von Galler (1987) und Rendtel (1995), an das Design eines komplexen Haushaltspanels angepasst. Damit sollen nicht nur gültige Schlüsse basierend auf den jeweiligen ersten Wellen der Teilstichproben, sondern auch gültige Querschnittsanalysen für alle Folgewellen (vgl. hierzu den Abschnitt Querschnittsgewichte ab Welle 2) sowie die unterschiedlichsten Längsschnittanalysen ermöglicht werden. Darüber hinaus zeigen neuere Arbeiten, dass die Gewichtung von Einheiten, basierend auf ihren (geschätzten) Beobachtungswahrscheinlichkeiten, nicht nur im design-basierten, sondern auch im modell-basierten Ansatz gül-

tige Schlüsse ermöglichen (z. B. Robins, Rotnitzky und Zhao 1994 und 1995, Wooldridge 2002a, 2002b und 2004).

Startwellengewichte

Die Erzeugung der Gewichte der jeweils ersten Welle einer Teilstichprobe besteht aus mehreren Schritten. Im ersten Schritt werden die Ziehungswahrscheinlichkeiten aus dem jeweils gewählten Auswahlverfahren abgeleitet. Die Auswahlverfahren der Teilstichproben variieren über die Teilstichproben des SOEP, grundsätzlich handelt es sich aber meist um ein zweistufiges Verfahren, bei dem zunächst sogenannte Primäreinheiten, etwa Stimmkreise, und aus diesen dann die Sekundäreinheiten, im Allgemeinen Haushalte, gezogen werden. Bei den Auswahlverfahren handelt es sich meist um systematisches und größenproportionales Ziehen mit festen Intervallen und Zufallsstart (z. B. Särndal, Swensson und Wretman 1992). Die Kehrwerte dieser Wahrscheinlichkeiten werden unter dem Variablennamen DESIGN (abgeleitet aus der Bezeichnung Designgewichte) im Rahmen der Standardweitergabe des SOEP ausgeliefert, die Identifikatoren der Primärziehungseinheiten unter dem Variablennamen PSU (*Primary Sampling Unit*) und die Abgrenzung der Schichtungen unter dem Variablennamen STRAT (*Stratum*) (Spiess 2001, Spieß und Kroh 2007, siehe auch Göbel et al. 2008).

Da nicht alle gezogenen Einheiten auch beobachtet beziehungsweise interviewt werden – waren dies in den ersten beiden Teilstichproben noch etwa 60 %, werden in neueren Teilstichproben wie in allen zufallsbasierten Erhebungen in Deutschland nur noch etwas mehr als 40 % der gezogen Einheiten beobachtet – ist in einem zweiten Schritt für diesen Ausfall zu korrigieren. Die prinzipielle Vorgehensweise besteht darin, auf der Basis der für alle gezogenen Einheiten vorliegenden Informationen Häufigkeitszellen zu bilden und auf der Basis der in jeder Zelle beobachteten Anteile beobachteter Einheiten deren Beobachtungswahrscheinlichkeiten, gegeben die Ziehung in die Stichprobe, zu schätzen. Produktbildung mit den Ziehungswahrscheinlichkeiten führt zu geschätzten Beobachtungswahrscheinlichkeiten. Um übermäßig große Gewichte zu vermeiden, die zu Effizienzverlusten und einer Sensitivität der gewichteten Schätzung bezüglich Ausreißern führen können (vgl. Abschnitt 4.2), werden die Kehrwerte dieser geschätzten Wahrscheinlichkeiten auf das 10fache des teilstichprobenspezifischen Median begrenzt.

Anschließend erfolgt eine Anpassung an Ränder des Mikrozensus mit den Variablen Region, Alter, Geschlecht, Haushaltsgröße und Nationalität (Pischner 2007a, 2007b). Damit sollen etwaige Unzulänglichkeiten bei der Umsetzung der Auswahlverfahren sowie der Schätzung der Antwortwahrscheinlichkeiten ausgeglichen werden. Allen zu befragenden Personen werden zunächst die Kehrwerte der geschätzten Beobachtungswahrscheinlichkeiten ihres Haushaltes zugewiesen, wobei allerdings für Zweitwohnsitze korrigiert wird, da sich mit einem Zweitwohnsitz die Ziehungswahrscheinlichkeit von Haushalten erhöht. Eine zweite Randanpassung auf Personenebene liefert dann die personenspezifischen Gewichte (Pischner 2007a, 2007b). Die so erzeugten Gewichte der jeweils ersten Welle werden für jede Teilstichprobe getrennt erzeugt und mit dem SOEP ausgeliefert. Auf der Ebene der Haushalte sind diese Variablen in der Datei HHRF und auf Personenebene in der Datei PHRF zu finden.

Längsschnittgewichte

Auch Längsschnittanalysen basieren auf Gewichten, in denen sich die Beobachtungswahrscheinlichkeiten der Einheiten zu den spezifischen Zeitpunkten widerspiegeln. Die Wahrscheinlichkeit eine natürliche Einheit, etwa einen Haushalt, in einer zweiten Welle zu beobachten, vorausgesetzt sie wurde bereits in Welle eins beobachtet, hängt davon ab, welche Einheiten der ersten Welle in der zweiten Welle beobachtet werden oder nicht. Werden nicht mehr alle Einheiten beobachtet, dann ist diese Wahrscheinlichkeit kleiner eins und muss auf Basis der bekannten Merkmale der Haushalte aus Welle eins geschätzt werden. Im SOEP kommen zur Schätzung solcher Wahrscheinlichkeiten Logitmodelle für binäre abhängige Variablen „Beobachtung versus Ausfall" eines Haushalts zum Einsatz. Eine zentrale Annahme ist dabei, dass diese Wahrscheinlichkeiten nur von Variablen abhängen, die für alle Einheiten beobachtet wurden. Dies trifft auf Merkmale der ersten Welle zu, bis auf Ausnahmefälle nicht jedoch auf Merkmale der Haushalte aus Welle zwei, da diese für die ausgefallene Population unbekannt sind. Im Allgemeinen handelt es sich somit bei den berücksichtigten Variablen um solche, die in der ersten Welle erhoben wurden oder um solche, die unabhängig von der Beteiligung der Einheiten in der zweiten Welle beobachtet werden können, wie etwa das Wohnumfeld der Personen, dass vom Interviewer beobachtet und mit den SOEP-Daten standardmäßig erfasst wird.

Im Folgenden etwas vereinfacht dargestellt ist die im SOEP gewählte Vorgehensweise zur Schätzung von Längsschnittgewichten von Welle eins auf Welle zwei ein zweistufiger Prozess. In einem ersten Schritt gilt es, den Kontakt zu den Einheiten aus Welle eins wieder herzustellen, und nur wenn diese Kontaktaufnahme erfolgreich ist, entscheidet sich in einem zweiten Schritt, ob es zu einer Antwort der entsprechenden Einheit in Welle zwei kommt oder nicht. Da der erste Schritt der Adressermittlung und Kontaktaufnahme mit einem Haushalt in Welle zwei plausiblerweise von anderen Merkmalen der Haushalte abhängt (z. B. deren Mobilitätsverhalten) als deren Bereitschaft auf Nachfrage einen Fragebogen auszufüllen (z. B. der Motivation), werden beide Schritte im SOEP getrennt mithilfe logistischer Regressionsschätzungen modelliert (Spieß und Kroh 2008). Das Produkt der beiden Wahrscheinlichkeiten liefert die sogenannten „Bleibewahrscheinlichkeiten", deren Kehrwerte auch als Bleibefaktoren bezeichnet werden.

Weiterhin sind im SOEP die unmittelbar gezogenen Einheiten Haushalte. Es sind daher zunächst die zweistufigen Beobachtungswahrscheinlichkeiten (von Welle 1 nach Welle 2) der Haushalte zu modellieren und zu schätzen. Dabei ist allerdings zu berücksichtigen, dass Haushalte (im Sinne einer oder mehrerer zusammen lebenden Personen) kurzlebige, in gewisser Weise „künstliche" Einheiten sind, die sich aufspalten oder die fusionieren können. Die Wahrscheinlichkeiten einen konkreten Haushalt in Welle zwei zu beobachten, gegeben dieser wurde in Welle eins beobachtet, hängt somit nicht nur von einem erfolgreichen Kontaktversuch und der Antwortneigung des Haushaltes, sondern auch davon ab, ob der Haushalt etwa aus einer Abspaltung oder Fusion entstanden ist. Eine detaillierte Darstellung der Vorgehensweise ist in Rendtel (1995) zu finden. Eine weitere Besonderheit gegenüber natürlichen Einheiten stellen in einen Haushalt hineinziehende Nicht-Original-Stichproben-Mitglieder dar, die die Antwortneigung des Haushaltes beeinflussen (können). Bei Nicht-Original-Stichproben-Mitgliedern handelt es sich um Personen, die in der Regel durch Zuzug erst in Welle zwei Teil eines bereits gezogenen Haushaltes aus Welle eins werden (vgl. Tabelle 3.3). Daher sind die Modelle der Beobachtungswahrscheinlichkeiten entsprechend anzupassen, wobei neben Haushalts- auch Personenmerk-

male berücksichtigt werden. Für eine ausführliche Darstellung siehe Spieß, Kroh, Pischner und Wagner (2008).

Diejenigen Haushalte, die in Welle 1 und in Welle 2 beobachtet wurden, erhalten als Wahrscheinlichkeit in Welle 2 beobachtet zu werden, das Produkt von Beobachtungswahrscheinlichkeit in Welle 1 mit der Bleibewahrscheinlichkeit in Welle 2. Der Kehrwert dieser Wahrscheinlichkeit – oder aber Produkt des Querschnittsgewichts der Vorwelle und des Bleibefaktors der aktuellen Welle – ergibt das Längsschnittgewicht der Einheit über beide Wellen.

Das beschriebene Vorgehen wird sequentiell auf jede weitere Welle angewandt und liefert damit die Beobachtungswahrscheinlichkeit für den Zeitraum zwischen der Start- und jeder folgenden Welle. Die Bleibefaktoren als Kehrwerte der Bleibewahrscheinlichkeiten auf Haushaltsebene, $HBLEIB, wobei $ für die jeweilige Welle steht, beziehungsweise auf Personenebene, $PBLEIB, sind Teil der Standarddatenweitergabe des SOEP. Auf der Basis dieser Informationen lassen sich Kennwerte in Längsschnittpopulationen (z. B. Rendtel 1995) oder Panelmodelle (z. B. Wooldridge 2004) schätzen, bei denen für Ausfälle und unterschiedliche Ziehungswahrscheinlichkeiten kompensiert wird.

Querschnittsgewichte ab Welle 2

Um gewichtete Querschnittsanalysen zu ermöglichen, werden mit der ersten Welle ($t = 1$) beginnend die Beobachtungswahrscheinlichkeit in t=1 und die Bleibewahrscheinlichkeit in $t = 2$ multipliziert. Die Kehrwerte dieser als geschätzte Beobachtungswahrscheinlichkeit in $t = 2$ interpretierbaren Produkte bilden die „Rohgewichte“ des Querschnitts in Welle $t = 2$. Mit derselben Begründung wie bei der Erzeugung der Gewichte der ersten Welle, werden diese Rohgewichte einer Anpassung an die aktuellen Ränder ausgehend vom Mikrozensus unterworfen. Die berücksichtigten Variablen sind Region, Alter, Geschlecht, Haushaltsgröße und Nationalität (Pischner 2007a, 2007b).

Wie einleitend beschrieben werden den in den Haushalten beobachteten Personen ebenfalls zunächst die Kehrwerte der Bleibewahrscheinlichkeiten zugeordnet. Die Kehrwerte bilden, nach Korrektur für den Zweitwohnsitz, wiederum die Ausgangsbasis für eine Randanpassung an die Alterstruktur der Personen in Privathaushalten am Hauptwohnsitz und zwar getrennt nach alten und neuen Ländern.

Gewichtungsvariablen in der Datenweitergabe

Tabelle 4 gibt einen Überblick über die mit dem SOEP ausgelieferten Gewichte. Dabei werden die Gewichte mit $xHRFy bezeichnet. Es bedeuten:

- $ = Wellenkennzeichen A,B,...,W für die Jahre 1984, 1985, ..., 2006.
- x = Unterscheidung nach Haushalten (x = H) und Personen (x = P).
- HRF kennzeichnet die Variable als Hochrechnungsfaktor.

y = eine Zusatzkennung, die die Art des Gewichts beschreibt:

Tabelle 4

Gewichte für Privathaushalte im Sozio-oekonomischen Panel für die Wellen A–W (1984–2006)

In den Gewichten enthaltene Stichproben und Eckdaten

Welle $	Jahr	$HHRF	$HHRF1	$HHRFALL	$HHRFy	Zahl der Privathaushalte Standard-Stichprobe	Zahl sämtlicher Haushalte	Hochgerechnete Privathaushalte in der Grundgesamtheit in 1 000
A	1984	AB	= 0	.	.	5853	5921	26076
B	1985	AB	AB	.	.	5238	5322	26367
C	1986	AB	AB	.	.	4991	5090	26739
D	1987	AB	AB	.	.	4920	5026	27006
E	1988	AB	AB	.	.	4719	4814	27402
F	1989	AB	AB	.	.	4602	4690	27793
G	1990	ABC	ABC	.	.	6722	6819	34848
H	1991	ABC	ABC	.	.	6581	6699	35256
I	1992	ABC	ABC	.	.	6564	6665	35700
J	1993	ABC	ABC	.	.	6537	6637	36230
K	1994	ABC	ABC	.	.	6459	6559	36695
L	1995	ABCD	ABC	.	.	6656	6768	36938
M	1996	ABCD	ABCD	.	D	6591	6698	37281
N	1997	ABCD	ABCD	.	D	6508	6617	37456
O	1998	ABCDE	ABCD	.	D	7359	7486	37532
P	1999	ABCDE	ABCDE	.	D	7905	7215	37794
Q	2000	ABCDEF	ABCDE	.	D und F	12905	13.078	38123
R	2001	ABCDEF	ABCDEF	.	D	11667	11783	38455
S	2002	ABCDEF	ABCDEF	ABCDEFG	D und G	11202	12308	38720
T	2003	ABCDEF	ABCDEF	ABCDEFG	D und G	10987	11910	38945
U	2004	ABCDEF	ABCDEF	ABCDEFG	D und G	10641	11642	39121
V	2005	ABCDEF	ABCDEF	ABCDEFG	D und G	10321	11294	39178
W	2006	ABCDEFH	ABCDEF	ABCDEFGH	D und G	11409	12361	39178

1 y = D: Zuwanderer, y = F: Ergänzungsstichprobe 2000, y = G: Hocheinkommensstichprobe.
2 Nur Haushalte mit positivem Gewicht.
3 Vorläufig: Die Daten basieren auf dem Mikrozensus 2005.

Quellen: Das Sozio-oekonomische Panel (Wellen A–W), eigene Darstellung.

- y = <leer>, also nicht besetzt, bezeichnet Standardhochrechnungsfaktoren. Standardgewichte umfassen sämtliche Teilstichproben des SOEP mit Ausnahme der Hocheinkommensstichprobe G und einzelne nach dem Schneeballsystem gezogene Haushalte in Sample D. Diese Gewichte sind für sämtliche Wellen verfügbar.
- y = 1 bezeichnet modifizierte Standardhochrechnungsfaktoren, in denen die Untersuchungseinheiten im Startjahr einer Teilstichprobe ein Gewicht von null erhalten (siehe Frick et al. 2006).
- y = ALL umfasst sämtliche erhobene Teilstichproben des SOEP, inklusive der Hocheinkommensstichprobe G.

- y = D kennzeichnet die isolierte Gewichtungsvariable für die gesamte Zuwanderer-Stichprobe D, inklusive der im Schneeballsystem gezogenen Haushalte.
- y = G kennzeichnet die Gewichtungsvariable für die Hocheinkommensstichprobe G.

In Tabelle 5 sind diese Zusammenhänge noch einmal tabellarisch zusammenfassend dargestellt. Die ausgewiesenen Fallzahlen und Ecksummen der Gewichte beziehen sich nur auf die Privathaushalte; Anstaltshaushalte, die auch – wie in den meisten amtlichen Stichproben – nicht repräsentativ im SOEP erfasst sind, bleiben in der Darstellung unberücksichtigt.

4.2 Varianzschätzung und effektive Fallzahlen

Die Bestimmung von Standardfehlern eines Parameters weicht im Fall komplexer Surveys wie dem SOEP häufig von der üblichen Varianzschätzung in einfachen Zufallsstichproben ab (vgl. z. B. Kalton 1977, Rust 1985, Wolter 1985). Seien es wie in den zuvor erwähnten Beispielen die Summen oder Anteile in einer endlichen konkreten Grundgesamtheit (etwa der in Privathaushalten lebenden Bevölkerung Deutschlands im Jahr 2007) oder aber ein Modell für Zusammenhänge in einer abstrakteren Grundgesamtheit (etwa der Effekt der Variable „Arbeitserfahrung“ auf eine abhängige Variable „logarithmiertes Einkommen“).

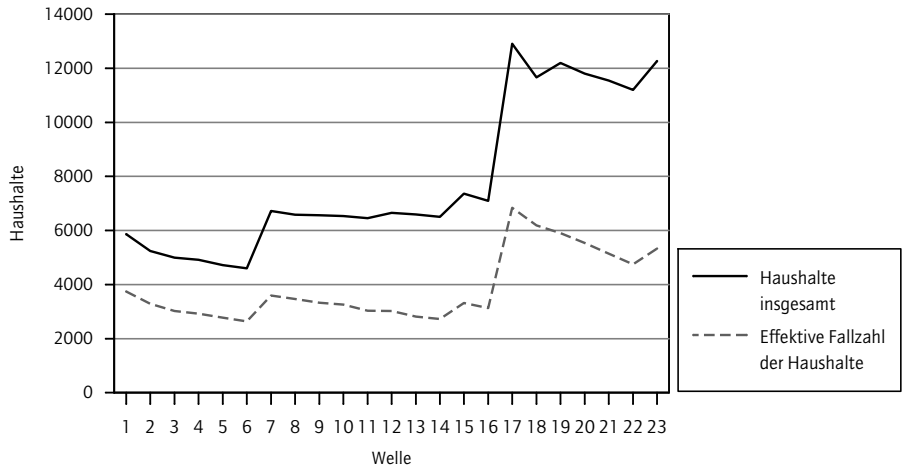
Die Verwendung der Längs- und Querschnittsgewichtung des SOEP dient zur Verhinderung von Verzerrungen der Schätzung durch das zuvor beschriebene mehrstufige und stratifizierte Ziehungsdesign, sowie durch die unvollständige Stichprobenausschöpfung und die Ausfälle von Analyseeinheiten im Längsschnitt. Gleichzeitig führt dieses Vorgehen jedoch bei korrekter Berücksichtigung der Komplexität des Designs zu Effizienzverlusten der Schätzung. Die effektive Stichrobengröße, die die Höhe des Standardfehlers bestimmt, entspricht dem Verhältnis aus tatsächlicher Anzahl an Beobachtungen und dem sogenannten Designeffekt. Dieser Designeffekt wiederum ist das Produkt zweier Komponenten, der Variation der Gewichte und der Klumpung der Stichprobe (Kish 1965 und 1995, Cochran 1977). Das heißt, dass bei gewichteten Analysen die effektive Stichprobengröße des SOEP umso kleiner ist und der Standardfehler einer Schätzung umso größer wird, je höher die Varianz der jeweiligen Gewichte ist beziehungsweise je stärker die Klumpung der jeweils verwendeten Teilstichprobe.

Nach 23 Wellen des SOEP liegen für 23 932 Haushalte (47 439 Personen) insgesamt 185 899 Haushaltsinterviews und 360 344 Personeninterviews vor. Jeder Haushalt wurde somit im Durchschnitt beinahe acht Mal befragt und immerhin 1 535 Haushalte (2 716 Personen) haben an allen Wellen teilgenommen, d. h. für sie liegen jeweils 23 konsekutive Haushalts- beziehungsweise Personeninterviews vor. In Abbildung 1. ist die Entwicklung der Fallzahlen abgetragen. Die obere Linie mit roten Quadraten zeigt, dass in den letzten Jahren circa 12 000 Haushalte in jeder Welle befragt wurden. Die zweite Zeitreihe darunter zeigt dagegen die effektiven Fallzahlen. Diese liegen mit Werten um 5 000 deutlich unter jenen der Zahl der faktisch erhobenen Haushalte. So ergab sich für die Welle 23 ein Quotient von 0,42; dies ist ein Maß für die Effizienz der SOEP-Stichprobe relativ zu einer reinen Zufallsstichprobe mit gleicher Anzahl an Beobachtungen.

Alternative Strategien zur Bestimmung von Standardfehlern in komplexen Surveys bestehen in Methoden des Resampling wie Jackknife und Bootstrapping (z. B. Rao und Wu

Abbildung 1

Entwicklung der Fallzahlen im SOEP nach 23 Wellen



Quelle: Das Sozio-oekonomische Panel, Wellen 1–23; eigene Berechnungen.

1988). Da das SOEP aus verschiedenen Teilstichproben mit jeweils recht unterschiedlichen Ziehungsdesigns besteht und aus wellenspezifischen Datensätzen mit unterschiedlichen Varianzen der jeweiligen Gewichtungsfaktoren, lässt sich keine für alle möglichen Anwendungen einheitliche Varianzschätzung vorgeben. Weil darüber hinaus für unterschiedliche Schätzer unterschiedliche Techniken vorgeschlagen wurden, sind diese abhängig von den Merkmalen der jeweils verwendeten Stichprobe und abhängig von der jeweiligen Analyse zu wählen (Särndal, Swensson und Wretman 1992). Eine vergleichsweise einfache Vorgehensweise über sogenannte „Random Groups" wird vom SOEP unterstützt. Dabei wird eine Stichprobe nachträglich in mehrere (möglichst unabhängige) Teilsstichproben unterteilt, sodass jede Teilstichprobe als Realisation des ursprünglichen Ziehungsverfahrens mit kleinerem Stichprobenumfang angesehen werden kann. Mit dem SOEP wird im File DESIGN die Variable RGROUP ausgeliefert, die die Unterteilung in acht Unterstichproben ermöglicht. Die Berechnung von Varianzen und Konfidenzintervallen auf Basis der *Random Groups* ist in Rendtel (1995) für das SOEP und allgemein in Wolter (1985) beschrieben.

4.3 Kompensation fehlender Werte

Die im Abschnitte 4.1 beschriebenen Gewichte können zur Kompensation fehlender Einheiten verwendet werden. Neben Einheiten, die gar nicht beobachtet werden (*Unit-Non-Response*), entweder weil sie sich von Beginn an weigern an der Befragung teilzunehmen oder weil sie zu einem späteren Zeitpunkt ausfallen, werden bei ansonsten auskunftsbereiten Einheiten häufig einzelne Variablen nicht beobachtet (*Item-Non-Response*). Ein möglicher Ansatz zur Kompensation ist die Ersetzung fehlender Werte oder „Items" durch „plausible" Werte. Diese Technik wird auch als Imputation bezeichnet.

Sogenannte Single-Imputations-Techniken ersetzen jeden fehlenden durch einen plausiblen imputierten Wert. Von einem statistischen Standpunkt aus betrachtet, wird der nicht beobachtete Wert durch einen plausiblen Wert geschätzt. Die entsprechenden (statistischen) Modelle zur Erzeugung der Schätzwerte sind sorgfältig zu formulieren und zu schätzen, denn von der Qualität der Schätzwerte hängt die Qualität der letztlich interessierenden Schlussfolgerungen ab. Ein Problem der Imputationstechniken ist, neben der Wahl eines geeigneten Vorhersagemodells, die Frage, wie mit der Unsicherheit in den Schätzwerten oder Vorhersagen umzugehen ist. Bei Single-Imputation-Techniken sind die Standardfehler der Schätzer, auf denen die inhaltlich interessierenden Aussagen etwa über eine endliche Grundgesamtheit basieren, anzupassen. Dies ist nicht trivial und entsprechende Vorschläge existieren nur für einzelne, vergleichsweise einfache Analysen und sind bisher in den Standardstatistikpaketen nicht verfügbar.

Die Methode der mehrfachen oder multiplen Imputation basiert dagegen auf der Idee, jeden fehlenden Wert durch mehrere Vorhersagen dieses Wertes zu ersetzen. In der Variation dieser Prädiktionen soll sich die gesamte, mit der Prädiktion verknüpfte Unsicherheit widerspiegeln. Sind die Imputationen geeignet im Sinne von Rubin (1987, 1996), dann lassen sich mehrfach imputierte Datensätze mit Standardanalyseverfahren auswerten. Die Schätzergebnisse basierend auf den mehrfach ergänzten Datensätzen können anschließend auf einfache Weise kombiniert und wie üblich interpretiert werden. Zur Erzeugung multipler Imputationen steht inzwischen eine Reihe von zum Teil frei verfügbaren Programmen zur Verfügung.[14] Auch die Auswertung solcher Datensätze wird zunehmend durch aktuelle Statistikpakete unterstützt. Allerdings ist die Erzeugung geeigneter multipler Imputationen in komplexen Datensätzen wie dem SOEP mit zurzeit existierenden Programmen kaum möglich. Die weitreichende multiple Imputation fehlender Werte im SOEP wird eine Aufgabe der näheren Zukunft sein; erste Ansätze liegen bereits auf Basis der Imputation fehlender Vermögensangaben aus 2002 (Frick und Grabka 2007) und ab der Datenweitergabe 2008 auch für das monatliche Haushaltsnettoeinkommen vor.

5 Datenstruktur und Generierung

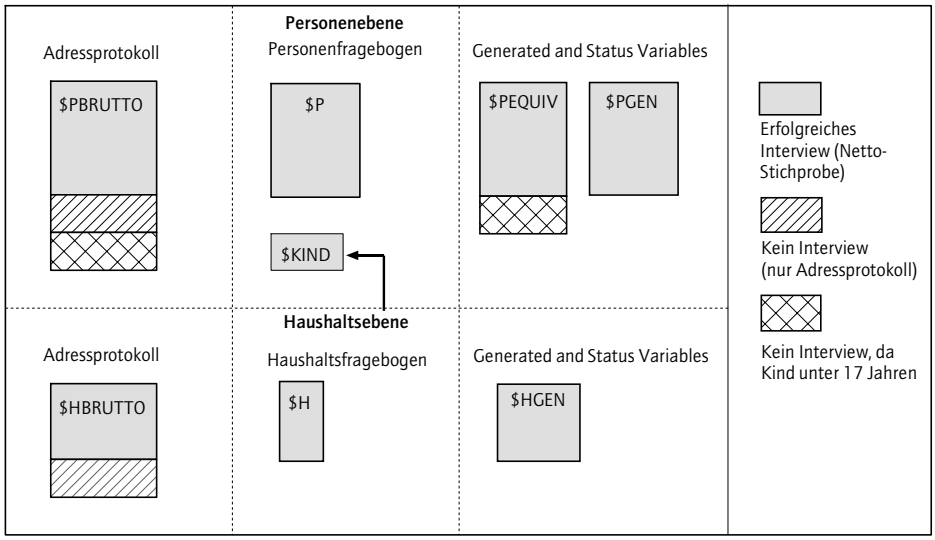
5.1 Allgemeine Datenstruktur

Der SOEP-Datensatz besteht mit der Datenweitergabe 2006 aus 282 unterschiedlichen Datensätzen (Files), die zusammen über vier Millionen Beobachtungen enthalten und in denen über 37 000 Variablen gespeichert sind. Im Grundsatz werden alle direkt erhobenen Personen- und Haushaltsdaten ($P und $H)[15] sowie die damit korrespondierenden Feldinformationen[16] ($PBRUTTO und $HBRUTTO) Jahr für Jahr als Querschnittsdatensätze abgelegt und als solche auch nahezu unverändert, allerdings nach detaillierten Datenprüfungen, an die Nutzer weitergegeben (vgl. oben Kapitel 2). Eine vereinfachte Übersicht zur Datenstruktur im Querschnitt findet sich in Abbildung 2. Von der allgemeinen Wei-

14 Frei verfügbare Programme sind z. B. IVEware oder MICE. In den letzten Jahren sind aber auch in den gängigen Statistikpaketen Stata und SAS-Prozeduren zur multiplen Imputation integriert worden.
15 Das Dollar-Zeichen ($) steht hierbei für den wellenspezifischen Präfix.
16 Feldinformationen liegen für alle Haushalte und Personen vor, die zur wellenspezifischen Bruttopopulation gehören. Diese setzt sich zusammen aus der Vorjahrs-Population abzüglich der verstorbenen, ausgewanderten und jenen Personen, die bereits im letzten Jahr endgültig verweigert haben. Vergrößert wird die Bruttopopulation durch seit der Vorwelle Neugeborene sowie sonstige Personen, die in bestehende SOEP-Befragungshaushalte eingezogen sind bzw. in durch Abspaltung neu entstandenen Befragungshaushalten leben.

Abbildung 2

Übersicht Datenstruktur



$: Wellenspezifikation: A, B, C … X für Dateinamen.

Quellen: Das Sozio-oekonomische Panel (Wellen A–W), eigene Darstellung.

tergabe ausgeschlossen sind lediglich Klartextangaben (z. B. Berufe und Vornamen) und kleinräumige Regional- oder Nachbarschaftsinformationen, um Re-Identifikationsmöglichkeiten von vornherein zu minimieren. Derartige Regional-Informationen können unter spezifischen Datenschutzvorkehrungen vor Ort am DIW Berlin beziehungsweise im Rahmen von Datenfernabfragen genutzt werden (siehe Abschnitt 7).

Informationen zu Kindergarten und Schulbesuch der noch nicht persönlich befragten Kindern – das Befragungsalter beginnt nach dem 16. Lebensjahr – werden auf Haushaltsebene erfasst und anhand der für alle Personen im Haushalt bereitgestellten Personeneinträge ($PBRUTTO) als disaggregierte Kinderinformationen ebenfalls Jahr für Jahr personenbezogen ($KIND) abgelegt. Diese Daten sind insofern ein Sonderfall, da das SOEP im Regelfall keine Proxy-Interviews zulässt.

Die Variablenbezeichnung verweist im SOEP bei den $P und $H Files direkt auf die Fragebogennummer und stellt so den direkten Bezug zum Erhebungsinstrument sicher. Darüber hinaus werden aus den jährlich abgelegten Personen- und Haushaltsdaten über die Zeit vergleichbare Datensätze generiert ($PGEN, $HGEN, $PEQUIV, $PKAL). Die Variablen dieser Datensätze sind mit über die Zeit einheitlichen Namen gekennzeichnet.

Die einfache Rechteckstruktur der jahresbezogenen Daten (Untersuchungseinheiten x Variablen [N x V]) gewährleistet die einfache und direkte Kommunikation zwischen Datenproduzenten und Datennutzern. Die direkt erhobenen jahresbezogenen Originaldaten (in den Dateien $P und $H) bleiben im Regelfall unverändert erhalten, wohingegen die daraus abgeleiteten beziehungsweise generierten Daten infolge verbesserter und längsschnittkon-

sistenter Algorithmen (inklusive Imputationen) laufend auch rückwirkend überarbeitet und modifiziert werden.

Zur Unterstützung von Längsschnittanalysen werden zwei weitere zentrale Datensätze bereitgestellt, die jeweils die Gesamtheit aller im SOEP je erfassten Personen beziehungsweise Haushalte enthalten. Für die Personenebene wird dieser Datensatz mit PPFAD bezeichnet und für die Haushaltsebene HPFAD. Beide Datensätze enthalten weitere zentrale Variablen zur Stichprobenabgrenzung wie z. B. Befragungsstatus, Geschlecht oder Region.

Unter anderem aufgrund der begrenzten Anzahl geschulter Interviewer (es werden derzeit jährlich etwa 500 Interviewer eingesetzt), können die SOEP-Daten nicht an einem Stichtag (beziehungsweise für einen einzigen Berichtstag) erhoben werden. Die Feldarbeit zieht sich daher über mehrere Monate hin. Der Einsatz eines festen und intensiv geschulten Interviewerstabes fördert entscheidend die Stabilität der Untersuchungspopulation, mit anderen Worten, der Anteil der erfolgreich wiederholten Teilnahme beziehungsweise Befragung kann so möglichst hoch gehalten werden (vgl. den Beitrag von Rosenbladt in diesem Heft).[17] Aufgrund dieser saisonalen Effekte kann das SOEP auch für unterjährige Analysen anhand der abgelegten und bereitgestellten Informationen zum Befragungsmonat und -tag genutzt werden. Freilich stehen aufgrund des Schwerpunkts der Feldarbeit im ersten Quartal nur für dieses pro Monat oder Woche (teilweise sogar pro Tag) genügend Stichtags- beziehungsweise Stichwochenfälle zur Verfügung. Im Allgemeinen sind im Laufe des März bereits rund 50 % aller Interviews erfolgreich durchgeführt. Beispielhafte Analysen, die diesen zeitlich variierenden Bezug der SOEP-Daten nutzen, finden sich für Fragen zu „Sorgen“ auf Basis einer Verknüpfung der SOEP-Daten mit aktuellen Medienthemen (Dittmann 2005) oder zum Einfluss der Tschernobyl-Katastrophe auf die Lebenszufriedenheit der Befragten (Berger 2007).

Tabelle 5

Zahl der Fälle mit vollständigen monatlichen Erwerbskalendarien nach Stichproben und Zahl der erfassten Monate für die ersten 23 Wellen

Erfasste Dauer Stichproben	276 Monate	Mindestens 240 Monate	Mindestens 180 Monate	Mindestens 120 Monate	Mindestens 60 Monate	Mindestens 12 Monate
Insgesamt	2 704	3 611	7 216	10 945	25 758	43 593
A	2 278	2 969	4 166	5 797	8 174	12 076
B	426	642	1 106	1 817	2 789	4 440
C	.	.	1 944	2 790	4 028	5 869
D	.	.	.	541	856	1 366
E	.	.	.	.	1 309	2 203
F	.	.	.	.	7 110	12 184
G	.	.	.	.	1 492	2 839
H	.	.	.	.	.	2 616

Quellen: SOEP (Wellen A–W); eigene Berechnungen.

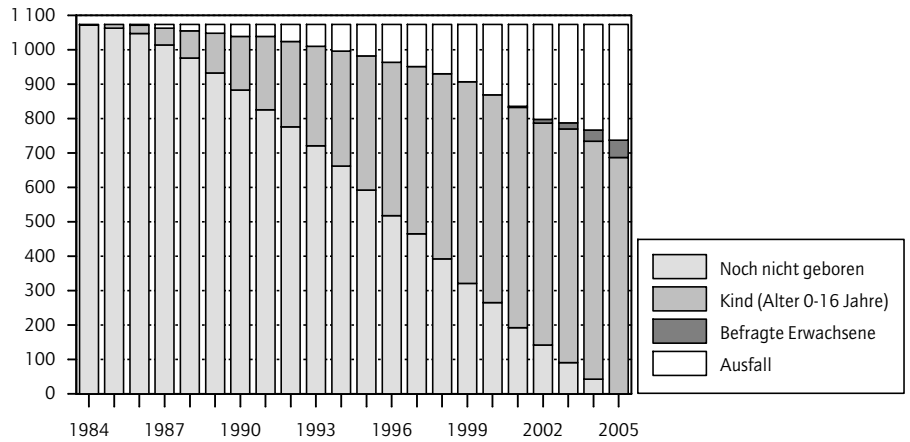
17 So weit möglich werden die jeweiligen Haushalte immer durch den gleichen Interviewer wieder befragt..

Freilich werden in der SOEP-Befragung nicht nur Daten mit Bezug zu einem Stichtag beziehungsweise zu einer normalen Woche erhoben, sondern auch retrospektiv über das vergangene Kalenderjahr. Darauf aufbauend werden differenzierte Jahreseinkommen konstruiert (siehe den folgenden Abschnitt 5.2). Außerdem werden auf Basis von „Kalender-Angaben“ zu Aktivitäten, die typischerweise auf Monatsbasis erfasst werden (z. B. Erwerbstätigkeit, Ausbildung, Wehr- oder Zivildienst) sogenannte „Spells“ erzeugt, die die zeitliche Lage und Dauer dieser Aktivitäten beschreiben und für Verweildaueranalysen geeignet sind (z. B. zur Dauer von Arbeitslosigkeitsperioden). Im SOEP stehen über den kompletten Erhebungszeitraum von Januar 1983 (erhoben in der ersten Welle im Frühjahr 1984 für das vorangegangene Kalenderjahr) bis Dezember 2006 für maximal 276 Kalendermonate Aktivitätsangaben (Spelldaten) zur Verfügung (vgl. Tabelle 5). Für circa 2 700 Personen liegen diese Daten lückenlos für den kompletten Zeitraum vor. Für fast 11 000 Personen liegen immerhin Daten über mindestens 120 Monate (10 Jahre) vor und für mehr als 25 000 Personen gibt es Kalendarien, die wenigstens 60 Monate umfassen. Eine Übersicht über die Anzahl der unzensierten Spells findet sich in Goebel et al. 2008.

Die besondere und mit der Laufzeit eines Haushaltspanels steigende Aussagekraft für die Analyse von Ereignissen und Lebensläufen wird z. B. auch an der Geburt von „SOEP-Enkeln“ deutlich (Abbildung 3). Dies sind im SOEP geborene Kinder, deren Eltern als Befragte am SOEP teilnahmen und von denen auch mindestens ein Großelternteil in der SOEP-Stichprobe enthalten ist. Abbildung 3 zeigt, dass seit 1984 mehr als 1 000 Kinder im SOEP erfasst wurden, deren Eltern und Großeltern schon selbst an der SOEP-Befragung teilgenommen haben; einige sind nach 16 Jahren (2001) bereits selbst ins Befragungsalter „hineingewachsen“, sodass hier prospektiv erhobene Informationen von drei Generationen vorliegen. Im Jahr 2005 sind noch mehr als zwei Drittel der erfassten Enkel im Erhebungsbestand.

Abbildung 3

Enkelkinder im SOEP



Quelle: SOEP (Wellen A–V); eigene Berechnungen.

5.2 Generierte Einkommen

Ein wichtiger Befragungsinhalt des SOEP, der in jeder Welle erfasst wird, ist die Einkommenssituation der Personen in Privathaushalten. Für längsschnittliche Analysen ist dabei die intertemporale Vergleichbarkeit der entsprechend aufbereiteten Einkommensvariablen im SOEP von zentraler Bedeutung. Jedoch mussten im Verlauf der vergangenen nahezu 25 Jahre die zu Grunde liegenden Fragen verschiedentlich auf geänderte Rahmenbedingungen angepasst werden oder wurden aufgrund von konzeptionellen Neuausrichtungen variiert.

Als Beispiel seien hier die Einkommen aus gesetzlichen, betrieblichen oder privaten Renten genannt, die in den ersten beiden Jahren des SOEP nicht einzeln, sondern nur als zu aggregierte Information im Sinne der Summe aller eigenen oder abgeleiteten Witwenbeziehungsweise Waisenrenten erfragt wurden. Die detaillierte Erfassung der Renteneinkommen, die nun explizit zwischen den einzelnen Rententrägern (GRV, Knappschaft, Kriegsopferversorgung etc.) unterscheidet, wurde ab der dritten Welle des SOEP (1986) vollzogen.

Eine einschneidende Neukonzeption der Erfassung der Einkommen im SOEP erfolgte mit der Erhebung des Jahres 1995. Bis dahin wurde mit Ausnahme der aktuellen Erwerbseinkommen und des Haushaltsnettoeinkommens im Befragungsmonat ausschließlich Einkommen retrospektiv für das vorangegangene Kalenderjahr erfragt. Für verschiedenste Fragestellungen ist aber ein Bezug auf vergangene Einkommen unzureichend. Seit der Erhebungswelle 1995 wird daher im Haushalts- und Personenfragebogen sowohl Einkommen retrospektiv für das vergangene Kalenderjahr als auch im laufenden Befragungsmonat erhoben. Um die Befragungsdauer nicht zu verlängern, wurde als Kompensation für diese zusätzlichen Fragen die dahin ebenfalls erhobene exakte zeitliche Lage der Einkommen im Vorjahr, in Form eines Einkommenskalendariums, aufgegeben.

Aufbauend auf diesen Originalinformationen des SOEP wurden in den vergangenen Jahren eine Reihe nutzerfreundlich aufbereiteter Variablen auf Personen- und Haushaltsebene generiert. Ein wesentlicher Vorteil der generierten Einkommensinformationen im SOEP ist die vollständige Imputation fehlender Antwortangaben aufgrund von *Item-Non-Response* (vgl. Grabka und Frick 2003, Frick und Grabka 2005). Zudem werden die generierten Einkommen einheitlich in Euro ausgewiesen, um Fehler durch eine fehlende Konvertierung von Originalinformationen, die bis 2001 in DM erhoben wurden, zu vermeiden (Grabka 2007).

Das aktuelle monatliche Haushaltsnettoeinkommen ist in den Datensätzen $HGEN mit einem einheitlichen Variablennamen (HINC$$) und in Euro durchgehend ab 1984 verfügbar. Mit der Datenweitergabe 2008 wird es auch eine multiple imputierte Version (vgl. Abschnitt 4.3) dieser zentralen Variable ab 1995 geben.

Eine besondere Stellung innerhalb der generierten Einkommensinformation nehmen die generierten Jahreseinkommen des Vorjahres ein, da diese explizit auch für den internationalen Vergleich konzipiert wurden. Die Definition dieser Einkommenskonstrukte orientiert sich an den Empfehlungen der Canberra-Group on Household Income Statistics (Canberra Group 2001). Die Generierung dieser Einkommen beruht auf einer Kooperation des SOEP mit Richard Burkhauser von der Syracuse University (heute Cornell University). Aus die-

ser Kooperation entstand Ende 1994 die erste Version des „PSID-GSOEP Equivalent Data File“ als empirische Basis für eine Reihe vergleichender (Einkommens-)Analysen für die USA und Deutschland (vgl z. B. Burkhauser, Frick und Schwarze 1997). Im Rahmen der Vorarbeiten zu diesem bis dato neuen Datenfile hatte Johannes Schwarze ein Steuer- und Sozialabgabensimulationsmodell für die Daten des SOEP entwickelt, das das verfügbare Haushaltsnettoeinkommen generierte (vgl. Schwarze 1995). Der PSID-GSOEP Equivalent File ist im Verlauf der Jahre um weitere Panelstudien ergänzt worden. Der heute als „Cross-National-Equivalent-File“ (CNEF) bekannte Datensatz umfasst derzeit neben den Panelstudien der USA und Deutschland auch solche aus Kanada, Großbritannien, Australien und der Schweiz (für eine Einführung in den CNEF und die entsprechende Variablenliste siehe Frick, Jenkins, Lillard, Lipps und Wooden 2007 und in diesem Heft).

Der deutsche Teil des CNEF ist in Form der wellenspezifischen $PEQUIV-Files Bestandteil der regelmäßigen Datendistribution (siehe Grabka 2007). Als besonderer Service für die Nutzer des SOEP beinhalten die $PEQUIV-Files neben den aggregierten Einkommensinformationen aus dem CNEF auch detaillierte Variablen separat für jede Einkommensart.

Eine Besonderheit in der Historie der $PEQUIV-Files besteht darin, dass in diesem Datenbestand bereits frühzeitig dem Aspekt von nicht-monetären Einkommenskomponenten Rechnung getragen wurde. So wurde bereits mit der ersten Version des PSID-GSOEP Equivalent File der Mietwert selbst genutzten Wohneigentums (*imputed rent*) geschätzt. Diese fiktive Einkommenskomponente wurde in den vergangenen Jahren mehrfach aktualisiert und die zuvor verwendete Approximation auf Basis von Selbstangaben des Haushaltsvorstandes durch ein regressionsbasiertes Schätzverfahren ersetzt. Seit der Datenweitergabe des Jahres 2006 berücksichtigen diese fiktiven Einkommensvorteile nicht mehr allein Eigentümerhaushalte (Frick und Grabka 2003), sondern – in Einklang mit den Empfehlungen von Eurostat (vgl. Frick, Göbel und Grabka 2007) – auch fiktive Einkommensvorteile subventionierter Mieter (Sozialbauwohnung sowie Mieter in mietfrei oder verbilligt überlassenem Wohnraum).

5.3 Datenformate zur Speicherung von Längsschnittinformationen

Liegen verschiedene Personen- oder Haushaltsinformationen über die Zeit vor, so handelt es sich um eine dreidimensionale Datenstruktur (Untersuchungseinheiten x Variablen x Zeit [N x V x T]). Der zeitliche Bezug kann datentechnisch auf unterschiedliche Weise operationalisiert werden – im SOEP finden drei unterschiedliche Formate Anwendung: wide-format, long-format und spells (vgl. Tabelle 6).

Beim *wide-format* werden zu jedem Jahr die Zahl der Variablen um eine weitere jahresspezifische Variable zeilenweise ergänzt, d.h., der Datensatz wird „verbreitert“. Diese Form erlaubt die direkte Anwendung von Datenmodifikationen und Rechenoperationen der Variablen im Längsschnitt. Die Population umfasst bei diesem Format die Gesamtheit aller im Zeitraum erfassten Einheiten und wächst in jedem Jahr nur um die Zahl der erstmalig hinzugekommenen Personen oder Haushalte an (NT). Typische Anwendungsbeispiele für dieses Datenformat im SOEP sind die Datensätze PPFAD und HPFAD, die jeweils die Gesamtheit aller jemals kontaktierten Personen beziehungsweise Haushalte umfassen.

Tabelle 6

Datenstruktur und Zeitbezug

Zeitbezug	Population	Population x Zeitbezug
Unverbundene Querschnitte	Pers./HHe zum Zeitpunkt t	[N x V] t1 … tn
WIDE-Format	Pers./HHe im Zeitraum T	NT x Vt1 … Vtn
LONG-Format	Pers./HHe gepoolt; Beobachtungs-Einheiten im Zeitraum T	NT x V
SPELL	Zustand (je Pers./HH) im Zeitraum T	NTZ x Ve; Te1, Ten

Quelle: Eigene Darstellung.

Beim *long-format* werden im Unterschied dazu die Variablen nicht einzeln jahresweise ergänzt, sondern über die Zeit „gepoolt". Die jahresspezifischen Informationen ergeben sich so als einzelne Schichten im über die Zeit kumulativ abgelegten Datenbestand. Der Variablenname bezieht sich in diesem Fall nicht mehr nur auf das einzelne Jahr, sondern auf den gesamten Zeitraum; dazu müssen die Daten zuvor vergleichbar aufbereitet sein. Bei dieser Darstellung ist die Zeit (Erhebungsjahr) als zusätzliche Variable erforderlich. Die Zahl der Variablen bleibt jedoch bei diesem Format immer gleich, wohingegen die Population sich jeweils um die gesamte Zahl der pro Jahr erfassten Untersuchungseinheiten erhöht, d.h., der Datenbestand wird um die neuen Interviews „verlängert" (N·t). Typische Anwendungen im SOEP sind die intern gehaltenen Files mit Klarschriftangaben zu Bildungs-, Berufs- und Branchenvercodungen (Berufe, Branche) sowie neuerdings die aufbereiteten Angaben zur Gesundheit (HEALTH); aber auch jahresweise erhobene Informationen aus DJ[18], Greifkraftmessung oder auch zu Vermögenssitutation werden in diesem Format aufbereitet.

Bei der Abbildung der Daten im spell-format werden Zustände oder Ereignisse (Z) gezählt. In derselben Weise, wie jedem Haushalt ein oder mehrere Personen zugeordnet werden, werden pro Person unterschiedliche Zustände zugewiesen. Die zeitliche Dauer wird durch die Angabe des Beginns und Endes für jeden Zustand kontinuierlich erfasst. Dieser Datentyp unterstützt insbesondere auf kontinuierliche Angaben abhebende ereignisanalytische Verfahren, die den Wechsel von Zuständen analysieren. Anwendungsbeispiele im SOEP sind soziodemografische Informationen zum Erwerbsverlauf (ARTKALEN, PBIOSPE) und Familienstandsänderungen (BIOMARSY, BIOMARSM).

Die derzeitige Datenstruktur und -distribution des SOEP umfasst so einerseits die Weitergabe der möglichst einfach strukturierten jährlichen Querschnittsfiles sowie andererseits eine Reihe von zeitraum-bezogenen Daten mit je nach Anwendungsbezug unterschiedlichen Formaten (siehe Übersicht in Tabelle 5.2): Auf der Ebene der Ursprungshaushalte werden – soweit verfügbar – die Bruttofiles zur Stichprobenziehung sowie Design-Informationen (z. B. zur Stichprobenziehungswahrscheinlichkeit) bereitgestellt. Auf Haushaltsebene werden wellenübergreifend HPFAD und HHRF, regionalspezifische Informationen sowie (bis 1995) Verlaufsinformationen zu Sozialhilfekarrieren (SOZKALEN) bereitgestellt. Personendaten werden auf verschiedenen Erhebungsebenen (alle Haushaltsmitglieder in den

18 DJ steht für „Denksport & Jugend", den Kognitionstest für Jugendliche, der als Ergänzung zum Jugendfragebogen seit 2005 erhoben wird und ab 2009 in die Datenweitergabe kommen soll.

Datensätzen $PBRUTTO oder $PEQUIV; nur Befragungspersonen in $PGEN; Kinder in $KIND, etc.) mit zeitübergreifenden Informationen ausgeliefert. Ereignisse – die kleinste Erhebungseinheit im SOEP – werden sowohl jahres- als auch monatsweise geführt.[19]

6 Dokumentation

Ein komplexer Datensatz wie das SOEP wird für externe Nutzer erst dann brauchbar, wenn eine extensive Dokumentation der Daten vorliegt. Für das SOEP ist die komplette Dokumentation über die Internetseite *www.diw.de/soep* frei zugänglich. Diese Seite beinhaltet unter anderem das „SOEP Desktop Companion" als zentrale Einführung in das SOEP (Haisken-DeNew und Frick 2005), eine genaue Dokumentation aller generierten Variablen inklusive der nutzerfreundlich aufbereiteten Biographiedaten, die eingesetzten Fragebögen, Methodenberichte des Feldinstituts TNS Infratest Sozialforschung, die Datenbank SOEPlit mit bibliographischen Hinweise auf SOEP-basierte Publikationen, und insbesondere die interaktive Webanwendung SOEPinfo.

Die Übersicht über die im SOEP über die Jahre hinweg insgesamt verfügbaren Informationen, die Itemkorrespondenzen der im Zeitverlauf erfassten Variablen sowie interaktiv erstellbare Syntax-Codes für die Statistikpakete SAS, SPSS und Stata werden über das Dateninformationssystem SOEPinfo bereitgestellt (vgl. oben Kapitel 2). Infolge der unterschiedlichen Erhebungseinheiten, der vielfältigen Erhebungsinstrumente sowie der im Laufe der Jahre kumulierten Vielfalt an Informationen ist die Komplexität der SOEP-Daten nicht nur für neue Nutzer erheblich angewachsen. SOEPinfo ist somit ein wichtiges Instrument bei der Dokumentation der SOEP-Daten und für jeden über die Internetadresse *http://panel.gsoep.de* zugänglich.

Während der Zugang zu den SOEP-Mikrodaten aus Datenschutzgründen nur im Rahmen eines Datenweitergabevertrages möglich ist, bildet SOEPinfo eine eigene Datenbank mit Informationen über die originalen SOEP-Daten. Es ist also eine Art Metadatenbank und ist datenschutzrechtlich unbedenklich, da keinerlei Zugriff auf die Mikrodaten möglich ist.

SOEPinfo ist heute aus Nutzersicht hauptsächlich eine Webanwendung, mit der die oben genannten Metadaten interaktiv erfragt und kombiniert werden können. Zur Erlangung der Informationen werden verschiedene Möglichkeiten geboten. So kann nach Variablennamen direkt (auch mit Mustern) gesucht werden. Die gefundenen Variablennamen können in einer Liste, einem sogenannten Warenkorb (Basket) zwischengespeichert und für weitere Aktionen, einzeln oder zusammen, verwendet werden. Variablenkorrespondenzen können über eine Themenliste, eine Suche über die Labels oder über den Warenkorb erschlossen werden. Ein weiterer wichtiger Zugang sind die Original-Fragebögen, die mit Variablennamen und direkt „klickbaren" Links angereichert wurden. Aus einem Fragebogen können eine oder mehrere Variablen direkt dem Warenkorb hinzugefügt werden oder aus dem Warenkorb kann direkt zu einer Variablen in einem Fragebogen gesprungen werden.

19 Für die Verarbeitung der im Spell-Format abgelegten Daten liegen spezielle Hilfsprogramme zur deutlichen Vereinfachung der Verarbeitung vor, vgl. dazu Göbel et al. (2008).

Eine besonders nutzerfreundliche Funktionalität bietet sich dem SOEPinfo-Anwender mit der Möglichkeit, aus den von ihm im Warenkorb gesammelten Variablen einen Syntaxquelltext für verschiedene Statistikprogramme zu erstellen. Mit dem generierten Syntaxquelltext werden nicht nur die im Warenkorb ausgewählten Variablen aus den einzelnen Datendateien herausgezogen und in einer einzigen neuen Datei zusammengefasst, mit der der Anwender dann seine Analysen erstellen kann, sondern es werden ebenfalls automatisch die jeweils benötigten Variablen zur Stichprobenabgrenzung (*balanced* versus *unbalanced design*, Personen versus Haushaltsebene, Region, Geschlecht etc.) und zur Gewichtung und Hochrechnung der Mikrodaten mit einbezogen. Derzeit werden die Statistikprogramme SPSS, Stata und SAS unterstützt.

7 Verknüpfungsmöglichkeiten mit regionalbezogenen Kontext-Informationen

Das SOEP bietet eine Vielzahl an Möglichkeiten, regionalbezogene Informationen bis hin zu „Nachbarschaftsdaten" bei der Analyse zu berücksichtigen. Mithilfe der regionalen Zuordnung des Haushaltes ist es möglich regionale Indikatoren auf der Ebene der Bundesländer (NUTS-1), der Raumordnungsregionen (oder NUTS-2), der Kreiskennziffer, des amtlichen Gemeindeschlüssels, der Postleitzahlen und der „Straßenabschnitte" an das SOEP zuzuspielen. Da die vorliegenden Verknüpfungsschlüssel (mit der Ausnahme der Postleitzahlen und Straßenabschnitte) die offiziellen geographischen Einheiten der Bundesrepublik Deutschland beschreiben, sind Verknüpfungen mit Daten der amtlichen Statistik problemlos möglich.[20] Auf Grund der damit einhergehenden erhöhten datenschutzrechtlichen Sensibilität der Daten müssen je nach Ebene der Regionalinformationen unterschiedliche Sicherheitsvorkehrungen eingehalten werden.

Der SOEP-Standarddatensatz, der mit Abschluss eines Datenweitergabevertrages erhältlich ist, enthält als regionale Information lediglich eine Ost-West-Unterscheidung und das Bundesland.[21] Eine genauere Beschreibung des Regionstyps ist im Rahmen des Standarddatensatz über Gemeindetyp (Boustedt oder BIK)[22] und politische Gemeindegrößenklasse möglich und kann ggf. mit einem getrennt zu beantragendem Passwort genutzt werden.

Die Identifikation der Raumordnungsregionen (NUTS-2 Level) muss aus datenschutzrechtlichen Gründen gesondert bei der SOEP-Gruppe angefordert werden. Nach Abstimmung eines speziellen Datenschutzkonzeptes durch die Datenschutzbeauftragten des Nutzers und des DIW Berlin werden die zusätzlichen Daten dem Antragsteller zur Nutzung an seinem Arbeitsort zur Verfügung gestellt. Tiefer gegliederte regionale Schlüssel können aus datenschutzrechtlichen Gründen nicht außerhalb des DIW Berlin bereitgestellt werden.

20 Zum Beispiel gibt das Statistische Bundesamt eine DVD mit dem Titel „Statistik regional" heraus und das Bundesamt für Bauwesen und Raumordnung die INKAR CD (Indikatoren und Karten).
21 Bei der Variablen „Bundesland" werden bis zum Jahr 2000 auf Grund von zu geringen Fallzahlen das Saarland und Rheinland-Pfalz zusammengefasst.
22 Die sogenannten Boustedt Regionen (benannt nach ihrem Erfinder Olaf Boustedt 1972) unterscheiden Kernstädte, Ergänzungsgebiet, Verstädterte Zone und Randzone. Die BIK Regionen sind die Nachfolgeklassifikation von Boustedt mit einer Erweiterung auf die neuen Bundesländer und unterscheiden Ballungsräume, Stadtregionen, Mittelzentrengebiete und Unterzentrengebiete.

Es besteht jedoch zum einen die Möglichkeit Gastarbeitsplätze am DIW zu nutzen (hierbei können alle verfügbaren Regionalinformationen genutzt werden) oder per Fernrechen-Zugang (SOEPremote) zusätzlich auf die Kreiskennziffern zuzugreifen. Bei der Nutzung von SOEPremote hat der Nutzer keinen Zugang zu den Daten, sondern schickt seine speziell aufbereitete Stata Syntax[23] an eine eigens eingerichtete E-Mail-Adresse. Dort wird dieses Programm automatisch datenschutzrechtlich geprüft und erst nach erfolgreicher Prüfung an einen besonders gesicherten Rechner im DIW weitergeleitet. Dieser bewältigt die eigentliche Analyse und sendet das Ergebnis wieder an den ersten Rechner, der es dem Nutzer wiederum per E-Mail zurückschickt.

Eine besondere Innovation in der Beschreibung der kleinräumlichen Umgebung der Haushalte („Nachbarschaften") basiert auf einer Kooperation des SOEP mit microm (microm Micromarketing-Systeme und Consult GmbH). Mithilfe der nur beim Befragungsinstitut Infratest Sozialforschung in München vorliegenden Adresse wurden die Mikrodaten des SOEP mit den mikrogeografischen Daten der microm angereichert. Diese Ergänzung kleinräumiger Wohnumfeldinformationen sowie soziodemografischer und konsumrelevanter Daten unterstützt vielfältige neue Analyseansätze (vgl. Schräpler et al. 2007). Diese Verknüpfung ist datentechnisch bereits erfolgt und der kombinierte Datensatz kann im DIW Berlin an einem der Gastarbeitsplätze genutzt werden. Eine Dokumentation des Datensatzes findet sich in Goebel et al. (2007).

8 Ausblick

Die Erhebung und Aufbereitung der SOEP-Daten war und ist immer wieder laufenden Änderungen unterworfen. Dabei ist es notwendig, das gesamte Instrumentarium im engen Verbund mit dem Erhebungsinstitut und den Datennutzern permanent an neue Anforderungen anzupassen. Im Zuge der zunehmenden Verbreitung kommerzieller Umfragen ist auch die Teilnahmebereitschaft an wissenschaftlichen Surveys in den letzten Jahren weiter gesunken. Für die SOEP-Anwendung haben sich eine eingehende Panelpflege sowie möglichst offene Erhebungsformen mit individuell von den Befragten selbst wählbaren Befragungsmodi (*mixed mode*) als wichtige Komponenten erwiesen, um die Befragungsbereitschaft langfristig aufrechtzuerhalten (vgl. auch den Beitrag von Rosenbladt in diesem Heft).

Die Mehrzahl der Interviews wird immer noch von Interviewern, aber inzwischen nicht mehr mit gedruckten Fragebögen (PAPI), sondern computer-unterstützt (CAPI) durchgeführt. CAPI-Befragungen erlauben auch das Einbeziehen von erweiterten Befragungstechniken wie die Erfassung von schrittweise eingeschränkten Angaben zu Einkommens- und Vermögenskategorien, wenn die Beantwortung der offenen Frage verweigert wird oder die Erfassung von langen Antwortlisten und Zeitmessungen bei Kognitionserhebungen. Auch webbasierte Erhebungen werden bereits getestet. Neben indirekten Befragungsmethoden, bei denen theoretische Konstrukte über Fragebatterien (Skalen) ermittelt werden, werden im SOEP in der Haupterhebung (zum Teil bisher nur in Pilotstudien und Pretests) auch direkte Messungen im Interview eingesetzt (Greifkraft, Körperumfang, Verhaltensexperimente, Test bei Kleinkindern), oder es werden Interviewdaten exemplarisch mit

23 Derzeit ist SOEPremote nur mit dem Statistikpaket Stata nutzbar, eine Beschreibung zur Erstellung von Stata Jobs für SOEPremote findet sich in Goebel (2005).

externen meist prozessproduzierten Informationen angereichert (institutionelle Angaben zu Arbeitsplatz, Schule; kleinräumige Regionalindikatoren; eventuell weitergehende medizinische Merkmale). Für derart reichhaltige Personeninformationen ist natürlich ein grundlegendes Vertrauensverhältnis zur integren Durchführung der Studie von Seiten der Befragten unerlässlich.

Auch auf Seiten der Nutzer hat sich die Kompetenz im Umgang mit komplexen Daten deutlich erhöht. Anwendungen, bei denen SOEP-Daten mit unterschiedlichen Zeitformaten und Erhebungseinheiten miteinander verknüpft werden, sind inzwischen keine Seltenheit mehr. Aus den mannigfaltigen Erhebungsinformationen werden immer komplexere Daten bereitgestellt (Biografiemerkmale, Arbeitsmarkterfahrung, Gesundheitsfaktoren). Generierte Daten werden zunehmend mit (multiple) imputierten Angaben bereitgestellt, auf die zum Teil mit eigens für imputierte Files ausgelegte Programmmodule in den einzelnen Statistikpaketen zugegriffen werden kann. So enthalten z. B. die auf Personen- und Haushaltsebene aufbereitete Vermögensdaten P/HWEALTH fünf Datenvarianten und die Angaben zum monatlichen Haushaltsnettoeinkommen werden zehn Datenvarianten für die fehlenden Werte enthalten.

Die eher konservative Form der Weitergabe der Daten als einfache unverbundene Querschnittsfiles wird zunehmend durch die Bereitstellung zeitübergreifender Biografiedaten ergänzt. Die Bereitstellung auch der jährlichen Befragungsdaten in kompakten zeitübergreifenden Formaten (long format) wird derzeit geprüft und weiter vorbereitet.

Die Auslieferung der SOEP-Mikrodaten erfolgt seit 2006 auf DVD. Direkt vom Web komplett herunterladbare Datenlieferungen, wie derzeit bereits bei vergleichbaren Daten in den USA und Großbritannien möglich, sind in Deutschland aus Datenschutzgründen noch nicht vorgesehen.

Literaturverzeichnis

Berger, Eva M. (2007): The Power of Monthly Data in the GSOEP: How the Chernobyl Catastrophe Affected People's Life Satisfaction and Environmental Concerns. SOEP Paper. 73.

Burkhauser, Richard V., Joachim R. Frick und Johannes Schwarze (1997): A Comparison of Alternative Measures of Economic Well-Being for Germany and the United States. *Review of Income and Wealth*, 43 (2), 153–171.

Canberra Group (2001): Expert Group on Household Income Statistics: Final Report and Recommendations. Ottawa.

Cochran, William G. (1977): *Sampling Techniques*. Bd. 3. New York, Wiley.

Dittmann, Jörg (2005): Forschungsbericht über die prototypische Verknüpfung des SOEP mit Medien-Tenor-Daten. DIW Berlin Research Notes. 6.

Frick, Joachim R., Jan Goebel und Markus M. Grabka (2007): Assessing the Distributional Impact of "Imputed Rent" and "Non-Cash Employee Income" in Microdata: Case Studies Based on EU-SILC (2004) and SOEP (2002): In: Eurostat (Hrsg.): *Comparative EU statistics on Income and Living Conditions: Issues and Challenges. Proceedings of the EU-SILC Conference, Helsinki, 6-8 November 2006*. Luxembourg, European Communities, 117–142.

Frick, Joachim R., Jan Goebel, Edna Schechtman, Gert G. Wagner und Shlomo Yitzhaki (2006): Using Analysis of Gini (ANoGi) for detecting whether two sub-samples represent the same universe: The German Socio-Economic Panel Study (SOEP) Experience. *Sociological Methods & Research*, 34 (4), 427–468

Frick, Joachim R. and Markus M. Grabka (2003): Imputed Rent and Income Inequality: A Decomposition Analysis for the Great Britain, West Germany and the U.S. *Review of Income and Wealth*, 49 (4), 503–537.

Frick, Joachim R. und Markus M. Grabka (2005): Item-Non-Response on Income Questions in Panel Surveys: Incidence, Imputation and the Impact on Inequality and Mobility. *Allgemeines Statistisches Archiv*, 89 (1), 49–60.

Frick, Joachim R., Markus M. Grabka und Jan Marcus (2007): Editing and Multiple Imputation of Item-Non-Response in the 2002 Wealth Module of the German Socio-Economic Panel (SOEP): SOEPpapers. 18.

Frick, Joachim R., Stephen P. Jenkins, Dean R. Lillard, Oliver Lipps und Marc Wooden (2007): The Cross National Equivalent File (CNEF) and its Member Country Household Panel Studies. *Schmollers Jahrbuch*, 127 (4), 627–654.

Frick, Joachim R., Peter Krause und Jürgen Schupp (1992): SIR and the GERMAN SOCIO-ECONOMIC PANEL STUDY (SOEP): ESF-Working Paper. 6.

Galler, Heinz Peter (1987): Zur Längsschnittgewichtung des Sozio-oekonomischen Panels. In: Krupp Hans-Jürgen und Ute Hanefeld (Hrsg.): *Lebenslagen im Wandel: Analysen 1987, Band 2 der Reihe: Sozio-oekonomische Daten und Analysen für die Bundesrepublik Deutschland.* New York, Frankfurt a.M., Campus, 295–317.

Gerhards, Jürgen und Silke Hans (2006): Zur Erklärung der Assimilation von Migranten an die Einwanderungsgesellschaft am Beispiel der Vergabe von Vornamen. DIW Discussion Paper. 583.

Goebel, Jan (2005): Job Submission Instructions for the SOEPremote System at DIW Berlin. Download unter: www.diw.de/documents/dokumentenarchiv/17/44069/soepremote2005.pdf (Stand 21. Mai 2008).

Goebel, Jan, Peter Krause, Rainer Pischner, Ingo Sieber und Gert G. Wagner (2008): Daten- und Datenbankstruktur der Längsschnittstudie Sozio-oekonomisches Panel (SOEP). DIW Data Documentation. 28.

Goebel, Jan, C. Katharina Spieß, Nils R. J. Witte und Susanne Gerstenberg (2007): Die Verknüpfung des SOEP mit MICROM-Indikatoren: Der MICROM-SOEP Datensatz. DIW Data Documentation. 26.

Grabka, Markus M. (2007): Codebook for the $PEQUIV File 1984-2006 - CNEF Variables with Extended Income Information for the SOEP. DIW Berlin Data Documentation. 21.

Grabka, Markus M. und Joachim R. Frick (2003): Imputation of Item-Non-Response on Income Questions in the SOEP 1984–2002. DIW Research Notes. 29.

Haisken-DeNew, John P. und Joachim R. Frick (2005): Desktop Companion to the German Socio-Economic Panel Study (GSOEP), Version 8.0 – Update to Wave 21, DIW Berlin. Download unter: www.diw.de/deutsch/soep/service_dokumentation/handbuch_(dtc)/27193.html (Stand 21.Mai 2008).

Horwitz, Daniel G. and D. J. Thompson (1952): A Generalisation of Sampling without Replacement From a Finite Universe. *Journal of the American Statistical Association*, 47, 663–685.

Kalton, Graham (1977): Practical Methods for Estimating Survey Sampling Errors. *Bulletin of the International Statistical Institute*, 47 (3), 495–514.

Kish, Leslie (1965): *Survey Sampling*. New York, Wiley.

Kish, Leslie (1995): Methods for Design Effects. *Journal of Official Statistics*. 11 (1), 55–77.

Krause, Peter, Rainer Pischner und Gert G. Wagner (1993): Optimale Verarbeitung von Längsschnittdaten – Das Beispiel des Sozio-ökonomischen Panels (SOEP). DIW Diskussionspapier. 75.

Krause, Peter und Gert G. Wagner (1991): Datenhaltung bei sozialwissenschaftlichen Panel-Studien. Sfb 187-Arbeitspapier zum Workshop „Datenbank", Ruhr-Universität Bochum.

Pischner, Rainer (2007a): Die Querschnittsgewichtung und die Hochrechnungsfaktoren des Sozio-oekonomischen Panels (SOEP) ab Release 2007 (Welle W). DIW Data Documentation. 22.

Pischner, Rainer (2007b): Die Querschnittsgewichtung und die Hochrechnungsfaktoren des Sozio-oekonomischen Panels (SOEP) ab Release 2007 (Welle W) – Modifikationen und Aktualisierungen. Download unter: www.diw.de/deutsch/soep/service_dokumentation/dokumentation/32046.html. Stand 21.Mai 2008.

Rao, J. N. K. und C. F. J. Wu (1988): Resampling Inference with Complex Survey Data. *Journal of the American Statistical Association*, 83 (401), 231–241.

Rendtel, Ulrich (1995): *Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität*. Frankfurt a. M., New York, Campus.

Rendtel, Ulrich und Ulrich Pötter (1993): Über Sinn und Unsinn von Repräsentativstudien. *Allgemeines Statistisches Archiv*, 77 (3), 260–280.

Robins, James M., Andrea Rotnitzky und Lue Ping Zhao (1994): Estimation of Regression Coefficients when some Regressors are not always observed. *Journal of the American Statistical Association*, 89 (427), 846–866.

Robins, James M., Andrea Rotnitzky und Lue Ping Zhao (1995): Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association*, 90 (429), 106–121.

Rubin, Donald B. (1987): *Multiple Imputation for Nonresponse in Surveys*. New York, Wiley.

Rubin, Donald B. (1996): Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91 (434), 473–489.

Rust, Keith F. (1985): Variance Estimation for Complex Estimators in Sample Surveys. *Journal of Official Statistics*. 1 (4), 381–397.

Särndal, Carl-Eric, Bengt Swensson und Jan Wretman (1992): *Model Assisted Survey Sampling*. New York, Springer.

Schräpler, Jörg-Peter (2004): Respondent Behavior in Panel Studies – A Case Study for Income Nonresponse by Means of the German Socio-Economic Panel (SOEP). *Sociological Methods & Research*, 33 (1), 118–156.

Schräpler, Jörg-Peter (2007): A Study of Mode-Effects of a Change from PAPI to CAPI. *Schmollers Jahrbuch*, 127 (1), 113–125.

Schräpler, Jörg-Peter, Jürgen Schupp und Gert G. Wagner (2008): Who Are the Nonrespondents? An Analysis Based on a New Subsample of the German Socio-Economic Panel (SOEP) including Microgeographic Characteristics and Survey-Based Interviewer Characteristics. DIW Research Note (in Vorbereitung).

Schräpler, Jörg-Peter und Gert G. Wagner (2000): Das Verhalten von Interviewern - Darstellung und ausgewählte Analysen am Beispiel des „Interviewer-Panels" des Sozio-oekonomischen Panels. *Allgemeines Statistisches Archiv*, 85 (1), 45–66.

Schräpler, Jörg-Peter und Gert G. Wagner (2005): Characteristics and Impact of Faked Interviews in Surveys – An Analysis of Genuine Fakes in the Raw Data of SOEP. *Allgemeines Statistisches Archiv*, 89 (1), 7–20.

Schwarze, Johannes (1995): Simulating the Federal Income and Social Security Tax Payments of German Households Using Survey Data. Cross-National Studies in Aging Program Project Paper. 19.

Spieß, Martin (2005): Derivation of Design Weights: The Case of the German Socio-Economic Panel (SOEP). DIW Data Documentation. 8.

Spieß, Martin and Martin Kroh (2007): Documentation of the Dataset DESIGN of the Socio-Economic Panel Study (SOEP). Download unter: *www.diw.de/deutsch/soep/service_dokumentation/dokumentation/32046.html* (Stand 21. Mai 2008).

Spieß, Martin and Martin Kroh (2008): Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) 1984–2006. DIW Data Documentation. 27.

Spieß, Martin, Martin Kroh, Rainer Pischner und Gert G. Wagner (2008): On the Treatment of Non-Original Sample Members in the German Household Panel Study (SOEP)-Tracing and Weighting. DIW Data Documentation.30.

Wagner, Gert G., Joachim R. Frick und Jürgen Schupp (2007): The German Socio-Economic Panel Study (SOEP) – Scope, Evolution and Enhancements. *Schmollers Jahrbuch*, 127 (1), 139–169.

Wolter, Kirk M. (1985): *Introduction to Variance Estimation*. New York, Springer.

Wooldridge, Jeffrey M. (2002a): Econometric Analysis of Cross Section and Panel Data. Cambridge, MA, The MIT Press.

Wooldridge, Jeffrey M. (2002b): Inverse Probability Weighted M-Estimators for Sample Selection, Attrition and Stratification. *Portugese Economic Journal*, 1 (2), 117–139.

Wooldridge, Jeffrey M. (2004): Inverse Probability Weighted Estimation for General Missing Data Problems. CeMMAP Working Papers. CWP05/04.