SOEP Documentation

Release v31

German Socio-Economic Panel (SOEP)

August 29, 2016

Contents

Note: This is a development version. Please visit http://about.paneldata.org/soep/dtc/ for the current version of our Desktop Companion (DTC).

CHAPTER 1

Quick start guide

The aim of this text is to provide users with an overview of the most significant features of the SOEP. At this point, the data themselve is not necessary - we actually propose that first-time users at least skim through this text before opening any of the SOEP datasets. We start by giving an overview of the contents, so any user will qickly have an idea whether the SOEP is useful for her/his research question - or may get new ideas on what to research. A documentation of the (still ongoing) developments in the SOEP over the last 30 years is presented with respect to the most relevant changes in the target population as well as the survey design. The structural elements, common to all data sets, are explained in the fourth section. Here we not only document the structure used since the beginning (now termed "SOEPclassic"), but also give an overview of the more recently introduced "SOEPlong" datasets.

Besides the SOEPcore study (in the two formats "SOEPclassic" and "SOEPlong") the SOEP consists in recent years of even more studies, which are more or less closely connected to the SOEPcore Study. However this brief introduction focuses only on SOEPcore.

1.1 Contents of the study

The SOEP started in 1984 as a longitudinal survey of private households in the Federal Republic of Germany. The central aim then and now is to collect representative micro-data to measure stability and change in living conditions by following a micro-economic approach enriched with variables from sociology and political science (influenced by the "Social Indicator" movement). Therefore the central survey instruments are a houeshold questionnaire, which is responded by the head of a household and an individual questionnaire, which each household member is intended to answer. Furthermore beginning with 1997, there are wave-specific $LELA files (Lebenslauf - engl. life course) containing the biography information as collected in the respective year.

A rather stable set of core questions is asked every year covering the most essential areas of interest of the SOEP:

- population and demography
- education, training, and qualification
- labor market and occupational dynamics
- earnings, income and social security
- housing
- health
- household production
- preferences and values
- satisfaction with life in general and certain aspects of life.

Additionally, yearly topical modules enhance the basic information in (at least) one of these areas by asking detailed questions as documented in Table 1 and Figure 1. These modules for the main part appear in the personal questionnaires; only some of them are additions to the household questionnaire. Starting in the year 2001, the data have become even richer by including several different health measures and well-known psychological concepts as well as age specific questionnaires.

Table 1.1: Overview of Supplementary Questionnaires 1986-2012

Year	Wave	Topic	
1986	C	Residential environment and neighborhood	
1987	D	Social security	transition to retirement
1988	E	Household finances and wealth	
1989	F	Further occupational training and professional qualifications	
1990	G	Time use and time preferences; Labor market and subjective indicators	
1991	H	Family and social networks	
1992	I	Social security (2nd measurement)	
1993	J	Further occupational training (2nd)	
1994	K	Residential environment and neighborhood (2nd); Working conditions; Expectations for the future	
1995	L	Time use (2nd)	
1996	M	Family and social networks (2nd)	
1997	N	Social security (3rd)	
1998	O	Transportation and energy use; Time use (3rd)	
1999	P	Residential environment and neighborhood (3rd); Expectations for the future (2nd)	
2000	Q	Further occupational training (3rd)	
2001	R	Family and social networks (3rd)	
2002	S	Wealth and assets (2nd); Social security (4th); Health (SF12	BMI)
2003	T	Transportation and energy use (2nd); Trust; Time use (4th)	
2004	U	Residential environment and neighborhood (4th); Further occupational training (4th); Risk aversion; Health (2nd)	
2005	V	Expectations for the future (3rd); Big Five; Reciprocity	
2006	W	Family and social networks (4th); Working conditions (ERI); Health (3rd); Grip strength	
2007	X	Wealth and assets (3rd); Social security (5th)	
2008	Y	Further occupational training (5th); Health (4th); Grip strength (2nd); Trust (2nd); Time use (5th)	
2009	Z	Residential environment and neighborhood (5th); Risk aversion (2nd); Big Five (2nd); Globalization and transnationalization; Diseases	
2010	BA	Consumption and saving; Reciprocity (2nd); Health (5th); Grip strength (3rd)	
2011	BB	Family and social networks (5th); Working conditions (ERI) (2nd); Diseases (2nd)	
2012	BC	Wealth and assets (4th); Social security (6th); Health (6th); Grip strength (4th)	
2013	BD	Big Five (3rd); Trust (3rd); Loneliness; Working conditions (ERI) (3rd); Diseases (3rd)	
2014	BE	Health (7th); Risk aversion (3rd); Globalization and transnationalization (2nd); Residential environment and neighborhood (6th)	

R Code to create figure.

Since the year 2000, youths (turning 17 during the survey year) form a new group of respondents with a specific questionnaire suited to their situation. The questions cover their situation at home, including the relationship to their parents and friends. School and job aspirations are a major part, while some of the psychological measures available for the adults (e.g. Big Five, risk aversion) are also taken. Overall, the youth questionnaire provides a broad overview of the individual's situation at a very interesting and potentially influencial point in their life.

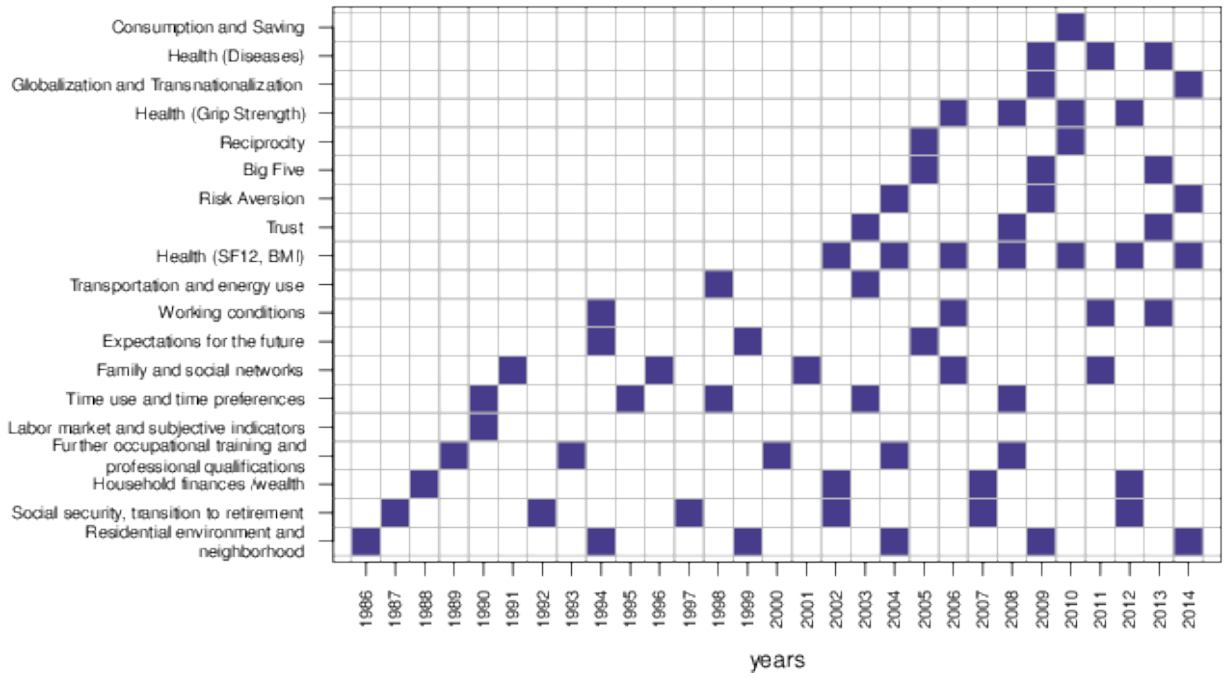


Fig. 1.1: Overview of Supplementary Topical Modules

Since 2003 SOEP also asks parents about their young children, by implementing age specific questionnaires. In 2003, a first questionnaire was added for infants and very young children born during the current or previous survey year. Since then, four additional questionnaires have been added for children in different age groups. In 2012, parents were asked about their children turning 10 during the current survey year for the first time. The topics in these questionnaires vary with the age of the child - for an overview, see Table 2.

Table 1.2: Overview of Proxy Questionnaires for Young Children

Questionnaire	Age Group (years)	First used in	Respondent	Topics
Mother-Child A	0-1	2003	Mother	Child Health; Circumstances of Pregnancy; Child Care
Mother-Child B	2-3	2005	Mother (Father)	Child Health; Child Care; Compentecies
Mother-Child C	5-6	2008	Mother (Father)	Strengths and Difficulties (SDQ); Child Health; Child Care; Activities
Parent-Child D	7-8	2010	Mother and Father	Educational Goals; Upbringing Styles and Goals; Child Care
Mother-Child E	9-10	2012	Mother (Father)	Educational Goals and School Performance; Activities; Child Health; Strengths and Difficulties (SDQ); Child Care

(Note that for all questionnaires except the Parent-Child D the mother is the first choice to fill out the questionnaire, whereas the father is meant to answer only if the mother is not available.)

Especially Table 1 shows that the SOEP evolved in various directions over the course of its existence. New topics will continue to be introduced in future waves of data collection, depending on what is important for the scientific community.

1.2 Target population and samples

The target population covered in the SOEP is defined as the residential population living in private households within the current boundaries of the Federal Republic of Germany (FRG). Because of changes in these boundaries (in 1990) and changes in the residential population due to migration, various adaptations have been applied to the initial sampling structure to keep the sample's representativity. In addition, certain groups have been oversampled to increase the statistical power.

In 1984, the survey started with a sample covering the entire population in then West Germany (FRG), where the five biggest groups of foreigners (the so-called "guestworkers") were oversampled.

> The institutionalized population, in the true sense of the word (hospitals, nursing homes, military installations) is generally not representatively included in new samples. E.g. in 1984 only 57 institutionalized households are included. Later, however, persons from the initial households who have taken up residence temporarily or permanently in institutions of this kind are followed. For a detailed description of the problems in covering this population in the SOEP, see Hanefeld (1987).

The SOEP was expanded to the territory of the German Democratic Republic in June 1990, only six months after the fall of the Berlin Wall. A further addition in 1994/95 was a sample of migrants who came to Germany after 1984, to take the influx of ethnic Germans from former Soviet countries into account. Two samples respresentative of the entire population in Germany were added in 1998 and 2000, to counter effects of panel attrition and to increase the overall sample size. In 2002, a high income sample was added, while in 2006 and 2009, additional refreshment samples were drawn.

To increase the overall sample size SOEP has started adding refreshment samples in 2011. While the first (in 2011) and second (2012) extensions are representative of the whole population, the third (2013) is supposed to explicitly cover migrants. For the fourth extension in 2014, the related study "Families in Germany", covering mainly families, will be integrated into the SOEP.

The different samples in the SOEP are identified by letters: sample "A" refers to the German sample drawn in 1984, "C" to the East Germans from 1990, and so on. Even though these samples are kept separate, the respondents recieved identical questionnaires for the most part and distinctions by sample are usually not be necessary in an analysis.

> However, one of the ideas of SOEP is, that the users have full information available about survey methodological issues and survey design. Which means in this case that you can of course identify the corresponding sample for each observation. In the following section, we present details on each of the samples, which - unless stated otherwise - are multi-stage random samples with regional clusters. The respondent's households are selected by random-walk routines. For an extensive discussion on sampling (and weighting) see @Kroh2012.

1.2.1 The SOEP samples in detail

Sample A "Residents in the FRG" covers persons in private households with a household head, who does not belong to one of the main foreigner groups of "guestworkers" (i.e. Turkish, Greek, Yugoslavian, Spanish or Italian households). Because only a few foreigners are in Sample A it is often called the "West German Sample" of the SOEP. In 1984 it covered 4,528 households with a sampling probability of about 0.0002.

Sample B "Foreigners in the FRG" adds persons in private households with a Turkish, Greek, Yugoslavian, Spanish or Italian household head, which in 1984 constituted the main groups of foreigners in the FRG. Compared to Sample A the population of Sample B is oversampled with a sampling probability of about 0.002. The first wave included 1,393 households in Sample B.

Sample C "German Residents in the GDR" consists of persons in private households where the household head was a citizen of the German Democratic Republic (GDR). This meant that approximately 1.7% of the residential population in the GDR in June 1990 was excluded from the sample as foreigners (who were mostly institutionalized). All in all, 2,179 households represent the starting size of this sample with a sampling probability of about 0.0005.

Sample D "Immigrants" started in 1994/95 with two different samples. In 1994, the first sample D1 had 236 households and in 1995, the second sample D2 had 295 households, leading to a total of 531 households (D1 and D2) in 1995. This sample consisted of households in which at least one household member had moved from abroad to West Germany after 1984. The sampling probability is about 0.0002.

Sample E "Refreshment" was added in 1998, selected from the entire population of private households in Germany. The households were chosen independently from the ongoing panel and its subsamples A through D, with the targets of increasing the number of observations of the general population and preserving its representativity. The selection scheme used for sample E essentially resembles the one used in subsample A. The number of households in the first wave of subsample E was 1,060, with a sampling probability of about 0.00005. With the data distribution of 2012, parts of subsample E have been extracted into the SOEP Innovation Sample.

Sample F "Refreshment" was selected independently from all other subsamples from the population of private households in 2000. The selection scheme was slightly altered compared to the previous addition in Sample E: while the 'German' households (all adults greater or equal 16 in the household have German nationality) were selected with a sampling probability of 0.00028, the 'non-German' households (at least one adult does not have German nationality) where oversampled with a probability of 0.0005. Overall, the number of added households in subsample F's first wave amounts to 6,043.

Sample G "High Income" entered the SOEP in 2002 independently from all other subsamples. The original selection scheme required that the responding households had a monthly income of at least DM 7,500 (EUR 3,835), which - due to the lack of an adequate sampling frame - were identified using a screening procedure. This sample of overall 1,224 households increased the potential for analyses in the high income areas, which previously were difficult to conduct because of low case numbers. The derived sampling probability is about 0.0014. Starting with Wave 2 in 2003, the selection scheme for this subsample was changed such that only households with a net monthly income of at least EUR 4,500 were followed.

Sample H "Refreshment" started in 2006 as a random sample, again independently of all previous subsamples, covering all residential households in Germany. The addition of 1,506 households was drawn with a sampling probability of 0.0001.

Sample I "Incentive sample" started in 2009, where in the first wave, a new incentive scheme was tested to increase participation rates (see also [sec:PanelCare]. The sampling was independent of all other SOEP-samples, adding a total number of 1,531 households to the SOEP. Their sampling probability was 0.00013. This sample remained in the main data distribution for its first two waves (i.e. 2010 and 2011, or waves Z and BA). With the data distribution of 2012, subsample I has been extracted into the SOEP Innovation Sample.

Sample J "Refreshment sample" started in 2011 as a random sample that was drawn independently of all previous subsamples, covering the residential households in Germany. The addition of 3,136 households was drawn with a sampling probability of 0.0002.

Sample K "Refreshment sample" started in 2012 as a random sample, drawn independently of all previous subsamples, covering the residential households in Germany. The addition of 1,526 households was drawn with a sampling probability of 0.0001.

In 2013 a new **migration sample** was added with around 2,700 households drawn by using register information of the German Federal Employment Agency.

1.2.2 Eligibility and follow-up

As mentioned, the SOEP's goal is to be representative of the residential population of Germany. All household members 16 and older are eligible for a personal interview, starting with the youth questionnaire at that age, followed by "regular" person questionnaires thereafter. As years go by, the children of the first wave reach age-eligibility and

become panel members. If they move out and form their own families, they and their new families are still part of the survey. “New” persons become part of the SOEP population due to birth or residential mobility. In case a person enters a SOEP household after the initial wave, this person is asked to fill out the regular person questionnaire if age-eligible, or will be asked to participate once old enough. Thus in the absence of panel attrition the SOEP would be a self-sustaining survey.

The concept of how to follow the respondents and sample members over time is important for the representativeness of the study. The basic principle for follow-up in the SOEP is that all persons participating in a wave of any subsample are to be surveyed in the following years as long as they stay within the boundaries of Germany. This rule also extends to respondents who entered a SOEP-household after the first wave due to residential mobility or birth. If there is a “split-off”, i.e. people move out of the household they were last interviewed in, the members of the new household receive a new household identifier. Table 3 conceptualizes how new sample members and households are realized in the SOEP. Figure 2 shows that as a result of the follow-up concept, up to , several thousand “new” households became part of the SOEP population. The weighting scheme takes into account this complete “follow-up” (see @Kroh2012).

Persons or households who could not be interviewed in a given year are termed “temporary drop-outs”. These are followed until there are two consecutive waves of missing interviews for all household members or a final refusal of the complete household. In the case of a cooperation after a temporary drop-out, the respondent is asked to fill out an additional short questionnaire on central information on employment and demographics during the year of absence.

Table 1.3: Changes to the Sample: Respondents and Households

	Existing Households	New Households
Existing Persons	“classic case”: without change of address entire household moves	Move-out
New Persons	Birth Move-in	Move-in or birth into move-out household

R Code to create figure.

1.2.3 Development of sample sizes

Individuals who refuse participation or are not available for an interview are kept in the so-called “gross” sample of the study as long as they continue to live in households with at least one participating person. Once the entire household declines to respond in two consecutive waves of data collection, all individuals from the household are removed from the SOEP. Table 4 shows the starting sample sizes of samples A through J, the years when the samples were first collected, as well as the percentage of those persons who were eligible for an interview but declined participation (“partial unit non-response”, PUNR) in the first wave. Figure 3 illustrates the development of the number of successful person interviews since 1984. The reduction in the population size for all individual samples is mainly the result of person-level drop-outs, refusals, moving abroad, etc. However, due to new persons moving into already existing households, and children reaching the minimum respondent’s age of 16, and thereby increasing the sample size, this negative development is offset somewhat.

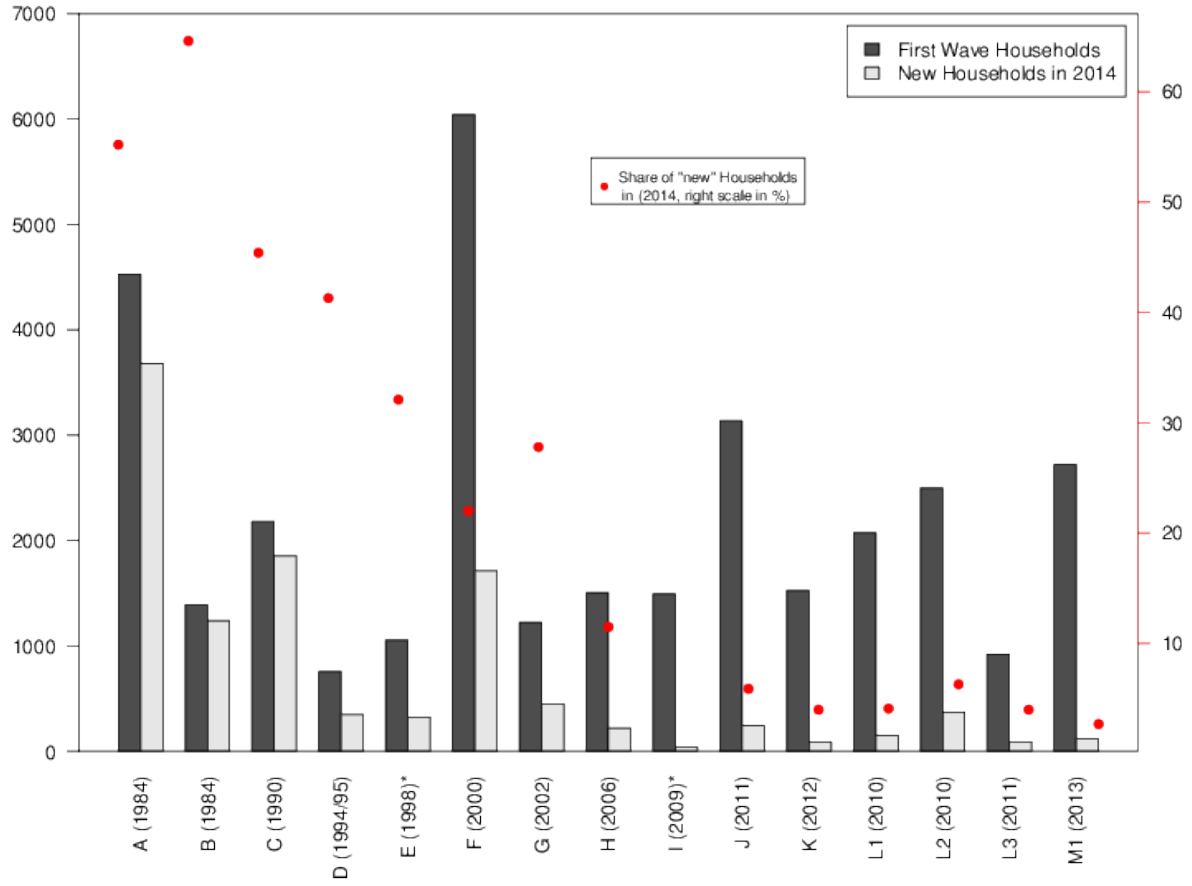


Fig. 1.2: Old and New Households in the SOEP

Table 1.4: Starting Sample Size of the SOEP Samples

Sample	Year	Households (net)	Persons (gross)	Respondents (net)	Partial Unit Non-response (percent)	Children (gross)
A	1984	4,528	11,422	9,076	0.6	2,290
B	1984	1,393	4,830	3,169	0.7	1,638
C	1990	2,179	6,131	4,453	1.9	1,591
D1	1994	236	733	471	2.9	248
D1/D2	1995	541	1,668	1,078	6.1	517
E	1998	1,057	2,446	1,910	3.5	466
F	2000	6,043	14,510	10,880	5.5	2,991
G	2002	1,224	3,538	2,671	6.1	693
H	2006	1,506	3,407	2,616	6.0	623
I	2009	1,495	3,428	2,432	13.4	620
J	2011	3,136	6,873	5,161	9.9	1,147
L1	2010	2,074	7,939	3,770	6.7	3,900
L2	2010	2,500	9,063	4,227	5.1	4,611
L3	2011	924	3,645	1,487	4.2	2,092
K	2012	1,526	3,286	2,473	9.2	563
M1	2013	2,723	8,522	4,964	17.8	2,481

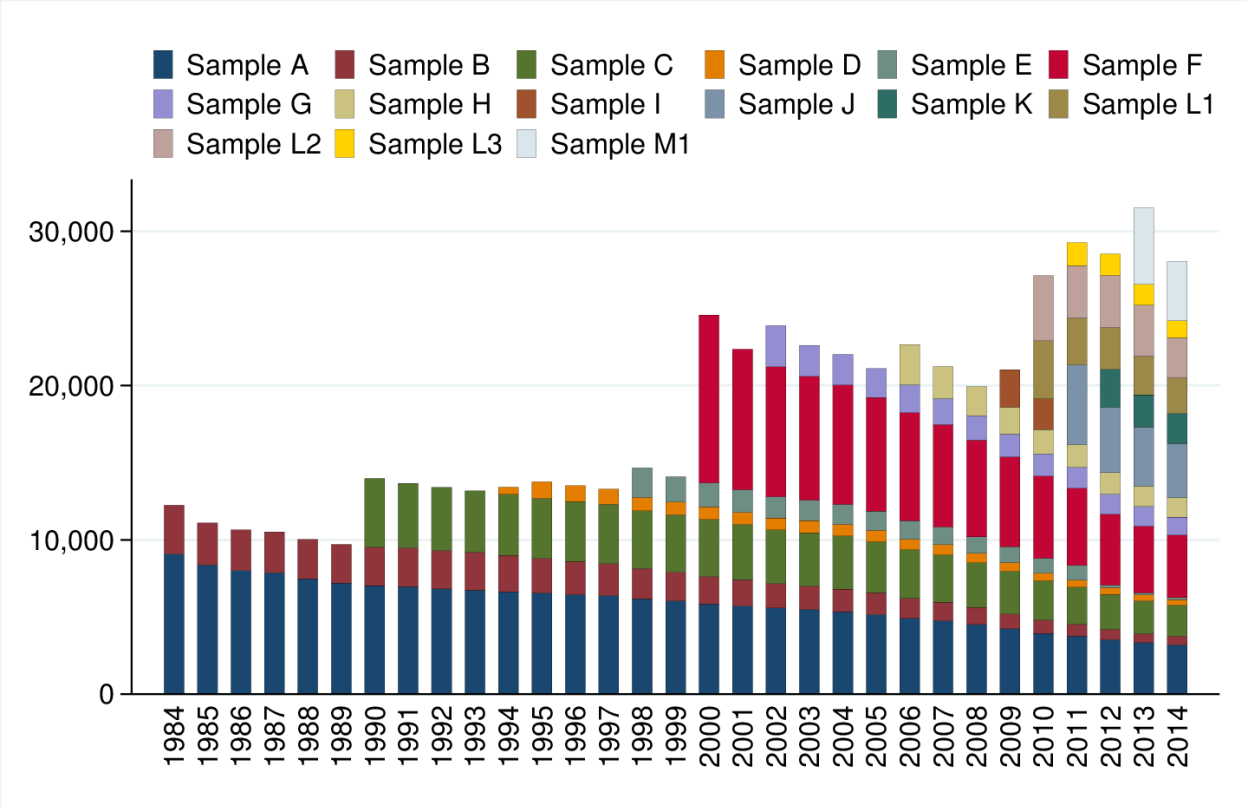


Fig. 1.3: Cross-Sectional Development of Sample Size (Respondents): Samples A-K

Stata Code to create figure.

This cross-sectional view is insufficient when examining the longitudinal development of the sample, which is influenced by different demographic and field-work related factors. As already shown in Table 3, demographic reasons for entering the panel are birth and residential mobility. Analogously, the demographic reasons for a panel exit are

death and moving abroad. Fieldwork related reasons are different, in that they relate to the interaction between the interviewer and the responding household. Respondents are either not reached for an interview (non-contact) or they decline to participate for the current year. Figure 4 illustrates the longitudinal development of first-wave respondents in 1984, as well as their children, of samples A and B.

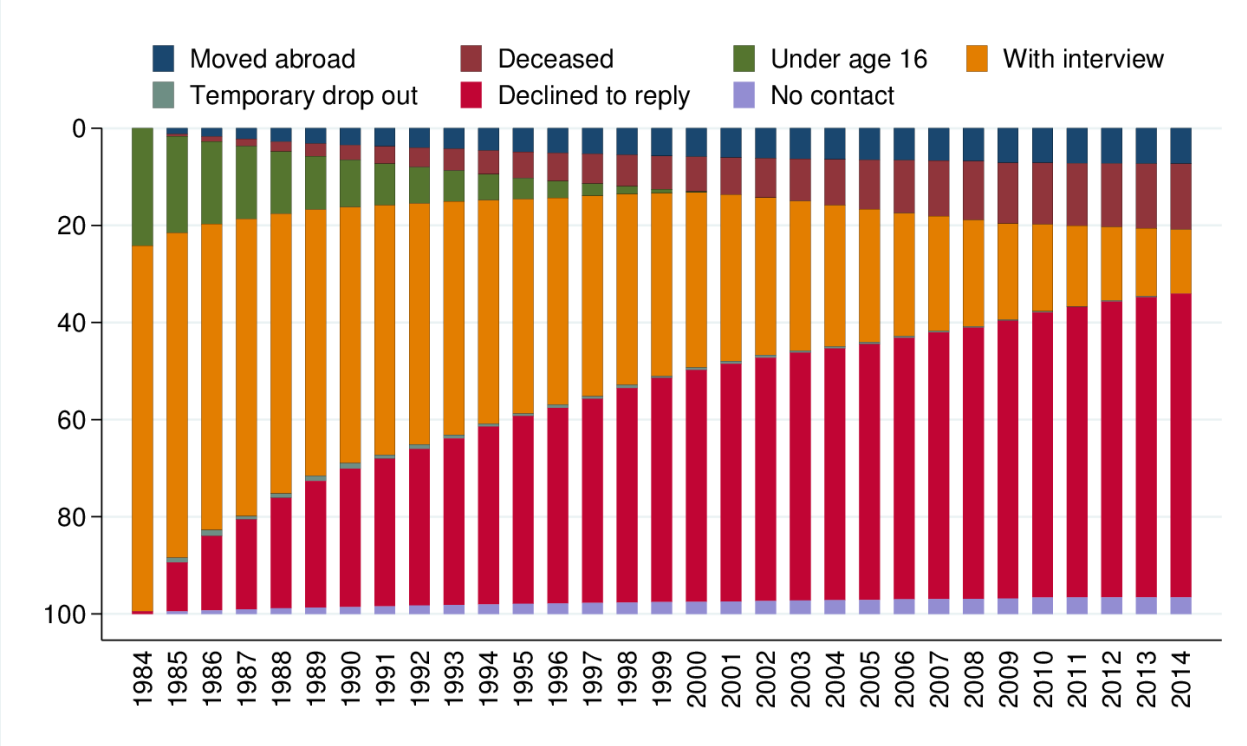


Fig. 1.4: Longitudinal Development of the 1984 Population

Stata Code to create figure.

1.3 Survey design

1.3.1 Survey instruments

The interview methodology of the SOEP is based on a set of pre-tested questionnaires for households and individuals. Interviewers try to obtain face-to-face interviews with all members aged 16 years and over of a given survey household. Thus, there are no proxy interviews for adult household members. Additionally, one person (the so called "head of household") is asked to answer a household related questionnaire covering information on housing, housing costs, and different sources of income (e.g. social transfers like social assistance or housing allowances). This questionnaire also covers some questions on children in the household up to the age of 16, mainly concerning their attendance in day care, kindergarten and school.

The questions in the SOEP are in principle identical for all particpants of the survey to ensure comparability across the participants within any given year (of course, there are differences across years. There are a few exceptions to this rule, which are due to different requirements in the target population. Up to 1996 the questionnaires for the foreigner's sample (B) and immigrant sample (D) covered additional measures of integration or information on re-migration behavior. Between 1990 and 1992, i.e. during the first years of the German unification process, the questionnaire for the East German sample (C) also contained some additional specific variables. Since 1996, all questionnaires are

uniform and completely integrated for all main SOEP samples. The related studies use SOEP related content, but also have specific questions, so the contents may differ to various degrees in every year.

Another type of difference in questionnaires is implemented because first time respondents are not treated identically to those with a repeated interview, since some information does not have to be asked every year unless a change occurred. Additionally, each respondent is asked to fill out a biography questionnaire covering information on the life course up to the first SOEP interview (e.g. marital history, social background, and employment biography).

Additional information - not provided directly by the respondents - can be obtained from the so-called "address logs", which are stored for every year in the `PBRUTTO` and `HBRUTTO` files. Every address log is filled in by the interviewer even in the case of non-response, thus providing very valuable information, e.g. for attrition analyses. For researchers interested in methodological issues these data also contain information on the field work process, e.g. the number of contacts, reason for eventual drop-outs, or the interview mode. For successfully contacted households, the address logs cover the size of the household, some regional information, survey status etc., while the individual data for all household members include the relation to the household head, survey status of the individual and some demographic information.

1.3.2 Survey concepts

Measuring stability and detecting changes means to repeat (almost) identical measures over time. Furthermore, the SOEP-questions capture stability and change by varying with regard to the time dimension, asking about events in the past, the present, and the future. Conceptually, different measurements of time are used:

- Questions about a point in time (present) e.g. current employment status or current levels of satisfaction
- Single retrospective questions on certain events in the past e.g. how often did you change your job during the last ten years?
- Retrospective life event history since the age of 15 (in the past) e.g. employment or marital history
- Monthly calendar information on income and labor market participation (in the past) e.g. employment status January through December last year
- Questions concerning a period of time (in the past) e.g. demographic changes since the last interview like marriage or death of spouse
- Questions concerning future prospects (future) e.g. satisfaction with life five years from now, or job expectations

1.3.3 Survey modes

The SOEP uses several different modes to collect the data. Originally, the respondent's answers were recorded by an interviewer who filled in a paper questionnaire, the so called pen-and-paper interview or PAPI. The personal contact between interviewer and respondent is important for the success of the survey; however, before losing a respondent because of a scheduling conflict between interviewer and respondent, the SOEP allows mailing in the questionnaire starting from the second wave of subsamples A-I. This concept does not resemble the concept of a regular mail survey, because the interviewer still keeps the personal contact with the household and schedules appointments with its respondents if possible. Starting with subsample J, only the computer assisted mode (CAPI) is allowed, and thus mailing in the questionnaires is no longer possible.

While the interviewer is in the household she/he directly conducts an interview with any household member, but can also hand out a questionnaire to other household members, who fill it in with or without her/his help (self-administered questionnaires, SAQ). This is much more time efficient for the interviewer, because household members can work in parallel on their questionnaires.

In 1998, interviews were conducted with computers for the first time, in computer-assisted personal interviews, or in CAPI mode. Compared to PAPI, CAPI is much more efficient in transferring the data into an electronic format, which was an important asset especially with the extensions of the panel starting in the year 2000. The CAPI mode was

first conducted in parallel to the PAPI mode, meaning that interviewers and respondents were free to chose how they wanted to do the interview. This was important for the "older" sample members (responents as well as interviewers), who were used to the PAPI concept. Only in the most recent samples (starting in subsample J), CAPI is the only mode. Figure 5 depicts the development of modes up to 2011, showing that the CAPI mode has gained importance since its implementation.

Since the questionnaires have to be identical in both modes, the CAPI implementation is relatively simple compared to what would be technically feasible. For example, the SOEP basically does not use any form of dependent interviewing (i.e. referring to respondent data from previous waves), because this cannot be easily implemented in the PAPI-mode. Also, the filtering structure is very simple in the SOEP, because any respondent must be able to follow the interview path on her/his own on paper. Still, some technical features like the control of value ranges (e.g. month of birth, year of first marriage) or the randomization of scale items are implemented in the CAPI version of the questionnaire.

In the future, new modes will be introduced into the SOEP as they develop. The computer-assisted web interview (CAWI) is close to implementation, it will, however, not be used as a replacement of the current CAPI and PAPI modes, but rather as an extension the respondents may use similar to the mail-in or self-administered questionnaires. The core interview concept of the SOEP survey, the personal contact between respondent and interviewer, will not change.

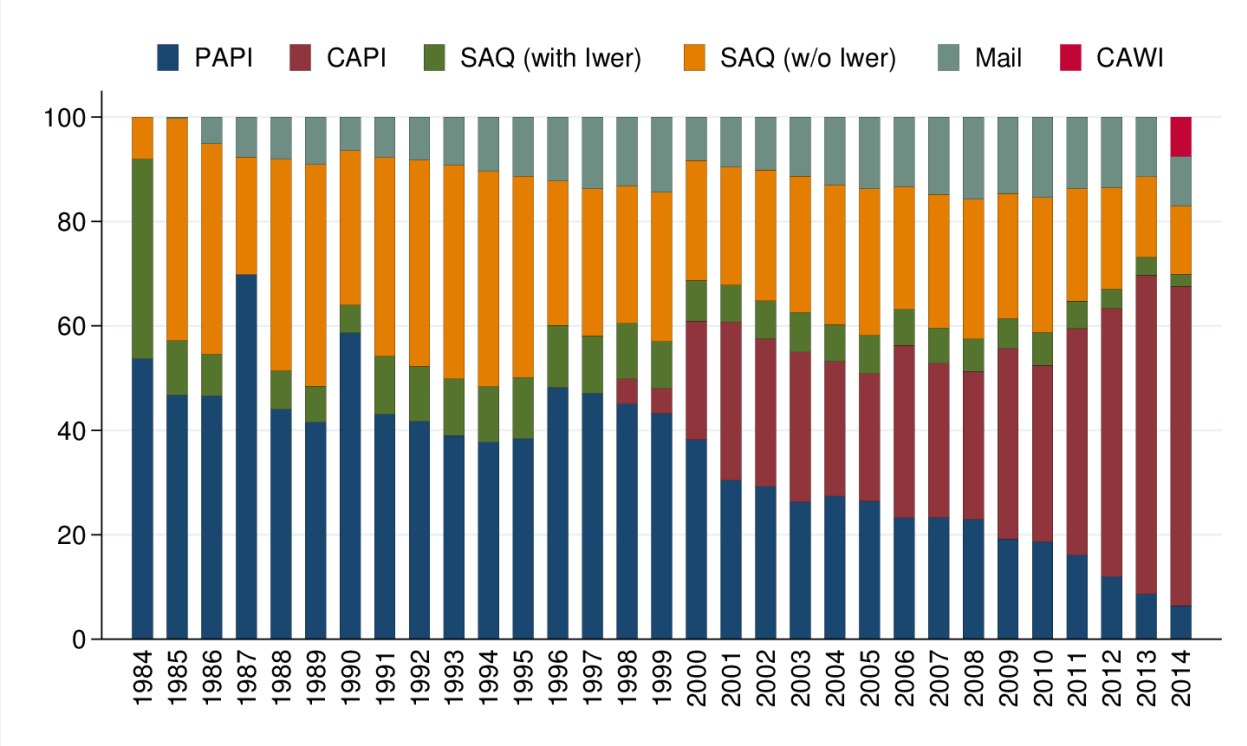


Fig. 1.5: Use of Different Interviewe Modes since 1984

Stata Code to create figure.

1.3.4 Panel care

To cope with panel attrition and to keep the the longitudinal response rates at high levels, the SOEP has implemented so-called "panel care" efforts to maintain the personal contact between respondents and the survey. Panel care can be divided into incentives directly given to the respondent and other measures undertaken to keep the respondent in the study.

The study has honored the respondents with gifts and tokens of appreciation from the very beginning. For the most part, these gifts are small in-kind incentives like flowers, for which the interviewers have their own budget. In addition, the interviewers are asked to hand out a brochure with recent results from the study. Up to 2007, the respondents also received a lottery ticket as a thank you upon completion of the interview. The lottery collects money for social projects in Germany. Since 2008, the lottery ticket is included in the contact letter which is sent out about two weeks prior to the interview. It is thus given unconditionally, as long as the person has participated in the previous wave. After any successful interview, the respondent receives a thank you letter from the field work organisation, which also inlcudes a stamp for a regular letter.

In 2009, different incentive schemes were tested in the new subsample I to increase the first-wave response rates. The basic experiment included four randomized groups of households: (1) those with the default setup of the conditional lottery ticket; (2) those with a "low" cash incentive involing 5 Euros per household and 5 Euros per adult respondent; (3) those with a "high" cash incentive involing 5 Euros per household and 10 Euros per adult respondent; and (4) those with a choice between a "low" cash incentive and a lottery ticket. The results showed slightly higher response rates in the cash groups, although the extra money in group (3) did not pay of. (Further results will be published on our website as soon as possible.) Additional work is done by the field work agency: Addresses are kept up to date throughout the year in order to be informed about residential mobility. This is achieved for example by sending out a brochure containing some results based on previously collected data, or saisonal greeting cards.

In addition, the face-to-face interview ensures a personal relationship, which increase the likelihood to stay in the survey. Thus, keeping the same interviewer over time is one important goal - some of the respondents have indeed had the same interviewer since the beginning in 1984.

1.4 Principles of data structure

The SOEP started with a basic data structure that is now termed SOEPclassic. The majority of datafiles still follow this basic principle, which is explained in the next section. Since 2012, there is a new concept of SOEP called SOEPlong, where the focus is on the longitudinal nature of the SOEP.

1.4.1 SOEPclassic

SOEPclassic contains a multitude of different datasets (in the data distribution of 2012, v28, there were over 300 different files). To get an overview of the data, a somewhat simplified categorization helps: there are datafiles which describe the **development** of the sample, such that the user knows which person or household was part of the interviewed sample in any given year. Then there are **wave specific original** data files, which contain the data from each year's questionnaires without any changes except for very basic consistency checks. To help the user with the data, there also are **wave specific generated** data. These contain consistently coded variables across all waves with common names, such that the users can easily use this information when combining datasets across waves. The SOEP also provides various data on the respondent's background, called **biographical data**. These can conceptually be separated into biographical data which are unchanging (such as information on parent's education, or data from the mother-child questionnaires) and data which may be updated through changes in a respondent's life (such as new children in the birth biography, or a job change in the job history). There are also some files in SOEPclassic, which are longitudinal in their nature, containing information from several years in one file, or - in the case of the multiple imputations (MI-HINC) - contain several observations per household for one year. Finally, there are some files which cannot be easily categorized - some are one-time datasets, some provide information about the interviewers, some about respondents outside of Germany.

One of the biggest assets of the SOEP data is their longitudinal nature, i.e. repeated observations of the same unit (person or household) over time. There are two datasets which should be the building block of any analysis, as they allow to define longitudinal populations very easily: PPFAD and HPFAD. HPFAD includes all households which have been interviewed successfully at least once. Similarly, PPFAD contains all persons who have ever lived in a household that has participated in the SOEP, i.e. that has been captured in HPFAD, including non-respondents and children. Both datafiles contain one record per household or person, respectively, with wave-specific variables for each year's survey

status. In addition to some time-invariant information (like gender, year of birth, migrant status), these files contain all necessary identifiers to combine other files with PPFAD and HPFAD.

Although they provide essential information, PPFAD and HPFAD alone are of little use for actual analyses. The most often used sources for additional information in SOEPclassic are the cross-sectional data files provided in each survey year (or "wave") (see Figure 6 for those cross-sectional data files included for all waves).

Each wave is identified by letters of the alphabet: the first wave in 1984 is wave "A", 1985 is wave "B", and so on, up to BB in 2011. To simplify notation, the sign is used, when all waves of one group of datasets are referred to. For example, H refers to all household level datasets AH to BBH. For each year of SOEP data there are single data files for households (e.g. H) as well as for individual respondents (e.g. P) and children (e.g. $KIND) based on interview information. These observations make up the "net" population, with each of these files containing as many records as interviews could be conducted. Additional data files with a limited number of variables based on the "address log" constitute the "gross" number of households and persons, i.e. all households and their members which were eligible for an interview in any given year.

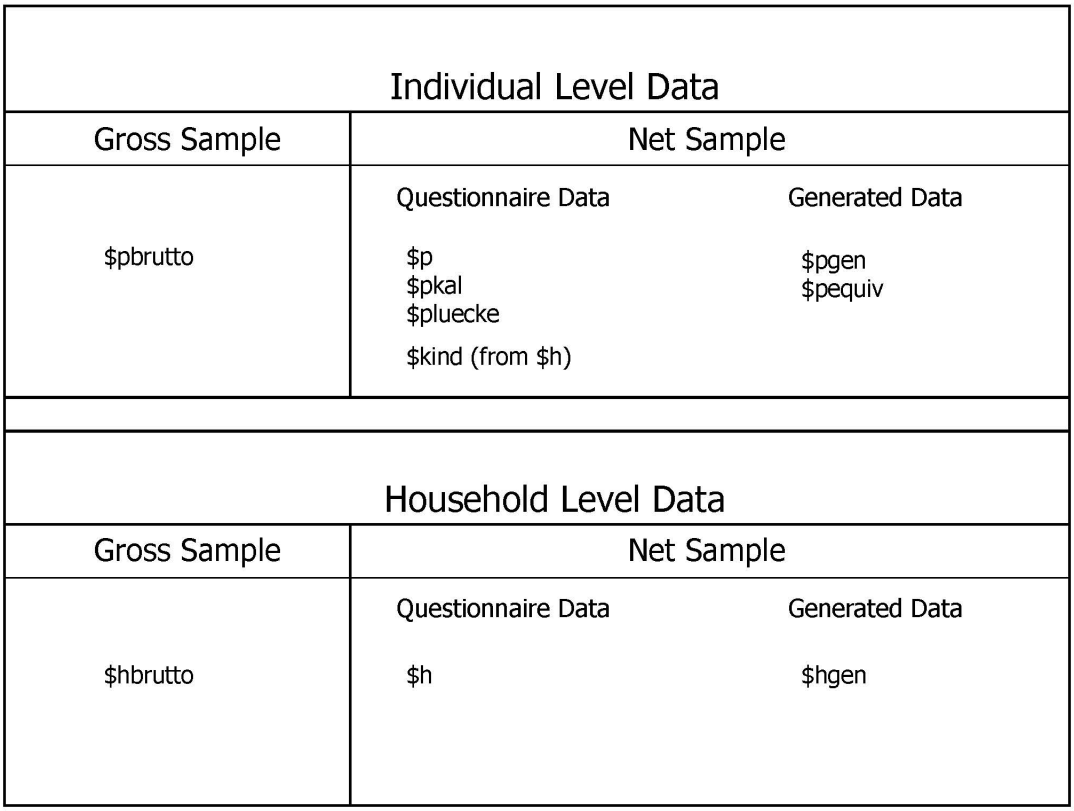


Fig. 1.6: Cross-Sectional Data Files Included in Every Wave

Each new sample is integrated in the old scheme - hence, for Sample C (East Germans), the first wave of data gathered in 1990 is still labeled "G", as in the original West German sample. Similar, Sample F starts in 2000 with the letter "Q", and so on. Even though there are many different samples in the SOEP, for the most part there is only one dataset for each year and topic. For example, the personal interviews recorded in any wave are jointly available for all samples in the P files. These considerations apply identically to the generated data files, such as PGEN (user friendly data on the individual level) and **$HGEN** (user friendly data on the household level).

In addition to the cross-sectional datasets there are datasets which are not wave-specific. These include spell data,

which are organized by person and spelltyp (such as ARTKALEN or PBIOSPE. Other files which are not wave-specifc belong to some biography data, e.g. the data on the first job (**BIOJOB**, or the data on the fertility history (**BIOBIRTH**)).

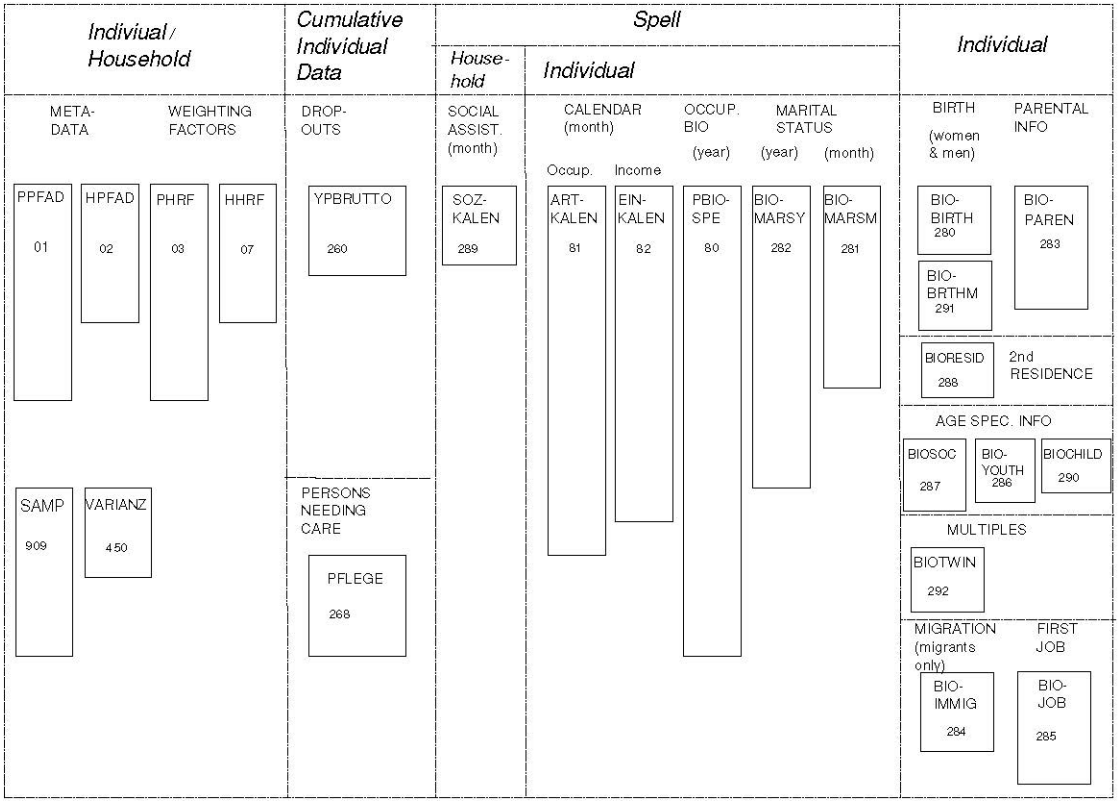


Fig. 1.7: Longitudinal Data Files

Because of the overall data structure with data on different observational levels, any analysis requires the combination of data using matching or merging procedures. These merging procedures need identifiers such that a combination of datasets becomes feasible. The central individual identifier across time is **PERSNR**, which is fixed over time (and of course datasets). Since a person might change the household in which he or she lives at any point in time, yearly household identifiers called **HHNRAKT** are necessary. The exact same information is also stored in **$HHNR**, allowing easier matching depending on the dataset used. Finally, each individual (respondents as well as children) can be traced back to be a member of or a split-off from an original household of the very first wave. This household's ID, which is fixed no matter how often a person changes the household in the course of time, is called **HHNR**. All these identifiers are included in the above mentioned master file **PPFAD** with the wave-specific household identifiers named **AHHNR** (for wave 1), **BHHNR** (wave 2), ..., **BBHHNR** (wave 28). Figure 8 provides a schematic overview of gross and net samples an how they relate to cross-sectional and longitudinal data.

Variable names in the SOEPclassic data files follow basic conventions: First, there are datasets with "speaking" variable names, where the variable name itself conveys something about the information stored in this variable. Most generated datasets follow this convention - e.g. the variable **PARTNR** in the datasets **$PGEN** contains the person identifier for the respondent's partner. Second, there are variable names which do not "speak", but remain identical across the waves up to a wave identifier, e.g. the variable I11102 in the **PEQUIV** datasets always contains post-government incomes. Finally, for the original datasets such as **H, P and KIND**, the variable names are set up "around" the unit of analysis (individual - "p", household - "h", and child - "k") and show before this indicator the wave in which the data were collected and after it the reference of where in the original survey instrument the question can be found (see Figure 9 for an overview). For example, in the dataset AP, the variable **AP06** refers to the person questionnaire in wave "A" (1984), question 6.

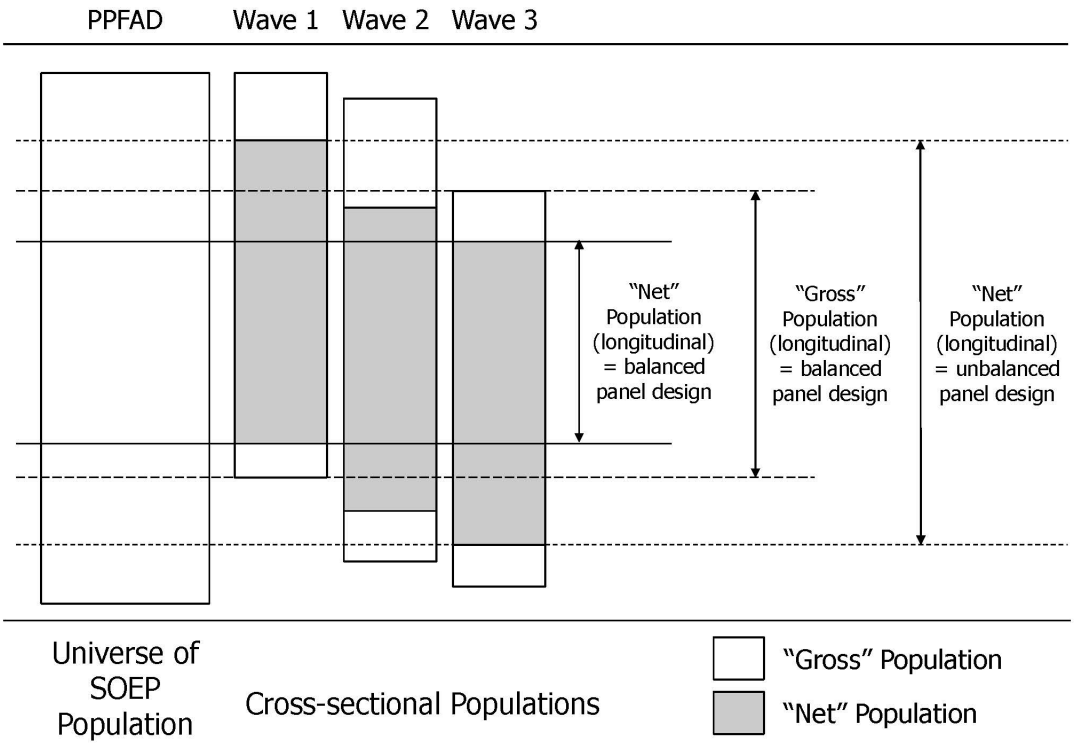
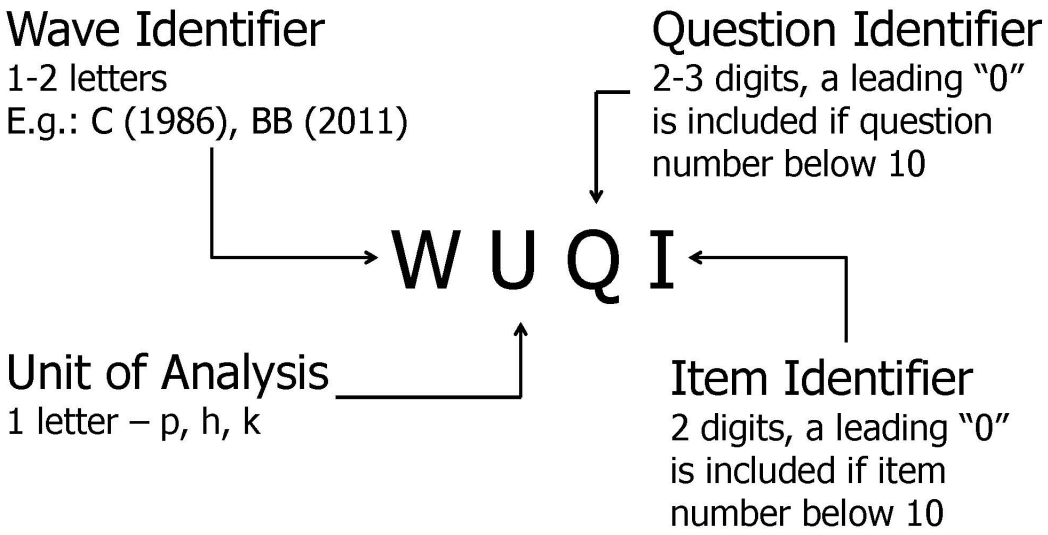


Fig. 1.8: Relationships between Cross-Sectional and Longitudinal Populations



Examples

AP06	Person questionnaire, wave A, question 6
TH1603	Household questionnaire, wave T, question 16, item 3
LP10312	Person questionnaire, wave L, question 103, item 12
BAP15604	Person questionnaire, wave BA, question 156, item 4
VK65	Children in Household questionnaire, Wave V, question 65

Fig. 1.9: Variable Naming Conventions for H, P and KIND

Table 1.5: Variable Names

Digit	Meaning
1	Wave (A for 1984, B for 1985 ... ; according to West samples) e.g. the "A" in AP06
2	Unit of analysis (H=household, P=person) e.g. the "H" in AH27
3-4	Number of question in original survey instrument (questionnaire) e.g. the "57" in AP57
5 or 7	indicating sample specific question (A=sample B, O=sample C due to the fact that "A" is the first letter of the German word Auslaeander which means foreigner and "O" is the first letter of Ostdeutscher which means East German) e.g. the last "A" in AP62A, or the letter "O" in HP42O
or 5	indicating questions in different versions of the questionnaire for first-time or new respondents (Blue version of the questionnaire) and those who have already been interviewed before (Green version) - only for 1985-1993 e.g. the "G" in BP27G06, or the "B" in DH26B01
or 2 thru 8	text for variables in PBRUTTO, HBRUTTO, PGEN, and HGEN files e.g. BHHGR, the household size in wave 2
1 thru 8	text for variables in PGEN and PEQUIV e.g. PARTNR88, the PERSNR of partner, wave 5 e.g. I1110204 , annual post-government income in wave 21

Note that starting with wave BA in 2010, the variable names change accordingly, i.e. an extra digit needed to be added for all variables names since then. As the space restriction to 8 digits is not an issue for modern computers anymore, future releases of the data may introduce new rules of naming the data.)

1.4.2 SOEPlong

"SOEPlong" is a highly compressed, easily analyzed version of the SOEP data that, according to numerous enthusiastic users, is much simpler to handle than the usual version. The data are no longer provided as wave- specific individual files but rather pooled across all available years (in "long" format). An overview about the connection of the datafile between the two SOEP formats are avvilabel in Table 6. In some cases, variables are harmonized to ensure that they are defined consistently over time.

For example, the income information provided up to 2001 is given in euros, and categories are modified over time when versions of the questionnaire have been changed. All these modifications are clearly documented and described for ease of understanding. In the case of recoding or integration of data (for example, datasets specific to East German or foreign populations), documentation is generated automatically and all modified variables are provided in their original form as well.

SOEPlong thus provides a well-documented compilation of all variables and data that is consistent over time. It thereby significantly reduces both the number of datasets and the number of variables. Our main structural tracking files **PFAD** and **PHRF** are also provided in a merged "long" form—in other words, weighting factors are a ready integrated into PPFADL and HPFADL.

And for the first time, a beta version of so-called "enumerated weights" [PHRFE] is provided in PPFADL, particularly for the analysis of household characteristics on the individual level.

Further variables included in the "long" format of the PFAD dataset are:

- The variable **IYEAR** (interview year) - which corresponds to the variable **DATUMY** in **HBRUTTO** - to mark the actual interview year, supplementing the variable **SYEAR** (survey year, referring to the reference year for the survey instrument).
- In addition, the generated partner IDs (**PARID**, **PARTNER**) are also included in **PPFADL** (to supplement the corresponding **PGEN** variables, allowing partners to be identified in households where one partner could not be interviewed).

A further addition to the "long" format of the SOEP data are the cumulative original data from the biographical questionnaire from the **BIO** dataset.

Preparation of the SOEPlong format also includes all datasets that are provided regularly as cross-sectional files. In generating the individual and household data in the SOEPlong format from the original survey data, comprehensive information is also generated from the cross-sectional variables documenting the long variables over time. This allows users to see what adaptations had to be made in variables over time and verify the variables' comparability.

Table 1.6: Matching of SOEPlong and cross-sectional datasets from SOEPclassic

SOEPlong	SOEPcore
ppfadl	ppfad, phrf
hpfadl	hpfad, hhrf
pbrutto	
hbrutto	
pl	ap, ..., zp, bap, bbp, ...
hl	ah, ..., zh, bah, bbh, ...
kidl	kidlong (akind, ..., zkind, bakind, bbkind, ...)
pgen	apgen, ..., zpgen, bapgen, bbpgen, ...
hgen	ahgen, ..., zhgen, bahgen, bbhgen, ...
pkal	apkal, ..., zpkal, bapkal, bbpkal, ...
pequiv	apequiv, ..., zpequiv, bapequiv, bbpequiv, ...
bio	biolela, mlela, ..., zlela, balela, bblela, ...

The "long" data are being provided, as in previous years, for users of different data formats: Stata, SPSS, and SAS (and in the unlabeled ASCII format). In addition, an English version of all datasets is being made available.

The first complete documentation on the SOEPlong format is the first content being made available in the new version SOEPinfo ("DDI on Rails"). It also is provided there in graphic form.

1.4.3 Missing conventions

Survey variables might be missing, i.e. without a valid code or value for different reasons. In the SOEP, negative values are not valid for any variable, but are used instead to code different reasons for missing information. There are two distinctions for missing values: they may originate in the respondent's answer or in the survey design. The respondent may refuse or not know an answer or she may report invalid values on the one hand, and the interview design may exclude respondents with certain characteristics from some questions on the other (e.g. men will never be asked if they are pregnant). The following codes apply both for SOEPclassic and SOEPlong, also shown in Table 7:

- A person might refuse to answer a question, which happens more often in sensitive questions (e.g. income related questions), or may just not know the answer to a question. In such a case, the missing code is "-1" for "no answer / don't know". Note that the SOEP does not distinguish between the refusal to answer and a true "don't know".
- Information may be missing when a question is not asked because it is not relevant for a specific person, e.g. owner-occupiers will not be asked about the amount of rent they pay. In such cases, the question "Does not apply" to this person, and the variable receives a code of "-2".
- Sometimes invalid answers are encountered, when respondents fill out a PAPI interview themselves or the interviewer mistypes an answer, e.g. persons cannot work more than 168 hours a week. In such a case, multiple checks are carried out, and if the inconsistency remains, the variable is recoded "-3 Implausible value".
- Some questions contain multiple answer possibilities, where the respondents are asked to pick one and only one answer. In the SOEP PAPI instruments, sometimes respondents ignore this request and provide more than one answer, e.g. they mark "very good" and "good" when asked about their current health status. In such cases, if the correct answer cannot be determined from the questionnaire itself, the code "-4 Invalid Multiple Answers" is given to this variable.

- With the extension of the SOEP in recent years, entirely new samples have been added to the core. In these samples, sometimes questions are left out completely, e.g. to shorten the questionnaire or because the focus of the sample is different as in some of the related studies. In such a case, the variable will be set to "-5 Not included in this version of the questionnaire" for an entire subsample.
- With the use of CAPI, recent developments include an "integrated" person questionnaire, i.e. the biography part and the "regular" part of the questionnaire are asked as one. Some of the questions in the biography part are repeated in the regular part. While in the PAPI mode, the respondent will answer the same question twice, the CAPI allows to filter the respondent around the question if it has already been asked. These cases are very rare - if they occur, they receive a code "-6 Version of questionnaire with modified filtering".

Table 1.7: Missing Values

Code	Meaning
-1	no answer / don't know
-2	does not apply
-3	implausible value
-4	Inadmissable multiple response
-5	Not included in this version of the questionnaire
-6	Version of questionnaire with modified filtering
-8	Question not part of the survey program this year (long format only)