Deutsches Institut für
Wirtschaftsforschung

DIW BERLIN

www.diw.de

SOEP The German Socio-Economic Panel Study

SOEPpapers
on Multidisciplinary Panel Data Research

89

Jan Göbel, Peter Krause, Rainer Pischner, Ingo Sieber, Gert G. Wagner

Daten- und Datenbankstruktur der Längsschnittstudie Sozio-oekonomisches Panel (SOEP)

Berlin, Februar 2008

SOEPpapers on Multidisciplinary Panel Data Research
at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPpapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPpapers are available at
http://www.diw.de/soeppapers

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: Uta Rahmann | urahmann@diw.de

Jan Göbel *
Peter Krause *
Rainer Pischner *
Ingo Sieber *
Gert G. Wagner **

Daten- und Datenbankstruktur der Längsschnittstudie Sozio-oekonomisches Panel (SOEP)

Berlin, Februar 2008

* DIW Berlin, Längsschnittstudie SOEP

** DIW Berlin, Längsschnittstudie SOEP und TU Berlin, gwagner@diw.de

Inhaltsverzeichnis

Verzeichnis der Tabellen und Abbildungen

1 Einleitung

Die Längsschnittstudie SOEP wurde geschaffen, um sozial- und wirtschaftswissenschaftliche Fragestellungen im Rahmen von Haushalten Lebensläufen und lebenslaufbezogenem Verhalten analysieren zu können (vgl. Krupp 2007). Da dabei auf Lebensverläufe nicht nur aus ökonomischer, sondern vor allem aus soziologischer Sicht geschaut wurde (und wird), wurden die erhobenen Daten zur Lebenszufriedenheit auch für verhaltenswissenschaftliche (psychologische) Forschungsfragen interessant; deswegen werden seit 2002 die Erhebungsinstrumente des SOEP systematisch um verhaltenswissenschaftliche Konzepte ergänzt (vgl. Wagner et al. 2007).

Durch die Vielzahl sozial- und verhaltenswissenschaftlicher Fragestellungen, die mit den SOEP-Daten wissenschaftlich bearbeitet werden, ist nicht nur die „lebenslaufbezogene“ Konzeption des SOEP immer mehr in den Vordergrund gerückt, sondern auch die Mikrodaten selbst werden immer mehr im Hinblick auf „Event-Analysen“ nutzerfreundlich aufbereitet.

Haushaltspanels wurden vorwiegend zum Zwecke der Dynamik der Einkommensverteilung und –armut begonnen; die erste Studie – PSID – führt diese Fragestellung sogar im Namen: Panel Study of Income Dynamics. Entsprechend liegen sehr viele Veröffentlichungen vor, die für Veränderungen der Gesamtbevölkerung (Querschnitte) repräsentativ sind. Es handelt sich dabei um eine Spannweite von internationalen Spitzenzeitschriften (vgl. Beaudry/Green 2003) bis hin beispielsweise zum Armuts- und Reichtumsbericht der deutschen Bundesregierung. Aber mit dem Vorliegen der entsprechenden Panel-Daten, die lange Teile von individuellen Lebensläufen im Familien- und Haushaltskontext beobachten, hat ein ganz anderer Typus von Fragestellungen und Veröffentlichungen zugenommen: die Analyse von Lebensverläufen aus Sicht einzelner Gruppen und Kohorten. Hier seien einige wenige ausgewählte Beispiele von jüngsten Längsschnittsstudien aus verschiedenen Bereichen von Sozial- und Wirtschaftswissenschaften und der Psychologie genannt, die auf Basis von SOEP-Daten der Frage der Lebensverläufe der Befragten und ihrem Zusammenspiel mit anderen Lebensverläufen nachgehen und wie einzelne Befragte mit den Folgen von Lebensverläufen umgehen: Familien-Dynamik (Tamm 2005), Verlassen der elterlichen Wohnung (Scherger 2007), Heirat (Diener et al. 2006; Lucas/Clark 2006; Zimmermann/Easterlin 2006; Stutzer/Frey 2006), Arbeitsteilung im Privathaushalt (Cooke 2007), Eintritt in die und Austritt aus der Arbeitslosigkeit (Lucas et al. 2004; Lucas 2005; Jürges 2007; Romeu Gordo 2006), Übergang in den Ruhe-

stand (Börsch-Supan/Jürges 2006), Scheidung (Andreß et al. 2003), Dynamik ehrenamtlicher Tätigkeiten (Erlinghagen 2007), Migration (Juerges 2006), Verwitwung (Burkhauser et al. 2005), und Tod (Gerstorf et al. 2007). Das SOEP ist damit nicht nur ein Haushaltspanel, sondern zugleich eine Kohortenstudie. Seine Daten- und Datenbankstruktur ist entsprechend kompliziert und wird hier dokumentiert.

2 Konzeption und Stichprobendesign des SOEP

Das SOEP ist eine wissenschaftsgetragene Längsschnittserhebung bei Personen und ihren Haushalten („Haushaltspanel“). Diese Erhebung wird von wissenschaftlichen Fragen des Theorietestens und der Politikberatung bestimmt und nicht von amtlichen und politischen Fragestellungen. Als wissenschaftsgetragene Erhebung ist das SOEP Teil einer weltweiten Forschungs-Infrastruktur, die Längsschnittsdaten für die Analysen von Personen und Haushalten zur Verfügung stellt (vgl. Butz und Torrey 2006, Frick et al 2007). Neben ähnlichen Erhebungen in Industrieländern (so z. B. PSID in den USA, BHPS in Großbritannien, HILDA in Australien und SHP in der Schweiz und SHARE (beschränkt auf 50-Jährige und Ältere) in etlichen EU-Staaten, sind wissenschaftsgetragene Haushaltspanels auch in Russland (RMLS)[1] und z. B. der Ukraine (UMLS)[2] zu finden. Für China ist eine entsprechende wissenschaftsgetragene Studie in Planung.

Im SOEP gibt es eine Vielzahl von Erhebungsinstrumenten, um Personen im Kontext ihrer Familie und ihres Haushaltes verfolgen zu können. Das SOEP erhebt möglichst lange Ausschnitte aus Lebensläufen mit Hilfe einer einmal pro Kalenderjahr stattfindenden Befragung. In zufällig und repräsentativ ausgewählten Haushalten werden alle Erwachsenen (17-jährige und Ältere) mit Hilfe von Personenfragebögen direkt befragt. Zusätzlich, werden von der „Hauptauskunftsperson“ im Haushaltsfragebogen Merkmale über den gesamten Haushalt (z.B. die Wohnung) erfragt. Hinzu kommen spezielle Fragebögen zum Lebenslauf bei neu erfassten Personen oder Fragebögen, mit denen Mütter Angaben über ihre (kleinen) Kinder machen.

Da das SOEP im Jahr 1984 in Westdeutschland begonnen wurde, werden nach dem Erhebungsjahr 2008 viele Befragte bereits 25 mal befragt worden sein (etwa 2500 Personen). Die in Ostdeutschland 1990 hinzugekommenen Befragten werden im Jahr 2009 zum 20. mal befragt werden (etwa 1500 Personen).

Während in der Psychologie und Medizin übliche Kohortenstudien einzelne Lebensläufe von der Geburt an verfolgen, so liefert das SOEP aufgrund der Erhebung im Haushaltskontext, bereits Informationen für die Zeit vor der Geburt, nämlich Informationen über das Leben der

[1] http://www.cpc.unc.edu/projects/rlms/

[2] http://www.iza.org im Forschungsbereich „Arbeitsmärkte in Transformations- und Schwellenländern“.

Mutter (und auch des Vaters, wenn er – was meist der Fall ist – mit der Mutter zusammenlebt). Für ein Kind, das in einen „SOEP-Haushalt“ hineingeboren wird, wird anschließend dessen Weg durch die Kindheit und Jugend verfolgt (über den Haushaltsfragebogen und spezielle Fragebögen, die an Mütter gerichtet sind). Ab dem 17. Lebensjahr wird ein Jugendlicher zum persönlichen Befragten (bei der erstmaligen Befragung wird – seit 2001 – ein spezieller Jugendfragebogen eingesetzt und seit 2006 wird die kognitive Leistungsfähigkeit mit einem halbstündigen Instrument getestet.

Während des Erwachsenenlebens werden zu einer Vielzahl von Bereichen der Zeitverwendung Indikatoren (z.B. Aus- und Weiterbildung, Kinderbetreuung, Erwerbstätigkeit, Arbeitslosigkeit, ausgewählte Freizeitaktivitäten) und subjektive Outcomes (z.B. Lebenszufriedenheit, Sorgen) erhoben. Schließlich werden auch automatisch die letzten Lebensjahre betrachtet und im Falle eines Lebens in einem Alten-/Pflegeheim wird zumindest die Tatsache dieses Umzugs erhoben (während Erhebungen in Pflegeheimen meist nicht realisierbar sind). Aufgrund der Erhebung im Haushaltskontext fallen schließlich noch nach dem Tode eines Befragten Informationen an, nämlich über Hinterbliebenenrenten, Erbschaften und die Lebenszufriedenheit von Hinterbliebenen.[3]

Realisiert wird dieses Konzept der Erhebung von Lebensläufen durch die Ziehung einer Haushaltsstichprobe, bei der alle Personen in diesen Haushalten (genauer: Privathaushalten) selbst Erhebungs-Einheit werden. Kinder werden freilich erst mit dem Erreichen des 17. Lebensjahres selbst persönlich befragt. Auch ungeborene Kinder gehören virtuell zu dieser Stichprobe. Ziehen Personen aus einem Befragungshaushalt aus, so werden diese weiter verfolgt und Personen, die mit ihnen zusammenziehen auf Dauer in das SOEP einbezogen. Durch den Einbezug von Nicht-Original-Stichproben-Mitgliedern (vgl. dazu Tabelle 3.1) entsteht im Prinzip eine Schneeballstichprobe. 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 i.d.R. nicht größer, da den Schneeball-Befragten jene gegenüberstehen, die nicht mehr bereit sind am SOEP weiter teil-

[3] Mit Hilfe eines speziellen Fragebogens für Hinterbliebene sollen ab 2008 die Informationen nach einem Sterbefall noch verbessert werden (über die letzte Lebensphase des verstorbenen SOEP-Teilnehmers und über die Trauerarbeit).

zunehmen („Panel-Ausfälle“). Auch diese Ausfälle werden im Laufe der Panel-Laufzeit durch Gewichtungen berücksichtigt, die unverzerrte Rückschlüsse auf die Grundgesamtheit zulassen.

Die Erhebung von Lebensläufen im Haushaltskontext erlaubt es, dass die jahresbezogenen Erhebungsdaten, für die Erhebungsjahre zwischen 1984 und 2007 und künftiger Jahre auf die Grundgesamtheit aller Personen und (Privat)Haushalte in Deutschland hochgerechnet werden können. Dadurch wird zum Beispiel erst die Analyse der Dynamik von relativer Einkommensarmut, die sich am mittleren Wert (Median) der Bevölkerung misst, möglich.

Das Weiterverfolgungskonzept des SOEP, das von allen später begonnenen Haushalt-Panels übernommen wurde,[4] bildet die endogene Bevölkerungsdynamik vollständig ab, da SOEP-Teilnehmer, die sterben, keine unerwünschten bzw. verzerrenden Ausfälle darstellen, sondern die Bevölkerungsdynamik in der Stichprobe nachvollziehen. Freilich wird durch dieses Konzept (wie auch bei jeder Kohorten-Studie) Zuwanderung nicht erfasst, wenn die Zuwanderung nicht in einem bestehenden Haushalt erfolgt (während Zuwanderung im Zuge von Familiennachzug im bestehenden Haushalt automatisch einbezogen wird). Deswegen müssen – zur Sicherstellung der Querschnittsrepräsentativität und z.B. der Erfassung der Einkommensverteilung idealerweise jährlich Zusatzstichproben für Zuwanderer gezogen werden. Faktisch geschah dies einmal gezielt (1994/95); seither fand dies nur im Rahmen repräsentativer Auffrischungsstichproben für die gesamte Bevölkerung in den Jahren 1998, 2000 und 2006 statt.

[4] Mit Ausnahme der umfassenden Weiterverfolgung aller Nicht-Original-Stichproben-Teilnehmer; die anderen Panels verfolgen nur geschiedene Elternteile/Ex-Partner weiter, wenn gemeinsame Kinder geboren wurden.

3 Die Stichproben des SOEP

Mit der 2006 gestarteten Ergänzungsstichprobe H umfasst das Sozio-oekonomische Panel (SOEP), das 1984 begonnen wurde, nunmehr acht Teil-Stichproben und – für die Stichproben A und B – 23 auswertbare Wellen (1984 – 2006).[5] Ein Ende der Erhebungen ist nicht geplant.

Die SOEP-Respondenten wurden haushaltsweise für das SOEP rekrutiert. Der Modus der Kontaktaufnahme ist bei allen Teilstichproben gleich: ein Schreiben der Feldarbeits-Organisation (TNS Infratest München), mit dem das SOEP und ein Interviewer angekündigt werden.[6]

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 (wird aber faktisch erst ab Welle 7 (1990) tatsächlich nennenswert genutzt und liegt nach 22 Wellen im Sample A bei 20%; nach 5 Wellen in Sample F bei ca. 6%).

Die verschiedenen Befragungs-Modi sind für jeden einzelnen Record codiert und im Standard-Datensatz abgelegt und auswertbar. Allein aufgrund der verschiedenen Befragungs-Modi ist das SOEP für surveymethodische Fragen eine wahre Fundgrube (vgl. Schräpler 2007). Hierzu kommt die Identifikation eines jeden einzelnen Interviewers (vgl. Schräpler und Wagner 2000), so dass potenzielle Interviewereffekte leicht analysierbar sind (vgl. Schräpler 2004).[7]

In Spezial-Datensätzen sind auch Angaben zu Haushalten vorhanden, die sich weigerten, an der Erhebung teilzunehmen („Brutto"-Daten) oder die im Laufe der Zeit 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.

---

[5] Eine aktuelle Beschreibung des SOEP findet sich bei Wagner, Frick, Schupp (2007).

[6] Die Adressen wurden 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).

[7] Da ein nicht-experimentelles Design vorliegt, sind Befragungsartefakte zwar kontrollierbar, aber eine kausale Zuschreibung auf Einzeleffekte bedarf der Setzung von Annahme.

Ab dem Jahr 2008 werden auch die Mikro-Daten einer 2007 durchgeführten Drop-Out-Erhebung allgemein verfügbar sein. Dann werden auch die Daten einer 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 6.5 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/Hans 2006).

Tabelle 3.1
Zahl der Haushalte im Sozio-oekonomischen Panel im Jahr 2006 nach Stichprobe und Stichprobenstatus der Befragten

Status	Haushalte Insgesamt	Nur OSM-Haushalte*)	Gemischte Haushalte	Nur NOSM-Haushalte**)	Nur OSM-Haushalte*)	Gemischte Haushalte	Nur NOSM-Haushalte**)
Stichprobe	Zahl der Haushalte				Status-Anteile in %		
Insgesamt	12361	9490	2317	554	76,8	18,7	4,5
A	2821	1572	950	299	55,7	33,7	10,6
B	655	392	223	40	59,9	34,0	6,1
C	1717	1123	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	3895	3394	450	51	87,1	11,6	1,3
G	859	786	69	4	91,5	8,0	0,5
H	1506	1506	.	.	100,0	.	.

*) OSM-Haushalte: (Original Sample Member): Haushalte, die keine zugezogenen Hausmitglieder enthalten

**) NOSM-Haushalte: (Non Original Sample Member): Haushalte, die keine Mitglieder mehr des Ursprungshaushaltes enthalten.

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

4 Hochrechnung des SOEP

Die Vielzahl der Teilstichproben bedeutet: für deskriptiv angelegte Analysen ist eine Hochrechnung für den SOEP-Datensatz unbedingt erforderlich, da die Ziehungswahrscheinlichkeiten der einzelnen Teilstichproben sich vom Design her unterscheiden (d. h. die jeweiligen Beobachtungen in den einzelnen Teilstichproben unterschiedlich viele Personen / Haushalte in der jeweiligen Grundgesamtheit repräsentieren). Hinzu kommt ein unterschiedliches Befragtenverhalten nach Welle 1 (selektive Attrition), wobei ebenfalls Unterschiede zwischen den Teilstichproben bestehen (z.B. aufgrund höherer Emigrationswahrscheinlichkeiten in den Stichproben B und D). Übersicht 4.1 zeigt die jeweilige repräsentative Grundgesamtheit der bisherigen acht Stichproben:

Übersicht 4.1
Die Stichproben des Sozio-oekonomischen Panels

Stichprobe A: (Deutsche) Haushalte[8] in der Bundesrepublik Deutschland (Hauptstichprobe, Start 1984)
Stichprobe B: Ausländische Haushalte[9] 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)

* 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.

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

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

4.1 Prinzip der Hochrechnung

Die Gesamtstichprobe des SOEP ist ungewöhnlich komplex. Ihre Gewichtung und Hochrechnung kann nur auf Grundlage eines konsistenten, theoretisch abgesicherten Konzeptes erfolgen. Es ist nicht möglich, an dieser Stelle detailliert hierauf einzugehen. Es erfolgt lediglich eine Beschreibung der wesentlichen Komponenten dieses Konzeptes mit entsprechenden Literaturhinweisen.[10]

Allein aufgrund der unterschiedlichen Stichprobengrößen und den sich daraus implizit ergebenden Ziehungswahrscheinlichkeiten ist eine Gewichtung der verschiedenen Teilstichproben des SOEP notwendig, wenn man alle Beobachtungen gemeinsam auswerten will. Diese Unterschiede werden durch „Designgewichte“ berücksichtigt (vgl. Spiess 2001), deren Größe und Verteilung in Tabelle 4.1.1 beispielhaft dargestellt wird. Hinzu kommen jedoch noch Verweigerungen in den ersten und auch weiteren Wellen.

Die Schätzung von Gewichten bzw. Hochrechnungsfaktoren erfolgt – wie international üblich – nach dem Ansatz von Horvitz und Thompson (1952) prinzipiell über den Kehrwert der Auswahlwahrscheinlichkeit der jeweiligen Stichprobeneinheit. Hochrechnungsfaktoren unterscheiden sich von Gewichten lediglich durch einen Skalarmultiplikator: Gewichte geben den Stichprobenumfang wieder, Hochrechnungsfaktoren hingegen den (geschätzten) Populationsumfang in der Grundgesamtheit an (Bevölkerung in Privathaushalten beziehungsweise Zahl der Privathaushalte am Hauptwohnsitz in Deutschland). Während die nach Horvitz und Thompson berechnete Summe der Hochrechnungsfaktoren der Zahl der Einheiten der Grundgesamtheit entspricht, stimmt die Summe der Gewichte mit der Fallzahl der Stichprobe überein.

[10] Die Frage, ob überhaupt eine Stichprobe hochgerechnet werden soll, wird hier nicht diskutiert werden. Siehe zu diesem Thema Rendtel und Pötter (1993)

Tabelle 4.1.1
Designgewichte im Sozio-oekonomischen Panel

Stichprobe/ Teilstichprobe	Kurz-Beschreibung der Stichprobe (Startjahr)	Kennziffern für die Designgewichte von Haushalten mit *positiven* Querschnittsgewichten		
		Mittelwert	Standardabweichung	Fallzahl im Startjahr
A	Deutsch (1984)	3344	0	4528
B	Ausländisch (1984)	546	305	1393
B1	Türkisch	772	332	397
B2	Italienisch	624	250	194
B3	Griechisch	355	158	196
B4	Jugoslawisch	487	211	306
B5	Spanisch	261	117	200
C	DDR (1990)	1900	0	2179
D	Zuwanderer (1994/1995)	3279	331	252 *
D1	Übersiedler	2946	0	127 *
D2	Aussiedler	3405	0	44 *
D3	Sonstige	3693	0	81 *
E	Ergänzung (1998)	19081	0	1056
F	Ergänzung (2000)	3406	402	6052
F1	Deutsch	3519	0	5607
F2	Ausländisch	1980	0	445
G	Hocheinkommenshaushalte mit Haushaltsnettoeinkommen > 4500 € (2000)	924	148	998 *
G1	West: < 5113 €	754	0	483 *
G2	Ost: < 5113 €	638	0	55 *
G3	West: >= 5113 €	525	0	419 *
G4	Ost: >= 5113 €	490	0	41 *
H	Ergänzung (2006)	10432	0	1506

* Folgende Stichproben enthalten weitere Haushalte ohne positive Gewichte für die Querschnittsgewichtung: 122 Haushalte in D, davon 49 in D1, 37 in D2, 36 in D3. Diesen Haushalten konnten kein Designgewicht zugeordnet werden, da ihre Auswahlwahrscheinlichkeit nicht zu bestimmen ist. (vgl. Abschnitt 3.3) 226 Haushalte in G, davon 152 in G1, 23 in G2, 43 in G3 und 8 in G4. Diese Haushalte überschritten in der zweiten Welle nicht die vorgeschrieben Einkommensgrenze von 4500 €. (vgl. Abschnitt 3.6)

Quelle: Das Sozio-oekonomische Panel. Release 2007; eigene Berechnungen.

Horvitz und Thompson stellten ihr Konzept indes nur für reine Querschnittsdaten vor. Stichprobeneinheiten sind beim SOEP primär die Haushalte, sekundär die Personen, die in diesen Haushalten leben. Der Ansatz von Horvitz und Thompson wurde von Galler (1987) für Paneldaten weiterentwickelt, indem die Schätzung von Auswahlwahrscheinlichkeiten für die Startwelle um die Schätzung der Verbleib- bzw. Antwortwahrscheinlichkeit für die folgenden Wellen erweitert wurde. Hierzu muss zunächst die Wahrscheinlichkeit einer erneuten Kontaktaufnahme, darauf folgend die Wahrscheinlichkeit einer erneuten Antwortgewährung be-

stimmt werden. Eng verknüpft mit dieser Aufgabe ist die Analyse des Ausfallverhaltens der Stichprobeneinheiten von Welle zu Welle.[11] Dieses Konzept ermöglicht es, sowohl Querschnittsgewichte ab Welle 2 jeder Stichprobe sowie Längsschnittgewichte theoretisch konsistent und praktisch zufriedenstellend zu ermitteln.

Der Gewichtungsprozess im SOEP stellt sich somit – pro Teilstichprobe – vereinfacht wie folgt dar[12]:

Nur Welle 1:

- Ermittlung der Startgewichte neuer Teilstichstichproben.

Ab Welle 2:

- Schätzung der Kontakt- und Antwortwahrscheinlichkeiten der verbliebenen Haushalte, über die vorläufige Gewichte der aktuellen Welle bestimmt werden.
- Schätzung von Startgewichten für seit der Vorwelle neu entstandene Haushalte in den Alt-Stichproben.

4.2 Querschnittsgewichtung und Randanpassung

Die wellenspezifischen Hochrechungs- und Gewichtungsfaktoren garantieren nicht, dass wichtige Ecksummen mit denen der amtlichen Statistik übereinstimmen. Dies ist z. B. allein wegen der Stichprobenfehler beim SOEP und beim Mikrozensus immer der Fall. Hinzu kommen auch systematische Probleme; so werden beim SOEP nicht jährlich Zuwanderer, die neue Haushalte gründen, erfasst. Um eine formale Konsistenz sicherzustellen, wird das SOEP jährlich an die jeweiligen Daten des Mikrozensus (MZ) angepasst, so dass die SOEP-spezifische Verteilung nach Region, Alter, Geschlecht, Haushaltsgröße und Nationalität in den Eckdaten derjenigen des Mikrozensus entspricht.

4.3 Längsschnittgewichtung

Die Schätzung der Wahrscheinlichkeit, befragte Haushalte wiederzufinden, sowie die Wahrscheinlichkeit, dass diese auch wieder ein Interview gewähren, bilden die Grundlage der Längsschnittgewichtung. Bei der Längsschnittgewichtung tritt also an die Stelle der Aus-

[11] Siehe hierzu z.B. Pannenberg (2000) sowie Kroh und Spiess (2006) .

[12] Vgl. hierzu auch Rendtel (1995).

wahlwahrscheinlichkeit die Wahrscheinlichkeit der erneuten Kontaktaufnahme[13]. Das Produkt aus Kontaktwahrscheinlichkeit und Antwortwahrscheinlichkeit wird als Bleibewahrscheinlichkeit bezeichnet. Somit ergibt sich das Längsschnittgewicht einer Welle t [14] aus dem Produkt von Querschnittgewicht der Welle t-1 und der reziproken Bleibewahrscheinlichkeit der Welle t. Geschätzt werden die Bleibewahrscheinlichkeiten für den SOEP-Datensatz im Rahmen der bereits erwähnten umfangreichen Ausfallanalyse, deren Betrachtung hier zu weit führen würde.

In einen Haushalt hineinziehende Nicht-Original-Stichproben-Mitglieder (Non-OSM) erhalten nach einem Verfahren, das bei Rendtel (1995) näher beschrieben wird, anhand ausgewählter Haushaltsmerkmale ein Startgewicht zugewiesen.

4.4 Personengewichtung

Ausgehend von den Haushaltsgewichten werden Personengewichte bestimmt. Da grundsätzlich sämtliche Personen, die mindestens 16 Jahre alt sind, an der Befragung teilnehmen sollen, gilt für repräsentative Stichproben prinzipiell die Gleichheit von Haushalts- und Personengewicht. Somit können die geschätzten Haushaltsgewichte zunächst 1:1 auf sämtliche Personen des Haushalts (einschließlich deren Kinder) übertragen werden. Anschließend erfolgen zwei Korrekturen:

Zum einen werden in der jeweiligen Startwelle Personen, die einen zweiten Wohnsitz haben, nur mit der Hälfte ihres eigentlichen Gewichts versehen, da sie eine doppelte Auswahlwahrscheinlichkeit besitzen. Zum anderen werden die Personengewichte - getrennt nach alten und neuen Ländern – an die Alterstruktur der Personen in Privathaushalten am Hauptwohnsitz angepasst.

[13] siehe hierzu Rendtel, 1995.

[14] In der ersten Welle einer Panelerhebung sind Querschnittsgewichtung und Längsschnittgewichtung definitionsgemäß identisch.

4.5 Standard-Hochrechnungsfaktoren

Zum besseren Verständnis werden die Bezeichnungen für die Gewichte vorab erläutert: Jedes Gewicht wird 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

y = <leer>, also nicht besetzt, bezeichnet Standardhochrechnungsfaktoren. Standardgewichte umfassen sämtliche Samples mit Ausnahme der Hocheinkommensstichprobe G. Diese Gewichte sind für sämtliche Wellen verfügbar.

y = 1 bezeichnet modifizierte Standardhochrechnungsfaktoren

Diese sind normalerweise identisch mit den Standardgewichten; allerdings werden die Gewichte von Stichproben in ihrer ersten Welle auf Null gesetzt. Zur erstmals mit der Datenauslieferung 2007 durchgeführten Null-Setzung der Querschnittsgewichte der samplespezifischen ersten Wellen sei angemerkt: Es ist inzwischen bekannt, dass neue Befragte bei der Erhebung von Merkmalen wie z.B. Lebenszufriedenheit und Haushaltseinkommen in den ersten Wellen eines Panels signifikante Lerneffekte zeigen.[15] Deswegen empfiehlt es sich, erste Wellen nicht in deskriptive Analysen einzubeziehen.[16] Durch das Nullsetzen der Hochrechnungsfaktoren geschieht dies gewissermaßen automatisch[17] für alle Teilstichproben - nur nicht für Sample C, da in der damaligen DDR das Einkommen sehr einfach strukturiert war und die Antwortbereitschaft in der ersten Befragungswelle außerordentlich hoch war; deshalb erscheint auch eine Analyse in der ersten Welle hier sinnvoll.

y = ALL, umfasst sämtliche erhobene Stichproben des SOEP

y = D kennzeichnet die isolierte Zuwanderer-Stichprobe D

y = G kennzeichnet die isolierte Hocheinkommensstichprobe G

[15] Vgl. Frick et al. (2006)

[16] Erste Welle-Effekte können indes mit multivariaten Verfahren kontrolliert werden.

[17] Dies hat zur Konsequenz, dass mit Hilfe von $HRF1 keine Analysen für das Jahr 1984 möglich sind.

In der folgenden Tabelle 4.5.1 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.

Tabelle 4.5.1
Hochrechnungsfaktoren 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 in der Standard-Stichprobe**	Zahl sämtlicher Haushalte**	Hochgerechnete Privathaushalte in der Grundgesamtheit in Tsd.
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***

* y=D: Zuwanderer. y=F Ergänzung 2000 y=G: Hocheinkommensstichprobe.

** Nur Haushalte mit positivem Gewicht.

*** vorläufig. Die Daten basieren auf dem Mikrozensus 2005.

Tabelle 4.5.2
Hochrechnungsfaktoren für Personen in Privathaushalten im Sozio-oekonomischen Panel für die Wellen A – W (1984 – 2006)
– In den Gewichten enthaltene Stichproben und Eckdaten –

Welle $	Jahr	$PHRF	$PHRF1	$PHRFALL	$PHRFy *	Zahl Personen in Privathaus-halten in der Standard-Stichprobe**	Zahl sämtli-cher Personen **	Hochgerechnete Zahl Personen in Privathaushalten am Hauptwohnsitz in der Grundge-samtheit in Tsd.
A	1984	AB	= 0	.	.	16099	16173	60501
B	1985	AB	AB	.	.	14443	14508	60160
C	1986	AB	AB	.	.	13742	13804	60247
D	1987	AB	AB	.	.	13496	13563	60404
E	1988	AB	AB	.	.	12817	12872	60587
F	1989	AB	AB	.	.	12393	12443	61094
G	1990	ABC	ABC	.	.	18165	18254	78678
H	1991	ABC	ABC	.	.	17756	17844	79015
I	1992	ABC	ABC	.	.	17350	17429	79624
J	1993	ABC	ABC	.	.	16993	17072	80320
K	1994	ABC	ABC	.	.	16642	16715	80588
L	1995	ABCD	ABC	.	.	17246	17345	80787
M	1996	ABCD	ABCD	.	D	16860	16942	81010
N	1997	ABCD	ABCD	.	D	16474	16570	81216
O	1998	ABCDE	ABCD	.	D	18132	18234	81110
P	1999	ABCDE	ABCDE	.	D	17382	17487	81208
Q	2000	ABCDEF	ABCDE	.	D	30587	30764	81368
R	2001	ABCDEF	ABCDEF	.	D	27762	27920	81466
S	2002	ABCDEF	ABCDEF	ABCDEFG	D und G	26119	29072	81690
T	2003	ABCDEF	ABCDEF	ABCDEFG	D und G	25202	27868	81734
U	2004	ABCDEF	ABCDEF	ABCDEFG	D und G	24372	26916	81704
V	2005	ABCDEF	ABCDEF	ABCDEFG	D und G	23292	25638	81639
W	2006	ABCDEFH	ABCDEF	ABCDEFGH	D und G	25213	27442	81639***

* y=D: Zuwanderer. y=F Ergänzung 2000 y=G: Hocheinkommensstichprobe.

** nur Personen mit positiven Gewicht.

*** vorläufig. Die Daten basieren auf dem Mikrozensus 2005.

Eine Umstellung innerhalb der amtlichen Statistik und die Einbeziehung der Stichprobe H wurde zum Anlass genommen, den Stichprobenrahmen leicht zu modifizieren. Außerdem

wurde die Abgrenzung der sog. Standard-Hochrechnungsfaktoren rückwirkend ab 1984 geändert.[18]

Seit dem Erhebungsjahr 2000, also mit Einführung der Stichprobe F, wurden die Standardhochrechnungsfaktoren im Prinzip auf der Basis von vier Hochrechnungsrahmen ermittelt. Getrennt nach alten und neuen Ländern und noch einmal unterteilt nach den Altstichproben A-E und der Ergänzungsstichprobe F. Eingeflossen in die Hochrechnung waren jeweils 22 Restriktionen, wobei deren fünf auf die Haushaltsgröße und 17 auf Ausländeranteil, Geschlecht und Altersklassen entfielen.

Beginnend mit Welle V wurde zwar auf eine gesonderte Hochrechnung für die Altstichproben A-E vs. Ergänzungsstichprobe F verzichtet, deren Stichprobenanteile werden in der Klassifizierung nach Haushaltsgröße indes weiterhin berücksichtigt. Somit gibt es einen stichprobenspezifischen Teil und einen von den Stichproben unabhängigen Teil des Hochrechnungsrahmens.

Dieses Vorgehen ermöglicht es, bei Bedarf weiterhin getrennte Auswertungen z.B. für Stichprobe F für das Jahr 2005 vorzunehmen:

Zunächst setzt man sämtliche Gewichte für Haushalte, die nicht der Stichprobe F entstammen auf Null. Die verbleibenden Gewichte müssen anschließend nur noch über einen einfachen Dreisatz an die Ecksumme (Zahl der Haushalte insgesamt) angepasst werden. Die Gewichte sind so konstruiert, dass die Struktur nach Haushaltsgröße erhalten bleibt. Die übrigen Restriktionen werden dagegen nur annähernd erfüllt; an dieser Stelle müssen leichte Abweichungen von den Ecksummen hingenommen werden.

[18] Das Statistische Bundesamt weist seit dem Jahr 2005 Angaben für das Land Berlin nicht mehr nach West und Ost getrennt aus. Dies führte zwangsläufig zu leichten Modifizierungen des Hochrechnungsrahmens, vgl. Pischner (2007, S. 2).

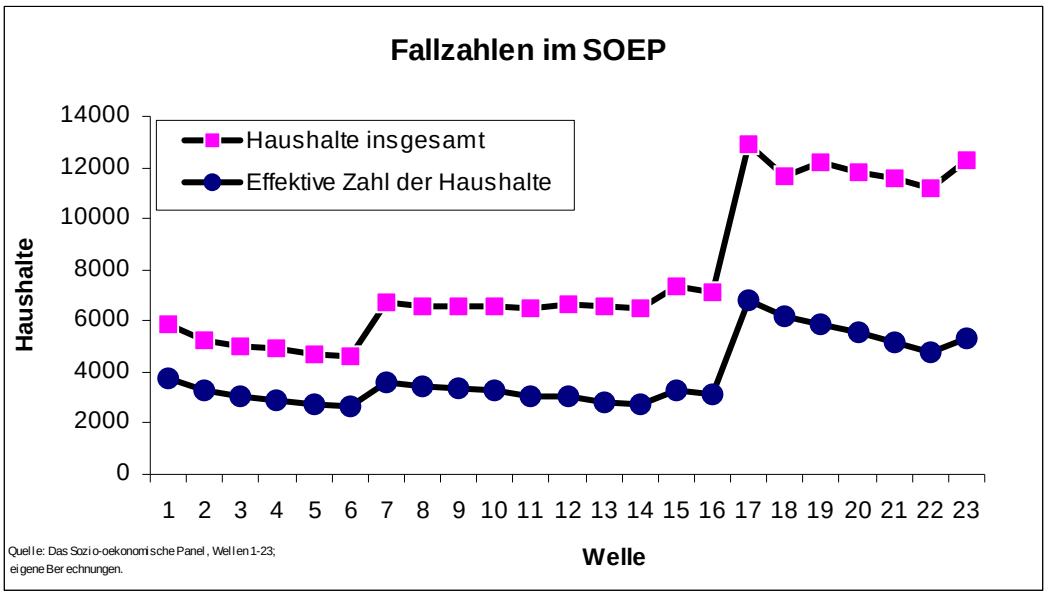
5 Fallzahlen

Das SOEP wird in Form von jährlich einmal erfragten und gemessenen Daten erhoben und die Daten der verschiedenen Erhebungswellen sind damit zuerst einmal eine Abfolge von Querschnittsdaten. So liegen nach 23 Wellen für 23.932 Haushalte (47.439 Personen) insgesamt 185.899 Haushaltsinterviews und 360.344 Personeninterviews vor. D.h. jeder Haushalt wurde im Durchschnitt beinahe acht Mal befragt und für immerhin 1535 Haushalte (2716 Personen) gibt es für jede Welle, d. h. 23 Mal, ein Haushalts- bzw. Personeninterview.

Je mehr Interviews es für eine Frage gibt, desto höher ist die Genauigkeit deskriptiver Statistiken bzw. desto enger sind die abzuleitenden Konfidenzintervalle für die Schätzwerte. Allerdings wird diese – eigentlich triviale – Aussage relativiert durch die Tatsache, dass sich die Varianz der Hochrechnungsfaktoren, die immer in der beschreibenden Statistik anzuwenden sind, ebenfalls auf die Breite des Konfidenzintervalls niederschlägt. Und durch die unterschiedlichen Auswahlsätze und die differentiellen Auswahlwahrscheinlichkeiten im Laufe der Panel-Laufzeit steigt diese Varianz an. Je höher die Varianz der Hochrechnungsfaktoren ist, desto geringer ist die Effektivität der zu Grunde liegenden Fallzahlen und desto breiter sind die abzuleitenden Konfidenzintervalle. Um diese Auswirkungen deutlicher zu beschreiben, werden sog. „Effektiven Fallzahlen“ berechnet, die die erhobenen Fallzahlen so reduzieren, dass ihre Zahl derjenigen entspricht, als hätten die Gewichte eine Varianz von Null.

In Abbildung 5.1. ist die Entwicklung der Fallzahlen abgetragen. In den letzten Jahren wurden ca. 12.000 Haushalte insgesamt in jeder Welle befragt. Die zweite Zeitreihe darunter zeigt dagegen die effektiven Fallzahlen. Sie liegen mit Werten um 5.000 deutlich unter denen der erhobenen Haushalte. So ergab sich für die Welle 23 ein Quotient von 0,42; dies ist ein Maß für die Effizienz der Stichprobe, das Werte zwischen 0 und 1 annehmen kann. Es bemisst den Preis, der für die Zusammenführung der Stichproben und für die Randanpassung gezahlt werden musste, um die Erhebung repräsentativ zu gestalten.

Abbildung 5.1: **Entwicklung der Fallzahlen im SOEP nach 23 Wellen**



Da das SOEP nicht an einem einzigen Stichtag (bzw. Berichtstag) erhoben wird, sondern die Feldarbeit sich über mehrere Monate hinzieht, um die Stabilität der wiederholten Teilnahme bzw. Befragung möglichst hoch zu halten, kann das SOEP auch für unterjährige Analysen anhand von Tagesangaben (z. B. Zufriedenheit mit dem Leben) bzw. Wochenangaben (z. B. Arbeitszeit) 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- bzw. Stichwochenfälle zur Verfügung. Normalerweise sind im Laufe des März bereits rund 50% aller Interviews abgeschlossen.

Tabelle 5.2 zeigt, dass über 43.000 Beobachtungen von Personen von mindestens 12 Monaten vorliegen, die ohne eine Lücke in Form von Monatsangaben erfasst sind. Auf der anderen Seite liegen für immerhin 2704 Fälle vollständige Monatsangaben – ohne jede Lücke – für 23 Jahre (276 Monate) vor. Ein Zeitraum von mindestens 15 Jahren (180 Monate) wird für 7.216 Personen erfasst. Der Anhang zeigt eine beispielhafte Analyse (vgl. Berger 2007).

Tabelle 5.2
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	2704	3611	7216	10945	25758	43593
A	2278	2969	4166	5797	8174	12076
B	426	642	1106	1817	2789	4440
C	.	.	1944	2790	4028	5869
D	.	.	.	541	856	1366
E	.	.	.	.	1309	2203
F	.	.	.	.	7110	12184
G	.	.	.	.	1492	2839
H	.	.	.	.	.	2616

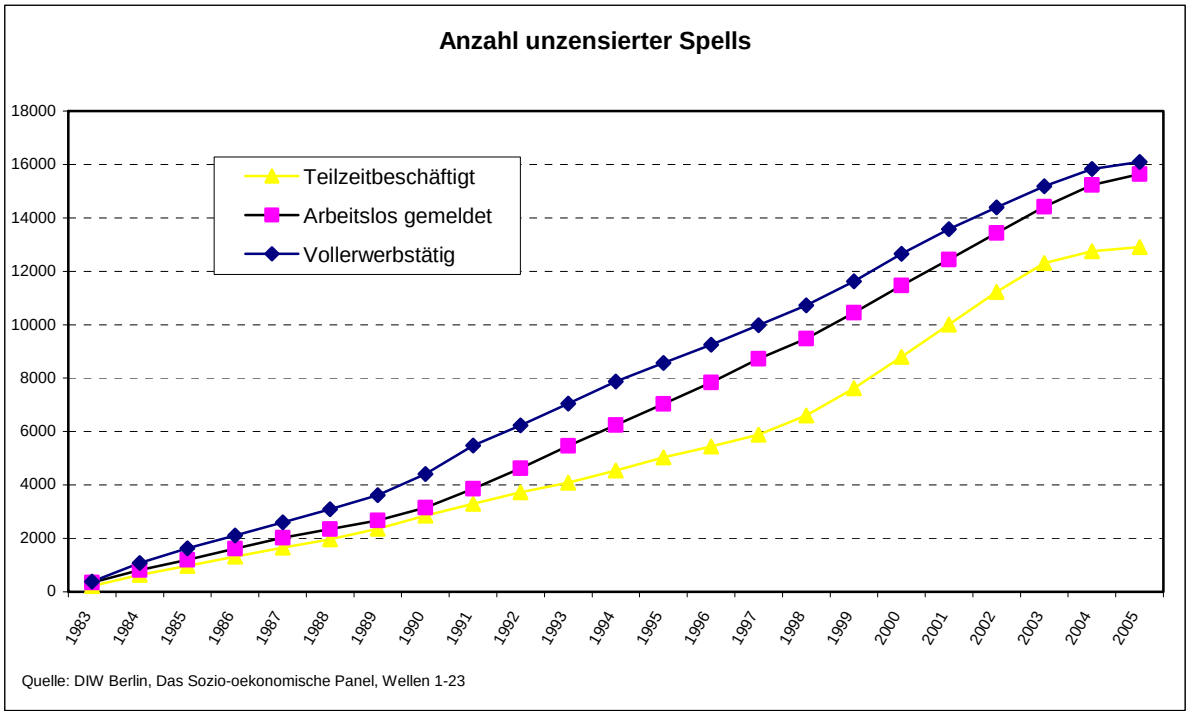
Quelle: Das Sozio-oekonomische Panel (Wellen Q - W); eigene Berechnungen.

Freilich werden zu einem Befragungs-Zeitpunkt nicht nur zum Stichtag bzw. einer normalen Woche Daten erhoben, sondern auch retrospektiv über das vergangene Kalenderjahr. Nur so können z. B. differenzierte Jahreseinkommen konstruiert werden (vgl. Grabka 2007; Frick/Grabka 2003). Außerdem kann aufgrund von „Kalender-Angaben“ zu Aktivitäten, die pro Monat ausgeübt werden (z. B. Erwerbstätigkeit,, Ausbildung, Wehr- oder Zivildienst), Daten über „Spells“, d. h. zeitliche Lage und Länge dieser Aktivitäten erzeugt werden. Damit stehen seit Beginn des Erhebungszeitraums (Januar 1983 – erhoben in der ersten Welle im Frühjahr 1984) bis Dezember 2006 für maximal 276 Kalendermonate Aktivitätsangaben (Spelldaten) zur Verfügung. Für 2.704 Personen liegen diese tatsächlich ohne Missings für diese 276 Monate vor. Für fast 11.000 Personen liegen für mindestens 120 Monate (10 Jahre) und für mehr als 25.000 Personen gibt es Kalendarien die wenigsten 60 Monate umfassen.

Eine weitere für (potentielle) Nutzer der SOEP-Daten aussagekräftige Darstellung der Reichhaltigkeit und Aussagekraft der Spell-Daten ergibt sich, wenn man die Zahl nicht-zensierter Spells darstellt, d. h. die Zahl von Aktivitäts-Episoden, deren Beginn und deren Ende beobachtet wurde (die also weder links- noch rechtszensiert sind). Wenn z. B. ein Befragter in den Erhebungsjahren 1990 und 1991 für die Monate 7/1990 bis 3/1991 Arbeitslosigkeit als Aktivität angibt, und er davor erwerbstätig und danach im Ruhestand war, dann ist das ein Spell von 9 Monaten Dauer für den Aktivitäts-Typus Arbeitslosigkeit. Abbildung 5.3 zeigt

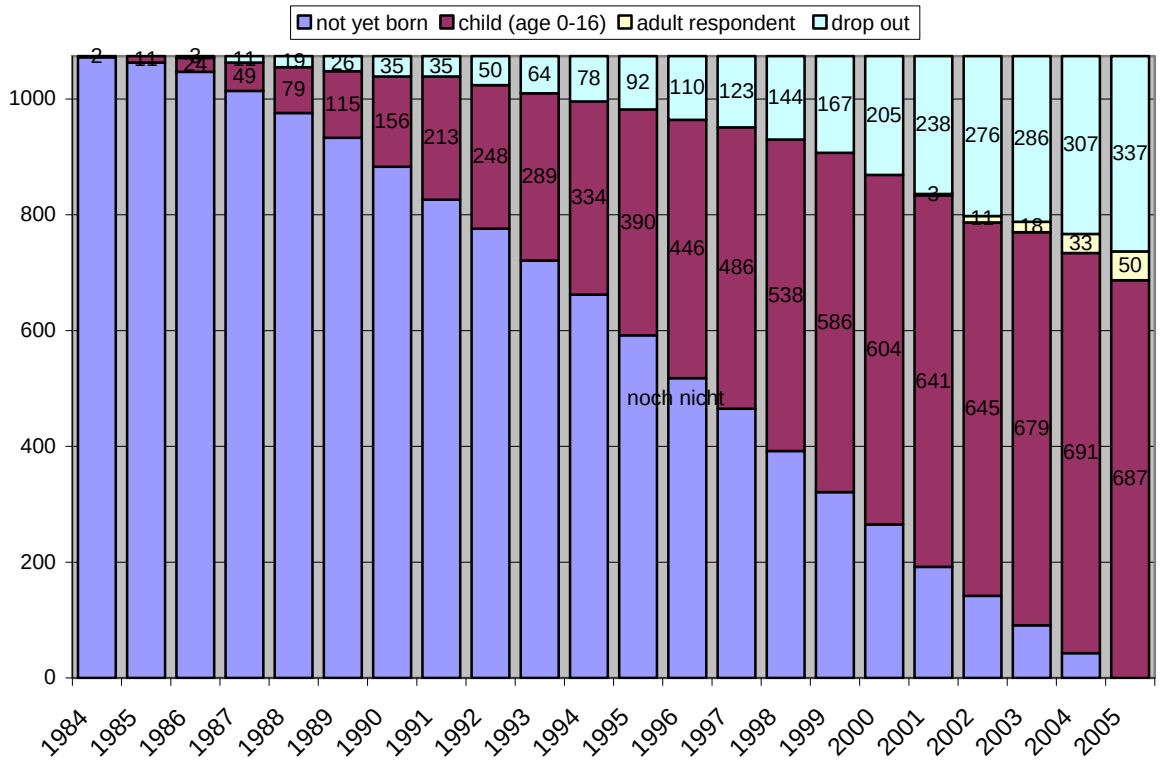
beispielhaft für drei Aktivitäten die kontinuierliche Zunahme der unzensierter Spells auf Monatsbasis.

Abbildung 5.3
Anzahl unzensierter Erwerbstätigkeit-Spells



Die besondere Power eines Haushaltspanels für die Analyse von Ereignissen und Lebensläufen wird z. B. auch an der Geburt von „SOEP-Enkeln“ deutlich. Dies sind Kinder, deren Eltern am SOEP teilnehmen und von denen auch ein Großelternteil in der SOEP-Stichprobe enthalten ist. Abbildung 5.4 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“. Im Jahr 2005 sind noch mehr als zwei Drittel der erfassten Kinder im Erhebungsbestand.

Abbildung 5.4
Enkelkinder im SOEP



6 Datenstruktur

Seit der ersten Erhebung der Daten des Sozio-oekonomischen Panels (SOEP) im Jahr 1984 wurde die Erhebungsweisen (Stichproben, Ziehungsmethoden und Fragebögen) wie auch die Verarbeitungsweisen (Datenstruktur, Distribution und Auswertungsmethoden) sukzessive verändert und permanent erweitert. Mit der Verfügbarkeit von mehr als 20 Wellen gehören nunmehr vergleichbar aufbereitete Zeitreihen- und Längsschnittdaten zum Kernbestand der SOEP-Anwendungen und der Datenweitergabe.

Der SOEP-Datensatz besteht mit der Datenweitergabe 2006 aus 282 unterschiedlichen Datensätzen (Files), die zusammen über 4 Millionen Beobachtungen enthalten, in denen 37.686 Variablen gespeichert sind. Im Grundsatz werden alle direkt erhobenen Personen- und Haushaltsdaten (_P und _H)[19] sowie die damit korrespondierenden Feldinformationen[20] (_PBRUTTO und _HBRUTTO) Jahr für Jahr als Querschnittsdatensätze abgelegt und als solche auch nahezu unverändert an die Nutzer weitergegeben. Von der allgemeinen Weitergabe ausgeschlossen sind lediglich Klartextangaben und kleinräumige Regional- Nachbarschaftsinformationen, um Reidentifikationsmöglichkeiten von vornherein auszuschließen. Derartige Informationen können nur unter spezifischen Datenschutzvorkehrungen vor Ort ausgewertet werden (siehe Kapitel 6.5).

Informationen zu Kindergarten und Schulbesuch zu den noch nicht direkt 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 sie Informationen von nicht direkt befragte Personen (Kinder unter 16 Jahren) enthalten. Die zugehörigen Fragen über diese Kinder (z.B. Art der besuchten Schule) sind zwar vom Haushaltsvorstand beantwortete worden, die Daten werden jedoch auf die Personenebene (der Kinder) transferiert weitergegeben.

[19] Der Unterstrich (_) steht hierbei für den wellenspezifischen Präfix.

[20] Feldinformationen liegen für alle Haushalte und Personen vor, die zur Jahres-spezifischen Bruttopopulation gehören, diese setzt sich zusammen aus der Vorjahrs-Population, minus den Gestorbenen, diejenigen die in das Ausland verzogen sind und auch ohne jene Personen die im letzten Jahr endgültig verweigert haben. Hinzu kommen jedoch Neugeborene und sonstige neu im Haushalt lebende Personen.

Die Variablenbezeichnung verweist im SOEP bei den Interviewdaten 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 vergleichbar aufbereitete Daten mit einheitlichen Namen generiert (_PGEN, _HGEN, _PEQUIV, _PKAL), die ebenfalls als jahresbezogene Datensätze zur Verfügung stehen. 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 bleiben unverändert erhalten, wogegen die daraus abgeleiteten generierten Daten sich infolge veränderter Fragestellungen und Imputationen in jedem Jahr ständig auch rückwirkend bearbeitet und erneuert werden.

6.1 Erhebungsinstrumente

Die Zahl und Vielfalt der eingesetzten Erhebungsinstrumente hat sich seit dem Jahr 2000 deutlich gesteigert. 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 2-3-jährigen Kinder erhoben. In den nächsten Jahren sollen durch weitere Befragungsinstrumente auch das Vorschul- und Schulalter noch detaillierter abgebildet werden (5-6-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 gemessen (siehe Tabelle 6.1.1).

Tabelle 6.1.1
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	YPBRUTTO
Tod	.	Haushaltsvorstand/ Nachrecherche	P	YPBRUTTO, PPFAD
Erinnerung und Renten	.	Hinterbliebener Partner	P-Partner	$P

Bei den Erwachsenen werden neuerdings durch das Einbeziehen von Kognitions- und Greifkrafttests, Recherchen zum Wegzug ins Ausland sowie experimentelle Testsituationen spezifische Verhaltensweise 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.

Tabelle 6.1.2
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)
	Mutterkind-Frabogen III (Kinder im Alter von 5-6)
Andere Erhebungsformen und Tests	
Greifkrafttest	Kognitive Leistungsfähigkeit von 17-Jährigen
Kognitionstests	
Weitere Datenanreicherungen und Recherchen	
Experimente	Verbleibstudie bei Ausfällen
Kleinräumige Regionalinformationen	Wegzug ins Ausland

6.2 Identifikatoren und Populationsabgrenzung

Die Verknüpfung der Vielzahl an Querschnittsdaten erfolgt über Identifikatoren, die auf Personen- und Haushaltsebene zentral (_PPFAD, _HPFAD) abgelegt sind. Personenbezogene Daten enthalten immer auch Identifikatoren der höher aggregierten Einheiten – auf diese Weise können Haushalts- und Personeninformationen durch Aggregation oder Disaggregation flexibel miteinander verknüpft werden. Im SOEP werden pro Erhebungsjahr drei grundlegende Identifikatoren unterschieden, die sich bei spezifischen Datensätzen noch weiter differenzieren: die Case-Id, die aktuelle Haushaltsnummer, sowie die unveränderliche Personennummer.

Die **Case-ID** [HHNR] ist die grundlegende Ziehungseinheit des SOEP. Sie umfasst die zuerst vergebene Haushaltsnummer für jeden Haushalt eines jeden Samples und ist im ersten Erhe-

bungsjahr identisch mit der aktuellen Haushaltsnummer. Die Case-Id ermöglicht nicht nur, Personen- und Haushaltsabspaltungen auf die ursprüngliche Ziehungseinheit zurückzuverfolgen; auf dieser Ebene werden auch unter Heranziehen der Sample-Point-information Varianzschätzer zur Bestimmung von Konfidenzbändern (Randomgroup - / Jackknife - / Bootstrap – Verfahren) gebildet.

Die **wellenspezifische aktuelle Haushaltsnummer** (HHNRAKT bzw. _HHNR) ist erforderlich, um Personendaten mit den jeweils aktuellen Haushaltsinformationen zu verknüpfen. Infolge von Aus- und Umzügen ist dieser Schlüssel jedoch für die Personen über die Zeit variabel.

Die **unveränderliche Personennummer** ist der zentrale Identifikator für personenbezogenen Verknüpfungen über die Zeit. Er wird zusammen mit den anderen Primärschlüsseln für jede Person beim Eintritt in den Datenbestand des SOEP generiert, unabhängig davon, ob diese Person einen Fragebogen ausgefüllt hat oder nicht, also auch für Kinder ab 0 Jahren.

Für spezifische zeitabhängige Datenformate (s.u.) sind weitere Identifikatoren wie das Erhebungsjahr oder die fortlaufende Nummer des Ereignisses pro Person (Spellnummer) notwendig, um die Daten auf verschiedenen Ebenen eindeutig miteinander verknüpfen zu können.

Tabelle 6.2.1
Identifikatoren im SOEP

Ids	**Beschreibung**	**Anwendung**
Case-Id [HHNR]	Unveränderliche HH-Nummer (Zuerst vergebene HH-Nummer je Sample)	bezeichnet den ursprünglich gezogenen Haushalt, aus dem sich nach Abspaltung alle weiteren Haushalte ableiten
HID [HHNRAKT / _HHNR]	Aktuelle HH-Nummer	gewährleistet die Verknüpfung zwischen Haushalten sowie von Personen und Haushalten im jeweiligen Jahr
PID [PERSNR]	Unveränderliche Personennummer	erlaubt die Verknüpfung von Personeninformationen über die Zeit
[SPELLNR]	weitergehende Differenzierungen der Population nach Ereignissen	erlaubt die Verknüpfung von Personen- und Ereignisdaten
SVYYEAR [ERHEBJ]	Erhebungsjahr	Zusätzlicher Verknüpfungsschlüssel bei zeitabhängigen Datenformaten (long-form).

Die wellenübergreifenden Dateien PPFAD und HPFAD, die im Zuge der Datenaufbereitung generiert und jährlich erweitert werden, beinhalten neben den Identifikatoren auch zeitlich variable Angaben zum Einsatz von Erhebungsinstrumenten und zur Verfügbarkeit von personen- oder haushaltsbezogenen Informationen sowie zeitinvariante Angaben (Stichprobenzugehörigkeit, Geschlecht, Geburtsjahr). Zudem werden für jede Person deren Eintritt und Erstbefragung festgehalten sowie deren Austritt und Letztbefragung laufend aktualisiert. In gleicher Weise sind auch die Hochrechnungsfaktoren (PHRF, HHRF) abgelegt.

Die Populationsabgrenzung erfolgt mit Hilfe der Variablen _NETTO und _POP: Erstere verweist auf die Art der verfügbaren Personeninformationen (Gehört die Person im aktuellen Jahr zur Panelpopulation? Liegt ein Personeninterview vor? Gab es weitere Fragebögen, die von der Person zusätzlich ausgefüllt wurden? Liegt nach einem temporären Ausfall ein nacherhobener Lückefragebogen vor? etc.); die Variable _POP grenzt zudem die Personen in Privathaushalten von der im Zuge der weiteren Panelbefragung erfassten Anstaltsbevölkerung (z.B. Altenheim) ab und differenziert nach der Nationalität der Bezugsperson im Haushalt.

Neben den wellenübergreifenden Files [PPFAD und HPFAD], die vor allem der Steuerung der über die Jahre hinweg erfassten Querschnittsdaten dienen, werden auch weitere Informationen von vornherein mit zeitlichem Bezug abgelegt. Die einfachste Form bilden dabei die Biografiedaten, die überwiegend nach inhaltlichen Schwerpunkten untergliedert (eigene Kindheit, soziale Herkunft, demografische Ereignisse, Zuwanderung, Eltern) als über die Zeit kumulativ erhobene Personendaten (BIOSOC, BIOJOB, BIOIMMIG, BIOPAREN) bereitgestellt werden. Andere ebenfalls kumulativ abgelegten Biografieinformationen werden mit eigenen Befragungsinstrumenten erfasst (Mutter-Kind-Fragebogen I+II: BIOAGE01, BIOAGE03; Jugendfragebogen: BIOAGE17), direkt aus den Personendaten abgeleitet (HEALTH) oder aber unter Heranziehen der laufenden Personeninformationen jährlich aktualisiert als Ereignisse (Spells) abgelegt (PBIOSPE, ARTKALEN, BIOMARSM, BIOMARSY).

6.3 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 6.3.1).

Bei dem wide-format werden zu jedem Jahr die Zahl der Variablen um eine weitere jahresspezifische Variable zeilenweise ergänzt. 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. Typische Anwendungsbeispiele für dieses Datenformat sind PPFAD und HPFAD.

Bei dem 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 zusätzlicher Identifikator 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. 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, Greifkraft oder auch zu Vermögensbilanzen werden in diesem Format gepflegt.

Bei der Abbildung der Daten im spell-format werden Zustände 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 sozio-demografische Informationen zum Erwerbsverlauf (Artkalen, PBIOSPE) und Familienstandsänderungen (Biomarsy, Biomarsm).

Tabelle 6.3.1
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

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 6.3.2): Auf der Ebene der Cases werden – soweit verfügbar – die Bruttofiles zur Stichprobenziehung sowie Design-Informationen zur Ziehungswahrscheinlichkeit bereitgestellt. Auf Haushaltsebene werden wellenübergreifend HPFAD und HHRF, regionsspezische Informationen sowie (früher) Verlaufsinformationen zu Sozialhilfekarrieren bereitgestellt. Personendaten werden auf verschiedenen Erhebungsebenen (Bevölkerung insgesamt, Befragungspersonen, Kinder, etc.) mit zeitübergreifenden Informationen ausgeliefert. Ereignisse – die kleinste Erhebungseinheit im SOEP – werden sowohl jahres- als auch monatsweise geführt.

Tabelle 6.3.2
Übersicht über Datenstruktur und Zeitbezug im SOEP

Zeit-Format:	**Unverbundene Querschnitte**	**WIDE-Format**	**LONG-Format**	**SPELL-Format**
Population ...	Population zum Zeitpunkt t	Population im Zeitraum T	Beobachtuns-Einheiten in T	Ereignisse in T
Cases	HBRUTT__	SAMP; Varianz		
HH-alle	_HBRUTTO	HPFAD, HHRF; KKZ, ROR, GGKBOU	REGION	
HH-real.	_H; _HGEN			Sozkalen
Pe-alle	_PBRUTTO	PPFAD, PHRF		[Eintritt, Austritt]
Pe-Befragte	_P; _PGEN; _PKAL		Berufe, Branche; Bildung; Klartext	[Erstbefr, Letztbefr]; BIO...
Pe-Kinder	_KIND			
Pe-Ausfälle		Ypbrutto		
Ereignisse-(J)				PBiospe; BioMarsY
Ereignisse-(M)				Artkalen; BioMarsM

Die Übersicht über die im SOEP über die Jahre hinweg insgesamt verfügbaren Informationen, die Itemkorrespondenzen der im Zeitverlauf erfassten Variablen sowie vorgefertigte Programme zu deren selbständiger Aufbereitung und Analyse werden über ein weiteres Datensystem – SOEPinfo – bereitgestellt. Infolge der unterschiedlichen Erhebungseinheiten, der vielfältigen Erhebungsinstrumente sowie der im Laufe der Jahre kumulativ anwachsenden Vielfalt an Informationen ist inzwischen die Komplexität bei der Verarbeitung der Daten nicht nur für neue Nutzer erheblich angewachsen (vgl. Abschnitt 6.6 unten).

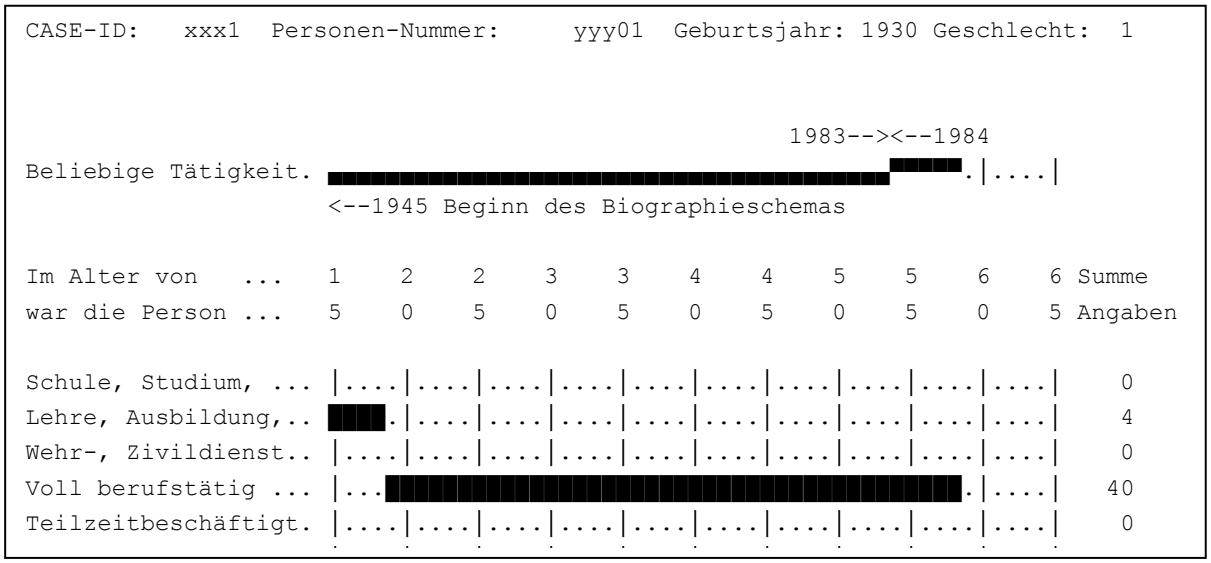
6.4 BIOSCOPE und NEWSPELL

Biographien werden in der Datenbank des SOEP als Spelldaten abgelegt. Dazu werden für jede Person oder jeden Haushalt der Beginn und das Ende eines definierten Zustands abgelegt. Diese Darstellungsform ist vorteilhaft, da auf diese Weise unkompliziert auch sich überlappende Zustände beschrieben werden können. So einfach aber diese Darstellungsform ist,

die Auswertung von Spelldaten ist nicht trivial, da diese keine sog. Rechteckdaten sind, die einfach zu analysieren wären. Wie kompliziert ein Datensatz für eine Person sein kann, lässt sich sehr einfach mit dem Tool BIOSCOPE optisch nachvollziehen. Dieses Programm zeigt für jede Person seinen auf Jahresbasis beruhenden Aktivitätskalender an. Es wird standardmäßig mit den Daten weitergegeben.

Zwei Beispiele: Das erste Beispiel zeigt den Idealfall eines Kalendariums[21]. Es ist geschlossen und jedes Altersjahr ist nur einmal mit einer Tätigkeit besetzt.

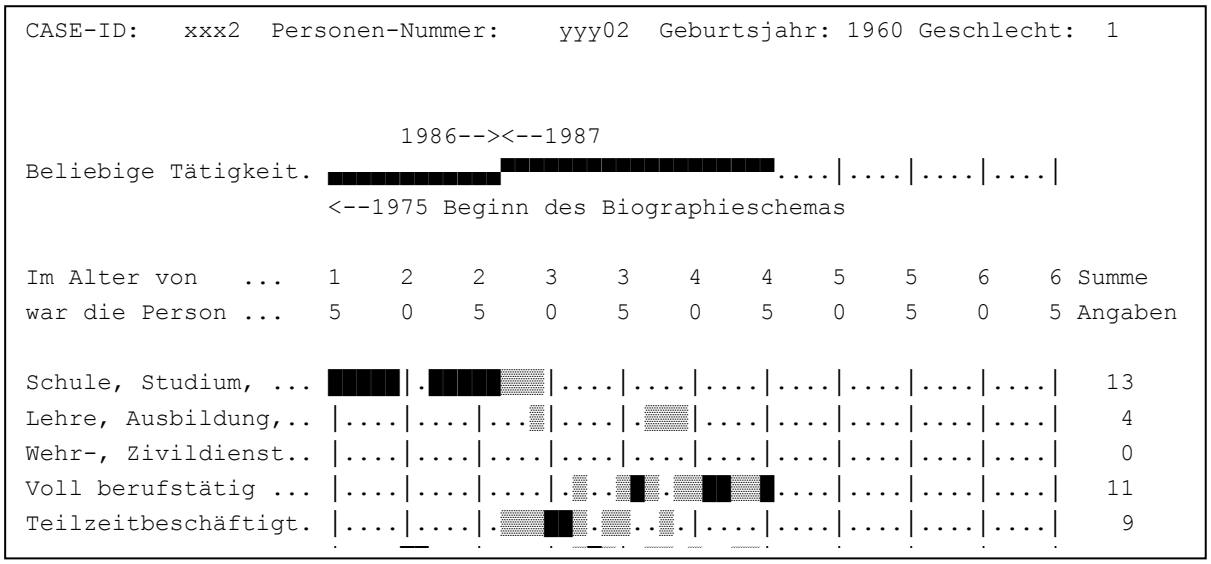
Abbildung 6.4.1
Idealfall Kalendarium



Das zweite Beispiel zeigt eine Erwerbsbiographie, die zwar vollständig ist, aber eine Vielzahl von Doppelnennungen für einzelne Jahre anzeigt. Nach den Angaben des Befragten war dieser im Alter von 38 Jahren in Ausbildung, teilzeitbeschäftigt und auch arbeitslos gemeldet. Eindeutige Ereignisse sind durch schwarze, Mehrfachnennungen durch graue Blöcke gekennzeichnet.

[21] Die Darstellung ist weitgehend selbsterklärend. Im Programm BIOSCOPE wird sie ausführlich erläutert.

Abbildung 6.4.2
Uneindeutiges Kalendarium



Eine Analyse dieses Schemas ist nicht einfach. Hier setzt das Programm NEWSPELL an.

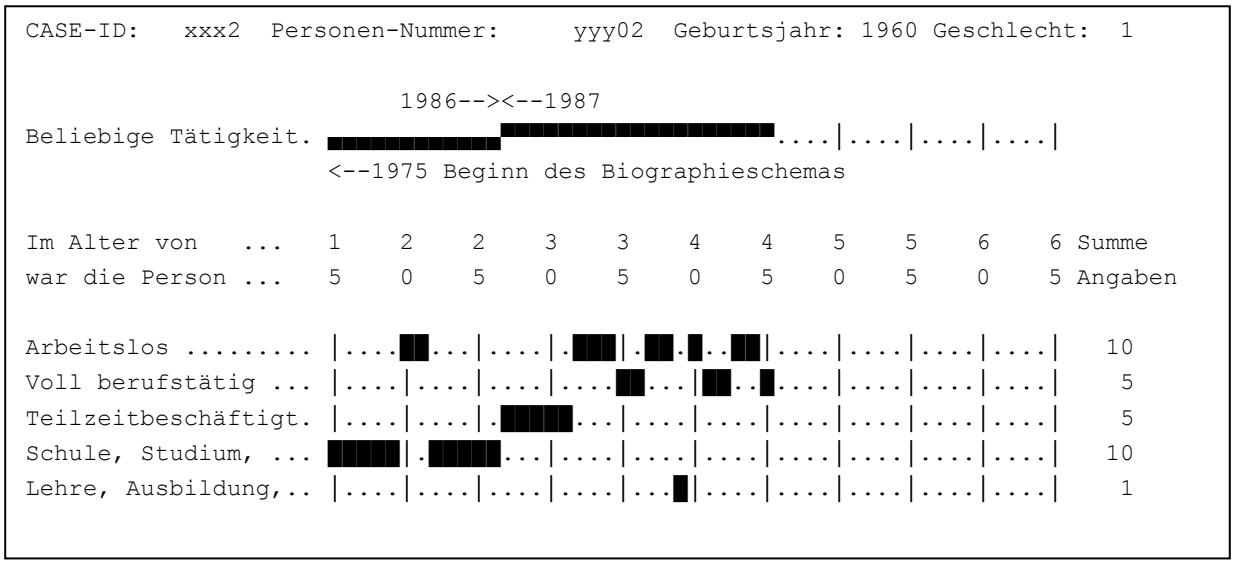
NEWSPELL ermöglicht die Manipulation von Spelldaten. Seine wesentlichsten Eigenschaften sind:

1. *Eindeutigkeit*: Die Erzeugung eindeutiger Kalendarien durch Eliminierung von Doppelnennungen unter Berücksichtigung einer vom Anwender festgelegten Prioritätsskala.
2. *Aggregation*: Zwei oder mehrere Ereignisse können zu einem einzigen zusammengefasst werden.
3. *Disaggregation:* Zwei oder mehrere gleichzeitig auftretende Ereignisse können als eigenständiges Ereignis neu definiert werden

Zum Einsatz von NEWSPELL ist lediglich ein einfaches Kommando-File zu erstellen. Auf der Weitergabe-DVD ist eine ausführliche Dokumentation von NEWSPELL zu finden.

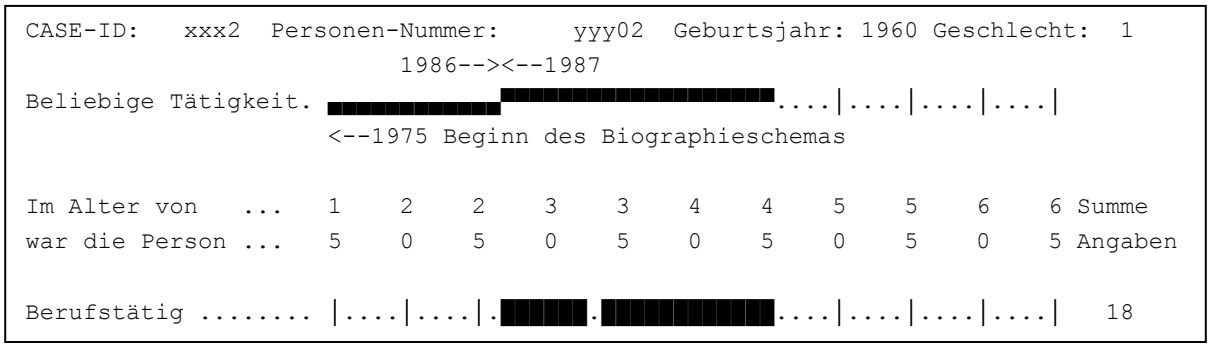
Ausgehend vom Beispiel in Abbildung 6.4.2 ist lediglich die Hierarchie der Ereignisse in der angegebenen Reihenfolge festgelegt worden und dadurch ein eindeutiges Spellsystem erzeugt worden.

Abbildung 6.4.3
Bereinigtes Kalendarium 1



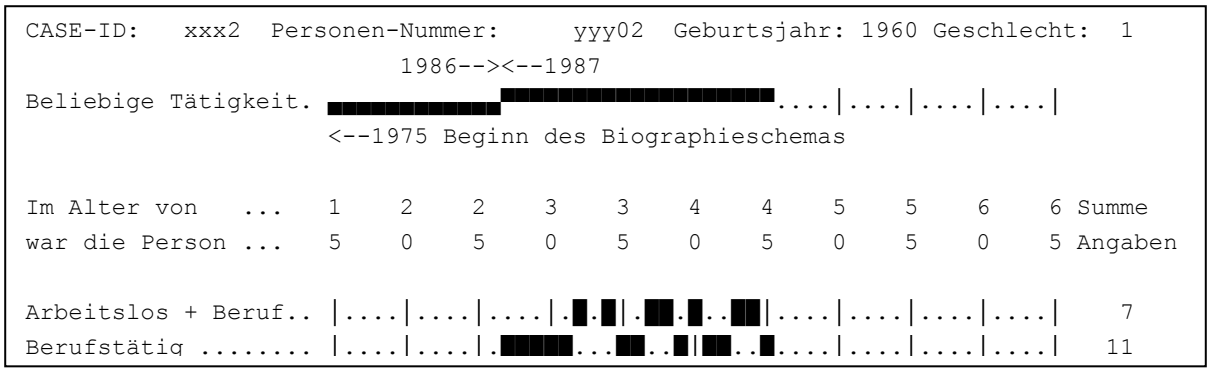
Im nächsten Beispiel sind die Ausbildungstätigkeiten und die Erwerbstätigkeiten zusammengefasst worden.

Abbildung 6.4.4
Bereinigtes Kalendarium 2



Sind das gleichzeitige Auftreten von Erwerbstätigkeit und Arbeitslosigkeit von Interesse, kann aus den Ursprungsdaten das nachstehende Kalendarium erzeugt werden.

Abbildung 6.4.5
Bereinigtes Kalendarium 3



Die Ergebnisse können wieder als Spelldateien ausgegeben werden. Da die erzeugten Kalendarien immer eindeutig sind, ist aber auch eine Ausgabe in Zeitreihenform möglich, die eine einfache Weiterverarbeitung gestatten.

Beide Tools sind MS-DOS-Programme, können aber ohne Schwierigkeiten unter Windows angewendet werden. Während BIOSCOPE als eigenständiges Programm aufgerufen werden sollte, kann NEWSPELL von Statistik-Programmen wie z.B. STATA aufgerufen werden.[22]

6.5 Verknüpfungsmöglichkeiten mit regionalbezogenen Kontext-Informationen

Das SOEP bietet eine Vielzahl an Möglichkeiten, regionalbezogene Informationen bis hin zu „Nachbarschaftsdaten“ bei der Analyse zu berücksichtigen. Mit Hilfe der regionalen Zuordnung des Haushaltes ist es möglich regionale Indikatoren auf der Ebene der Bundesländer (NUTS-1), der Raumordnungsregionen (oder NUTS-2), der Kreiskennziffer, 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.[23] Auf Grund der damit einhergehenden

[22] Es wird beispielsweise mit STATA ein Eingabefile für NEWSPELL erzeugt, NEWSPELL aufgerufen und anschließend werden die Ergebnisfiles von STATA weiterverarbeitet.

[23] 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).

erhöhten datenschutzrechtlichen Sensibilität der Daten müssen je nach Ebene der Regionalinformationen unterschiedliche Sicherheitsvorkehrungen eingehalten werden.

Der Standarddatensatz, der mit Abschluss eines Datenweitergabevertrages erhältlich ist, enthält als regionale Information lediglich eine Ost-West-Unterscheidung und das Bundesland.[24] Eine Beschreibung des Regionstyps liegt im Standarddatensatz als Gemeindetyp (Boustedt oder BIK) und als politische Gemeindegrößenklasse vor.

Die Raumordnungsregionen oder der NUTS-2 Level muss aus datenschutzrechtlichen Gründen gesondert bei der SOEP-Gruppe angefordert werden. Wenn ein spezielles Datenschutzkonzept von beiden Datenschutzbeauftragten (des Nutzers und des Datengebers, d.h. des DIW Berlin) angenommen wurde, werden die zusätzlichen Daten dem Nutzer zur Verfügung gestellt und er kann an seinem Arbeitsort die Daten analysieren. Tiefer gegliederte regionale Schlüssel können aus datenschutzrechtlichen Gründen nicht dem einzelnen Wissenschaftler an seinem Arbeitsort bereitgestellt werden.

Es besteht jedoch zum einen die Möglichkeit Gastarbeitsplätze am DIW zu nutzen (hierbei könne alle regionalen Informationen genutzt werden) oder per Fernrechen-Zugang (SOEPremote) zusätzlich die Kreiskennziffern zu nutzen. Bei der Nutzung von SOEPremote hat der Nutzer keinen Zugang zu den Daten, sondern schickt seine speziell aufbereitete Stata Syntax[25] an eine eigens eingerichtete E-Mail Adresse. Dort wird dieses Programm automatisch auf Verletzungen des Datenschutzes 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 zurückschickt.

Eine besondere Innovation in der Beschreibung der kleinräumlichen Umgebung der Haushalte („Nachbarschaften“) ist dem SOEP durch eine Kooperation mit microm (microm Micromarketing-Systeme und Consult GmbH) gelungen. Mit Hilfe der Adresse wurden die Daten des SOEP mit den mikrogeographischen Daten der microm angereichert. Die Anreicherung des SOEP mit derartigen geographischen Daten stellt eine Erweiterung des SOEP dar, da das Hinzunehmen von feinräumigen Wohnumfeldinformationen sowie soziodemographischen und konsumrelevanten Daten vielfältige neue Analyseansätze erlauben (vgl. Schräpler et al.

[24] Bei den Bundsländern ist bis zum Jahr 2000 auf Grund von zu geringen Fallzahlen das Saarland Rheinland-Pfalz zugeordnet.

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.

6.6 Dokumentation und Metadaten

Ein komplexer Datensatz wie das SOEP ist für einen potentiellen Nutzer nur dann sinnvoll zu nutzen, 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 eine einführende Zusammenfassung (das „SOEP Desktop Companion“), die genaue Dokumentation aller generierten Variablen, die genutzten Fragebögen, Methodenberichte des Feldinstituts 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 vorgefertigte Programme zu deren selbständiger Aufbereitung und Analyse werden über das Datensystem – SOEPinfo – bereitgestellt. Infolge der unterschiedlichen Erhebungseinheiten, der vielfältigen Erhebungsinstrumente sowie der im Laufe der Jahre kumulativ anwachsenden Vielfalt an Informationen ist inzwischen die Komplexität bei der Verarbeitung der Daten nicht nur für neue Nutzer erheblich angewachsen.

Auch SOEPinfo ist somit ein wichtiges Instrument bei der Dokumentation der SOEP-Daten und für jeden über die Internetadresse http://panel.gsoep.de zugänglich. Es enthält in einer nutzerfreundlich aufbereiteten Form zahlreiche Informationen über die Struktur und Inhalte der Daten, wie Datentypen der Variablen, Labelinformationen, sowie Häufigkeitsauszählungen. Darüber hinaus werden aber auch Relationen zwischen einzelnen Variablen und Datensätzen über die Zeit abgebildet. Die Relationen zwischen den Variablen ist die sogenannte Variablen- oder Itemkorrespondenz. In dieser Korrespondenz werden Variablen, die wiederholt in den einzelnen Jahren erfragt wurden unter entsprechenden Items zusammengefasst.

Weil der Zugang zur originalen SOEP-Datenbank geschützt und nur beschränkt möglich ist, bildet SOEPinfo eine eigene Datenbank mit Informationen über die originalen SOEP-Daten.

[25] Derzeit ist SOEPremote nur mit Stata benutzbar.

Es ist also eine Art Metadatenbank und hat aus Datenschutzgründen keinen direkten Zugriff auf die erhobenen SOEP Daten.

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 so genannten Warenkorb (Basket) zwischengespeichert und so später 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 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.

Eine besonders nutzerfreundliche Funktionalität bietet SOEPinfo dem 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 automatisch die jeweils benötigten Variablen zur Stichprobenabgrenzung und Hochrechnung mit einbezogen. Zur Zeit werden die Statistikprogramme SPSS, Stata und SAS unterstützt.

7 Ausblick

Die SOEP-Mikrodaten, die Teil der deutschen und internationalen „Forschungs-Infrastruktur“ sind, werden intensiv genutzt. Bislang liegen über 4500 registrierte Publikationen vor; jährlich kommen inzwischen etwa 400 hinzu. Um Nutzern der Mikrodaten und an Ergebnissen Interessierten einen konzentrierten Überblick über neueste Forschungsarbeiten zu geben, wurde 2007 eine Pre-Print-(Diskussionspapier)-Reihe eingerichtet, die unter www.diw.de/soeppapers abrufbar ist.

Die SOEP-Mikrodaten sind auch mit anderen Paneldaten vergleichend auswertbar, d. h. es sind auf Basis der Mikrodaten internatonale Vergleiche möglich. Um derartige Analysen, die sehr komplex sind, technisch wesentlich zu erleichtern, wurde – an der Cornell University in den USA – das „Cross National Equivalent File“ (CNEF) geschaffen, das gegenwärtig über zwei Millionen Mikrodatensätze für die USA, Kanada, Australien, Großbritannien, die Schweiz und Deutschland enthält (vgl. Frick et al. 2007).

Literatur

Andreß, H.-J.; B. Borgloh; M. Güllner; K. Wilking (2003): Wenn aus Liebe rote Zahlen werden – Über die wirtschaftlichen Folgen von Trennung und Scheidung. Wiesbaden.

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, Berlin.

Beaudry, P.; Green, D.A. (2003): Wages and Employment in the US and Germany: What Explains the Differences? American Economic Review 93 (3)

Börsch-Supan; H. Jürges (2006): Early Retirement, Social Security and Well-Being in Germany. NBER Working Paper 12303, Cambridge: National Bureau of Economic Research (NBER).

Burkhauser, R.V.; P. Giles; D.R. Lillard; J. Schwarze (2005): Until Death Do Us Part: An Analysis of the Economic Well-Being of Widows in Four Countries. Journal of Gerontology, Series B - Social Sciences. 60 (5): 238-246.

Butz, W. P. and Boyle Torrey, B. (2006): Some Frontiers in Social Science, Science 312: 1898-1900

Cooke, L. P. (2007): Persistent Policy Effects on the Division of Domestic Tasks in Reunified Germany, Journal of Marriage and Family. 69 (4): 930-950

Diener, E.; R. E. Lucas; C. N. Scollon (2006): Beyond the hedonic treadmill: Revising the adaptation theory of well-being. American Psychologist, 61: 305-314.

Erlinghagen, M. (2007): Soziales Engagement im Ruhestand: Erfahrung wichtiger als frei verfügbare Zeit, DIW Wochenbericht. 74 (39): 565-570

Frick, J. R. und Grabka, M. M. (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, J. R., Jenkins, J. P., Lillard, D. R., Lipps, O., und Wooden, M. (2007) The Cross National Equivalent File (CNEF) and its Member Country Household Panel Studies, Schmollers Jahrbuch127(4) (in print).

Frick, J.R., Goebel, J., Schechtman, E., Wagner, G.G., Yitzhaki, S. (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

Galler, H.P. (1987) Zur Längsschnittgewichtung des Sozio-oekonomischen Panels. In: Krupp H-J, Hanefeld U (Hrsg.) Lebenslagen im Wandel: Analysen 1987, Band 2 der Reihe: Sozio-oekonomische Daten und Analysen für die Bundesrepublik Deutschland, Verlag, Frankfurt, 295-317

Gerhards, J. und Hans, S. (2006): Zur Erklärung der Assimilation von Migranten an die Einwanderungsgesellschaft am Beispiel der Vergabe von Vornamen. DIW Discussion Paper No. 583. Berlin: German Institute for Economic Research (DIW Berlin)

Gerstorf, D.; N. Ram; R. Estabrook; J. Schupp; G.G. Wagner; U. Lindenberger (2007): Life Satisfaction Shows Terminal Decline in Old Age: Longitudinal Evidence from the German Socioeconomic Panel Study. Manuscript under review.

Goebel, J.; Spieß, C. K.; Witte, N. R. J.; Gerstenberg, S. (2007) Die Verknüpfung des SOEP mit MICROM-Indikatoren: Der MICROM-SOEP Datensatz. Berlin : DIW, Data Documentation 26.

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

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

Jürges, H. (2006): Gender Ideology, Division of Housework, and the Geographic Mobility of Families. Review of Economics of the Household, 4 (4): 299-323.

Jürges, H. (2007): Unemployment, life satisfaction and retrospective error. Journal of the Royal Statistical Society, Series A - Statistics in Society, 170 (1): 43-61.

Kroh, Martin and Spieß, Martin (2006): Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) (1984 until 2005), DIW Berlin Data Documentation 15.

Krupp, H.-J. (2007): Das Sozio-oekonomische Panel - Wie es dazu kam. In: Schwarze, Johannes; Räbiger, Jutta und Thiede, Reinhold (Hrsg.): Arbeitsmarkt- und Sozialpolitikforschung im Wandel - Festschrift für Christof Helberger zum 65. Geburtstag. Hamburg: Verlag Dr. Kovac: 15-39

Lucas, R.E. (2005): Time Does Not Heal All Wounds: A Longitudinal Study of Reaction and Adaptation to Divorce. Psychological Science, 16 (12): 945-950.

Lucas, R.E.; A. Clark; Y. Georgellis; E. Diener (2004): Unemployment alters the set point of life satisfaction. Psychological Science, 15: 8-13.

Lucas, R.E.; A.E. Clark (2006): Do People Really Adapt to Marriage? Journal of Happiness Studies, 7 (4): 405-426.

Pannenberg, Markus (2000): Documentation of the Sample Sizes and Panel Attrition in the German Socio-Economic Panel (GSOEP), DIW Diskussionspapier No. 196.

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

Rendtel, U. (1995) Lebenslagen im Wandel: Panelausfälle und Panelrepräsentativität. Campus, Frankfurt/New York

Rendtel, U.; Pötter, U. (1993): Über Sinn und Unsinn von Repräsentativstudien. Allgemeines Statistisches Archiv (AStA), 77 (3): 260-280

Romeu Gordo, L. (2006): Effects of short- and long-term unemployment on health satisfaction: evidence from German data. Applied Economics, 38 (20): 2335-2350.

Scherger, Simone (2007): Destandardisierung, Differenzierung, Individualisierung - Westdeutsche Lebensläufe im Wandel. Wiesbaden: VS Verlag für Sozialwissenschaften.

Schräpler, J.-P., Schupp, J., Wagner, G.G. (2007) 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 Berlin Research Note (in Vorbereitung).

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

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

Schräpler, Jörg-Peter and Wagner, Gert G. (2005): Characteristics and impact of faked interviews in surveys - An analysis of genuine fakes in the raw data of SOEP. In: Allgemeines Statistisches Archiv (AStA), Jg. 89, Heft 1, S. 7-20.

Schräpler, Jörg-Peter (2007): A Study of Mode-Effects of a Change from PAPI to CAPI. In: Schmollers Jahrbuch (Proceedings of the 7th International Socio-Economic Panel User Conference (SOEP2006), ed. by Ferrer-i-Carbonell, Ada; Grabka, Markus M. and Kroh, Martin), Jg. 127, Heft 1, S. 113-125.

Spieß, M. (2001) Derivation of design weights: The case of the German Socio-Economic Panel (GSOEP), DIW-Materialien / Research notes No. 5

Spieß, M. ; M.Kroh ; R. Pischner; G. Wagner (2008): On the Treatment of Non-Original Sample Members in the German Household Panel Study (SOEP)—Tracing and Weighting. Data Documentation des DIW Berlin. Berlin (in Vorbereitung).

Stutzer, A.; B.S. Frey (2006): Does marriage make people happy, or do happy people get married? The Journal of Socio-Economics, 35 (2): 326-347.

Tamm, M. (2005): The Effect of Poverty on the Health of Newborn Children - Evidence from Germany. RWI Discussion Paper, No. 33, Essen.

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

Zimmermann, A.C.; R.A. Easterlin (2006): Happily Ever After? Cohabitation, Marriage, Divorce, and Happiness in Germany. Population and Development Review, 32 (3): 511-528.

Anhang :
Beispielhafte Analyse unterjähriger Informationen im SOEP

Von Eva Berger

Tabelle A.1 stellt die Ergebnisse einer Analyse dar, die beispielhaft die Möglichkeiten aufzeigt, die das SOEP aufgrund seiner großen Fallzahl und der sich über einen längeren Zeitraum hinziehenden Feldzeit bietet. Analysiert wird die Wahrscheinlichkeit, dass ein Befragter "Große Sorgen um die Umwelt" angibt – und zwar vor und nach dem Reaktorunfall in Tschernobyl am 26. April 1986, der zwei Tage später, am 28. April 1986, bekannt wurde. Die Regression wurde mit den für die Erklärung von Besorgnis üblichen sozio-ökonomischen Variablen durchgeführt. In den Ergebnissen zeigt sich der Einfluss des Interviewzeitraums (vor und nach Tschernobyl) hoch signifikant; die besagte Wahrscheinlichkeit ist nach Tschernobyl um 11,4 % höher. Auch weitere Analysen, die die Interviewzeiträume auf einzelne Monate ausdehnen, zeigen, dass in den beiden Monaten unmittelbar nach dem Unfall die Wahrscheinlichkeit, dass große Sorgen angegeben werden, ca. um 7.3 % ansteigt (Mai und Juni 1986; mit April als Referenzkategorie). Die Ergebnisse sind hoch signifikant. Im Juli und August ist die Zahl der Beobachtungen so gering, dass sich kein signifikanter Effekt mehr zeigen kann (171 bzw. 30 Fälle). Gleiches gilt für den Januar 1987 (4 Fälle). Danach zeigen sich im Februar und März 1987 – wahrscheinlich aufgrund der anlaufenden Berichterstattung über das Ein-Jahres-Jubiläum – ein Anstieg der Sorgen im Vergleich zum Niveau unmittelbar vor der Katastrophe um etwa 18%.

Tabelle A.1
Einfluss des Reaktorunfalls in Tschernobyl auf die Einstellung zur „Sorge um den Umweltschutz“

– Ergebnisse einer logistischen Regressionsanalyse –

Anzahl der Beobachtungen 25367	Große Sorgen um Umweltschutz => 1, sonst => 0	Marginalen Effekte und Signifikanzniveau
Tschernobyl-Dummy	Interview-Zeitraum 1.1.85-27.4.86 => 0 28.4.86-31.12.87 => 1	+0,114 **
	Geschlecht: männlich	-0,032 **
	Alter	+0,002
	Alter * Alter	-0,000 **
	Monatliches Netto-Haushaltseinkommen (logarithmiert.)	-0,015 **
	Behinderung durch Gesundheitszustand	+0,046 **
Familienstand (Referenzkategorie: ledig)	Verheiratet zusammenlebend	-0,024 *
	Verheiratet getrennt lebend	-0,121 **
	Geschieden	+0,005
	Verwitwet	+0,006
Kinderzahl (Referenzkategorie: keine Kinder unter 16 Jahre)	Ein Kind im Haushalt	-0,042 **
	Zwei Kinder im Haushalt	-0,063 **
	Mehr als zwei Kinder im Haushalt	-0,129 **
Erwerbsstatus	Vollzeit beschäftigt	-0,021 **
	Teilzeit beschäftigt	+0,028 *
	In Ausbildung	+0,161 **
	Geringfügig bzw. unregelmäßig beschäftigt	+0,063 **
	Wehr-/ Zivildienst	+0,110 **
	Arbeitslos gemeldet	+0,026
Ausbildungsabschluss (Referenzkategorie: abgeschlossene Berufsausbildung)	Hochschulabschluss	+0,139 **
	Kein Berufsabschluss	-0,124 **

* Signifikant auf 95%-Niveau ** Signifikant auf 99%-Niveau

Log Likelihood = -16852,161 LR Chi² = 1381,04 Pseudo-R² = 0,0394

Quelle: Das Sozio-oekonomische Panel (Wellen 1 – 23); eigene Berechnungen