DTC

Desktop Companion to the German Socio-Economic Panel (SOEP)

Edited by
John P. Haisken-DeNew (RWI Essen) &
Joachim R. Frick (DIW Berlin)

Version 8.0 - Dec 2005
Updated to Wave 21 (U)

Forword

This documentation is intended to give novice users a "jump start" in understanding the SOEP, its structure, depth, and research potential. Further, after having read the practical section, where retrievals in Stata, SAS, and SPSS are demonstrated, users should be then able at least to run a rudimentary retrieval.

This is the compilation of many, many, working papers, overhead slides, example programs, and lecture notes, that have been prepared over the years by the members of the SOEP crew. We have combined these various sources of information, to make it easier for users to work with the SOEP. This collection of information is intended to be *the main reference guide*, and a practical companion in basic understanding and implementation of the SOEP. The Cross-National Equivalent file is only briefly described in this companion.

Our thanks go to our fellow contributors **Markus Grabka**, **Peter Krause**, **Markus Pannenberg**, **Rainer Pischner**, **Ulrich Rendtel**, **Jürgen Schupp**, **C. Katharina Spiess**, **Martin Spiess**, and **Gert G. Wagner** for their support in keeping this documentation current.

We know that starting to use any new dataset, it is a difficult challenge and this is especially true given the complexity of the panel data. We hope however that this introduction will help. The SOEP is certainly the best micro panel data set in Germany (and certainly the longest running), and we believe it is well worth getting to know. This reference owes its length to users' requests of "how to get started". We always welcome any feedback, or information on ways that we can improve this documentation.

John P. Haisken-DeNew (`jhaiskendenew@rwi-essen.de`)
Joachim R. Frick (`jfrick@diw.de`)

DTC Revision History

Version 8.0 - Dec 2005

- Wave U (2004) and wave T (2003) specific updates
- Added info on "SOEP Menu"
- Added *BIOCHILD*, *BIOBRTHM*, *BIORESID*, *BIOTWIN*
- New Generated variables *$PGEN*, *$HGEN* and *PPFAD*
- Weighting updated
- SOEP Best Papers "GSOEP 2002-2004" added
- SOEP Best Papers presented "GSOEP 2004 Conference" added
- Retrieval algorythm for person based retrievals changed

Version 7.0 - Sep 2003

- Wave S (2002) specific updates
- Updates to *BIOYOUTH*, *BIOPAREN*, *$PGEN*
- *BIOSOC* added
- New Generated Education Variables: ISCED (1997) and CASMIN
- Weighting updated

Version 6.0 - Dec 2002

- Wave R (2001) specific updates
- *BIOYOUTH* added
- Weighting updated
- Generated Variables redone, specifically Industry
- New Industry NACE coding
- New Occupation ISCO88 coding
- SOEP Best Papers "GSOEP 2002" added

Version 5.0 - Nov 2001

- Wave Q (2000) specific updates
- Sample F and Weighting added
- Retrievals chapter rewritten: Stata [7.0], SAS [8.2], SPSS [10.X]
- TDA documentation removed
- Generated Variables redone, specifically Education
- New Occupation ISCO88 coding from Infratest, *OPGEN-QPGEN*

- New Industry NACE codes in *PP* and *QP*
- Documentation for *VARIANZ* added
- SOEP Best Papers 1999-2001 added

Version 4.0 - Nov 2000

- Wave P (1999) specific updates
- Macro Information at the Region level added
- SOEP Literature lists added

Version 3.0 - Mar 2000

- Wave O (1998) specific updates
- *BIOIMMIG*, *BIOJOB* description added
- General down-sizing of non-core information

Version 2.2 - Sept 1998

- Wave N (1997) specific updates
- New introductory retrieval section
- Advanced SAS and SPSS part moved to appendix
- $SYR description removed (dropped from data set)
- *SOZKALEN*, *PFLEGE*, *$PKAL* description added

Version 2.1 - Mar 1998

- Primer of German Institutions (Appendix) added
- PSID-GSOEP Equivalent File Information (Appendix) added
- USA Mailing List Info added
- Industry/Occupation/Other Text Codings added
- GSOEP User Contract / Data Secrecy added

Version 2.0 - Sept 1997

- Wave M (1996) specific Updates
- Installation chapter removed
- Item Correspondence List removed
- Ensuring Continued Respondent Participation added
- *$SYR* files detailed variable description added
- Biography Data Files added

- Stata [4.0] Retrievals added
- TDA [6.1] Retrievals added
- Newspell added
- Mailing List Information updated
- Sample D Weighting updated

Version 1.0 - Sept 1996 : Original Version

SOEP Address and Telephone Information

Mailing Address :

SOEP
14191 Berlin
Germany

Street Address :

SOEP
DIW Berlin
Koenigin-Luise-Strasse 5,
14195 Berlin
Germany

General Information :

Fax: +49 / 30 / 89789-109 (Attn: SOEP Secretary)
Fax: +49 / 30 / 89789-200 (Attn: SOEP Secretary)
WWW: `http://www.diw.de/english/sop/`
Hotline: `soepmail@diw.de`

Table 1: The German Socio-Economic Panel Study: Core Staff

Core Staff	Service	Research	Email Address	Telephone +49 / 30 /
Gert G. Wagner	**Director**	Survey Methodology, Behavioral Economics	gwagner@diw.de	89789-283
Joachim Frick	**Data Operations Manager** DTC	Migration, Income Distribution, Housing Int'l Panel Data	jfrick@diw.de	89789-279
Jürgen Schupp	**Survey Manager**	Labor Economics, Time Allocation, Survey Research, Social Indicators	jschupp@diw.de	89789-238
Silke Anger	Labor market vars	Labor	sanger@diw.de	89789-526
Jan Goebel	Data Distribution	Imputation, Income distribution	jgoebel@diw.de	89789-377
Markus Grabka	Cross-National Equivalent File	Income Distribution	mgrabka@diw.de	98789-339
Elke Holst	Newsletter Panel Maintenance	Labor Economics, Female LFP	eholst@diw.de	89789-281
Bettina Isengard	Ultra Rich / Ultra Poor Education variables	Comparative Studies of Poverty and High Income	bisengard@diw.de	89789-284
Peter Krause	Data Management	Poverty, Income Distribution	pkrause@diw.de	89789-690
Martin Kroh	Weighting, Political Science Variables	Political Science, Statistics	mkroh@diw.de	89789-678
Rainer Pischner	Data Distribution, Weighting	Spell Data	rpischner@diw.de	89789-319
Ingo Sieber	Perl, JavaScript SOEP-Databank	SOEPINFO, SOEPLIT, Data Distribution	isieber@diw.de	89789-260
C. Katharina Spiess	Regional Data	Social Policy, Child Care	kspiess@diw.de	89789-254
Martin Spiess	Weighting, Imputation	Weighting, Imputation, Non-Response	mspiess@diw.de	89789-291

Table 2: The German Socio-Economic Panel Study: Support Staff

Support Staff	Service	Area of Concentration	Email Address	Telephone +49 / 30 /
Debbie Bowen	Office Assistance	Translations	dbowen@diw.de	89789-332
Michaela Engelmann	Office Assistance	HOTLINE, User Service	mengelmann@diw.de	89789-292
Gabi Freudenmann	Office Assistance	Secretariat	gfreudenmann@diw.de	89789-402
Christine Kurka	Office Assistance	Organization	ckurka@diw.de	89789-283
Uta Rahmann	Office Assistance	SOEPLIT, Web Texts	urahmann@diw.de	89789-287

Contents

Chapter 1

Overview of the SOEP

by Joachim R. Frick, John P. Haisken-DeNew, Martin Spiess and Gert G. Wagner

1.1 What is the Desktop Companion ?

The SOEP Desktop Companion (DTC) is intended to give the reader a very detailed understanding of the structure of the German Socio-Economic Panel. **Chapter 1** gives an overview of the SOEP data and available documentation resources. **Chapter 2** explores extensions to the original questionnaire data, developed by the SOEP group and by the Cornell "Scientific Use / Cross-National Equivalent File" group at Cornell University. **Chapter 3** describes the biography information collected in the life history questionnaires and the user-friendly data structure in which this information has been prepared. **Chapter 4** looks at effective data retrieval writing in Stata, SAS, SPSS with "hands-on" examples. **Chapter 5** describes the weighting and sampling methods used in the SOEP. A subject index at the back allows the reader to quickly find information using a keyword search.

The SOEP Desktop Companion (DTC) should be on every SOEP user's desktop, as the name suggests and is updated with every new data release. Armed with the **DTC** and the online variables information web program **SOEPINFO**, a novice SOEP user should be able to make significant research headway in a short period of time and an experienced user will benefit from the detailed reference.

1.2 Introduction

The aim of this chapter is to provide first-time users with a brief overview of the contents and design of the SOEP. Specific areas which typically cause problems for first-time users will be dealt with in some detail[1].

The micro-data of this survey are available for scientific research all over the world. However, due to strict data protection laws, researchers outside the European Union are allowed to process 95% of the original sample, only. Figures and numbers shown in the following description are based on the full 100% sample (if given, information for the 95% version is in brackets).

1.3 Contents of the Study

The SOEP was started in 1984 as a longitudinal survey of private households and persons in the Federal Republic of Germany. The central aim of this panel study is to collect representative micro-data on persons, households and families in order to measure stability and change in living conditions by following principally a micro-economic approach enriched with sociology and political science variables, mainly determined by the "Social Indicator" movement.

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

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

Additionally, as a yearly topical module, the basic information in one of these areas is enlarged by detailed questions as seen in Table 1.1.

In order to measure change and stability across time, the SOEP-questions are targeted at different dimensions of time (past, present and future) using also different measurements of time (information at a given point of time, periodical information, calendar information, life history information).

[1] This short introduction does not come close to covering all the details necessary for exploiting the full richness of SOEP data. For more detailed sources of information see the list of technical and substantive papers in the Reference section.

Table 1.1: Special Topics Modules

Year	Wave	Sample	Topic
1984	A/1	A B	Employment biography since age 15 (Bio)
1985	B/2	A B	Marriage and family biography (Bio)
1986	C/3	A B	Social origins (Bio), first job (Bio), neighborhood
1987	D/4	A B	Social security, early retirement, persons requiring care, and child care
1988	E/5	A B	Assets
1989	F/6	A B	Further education or training and qualification
1990	G/7	A B C	Use of time and preferences Base questions (labor market + subjective indicators)
1991	H/8	A B C	Family and social services Family and social services (shortened version plus repetition of subjective indicators and labor market indicators of wave 1 base questions)
1992	I/9	A B C	Social security and poverty (partly repetition of Wave 4 (W)) Social security and poverty (partly repetition of Wave 4 (W) labor market indicators and biographical information (Bio)
1993	J/10	A B C	Further education or training (shortened repetition of Wave 6) Further education or training, labor market
1994	K/11	A B C D1	Neighborhood, values, and expectations Same as Wave 11 plus immigration history and biography
1995	L/12	A B C D1 D2	Partial repetition of Wave 7 - use of time and preferences, increased range of income questions Same as Wave 12 plus immigration history and biography
1996	M/13	A B C D	Repetition of social network questions (Wave 8)
1997	N/14	A B C D	Social security and poverty (repetition of Wave 9)
1998	O/15	A B D C E	Ecology and environmental behavior (indirect taxation)
1999	P/16	A B D C E	Neighborhood, Values, Expectations
2000	Q/17	A B D C E F	Further Education Training, Labor Market
2001	R/18	A B D C E F	Social Networks, Working Conditions
2002	S/19	A B D C E F G	Assets (see also Wave 5), Repetition Social Security (Wave 14)
2003	T/20	A B D C E F G	Ecology and environmental behavior (Wave 15)
2004	U/21	A B D C E F G	Further Education Training (Wave 6, 17) Neigbourhood (Wave 16), Risk Awareness
2005	V/22	A B C D E F G	Use of Time and Preferences (see also Waves G, L)
2006	W/23	A B C D E F G	Family and social services, networks (see also Wave H, M, R)

Note:
W: West German Sample: A,B
O: East German Sample: C
D1, D2: Immigrant Sample: D
E: Refreshment Sample: E
F: Innovation Sample: F
G: High Income Sample: G

- Questions about a point of time (present)
  e.g. current employment status or current levels of satisfaction
- Single retrospective questions on certain events in the past (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 on income and labor market related issues (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

1.4 Target Population and Sampling

In order to start the survey in early 1984, the original SOEP sample was drawn in 1983. The SOEP was expanded to the territory of the German Democratic Republic in June 1990, only six months after the Berlin wall fell. Thus, the target population to be represented by the SOEP is defined firstly as the residential population of the FRG in 1984 including West Berlin, and secondly as the German residential population in the GDR (including East Berlin) in June 1990.

In the FRG, selected foreigner groups were oversampled in the study. The sampling probability for the eastern sample is also larger than the probability for the main sample in West Germany. Those different sampling probabilities were chosen to make sure that the number of cases in the sample are large enough for analyses of the three samples on their own. For a more detailed description, see Chapter 5.

The institutionalized population, in the true sense of the word (hospitals, nursing homes, military installations) was not representatively included in the first wave[2]. Later, however, persons from the initial households who had taken up residence temporarily or permanently in institutions of this kind were followed.

SOEP samples

Sample A "Residents in the FRG" covers persons in private households with a household head who does not belong to the main foreigner groups of "guestworkers" (i.e. household heads who are Turkish, Greek, Yugoslavian, Spanish or Italian). Because only a few foreigners are in Sample A it is often called the "West German Sample" of SOEP. In 1984 it covered 4528 households (4298 in the 95% Scientific Use Version) with a sampling probability of about 0.0002.

Sample B "Foreigners in the FRG" covers persons in private households with a Turkish, Greek, Yugoslavian, Spanish or Italian household head. Compared to Sample A the population of Sample B is oversampled and started with 1393 households (1326 in the 95% Scientific Use Version). The sampling probability was about 0.0008.

Sample C "German Residents in the GDR" covers persons in private households where the household head was a GDR citizen. 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

[2]See Section 5.4.2 below. In 1984 there are only 15 institutionalized households in Sample A and 42 in Sample B. For a detailed description of the problems in covering this population in the SOEP, see Hanefeld (1987)

institutionalized). All in all, 2179 households (2071 in the 95% Scientific Use Version) represent the starting size of this sample with a sampling probability of about 0.0004.

Sample D "Immigrants" started in 1994/95 in two different samples. In 1994, the first sample D1 had 236 households and in 1995, the second sample D2 had 295 households, making in 1995 a total 522 households (D1 and D2). 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" In 1998, a new sample was selected from the population of private households in Germany. The new sample, also denoted as subsample E, was selected independently from the ongoing panel (subsamples A through D). The selection scheme used for sample E essentially resembles the scheme also used in selecting subsample A. The number of observed and valid private households in subsample E in 1998 was 1067, covering a total of 1932 successfully interviewed persons aged 16 and older. The number of children within these households in 1998 was 468. The sampling probability is about 0.00003.

Sample F "Innovation" Subsample F was selected independently from all other subsamples from the population of private households in 2000. The selection scheme was essentially the same as for selecting subsample A and E, however, with one exception. Within each PSU, 24 households were selected according to the same scheme as in subsamples A and E. However, 'German' households (all adults aged $\geq$ 16 having German nationality) were selected mainly using the first 12 addresses within each PSU (although a few were selected from the second 12 addresses as well). The 'non-German' households (at least one adult has not the German nationality) were selected using all the 24 addresses. The number of observed and valid private households in subsample F in 2000 was $6,052$. These 6052 households covered $2,993$ kids (age < 16) and $11,532$ adult persons, valid interviews are available for $10,890$ of the adult persons. The sampling probabilities are approximately 0.00028 for 'German' households and 0.0005 for 'non-German' households.

Sample G "Oversampling of High Income" Subsample G was selected independently from all other subsamples from the population of private households in 2002. The original selection scheme required that the responding household had a monthly income of at least DM 7,500 (EURO 3835). The number of observed and valid private households in subsample G in 2002 was $1,224$. These $1,224$ households covered 693 children

(age < 16) and $2,845$ adult persons, valid interviews are available for $2,671$ of the adult persons. Starting in Wave 2, the selection scheme was changed in such a way that only households with a net monthly income of at least EUR 4,500 were followed.

1.5 Survey Design

All samples of SOEP are multi-stage random samples which are regionally clustered. The respondents (households) are selected by random-walk. For further details, see Chapter 5.

1.5.1 Interview Methodology and Survey Instruments

The interview methodology of the SOEP is based on a set of pre-tested questionnaires for households and individuals. Principally an interviewer tries to obtain face-to-face interviews with all members of a given survey household aged 16 years and over. Thus, there are no proxy interviews for adult household members. Additionally one person (head of household[3]) 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 covers also some questions on children in the household up to 16 years of age, mainly concerning attendance at institutions (kindergarten, elementary school, etc.)[4] .

There are different versions of the questionnaires. First, the questionnaires for the foreigner's sample (B) and immigrant sample (D) cover additional measures of integration or information on re-migration behavior. Secondly, in 1990 and the first years of the German unification process, the questionnaire for the East German sample (C) also contained some additional specific variables (since 1992 there are no longer different questionnaires (and data files) for East and West Germany). Since 1996, the questionnaires are uniform and completely integrated for all samples (and as such, the foreigner specific data files *APAUSL*,...*LPAUSL* are no longer separate, starting with wave M or 1996). Thirdly, there is a need to differentiate between first time respondents and those with a repeated interview, since some information does not have to be asked every year, unless a change occurred. Additionaly each respondent is asked to fill out a biography questionnaire covering information on the life

[3]The head of the household is defined as the person who knows best about the general conditions under which the household acts and is supposed to answer this questionnaire in each given year. This reduces the risk of longitudinal inconsistencies.

[4]Although information on children is gathered only in the household questionnaire, the SOEP micro-data offers individual records for each child. This allows for individual longitudinal analyses even of those persons who are not yet of the respondents age.

course (e.g. marital history, social background, employment biography etc.). For further details see Chapter 3.

Additional information can be obtained from the so-called "address log". This is filled in by the interviewer even in case of non-response, thus providing very valuable information for attrition analyses. For researchers interested in methodological issues this data also contains information on the process of the field work, e.g. the number of contacts, reason for eventual drop-outs, or the interview method. For successfully contacted households, the address log covers the size of the household, some regional information, survey status etc., while the individual data for all household members includes the relation to the household head, survey status of the individual and some demographic information.

1.5.2 The Follow-Up Concept

One of the most crucial features of a longitudinal survey to cope with problems of representativeness is the concept, according to which respondents are traced across time.

Since in the SOEP all household members are to be interviewed individually once they reach the age of 16, the next generation is automatically taken into account. In principle, all persons who took part in the very first wave of the survey as well as their children whenever born, are to be surveyed in the following years. In case of residential mobility, the person is to be followed within the survey territory (Federal Republic of Germany). Third persons moving into an existing SOEP household are to be surveyed, or "followed-up" even in case of subsequently leaving that household[5]. The weighting scheme takes into account this "follow-up" of everybody.

Temporary drop-outs or persons and households which could not be successfully interviewed in a given year are followed until there are two consecutive temporary drop-outs of all household members or a final refusal. In the case of a successful interview after a drop-out, there is also a small questionnaire including questions on central information which is missing for the year of the drop-out (e.g. employment status).

"New" persons become part of the SOEP population due to birth, or residential mobility. Those persons living in SOEP households, who then move out or "split-off" into new households, are still followed, but under a new household identifier. See Table 1.2.

Fig. 1.1 shows that as a result of the follow-up concept, up to 2004, several thousand "new" households were added to the SOEP population.

[5]This had not been the case up to 1988.

Table 1.2: The Emergence of New Households

	Households	
Persons	Old	New
Old	• "classic case": without change of address • entire household moves	• Move-out
New	• Birth • Move-in	• Birth • caused by split-offs of old persons from old households

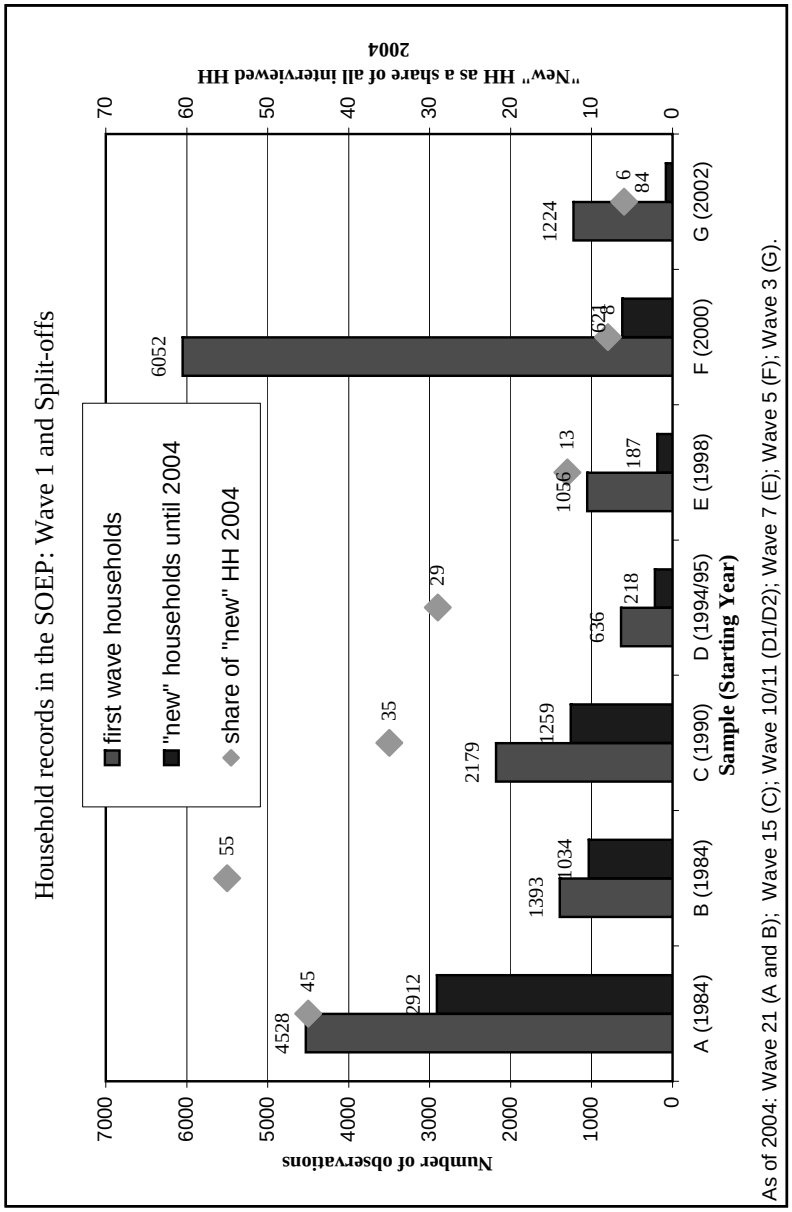


Figure 1.1: Old and New Households in the SOEP (100% Sample)

1.6 Development of Sample Size

Table 1.3 shows the starting samples sizes of the samples A through E, and the years when the samples were first collected. Figure 1.2 illustrates 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.

Table 1.3: Starting Sample Size

Sample	Year	Households (net)	Persons (gross)	Respondents (net)	Children (net)
			100% Sample		
A and B	1984	5921	16205	12245	3915
C	1990	2179	6131	4453	1591
D1	1994	236	733	471	248
D1/D2	1995	522	1665	1078	517
E	1998	1067	2470	1923	468
F	2000	6052	14525	10890	2993
G	2002	1224	3538	2671	693
			95% Sample		
A and B	1984	5624	15397	11610	3711
C	1990	2071	5818	4229	1510
D1	1994	225	696	451	231
D1/D2	1995	497	1584	1027	488
E	1998	1014	2342	1827	448
F	2000	5750	13772	10324	2838
G	2002	1163	3359	2536	653

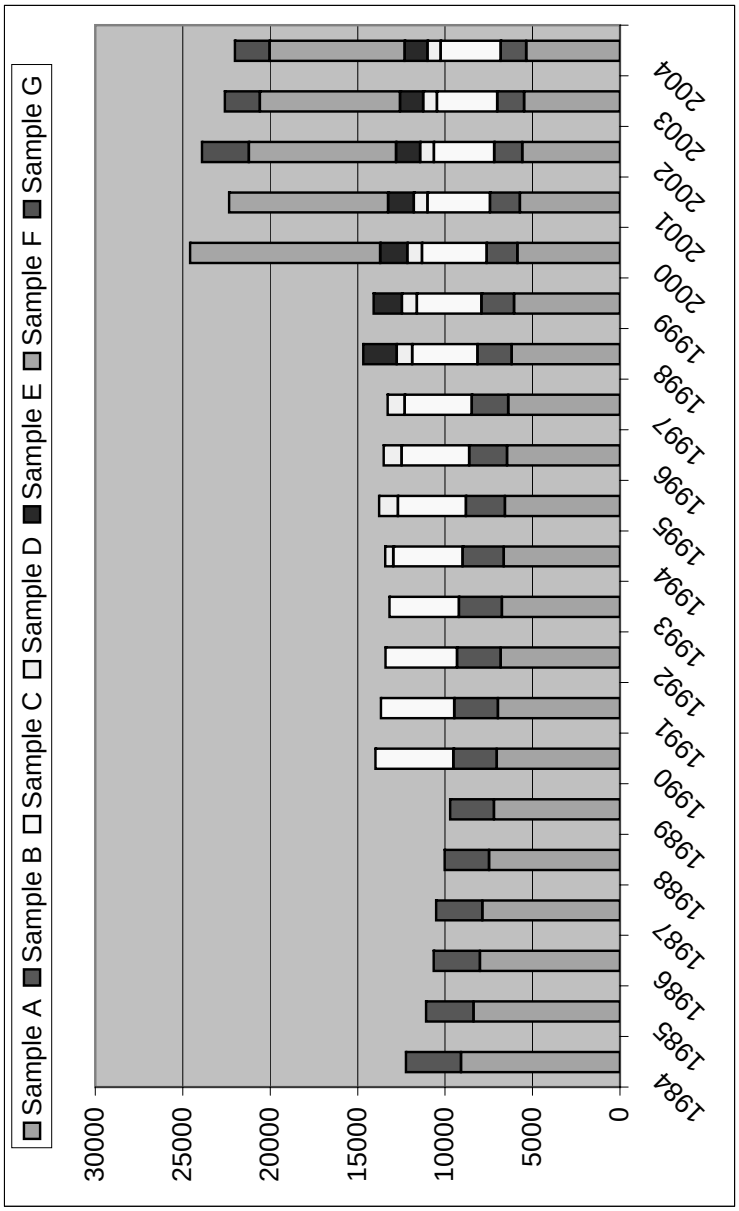


Figure 1.2: Cross-Sectional Development of Sample Size: Samples A-G

However, this cross-sectional view is insufficient when examining the longitudinal development of the sample, which is influenced by demographic and field-work related factors as outlined below:

Determinants of the Sample Development:

1. Demographic factors
   - Persons exit by:
     - Death
     - Moving abroad
   - Persons enter by:
     - birth
     - moving into a SOEP household from somewhere else in Germany or from abroad
     - reaching age of 16 years (minimum respondents age)
     - new households and persons from a split of at least one old person from an old household
2. Field-work related factors (2 stages)
   - making a successful contact to a given household
   - realizing a successful interview
   - social groups which are hardest to contact
     - single person households
     - residentially mobile households and persons
     - young adults leaving parental home

However, in order to improve response rates, the SOEP has implemented a respondent-incentive program such that small "bonuses" or gifts are given, and every effort is made to maintain the personal contact between respondents and the survey.

Panel care:

- For each successful interview, each respondent
  - receives a small gift related to the yearly topical module
  - takes part in a monthly nationwide lottery.
- Addresses are kept up to date by the field work agency throughout the entire year in order to be informed about residential mobility; for example by sending them a brochure containing some results based on last years data.
- The interview situation (face-to-face) ensures a personal relationship, which makes it harder to withdraw from the survey. Thus, the stability of the interviewer over time is very crucial.

Fig. 1.3 illustrates the longitudinal development of first-wave respondents in 1984, as well as their children, of samples A and B.

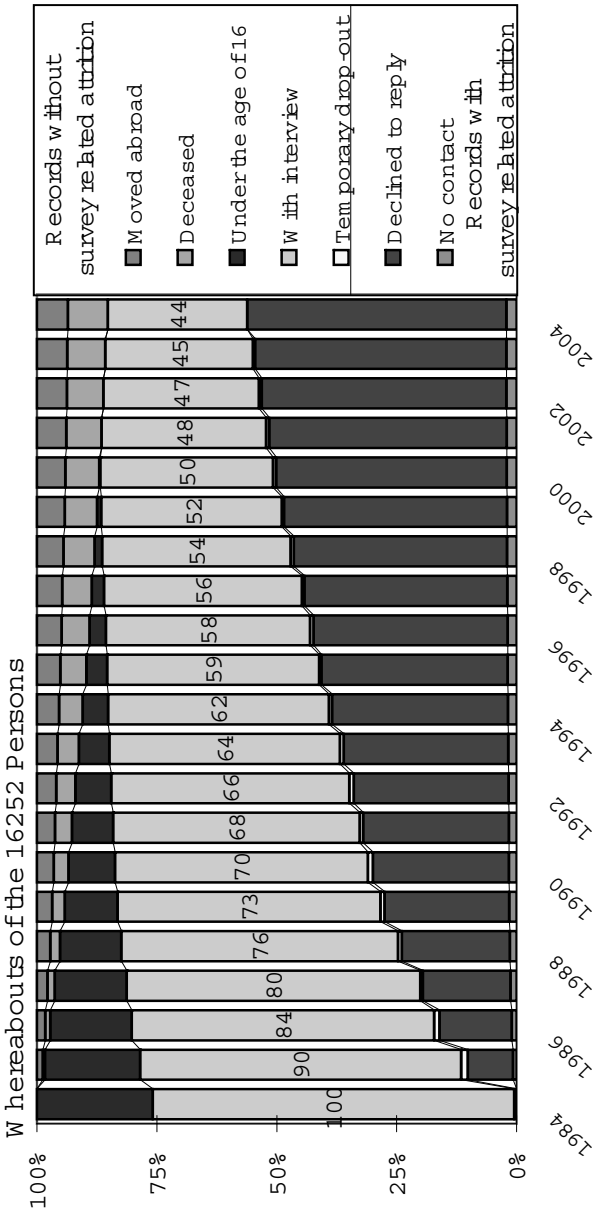


Figure 1.3: Longitudinal Development of the 1984 Population

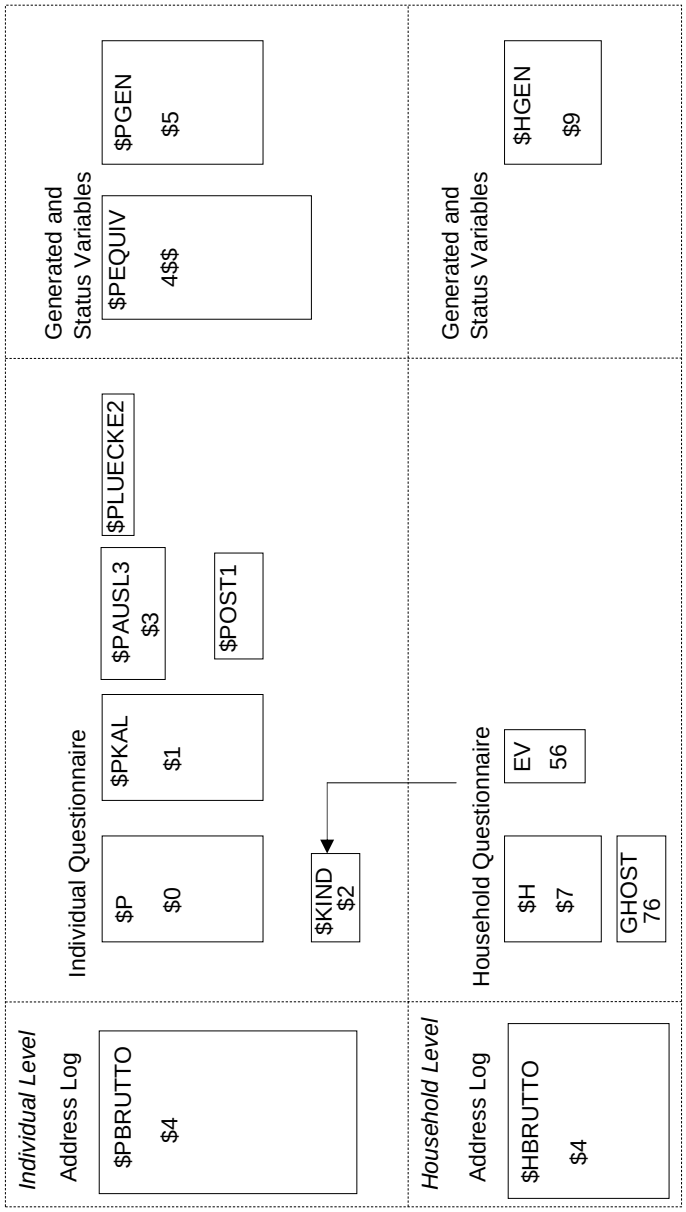
1.7 Principles of the Data Structure

In the SOEP, the entire data processing, including the data structure as well as variable naming, is very closely connected to the way the information was obtained. But some **very important variables** are independent of the questionnaire. Especially the variables for gender and the year of birth. These variables cannot change over time but they are asked (for control purposes) year by year. You find that information which is longitudinally checked in the file *PPFAD* under the variable names `SEX` and `GEBJAHR`. For more details see the description of this file below.

1.7.1 Cross-Sectional and Longitudinal Data Files

Since interviews are carried out on a yearly basis, the principle data structure is cross-sectional (See Fig. 1.4). For each year (or wave) of SOEP data there are single data files for households (*$H*) as well as for individual respondents (*$P*) and children (*$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" (see above), constitute the "gross" number of analyzable households and persons[6]. The naming of the data files is wave-specific, starting with A for the first wave in 1984, B for 1985, ... , U for 2004. For Sample C (East Germans), the first wave of data was gathered in 1990 (wave G in terms of the West German sample). For Sample D (Migrants to Germany since 1984), the first subsample was collected in 1994 (D1) and the second in 1995 (D2). Sample E (Refreshment Sample) which started in 1998, sample F (Innovation Sample) which started in 2000 (Wave Q), and Sample G (High Income) which started in 2002 (wave S) also include Germans and foreigners.

[6]The difference in the number of observations between gross and net population is made up by drop-outs and refusals.



$: Wave specification: A, B, C... U for file names; 1, 2, 3 ... 21 for file numbers.
1 Waves G and H only; 2 Waves B through Q only; 3 Waves A through L only

Figure 1.4: The Cross-Sectional Data Structure

Some variables which are often used, are not asked directly but must be calculated (for example the highest level of education). In order to make the database more user-friendly, the most important variables of this kind are calculated by the SOEP-group and stored in the database. You will find those "generated" variables in files with the extension *GEN*. Thus, you find generated information for persons (respondents) in the files *$PGEN* and information related to households in files named *$HGEN*. Please check the variables which are provided in those files, discussed further in Chapter 2, *before* you start your analysis! Please have a look at Table 2.4 on p. 67 and Table 2.5 on p. 68 *before* you write your first retrieval !

In order to facilitate the definition of longitudinal populations, the SOEP provides data files encompassing every individual respondent and child (file *PPFAD*) and any household (file *HPFAD*) ever contacted in the survey. (See Fig. 1.5) These files contain one record per unit of analysis and wave-specific variables indicating the survey status and necessary identifiers (see Chapter 2).

In order to control for non-random selection due to the sampling design and attrition, weighting factors are provided in the two files, *PHRF* at the person level and *HHRF* at the household level. Both contain information from all waves (see Chapter 5).

Cumulating drop-outs across all waves, *YPBRUTTO* contains information on the reason for temporary or permanent drop-out at the individual level. See Section 2.5 for more details.

Biography information at the individual level can be found in *BIOPAREN* concerning parental information and *BIOBIRTH* concerning birth information from women and *BIOBRTHM* containing birth information from male respondents. *BIOJOB* contains information on the first occupation and *BIOIMMIG* on the information related to immigration. *BIOYOUTH* contains biography information collected from 16-17 year olds, replacing the usual adult biography instrument. *BIOSOC* contains biographical information on the socialization process. *BICHILD* contains information from the "Mother and Child" newborn questionnaire. *BIOTWIN* identifies the unique person numbers for twins, triplets and quads. *BIORESID* contains information on second residency. (see Chapter 3).

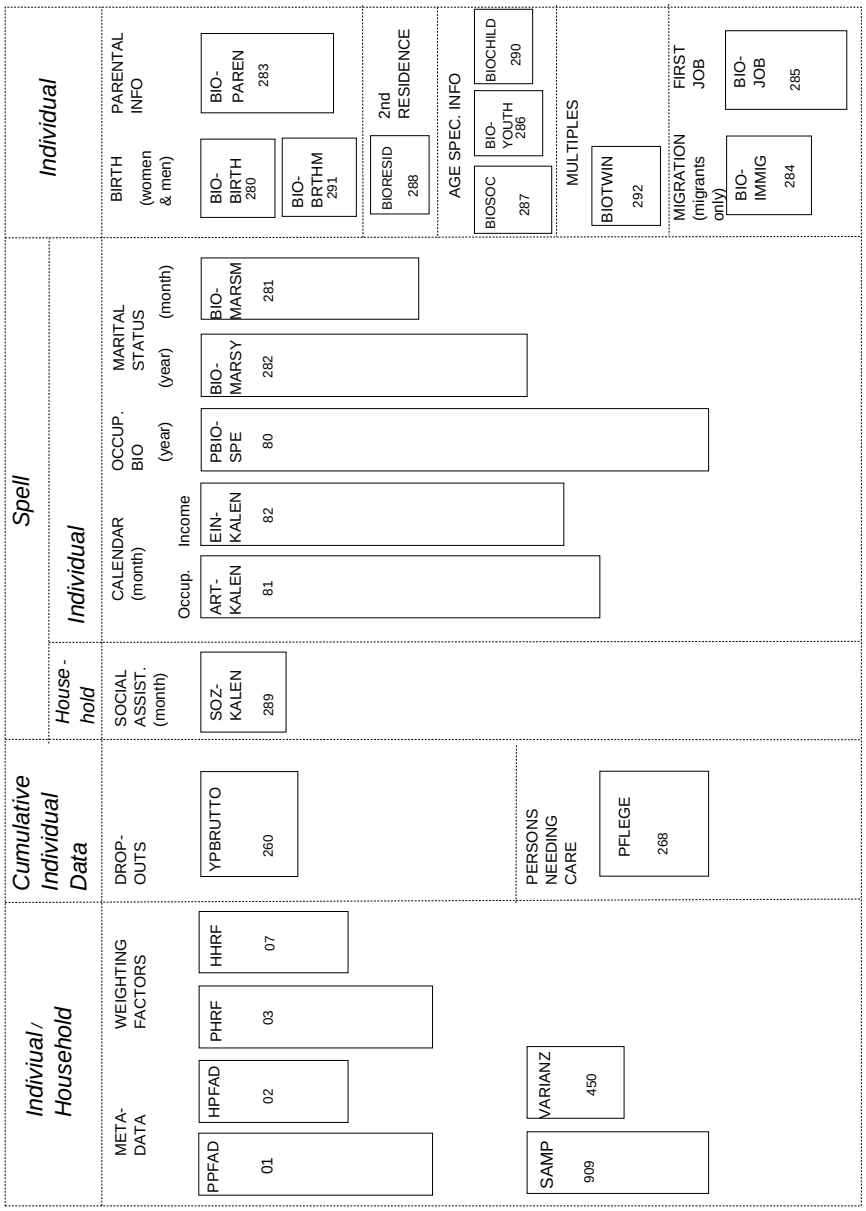


Figure 1.5: Longitudinal Data Files

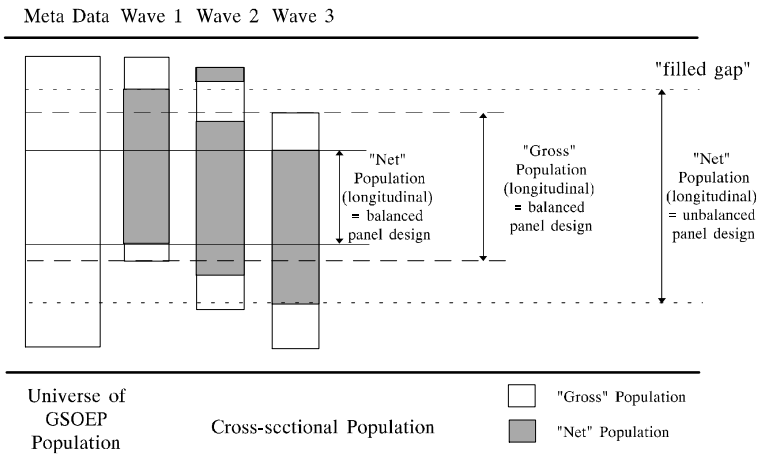


Figure 1.6: Cross-Sectional and Longitudinal Populations

Besides these files on individual and household level, the SOEP provides different kinds of biographical information per individual, which in the data base is stored as spell data. For each spell, there is a definition of the spell type, begin and end point, and the censoring status, indicating if a given employment or income spell is censored (left and/or right) or uncensored.

This information covers:

- the employment history over the entire life span from 15 up to a maximum of 65 years of age (file *PBIOSPE*).
- an activity calendar with up to eleven different activities (including full-time employment, unemployment, house-keeping, etc.) based on monthly data which is gathered in each interview for last year's January through December (file *ARTKALEN*).
- an income calendar with different sources of income (including income from employment, pension, unemployment assistance, etc.) based on monthly data which is gathered in each interview for last year's January through December until 1994 (file *EINKALEN*).
- marital history files, differentiated between yearly spells (*BIOMARSY*) since birth and monthly spells (*BIOMARSM*) since the preceding year of the first realized interview (for those in the first wave in 1984, this would be since January 1983).

Because of this decomposed data structure, cross-sectional as well as longitudinal analyses normally require matching or merging procedures. (See Fig. 1.6) In order to guarantee a unique match of individuals across time (longitudinal) and a perfect link between household data and the members living in this household, a set of identifiers is required. The central individual identifier across time is the `PERSNR`. This unique information is fixed, thus by definition does not change over time. Since a person might change

the household in which he or she lives at a given point in time, there are yearly household identifiers called HHNRAKT. The exact same information is also stored as $HHNR, allowing easier matching for some statistical software packages. 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[7]. 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), ..., UHHNR (wave 21).

[7] For example HHNR is used to divide at random the SOEP in different subpanels. These subpanels are used to create the 95%-Scientific-Use-Version of SOEP, using the variable RGROUP20, which divides the first waves into 20 equally sized groups. One of these groups is removed for the 95% version.

1.7.2 Variable Names and Missing Values

Survey variables might be missing due to different reasons. A person might refuse to answer to a question (e.g. income related questions) or just does not know an answer. These cases of item-non-response must be differentiated from a second reason of missing information, which occurs when a question does not apply to a household or person, e.g. the amount of rent to be paid just cannot be valid information to ask of owner-occupiers. The SOEP differentiates three kinds of missing values, as described in Table 1.4.

Table 1.4: Missing Values

Code	Meaning
-1	no answer / don't know: item-non-response
-2	does not apply
-3	after intensive checks a given value was found to be implausible and was finally deleted (to be interpreted like -1)

Variable names are set up in a way that indicates the year in which the data was collected, the unit of analysis, and additionally gives a reference to the original survey instrument. The following rules apply for all variable names in the SOEP of up to eight digits. See Table 1.5.

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`
5-6	Number of item in original question e.g. the "01" in `AP3301`
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 *Ausländer* 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

1.8 Overview of Weighting

For background reference for this section see also Chapter 5. The goal of any sample is to draw conclusions from the sample and apply them to the "recorded" target population. Due to different sampling probabilities, non-response in the first wave and attrition in the course of time, a weighting ("projection") of the sample cases is required in order to be able to infer the case numbers of the target population. For a panel like the SOEP we must distinguish two steps of weighting:

- Cross-sectional weighting
- Weighting of longitudinal populations.

Person-related weighting variables are stored in file *PHRF* (respondents and children). Household related weighting variables are stored in file *HHRF*.

The file *PHRF* contains one record for each person who was ever listed in the panel database. The entries are ordered by the original household ID (`HHNR`) and the person identification number (`PERSNR`). File *HHRF* contains one record for each household that was ever listed in the panel database. The entries are ordered by for the original household ID (`HHNR`) and household identification number (`HHNRAKT`).

All weighting variables are initialized to zero. Therefore the values of the variables are valid for all persons and households even if they did not participate in a particular panel wave or stopped participating altogether.

1.8.1 Cross-Sectional Weighting

The following variables may be used to weight cross-sections of waves 1 (wave A) through 21 (wave U) [8]

- In *PHRF*:
  `APHRF, BPHRF, CPHRF, ... UPHRF`
- In *HHRF*:
  `AHHRF, BHHRF, CHHRF, ... UHHRF.`

[8] An alternative weighting variable is available to users who do not want to use the weighting scheme that is based to some extent on fitting to population marginals. This alternative weighting variable takes into account solely the design selection probabilities of the sampling procedure and the regional characteristics of missing and participating households at the start of the panel. It is labeled `APDESREG` (in file *PHRF*) and `AHDESREG` (in file *HHRF*). In the first wave it can be used directly to estimate population totals. In all other waves it can be used to estimate population proportions but they do not take into account the ongoing attrition process. Thus we do not recommend these variables for substantial analysis of wave 2 and the following waves.

Results of Weighting: For sample A and B the statistical cross-sectional population is the population residing in primary residences in private households. Excluding persons residing in institutions, the weighting factors add up in each wave to the corresponding populations in the Micro-Census (a 1% sample of German households).

For sample C, the population adds up to the resident population.

Since subsample D was selected from a part of the population from which subsamples A–C were selected, from the 1995 on the cross-sectional weights of subsamples A–C were multiplied by 0.949. If, from 1995 until 1998, cross-sectional analyses are to be carried out and subsamples A–D are used then the standard cross-sectional weights can be used. However, if only subsamples A–C are used then, from 1995 until 1997 (waves L through N), the standard weights have to be multiplied by 1.053.

In 1998, a new sample was selected from the population of private households in Germany. The new sample, also denoted as subsample E, was selected independently from the ongoing panel (subsamples A through D). The selection scheme used for sample E essentially resembles the scheme also used in selecting subsample A. Therefore, the design weights are calculated in the same way as the wave one design weights for subsample A elements. For a detailed description of how "pure" design weights are derived, see Spiess (2000). Since subsample E was selected from the same population as subsamples A–D, cross-sectional weights had to be modified in a way to enable cross-sectional analysis of the whole sample (A–E) without any extra effort. As a consequence, if from 1998 (wave O) on (until 1999, wave P, see below), cross-sectional analyses are carried out using all the subsamples A–E, then the standard cross-sectional (individual as well as on household level) weights can be used. However, if only subsamples A–D are used then the weights have to be multiplied by 1.25. Correspondingly, if only subsample E is used, then the weights should be multiplied by 5. The derivation of these factors are described in more detail in Spiess and Rendtel (2000). Consequently, if only subsamples A–C are used, then the weights have to be multiplied by $1.316 \approx 1.25 \times 1.053$

Subsample F was selected independently from all other subsamples from the population of private households in Germany in 2000. With one exception, the selection schema was essentially the same as for selecting subsample A and F (c.f. section 1.4; for details, see Spiess, 2001). In the same way as described above, if from 2000 on (wave Q) all subsamples, i.e. A–F are used, then the standard weights can be used. If, however, only subsamples A–E are used, then the corresponding cross-sectional weights have to be multiplied by 1.82.

If only subsample F is used, the corresponding cross-sectional weights have to be multiplied by 2.22[9]. The derivation of these factors are given in Spiess (2001b). Accordingly, if only subsamples A–D are used then the weights should be multiplied by $2.27 \approx 1.82 \times 1.25$, if only subsamples A–C are used, the weights should be multiplied by $2.393 \approx 1.82 \times 1.25 \times 1.053$.

For estimation purposes it is recommended to use all subsamples, i.e. subsample A through F. If design-based estimators are calculated, then the standard cross-sectional weights `$HHRF` and `$PHRF` can be used. The Cross-Sectional weights for the waves G through N have been revised. The algorithm is described in Pischner (2000). See also Spiess and Rendtel (2000).

Starting with data released in 2002, there are some notable conceptual changes in the cross-sectional weighting scheme:

- For the initial weighting factors of sample F, more detailed design information is used, taking into account the sample-point specific realization rate of potential addresses.
- For each sample, A-F, the initial weights are "top trimmed" at a value of 10 times the respective median. This reduces the variation of weighting factors in general and the impact of outliers.
- Finally when adjusting these initial weights to marginal distributions of external statistics, i.e. from the German Mikrozensus, we now use more detailed age categories. This adjustment is made separately for East and West Germany, as well as for Samples A-E and Sample F.

Subsample G (Over-representation of high income households) is unique in that it does not have an analogous benchmark in any other major survey, beit panel or cross-section. That is why Sample G is not included in the overall standard weighting scheme of SOEP. However in separate variables integrated weights for samples A through G, starting in wave 2002 are provided: `$HHRFAG` and `$PHRFAG` at household and person levels respectively. Additionally G specific weighting factors exist: (`SHHRFG` and `SPHRFG` in files *HHRF* and *PHRF*, respectively).

[9] However there are already prepared cross-sectional weights which the user can immediately use for this purpose, e.g. `QPHRFAE` for sample A-E in contrast to `QPHRFF` for just Sample F.

1.8.2 Longitudinal Weighting

Longitudinal weighting is more difficult because the number of "longitudinal populations" gets bigger and bigger wave by wave. For example after five panel waves there are 10 possible longitudinal samples. Table 1.6 displays these longitudinal samples.

Table 1.6: Forming Longitudinal Samples

		Ends in wave				
		A	B	C	D	E
Starts in wave	A	x				x
		x				x
		x				x
		x				x
	B		x			x
			x			x
			x			x
	C			x		x
				x		x
	D				x	x
	E					

There is a flexible way of weighting longitudinal samples. The accompanying weighting factors can be easily determined by the use of the staying probabilities, i.e. the probability that a person or household participates in the named wave and also participated in the previous wave. These probabilities are calculated by the SOEP-group and the reciprocal of the "staying probability" is stored in the following variables:

- In *PHRF*:
  `BPBLEIB, CPBLEIB, DPBLEIB, ... , UPBLEIB`
- In *HHRF*:
  `BHBLEIB, CHBLEIB, DHBLEIB, ... , UHBLEIB.`

For example with the help of those variables, a longitudinal sample from wave 5 (wave E), to wave 21 (wave U) can be constructed. The weighting factor referring to persons can be labeled `EUPHRF` (but is not stored in file *PHRF*). The variable `EUPHRF` can be calculated as follows:

- `EUPHRF = EPHRF * FPBLEIB * GPBLEIB * HPBLEIB * ... * UPBLEIB.`

In general the weighting factor for a longitudinal sample can be calculated as the product of the weighting factor of the start wave and all the "reciprocal staying factors" to the end of the longitudinal sample.

Results of Longitudinal Weighting: The longitudinal sample of the statistical population from wave X to wave Y includes the following persons:

- Persons who belong in the cross-sectional statistical population in wave X
- Persons who remain in the survey area through wave Y (have not moved abroad or died). This longitudinal sample of the population is the natural statistical population. It is used for studying the life-course of panel members who participated in the panel from the first wave to the last wave of a given longitudinal period.

Therefore the longitudinal sample always contains only a part of the cross-section of the first wave. The difference is due to losses from migratory movements and deaths occurring over the course of successive waves. The population definition for migratory movements is as follows: the longitudinal section of the statistical population equals the cross-section of the statistical population minus the respondents who died. The population definition for respondents who died is as follows: the longitudinal section of the statistical population equals the cross-section of the statistical population minus the losses due to migratory movements.

1.9 SOEP Detailed Information

Further SOEP information sources, most of which are included on the CD distribution, include:

- SOEPINFO-WWW is the comprehensive documentation and analysis tool of the SOEP, containing the item correspondence list, questionnaires and frequencies of all variables across all waves. This gives a quick overview of all questions and generated variables. The item correspondence list is most important for longitudinal analysis. This allows the user to click on variables, or rows of variables in an item correspondence list, thereby creating a list of variables. With this list, the user can output frequencies, an item correspondence list, and generate automatically a syntactically correct command file in SPSS [10.0], SAS [8.2], and Stata [9.0]. This is intended as an aid in getting going quickly. Of course, the user will have to do some amount of programming himself, depending on the complexity of the retrieval. This version of SOEPINFO-WWW requires the browser MS Internet Explorer [5.5] or newer. SOEPINFO-WWW is bilingual: German and English. Click on `http://panel.gsoep.de/soepinfo/` for more information.
- German and English Translation of household and individual questionnaires for all waves.
- SOEPLIT-Win is an interactive database search program for Windows of all publications based on SOEP data, with a choice of German or English. Click on `http://www.diw.de/english/sop/soeplit/` for more information.
- SOEPLIT-WWW like the Windows version, but is available over the Internet, on the SOEP homepage. Included here are working papers, journal articles, books, monographs, collected volumes. Some "gray literature" has been left out in the WWW-version, such as citations for newspaper articles or official internal government reports. The data is in REFER format, also available on the CDROM. Click on `http://panel.gsoep.de/soeplit/` for more information.
- BIOSCOPE is a graphical viewer for the biography data in SPELL format. It is a program written for MSDOS.
- NEWSPELL is a program written for MSDOS (and also Windows) allowing the user to define his own spells. Especially for overlapping spells, this is very useful, as the user can define his own priority rules as to which event dominates in the case of overlapping spells. Optional output includes time-series and spell data.
- Mailing List Server at DIW, Berlin for questions, problem solving, contact with other users, mostly in German.
- The email address for SOEP technical questions and ordering a German version of the data is soepmail@diw.de. To order the SOEP international scientific use version, please contact GSOEP@cornell.edu. The email address for CNEF related questions is CNEF@cornell.edu.

External SOEP related packages can be obtained from:

- "SOEP Menu" is an external and separate data retrieval tool that eases data extractions from the German Socio-Economic Panel with Stata SE written by John Haisken-DeNew. See `http://www.soepmenu.de` and Haisken-DeNew (2005) for more details.

  John Haisken-DeNew. See `http://www.soepmenu.de` for more details.

- Ulrich Kohler, WZ Berlin
  - Random Group Variance Estimator "rgroup"
    In Stata, type: "`net search rgroup`", then install "rgroup"
  - Stata and SOEP related resources:
    `http://www.wz-berlin.de/~kohler/lehre/index.html`

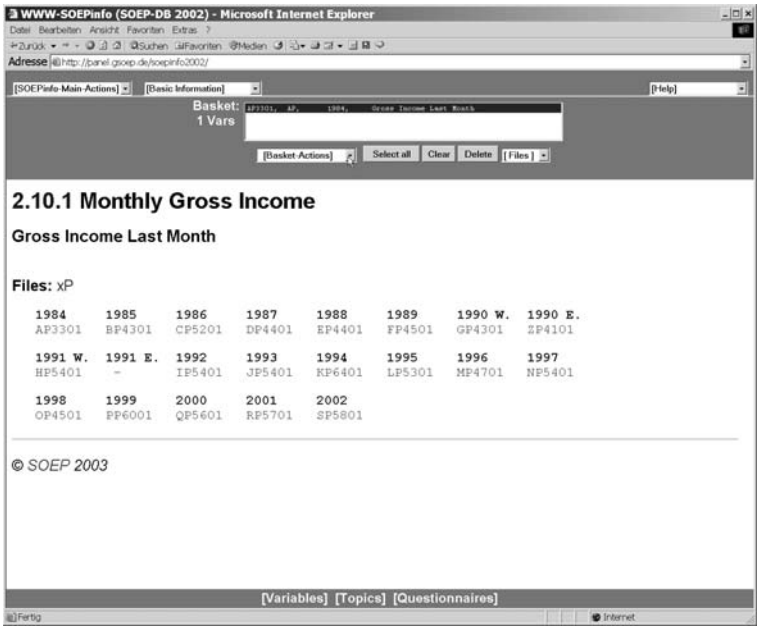


Figure 1.7: SOEPINFO: Item Correspondence List

1.9.1 SOEPINFO

SOEPINFO is a program written by Ingo Sieber and John Haisken-DeNew for the "MS Internet Explorer". It allows the researcher to search online for information on almost all variables in the SOEP. For instance, one can search by explicit variable name, if it is known, or by thematic category. Due to the dynamic nature of the SOEP, variable names over time are not consistent, but rather correspond to the order of the variables in the respective questionnaires. Some variables have been additionally generated as well, leading to the difficult task of somehow grouping variables over time. Thus, if one knows that the Gross Wage in 1984 is the variable `AP3301`, then one can find out what the variables for Gross Wage in all other years are called. This is very useful for panel studies, such that the work of selecting variables is done once.

Thus in Figure 1.7, the variables for the West German sample are: `AP3301`, `BP4301`, `CP5201`, `DP4401`, `EP4401`, `FP4501`, `GP4301`, `HP5401`, `IP5401`, `JP5401`, `KP6401`, `LP5301`. For the East German sample, for 1990 the variable is `ZP4101`, followed by `HP5401`, `IP5401`, `JP5401`, `KP6401`, `LP5301`, `LP5301`, `MP4701`, `NP5401`, `OP4501`, `PP6001`, `QP5601`, etc for both east and west. Having this program "online" obviates the need to have many many printed manuals, and allows quick and easy searches, where printed manuals are very cumbersome.

An additional feature of the program is that frequencies can be viewed for

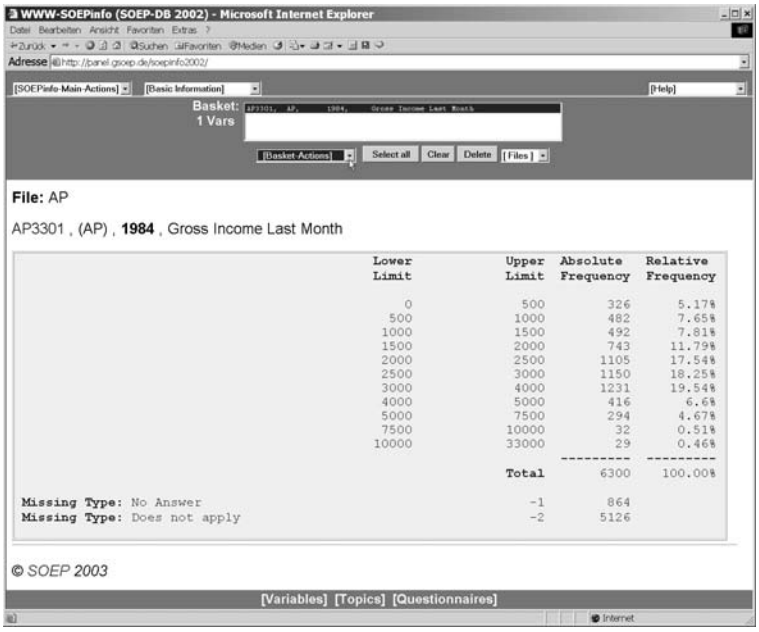


Figure 1.8: SOEPINFO: Frequencies

almost all variables. Continuous variable typically have frequencies for ranges of values, and discrete variables have the usual frequencies. For example, our variable `AP3301` is depicted in Figure 1.8. By adding the variable `AP3301` to the "basket" above, and selecting "Frequencies" and then "Submit", frequencies information is displayed. For 6300 cases there is valid information, whereas 864 did not give an answer (item-non-response) and were coded "-1" and for 5126 cases, the question did not apply (they were not working) and were coded "-2". This is very useful when first starting research projects, to scout out what the data can deliver. If you are interested in the number of self-employed females working in the agricultural industry over the age of 55, you might want to first check, if there are any workers in the agricultural industry to begin with, *before* you write your retrieval. This has, on countless occasions, saved many many hours of unnecessary work for the people who use SOEPINFO regularly!

Nevertheless, using SOEPINFO only, is not an adequate method of finding out more involved questions such as, "how many employed females over the age of 55 gave an interview ?" To do this, you would have to write a retrieval and examine the actual data directly. See Chapter 4 for more details.

In Figure 1.9, the "META" variable information is clearly illustrated. Further, some variables have been generated to handle information concerning

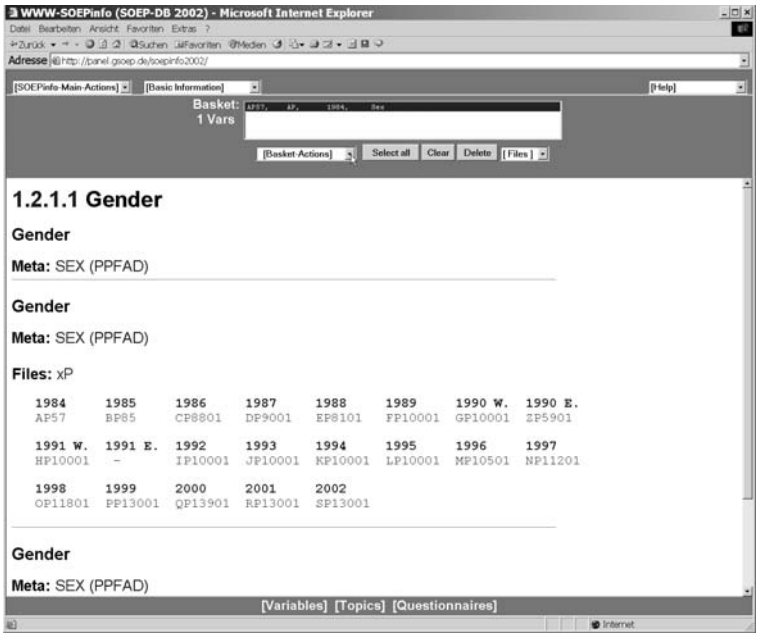


Figure 1.9: SOEPINFO: Meta Information

several waves, and/or longitudinally checked. An example of this is the variable `SEX` in the file *PPFAD*, as shown in the META column. The following gender indicators `AP57`, `BP85`, `CP8801` etc, come from the *$P* files. However, `SEX` is longitudinally checked, and is the *better source of information*. The variables coming from the *BIO* files are almost all META variables, as the information is based on several time periods and/or has no time dimension. Longitudinal weighting factors follow this pattern as well.

1.9.2 "SOEP Menu" for Stata SE

"SOEP Menu" is an external and separate data retrieval tool that eases data extractions from the German Socio-Economic Panel with Stata SE written by John Haisken-DeNew[10]. SOEP Menu is directly combined into Stata SE, allowing a seamless interaction between the micro data and the statistics package. The user can open SOEP data files by drop down menus.

The most important feature of SOEP Menu is that one can select *not just* single variables, but rather entire *vectors* of variables, called item-correspondences all at once[11]. For instance, one can select wage information for all years, say years 1984 through 2004. Special cleaning programs written in Stata called "plugins" can clean a particular item-correspondence and make it time and/or content consistent. Groups of item-correspondences can be stored as projects. Groups of projects can be stored as libraries. This method of organizing the projects and plugins allows for a modular administration, facilitating knowledge transfer and group work.

Data can be retrieved by mouse-click, providing rectangularized data in wide and long format. All programs used are available in source Stata code which allows complete transparency of content. All commands used in the generated retrieval are documented in a full functional retrieval DO file, capable of recreating the identical retrieval at any time[12].

The idea behind the tool is that because of the intrinsically longitudinal nature of the data, one is interested NOT in retrieving a variable in a particular wave, but rather in retrieving the variable for several waves, i.e. an item-correspondence list. An additional complicating factor is that in the SOEP, the variable naming algorythm used reflects the order of the questions in the questionnaire, NOT reflecting any particular content. This obviously implies a changing variable name over time. See Haisken-DeNew (2001) for more information on this. Thus, if one opens a SOEP data file and one finds a variable of interest, one clicks on the variable and information for the entire item correspondence is also collected and added to a "soep project". Straightforwardly, the object is to collect items and save them into the project, allowing an automatic data retrieval.

[10] See `http://www.soepmenu.de` and Haisken-DeNew (2005) for more details. Persons interested in using SOEP Menu are required to complete and sign a user contract and are required to make a donation directly to UNICEF in the amount of EUR 10 per licenced user. Further, users are required to cite the use of SOEP Menu in their projects.

[11] This item-correspondence information is provided by Ingo Sieber and Jan Göbel at the DIW.

[12] Although very different in its implemetation and historical development, SOEP Menu is a logical extension of the features built into the web version of "SOEPinfo" established by Ingo Sieber and John Haisken-DeNew at the SOEP, however SOEP Menu runs only in Stata. It cannot be used with SAS(R) or SPSS(R). However, data generated with SOEP Menu can be *exported* to these formats.

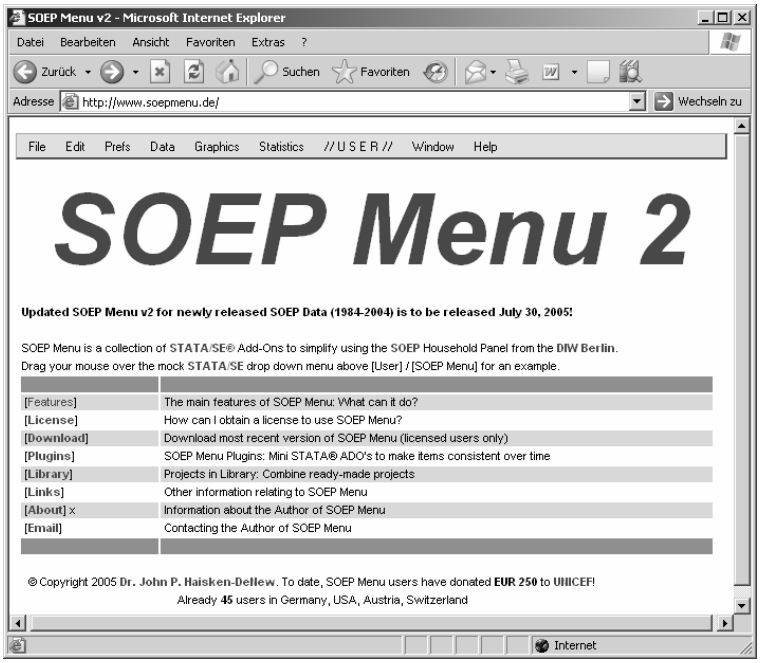


Figure 1.10: SOEP Menu Website

SOEP Menu allows one to create a "`soep project`" by various means. One can directly open a SOEP data file and add items. One can simply append other soep projects. If a SOEP data file is in memory, one can either use the drop-down-menu tool or one can create a "browse page". The drop-down-menu tool allows one to view all items found in a dataset, to search for items based on the Stata "lookfor" command. There are extensive search functions available allowing one to search according to arbitrary keywords or JEL[13] keywords: within a file, within a wave, or throughout the entire SOEP data distribution. Once items have been found, they can be added to the project. The project can be named, renamed, saved, resaved, deleted etc. The saved soep projects are saved as Stata dta files with the filename ending "*∗.soep*". If one saves projects in a modular fashion, one can create quickly full projects, e.g. *wages.soep*, *firm.soep*, and *humancapital.soep* get appended together to create *labor.soep*.

Assuming that a project is complete (the user has found all the items of interest), then by clicking on a button, the retrieval can be executed. As new releases of the SOEP micro data become available, the user can "automat-

[13]See `http://www.aeaweb.org/journal/jel_class_system.html` for more information on the JEL classification scheme.

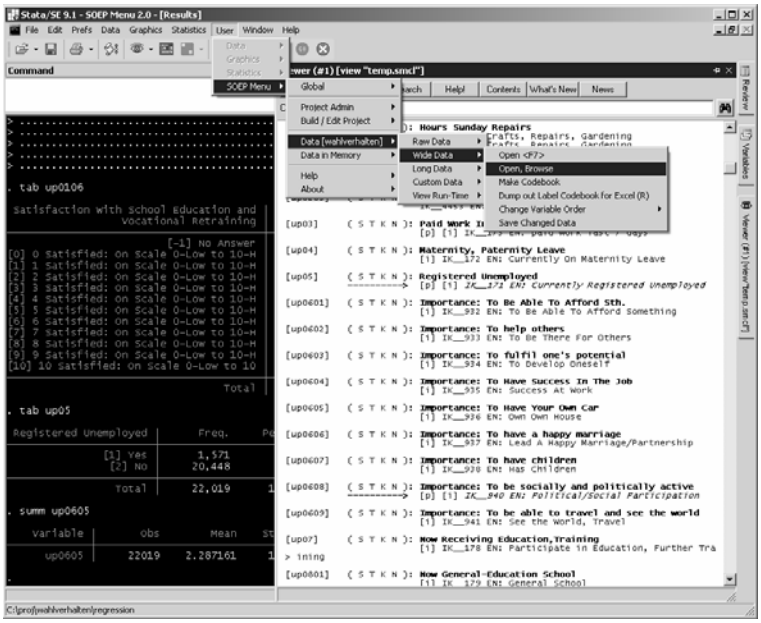


Figure 1.11: SOEP Menu for Stata/SE in Action

ically" update his projects to include the lastest wave of information. This obviously requires that the SOEP data be located on a local or network disk that the user's computer has access to.

As a user of panel data, one would eventually like the SOEP data to be in "long" form. SOEP Menu retrieves in both manners, wide and long using the command reshape. This requires some standardization of variable names over time. Indeed, items in the SOEP have been given a serial number. Thus wide variables from the item number `2277` would be `ap2277x`, `bp2277x`, ..., `up2277x`. The first letter indicates the wave `[a-u]`, the second letter `[p,h]` indicates whether at the person or household level, and "`x`" at the end indicates a SOEP Menu variable. Thus after the reshape command has be executed, the first letter indicating the wave is dropped to create the long variable name for example, `p2277x`. The user can access the retrieved data at the "wide", "long" and "custom" level.

Using the SOEP, one immediately is confronted with the fact that values of variables in any particular item may change over time. For instance if `1=yes` and `2=no` in 1984, there is no particular reason to assume this will remain for all years following. The following year may have `1=yes, 2=maybe, 3=no`, problematic for long data sets. Hence, using "plugins", allows the user to "clean up" these value inconsistencies. A plugin is simply an add-on in its

own right with the same name as the long variable name, e.g. *p2277x.ado*, and is executed whenever that item is selected in the retrieval. For instance, one can use a plugin to deflate income measures with a price index automatically. Theoretically, there should be a plugin for each of the approximately 4,500 items, which would guarantee that all values would be consistent over time. In fact, if there is no plugin available for an item, value labels are stripped from the long variable to ensure that the user examines the item consistency over time.

The Stata files used with SOEP Menu have been preloaded with both German and English variable labels and values labels. The user can at any time switch between languages. Indeed, one can have an English version of the SOEP on disk and create a fully German labelled retrieval.

If for some reason, the data just created should be exported into another format, one can use several commercial tools such as StatTransfer(R) or DBMSCOPY(R), or several tools made available here. SOEP Menu allows exporting data to SPSS(R), SAS(R), LIMDEP(R), GAUSS(R), MS Excel(R) and also SQL systems. If the data is in memory, then it can be exported. In addition, the entire SOEP data set can be dumped into SQL files for import in MySQL(R). Where variable labels and value labels have a relevant meaning in the exported format, they are kept. For instance, Limdep and GAUSS do not have a concept of value labels and MS Excel does not have a concept of neither value nor variable labels.

As mentioned, SOEP Menu requires a standardized naming scheme to reshape from wide to long format. This creates variable names of the following sort: `p2777x`. This is perhaps not so intuitive however SOEP Menu allows users to have the option of customizing SOEP Menu to rename SOEP Menu variables with other arbitrary names. Thus, whenever the SOEP Menu variable `p2777x` is retrieved in a "long" file, SOEP Menu automatically creates an additional "custom" file in which SOEP Menu variables have been renamed to something perhaps more intuitive such as "`wage`" or "`labor_income`". The renaming feature is only activated if the user explicitly requests it. The user can provide custom names for variables extracted. Should a variable be found in the list of variables to be renamed, the variable is renamed according to the instructions of the user. The exact renaming executed in the retrieval is documented in the generated retrieval DO file. Thus, there is unambiguous information as to the origins of each and every variable.

If you are a Stata/SE user of the SOEP, SOEP Menu will certainly save you a tremendous amount of time. However, if you decide to write your data retrievals from first principles in Stata, this is described later in Chapter 4.

SOEP Menu Main Features

Creating SOEP Menu Projects (*∗.soep*)

- Easily extract data from German Socio-Economic Panel (SOEP), simply by using your mouse and clicking.
- More than 240 original Stata files in the SOEP, no problem! SOEP Menu merges files automatically!
- You can open data files: SOEP Menu autoamtically scans the data files for "meta" data and creates browse pages.
- You can select vectors of variables, called items, i.e. the same variable for all years at once.
- You create SOEP Menu Projects by collecting items together. These items can be stored as a SOEP Menu project file.
- The SOEP Menu project files can be stored separately and appended together in a modular manner.
- Projects can be automatically updated as new releases of SOEP become available.
- Project files (*∗.soep*) can be legally shared over the internet as they contain only "meta" information, not micro data.

Creating SOEP Menu Libraries

- A collection of SOEP Menu Projects is called a library.
- There are private, public and internet libraries available to the user.
- A "private library" is on the user's own disk.
- A "public library" is a shared disk on a network.
- An "internet library" is a collection of contributed SOEP Menu projects available to everyone.
- Using libraries allows easiest project creation. The work is done once and only once!

Using SOEP Menu Plugins

- SOEP Menu allows for item-specific plugins to ensure consistency over time.
- Around 1000 items have plugins already! Users are invited to contribute to the plugin knowledge base.

Types of Output Data

- Raw: Data are merged together in wide format.
- Wide: Raw data is checked and/or adjusted for longitudinal consistency with plugins.
- Long: Wide data are transposed to Long data (person year observations).
- Custom: Long data variables are renamed to user specific "speaking names", like "wage".

Exporting Data to Other Formats

- Once you have created a SOEP retrieval, you can export the output data to other formats:
- SPSS (R): Creates an ASCII file and an SPSS syntax file to read in the data, maintaining all labels.
- SAS (R): Creates an ASCII file and an SAS command file to read in the data, maintaining all labels.
- LIMDEP (R): Creates an ASCII file and a LIMDEP command file to read in the data, documenting all labels.
- GAUSS (R): Creates an ASCII file and a GAUSS AtoG command file to read in the data.
- EXCEL (R): Creates a separated ASCII file for import into EXCEL.
- MySQL (R): Creates an ASCII file and a MySQL command file to read in the data, maintaining all labels.

Multilingual Labels

- SOEP Menu supports English and German value and variable labels. Users can switch instantaneously.
- Projects can be created in one language, shared with a research partner and switched to another language.

1.9.3 BIOSCOPE

For those researchers interested in spell data, that is micro data about events or the length of time events occurred, like unemployment, training, retirement etc., the program, written by Rainer Pischner for use under MSDOS called BIOSCOPE, can be extremely helpful. This program allows graphical viewing of these events for each person in the sample. The program uses retrospective information collected in in the first wave (1984) and updates that information using the monthly calendar information for all waves thereafter. See Chapter 2 for more details. For example, in Table 1.7 the person 201, who is a women, born in 1926, was in schooling from age 15-19, then from 20-24 did some kind of training, and then worked full-time for 5 years. From the age of 30 to 39, she was a housewife, and then at the age of 40 retired. At the age of 56 and 57 she had overlapping retired and housewife spells.

Table 1.7: BIOSCOPE: Graphical Spell Data Representation

```
+-------------------------------------------------------------------------------+
|                                                                               |
| BIOSCOPE                                                                      |
|                                                                               |
|  CASE-ID:     27  Personal-Number:      201   Year of birth:1926    Sex:  2   |
|                                                                               |
|                                                                               |
|                                                        1983-->|<--1984        |
| Any Activity........ xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxHHHHHHHH  |
|                      <--1941 Start of Working Life History                    |
|                                                                               |
| At Age of      ...   1    2    2    3    3    4    4    5    5    6    6 Sum of |
| the person was ...   5    0    5    0    5    0    5    0    5    0    5 events |
|                                                                               |
| School, University.. XXXXX|....|....|....|....|....|....|....|....|....|    5   |
| Training etc........ |....XXXXX|....|....|....|....|....|....|....|....|    5   |
| Military Service.... |....|....|....|....|....|....|....|....|....|....|    0   |
| Full employed....... |....|....XXXXX|....|....|....|....|....|....|....|    5   |
| Part time employed.. |....|....|....|....|....|....|....|....|....|....|    0   |
| Unemployed.......... |....|....|....|....|....|....|....|....|....|....|    0   |
| Housewife,-man...... |....|....|....XXXXXXXXXX|....|....|....|.OO.|....|   12   |
| Retire.............. |....|....|....|....|....XXXXXXXXXXXXXXXXXXOOXXXXXXX   26   |
| Other activity...... |....|....|....|....|....|....|....|....|....|....|    0   |
|                                                                               |
|                                                                               |
| Input:<Personal-Number>,<+n>,-<n>,<Up>,<Dn>,<PgUp>,<PgDn>,<ESC> 201           |
|                                                                               |
+-------------------------------------------------------------------------------+
```

1.9.4 SOEPLIT

Finally, given that the SOEP has been running for so many years now, the odds of wanting to research an area where others have already researched with the SOEP is very large. So why re-invent the wheel ? Have a look at what others have done, what pitfalls or problems others have had when analyzing the data. The program SOEPLIT, as seen in Figure 1.12, will help you to do exactly this. It lists all known SOEP publications, and these can be searched as a database by keywords, data, authors etc. This information can be printed out in ASCII format, and easily integrated into your new research paper in the references section.

Just a reminder: All users, upon signing the SOEP users contract, are obliged to send a copy of their SOEP papers to the DIW. At the latest, this should occur at the working paper stage. Should the papers be finally published (and using the SOEP, of course they will be!), then a copy of the final published version must also be sent to the DIW. This way, all papers, books, etc., can be catalogued. Please be fair, and send us your work. The sooner you send us your papers, the faster others will cite you, and make you famous.

There is also a Windows version of SOEPLIT available from the CDROM distribution and also for download at the SOEP homepage. Figure 1.13 illustrates the selective search input mask and Figure 1.14 illustrates the data browsing capabilities of the Windows version.

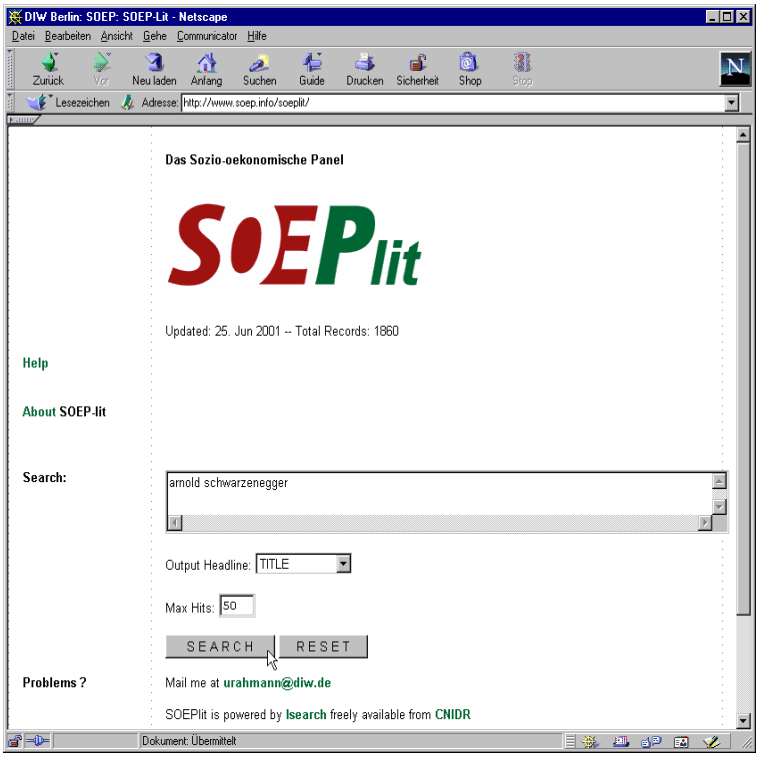


Figure 1.12: SOEPLIT: Working Papers / Books / Journal Articles

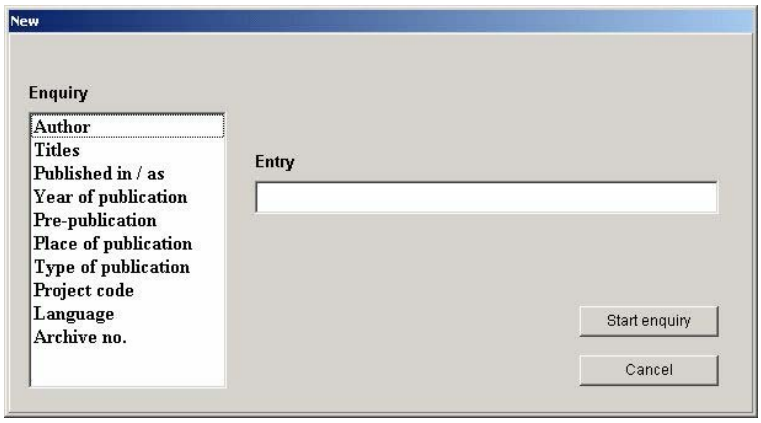


Figure 1.13: SOEPLIT-Win: Search Mask

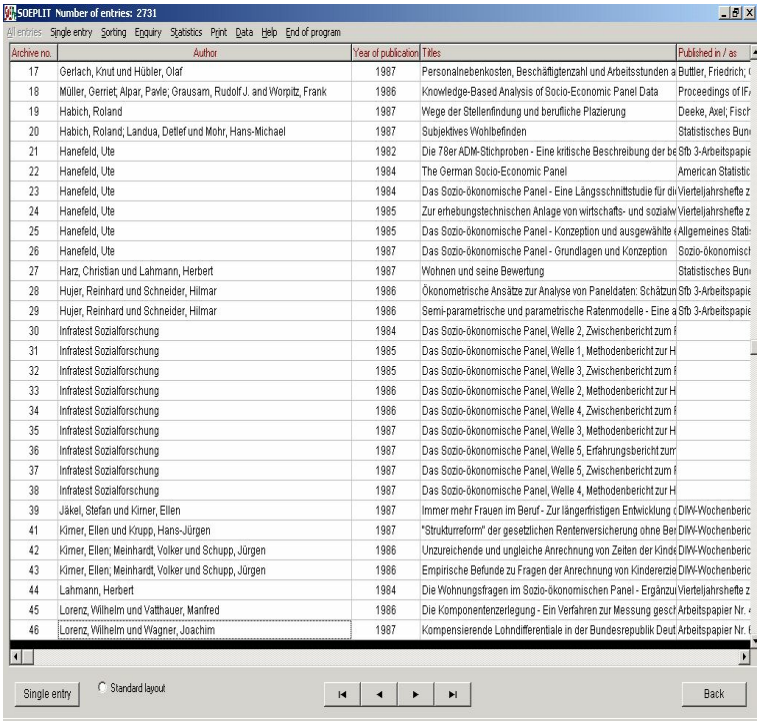


Figure 1.14: SOEPLIT-Win: Open Search

Tables 1.8 through 1.10 list the winners of the prizes for "Best papers presented at the SOEP conferences" and "Best SOEP-Papers Published" respectively. The list clearly indicates a diverse mix of topics in economics, sociology, geography, etc; showing the range of topics that can be covered using the SOEP data.

Table 1.8: Best SOEP Papers Presented

Best papers presented at the GSOEP2004 conference	
Frijters, Haisken-DeNew, and Shields (2004a)	How Well Do Individuals Predict Their Future Life Satisfaction? Evidence from Panel Data Following a Nationwide Exogenous Shock
Geishecker and Goerg (2004)	International Outsourcing and Wages: Winners and Losers
Siedler (2004)	Is the receipt of social assistance transmitted from parents to children? Evidence from German panel data
Best papers presented at the GSOEP2002 conference	
Biewen (2002)	The Covariance Structure of East and West German Incomes and its Implications for the Persistence of Poverty and Inequality
Pannenberg (2002)	Long - Term Effects of Unpaid Overtime: Evidence for West Germany
Schraepler (2002)	Respondent Behavior in Panel Studies - A Case Study for Item-Nonresponse by Means of the GSOEP
Best papers presented at the GSOEP2000 conference	
Clark, Diener, and Georgellis (2000)	Lags and Leads in Life Satisfaction: A Test of the Baseline Hypothesis
Kreyenfeld (2000)	Timing of First Births in East Germany after Unification
Margolis, Simonnet, and Vilhuber (2000)	Early Career Experiences and Later Career Outcomes: A Comparison of the United States, France and Germany

Table 1.9: Best SOEP Papers Published

Best SOEP-Papers 2003-2004	
DiPrete and Engelhardt (2004)	Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation
Bishop, Chow, and Zeager (2003)	Decomposing Lorenz and Concentration Curves
Constant and Massey (2002)	Return Migration by German Guestworkers: Neoclassical versus New Economic Theories
Frijters, Haisken-DeNew, and Shields (2004b)	Investigating the Patterns and Determinants of Life Satisfaction in Germany Following Reunification
Gangl (2004)	Welfare States and the Scar Effects of Unemployment: A Comparative Analysis of the United States and West Germany
VanKerm (2004)	What Lies Behind Income Mobility? Reranking and Distributional Change in Belgium, Western Germany and the USA
Horstkötter and Hübner (2004)	*Verteilter Wohlstand*
Borchard-Tuch (2004)	*Unzufrieden ohne Job*
Best SOEP-Papers 2002-2003	
Lucas, Clark, Georgellis, and Diener (2003)	Reexamining Adaptation and the Set Point Model of Happiness: Reactions to Changes in Martial Status
Dustmann and vanSoest (2002)	Language Fluency and Earnings: Estimation with Misclassified Language Indicators
McGinnity (2002)	The Labour Force Participation of the Wives of Unemployed Men
Schmid (2002)	*Was heisst schon solidarisch? Auch unter der rot-grünen Regierung bleibt die Kluft zwischen Arm und Reich in Deutschland bestehen*

Table 1.10: Best SOEP Papers Published Continued

Best SOEP-Papers 1999-2001	
Fabig (1999)	Income Mobility and the Welfare State: An International Comparison with Panel Data
Goodin, Dirven, Headey, and Muffels (1999)	The Real Worlds of Welfare Capitalism
Huebler and Koenig (1999)	*Betriebliche Weiterbildung, Mobilitaet und Beschaeftigungsdynamik - Empirische Untersuchungen mit Individual- und Betriebsdaten*
Kraft (2001)	Unemployment and the Separation of Married Couples
Oxley, Dang, and Antolin (2001)	Poverty Dynamics in Six OECD Countries
Siebern (2000)	Better LATE? Instrumental Variables Estimation of the Returns to Job Mobility during Transition
Szydlik (2000)	*Lebenslange Solidaritaet? Generationenbeziehungen zwischen erwachsenen Kindern und Eltern*
Best SOEP-Papers 1984-1998	
Blau and Riphahn (1999)	Labor force transitions of older married couples in Germany
Buechel and Falter (1994)	*Der Einfluss von Langzeitarbeitslosigkeit auf die Parteibindung in der Bundesrepublik Deutschland* (Long-Term Unemployment and Political Party Preferences)
Clark, Deurloo, and Dieleman (1997)	Entry to Home-ownership in Germany: Some comparisons with the United States
Winkelmann and Winkelmann (1993)	Why Are the Unemployed So Unhappy? Evidence from Panel Data
Hunt (1999)	Has Work-Sharing Worked in Germany?
Juerges (1998)	*Beruflich bedingte Umzuege von Doppelverdienern - Eine empirische Analyse mit Daten des SOEP* (Dual Earner Moves for Job Reasons)
Lechner (1998)	Training the East German Labour Force

1.9.5 Recoding SPELL data with newspell.exe

Newspell is a program written for MSDOS by Rainer Pischner. It allows one to redefine spells, and select appropriate time intervals. The program is called with two arguments: the command/parameter file and the current password for the SOEP distribution.

```
c:\gsoep\newspell nspell.cmd  <password>
```

The example command file `nspell.cmd` contains all of the parameters that **newspell.exe** requires. Here an example is displayed. Using the `PI=` command, the input data file directory is defined (with the ZIP archived data files, or plain ASCII files). The `NI=` command defines the specific spell data file to be used. There are currently several SOEP data files with information in spell format: *ARTKALEN*, *EINKALEN*, *PBIOSPE*, *BIOMARSY*, *BIOMARSM* and *SOZKALEN*. The output directory for the output data files and log file is defined by `PO=`. Using the `NT=` command, the time series output file name is defined, and with `NS=` the new spell output file name is defined. Using the `%=` command defines which sample to use. Currently this is the 100% research sample. The command `NL=` defines the output log file. The real action of redefining spell types is when the user types something like `1=1 or 2 or 3`, thereby redefining 1, 2 and 3 to become 1. The user can define starting and ending periods as well with the `NB=` and `NE=` commands. The units of these periods depends on the spell file at hand.

For more detailed information, see the extensive NEWSPELL documentation listed in the *INDEX.HTM* file on the CDROM.

Table 1.11: A Sample **newspell.exe** Command File

```
                                        /* === Contents of "nspell.cmd" ===
PI=M:\DATA\GSOEP                        /* input data file directory
NI=ARTKALEN.DAT                         /* spell file input
                                        /* (ARTKALEN, EINKALEN, SOZKALEN
                                        /* PBIOSPE, BIOMARSY, BIOMARSM)
PO=G:\NEWSPELL\TEST                     /* output directory
NT=ARTKAL_1.TIM                         /* Time Series Output File Name
NS=ARTKAL_1.SPL                         /* Spell Form Output File Name
%=5                                     /* Sample: 5, 50, 100
NL=ARTKAL_1.LOG                         /* Logfile
                                        /* redefine spell types
1=1 or 2 or 3 or 4 or 11                /* new = old
2=5 or 6 or 7 or 8 or 9 or 10 or 12     /* new = old
NB=73               /* begin month
NE=144              /* end month
```

1.9.6 Mailing List Server in Berlin, Germany

At the SOEP in Berlin there is also a mailing list. These emails tend to be in German, but are often in English. To join or **subscribe** to the mailing list, send an email to:

`sympa@list.diw.de`

with the subject containing only **subscribe soep-l**.
Only once you have registered with the list, can you **send** an email to the entire mailing list by sending an email to the following address:

`soep-l@list.diw.de`

The experience gained with the mailing list so far has been very positive. This is by far the fastest method of announcing bug-fixes, new services and solving user problems. Our motto is: *There are no stupid questions !* If you are stuck and cannot go any further, chances are that others are (or have been) in the same situation, and this invaluable experience can be drawn on. Although we have tried to improve documentation dramatically, sometimes there are unfortunately still some ambiguities. This is a perfect way to help us keep our documentation up-to-date and accurate. With the mailing list, when an answer to the problem at hand is found, *everyone* gets to share in the knowledge.

Chapter 2

Survey Extensions

by Joachim R. Frick, John P. Haisken-DeNew, Peter Krause, Rainer Pischner, Jürgen Schupp and C. Katharina Spiess

In this chapter, some extensions to the questionnaire data will be explored. The longitudinal files *PPFAD* and *HPFAD*, containing longitudinally checked information important for any panel study, will be described. For those interested in duration or event analysis concerning labor market related issues, the spell data in *PBIOSPE*, *BIOMARSM*, *BIOMARSY*, *ARTKALEN*, *EINKALEN* and *SOZKALEN* will be examined. As many variables are not explicitly asked every year, but rather only when something has changed, generated variables have been created to fill in missing values. The variables in *$PGEN*[1] and *$HGEN* will be explained.

2.1 Basic Information

2.1.1 Person-Level Longitudinal File: *PPFAD*

This file includes all members of all households ever contacted in the SOEP including respondents, children, and even those who never gave an interview. For each person it includes the current household identifiers for each wave

[1]The Dollar-Sign ($) used in a variable name or data file represents a "wild-card", and stands for each and every wave (A-U). Thus *$PGEN* stands for *APGEN*, *BPGEN*, ..., *UPGEN*. See Table 1.5 for more information on the naming of variable names.

`$HHNR` as well as wave-specific variables concerning the survey status `$NETTO`. Thus *PPFAD*'s central function is to facilitate the definition of longitudinal populations.

Table 2.1: Example for `$NETTO` variables:
`QNETTO`: Survey status 2000

Label	Value	Frequency 100%	Frequency 95%
in *YPBRUTTO*	0	217	206
in *QP*	1	24,586	23,341
in *QKIND*	2	6,659	6,295
in *QPBRUTTO*	3	2,640	2,475
in *QPLUECKE*	4	190	179
	Total	34,292	32,496
Missing*	-2	16,147	15,347
Total in *PPFAD* (as of 2001)		50,439	47,843

* Individuals who did not yet enter the sample or who already left.

Additionally *PPFAD* contains the longitudinally checked information on gender and year of birth for each individual. For the sake of consistency this information should be used in longitudinal analyses, since it can not be taken for granted that the corresponding yearly, cross-sectional information is perfectly stable.

Table 2.2: List of variables in *PPFAD*

Variable Name	Meaning
HHNR	original household identifier (case) from wave 1
PERSNR	unique individual identifier
PSAMPLE	sample identifier
SEX	gender (longitudinally verified)
GEBJAHR	year of birth (4 digit) longitudinally verified
GEBMONAT	month of birth (2-digit) longitudinally verified
TODJAHR	year of death (4-digit)
TODINFO	source of information to compute year of death
EINTRITT	year in which individual entered the survey (4 digit)
AUSTRITT	year in which individual left the survey (4 digit)
ERSTBEFR	year in which first individual interview was conducted (4 digit)
LETZTBEF	year in which last individual interview was conducted (4 digit)
IMMIYEAR	year of immigration to Germany
GERMBORN	whether German born or not
LOC1989	location in 1989
CORIGIN	country of origin
AHHNR	household identifier 1984
BHHNR	household identifier 1985
CHHNR	household identifier 1986
$HHNR	household identifier ...
UHHNR	household identifier 2004
ANETTO	survey status 1984
BNETTO	survey status 1985
CNETTO	survey status 1986
$NETTO	survey status ...
UNETTO	survey status 2004
HSAMPREG	Region in which household lives (West or East Germany) 1991
ISAMPREG	Region in which household lives (West or East Germany) 1992
JSAMPREG	Region in which household lives (West or East Germany) 1993
$SAMPREG	Region in which household lives (West or East Germany) ...
USAMPREG	Region in which household lives (West or East Germany) 2004
APOP	Population Indicator 1984
BPOP	Population Indicator 1985
CPOP	Population Indicator 1986
$POP	Population Indicator ...
UPOP	Population Indicator 2004

2.1.2 Household-Level Longitudinal File: *HPFAD*

The *HPFAD* files are especially important for matching longitudinally household information. This is explained more in Chapter 4. For those interested in looking at East-West migration, the *HPFAD* has the variables `$SAMPREG` for waves H through U. As a household by design *always* stays in its original sample (`HSAMPLE`), even if the household physically moves from East to West or vice-versa, the *HPFAD* information will be very useful.

Table 2.3: List of variables in *HPFAD*

Variable Name	Meaning
`HHNR`	original household identifier (case) from wave 1
`HHNRAKT`	unique household identifier
`HSAMPLE`	sample identifier
`AHHNR`	household identifier 1984
`BHHNR`	household identifier 1985
`CHHNR`	household identifier 1986
`$HHNR`	household identifier ...
`UHHNR`	household identifier 2004
`AHNETTO`	survey status 1984
`BHNETTO`	survey status 1985
`CHNETTO`	survey status 1986
`$HNETTO`	survey status ...
`UHNETTO`	survey status 2004
`HSAMPREG`	Region in which household lives (West or East Germany) 1991
`ISAMPREG`	Region in which household lives (West or East Germany) 1992
`JSAMPREG`	Region in which household lives (West or East Germany) 1993
`$SAMPREG`	Region in which household lives (West or East Germany) ...
`USAMPREG`	Region in which household lives (West or East Germany) 2004
`AHPOP`	Population Indicator 1984
`BHPOP`	Population Indicator 1985
`CHPOP`	Population Indicator 1986
`$HPOP`	Population Indicator ...
`UHPOP`	Population Indicator 2004

2.2 Generated & Status Variables: *$PGEN, $HGEN*

There are some problems with using only the variables in the dataset that are directly likned to the original survey questions. Some of the SOEP information is gathered in the course of the first interview. Some information is asked only in separate questions for old and new households and individuals. In some cases old respondents are only asked about changes since last year's interview, while new respondents have to fill in their current status. Respondents in different subsamples might be asked the same information in different questions.

The use of "status variables" is intended to solve this problem. In all these cases, the originally collected information is stored in different variables. In order to minimize computing efforts for the user, the SOEP provides yearly status variables on individual and household level, which integrate this information in a common variable showing the current status. Thus, there is just a re-organization of already existing data, without any assumption or normative setting in the generating process.

In addition to the above mentioned status variables the SOEP provides some generated, partially assumption-based variables for households and individuals.
For example,

- status variable `$WOHNFL` in household file *$HGEN*:
  The size of the house/apartment (in square meters) in which a household lives is asked in the first interview as well as after each residential move. An old household remaining at the old address is asked only if there were changes in the house/apartment due to reconstruction, etc. Thus, the 1994 information for a given household might have been gathered in 1984 or in any year since then. The generated variable `$WOHNFL` carries on this old, but still valid, information until it has to be actualized due to a change. If there was a change, but the new information is missing, `$WOHNFL` is defined as missing (Code -1).

- generated variables `PARTZ$$` and `PARTNR$$` in *$PGEN*:

Each individual respondent is asked whether or not he/she lives with a partner. Variables used to describe such a relationship in the SOEP-data are marital status, changes in marital status since last year, the relationship to the head of the household, and information taken from the family and marriage biography. While the variable `PARTZ$$` defines the legal status of the partnership (legally married or cohabiting), the variable `PARTNR$$` contains the unique individual identifier of the partner. This variable allows researchers to merge individual characteristics of both persons in a partnership.

As seen in Table 2.4, with the data release 2002, job related information on industry and occupation is coded according to the NACE and ISCO88 classification schemes respectively. Prestige scores have been updated accordingly, based on the ISCO88 indicator, which is now available for all years.

As of wave 21 (2004), monthly labor income measures `LABGRO$$` (gross) and `LABNET$$` (net) are also avaialble as fully imputed values (in case of item-nonresponse) in Euro for all years.

Table 2.4: List of variables in the cross-sectional file *SPGEN*

Variable Name	Meaning	ID or Status or Generated
HHNR	original household identifier (case) from wave 1	ID
HHNRAKT	current household identifier	ID
SHHNR	current household identifier	ID
PERSNR	unique individual identifier	ID
ERWTYP02	employment status	G
ERLJOB02	working in the original job?	S
BETR02	size of employer	S
OEFFD02	public sector	S
AUSB02	educational requirements of job	S
PARTZ02	kind of relationship to partner	G
PARTNR02	unique individual identifier of partner	G
NATION02	nationality	S
SPSBIL	highest school degree received	S
SPBBIL01	highest occupational degree received	S
SPBBIL02	university degree	S
SPBBIL03	no occupational degree	S
SPSBILA	highest school degree received abroad (Sample B)	S
SPBBILA	highest occ. degree received abroad (Sample B)	S
SPSBILO	highest school degree (Sample C)	S
SPBBILO	highest occ. degree received (Sample C)	S
SFAMSTD	marital status	S
SBILZEIT	institutional years necessary to receive current degree of education	G
SERWZEIT	years with current employer	G
STATZEIT	average actual work hours / week	S
SVEBZEIT	contracted work hours / week	S
SUEBSTD	overtime last week	S/G
LFS02	labor force status	S/G
IS8802	ISCO88-4-digit	S
ISEI02	ISEI-Status88 according to Ganzeboom	S
MPS02	Magnitude Prestige Scale (based on KLAS94)	S
NACE02	NACE industry codes	S
SIOPS02	Treiman Standard Int. Occ. Prestige Score	S
EGP02	Erikson and Goldthorpe Class Category	S
KLAS02	Classification of occupation (Statistical Office)	S
AUTONO02	Autonomous Decision Making at Work	G
STIB02	Job Type and Level	G
ISCED02	Highest Completed Schooling ISCED-1997	G
CASMIN02	Highest Completed Schooling CASMIN	G
MONTH02	Month Of Interview	G
MODE02	Interview Method	G
LABGRO02	Monthly Gross Labor Market Income	G
LABNET02	Monthly Net Labor Market Income	G
IMPGRO02	Impute flag: Gross Labor Market Income	G
IMPNET02	Impute flag: Net Labor Market Income	G

Table 2.5: List of variables in the cross-sectional file *SHGEN*

Variable Name	Meaning	ID or Status or Generated
HHNR	original household identifier (case) from wave 1	ID
HHNRAKT	current household identifier	ID
SHHNR	current household identifier	ID
SEINZUG	year moved into house or apartment	S
SBAUJ	year of construction (classified)	S
SRENOV	degree to which dwelling needs repair	S
SWOHNFL	size of housing unit in square-meter	S
SWOHNR	no. of rooms in dwelling	S
SWGURT	evaluation of apartment-size	S
SAUS1	kitchen	S
SAUS2	bath, shower	S
SAUS3	toilet in dwelling	S
SAUS4	central heating	S
SAUS5	balcony, porch	S
SAUS6	cellar	S
SAUS7	garden	S
SAUS8	hot water, boiler	S
SAUS9	telephone	S
SEIGEN	owner-occupier or tenant	S
SERWERB	type of acquisition	S
SFOERD	support by public loans	S
SMIETE	monthly rent in DM	S
SNOMIET	don't have to pay rent	S
SMURT	evaluation of rent to be paid	S
SSOZIAL	social housing	S
SBILLIG	rent reduced by owner ?	S
SKOSTEN	monthly cost for hot water, heating	S
STYPHH1	household typology (1-digit)	G
STYPHH2	household typology (2-digit)	G
SMIETEG	generated gross cold rent in DM	G
SHEIZG	generated heating and warm water costs in DM	G
HMONTH02	Month Of Interview	G
HMODE02	Interview Method	G
HINC02	Household income in Euro	G
AHINC02	Household income in Euro (Adjusted for under-reporting)	G

2.2.1 Generated Schooling Variables

The variable `$BILZEIT` in the *$PGEN* files is the generated years of education, including schooling and occupational education. This variable description has been included directly in this handbook, as users have repeatedly asked questions concerning its definition and generation. Further description is found in Schwarze (1991) and Helberger (1988). The following statements describe the standard computation for West German years of schooling (including years of secondary occupational education).

As it can be seen the code is not very differentiated. For example, special schools for health care training and other are included in the *Fachschule* label. However, the present code is the most common in Germany when earnings functions based on human capital theory are estimated. The variable `$BILZEIT` is computed for all samples. The years-of-schooling mapping is based on the typical average number of years required to achieve a particular degree or certificate, e.g. 13 years for *Abitur*. In order to avoid under-counting years of schooling, `$BILZEIT` is set to missing for those persons who have incomplete schooling attainment information, i.e. missing values, even if valid occupational training information is available. For schooling as well as occupational training only the *highest* degree achieved is included in the calculation. This implies a range of valid values from 7 to 18 years for `$BILZEIT`.

```
** ********************************************************
** years of education= schooling + occupational_training ***
** schooling                                               *
**     no degree                        = 7   years        *
**     lower school degree              = 9   years        *
**     intermediary school              = 10  years        *
**     degree for a professional coll.  = 12  years        *
**     high school degree               = 13  years        *
**     other                            = 10  years        *
** additional occupational training (includes universities)*
**     apprenticeship                   = 1.5 years        *
**     technical schools (incl. health) = 2   years        *
**     civil servants apprenticeship    = 1.5 years        *
**     higher technical college         = 3   years        *
**     university degree                = 5   years        *
```

```
** ************************************************************
```

The component parts of the generated years-of-schooling variable are described below. All respondents are asked for their primary and secondary education when they fill out the individual questionnaire for the first time (blue version; since 1994 Biographical Questionnaire). The generated variable integrates the foreign version (`$PSBILA`) and the East German version (`$PSBILO`) into the West German categories. This information is repeated and updated for changes.

```
$PSBIL  'Primary and Secondary Education'
(1)'Hauptschulabschluss'
(2)'Realschulabschluss'
(3)'Fachhochschulreife'
(4)'Abitur'
(5)'Anderer Abschluss'
(6)'Ohne Abschluss verlassen'
(7)'Noch kein Abschluss'
```

In addition to variable `$PSBIL` primary and secondary education is given separately for East Germans. This variable is only updated until 1991 (as this is the last year, when an East German school degree could be achieved). However new respondents which join the sample later are still asked for their education degree in the former GDR. This information is included in `$SPBILO`.

```
$PSBILO 'Primary and Secondary Education'
(1)'Abschluss 8.Klasse, POS'
(2)'Abschluss 10.Klasse, EOS'
(3)'Abitur'
(4)'Anderer Abschluss'
(5)'Ohne Abschluss verlassen'
(6)'Noch kein Abschluss'
```

In addition to variable `$PSBIL` the last education degree from abroad is asked separately (mainly for people in sample B and D). This information is included in `$SPBILA`.

```
$PSBILA    'Primary and Secondary Education Abroad'
(1)'Pflichtschule o. Abschl.'
(2)'Pflichtschule m. Abschl.'
(3)'Weiterfuehrende Schule'
```

All respondents are asked for their vocational training when they fill out the individual questionnaire for the first time (blue version; since 1994 Biographical Questionnaire). The generated variable integrates the foreign version (`$PBBILA`) and the East German version (`$PBBILO`). For the creation of the variable the originally independent categories are combined so that they match with the question for changes in vocational training. This individual information is repeated and updated for changes every year.

```
$PBBIL01   'Vocational Training'
(1)'Lehre'
(2)'Berufsfachschule, Gesundheitswesen'
(3)'Schule Gesundheitswesen (bis 99)'
(4)'Fachschule, Meister'
(5)'Beamtenausbildung'
(6)'Sonstige Ausbildung'
```

In addition to $PBBIL01 the vocational training is given separately in the East German version. This variable includes updates until 1991 in East Germany (Sample C) and updates for new respondents who grew up in the former GDR. For all others, this information is simply repeated every year.

```
$PBBILO 'Vocational Training in East-Germany'
(1)'Berufsausb.,Facharbeiter'
(2)'Meisterabschluss'
(3)'Ingenieur- und Fachschulabschl.'
(4)'Sonstige Ausbildung'
```

The university degree is asked for all respondents in all samples when they enter the panel population. The categories (4-6) cover people, who have not achieved a West German occupational degree. This information is repeated and - if changed - updated every year.

```
$PBBIL02   'University Degree'
(1)'Fachhochschule'
(2)'Universitaet, TH'
(3)'Hochschule im Ausland'
(4)'Ingenieur-Fachschule (Ost)'
(5)'Hochschule (Ost)'
```

Together with the questions for vocational training ($PBBIL01) and university degree ($PBBIL02), all respondents in all samples are asked, whether they have not finished any degree, or are still in the educational system. If any information on vocational training is valid, variable $PBBIL03 has the missing code -2.

```
$PBBIL03   'No Degree in Vocational Training'
(1)'Kein Berufsabschluss'
(2)'Lehre'
(3)'Studium';
```

In addition to variable $PBBIL01 the original version of vocational training abroad is saved separately for all respondents every year.

```
$PBBILA    'Vocational Training Abroad'
(1)'Angelernt'
(2)'Betriebl. Ausbildung'
(3)'Berufsbild. Schule'
(4)'Hochschule'
(5)'Sonstiges'
```

The previously mentioned education variables are the result of carried over information over various previous waves, and updated whenever changes were reported to have taken place. Since 2001, the updates are based on the explicit survey of all respondents with respect to their highest educational attainment in 2000, starting with the variable QP103 onwards.

2.3 The CNEF and *$PEQUIV*

The Cross-National Equivalent File is created by Cornell University, in close cooperation with DIW-Berlin, ISER-Essex and StatsCan-Ottawa, consisting of variables from the German SOEP, American PSID, Canadian SLID and British BHPS, based on common definitions. The income variables are all annualized, meaning that the typical German SOEP variables asking about monthly income components have been transformed. The Equivalent File itself cannot be described here due to its complexity, but a listing of the generated variables is included here. The Equivalent File variable names are identical across datasets, adding to ease of use. The reader is referred to the standard Equivalent File documentation in Burkhauser, Butrica, Daly, and Lillard (2001) to further information (all used original variables names from the data sets are included with the algorithms). The codebooks are available at http://www.human.cornell.edu/che/PAM/Research/Centers-Programs. The file structure of the CNEF is such that for each wave (year) and each country, there is a separate file.

For ease of use, the German portion of the cross-national equivalent file has been included in the regular distribution of the SOEP data, both for the German and international distribution. In addition, the regular matching variable indicators `HHNR`, `HHNRAKT`, `$HHNR` and `PERSNR` have been added (in addition to the already existing equivalent file matching variables such as `X11101LL`). The German portion is found in the files *$PEQUIV*, available starting 1984 (wave A) onward. The sampled population includes adult respondents, adult non-respondents and children in households with an interview, corresponding to the SOEP population defined by `$NETTO` ≥ 1 and `$HNETTO` $= 1$.

For the first two waves of the East German sample, the CNEF information is missing. This is due to the different way of asking the information in 1990/91 and the different currency (the East-German Mark as opposed to the West-German DM). Missing income information due to item non-response in the SOEP-CNEF data is imputed based on the row-and-column imputation procedure suggested by Little and Su (1989) if longitudinal data is available, otherwise by adequate cross-sectional imputation techniques (e.g., wages are imputed based on Mincer-type regression models). The application of the Little and Su (1989) procedure to the SOEP is described in Butrica (1997). See Frick and Grabka (2003) for a more detailed description of incidence of item-non-response in SOEP and cross-sectional imputation procedures. Sample G has been include included in the CNEF and the *$PEQUIV*-files, however the corresponding SOEP standard weights for this subsample have been set to zero.

The reason for integrating the German portion of the CNEF into the standard SOEP data distribution is to make matching easier. The *$PEQUIV* files

are also included in SOEPINFO, which will allow automatic matching in SAS, SPSS and Stata. Starting in 2004, the *$PEQUIV* files also contain separate imputed income components with corresponding impute flag indicators. Thus not only are the total income indicators imputed, but also all individual inputs and components, such as for individual labor income `I11110$$`, which includes first and second job employment, year-end bonuses, etc.

Table 2.6: List of variables in *UPEQUIV*

Variable	Meaning
HHNR	Original Household Number
HHNRAKT	Current Wave HH Number (=UHHNR)
UHHNR	Current Wave HH Number (=UHHNR)
PERSNR	Never Changing Person ID
X11101LL	Person Identification Number
D11102LL	Gender of Individual
X11104LL	Subsample Identifier
X1110204	HH Identification Number
X1110304	Individual in HH at Survey
X1110504	Individual responded to Survey
D1110104	Age of Individual
D1110304	Race of HH Head
D1110404	Marital Status of Individual
D1110504	Relationship to HH Head
D1110604	Number of Persons in HH
D1110704	Number of Children in HH
D1110804	Education With Respect to High School
D1110904	Number of Years of Education
D1111004	Disability Status of Individual
D1111104	Satisfaction With Health
E1110104	Annual Work Hours of Individual
E1110204	Employment Status of Individual
E1110304	Employment Level of Individual
E1110404	Primary Activity of Individual
E1110504	Occupation of Individual
E1110604	1 Digit Industry Code of Individual
E1110704	2 Digit Industry Code of Individual
I1110104	HH Pre-Government Income
I1110204	HH Post-Government Income
I1110304	HH Labor Income
I1110404	HH Income From Asset Flows
I1110504	HH Imputed Rent
I1110604	HH Private Transfers
I1110704	HH Public Transfers
I1110804	HH Social Security Pensions
I1110904	Total HH Taxes
I1111004	Individual Labor Earnings
I1111104	HH Federal Taxes
I1111204	HH Social Security Taxes
I1111304	HH Post-Government Income (TAXSIM)
I1111404	Total HH Taxes (TAXSIM)
I1111504	HH State Taxes (TAXSIM)
I1111604	HH Federal Taxes (TAXSIM)
I1111704	HH Private Retirement Income
I1120104	Impute HH Pre-Government Income
I1120204	Impute HH Post-Government Income
I1120304	Impute HH Labour Income
I1120404	Impute HH Income From Asset Flows
I1120604	Impute HH Private Transfers
I1120704	Impute HH Public Transfers

Table 2.7: List of variables in *UPEQUIV* Cont'd

Variable	Meaning
I1120804	Impute HH Social Security Pensions
I1121004	Impute Individual Labor Earnings
I1121704	Impute HH Private Retirement Income
W1110104	X-Sectional Weight - Respondent Individual
W1110204	HH Weight
W1110304	Longitudinal Weight - Respondent Individual
W1110504	Individual Weight - Immigrant Sample
W1110604	HH Weight - Immigrant Sample
W1110704	X-Sectional Weight - Enumerated Individual
W1110804	Longitudinal Weight - Enumerated Individual
W1110904	Population Factor for w11103
W1111004	Population Factor for w11107
W1111104	Population Factor for w11108
H1110104	Number of hh members age 0-14
H1110204	Number of hh members age 15-18
H1110304	Number of hh members age 0-1
H1110404	Number of hh members age 2-4
H1110504	Number of hh members age 5-7
H1110604	Number of hh members age 8-10
H1110704	Number of hh members age 11-12
H1110804	Number of hh members age 13-15
H1110904	Number of hh members age 16-18
H1111004	No. hh members 19 and above or 16-18,ind
H1111204	Indicator-wife in HH
Y1110104	Consumer Price Index
L1110104	State of Residence
L1110204	Region
I1111804	Household Windfall Income
I1121804	Impute Household Windfall Income
H1111104	Indicator - Head in HH
I1120504	Impute HH Imputed Rental Value
I1120904	Impute Total HH Taxes
W1110404	Population Factor for W11101
E1120104	Impute Annual Work Hours of Individual
IJOB104	Wages,Salary from main job
IJOB204	Income from secondary employment
ISELF04	Income from self-employment
IOLDY04	old-age,disability and civil serv. pensions
IWIDY04	widows and/or orphans pension
IUNBY04	Unemployment benefit
IUNAY04	Unemployment assistance
ISUBY04	Subsistence allowance
IERET04	Old-age transition benefit
IMATY04	Maternity benefit
ISTUY04	Student grants
IMILT04	Military/community service pay
IALIM04	Alimony
IELSE04	Private Transfers received
ICOMP04	Company pension (surviving dependants c.p.)
IPRVP04	Private pension (old-age,accid.,disability)

Table 2.8: List of variables in *UPEQUIV* Cont'd

Variable	Meaning
I13LY04	13th monthly salary
I14LY04	14th monthly salary
IXMAS04	Christmas bonus
IHOLY04	Vacation bonus
IGRAY04	Profit-sharing
IOTHY04	Other bonuses
IGRV104	Retirement pay: stat. pension insurance
IGRV204	Widows pension: stat pension insurance
RENTY04	Income from rental and leasing
OPERY04	Operation, maintenance costs
DIVDY04	Interest, dividend income
CHSPT04	Child allowance
HOUSE04	Housing benefit
NURSH04	Compulsory long term care insurance
SUBST04	Social assistance(living expenses etc)
SPHLP04	Social assistance f. spec. circumstances
HSUP04	Housing support for owner-occupiers
FJOB104	Imp.flag:Wages,Salary from main job
FJOB204	Imp.flag:Income from secondary job
FSELF04	Imp.flag:Income from self-employment
FUNBY04	Imp.flag:Unemployment benefit
FOLDY04	Imp.flag:old-age,civil serv. pensions
FWIDY04	Imp.flag:widows/orphans pension
FUNAY04	Imp.flag:Unemployment assistance
FSUBY04	Imp.flag:Subsistence allowance
FERET04	Imp.flag:Old-age transition benefit
FMATY04	Imp.flag:Maternity benefit
FSTUY04	Imp.flag:Student grants
FMILT04	Imp.flag:Military/community service pay
FALIM04	Imp.flag:Alimony
FELSE04	Imp.flag:Private Transfers received
FCOMP04	Imp.flag:Company pension
FPRVP04	Imp.flag:Private pension(old-age,accid.)
F13LY04	Imp.flag:13th monthly salary
F14LY04	Imp.flag:14th monthly salary
FXMAS04	Imp.flag:Christmas bonus
FHOLY04	Imp.flag:Vacation bonus
FGRAY04	Imp.flag:Profit-sharing
FOTHY04	Imp.flag:Other bonuses
FGRV104	Imp.flag:retirement pay from stat.insurance
FGRV204	Imp.flag:widows pension from stat.insurance
FRENTY04	Imp.flag:Income from rental and leasing
FOPERY04	Imp.flag:Operation, maintenance costs
FDIVDY04	Imp.flag:Interest, dividend income
FCHSPT04	Imp.flag:Child allowance
FHOUSE04	Imp.flag:Housing benefit
FNURSH04	Imp.flag:Compuls. long term care insurance
FSUBST04	Imp.flag:Social assist.(living expenses ..)
FSPHLP04	Imp.flag:Soc. assist. for spec. circumstan.
FHSUP04	Imp.flag:Housing support f. owner-occupiers

Table 2.9: List of variables in *UPEQUIV* Cont'd

Variable	Meaning
ISMP104	Social miners insurance pension
ICIV104	Civil servant pension
IWAR104	War victim pension
IAGR104	Farmer Pension
IGUV104	Statutory accident insurance
IVBL104	Supplementary benefits for civil servants
ICOM104	Company pension
IPRV104	Private pension
ISON104	Other pension
ISMP204	Widows social miners insurance pension
ICIV204	Widows civil servant pension
IWAR204	Widows war victim pension
IAGR204	Widows farmer Pension
IGUV204	Widows statutory accident insurance
IVBL204	Widows supplement. benefits(civil servants)
ICOM204	Widows company pension
ISON204	Other widows pension
IPRV204	Widows private pension
M1110104	Overnight hosp stay
M1110204	Inpatient nights in hosp
M1110304	Work accident required treatment
M1110404	Frequency of sport or exercise
M1110504	Have had stroke
M1110604	High blood pressure/circulation problems
M1110704	Have or had diabetes
M1110804	Have or had cancer
M1110904	Psychiatric problems
M1111004	Arthritis
M1111104	Angina or heart condition
M1111204	Difficulties breathing
M1111304	Have trouble climbing stairs
M1111404	Need help or have difficulty bathing alone
M1111504	Dressing difficult alone
M1111604	Difficulty/need help getting in/out bed
M1111704	Need help with shopping
M1111804	Walk 10+ min alone difficult
M1111904	Housework difficult alone
M1112004	Health limits kneeling
M1112104	Health limits vigorous activities
M1112204	Body height
M1112304	Body weight
M1112404	Disability Status of Individual
M1112504	Satisfaction with Health
M1112604	Current Self-Rated Health Status
M1112704	Number of annual doctor visits

2.4 Temporary Drop-Outs: *$PLUECKE*

Temporary drop-outs ("gaps") can cause problems for longitudinal analyses. This is especially true for the employment and income data stored in the files *$PKAL* and the spell-oriented files *ARTKALEN* and *EINKALEN*. That is why the SOEP tries to fill in at least some of the central missing information. Persons who take part in the survey after such a temporary unit non-response are asked to complete:

- the normal (individual) questionnaire for the current wave, plus
- a small questionnaire covering information on the year previous to which the drop-out occurred. This covers questions on
  - job-related changes,
  - calendar of occupation and income (is taken into account in the spell-data),
  - education and qualification.

Note:
This additional data is stored in the cross-sectional files *$PLUECKE*. Persons with a completed "gap"-questionnaire are marked in the corresponding `$NETTO` variable in *PPFAD* with the code 4.

2.5 Individual Drop-Outs: *YPBRUTTO*

The wave-specific cross-section files *$PBRUTTO* encompass all individuals currently living in SOEP-households at a given point of time. These include respondents, children, and persons who refused to answer (unit-non-response).

In contrast to these files, *YPBRUTTO* cumulates across all waves all individuals who left the household they lived in last year or even the survey due to:

- moving to another household within the survey territory,
- moving abroad,
- death.

In contrast to the first alternative, moving abroad is not necessarily a final exit from the survey. For example the foreigners sample (B) encompasses persons temporarily returning to their home country, e.g. for military service reasons. When they return into the German household, they will be listed

according to their unique individual identifier PERSNR. The third group is a pooled set of household changing people.

Since the second and third alternatives might lead to a repeated documentation in *YPBRUTTO*, there is a second variable necessary to uniquely identify a record in this file. The variable ERHEBJ indicates the year in which the person left the household. Another major variable of interest in this file is YPZUG, indicating the reason for the drop-out.

2.6 Persons Needing Care (Invalids): *PFLEGE*

Since wave B (1985) the SOEP household questionnaire includes questions on household members in need of care. In order to support analyses on an individual level, this information has been restructured and stored in the cumulative file *PFLEGE*. For any person mentioned in the household questionnaire as needing care in a given year, now a single record in the file *PFLEGE* is provided. Since a person might show up in *PFLEGE* more than once, a second identifier ERHEBJ indicates the year in which this person has been mentioned.

Additional variables describe the intensity of care necessary (variables MAXGRAD and MULTGRAD) as well as which person provides the care (variable WERPFLGT, available for the years 1985 to 1990, only).

```
HHNR            Household identifier of first wave
PERSNR          Unique individual identifier
ERHEBJ          Survey Year

MAXGRAD         Maximum of care intensity
                (1=lowest ... 5=highest intensity)

MULTGRAD        Multiple responses for different categories of care needs
                (1. digit = lowest ... 5. digit = highest degree)

WERPLGT         Person(s) providing care
                multiple responses possible:
                        1. digit = community nurse / social worker
                        2. digit = friends
                        3. digit = neighbors
                        4. digit = relatives in the household
                        5. digit = relatives outside the household
                        6. digit = can be anybody
```

Table 2.10: Selected observations in *PFLEGE*

HHNR	PERSNR	ERHEBJ	MAXGRAD	MULTGRAD	WERPFLGT
43	401	1985	1	10000	100000
43	1201	1985	1	10000	10
213	2102	1993	1	10000	-2
213	2102	1994	1	10000	-2
213	2102	1995	5	10001	-2
213	2102	1996	1	10000	-2
779	7702	1991	-1	-1	-2
779	7702	1993	1	10000	-2
779	7702	1994	1	10000	-2
779	7702	1995	2	11000	-2
779	7702	1996	2	11000	-2
779	7702	1997	4	11010	-2

2.7 Social Assistance Spells: *SOZKALEN*

The file *SOZKALEN* provides spell data on receiving social assistance of households, defining begin, end, and censoring status of any period of receiving 3 different types of assistance (see variable `SPELLTYP`).

This file is set up, using information from the calendar, asked for the previous year (asked for the years 1992-2000). Thus, it contains information on a monthly basis, beginning with January of the year preceding the first interview and ending with December of the year preceding the latest interview (as of 2000: month 97 = January 1991, ... , month 204 = December 1999).

The file includes only households with at least one spell of social assistance, but since it covers the entire period in which these households took part in the survey, there are also spells defining periods without receiving social assistance.

```
Variable SPELLTYP

(1)   continuous living assistance
      (Laufende Hilfe zum Lebensunterhalt HLU)
(2)   assistance for special circumstances
      (Hilfe in besonderen Lebenslagen HbL)
(3)   one-time living assistance
      (einmalige Hilfe zum Lebensunterhalt) or item-non-reponse
(4)   no social assistance received
(99)  unit-non-response


Variable ZENSOR (R=right; L=left)
  (1)  uncensored
  (2)  R-censored
  (3)  R-(KA)-censored
  (4)  L-censored
  (5)  L-R-censored
  (6)  L-R-(KA)-censored
  (7)  L-(KA)-censored
  (8)  L-(KA)-R-censored
  (9)  L-(KA)-R-(KA)-censored
 (-2)  does not apply (Spelltyp 99 only)
(KA=censored because of item- or unit-non-response)
```

Table 2.11: Selected Observations in *SOZKALEN*

HHNR	HHNRAKT	SPELLNR	SPELLTYP	BEGIN	END	ZENSOR
35	35	1	1	108	108	2
35	35	2	4	97	107	4
167	167	1	4	97	108	6
167	167	2	4	121	156	8
167	167	3	99	109	120	-2
272	272	1	4	97	120	6
272	272	2	4	133	156	8
272	272	3	99	121	132	-2
280	280	1	3	97	108	4
280	280	2	4	109	156	2
310	310	1	1	97	144	5
400	400	1	4	97	108	6
400	400	3	99	109	120	-2

2.8 Regional Information

The SOEP provides several types of regional information. There are three different groups of regional variables, which are different in respect to the data protection procedures which are required to use them. The different groups cover the following variables:

Group 1: Federal State (`$BULA`), variables describing a household's neighborhood and infrastructure variables (stores, banks, doctors, day care centers, schools etc.)[2]: These variables are available to all users, who have signed the standard data protection contract. These variables are part of the usual yearly data release (SOEP-CD-Rom).

Group 2: Community Type (Boustedt), Community Size (Political), Community Type (BIK): These variables are part of the standard data release in Germany as well, but some additional data protection procedures have to be fulfilled. This group of variables is provided in the file *GGKBOU*, which can be accessed with an additional password only.

Group 3: Regional Units: The *Raumordnungsregionen* are a regional classification, below the level of the federal states and above the level of the counties. The classification of the *Raumordnungsregionen* changed in 1996, which is a remarkable problem for longitudinal analysis which cover a time period including the year 1996. For a short description of this change in the classification system, see Böltken (1996). These variables have to be ordered separately and are not part of the SOEP-CD Rom. To use these variables an extra data protection contract has to be signed. For users interested in analyses on the level of the Raumordnungsregionen there is a short memo available (please order the SOEP-Geocode manual via the SOEP hotline, soepmail@diw.de).

Users outside Germany can use the regional information described in Group 2 and Group 3, if they work *on site* in the data center at the DIW Berlin or at Cornell University.

[2]The infrastructure questions are not asked in every year, so far they have only been asked in 1986, 1994, 1999 and 2004.

Chapter 3

Calendar and Biography Extensions

by Joachim R. Frick and John P. Haisken-DeNew

Since 1984, representative biography information has been available in the standard distribution consisting of such subtopics as: individual job biography since the age of 15, marriage and youth biography, entrance into the job market, individual social background and immigration information. The purpose of this data is to provide firstly important background information for many different analyses, e.g. birth information in estimating labor supply for women, and secondly also information for use in stand-alone analysis, e.g. job histories or intergenerational educational attainment.

The following files are part of the standard data distribution: *BIOMARSM*, *BIOMARSY* concerning marital status, *BIOBIRTH* concerning birth information from women, *BIOBRTHM* containing birth information from male respondents, *BIOTWIN* containing person identifiers for twins and triplets, *BIOPAREN* concerning parental information and time independent information on social origin, *BIOIMMIG* concerning immigration specific information, *BIOJOB* concerning first and last job, *BIOYOUTH* concerning adolescence, *BIOSOC* concerning primary and secondary socialization, *BIOCHILD* with info on newborns, *BIORESID* concerning second residence.

3.1 Employment and Income Calendar Files: *$PKAL*

Ever since its start in 1984, the SOEP contained a calendar section asking about employment status and sources of income received as of January through December of the previous year.

The income section was changed in 1995, when a variable was introduced counting the number of months a given type of income was received, instead of asking the same information for each single month of the previous year.

Up to 1997 this information was available to the user

- In spell form in the files *ARTKALEN* and *EINKALEN* using information from the files *$PLUECKE* as well; due to the changes in the income section *EINKALEN* covers the period January 1983 to December 1994, only.
- As string variables in the cross-sectional files *$PGEN*; variables vary in length (12 digits and 24 digits).
- As a matrix for 12 months by up to 11 different employment status and types of income, respectively in the files *APKAL* to *GPKAL*; thus for the first 7 waves of SOEP, only.

There is one major exception for the time period covered in the calendar: In contrast to the West German subsamples (A and B) which run from January through December, in subsample C (East Germans) the respective calendars ran from July 1989 to June 1990 (in wave 1, 1990) and from July 1990 to March 1991 (in wave 2, 1991). Since 1992 the calendar covers the same time period for all subsamples including subsample D (immigrants), which started in 1994/95.

Starting with the data distribution of 1998 the SOEP provides an improved set of updated *$PKAL* files with two additional files specific for subsample C (due to the above mentioned difference in the period covered).

The following types of variables are available for employment status and type of income received, respectively:

- A variable indicating whether a given employment status or a type of income was valid for at least one month of the entire time period under consideration (yes/no)
- A variable counting the number of months (up to 12) in a given employment status or with having received of a given type of income
- Twelve two-digit dummy variables (one for each single month of the previous year) indicating whether a given employment status or type of income was valid for that specific month (January through December)
- A 24-digit string variable for each year concatenating the information given in the twelve month specific variables
- The average monthly amount of income received, for only those months that the income source was received (this type of variable is by definition not available for employment status information).

The following cross-sectional *$PKAL* files are available:

```
File-Name          Subsample

  APKAL            A B - - - - -
  BPKAL            A B - - - - -
  CPKAL            A B - - - - -
```

```
DPKAL           A B - - - - -
EPKAL           A B - - - - -
FPKAL           A B - - - - -
GPKAL           A B - - - - -
GPKALOST        - - C - - - -
HPKAL           A B - - - - -
HPKALOST        - - C - - - -
IPKAL           A B C - - - -
JPKAL           A B C - - - -
KPKAL           A B C D - - -
LPKAL           A B C D - - -
MPKAL           A B C D - - -
NPKAL           A B C D - - -
OPKAL           A B C D E - -
PPKAL           A B C D E - -
QPKAL           A B C D E F -
RPKAL           A B C D E F -
SPKAL           A B C D E F G
$PKAL           A B C D E F G
UPKAL           A B C D E F G
```

Availability of calendar information over time Variable types as described above:

```
(1) Information is valid for at least one month (yes/no)
(2) Number of months
(3) Twelve single month variables
(4) String variable
(5) Average monthly amount
```

	Employment				Income				
File	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)	(5)
APKAL	x	x	x	x	x	x	x	x	x
BPKAL	x	x	x	x	x	x	x	x	x
CPKAL	x	x	x	x	x	x	x	x	x
DPKAL	x	x	x	x	x	x	x	x	x
EPKAL	x	x	x	x	x	x	x	x	x
FPKAL	x	x	x	x	x	x	x	x	x
GPKAL	x	x	x	x	x	x	x	x	x
GPKALOST*	x	x	x	x	x	x	-	-	x
HPKAL	x	x	x	x	x	x	x	x	x
HPKALOST**	x	x	x	x	x	x	x	x	x
IPKAL	x	x	x	x	x	x	x	x	x
JPKAL	x	x	x	x	x	x	x	x	x
KPKAL	x	x	x	x	x	x	x	x	x
LPKAL	x	x	x	x	x	x	-	-	x
MPKAL	x	x	x	x	x	x	-	-	x
$PKAL	x	x	x	x	x	x	-	-	x
UPKAL	x	x	x	x	x	x	-	-	x

```
*   Time period covered is July 1989 through June 1990.
**  Time period covered is July 1990 through March 1991 (9 months, only).
```

Variable Names are defined as follows:

```
1. digit        Wave specifier (A for 1984, B for 1985, etc.)

2. digit        P = personal information (always P)

3. digit        Differentiating Employment and Income Data
                1       Employment Status
                2       Type of Income received

4. digit        Exact specification of Employment Status/Income

                if 3. digit = 1 (Employment Status)
```

```
                A       Full-time employment
                B       Part-time employment
                C       Vocational training
                D       Unemployed
                E       Retired
                F       Maternity leave
                G       School, college
                H       Military or civil service
                I       Housewife/housekeeping
                J       Other
                K       Short work hours
                L       Second job

                if 3. digit = 2 (Type of Income received)
                A        Employment income
                B        Self-employment income
                C        Second job income
                D        Old age pension
                E        Widows/ers pension
                F        Unemployment benefits
                G        Unemployment relief
                H        Subsistence allowance
                J        Maternity benefits
                K        Student aid / stipend
                M        Payments from outside household
                N        None of these

5.-6. digit
                if 3. digit = 1 (Employment status)
                01      employment status is valid for a given month (Y/N)
                02      number of months in a given employment status

                if 3. digit = 2 (Type of Income received)
                01      income information is valid for a given month (Y/N)
                02      number of months income source was received
                03      average amount received per month

5.-7. digit     Month specific variables
                01     January
                02     February
                ..
                11     November
                12     December

last digit      0 (the letter "O", as in Ostdeutschland)
                  = specific for Sample C (East Germans)
                  in files GPKALOST and HPKALOST
```

Examples for Variable Names concerning Employment Status

- Full time employment in 1987 (yes, no)

```
EP1A01      'Vollzeit Erwerbstaetig im Jahr  1987'
```

- Number of months with full time employment in 1987

```
EP1A02      'Vollzeit Erwerbstaetig N-Monate 1987'
```

- Full time employment in each single month 1987

```
EP1A001     'Vollzeit Erwerbstaetig Jan 1987'
EP1A002     'Vollzeit Erwerbstaetig Feb 1987'
EP1A003     'Vollzeit Erwerbstaetig Mar 1987'
EP1A004     'Vollzeit Erwerbstaetig Apr 1987'
EP1A005     'Vollzeit Erwerbstaetig Mai 1987'
EP1A006     'Vollzeit Erwerbstaetig Jun 1987'
EP1A007     'Vollzeit Erwerbstaetig Jul 1987'
```

```
EP1A008     'Vollzeit Erwerbstaetig Aug 1987'
EP1A009     'Vollzeit Erwerbstaetig Sep 1987'
EP1A010     'Vollzeit Erwerbstaetig Okt 1987'
EP1A011     'Vollzeit Erwerbstaetig Nov 1987'
EP1A012     'Vollzeit Erwerbstaetig Dez 1987'
```

- Full-time employment 1987 (24-digit String Variable)

```
EP1A        'Vollzeit Erwerbstaetig Jan-Dez 1987'
```

- Sample C specific variable for full-time employment Dec. 1989

```
GP1A012O    'Vollzeit Erwerbstaetig Dez 1989'
```

Examples for Variable Names concerning Type of Income

- Employment income in 1987 (yes, no)

```
EP2A01      'Lohn als Arbeitnehmer    Bezogen 1987'
```

- Number of months with employment income in 1987

```
EP2A02      'Lohn als Arbeitnehmer    Monate  1987'
```

- Average monthly employment income received in 1987

```
EP2A03      'Lohn als Arbeitnehmer    Betrag  1987'
```

- Employment income in each single month 1987

```
EP2A001     'Lohn als Arbeitnehmer    Jan 1987'
EP2A002     'Lohn als Arbeitnehmer    Feb 1987'
EP2A003     'Lohn als Arbeitnehmer    Mar 1987'
EP2A004     'Lohn als Arbeitnehmer    Apr 1987'
EP2A005     'Lohn als Arbeitnehmer    Mai 1987'
EP2A006     'Lohn als Arbeitnehmer    Jun 1987'
EP2A007     'Lohn als Arbeitnehmer    Jul 1987'
EP2A008     'Lohn als Arbeitnehmer    Aug 1987'
EP2A009     'Lohn als Arbeitnehmer    Sep 1987'
EP2A010     'Lohn als Arbeitnehmer    Okt 1987'
EP2A011     'Lohn als Arbeitnehmer    Nov 1987'
EP2A012     'Lohn als Arbeitnehmer    Dez 1987'
```

- Employment income 1987 (24-digit String Variable)

```
EP2A        'Lohn als Arbeitnehmer    Jan-Dez 1987'
```

- Sample C specific variable for employment income July 1990 through March 1991

```
HP2AO       'Lohn als Arbeitnehmer    Jul90-Mar91'
```

Coding of Employment Status and Type of Income received in the *$PKAL* files and in the spell data files (*ARTKALEN* and *EINKALEN*).

As mentioned above, most of the information from *$PKAL* is also available in the spell files *ARTKALEN* (employment status section) and *EINKALEN* (income section; running from January 1983 through December 1994), where the variable `SPELLTYP` differentiates the types of income and employment status.

- (1) Employment Status

$PKAL	ARTKALEN SPELLTYP	Meaning
A	1	Full-time employment
B	3	Part-time employment
C	4	Vocational training
D	5	Unemployed
E	6	Retired
F	7	Maternity leave
G	8	School, college
H	9	Military or civil service
I	10	Housewife/housekeeping
J	12	Other
K	2	Short work hours
L	11	Second job

- (2) Type of Income Received

$PKAL	EINKALEN SPELLTYP	Meaning
A	1	Employment income
B	2	Self-employment income
C	3	Second job income
D	4	Old age pension
E	5	Widows/ers pension
F	8	Unemployment benefits
G	9	Unemployment relief
H	10	Subsistence allowance
J	7	Maternity benefits
K	6	Student aid / stipend
M	11	Payments from outside household
N	12	None of these

Table 3.1: Selected Observations in *EPKAL*

(selected variables on full-time employment and employment income)

HHNR HHNR	HHNRAKT HHNRAKT	PERSNR PERSNR	EP1A01 EP2A01	EP1A02 EP2A02	EP2A03	EP1A001 ... EP1A012 EP2A001 ... EP2A012	EP1A EP2A
19	19	101	1	12		1 1 1 1 1 1 1 1 1 1 1 1	01010101010101010101
19	19	101	1	12	-1	1 1 1 1 1 1 1 1 1 1 1 1	01010101010101010101
19	19	102	2	-2		-2-2-2-2-2-2-2-2-2-2-2-2	00000000000000000000
19	19	102	2	-2	-2	-2-2-2-2-2-2-2-2-2-2-2-2	00000000000000000000
27	27	201	2	-2		-2-2-2-2-2-2-2-2-2-2-2-2	00000000000000000000
27	27	201	2	-2	-2	-2-2-2-2-2-2-2-2-2-2-2-2	00000000000000000000
27	60313	203	2	-2		-2-2-2-2-2-2-2-2-2-2-2-2	00000000000000000000
27	60313	203	2	-2	-2	-2-2-2-2-2-2-2-2-2-2-2-2	00000000000000000000
35	35	301	1	12		1 1 1 1 1 1 1 1 1 1 1 1	01010101010101010101
35	35	301	1	12	2000	1 1 1 1 1 1 1 1 1 1 1 1	01010101010101010101
35	35	302	1	9		1 1 1 1 1 1 1 1 1-2-2-2	01010101010101010000
35	35	302	1	9	2209	1 1 1 1 1 1 1 1 1-2-2-2	01010101010101010000
159	159	1501	1	8		-2-2-2-2 1 1 1 1 1 1 1 1	00000000010101010101
159	159	1501	1	8	760	-2-2-2-2 1 1 1 1 1 1 1 1	00000000010101010101

3.2 Introduction to Biography Data

SOEP's representative biography information available in the standard distribution consisting of such subtopics as: individual job biography since the age of 15, marriage and youth biography, entrance into the job market, individual social background and immigration information. The purpose of this data is to provide firstly important background information for many different analyses, e.g. birth information in estimating labor supply for women, and secondly also information for use in stand-alone analysis, e.g. job histories or intergenerational educational attainment.

Every person (starting with 16 years of age) answering a SOEP questionnaire, responds normally once to a set of retrospective biography questions. This was the practice for persons who have been participating in the SOEP since 1984, for the first 3 waves (1984-1986). Since 1988, all biography information, including job history, marriage, labor market entrance, and social background, has been collected with varying instruments and questionnaires. However in the span 1984-1986, due to unavoidable panel mortality, moving abroad etc., this lead to some complications, such that components of the biography information were collected in the years that persons had indeed participated, but was missing for the years in which persons had dropped out. Further, due to budgetary reasons, it was not possible to ask the complete biographical questions for the first wave of respondents from East Germany. In 1990 the first questionnaire was given to East Germans, and only in 1992 was the biography questionnaire applied. This lead to completely missing information for some persons: some had dropped out before 1992, and others had completely refused to answer the biography questions in 1992. Due to technical reasons, the biography information was not collected from individuals who were 16 and 17 in their first year of participation. It was assumed for this sample, that the little biography information (e.g. marriage and family biography, job history since the age of 15, and social background) could be reconstructed using person-level information from the current sample.

The following is a chronological outline of the many changes in connection with collecting biography relevant information for the period since 1984. The data collection differences, with respect to "timing" (i.e. when the questionnaire was given), the "instruments" used, and "extent" (i.e. the extent of biographical information asked), between and within the individual subsamples will be distinguished.

- In 1984, the yearly questionnaire focus was on the job history in the calendar form (from age 15 to the current age, up to a maximum of age 65) for the subsamples A and B. The calendar was a matrix, such that one column represented one year, with nine rows, one row for each of the

different job status states: still in school, job training, military or social service, full-time employed, unemployed, house-husband or housewife, retired and "other" (see Question 62 in the standard person questionnaire).

- In 1985, the yearly questionnaire focus for subsamples A and B was on retrospective marriage and family biography information (Questions 81-88 in the wave 2 "blue" and "green" standard person questionnaires). Among others, information was collected on number of children born (answered only by females) and if and when these children had left the household (i.e. "moved out"). Marriage start, current status, and possible reason for marriage dissolution for up to 3 marriages was asked. Persons were asked in which city/town they has spent their childhood, and when they had "left home".

- In 1986, the questionnaire focus was directed to biography data dealing with social background and job market entrance (Questions 10-13 and 80-87 in the "blue" and "green" standard questionnaires in wave 3) for subsamples A and B. For each person, the following information was asked: parents' year of birth and possible year of death, parents' schooling and job training, parents' current job status. For job market entrance information, the following was asked: age at first entry into the job market, and type of occupation at first entry, and also ages at any possible occupation changes.

- In 1987 there was no biographical information collected.

- In 1988 , the complete set of biography questions was asked of the first-time respondents in the "blue" person questionnaires. For those "newcomers" (starting in 1985) to the SOEP, who had missed entire or partial questions on biography, then this information was explicitly collected again. Only starting in 1988 was it possible to construct complete information concerning all three biographical topics for SOEP respondents, assuming they were still in the SOEP. However, children in the household, who had completed 16 years of age, were excluded from the second explicit questionnaire, as it could not be appropriately applied in this case. There were also difficulties to determine the exact minimum age of the persons in the second explicit questionnaire.

- In 1989 and 1990, the 1988 questionnaire form for the biographical indicators was maintained.

- Starting 1991, persons from the subsample A (West Germans) and subsample B (foreigners) answered biography questions in a separate questionnaire booklet.

- In 1990, subsample C (East Germans) was collected for the first time. In 1992, an additional biography questionnaire had been developed for the East Germans to correspond to their different terminology concerning occupation, education degrees, and biographies, and was asked of all subsample C persons. All three biography areas of concentration were asked in a separate retrospective questionnaire "Biography". It was identical in structure and content to the West German version, with only small exceptions. Some categories were extended or modified (e.g. corresponding to occupational standing, or learned job). Starting 1993, all those East German "newcomers" to the panel, this extensive set of biography questions was asked.

- In 1994, a separate biography questionnaire "Curriculum Vitae" was established for all 4 subsamples (A, B, C, D1/D2). Some questions were changed marginally, and some new questions were added, which were asked in some cases of only certain subsamples and not others. Logically, questions pertaining to immigration were asked only of subsample D.

- Since 1996, the biography questionnaire "CV" has been fully integrated for all subsamples (A-D). All questions are asked of all subsamples.

- In 1999, biography information was asked of Sample E respondents for the first time.

- In 2000, biography information was asked of all respondents aged 16 or 17, in a Youth "special focus" questionnaire *in place of* the standard Biography questionnaire for the first time, i.e. staring with birth cohort 1984. Of course, new entrants into households, aged 18 and over, would then still fill in a *regular* one-time Biography questionnaire.

- In 2001, the Biography Questionnaire "Lebenslauf" ('life history') was further expanded and now captures also more questions on school, i.e. marks, and activities during childhood. It was asked of Sample F respondents for the first time, using both questionnaires as described above: the "youth" instrument for those aged 16-17 and the regular version for those aged 18 and over.

- 2002: A new sample G is drawn, which is only targeted at high-income households, i.e. households with a monthly net household income of more than 7,500 DM (about 3,850 EUR).

- 2003: Persons from sample G answered the Biography Questionnaire for the "first" time. The new questionnaire "Mother and Child" was given to mothers of newborns (all samples), starting with the birth cohort 2002/03.

- 2004: The Biography Questionnaire was slightly expanded with questions concerning the "numbers of brothers and sisters" and "region a person lived in prior to reunification (East Germany, West Germany or Abroad)". The question on siblings is also asked in the Youth Questionnaire.

- Future prospects: With the questionnaire for new born children (see *BIOCHILD*) which has been developed and directed to their mothers in 2003 for the first time, the SOEP has started to survey the development of children from the very beginning of their life. There will be follow-up interviews to collect data about these children at specific ages which typically are associated with relevant decisions for their further individual development (e.g. at ages 3, 6 and 12), allowing for cohort analyses.

Especially with respect to the period 1984-1996, there were some problems in appropriately restructuring the various sources of data into wave-independent information:

- The biography questions asked were unfortunately not always consistent over time.
- The target population was not always consistent.
- The timing of biography questions with respect to the first-time response varied.
- The positioning of the biography questions within the questionnaire varied.

In principle, there are two broad classes of biography information in the biography data files, time-invariant and time-variant information. Time-invariant information include such variables as: year of first migration to Germany, year of entrance into job market, city/town growing up. Time-variant variables on the other hand include: marital status, number of births, job status or labor market participation. Whereas time-invariant information by definition does not change at each time period, time-variant information must be brought up-to-date using past information and relevant new information, as measured in the SOEP. Thus the yearly updates go through the following process: time-variant information must be

- collected for first-time respondents
- updated for old respondents by either carrying old information over, or by changing status indicators.

Time-invariant information must be collected for all first-time respondents in a current wave, and then appended to the current stock of old respondents. It is the SOEP's aim to provide all biography information, as far as it is collected and without information loss, in a user-friendly manner and as current

as possible, i.e. all time-variant information would correspond to the most recent realized person level interview.

In the documentation concerning the steps necessary to generate biography variables, the file *BIOLELA* may be referenced, which contains all SOEP collected biography information from the person, biography and CV-questionnaires. This file however is not part of the standard distribution as it is very complex in structure and is used as input in generating the distributed variables. It does not contain any information necessary for updating time-variant variables to the most current wave. Further, in this file there are often various redundant information sources, not allowing easy usage by the researcher. However, if desired by the individual researcher, this file can be distributed specially. When all biography related files are complete and in a user friendly form, the *BIOLELA* file will no longer be required.

Unless otherwise explicitly mentioned, the dollar sign ($) in variable names or file names refers to a wave-specific prefix (e.g. `$KMUTTI` in the file `$KIND`: meaning variables `AKMUTTI` through to `UKMUTTI` in the files *AKIND* through to *UKIND*), and a double dollar sign ($$) refers to a variable name suffix (e.g. `NATION$$` in *$PGEN* meaning: `NATION84` through to `NATION04` in the files `APGEN` through to `UPGEN`).

The Structure of Biography Information in the SOEP

- *PPFAD* with Year of Birth/Death, Gender, Year of Migration to Germany, Country of Origin, Location in 1989.

  Sort-ID: `PERSNR`

- *$PGEN* files with Nationality `NATION$$`, highest schooling and job training `$PSBIL`, `$PSBILO`, `$PSBILA`, `$PBBIL01`, `$PBBIL02`, `$PBBIL03`, `$PBBILA`, `$PBBILO`

  Sort-ID: `HHNRAKT` (or `$HHNR`), `PERSNR`

- *PBIOSPE* file with Job Status Calendar in spell form, updated with year-level aggreated monthly information from the *$P* files of the CURRENT wave for the PAST calendar year. (See also *ARTKALEN*).

  Sort-ID: `PERSNR`, `SPELLNR`

- *BIOMARSY* (yearly marital status) and *BIOMARSM* (monthly marital status) contains marriage information in spell format (begin, end, status). *$P* files are used to update the marital status information. Every change in marital status results in a new spell. This information is available on a yearly basis in *BIOMARSY* since the age of 16 for every

respondent (with interview), and on a monthly basis for every respondent (with interview) since January 1 of the calendar year preceeding a person's first SOEP interview.

Sort-ID: `PERSNR`, `SPELLNR`

- *BIOBIRTH* file: for all women with at least one interview since 1984, all births are registered using information directly from the mother, or from the *$PBRUTTO* files. The variables included are: number of births, year of birth, gender of child for up to 15 children. In the case that a child is identified in the SOEP population, then also the child's `PERSNR` is included.

  Sort-ID: `PERSNR` of the woman

- *BIOBRTHM* file: for all *men* with at least one interview since 2001, all births are registered using information directly from the *father*, or from the *$PBRUTTO* files. The variables included are: number of births, year of birth, gender of child for up to 15 children. In the case that a child is identified in the SOEP population, then also the child's `PERSNR` is included.

  Sort-ID: `PERSNR` of the *man*

- *BIOTWIN* file: contains the person numbers of those persons identified as being twin, triplets or quadruplets. Such persons must have identical year and month of birth, identical mothers and identical relationships to the head of household.

  Sort-ID: `PERSNR` of the child

- *BIOCHILD* file: contains information on newborn children up to the age of 15 months concerning birth weight and size, the experience of the mother in and around the pregnancy etc.

  Sort-ID: `PERSNR` of the child

- *BIOPAREN* file: for all persons having given at least one interview since 1984, information concerning parents is stored here: year of birth and death, schooling and job training degrees from the mother and father, job occupation and position of father when the respondent was 15 years old, and if the mother and father can be identified in the SOEP, the `PERSNR` of the mother and father. In addition, three variables on social origin are included.

  Sort-ID: `PERSNR`

- *BIOIMMIG* file: for all those persons, *excluding* those who: are born in Germany and have German nationality and have no valid *BIOIMMIG* information in any wave that they were observed. The variables deal with questions related to foreigners in (and migrants to) Germany. Specifically, questions concerning desire to return to the home country, the presence of relatives in the home country, reasons for coming to Germany, and conditions upon initial arrival in Germany.

  Sort-ID: `HHNRAKT`, `PERSNR`, `ERHEBJ`

- *BIOJOB* file: updated yearly, for all persons with any entry in any *$LELA*-file, even if no biographical data on employment were collected, consisting of samples A-E. The *$LELA*-data relevant for *BIOJOB* consists of: the age at entry into the working force, the type of occupation at entry (blue/white collar worker, self-employed, civil servant), the occupation at entry, changes of occupation and desired schooling/training certificate.

  Sort-ID: `PERSNR`

- *BIOYOUTH* file: collecting information on adolecents since 2000, this file is updated yearly by new sample members answering the "youth" version of the biography questionnaire. Information covers the relationship to parents, leisure activities, school performance and perspectives on career and family planning.

  Sort-ID: `PERSNR`

- *BIOSOC*: collecting information on socialization for all first time respondents aged 18 and over (see also *BIOYOUTH* for those aged 16 and 17). Information covers relationship with parents at age 15, detailed information on school and vocational qualification attained as well as perspectives on further qualification.

  Sort-ID: `PERSNR`

- *BIORESID*: collecting information on second residence.

  Sort-ID: `PERSNR`

For a detailed description of these biography files, please see the extended documentation in Frick and Schneider (2005).

Table 3.2: Biography Data in SOEP. Part 1 of 3

Biography Sub-topic	Number of Question in standard Biography Questionnaire (2004)	Comparable Questions in Youth Questionnaire (2004)	SOEP Target Population	SOEP Data File	Unit of Analysis	Updating Need and required Files	Availability Status (Wave U, 2004)
Place of birth	2, 3	55	All persons surveyed	*PPFAD*	Individual	No	Available
Year of immigration	4	58	For persons not born in Germany	*PPFAD*	Individual	No	Available
Immigration biography	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 15a	57, 59, 60, 61, 62, 63, 64, 65, 66a, 66b	For persons not born in Germany	*BIOIMMIG*	Individual	No	Available
Living in East or West Germany in 1989	16	–	All persons surveyed	*PPFAD*	Individual	No	Available
Place of childhood; Life at childhood residence; grew up with parents, Living together with parents	17, 17a, 19, 20	67a, 67b	All persons surveyed	*BIOPAREN*	Individual	No	Available
Brothers and Sisters	18	68	All persons surveyed	–	Individual	Yes	NOT Available
Parents living region, year of birth, year of death, nationality	21, 22, 23	71, 72, 73	All persons surveyed	*BIOPAREN*	Individual	Partly (year of death from *PPFAD*)	Available
Parents' school and occupational degree, their job + occupation as respondent was 15 years old	24, 25, 26, 27	74, 75, 76, 77	All persons surveyed	*BIOPAREN*	Individual	No	Available
Religious affiliation of parents	28	78	All persons surveyed	*BIOPAREN*	Individual	No	Available
Parents took care about efforts at school	29	39	All persons surveyed	*BIOSOC*	Individual	No	Available
Respondent's last school marks	30	35	All persons surveyed	*BIOSOC*	Individual	No	Available

Table 3.3: Biography Data in SOEP. Part 2 of 3

Biography Sub-topic	Number of Question in standard Biography Questionnaire (2004)	Comparable Questions in Youth Questionnaire (2004)	SOEP Target Population	SOEP Data File	Unit of Analysis	Updating Need and required Files	Availability Status (Wave U, 2004)
Relationship to parents during youth	31	13	All persons surveyed	*BIOSOC*	Individual	No	Available
Sport and activities during youth	32, 33, 34, 35	16, 20, 21, 24	All persons surveyed	*BIOSOC*	Individual	No	Available
Occupational biography	36	–	All persons surveyed	*PBIOSPE*	Spell	Yes (*$P*)	Available
Year and place of acquiring a school degree	37, 38, 41	26a	All persons surveyed	*BIOSOC*	Individual	No (although possible using (*$P*)	Available
Level of school degree	39, 40, 42	26b	All persons surveyed	*$PGEN*	Individual	Yes (*$P*)	Available
Number of foreign classmates in last attended school class	43	43	All persons surveyed	*BIOSOC*	Individual	No	Available
Target school degree	44, 45	27, 28	All persons surveyed	*BIOSOC*	Individual	No	Available
Attained vocational degree, year and place of attaining, certificate of degrees attained abroad	46, 47, 48, 49, 50, 51, 52	44, 45	All persons surveyed	*BIOSOC*	Individual	No (although partly possible using *$P*)	Available

Table 3.4: Biography Data in SOEP. Part 3 of 3

Biography Sub-topic	Number of Question in standard Biography Questionnaire (2004)	Comparable Questions in Youth Questionnaire (2004)	SOEP Target Population	SOEP Data File	Unit of Analysis	Updating Need and required Files	Availability Status (Wave U, 2004)
Target vocational degree	53, 54	46, 47	All persons surveyed	*BIOSOC*	Individual	No	Available
First job (age, occupational position, public sector, industry)	55, 56, 57, 58, 59, 60a, 60b	–	All persons surveyed	*BIOJOB*	Individual	Yes, if person previously did not work (*$P*)	Available
Occupational changes	61	–	All persons surveyed	*BIOJOB*	Individual	Yes	not updated yet (2003)
Last job (year, scope, public sector branch, occupational position)	62, 63, 64, 65, 66, 67	–	All persons surveyed	*BIOJOB*	Individual	Yes	not updated yet (2003)
Year since living personally in current apartment; second residence	68, 69	–	All persons surveyed	*BIORESID*	Individual	No	Available
Births	70	–	All women surveyed (since 2000 all persons surveyed)	*BIOBIRTH* and *BIOBRTHM*	Individual	Yes (*$P*, *$PBRUTTO*, *$KIND*)	Available
Family status (marriage biography)	71, 72	–	All persons surveyed	*BIOMARSY*	Spell	Yes (*$P*, *$PBRUTTO*)	Available
Military or alternative civilian service (only men) and voluntary service	73, 74	–	All persons surveyed	*BIOSOC*	Individual	No (although partly possible using *$P*)	Available
Youth	–	"Youth" Questionnaire	16 and 17 year old respondents	*BIOYOUTH*	Individual	No	Available
Newborns	–	"Mother and Child" Questionnaire	newborns	*BIOCHILD*	Individual	No	Available

3.3 Biography Spell Data

In Table 3.5 panel and spell data are compared. Panel data have one observation per person per year, whereas spell data have one observation per spell per person. It is quite conceivable that a person has only one spell over a given period, such as a male who is full-time employed. For a ten year period, there may be just the one spell "full-time employed". In panel data, the same person would have 10 observations, one per year. A person may have many spells over a time period, and even have overlapping spells, like working part-time and receiving a disability pension. Spell data is useful for looking at stays in a certain state, and transitions in and out of that state. The program LIMDEP from William Greene and SAS will handle some models as well.

Table 3.5: Comparing Panel and Spell Data

Comparison	Panel Data	Spell Data
Sample	N Individuals observed over time, 1 observation/person/year, Sample=$N \times T$	Person i may have j different spells, Each with varying duration Sample=$\sum_{i,j}$
Application	Wage, Employment, Job, Corrections for Unobserved Heterogeneity	Durations in a state: Unemployment, Maternity Leave, Changes from one state to another
Software	LIMDEP, STATA, SAS	LIMDEP, STATA, SAS

The spell data available in the SOEP is divided into the following parts:

- *PBIOSPE*: Activity Biography: Retrospective/Calendar, Yearly
- **bioscope.exe**: Graphical Activity Biography, Yearly
- *ARTKALEN*: Activity Calendar, Monthly
- *EINKALEN*: Income Calendar, Monthly
- *BIOMARSY*: Marital Status Yearly since Birth
- *BIOMARSM*: Marital Status Monthly since Jan 1983
- *SOZKALEN*: HH Social Assistance Monthly from Jan 1991 to Dec 1999

3.3.1 Biography: *PBIOSPE*

The possible activities are listed in Table 3.6. Both data files use not only information from the retrospective question "What were you doing every year since the age of 15 ?", but also starting in 1984, "What were you doing each month of last year ?" Thus two sets of information are joined together, leading to potential censoring problems. Spells missing information at the beginning are considered "left-censored". Conversely, spells missing ending information are considered "right-censored". The trick then becomes, how to join the retrospective questions with the calendar monthly data.

Table 3.6: Biography Spell Data: *PBIOSPE*

	Spelltyp	Spell Type
(1)	Schule,Studium	School/University
(2)	Lehre, Ausbildung	Apprenticeship/Training
(3)	Wehr-, Zivildienst	Military/Social Service
(4)	Voll berufstaetig	Full-time Employed
(5)	Teilzeitbeschaeftigt	Part-time Employed
(6)	Arbeitslos	Unemployed
(7)	Hausfrau, Hausmann	House-Husband/Wife
(8)	Im Ruhestand	Retired
(9)	Andere Taetigkeit	Other

Table 3.7 depicts some actual *PBIOSPE* data. The first two columns of the data are `HHNR` and `PERSNR`, followed by the spell number, and the spell type. Then the next two columns define the merged beginning and end spells, the next two use only retrospective biography (BIO) information to attain beginning and end information, followed by only calendar (CAL) information to attain beginning and end information. Using person 201 again, let us examine the structure of the data.

For spell number 1, spell type 1 occurred (school, university) from the age 15 to 19. This can only use information from the retrospective biography source, which is why there are missing values -2 for the begin and end calendar (CAL) columns.

For spell number 5, the woman was a housewife (spell type 7) from the age 57 to 58. As this happened sometime during the questionnaire period

(1983-onward), there is information from the calendar but not, by definition, from the retrospective biography information (-2).

For spell number 6, which is spell type 8 (retired), retrospective information has the woman being retired from age 40 to 58, but we know that using the calendar information she has been retired from age 57 to 65. Therefore these two pieces of information are joined to one continuous spell from age 40 to 65.

Most users will simply take the information from "BEGIN AGE", and "END AGE", but it is important to know how the data was generated.

Table 3.7: Data from *PBIOSPE*

HHNR	PERNSR	Spell Number	Spell Type	BEGIN AGE	END AGE	BEGIN AGE BIO	END AGE BIO	BEGIN AGE CAL	END AGE CAL	CENSOR	SPELLINF	YEAR	ERRCODE	SAMPLE
19	101	1	2	15	18	15	18	-2	-2	2	1	84	0	1
19	101	2	4	19	58	19	54	53	58	3	3	84	0	1
19	102	1	1	15	15	15	15	-2	-2	2	1	84	0	1
19	102	2	4	16	22	16	22	-2	-2	1	1	84	0	1
19	102	3	7	23	48	23	44	43	48	3	3	84	0	1
19	103	1	1	15	16	15	16	-2	-2	2	1	84	0	1
19	103	2	2	17	20	17	20	-2	-2	1	1	84	0	1
19	103	3	4	20	23	21	21	20	23	3	3	84	0	1
27	201	1	1	15	19	15	19	-2	-2	2	1	84	0	1
27	201	2	2	20	24	20	24	-2	-2	1	1	84	0	1
27	201	3	4	25	29	25	29	-2	-2	1	1	84	0	1
27	201	4	7	30	39	30	39	-2	-2	1	1	84	0	1
27	201	5	7	57	58	-2	-2	57	58	1	2	84	-2	1
27	201	6	8	40	65	40	58	57	65	3	3	84	16	1
27	202	1	1	15	25	15	25	-2	-2	2	1	87	-2	1
27	202	2	1	28	28	-2	-2	28	28	1	2	87	-2	1
27	202	3	4	27	31	27	31	28	30	3	4	87	-2	1
27	202	4	5	26	26	26	26	-2	-2	1	1	87	-2	1
27	202	5	5	30	30	-2	-2	30	30	1	2	87	-2	1
27	202	6	6	29	29	29	29	-2	-2	1	1	87	-2	1
27	202	7	7	29	29	-2	-2	29	29	1	2	87	-2	1

3.3.2 Biography: bioscope.exe and biosco95.exe

bioscope.exe is a program written for MSDOS by Rainer Pischner. This program allows graphical viewing of the *PBIOSPE* information, for each person in the sample. The program uses retrospective information collected in in the first wave (1984) and updates that information using an aggregated version of the monthly calendar information for all waves thereafter.

For example, in Table 3.8 the person 201, who is a women, born in 1926, was in schooling from age 15-19, then from 20-24 did some kind of training, and then worked full-time for 5 years. From the age of 30 to 39, she was a housewife, and then at the age of 40 retired. At the age of 56 and 57 she had overlapping retired and housewife spells. The program **biosco95.exe** views the 95% version of the SOEP data.

In Table 3.9, the data for **bioscope.exe** are explained. All data that can be viewed in Bioscope, can also be printed to an output file, for use in other econometrics packages. The file format is very simple, consisting of only 6 variables. First, the household and personal identifiers, and then an sequentially listed event indicator. Further, there is an event type indicator, corresponding to the first column of Table 3.8, and the begin age, and finally the end age for which the event lasted.

All persons have the event number "0". This is a special case, where the next variables are the gender (1=male, 2=female), the date of birth, and the date first interviewed. In Table 3.9, the person 201 (using the event number 0 information) is female, born in 1926 and first interviewed in 1984. Her first event (1) was in schooling from age 15-19, then (2) from 20-24 did some kind of training, and then (4) worked full-time for 5 years. From the age of 30 to 39, she was (7) a housewife, and then at the age of 40 retired (8). At the age of 56 and 57 she had overlapping retired and housewife spells.

Table 3.8: **bioscope.exe** and Graphical Representation

```
+-------------------------------------------------------------------------------+
|                                                                               |
| CASE-ID:     27  Personal-Number:      201   Year of birth:1926    Sex:  2    |
|                                                                               |
|                                                                               |
|                                                      1983-->|<--1984          |
| Any Activity........ ***********************************************oooooooo  |
|                      <--1941 Start of Working Life History                    |
|                                                                               |
| At Age of      ...   1    2    2    3    3    4    4    5    5    6    6 Sum of |
| the person was ...   5    0    5    0    5    0    5    0    5    0    5 events |
|                                                                               |
| School, University.. XXXXX|....|....|....|....|....|....|....|....|....|    5   |
| Training etc........ |....XXXXX|....|....|....|....|....|....|....|....|    5   |
| Military Service.... |....|....|....|....|....|....|....|....|....|....|    0   |
| Full employed....... |....|....XXXXX|....|....|....|....|....|....|....|    5   |
| Part time employed.. |....|....|....|....|....|....|....|....|....|....|    0   |
| Unemployed.......... |....|....|....|....|....|....|....|....|....|....|    0   |
| Housewife,-man...... |....|....|....XXXXXXXXXX|....|....|....|.OO.|....|   12   |
| Retire.............. |....|....|....|....|....XXXXXXXXXXXXXXXXXXOOXXXXXXX   26   |
| Other activity...... |....|....|....|....|....|....|....|....|....|....|    0   |
+-------------------------------------------------------------------------------+
```

Table 3.9: Bioscope Spell Data

HHNR	PERSNR	Spell Nr. = 0	Man/Woman	DOB	Date	Interpretation
		Spell Nr. > 0	Spell Type	Begin Age	End Age	
27	201	0	2	1926	84	Woman, born 1926
27	201	1	1	15	19	School, University
27	201	2	2	20	24	Training
27	201	3	4	25	29	Full employed
27	201	4	7	30	39	Housewife,-man
27	201	5	7	57	58	Housewife,-man
27	201	6	8	40	65	Retire

3.3.3 Activity Calendar: *ARTKALEN*

The *ARTKALEN* contains spells (monthly) for events starting in January 1983. This is in contrast to *PBIOSPE*, where spells were in yearly durations, and events previous to 1983 were included.

In Table 3.10, the question is asked of the respondents, and the activity item number corresponds to the order of possible items. The respondent would simply check off for each month, the appropriate activities. To generate the spells, all monthly calendars, from previous years as well, are used.

Table 3.10: Activity Calendar

74.	And now please think about the entire previous year, in other words about 1993: We have made a sort of calendar. On the left, we have written things that could have happened last year. Please go through the entire list and check each month, in which, for example, you were employed or unemployed, etc. Please make sure you answer for each month.											
1993	J	F	M	A	M	J	J	A	S	O	N	D
full-time employment, job creation measure	_	_	_	_	_	_	_	_	_	_	_	_
Short-time work	_	_	_	_	_	_	_	_	_	_	_	_
part-time or occasionally employed	_	_	_	_	_	_	_	_	_	_	_	_
vocational training, education, retraining	_	_	_	_	_	_	_	_	_	_	_	_
registered unemployed	_	_	_	_	_	_	_	_	_	_	_	_
retired, early retirement	_	_	_	_	_	_	_	_	_	_	_	_
maternity leave	_	_	_	_	_	_	_	_	_	_	_	_
in school/college	_	_	_	_	_	_	_	_	_	_	_	_
military/civilian service	_	_	_	_	_	_	_	_	_	_	_	_
housewife/househusband	_	_	_	_	_	_	_	_	_	_	_	_
other, namely _____	_	_	_	_	_	_	_	_	_	_	_	_

An examination of the some actual data from *ARTKALEN* reveals the simple structure of the data. Table 3.11 shows data from the first 5 observations of the data, each person separated by a horizontal line. Examining person 201 again, it is clear that for the time period January 1983 (month 1) to December 1994 (month 132), she had been retired (code 6). In addition from January 1983 (month 1) to December 1984 (month 24), she had been a housewife (code 10).

Table 3.11: Data from *ARTKALEN* (as of 1994)

HHNR	PERNSR	Spell Number	Spell Type	BEGIN MONTH	END MONTH	CENSOR	SAMPLE
19	101	1	1	1	72	5	1
19	102	1	10	1	72	5	1
19	103	1	1	1	48	5	1
27	201	1	6	1	132	5	1
27	201	2	10	1	24	4	1
27	202	1	1	13	20	4	1
27	202	2	1	27	28	1	1
27	202	3	1	31	31	1	1
27	202	4	1	35	36	1	1
27	202	5	1	39	48	2	1
27	202	6	3	37	38	1	1
27	202	7	8	21	24	1	1
27	202	8	10	25	26	1	1
27	202	9	10	29	30	1	1
27	202	10	10	32	34	1	1

3.3.4 Income Calendar: *EINKALEN*

The income calendar is used to gain information about sources of income throughout the year. The Cross-Nation Equivalent File from Cornell University uses exactly this information to generate *yearly* income variables. Further details are available in Section 2.3. Table 3.12 displays the question asked of the respondents concerning sources of income. The respondent checks off for each month all appropriate sources of income. There may of course be several sources of income for any given month.

Table 3.12: Income Calendar (only until 1995)

```
+-------------------------------------------------------------------------+
| Income, 1993                                                            |
|                                                                         |
| 75.     Here we have another calendar.  On the left are various         |
|         types of income.  Please go through the list.  If you yourself  |
|         were in receipt of the income mentioned, then check each        |
|         applicable month.                                               |
|                                                                         |
| 1993                                  Income  J F M A M J J A S O N D   |
| earnings/wages (including payments                                      |
|         for training, early- retirement)        _ _ _ _ _ _ _ _ _ _ _ _ |
| self-employed earnings                          _ _ _ _ _ _ _ _ _ _ _ _ |
| earnings from second job                        _ _ _ _ _ _ _ _ _ _ _ _ |
| old-age pension/invalid pension/company                                 |
|         pension (due to gainful employment)     _ _ _ _ _ _ _ _ _ _ _ _ |
| widower's/widow's/orphan allowance              _ _ _ _ _ _ _ _ _ _ _ _ |
| student grant, scholarship                      _ _ _ _ _ _ _ _ _ _ _ _ |
| maternity pay during maternity leave            _ _ _ _ _ _ _ _ _ _ _ _ |
| unemployment benefit                            _ _ _ _ _ _ _ _ _ _ _ _ |
| unemployment relief                             _ _ _ _ _ _ _ _ _ _ _ _ |
| support from employment office for                                      |
|         additional or re- training              _ _ _ _ _ _ _ _ _ _ _ _ |
| payments/support from persons who do not live                           |
|         in the household (including public                              |
|         maintenance support)                    _ _ _ _ _ _ _ _ _ _ _ _ |
| received no benefits of this type in this month _ _ _ _ _ _ _ _ _ _ _ _ |
|                                                                         |
| received no income at all in 1993 of these types _____                  |
+-------------------------------------------------------------------------+
```

Again we shall examine the person 201. Table 3.13 displays actual data for the first 5 individuals. The woman 201 received retirement benefits or pension from January 1983 (month 1) to December 1994 (month 132). From January 1983 (month 1) to December 1983 (month 12), and again from January 1986 (month 37) to December 1986 (month 48), she received payments from persons outside the household.

From 1984 until 1995, the questionnaire asked whether a person received

a certain source of income, and in which particular month(s) that income was received. Starting in 1996, only the *number* of months was asked, *not* which particular months that the income was received. As the information was retrospective for the previous year, the information contained only covers Jan 1983 (month=1) to Dec 1994 (month=144) and cannot be extended.

Table 3.13: Data from *EINKALEN*

HHNR	PERNSR	Spell Number	Spell Type	BEGIN MONTH	END MONTH	CENSOR	SAMPLE
19	101	1	1	1	72	5	1
19	102	1	12	1	72	5	1
19	103	1	1	1	48	5	1
27	201	1	4	1	132	5	1
27	201	2	11	1	12	4	1
27	201	3	11	37	48	1	1
27	202	1	1	13	20	6	1
27	202	2	1	27	28	1	1
27	202	3	1	31	31	1	1
27	202	4	1	35	48	2	1
27	202	5	12	25	26	7	1
27	202	6	12	29	30	1	1
27	202	7	12	32	34	1	1
27	202	8	99	21	24	-2	1

3.3.5 Yearly Marital Biography: *BIOMARSY*

The files *BIOMARSY* and *BIOMARSM* contain spell information on marital status of each respondent.

BIOMARSY is on a yearly basis, measuring begin and end of each marital status spell in years of age (`BEGIN` of the very first spell is 0). The variable marital status `SPELLTYP` has the following codings: (1) single, (2) married, (3) divorced, (4) widowed, (5) separated (no differentiation possible between divorced and widowed), (8) missing because of item-non-response and (9) missing because of unit-non-response.

Table 3.14 displays a sample portion of the data in the data file *BIOMARSY*. Person `101` from household number `19` was single until age 24, was married at 24 until age 28 when he divorced. At the age of 32 he remarried and stayed married until age 59. His wife, person number `102` appears to be his second wife. She married him at age 22 (when he was 32) and has remained married at age 49. Person number `103` appears to be the product of the second marriage, who at 24 years of age is still single.

All adult persons will have at least one entry in this data file, as by definition, all persons are single at birth. Thus, `SPELLNR=1` will always have a `SPELLTYP=1`.

Table 3.14: Data from *BIOMARSY* (as of 1996)

HHNR	PERSNR	SPELLNR	SPELLTYP	BEGIN	END	REMARK
19	101	1	1	0	24	0
19	101	2	2	24	28	0
19	101	3	3	28	32	0
19	101	4	2	32	59	0
19	102	1	1	0	22	0
19	102	2	2	22	49	0
19	103	1	1	0	24	0
27	201	1	1	0	27	0
27	201	2	2	27	39	0
27	201	3	3	39	70	0
27	202	1	1	0	31	0
27	203	1	1	0	36	8
35	301	1	1	0	24	0
35	301	2	2	24	33	0
35	302	1	1	0	23	0
35	302	2	2	23	32	0

3.3.6 Monthly Marital Biography: *BIOMARSM*

BIOMARSM is on a monthly basis using information from the yearly questions on marital status and marital status changes since the previous year up to and including the month of the most recent interview. Thus, begin and end of each marital status spell are given in months (month 1 = January 1983, ... , month 167 = November 1996, etc). There is at least one entry for each adult person, with the first spell being the status of the person in January of the year previous of the first interview.

Table 3.15 displays an excerpt from the data file *BIOMARSM*. In the household `35`, person number `301` was married in November 1984 until March 1993 (or for 123 months). After that this person's marital status is unknown. It appears that person number `302` is `301`'s wife, as the marital status information is identical.

Table 3.15: Data from *BIOMARSM* (as of 1996)

HHNR	PERSNR	SPELLNR	SPELLTYP	BEGIN	END	REMARK
19	101	1	2	1	75	0
19	102	1	2	1	75	0
19	103	1	1	1	51	0
27	201	1	3	1	158	0
27	202	1	1	13	51	0
27	203	1	1	13	158	8
35	301	1	1	1	23	0
35	301	2	2	23	123	0
35	302	1	1	1	23	0
35	302	2	2	23	123	0
43	401	1	4	1	39	0
51	501	1	4	1	110	0
60	601	1	1	1	59	0
60	601	2	2	59	165	0
60	602	1	1	1	92	0
60	602	2	2	92	160	0
78	701	1	4	1	39	0
86	801	1	1	1	27	0
94	901	1	1	1	159	0

3.4 Biography Individual Data

3.4.1 Parents: *BIOPAREN*

BIOPAREN contains all individuals with at least one interview starting 1984. Using data from the biography questionnaire and the yearly information from the *$P* and *$PGEN* files, this file contains information on parents, which can be used for intergenerational analyses.

The variables include: year of birth and year of death for both parents; schooling and occupational education of both parents; occupational status of the father, when respondent was 15 years old; religious membership of both parents; `PERSNR` of parents, if they could be identified in the SOEP.

Additional variables indicate whether the information on a parent is a proxy-information given by the child in the course of answering the biography questionnaire or was asked from the parents themselves. This information is used only in case of missing proxy-information.

Table 3.16: List of variables in *BIOPAREN*

Variable	Meaning
HHNR	Original Household Number
PERSNR	Never Changing Person ID
VNR	Person Number Father
MNR	Person Number Mother
VGEBJ	Year Of Birth Father
MGEBJ	Year Of Birth Mother
VTODJ	Year Of Death Father
MTODJ	Year Of Death Mother
VAORT91	Place Of Residence Father 91
MAORT91	Place Of Residence Mother 91
VAORT96	Place Of Residence Father 96
MAORT96	Place Of Residence Mother 96
VSBIL	Level Of Education Father
MSBIL	Level Of Education Mother
VBBIL	Vocational Training Father
MBBIL	Vocational Training Mother
VSINFO	Origin SBIL Father
MSINFO	Origin SBIL Mother
VBINFO	Origin BBIL Father
MBINFO	Origin BBIL Mother
VRELI	Religion Father
MRELI	Religion Mother
VBSINFO	Origin VBSTELL Father
ORTKINDH	Place Raised To Age 15
ORTKIND1	Still Lives In Town Where Raised
BIOYEAR	Year Of Biography Survey
VNAT	Nationality of Father
MNAT	Nationality of Mother
VAORT01	Place Of Residence Father 01
MAORT01	Place Of Residence Mother 01
MBSINFO	Origin VBSTELL Mother
VISCO88	FATHER ISCO88 - New Generation
MISCO88	MOTHER ISCO88 - New Generation
VISEI	FATHER ISEI-Status88 Ganzeboom (IS88)
MISEI	MOTHER ISEI-Status88 Ganzeboom (IS88)
VMPS	FATHER ISEI-Status88 Ganzeboom (IS88)
MMPS	MOTHER ISEI-Status88 Ganzeboom (IS88)
VNACE	FATHER Branch, NACE
MNACE	MOTHER Branch, NACE
VSIOPS	FATHER TREIMANS STANDARD INT.OCC.PR.SCORE (IS88)
MSIOPS	MOTHER TREIMANS STANDARD INT.OCC.PR.SCORE (IS88)
VEGP	FATHER ERIKSON,GOLDTHORPE Class Category (IS88)
MEGP	MOTHER ERIKSON,GOLDTHORPE Class Category (IS88)
VKLAS	FATHER StaBuA 1992 Job Classification
MKLAS	MOTHER StaBuA 1992 Job Classification
LIVING1	No. Of Years Living With Bio. Parents
LIVING2	No. Of Years Living With Single Mother
LIVING3	No. Of Years Living With Single Mother And Partner
LIVING4	No. Of Years Living With Single Father
LIVING5	No. Of Years Living With Single Father And Partner
LIVING6	No. Of Years Living With Other Relatives
LIVING7	No. Of Years Living With Foster Parents
LIVING8	No. Of Years Living In Home
VBSTELL	Job Position Father
MBSTELL	Job Position Mother
VSTREIT	Argue Or Fight With Father When Respondent 15
MSTREIT	Argue Or Fight With Mother When Respondent 15
VAORTAKT	Most recent whereabouts of Father
MAORTAKT	Most recent whereabouts of Mother
VAORTUP	Update timing - VAORTAKT
MAORTUP	Update timing - MAORTAKT

3.4.2 Births: *BIOBIRTH* and *BIOBRTHM*

BIOBIRTH contains all women with at least one interview between 1984 and onward. Using data from the biography questionnaire and the yearly information from the *$PBRUTTO* files from 1984 onward, this file contains information on births. The variables include: number of children ever born; year of birth and gender of up to 15 children; `PERSNR` of children if they could be identified in the SOEP.

Table 3.17: List of variables in *BIOBIRTH* and *BIOBRTHM*

Variable	Meaning
`HHNR`	Original HH number
`PERSNR`	Person Number of Woman
`BIOYEAR`	Year of Biography Information
`BIOVALID`	Status of Birth Biography Information 0: No Birth Information 1: Birth Information with Children 2: Birth Information without Children
`BIOAGE`	Age of Woman/Man when Birth Biography Information collected
`SUMKIDS`	Total Number of Children Born
`BIOKIDS`	Total Number of Children with Birth Biographies
`NEWKIDS`	Total Number of Children identified through *$PBRUTTO*, *$KIND*
`KIDGEB$$`	Year of Child's Birth (Child 01 to 15)
`KIDSEX$$`	Gender of Child (Child 01 to 15)
`KIDPNR$$`	Person Number of Child (Child 01 to 15)

BIOBRTHM contains birth information from *male respondents* with at least one interview since 2000. This information is analog to *BIOBIRTH*, however, only for men who joined the survey after 2000 when biography questionnaire was redesigned for sample F.

3.4.3 First Job Information: *BIOJOB*

The *$LELA*-data relevant for BIOJOB consists of (a) the age at entry into the working force (b) the type of occupation at entry (blue/white collar worker, self-employed, civil servant) (c) the occupation at entry (d) changes of occupation.

The purpose of BIOJOB is to create a file, which offers the user convenient access to biographical information on past job activities. Up to now all but two variables of BIOJOB are time-invariant. Information on occupational changes and on the age at the most recent change of occupation refer to the date of the respondents biography interview. The population is made up of all persons with an entry in any *$LELA*-file, even if no biographical data on employment were collected.

Table 3.18: List of variables in *BIOJOB*

Variable	Meaning
HHNR	original household identifier
PERSNR	unique individual identifier
BIOYEAR	year of biography / youth interview
AGEFJOB	age at first job
AGEINFO	information source AGEFJOB
NOJOB	never worked before the time of the interview
STILLFJ	still employed in first job
OCCFJOB	occupational position first job
FULLTIME	first job was a full-time or part-time job
FJBLUE	first job blue collar worker
FJSELFE	first job self-employed
FJSEFSIZ	number of employees FJSELFE
FJWHITE	first job white collar worker
FJCIVS	first job civil servant
ISCO88	International Standard Classification of Occupation 1988
STBA	classification of career according to the German Federal Statistical Office
EGP	Erikson and Goldthorpe's Class Category (EGP)
ISEI	Intl. Socio-Economic Index of Occupational Status, Ganzeboom (ISEI)
MPS	Magnitude Prestige Scale after Wegener
SIOPS	Treiman Standard Int. Occ. Prestige Scale
REQEDUC	required education for first job
CIVILSFJ	first job was in civil service
NACEFJ	NACE branch code first job
OCCMOVE	number of occupational changes
AGEATMV	age at most recent occupational change
INTEDUC1	highest intended educ. degree
INTEDUC2	second intended educ. degree
INTEDUC3	third intended educ. degree
INTEDUC4	fourth intended educ. degree
CURREMPL	employed at time of biography interview
YEARLAST	year of last employment
SCOPELJ	last job was a full-time or part-time job
CIVILSLJ	last job was in civil service
NACELJ	NACE branch code last job
OCCLJOB	occupational position last job
LJBLUE	last job blue collar worker
LJSELFE	last job self-employed
LJSEFSIZ	number of employees LJSELFE
LJWHITE	last job white collar worker
LJCIVS	last job civil servant

Important note: The data set *BIOJOB* has not been updated to wave U (2004), i.e., it contains the exact same information as in the 2003 release.

3.4.4 Migration Information: *BIOIMMIG*

The variables contained in *BIOIMMIG* deal with questions related to foreigners in (and migrants to) Germany. Specifically, questions concerning desire to return to the home country, the presence of relatives in the home country, reasons for coming to Germany, and conditions upon initial arrival in Germany. A complete list of variables is shown below. In addition to these variables, there are of course `CORIGIN`, `GERMBORN`, `IMMIYEAR` in the file *PPFAD* for *all individuals* (including children) as discussed already in Section 2.1.1.

The data available in *BIOIMMIG* are longitudinal, that is to say, the same variable name refers to different time periods, differentiated by the variable `ERHEBJ`. The data is stacked for each person, such that the unit of observation is a person-year. Thus for every person, there are as many observations as interviews given by this person. Much of the information was asked only once, and "carried" forth in the following years.

The sample in the dataset is defined by taking all available information and deleting all those persons who: (a) are born in Germany and (b) have German nationality and (c) have no valid *BIOIMMIG* information in any wave that they were observed.

As the data consists of person-year observations, if a person is excluded from the sample, then for all years. However if a person once belonged to the sample, then he is always included (say, even after receiving German citizenship).

The variables found in *BIOIMMIG* are created first using information from the SOEP biography files, the so-called *BIOLELA*, *$LELA* (starting with wave M) files. If valid information is found in the *$LELA* files for the given response year, then it is taken. Yearly valid update information is taken from the foreigner specific files *APAUSL* through *LPAUSL* and the foreigner specific questions in *MP* through *UP*. Starting with wave M, the foreigner specific variables are found in the regular *$P* files, as the questionnaire is identical for natives and foreigners. Sometimes there is competing information in the biography and regular yearly person questionnaires. The most recent valid information is taken to be correct. First the *$LELA* info is used and then updated with person questionnaire info.

The *BIOIMMIG* file can be used in cross-section or in panel. The usual matching variables are included.: `PERSNR` (Person Number), `HHNR` (Original HH Number), `HHNRAKT` (Current HH Number for survey year given in `ERHEBJ`), `ERHEBJ` (Year). The data is sorted by `HHNR`, `HHNRAKT`, `PERSNR` and `ERHEBJ` such that there are typically many person-year observations for every person. In that sense, the data are ready to be used/matched to a longitudinal dataset. However, simply by selecting on the appropriate year in `ERHEBJ`, the file can be used cross-sectionally as well. The variables correspond to those in the

Wave 13/M/1996 Biography Questionnaire.

Table 3.19: List of variables in *BIOIMMIG*

Variable	Meaning
PERSNR	Person Number
HHNR	Original HH Number
HHNRAKT	Current HH Number for ERHEBJ
ERHEBJ	Current Year / Year Answered
BIIMGRP	BI: Immigration Group
BIRESPER	BI: Residence Status
BICAMP	BI: Refugee Residence Y/N
BICAMPW	BI: Refugee Residence: Weeks
BICAMPM	BI: Refugee Residence: Months
BIWFAM	BI: Already had Family in Country
BIFAMC	BI: Contacts with Family in Germany
BIFAMCL	BI: Moved to Same City/Town as Family
BIRBETR	BI: Reason Migrate: Better
BIRMONEY	BI: Reason Migrate: Money
BIRFREE	BI: Reason Migrate: Freedom
BIRFAM	BI: Reason Migrate: Family
BIRPOOR	BI: Reason Migrate: Poor
BIRWAR	BI: Reason Migrate: War
BIRJUST	BI: Reason Migrate: Just So
BIROTHR	BI: Reason Migrate: Other
BIEXPR	BI: Expectations in Germany
BIEXPRLV	BI: Expectations: Find Apt
BIEXPRAC	BI: Expectations: Accepted by Coworker
BIEXPRAN	BI: Expectations: Accepted by Neighbor
BIRELH	BI: Family Abroad
BIRELHP	BI: Family Abroad: Parents
BIRELHGP	BI: Family Abroad: Grandparents
BIRELHC	BI: Family Abroad: Children
BIRELHBS	BI: Family Abroad: Brother/Sister
BIRELHDR	BI: Family Abroad: Distant Relatives
BIRELHSP	BI: Family Abroad: Spouse
BIRELHFR	BI: Family Abroad: Friends
BIRELHMI	BI: Persons abroad bring to Germany
BIRELHS2	BI: Spouse in Germany
BIRELHC2	BI: Underage Children not in Germany
BIGOBACK	BI: Go back home ?
BISTAY	BI: Desire to Stay in Germany
BISTAYY	BI: Years Desired to Stay in Germany
BISCGER	BI: Attended School in Germany
BISCGRAD	BI: Which Grade School
BISCGERC	BI: Attended Special Foreigner Prep Class
BISCGC	BI: Also German Pupils in Class
BISCGCF	BI: How many Pupils foreign
BISCGCFN	BI: Mix of Nationalities in Class

3.4.5 The Youth Questionnaire and *BIOYOUTH*

BIOYOUTH contains youth specific information collected in the Youth Questionnaire since 2000, starting with the birth cohort 1984. The population is made up by first time respondents aged 16 and 17 who answer the youth questionnaire instead of the standard biography questionnaire. Information covers relationship to parents, leisure time activities, school performance as well as information on personality characteristics. Numerous prospective questions about intentions for further education and training as well as career and family planning are included. Data collected in the youth questionnaire on "Immigration" and "Childhood and Parental Home" are not stored in *BIOYOUTH* but in the corresponding biographical files *BIOIMMIG* and *BIOPAREN*, respectively.

Table 3.20: List of variables in *BIOYOUTH*

Variable	Content of the Variable
HHNR	Original household identifier
HHNRAKT	Current household identifier
PERSNR	Personal identifier
BEFRPER	Respondent identifier
ERHEBJ	Survey year
BYGEBJAH	Year of birth
BYMNR	identifier of mother (taken from BIOPAREN; social, not necessarily biological relationship)
BYVNR	identifier of father (taken from BIOPAREN; social, not necessarily biological relationship)
	Residence
BYWOELT	Residing in parents‘ household (HH)
BYWOZIM	Own room
BYWOWEI	Additional apartment outside of parents' HH
	Jobs and Money
BYVDEIG	Own income
BYVDART	Type of income
BYJBFRUE	Worked before (on holiday or while in school)
BYJBALT	Age by first job (on holiday or while in school)
BYJBGRUN	Reason for working
BYTGELD	Allowance
BYTGELDW	Amount of allowance per week
BYTGELDM	Amount of allowance per month
BYSPAR	Saving money
BYSPARM	Amount saved every month
BYSPARUN	Sporadic saving
	Relationships
	Importance of various persons:
BYWIVA	Father
BYWIMU	Mother
BYWIBS	Brother, Sister
BYWIVW	Other related persons
BYWIFFR	Serious boy/girlfriend
BYWIBFR	Best friend
BYWILEHR	Teacher
BYWICLQ	Clique
BYWISON	Other persons
	Frequency of fights with:

```
BYSTRVA    Father
BYSTRMU    Mother
BYSTRBS    Brother, Sister
BYSTRFFR   Serious boy/girlfriend
BYSTRBFR   Best friend

BYBZO1MU   Talk with mother about personal experiences
BYBZO1VA   Talk with father about personal experiences
BYBZO2MU   Mother addresses problems
BYBZO2VA   Father addresses problems
BYBZO3MU   Mother asks opinion before a decision is made
BYBZO3VA   Father asks opinion before a decision is made
BYBZO4MU   Mother shows approval
BYBZO4VA   Father shows approval
BYBZO5MU   Solve problems together with mother
BYBZO5VA   Solve problems together with father
BYBZO6MU   Mother shows trust
BYBZO6VA   Father shows trust
BYBZO7MU   Mother asks opinion on family issues
BYBZO7VA   Father asks opinion on family issues
BYBZO8MU   Mother justifies decision
BYBZO8VA   Father justifies decision
BYBZO9MU   Mother shows love
BYBZO9VA   Father shows love

           Free time and Sport
           Frequency of free time activities:
BYFZFERN   TV, Video
BYFZPC     Computer games
BYFZMUSH   Listen to music
BYFZMUSS   Play music
BYFZSPRT   Do sports
BYFZTANZ   Dance, Theatre
BYFZTECH   Technical work, Programming
BYFZLESE   Read
BYFZEHRE   Volunteer activities
BYFZABH    Do nothing, hang around, day dream
BYFZMFFR   Spend time with boy/girlfriend
BYFZMBFR   Spend time with best friend
BYFZMCLQ   Spend time with clique
BYMUSSP    Actively make music
BYMUSART   Style of music made
BYMUSMW    Play music with whom
BYMUSALT   Age starting playing music
BYMUSUNT   Paid music lessons
BYSPRTTR   Participate in sports
BYSPRTAR   Favourite sport
BYSPRTAL   Age started favourite sport
BYSPRTMW   Where and with whom favourite sport
BYSPRTWE   Participation in competitions

           School
BYSCHBES   School attendance
BYSCHART   Type of general school attended
BYSCHEND   Last year of school
BYSCHABS   Type of school certificate
BYSCHZUK   Strive for further school certificate
BYSCHZAR   Type of further school certificate
BYFMD1     1. foreign language
BYFMD2     2. foreign language
BYSCHAUS   School attendance in foreign country
BYSCHPRI   Attendance in a private school

           Activities in school:
BYENKSPR   Class representative
BYENSSPR   School representative
BYENSZTG   School newspaper
BYENTHEA   Theatre, Dance group
BYENCHOR   Choir, Music
BYENSPRT   Sport group
BYENSONS   Other groups
BYENNEIN   No activities
```

BYZFINSG	Satisfaction with effort at school (overall)
BYZFDEUT	Satisfaction with effort in German
BYZFMATH	Satisfaction with effort in math
BYZFFMD1	Satisfaction with effort in 1. foreign language
BYEMPFEH	Recommendation after elementary school
BYNTDEUT	Last grade in German
BYNTMATH	Last grade in math
BYNTFMD1	Last grade in first foreign language
BYPTDEUT	Total points in German
BYPTMATH	Total points in math
BYPTFMD1	Total points in first foreign language
BYGSDEUT	Level of German at comprehensive school
BYGSMATH	Level of math at comprehensive school
BYGSFMD1	Level of first foreign language at comprehensive school
BYLKDEUT	Complementary / main subject in German
BYLKMATH	Complementary / main subject in math
BYLKFMD1	Complementary / main subject in first foreign language
BYKLWDJA	Class repeated
BYKLWD1	Class level 1. repeated
BYKLWD2	Class level 2. repeated
BYNACHHI	Paid tutor lessons
BYELKUEM	Parents care about efforts at school
BYELHAUS	Parents help with homework
BYELDIFF	Problems with parents because of effort at school
BYELABEN	Parents attend parents' evening
BYELSPRE	Parents go to parents' day
BYELLEHR	Parents go to see a teacher
BYELVERT	Active as parent representative
BYELNIDA	Parents do not participate in any of these activities
BYKLAUSL	Number of foreign classmates
	Education and Career plans
BYBAABGE	Vocational education, Internship, training
BYBABGJ	Vocational introductory year ("Berufsgrundschul? / Berufsvorbereitungsjahr")
BYBABEGL	Vocational integration training ("Berufl. Eingliederungslehrgaenge")
BYBALEH	Vocational education, apprenticeship ("Berufsausbildung, Lehre")
BYBABFS	Full-time vocational school/ School for public health ("Berufsfachschule / Schule des Gesundheitswesen")
BYBAPRAK	Internship ("Praktikum, Voluntariat")
BYZAJA	Vocational / university degree is aspired
	Type of aspired vocational / university degree:
BYZALEH	Apprenticeship ("Lehre")
BYZABFS	Full-time vocational school/ School for public health ("Berufsfachsule / Schule des Gesundheitswesen")
BYZAFSC	Technical school, school for master of a trade ("Fachschule, Meister-, Technikerschule")
BYZABEA	Training for civil servants (officer) ("Beamtenausbildung")
BYZABAK	Approved vocational academy ("anerkannte Berufsakademie")
BYZAFH	Advanced technical college ("Fachhochschule")
BYZAUNI	University
BYSLBALT	Desired age for financial independence
BYSLBHEU	Already financially independent
BYBWUNJA	Occupation is aspired
	Occupation categories, encoded:
BYKLAS	Classification of career according to the German Federal Statistical Office
BYISCO88	International Standard Classification of Occupation 1988 (ISCO88)
BYEGP	Erikson and Goldthorpe's Class Category (EGP)
BYISEI	International Socio-Economic Index of Occupational Status, Ganzeboom (ISEI)
BYSIOPS	Treiman's Standard International Occupational Prestige Scale (SIOPS)
BYMPS	Magnitude Prestige Scale after Wegener (MPS)
BYZBINF	Information level of planned career
BYZBELT	Influence of the parents on career choice
BYZBLAS	No specific career in mind
BYZBBES	Intensive thoughts about various careers

BYZBRAU	Still looking for a career
BYWBSICH	Secure job
BYWBEINK	High income
BYWBAUF	Promotion opportunities
BYWBANE	Established profession
BYWBFREI	Enough free time
BYWBINT	Interesting activities
BYWBSELB	Working independently
BYWBKONT	Contact with persons
BYWBGSL	Relevant to society
BYWBGSND	Healthy conditions at work
BYWBFAM	Flexibility for family
BYWBHELF	Help others
	Future
	Probability of future career related and private events:
BYWAAUSP	To be accepted for a desired apprenticeship / place at university
BYWAERFA	To complete Training/ university successfully
BYWAARBP	Job in desired career
BYWABERF	Job-related success
BYWAARBL	Longer unemployment
BYWAZURU	From family related reasons held back in career
BYWASELB	Self-employed
BYWAAUSL	Work in foreign country
BYWAHEIR	To marry
BYWAPART	Live together with partner (not married)
BYWAKID1	Have one child
BYWAKIDM	Have more than one child
	Attitudes and Opinions
BYGLPART	Happiness: live with/without partner
BYGLKIND	Happiness: with/without children
	Success in Federal Republic of Germany (FRG) from
BYEFFLEI	Studiousness
BYEFAUSN	Exploitation of others
BYEFINT	Intelligence
BYEFFAM	Family's origin
BYEFFACH	Technical know-how
BYEFGELD	Money
BYEFSABS	School education
BYEFHART	Being inconsiderate and hard
BYEFBEZ	Networking
BYEFPOLI	Political activities
BYEFMANN	Sex/ 'being a man'
BYEFINI	Being dynamic and taking initiative
BYESVERL	What happens in life, depends on me
BYESERRE	Did not reach, what I deserve
BYESGLUE	What you achieve, is a matter of luck
BYESAND	Others decide about my life
BYESHART	You have to work hard for success
BYESZWEI	By difficulties, doubt about own abilities
BYESSOZU	Chances are determined by social circumstances
BYESFAEH	Abilities are more important than efforts
BYESKNTR	Little control over events in my life
BYESENGA	Change of social circumstances through social/political activities
	Specification of Interview Situation
BYINTA	Type of interview
BYDAUER1	Duration of personal interview
BYDAUER2	Duration of interview filled out independently
BYANW	Presence of other persons
BYTAGIN	Day of the interview
BYMONIN	Month of the interview
BYINTNR	Identifier of the interviewer

3.4.6 Information on Socialisation: *BIOSOC*

In the year 2001 the standard supplementary Biography Questionnaire was extended to capture some specific questions on youth and early adulthood. Part of these questions are derived from the independent Youth Questionnaire which was started in 2000. Respondents aged 18 and over are now asked information concerning the relationship to their parents at age 15, school performance, last educational qualification attained, vocational qualification as well as the intentions regarding further educational and vocational qualification, etc. This data set will be complemented by new respondents entering the SOEP population year by year, thus increasing the number of observations potentially available for cohort analysis.

Table 3.21: List of variables in *BIOSOC*

Variable	Meaning
HHNR	Original household identifier
HHNRAKT	Current household identifier
PERSNR	Person identifier
BEFRPER	Respondent identifier
ERHEBJ	Survey year
BSGEBJAH	Year of birth
	School
BSELKUEM	Parents took care about efforts at school
BSNTDEUT	Last grade in German
BSNTMATH	Last grade in math
BSNTFMD1	Last grade in 1. foreign language
BSPTDEUT	Total points in German (last grade)
BSPTMATH	Total points in math (last grade)
BSPTFMD1	Total points in 1. foreign language (last grade)
BSGSDEUT	Level of German at comprehensive school (last grade)
BSGSMATH	Level of math at comprehensive school (last grade)
BSGSFMD1	Level of 1. foreign language at comprehensive school (last grade)
BSLKDEUT	Complementary / main subject in German (last grade)
BSLKMATH	Complementary / main subject in math (last grade)
BSLKFMD1	Complementary / main subject in 1. foreign language (last grade)
	Relationships to Parents, Sport and Activities during Youth
	Frequency of fights as respondent was 15 years old with:
BSSTRVA	Father
BSSTRMU	Mother
BSSPRTTR	Participated in sports during youth
BSSPRTAR	Favourite sport during youth
BSSPRTWE	Participated in competitions during youth
BSMUSSP	Played music or sang during youth
	School Attendance
BSSCHBES	Still at school
BSSCHEND	Year left school
BSSCHWO	Country of last school attendance
BSSCHLA	Federal State of last school attendance
BSKLAUSL	Number of foreign classmates
BSSCHZUK	Strive for further school certificate
BSSCHZAR	Type of further school certificate
	Attained and Planed Vocational Qualification
BSBADABG	Vocational / university degree acquired in Germany
	Type of vocational / university degree attained in Germany:

BSBADLEH	Apprenticeship ("Lehre")
BSBADBFS	Full-time vocational school / School for public health ("Berufsfachschule / Schule des Gesundheitswesens")
BSBADFSC	Technical school, school for master of a trade ("Fachschule, Meister-, Technikerschule")
BSBADBEA	Training for civil servants (officer) ("Beamtenausbildung")
BSBADFHA	Advanced technical college ("Fachhochschule") or approved vocational academy ("anerkannte Berufsakademie")
BSBADUNI	University degree
BSBADSON	Other vocational qualification
BSBADEND	Year of attaining vocational / university degree in Germany
BSBAAABG	Vocational / university degree acquired abroad
	Type of vocational / university degree attained abroad:
BSBAAFAN	Short-term training in a company
BSBAAFBA	Apprenticeship in a company
BSBAASCH	Vocational or professional school
BSBAAUNI	University degree
BSBAASON	Other vocational qualification
BSBAAEND	Year of attaining vocational / university degree abroad
BSBAAZEU	Certificate for abroad attained qualification
BSBAAZEA	Recognition of aboard attained certificate
BSZAJA	Vocational / university degree is aspired
	Type of aspired vocational / university degree:
BSZALEH	Apprenticeship ("Lehre")
BSZABFS	Full-time vocational school/ School for public health ("Berufsfachschule / Schule des Gesundheitswesens")
BSZAFSC	Technical school, school for master of a trade ("Fachschule, Meister-, Technikerschule")
BSZABEA	Training for civil servants (officer) ("Beamtenausbildung")
BSZABAK	Approved vocational academy ("anerkannte Berufsakademie")
BSZAFH	Advanced technical college ("Fachhochschule")
BSZAUNI	University degree
	Military and Voluntary Service
BSDIGEL	Military or alternative service done (only men)
BSDIART	Type of service (only men)
BSDIGRU	Reason for not serving (only men)
BSFSJ	Voluntary social service ("Freiwilliges Soziales Jahr")
	Specification of Interview Situation
BSINTA	Type of interview
BSDAUER1	Duration of personal interview
BSDAUER2	Duration of interview filled out independently
BSTAGIN	Day of the interview
BSMONIN	Month of the interview
BSINTNR	Identifier of the interviewer

3.4.7 Information on Twins and Triplets: *BIOTWIN*

The file *BIOTWIN* contains all twins that were ever identified within the SOEP. To be classified as a twin a person has to: (a) have exactly the same age as his or her sibling, (b) have a relationship to the head of the household that indicates that he or her and a second person are siblings and (c) has to have the identical mother (as far as a pointer to the mother is available).

Furthermore, it is not only twins that are recorded in the *BIOTWIN* data set, but also triplets or quadruple siblings. The following variables are stored within the *BIOTWIN* data set:

Table 3.22: List of variables in *BIOTWIN*

Variable	Meaning
HHNR	Original Household Number
PERSNR	Person Number 1. Sibling
PNRTWIN	Person Number 2. Sibling
PNRTRIP	Person Number 3. Sibling
PNRQUAD	Person Number 4. Sibling
PNRMOTH	Mother's Person Number
BIOMONOZ	Gender Combination Of Siblings

The central variable PERSNR is assigned to the sibling in the group with the lowest personal identifier. PNRTWIN and in rare occasions if available PNRTRIP or even PNRQUAD contains the personal identifier of second, and the third and fourth sibling respectively in the group. This means that every case in the data set consists of a group of twins (or triplets or quadruplets). The code -2 is assigned to PNRTRIP and/or PNRQUAD if a third or fourth sibling does not exist. PERSNR and PNRTWIN however should always contain valid codes.

The variable PNRMOTH shows the link to the mother of the group and is derived from the data set *$KIND*. It is identical to the variable $KMUTTI. The variable BIOMONOZ indicates if the sex of all the siblings is identical and this group therefore might be monozygotic (code 1). If the sex of the siblings differs, this code is set to 0.

The selection of twins within the SOEP for the data set BIOTWIN is based on the month of birth. As a woman might give birth at two different times in a year and social siblings need to be distinguished from biological siblings, the month of birth is of vital importance when determining the twin-groups. Therefore, only people with a) valid month of birth information and b) identical month of birth may be classified as twins. In a second step the relationship of these potential twins to the head of household is scanned ($STELL). If the relationship of both persons assures that they are siblings, then they are assumed to be twins. In a third step the pointer to the mother ($KMUTTI) is checked if available for both siblings to cross check the results of

the previous steps. This pointer is then transferred into the variable `PNRMOTH`. If the variable `$KMUTTI` is unavailable or incomplete the variable `PNRMOTH` is set to 1.

3.4.8 Occupancy and Second Residence: *BIORESID*

In 1994 questions with a focus on occupancy have been introduced in the Biographical Questionnaire asking for the duration of residence in the current dwelling, and on any second residence. Questions on the second residence were also asked before 1994, however, these were collected in the (blue version of the) Individual Questionnaire and therefore the corresponding variables are part of the *$P* files. Only the information surveyed in the Biographical Questionnaire is stored in the file *BIORESID*.

The information for the years 1994 and 1995 stem from the file *BIOLELA*. Information for later years are taken from the wave-specific data sets *$LELA*. In principle, SOEP respondents answer the Biography Questionnaire only once, so every person has only one record with wave-specific information in *BIORESID*. Due to fieldwork related reasons, some very few people have answered the Biography Questionnaire twice. For these cases, the first interview is taken as relevant for *BIORESID*. Further cases are dropped if their information stems from an interview held before the 1994.

The data set *BIORESID* is supplemented every year by new respondents filling in the supplementary Biography Questionnaire.

The information in *BIORESID* is treated as time invariant. Although, in principle, it is possible to update the information on occupancy for some individuals on the basis of more recent information, we abstain from doing so for selectivity reasons.

Table 3.23: List of variables in *BIORESID*

Variable	Meaning
HHNR	Original Household Number
HHNRAKT	Current Wave Household Number
PERSNR	Never Changing Person ID
ERHEBJ	Survey Year, 4-Digit
BRMOVEIN	Year Moved Into Current Residence
BRSECHOM	Second Home
BRSECREG	Second Home In W./E. Germany, Abroad
BRSECUSE	Use Of Second Home
BRSECWOR	Second Home At Place Of Work
BRINTA	Survey Instrument
BRINTNR	Interviewer Number

3.4.9 "Mother and Child" Questionnaire: *BIOCHILD*

Since 2003, starting with the birth cohort 2002/03, questions regarding the birth of a child have been integrated in the SOEP with the help of the "Mother and Child" questionnaire. The questionnaire is aimed at all women who, in the current survey year or the year before, gave birth to a child, as well as at women whose non-biological child was born in the time period mentioned above. This means that at the time of the survey, these children are either newborn or they have a maximum age of one and a half years.

The questionnaire is comprised of 19 questions covering four subjects: (a) pregnancy (b) body measurements and health of the child (c) change in living circumstances due to the birth of the child (d) circumstances surrounding the care of the child.

In the future, this additional questionnaire will be used for newborn cohorts on an annual basis. There are plans to follow and record the childrens development by surveying them with additional brief questionnaires at certain points in time.

The aim of the *BIOCHILD* data file is to observe the future generation of SOEP respondents, preferably from birth onwards. The data set is in the order of the unchanging person ID of the child, so that information from the *BIOCHILD* data set can be directly linked to the child files (*$KIND*). From a mothers point of view, the data set presents a detailed source of information on pregnancy and the changes in life experienced by women who have recently become mothers. The basis for *BIOCHILD* therefore consists of all individuals who are identified as children in the Mother and Child Questionnaire (`PERSNR`). With the help of the mothers unchanging person ID, information on the mothers in this data file can also be directly linked to the individual information on the mothers.

In the case of multiple children, the mother fills out the respective number of questionnaires, so that siblings can be identified by the identical unchanging person ID of the mother (`PERSNRM`), month of birth, and year of birth. The variables in *BIOCHILD* correspond to their structure in the user-friendly original variables from the "Mother and Child" questionnaire. Information is provided on the time and place of birth, as well as the childs height and weight at birth. The variable `BCKALTER` gives the age of the child at the time of the survey in months. With regard to the children, there is information on any disorders in their development, as well as the health of the child in its first three months of life. *BIOCHILD* also includes information on whether or not the pregnancy was planned, if the child is the first, and if this is a biological child. For mothers with a non-biological child, the questions are presented in an identical fashion, but questions on pregnancy are left out. Finally, there are questions on the mother's evaluation of the child's behaviour at the time

of the interview. The variables `BCVERAE1` to 8 refer to the mothers assessment of the new circumstances of life. *BIOCHILD* also contains information on the mother's current personal situation, i.e. whether there is a partner present, as well as the extent to which the child is cared for by people other than the main care provider.

Table 3.24: List of variables in *BIOCHILD*

Variable	Meaning
HHNR	Original Household Number
HHNRAKT	Current Wave HH Number
PERSNR	Never Changing Person ID Child
PERSNRM	Never Changing Person ID Mother
ERHEBJ	Survey Year
BCKGEBMO	Child - Month Of Birth
BCKGEBJA	Child - Year Of Birth
BCKALTER	Age Of Child At Time Of Survey
BCENTBIN	Place Where Delivery Took Place
BCSSW	Birth In Which Pregnancy Week
BCKGEW	Child - Weight At Birth In Grams
BCKGROE	Child - Height In Cm
BCKKOPF	Child - Head Circumference In Cm
BCKLETZU	Most Recent Child Med. Examination
BCKSTOER	Child Has Confirmed Disorder
BCKARZT	Child - Medical Assistance: Number Of Times
BCKKRHAU	Child - Hospital Stays First 3 Months In Days
BCKIZAHL	Newborn Is x Child
BCKLEIBL	Biological Child
BCSSPLAN	Planned/Unplanned Pregnancy
BCVATER	Father Lives In Household
BCUNTPA	Support From Partner
BCHAUPTB	Mother Is Main Person Providing Care
BCBEFIN1	Physical State: Last Third Of Pregnancy
BCBEFIN2	Physical State: First 3 Months After Birth
BCBEFIN3	Mental State: Last Third Of Pregnancy
BCBEFIN4	Mental State: First 3 Months After Birth
BCVERAE1	Life Circumstances Have Greatly Changed
BCVERAE2	Bringing Up Child Provides Happiness
BCVERAE3	Often At The End Of My Strength
BCVERAE4	Role As Mother Is Satisfying
BCVERAE5	Feel Overdemanded
BCVERAE6	Met New People Through Child
BCVERAE7	Role As Mother Is Limiting
BCVERAE8	Important To Provide Tenderness
BCBETRE1	Care From Spouse/Partner: Hrs Per Wk
BCBETRE2	Care From Grandparents: Hrs Per Wk
BCBETRE3	Care From Older Siblings: Hrs Per Wk
BCBETRE4	Care From Other Relatives: Hrs Per Wk
BCBETRE5	Care From Child Minder: Hrs Per Wk
BCBETRE6	Care From Creche: Hrs Per Wk
BCBETRE7	Care From Other Individuals: Hrs Per Wk
BCKGESU1	Worried About Child's Health
BCKGESU2	Child Is Generally Happy And Satisfied
BCKGESU3	Child Is Irritable/Crys Frequently
BCKGESU4	Child Is Hard To Console
BCKGESU5	Child Is Curious/Active
BCKSEX	Gender Of Child
BCBETRE8	No Other Care Apart From Mother

Chapter 4

Introduction to Data Retrievals

by John P. Haisken-DeNew

The intent of this chapter is to give the reader a short but helpful introduction to actually running a retrieval in the popular statistics packages found in universities. We will concentrate on SAS [8.x], SPSS[10.x], Stata [9.0].

A general introduction in matching will be given, followed by detailed examples in the various packages. The point here is that the user, after reading this chapter, should be able to at least run a rudimentary retrieval himself/herself. There are of course many, many details of the packages which cannot even be mentioned here, but assuming some basic knowledge of computer packages, it should be possible to "get started" quickly.

4.1 Matching & Data Files

The SOEP distribution, whether the German or the international Scientific Use Version, comes with a hierarchical structure. That is to say, the data are not in a single flat rectangular file, as is often the case with many datasets. The American PSID used to have this, but at some point, the physical limit with respect to the number of variables that can fit in a data file is reached. In fact, the PSID has recently changed its data structure, and it looks more like the SOEP ! Thus from the very beginning, the SOEP has been hierarchical. For every year/wave, and for every level of analysis, such as person, household, foreigner, generated, etc., there is a separate file. For example the file *AP* is from the year 1984 ("A") and is at the person level ("P"). The question which remains, is how to match information for an individual (for example)

over several years and over several data files (levels of analysis).

Table 4.1 lists data files, and matching strategies.

Thus all different record types can be matched together at the person level, using a matching strategy as described in Table 4.1. For example, one starts off with the master file at the person level *PPFAD*. Successively, all wave specific person information such as *AP*,..., *UP*, and *APGEN*,..., *UPGEN* etc., is matched to the master file. If household information is desired, this can also be matched. Thus, *AH*,..., *UH*, and *AHGEN*,..., *UHGEN* for example can be matched.

The data is setup such that the person ID or `PERSNR` is unique over time and individuals. However this is not necessarily true for the household identifier. Individuals can leave households, and found new ones. (Think of a son or daughter moving out of an existing household after school.) Thus matching at the household level is done with the *current* household ID or `UHHNR`. Matching of person level information to the person level master file is always done with the unique person identifier or `PERSNR`.

Table 4.1: Data Files and Matching

Filename	Description	Sort 1	Sort 2	Sort 3
PPFAD	Master Person File	PERSNR	-	
PHRF	Person-Level Weighting	PERSNR	-	
HPFAD	Master Household File	HHNR	HHNRAKT*	-
HHRF	Household-Level Weighting	HHNR	HHNRAKT*	-
$PBRUTTO	Person Gross File	PERSNR		
$P	Person	PERSNR		
$PAUSL	Person (Foreigner Only)	PERSNR		
$PGEN	Person Generated	PERSNR		
$PKAL	Person Calendar	PERSNR		
$PLUECKE	Person Gap	PERSNR		
$POST	Person (East Germans Only)	PERSNR		
$KIND	Person (Children)	PERSNR		
$PEQUIV	Person (adult and children)	PERSNR		
$HBRUTTO	Household Gross File	HHNR	$HHNR	-
$H	Household	HHNR	$HHNR	-
$HGEN	Household Generated	HHNR	$HHNR	-
GHOST	Household (East Germans Only)	HHNR	$HHNR	-
YPBRUTTO	Drop-Outs	PERSNR	ERHEBJ	
PFLEGE	Invalidity / Care	ERHEBJ		
ARTKALEN	Activity Calendar Monthly	PERSNR	SPELLNR	
EINKALEN	Income Calendar Monthly	PERSNR	SPELLNR	
PBIOSPE	Activity Calendar Yearly	PERSNR	SPELLNR	
BIOMARSM	Bio Marital Status Monthly	PERSNR	SPELLNR	
BIOMARSY	Bio Marital Status Yearly	PERSNR	SPELLNR	
SOZKALEN	Welfare	HHNR	HHNRAKT*	SPELLNR
BIOBIRTH	Bio Birth History of Women	PERSNR		
BIOBRTHM	Bio Birth History (from Men)	PERSNR		
BIOCHILD	Bio Child	PERSNR		
BIOCHILD	Bio Twin/Triplet/Quad	PERSNR		
BIOPAREN	Bio Parental Info	PERSNR		
BIOJOB	Bio First Job	PERSNR		
BIOIMMIG	Bio Migration	PERSNR	ERHEBJ	
BIOYOUTH	Bio Youth/Adolecents	PERSNR		
BIOSOC	Bio Socialization	PERSNR		
BIORESID	Bio 2nd Residency	PERSNR		

*Note: HHNRAKT must not be replaced by $HHNR in *SOZKALEN*, *HPFAD* and *HHRF*.

4.2 A Simple SOEP Retrieval

In this trivial example, we are interested in information from 1984 and 1985 at the person and household level. Below is a listing of those variables mentioned.

VARNAME	FILE	YEAR	LABEL
AP06	AP	1984	School-Leaving Degree
BP16	BP	1985	Employment Status
AH02	AH	1984	Change in HH comp Since Jan 1st Prev yr
BH01	BH	1985	Children under age 16 in HH

Looking at the filenames for the various variables that are being pulled out of the datasets, there are obviously various levels of data: person level data such as *AP* and *BP*, and household level data such as *AH* and *BH*. With different levels of data, different or rather "appropriate" matching strategies must be used.

The following sections will deal with the practical implementation of the above mentioned problem in SAS, SPSS and Stata, while giving some background information on the statistical packages themselves.

It is completely up to the individual user to select a statistical package that is appropriate to the analysis at hand. There are advantages and disadvantages to all statistical packages, however Stata seems currently to have more advantages in the area of pre-programmed panel estimators. This may be especially useful for novice programmers. See Table 4.2 for more information.

4.3 Stata [9.0]

A basic outline will be examined for a STATA [9.0] implementation. As of Wave M (1996), STATA binary (DTA) files are delivered as part of the regular data distribution, due to STATA's growing popularity in the economics community.

The same basic structure will be followed as for other packages, with some notable exceptions. STATA [7.0] namely does *not* allow selecting variables on merging, forcing the user to first make temporary files using the variables desired from each data file, and then merging these temporary files to the master file (*PPFAD* or *HPFAD*). As an aside, SPSS and SAS allow selecting variables and merging all in the same command. Another difference is in the way STATA handles memory. Thus in the following example, some variables will be taken out of *APGEN* and stored in a temporary file in an other working directory under the same filename with the `use ... using ...` command. Then the memory is cleared to allow moving on to the next step.

Table 4.2: Comparing Statistical Packages for SOEP

Feature	Stata [9.0]	SPSS [10.0]	SAS [8.2]
Used by SOEP Researchers	Yes	Yes	Yes
SOEP Workshop Schooling	Yes	Yes	Yes
Binary File Compatibility	Win/Unix	Win/Unix	Win/Unix
SOEP File Size On Disk (2003)	1 GB	1 GB	1.5 GB
Export Files	ASCII, XPT	POR, ASCII	XPT, ASCII
Platform	Win/Unix	Win, Unix limits & only SPSS[6.2]	Win/Unix
Programming Language	Yes	Yes	Yes
Matrix Language	Yes	Yes	Yes
Extensive Panel Estimators	Yes	Very Little	Some
Variable Labels	in data file	in data file	in data file
Value Labels	in data file	in data file	in *formats.sas7bcat*
Memory Management	attempts to load all data	piece-wise	piece-wise
One Step Pick and Merge	No	Yes	Yes

```
#delimit ;
set memory 40m;

use   hhnr ahhnr persnr
      apsbil apbbil01 apbbil02 apbbil03 apsbila apbbila
      afamstd abilzeit aerwzeit
using d:\stata\apgen;

sort  persnr ;
save  temp\apgen,replace ;
clear;
```

STATA also allows full cumulative merging, that is to say, if the user runs

```
use   ppfad;
keep  if lnetto==0;
sort  persnr;

merge persnr
using lp;

drop  if _merge==2;
drop  _merge;
```

then STATA creates a dataset having all observations that ever appeared in *PPFAD* AND/OR in *LP*. On merging STATA creates a merge status variable called `_merge`, and when this equals 2, this means that the observation came only from the `using` file, and thus should be discarded (otherwise you will have many many observations with system missings). Finally the variable `_merge` is dropped from the dataset, otherwise on the next merge, STATA will try to create the variable again, and will fail with an error message.

4.3.1 Person-Level Retrievals

In the following example, a person-level panel retrieval is done, using *PPFAD* (of course), several *$P*'s, *$PGEN*'s, and *$H*'s. The result is a rectangular file of a size determined by the selections made in *PPFAD* (rows) and the number of variables (columns) chosen from the regular data files along the way.

After merging, each temporary file is deleted to conserve disk space. **Warning: Do not forget to write the temporary file to another directory, otherwise you will write over your input file, and destroy it's original contents !** Stata differentiates between upper and lower case letters in the variable names. All variables in the SOEP distribution have been stored in *lower case*, and as such, must be referred to in lower case !

There are of course many ways to write retrievals. It is not claimed or even suggested that the ways described here are "optimal". If you find other ways that you feel more comfortable with, then use them instead.

```
/* ---------------------------------------------------------- */
/*  This command file was generated by SOEPINFO-WWW           */
/*  from the SOEP HOMEPAGE: http://www.diw-berlin.de/soep.    */
/*                                                            */
/*      !!! I M P O R T A N T  -  W A R N I N G !!!           */
/*  You alone are responsible for contents and appropriate    */
```

```
/*  usage by accepting the usage agreement.                      */
/* ------------------------------------------------------------ */
/*  To copy this command file to your own harddisk:             */
/*  Use your mouse or (CNTL-A) to select this text, then "copy" */
/*  (CNTL-C) then in your windows application, or in a text     */
/*  editor, "paste" (CNTL-V).                                   */
/* ------------------------------------------------------------ */
/*  Please report any errors of the STATA code generated here   */
/*  to Jan Goebel: jgoebel@diw.de                               */
/* ------------------------------------------------------------ */
/*                                                              */

#delimit;
log       using  temp/new, text replace;
set       more off;

/*        ----[ automatically pull PPFAD ]----------------------*/;

use       hhnr persnr sex gebjahr psample
          ahhnr bhhnr
          anetto bnetto
using     soep/ppfad;

/*        ----[ Data Structure and Unit of Analysis ]---*/;
keep      if anetto == 1 & bnetto == 1;

/*        ----[ Gender ]---------------------------------*/;
/*        -- male    = 1                                --*/;
/*        -- female  = 2                                --*/;
keep      if  ((psample == 1) & (sex == 1));

sort      persnr;
save      temp/ppfad, replace;
clear;

/*        ----[ automatically pull PHRF ]---------------- */;

use       hhnr   persnr  prgroup
          aphrf  bphrf
                 bpbleib
using     soep/phrf;

/*        ----[ tips for longitudinal weights ]------------------- */;
/*        create your own LONGITUDINAL person weights here.        */;
/*        e.g. longitudinal person weight from wave A to wave D.   */;
/*        take the starting wave cross-sectional weight (aphrf)    */;
/*        and multiply through by each FOLLOWING WAVE staying      */;
/*        factor, as in the following example:                     */;
/*        gen  adphrf=aphrf*bpbleib*cpbleib*dpbleib;               */;
/*        ---------------------------------------------------------- */;

sort      persnr;
save      temp/phrf, replace;
clear;


/*        ----[ automatically create PMASTER ]---------- */;

use       temp/ppfad;
merge     hhnr persnr
using     temp/phrf;
drop       if _merge == 2;
drop       _merge;
erase     temp/ppfad.dta;
erase     temp/phrf.dta;

sort      persnr;
save      temp/pmaster, replace;


/*        ----( pull AP )--------------------------- */;
```

```
use       hhnr ahhnr persnr
          ap06
using     soep/ap;
sort      persnr;
save      temp/ap, replace;
clear;

/*        ----( pull BP )---------------------------- */;

use       hhnr bhhnr persnr
          bp16
using     soep/bp;
sort      persnr;
save      temp/bp, replace;
clear;


/*        ----( pull AH )------------------------- */

use       hhnr ahhnr
          ah02
using     soep/ah;
sort      hhnr ahhnr;
save      temp/ah, replace;
clear;


/*        ----( pull BH )------------------------- */

use       hhnr bhhnr
          bh01
using     soep/bh;
sort      hhnr bhhnr;
save      temp/bh, replace;
clear;


/*        ----( merge together by person: ALL Waves )---------- */;

use       temp/pmaster;
erase     temp/pmaster.dta;

/*        ----( merge AP )----------------------------------- */;

sort      persnr;
merge     persnr
using     temp/ap;
drop      if _merge == 2;
drop      _merge;
erase     temp/ap.dta;


/*        ----( merge BP )----------------------------------- */;

sort      persnr;
merge     persnr
using     temp/bp;
drop      if _merge == 2;
drop      _merge;
erase     temp/bp.dta;


/*        ----( merge AH )----------------------------------- */;

sort      hhnr ahhnr;
merge     hhnr ahhnr
using     temp/ah;
drop      if _merge == 2;
drop      _merge;
erase     temp/ah.dta;
```

```
/*        ----( merge BH )------------------------------------ */;

sort      hhnr bhhnr;
merge     hhnr bhhnr
using     temp/bh;
drop      if _merge == 2;
drop      _merge;
erase     temp/bh.dta;


/*        ----( Done ! )-------------------------------------- */;

label     data "SOEPINFO: Magic at Work! http://www.diw.de/soep";
save      temp/new.dta, replace;
desc;
log       close;
```

4.3.2 Household-Level Retrievals

```
/* ------------------------------------------------------------ */
/*  This command file was generated by SOEPINFO-WWW             */
/*  from the SOEP HOMEPAGE: http://www.diw-berlin.de/soep.      */
/*                                                              */
/*      !!! I M P O R T A N T  -  W A R N I N G !!!             */
/*  You alone are responsible for contents and appropriate      */
/*  usage by accepting the usage agreement.                     */
/* ------------------------------------------------------------ */
/*  To copy this command file to your own harddisk:             */
/*  Use your mouse or (CNTL-A) to select this text, then "copy" */
/*  (CNTL-C) then in your windows application, or in a text     */
/*  editor, "paste" (CNTL-V).                                   */
/* ------------------------------------------------------------ */
/*  Please report any errors of the STATA code generated here   */
/*  to Jan Goebel: jgoebel@diw.de                               */
/* ------------------------------------------------------------ */
/*                                                              */

#delimit;
log       using  temp/new, text replace;
set       more off;

/*        ----[ automatically pull HPFAD ]----------------------*/;

use       hhnr     hhnrakt  hsample
          ahhnr    bhhnr
          ahnetto  bhnetto
using     soep/hpfad;

/*        ----[ Data Structure and Unit of Analysis ]-----------*/;
keep      if ahnetto == 1 & bhnetto == 1;

/*        ----[ Define Sample/Region ]--------------------------*/;
keep      if  (hsample == 1);

sort      hhnr hhnrakt;
save      temp/hpfad, replace;
clear;

/*        ----[ automatically pull HHRF ]---------------------- */;

use       hhnr   hhnrakt  hrgroup
          ahhrf  bhhrf
using     soep/hhrf;

sort      hhnr hhnrakt;
```

```
save      temp/hhrf, replace;
clear;


/*        ----[ automatically create HMASTER ]---------------- */;

use       temp/hpfad;
merge     hhnr hhnrakt
using     temp/hhrf;
drop       if _merge == 2;
drop       _merge;

drop      hhnrakt;
erase     temp/hpfad.dta;
erase     temp/hhrf.dta;

sort      hhnr ahhnr;
save      temp/hmaster, replace;


/*        ----( pull AH )------------------------- */

use       hhnr ahhnr
          ah02
using     soep/ah;
sort      hhnr ahhnr;
save      temp/ah, replace;
clear;


/*        ----( pull BH )------------------------- */

use       hhnr bhhnr
          bh01
using     soep/bh;
sort      hhnr bhhnr;
save      temp/bh, replace;
clear;


/*        ----( merge together by household: ALL Waves )---------- */;

use       temp/hmaster;
erase     temp/hmaster.dta;


/*        ----( merge AH )------------------------------------- */;

sort      hhnr ahhnr;
merge     hhnr ahhnr
using     temp/ah;
drop      if _merge == 2;
drop      _merge;
erase     temp/ah.dta;


/*        ----( merge BH )------------------------------------- */;

sort      hhnr bhhnr;
merge     hhnr bhhnr
using     temp/bh;
drop      if _merge == 2;
drop      _merge;
erase     temp/bh.dta;


/*        ----( Done ! )--------------------------------------- */;

label     data "SOEPINFO: Magic at Work! http://www.diw.de/soep";
save      temp/new.dta, replace;
desc;
log       close;
```

4.4 SPSS [10.X]

The following is a possibility for programming the retrieval in SPSS. It is not suggested that this method is particularly efficient, however it works reasonably well. It follows the structure of a STATA retrieval very closely. First the master file *PPFAD* or *HPFAD* is opened and the sample is defined by the `UNETTO` variables and for example `SEX` and/or `PSAMPLE`, saving this result in a temporary file, in a temporary work area. Then, appropriate weighting factors are selected and merged. Following this, the regular data retrieval process starts, such that in the first step, only those variables that are required are taken out of the appropriate data files and stored temporarily in a temporary work area. After all data has been pull out, then the temporary work files are matched to the master file one by one. The result is a rectangular file with the person as the unit of observation. The data set consists of columns of variables. It is of course possible to match household level information to person-level retrievals, whereby the sorting and matching is made on `HHNR` and `$HHNR`. At the end, all *temporary work files* are discarded as they are no longer required.

4.4.1 Person-Level Retrievals with HH-Info

```
*                                                                .
*       ------------------------------------------------------------ .
*        Please report any errors of the SPSS code generated here    .
*        to Jan Goebel: jgoebel@diw.de                               .
*       ------------------------------------------------------------ .
*         WARNING: SPSS for UNIX only allows up to 99 files to       .
*                  be accessed in a single retrieval. This is hard   .
*                  coded by SPSS and we cannot change this. You may  .
*                  have to run your retrieval in smaller blocks.     .
*       ------------------------------------------------------------ .
*                                                                    .

set     compression on.
set     header off.

*        -------------[ automatically pull PPFAD ]----------------.

get      file        = 'soep/ppfad.sav'
         /keep       = hhnr     persnr  sex   gebjahr  psample
                       ahhnr    bhhnr
                       anetto   bnetto.

*                      [ balanced / unbalanced design ]-----------.
select   if           ((anetto eq 1) and (bnetto eq 1)).

*                      [ define sample / region / gender ]--------.
select   if           ((psample eq 1) and (sex eq 1)).

sort     cases  by     persnr.
save     outfile     = 'temp/ppfad.sav'.
```

```
*        -------------[ automatically pull PHRF ]----------------.

get      file        = 'soep/phrf.sav'
         /keep       = hhnr   persnr  prgroup
                       aphrf  bphrf
                              bpbleib.

*        -------------[ tips for longitudinal weights ]----------.
*        create your own LONGITUDINAL person weights here.
*        e.g. longitudinal person weight from wave A to wave D.
*        take the starting wave cross-sectional weight (aphrf)
*        and multiply through by each FOLLOWING WAVE staying
*        factor, as in the following example:
*compute adphrf=aphrf*bpbleib*cpbleib*dpbleib.
*        ------------------------------------------------------.

sort     cases  by     persnr.
save     outfile     = 'temp/phrf.sav'.


*        -------------[ automatically create PMASTER ]----------.

match    files file  = 'temp/ppfad.sav'
              /table = 'temp/phrf.sav'
              /by      hhnr persnr.

sort     cases  by     persnr.
save     outfile     = 'temp/pmaster.sav'.


*        -------------( pull AP )-----------------------------.

get      file        = 'soep/ap.sav'
         /keep       = hhnr ahhnr persnr ap06.

sort     cases  by     persnr.
save     outfile     = 'temp/ap.sav'.


*        -------------( pull BP )-----------------------------.

get      file        = 'soep/bp.sav'
         /keep       = hhnr bhhnr persnr bp16.

sort     cases  by     persnr.
save     outfile     = 'temp/bp.sav'.


*        -------------( pull AH )-----------------------------.

get      file        = 'soep/ah.sav'
         /keep       = hhnr ahhnr ah02.

sort     cases  by     hhnr ahhnr.
save     outfile     = 'temp/ah.sav'.


*        -------------( pull BH )-----------------------------.

get      file        = 'soep/bh.sav'
         /keep       = hhnr bhhnr bh01.

sort     cases  by     hhnr bhhnr.
save     outfile     = 'temp/bh.sav'.


*        -------------( merge together by person: ALL Waves )----.

match    files file  = 'temp/pmaster.sav'
              /table = 'temp/ap.sav'
```

```
             /by      persnr.

match    files file  = *
             /table = 'temp/bp.sav'
             /by      persnr.


*        -------------( merge together by household: Wave A )----.
*        Just One $HHNR sort for all HH files per WAVE.

sort     cases  by     hhnr ahhnr.
match    files file  = *
             /table = 'temp/ah.sav'
             /by      hhnr ahhnr.


*        -------------( merge together by household: Wave B )-----.
*        Just One $HHNR sort for all HH files per WAVE.

sort     cases  by     hhnr bhhnr.
match    files file  = *
             /table = 'temp/bh.sav'
             /by      hhnr bhhnr.


*        -------------( save output file )-------------------------------.

file     label       "SOEPINFO: Magic at Work! http://www.diw.de/soep".
save     outfile     = 'temp/new.sav'.
desc     all.


*        -------------( discard temporary files )------------------------.

erase    file        = 'temp/pmaster.sav'.
erase    file        = 'temp/phrf.sav'.
erase    file        = 'temp/ppfad.sav'.
erase    file        = 'temp/ap.sav'.
erase    file        = 'temp/bp.sav'.
erase    file        = 'temp/ah.sav'.
erase    file        = 'temp/bh.sav'.

finish.
```

4.4.2 Household-Level Retrievals

Here we follow a similar procedure as used with person-level retrievals. Instead here we use *HPFAD* as the master file. and the variables `$HNETTO` and `HSAMPLE` to define the sample. Matching and sorting of household level files in done by `HHNR` and `$HHNR`. The only main exception to this rule is the matching of *HHRF* which **must** be done by `HHNRAKT`. At the end, all temporary work files are discarded.

```
*                                                                  .
*        ---------------------------------------------------------- .
*         Please report any errors of the SPSS code generated here    .
*         to Jan Goebel: jgoebel@diw.de                               .
*        ---------------------------------------------------------- .
*          WARNING: SPSS for UNIX only allows up to 99 files to       .
*                   be accessed in a single retrieval. This is hard   .
*                   coded by SPSS and we cannot change this. You may  .
```

```
*                  have to run your retrieval in smaller blocks.     .
*       ------------------------------------------------------------ .
*                                                                     .

set     compression on.
set     header off.

*       -------------[ automatically pull HPFAD ]-------------------.

get     file        = 'soep/hpfad.sav'
        /keep       = hhnr    hhnrakt   hsample
                      ahhnr   bhhnr
                      ahnetto bhnetto.

*                     [ balanced / unbalanced design ]----------.
select  if            (ahnetto eq 1) and (bhnetto eq 1).

*                     [ define sample / region ]----------------.
select  if            (hsample eq 1).

sort    cases  by     hhnr hhnrakt.
save    outfile     = 'temp/hpfad.sav'.


*       -------------[ automatically pull HHRF ]----------------.

get     file        = 'soep/hhrf.sav'
        /keep       = hhnr   hhnrakt  hrgroup
                      ahhrf  bhhrf.

sort    cases  by     hhnr hhnrakt.
save    outfile     = 'temp/hhrf.sav'.


*       -------------[ automatically create HMASTER ]----------.

match   files file  = 'temp/hpfad.sav'
             /table = 'temp/hhrf.sav'
             /by      hhnr hhnrakt.

sort    cases  by     hhnr ahhnr.
save    outfile     = 'temp/hmaster.sav' /drop=hhnrakt.


*       -------------( pull AH )-------------------------------.

get     file        = 'soep/ah.sav'
        /keep       = hhnr ahhnr ah02.

sort    cases  by     hhnr ahhnr.
save    outfile     = 'temp/ah.sav'.


*       -------------( pull BH )-------------------------------.

get     file        = 'soep/bh.sav'
        /keep       = hhnr bhhnr bh01.

sort    cases  by     hhnr bhhnr.
save    outfile     = 'temp/bh.sav'.



*       -------------( merge together by household )----------------.

match   files file  = 'temp/hmaster.sav'
             /table = 'temp/ah.sav'
             /by      hhnr ahhnr.


sort    cases  by     hhnr bhhnr.
match   files file  = *
```

```
             /table = 'temp/bh.sav'
             /by      hhnr bhhnr.

*        -------------( save output file )------------------------------.

file     label       "SOEPINFO: Magic at Work! http://www.diw.de/soep".
save     outfile     = 'temp/new.sav'.
desc     all.

*        -------------( discard temporary files )-----------------------.
erase    file        = 'temp/hmaster.sav'.
erase    file        = 'temp/hpfad.sav'.
erase    file        = 'temp/hhrf.sav'.
erase    file        = 'temp/ah.sav'.
erase    file        = 'temp/bh.sav'.

finish.
```

4.5 SAS [8.X]

SAS is a relatively standard package available at most universities. It is known for its advanced data handling facilities, which will come in very handy for those wishing to use SAS for SOEP retrievals. The first step in a SOEP retrieval is to define where the SOEP data are. For instance:

```
libname soep     'c:\gsoep17';
libname library  'c:\gsoep17';
libname mywork   'c:\temp';
```

The `libname soep` statement refers to the directory where the data are kept.

The `libname library` statement refers to the directory where the *formats.sas7bcat* file is located. This single file contains all value label information for the entire SOEP data distribution.

Further, if formats have been defined, and the formats file cannot be found, SAS will give an error message unless this has been specifically turned off with the following command:

```
options nofmterr;
```

The `libname mywork` statement defines a working area where temporary and work files are placed.

4.5.1 Person-Level Retrievals

The method used here in the SAS retrieval is that of the "omnibus" (or "milk-run") concept. The main file *PPFAD* is opened, and the most important matching variables are pulled out, i.e. the person ID `PERSNR` and the yearly

household IDs AHHNR-UHHNR. As well, the netto variables ANETTO-UNETTO allow one immediately to select a sample, restricted to those that actually took part in the survey in the particular years selected. It happens in many cases of course that a person was absent for one or more waves, or only started answering the questionnaire as "teenager" that just became an "adult", etc. Further, sample selections are made according to gender, sample and region of residence. Obviously it makes sense to select the sample at the beginning rather than at the end, thereby saving memory and temporary file disk space. The *PPFAD* data are then sorted by PERSNR to allow merging at the next step of the retrieval. Here all person level files are accessed and then all household level files. This is done to save on unnecessary sorting.

```
* ------------------------------------------------------------ *;
*  This command file was generated by SOEPINFO-WWW             *;
*  from the SOEP HOMEPAGE: http://www.diw-berlin.de/soep.      *;
*                                                              *;
*      !!! I M P O R T A N T  -  W A R N I N G !!!             *;
*  You alone are responsible for contents and appropriate      *;
*  usage by accepting the usage agreement.                     *;
* ------------------------------------------------------------ *;
*  To copy this command file to your own harddisk:             *;
*  Use your mouse or (CNTL-A) to select this text, then "copy" *;
*  (CNTL-C) then in your windows application, or in a text     *;
*  editor, "paste" (CNTL-V).                                   *;
* ------------------------------------------------------------ *;
*  Please report any errors of the SAS code generated here     *;
*  to Jan Goebel: jgoebel@diw.de                               *;
* ------------------------------------------------------------ *;

libname  soep    'soep';
libname  library 'soep';
libname  mywork  'temp';
options  compress=no ls=80 errors=1 nofmterr nodate nocenter ;


*        ----[ automatically pull PPFAD ]------------------*;

data     mywork.new;
         set soep.ppfad(keep=
             hhnr persnr sex gebjahr psample
             ahhnr bhhnr
             anetto bnetto);

*        ----[ data structure and unit of analysis ]------*;
if       anetto = 1 and  bnetto = 1;

*        ----[ gender ] ----------------------------------*;
*        ----- male    = 1                                 *;
*        ----- female  = 2                                 *;
if       psample = 1 and sex = 1;

run;
proc     sort; by persnr;
run;

*        ----[ automatically pull phrf ]-----------------*;

*        ----[ tips for longitudinal weights ]------------------- *;
*        create your own LONGITUDINAL person weights after        *;
*        the following SET SOEP.PHRF statement.                   *;
*        e.g. longitudinal person weight from wave A to wave D.   *;
*        take the starting wave cross-sectional weight (aphrf)    *;
*        and multiply through by each FOLLOWING WAVE staying      *;
*        factor, as in the following example:                     *;
```

```
*       adphrf=aphrf*bpbleib*cpbleib*dpbleib;                    *;
*       -------------------------------------------------------- *;

data    phrf;
  set   soep.phrf (keep=
        hhnr persnr prgroup
             aphrf  bphrf
                    bpbleib);
proc    sort; by persnr;
run;

data    mywork.new;
        merge mywork.new (in=present) phrf;
        by persnr;
        if present;
run;

proc    sort; by hhnr persnr;
run;


*       ----( pull/merge ap )--------------------------------*;

data    ap;
  set   soep.ap (keep=
        hhnr persnr ap06);
proc    sort; by persnr;
run;

data    mywork.new;
        merge mywork.new (in=present) ap;
        by persnr;
        if present;
run;


*       ----( pull/merge bp )--------------------------------*;

data    bp;
  set   soep.bp (keep=
        hhnr persnr bp16);
proc    sort; by persnr;
run;

data    mywork.new;
        merge mywork.new (in=present) bp;
        by persnr;
        if present;
run;

*       -------------------------------------------------------*;

proc    sort; by hhnr ahhnr;
run;


*       ----( pull/merge ah )--------------------------------*;

data    ah;
  set   soep.ah (keep=
        hhnr ahhnr ah02);
proc    sort; by hhnr ahhnr;
run;

data    mywork.new;
        merge mywork.new (in=present) ah;
        by hhnr ahhnr;
        if present;
run;

*       -------------------------------------------------------*;
```

```
proc     sort; by hhnr bhhnr;
run;


*        ----( pull/merge bh )---------------------------------*;

data     bh;
 set     soep.bh (keep=
         hhnr bhhnr bh01);
proc     sort; by hhnr bhhnr;
run;

data     mywork.new;
         merge mywork.new (in=present) bh;
         by hhnr bhhnr;
         if present;
run;


*        ----( Done ! )------------------------------------ *;

proc     contents;
proc     means;
run;
```

4.5.2 Household-Level Retrievals

Here we follow a similar procedure as used with person-level retrievals. Instead here we use *HPFAD* as the master file. and the variables `$HNETTO` and `HSAMPLE` to define the sample. Matching and sorting of household level files in done by `HHNR` and `$HHNR`. The only main exception to this rule is the matching of *HHRF* which **must** be done by `HHNRAKT`. At the end, all temporary work files are discarded.

```
* ------------------------------------------------------------ *;
*  This command file was generated by SOEPINFO-WWW             *;
*  from the SOEP HOMEPAGE: http://www.diw-berlin.de/soep.      *;
*                                                              *;
*      !!! I M P O R T A N T  -  W A R N I N G !!!             *;
*  You alone are responsible for contents and appropriate      *;
*  usage by accepting the usage agreement.                     *;
* ------------------------------------------------------------ *;
*  To copy this command file to your own harddisk:             *;
*  Use your mouse or (CNTL-A) to select this text, then "copy" *;
*  (CNTL-C) then in your windows application, or in a text     *;
*  editor, "paste" (CNTL-V).                                   *;
* ------------------------------------------------------------ *;
*  Please report any errors of the SAS code generated here     *;
*  to Jan Goebel: jgoebel@diw.de                               *;
* ------------------------------------------------------------ *;

libname  soep    'soep';
libname  library 'soep';
libname  mywork  'temp';
options  compress=no ls=80 errors=1 nofmterr nodate nocenter ;


*        ----[ automatically pull HPFAD ]------------------*;

data     mywork.new;
         set soep.hpfad(keep=
```

```
            hhnr hhnrakt hsample
            ahhnr bhhnr
            ahnetto bhnetto);

*       ----[ data structure and unit of analysis ]------*;
if      ahnetto = 1 and  bhnetto = 1;

*       ----[ region / sample ]---------------------------*;
if      hsample = 1;

run;
proc    sort; by hhnr hhnrakt;
run;


*       ----[ automatically pull hhrf ]-----------------*;

data    hhrf;
  set   soep.hhrf (keep=
        hhnr hhnrakt hrgroup
             ahhrf  bhhrf);
proc    sort; by hhnr hhnrakt;
run;

data    mywork.new;
        merge mywork.new (in=present) hhrf;
        by hhnr hhnrakt;
        if present;
run;

proc    sort; by hhnr ahhnr;
run;


*       ------------------------------------------------------*;

proc    sort; by hhnr ahhnr;
run;


*       ----( pull/merge ah )--------------------------------*;

data    ah;
  set   soep.ah (keep=
        hhnr ahhnr ah02);
proc    sort; by hhnr ahhnr;
run;

data    mywork.new;
        merge mywork.new (in=present) ah;
        by hhnr ahhnr;
        if present;
run;

*       ------------------------------------------------------*;

proc    sort; by hhnr bhhnr;
run;


*       ----( pull/merge bh )--------------------------------*;

data    bh;
 set    soep.bh (keep=
        hhnr bhhnr bh01);
proc    sort; by hhnr bhhnr;
run;

data    mywork.new;
        merge mywork.new (in=present) bh;
        by hhnr bhhnr;
        if present;
```

```
run;

*        ----( Done ! )---------------------------------- *;

proc     contents;
proc     means;
run;
```

Chapter 5

Sampling and Weighting

by Markus Pannenberg, Rainer Pischner, Ulrich Rendtel, Martin Spiess and Gert G. Wagner

Longitudinal data files as well as cross-sections can be constructed for the German Socio-Economic Panel. This can be done on either the individual (personal) or household level. In the SOEP database, different weighting variables for cross-sectional as well as for different kinds of longitudinal weighting are set aside for each person in the *PHRF*-file and for each household in the *HHRF*-file. This multiplicity of weighting variables may at first glance seem confusing, but they are necessary to properly weight all the possible samples of data. The number of possible longitudinal samples grows with the square of the available waves. In addition, the two levels of analysis (person and household) double the number of required weighting factors.

From a user's point of view, the theory of weighting and the procedures which calculate the different weights may be a little bit hard to understand. But the encouraging message is: the usage of the weighting variables which are in the data files is user friendly. In this chapter we give advice on using the weighting variables and the methodology of the weighting itself is explained. For a better understanding of the weighting, the important characteristics of the sampling rules are explained.

5.1 Target Population and Respondents

The target population to be represented by the SOEP is defined firstly as the residential population of the FRG in 1984 including West Berlin, secondly as the German residential population in the GDR (including East Berlin) in June 1990. In the FRG, selected foreign groups were oversampled in the study. The

sampling probability for the eastern sample is bigger than the probability for the main sample in West Germany. Those different sampling probabilities were chosen to make sure that the number of cases for the different groups in the sample are large enough for their analyses.

The different sampling probabilities are the first reason for weighting. The second reason is non-response, i.e. not willing to participate in the first wave and attrition in the subsequent waves.

5.1.1 Sampling

The original West German sample in 1984 was carried out separately for two populations:

Sample A "Residents in the FRG" covers persons in private households with a household head who is not a Turk, Greek, Yugoslavian, Spanish or Italian. Although, one should keep in mind that some non-Germans are included in this sample, Sample A is often called the "West German Sample" of SOEP.

Sample B "Foreigners in the FRG" covers persons in private households with a Turk, Greek, Yugoslavian, Spanish or Italian household head. Compared to Sample A the population of Sample B is oversampled in order to allow for stand-alone analyses of this population which was thought to be affected by additional drop-out behavior, due to re-migration.

Institutionalized persons in the true sense of the word (hospitals, nursing homes, military installations) were not representatively included in the first wave. Later, however, persons from the initial households who had taken up residence temporarily or permanently in institutions of this kind were followed.

Sample C "German Residents in the GDR" covers persons in private households where the household head is a GDR citizen. 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).

Sample D "Immigrants" was started in 1994 and covers persons in households where at least one household member immigrated to West Germany, since 1984 when the initial samples were drawn.

Sample E "Refreshment" was started in 1998 and covers persons in private households independently from the nationality of the household head. Sample E was selected independently from Samples A through D.

Sample F "Innovation" was started in 2000 and similar to subsample A–D and E covers persons in private households, where households with at least one adult household member not having German nationality had a higher selection probability compared with households where all the adult members had German nationality. Sample F was selected independently from Samples A through D and E.

Sample G "High income households" was started in 2002 and covers persons in private households. Households were selected into this sample only if their monthly household net income in 2002 was equal to or larger than 3835 Euro. Sample G was selected independently from Subsamples A – F.

Sample A "West German Residents" ("German Sample")

Sample A was intended to net 4,500 households. In the end the completed net sample contained 4,554 households. The ADM[1] master tape from 1982 served as a basis for collecting sample A. 584 sample points were randomly selected from it by means of a multi-stage stratified sampling procedure. The interviewer selected the households within the selected constituency according to the random-route procedure. Working from a given random start address the interviewer had to select every seventh household as a target household. Households whose household-head belonged to the definition of sample B were discarded.

Sample B "Foreigners in West Germany"

Sample B consists of five autonomous samples for the five numerically largest foreign nationality groups living as immigrants in the FRG in 1984. To facilitate detailed analyses, a sample of 1400 net cases was projected. Thus the sampling rate for this sample exceeds the rate for sample A. Anticipated out-migration rates were taken into account in setting a sampling rate which gave a high probability that after several waves a considerable number of 1000 cases are still in the sample.

Population B was selected from primary sampling units (PSUs) of counties and metropolitan areas. A random selection of PSUs was independently drawn for each nationality. Using immigrant registration records in each PSU, the respondents were then selected by probability sampling, i.e. systematic sampling with random start address. The household of the respondent selected in this manner then came into the sample, provided that the household head had the same citizenship as the selected respondent. In a number of counties and metropolitan areas - particularly in Baden-Württemberg - it was not possible to draw from the immigrant registration lists. The alternative solution here was to randomly select counties and then use the local residents' registration lists.

Some 80 PSUs were drawn for the (strongly overly-represented) Turks and 40 PSUs for each of the remaining nationalities. Some 20 addresses were then drawn from the registers for each PSU, some of which were used as "reserve addresses".

[1]ADM is the "Arbeitsgemeinschaft Deutscher Marktforschungsinstitute" (Working Group of the German Marketing Research Institutes).

The number of addresses used per sample point in the sample B show a stronger variation than in sample A. Substantially more addresses were false or contained no-longer eligible respondents, in which case a "reserve address" was to be used.

Sample C "German Residents in the GDR"

German unification was anticipated in the Spring of 1990. The size of the East-sample C was set to permit analyses for the GDR and the later new *Bundesländer*. A target of at least 2,000 households was set, implying a greater sampling rate than sample A. Some 2,179 households were ultimately interviewed.

Because access was granted to addresses from the central residents' file of the GDR, a different and better sample method than in samples A and B was possible[2]. In contrast to the ADM master sample in the old FRG, the sample frame is a probability selection of private addresses drawn from the central residents' data base (Pietzke 1991)[3]. That is to say, addresses are drawn by fixed steps with random start (issue date: March 14, 1990).

To design sample C of SOEP this master sample was used in the following ways:

- First a household-proportional allocation was calculated for 360 sample points which followed the stratification of the master sample according to county and community size.
- For each stratum the sample points which corresponded to this household-proportional allocation were then taken from the master sample by systematic with random start selection.
- Finally, for each of the available addresses for these sample points, a person 16 years of age or older was selected as a "start address". In order to produce a representative household sample (and to save costs and traveling time by lumping together the respondent addresses) the random-route method was chosen. Commencing with this start address, each interviewer was to list the households on a formally described and clearly defined random route. The start address itself was not to be surveyed.
- Ten private households were to be listed and recruited for panel participation[4] unless it turned out that while making contact, one of the listed addresses did not belong to the target population (or that the residence was vacant). In this case up to two substitute addresses could be listed and contacted. Every

[2]Insiders from the GDR social research organizations raised objections against taking addresses out of the central register, claiming that often at these addresses other people lived there than the ones who were registered there. Although this assertion is difficult to prove or disprove, in view of the housing shortage in the GDR it is plausible and probable. The quality of the SOEP, however, is in no way affected by this because the address sample was merely used to ensure a random and representative regional distribution of the respondent households only. Who finally lives there is in fact irrelevant for the random-route method.

[3]The ADM group first brought out a drawing frame for the new Bundeslaender using the election districts as a basis after the first national election in both Germanys in Dec.1990.

[4]A separate, preliminary address collection was not possible due to lack of time.

third household was a "target household" and thus to be recruited for the survey[5].

Sample D "Immigrants"

For a detailed description of the sampling of Sample D "Immigrants", see Infratest-SOEP-Gruppe (1996). The weighting scheme of this sample is described in Rendtel, Pannenberg, and Daschke (1997) and is therefore not included in this chapter. An overview is documented in Burkhauser, Kreyenfeld, and Wagner (1997).

Sample E "Refreshment"

In 1998, a new sample was selected from the population of private households in Germany. The new sample, also denoted as subsample E, was selected independently from the ongoing panel (subsamples A through D). The selection scheme used for sample E essentially resembles the scheme also used to select subsample A. Again, the data are collected in two stages and two phases within the first stage, where the first- and second-phase samples are selected using the scheme also used for selecting subsample A. Although there are slight differences in the selection of the second-stage sample, mainly due to testing a new survey instrument (using a laptop for the personal interviews, i.e. computer assisted personal interview (CAPI) vs. paper-and-pencil personal interviews (PAPI)), the selection scheme is very similar to the one used to select the second-stage sample of subsample A. Sample E was intended to net 1000 households. At the end the completed net sample contained 1979 households. Apart from the new survey instrument, subsample E differs from the other subsamples in that more information about nonresponding units is available. It is expected that the additional information selected allows methodological studies about nonresponding units. For more details, see von Rosenbladt and Stutz (1998).

Sample F "Innovation"

Starting 2000, subsample F, was selected from the population of private households in Germany. Like subsample E, it was selected independently from all other subsamples and the selection schema was essentially the same as for selecting subsample A and E, however, with one exception. Like A and E, subsample F was selected in two stages and two phases within the first stage.

[5]This three-address interval is routinely used in the random-route procedure by the ADM institutes. With the West samples A and B a wider interval was constructed (7 households) in order to attain a higher degree of independence for the households. This proved impossible in the GDR because the interviewers were unfamiliar with the random-route procedure to begin with, so it made no sense to burden them additionally with more distance to cover.

The difference between F and subsamples A and E was in the selection of households within PSUs as follows: First, the population was divided in to two parts: Those households with all adults (age $\geq$ 16) with the German nationality ("German" households) and those households where at least one adult does not have German nationality ("non-German" households). Within each PSU, 24 households were selected according in the same manner, as in subsamples A and E. However, "German" households were selected mainly using the first 12 addresses within each PSU. In fact, a few were selected from the second 12 addresses as well. The "non-German" households were selected using all the 24 addresses. Like in subsample E, in addition to the traditional interview technique (PAPI), CAPI was used and more information about non-responding units is available as compared to subsamples A–D. Sample F was intended to net 6000 households. At the end the completed net sample contained 6052 households. For more details, see (Rosenbladt, 2001).

Sample G "Oversampling of High Income Households"

The first wave of Sample G was started in 2002 and is a sample of the population of private households with a monthly income of at least EURO 3,835, sampled independently from all other subsamples of the SOEP. (It should be noted that starting from wave 2 in 2003, the income threshold was changed. Households with a net monthly household income below to EURO 4,500 were not to be followed. However, the corresponding population of "less rich" households in sample G remains part of the standard distribution of SOEP-data.) Sample G is selected from a larger sample, the "Infratest-Telefon-Master-Sample" (ITMS). The ITMS is a multi-stage stratified telephone sample of households, started in 2001, following ADM-standards with respect to the random-digit-dialing technique. Households from this sample were selected according to a stratified sample selection scheme into subsample G if they agreed to participate. As in the other SOEP subsamples, the data of wave 1 were surveyed by means of face-to-face interviews. For details, see Sozialforschung (2002).

5.1.2 Strata and cluster information

Three variables `STRAT1`, `STRAT2` and `SAMPOINT` were generated to give more information about the strata used to select the subsamples and allow the definition of clusters for the estimation of variances using program packages like STATA or SUDAAN, that use strata and cluster information at the first wave of each subsample. The variables `STRAT1`, `STRAT2` and `SAMPOINT` can be found in the file *VARIANZ*. A further variable that may be of interest is the identifier of the interviewers (`INTNR`) also located in the file *VARIANZ*. Due to German data security laws, the values of all three variables are randomly

assigned, however, under the restrictions described below.

The variable `STRAT1` identifies the strata from which the primary sample units (PSU's, similar to voting districts) were selected. For subsamples A,C,D,E and F these are given by regional strata, defined by Bundesland (federal state), Regierungsbezirk (administrative district) and Gemeindetyp (type of community). Note that these units may change over time. For subsample B, strata are defined by the nationality of the head of the household.

The strata in variable STRAT1 are coded as follows: Each *Bundesland* received a number between 20,000,000 and 180,000,000, which is the same for every first wave of each subsample, each *Regierungsbezirk* received a number between 1,000 and 999,000, differing over the first waves, and each *Gemeindetyp* received a number between 1 and 999, again differing over the first waves. These values were added to give the values of the variable `STRAT1`. The exception is for households in subsample B, where a number between 1 and 5 codes different strata.

Although the strata are different over the first waves of the different subsamples, they may be geographically the same or largely overlapping areas. This is, however, reflected in the values of the variable `STRAT1`: similar values identify geographically similar strata.

The variable `STRAT2` is missing for subamples A and C–F. Values are given for subsample B only. The values are generated in the same way as for subsamples A and C–F in `STRAT1`. However, these values do not reflect real strata used to select the sample, but merely identify geographically neighboring PSU's (artificial strata).

The variable `SAMPOINT` identifies the primary sampling units (PSU's) from which the households were selected. For subsample D there are no primary sample units, so the corresponding values of `SAMPOINT` are missing.

5.1.3 Results of Sampling in the 1st Waves

According to the sample plan, the original gross number of households in sample A encompassed 7,008 addresses. Of the 1,168 reserve addresses included therein, 158 were not used because in each respective sample point, a maximum response rate (9 or 10 households with completed interviews) had already been attained or seemed to be within reach. But in the sample points with weak response rates the sample was boosted with 1,129 addresses. So altogether 7,979 addresses were used.

In order to calculate the drop-out rate of the first wave, households that did not belong to the target population "Private Households Excluding the Separately-Interviewed Households in Sample B" had been subtracted from the total amount of start addresses. These addresses are defined as "quality-neutral drop-outs" and the result as "edited gross amount".

There was 5.8% quality-neutral attrition in sample A. The edited gross amount encompassed 7,519 addresses. The quality-neutral attrition is a result of the address procedure, namely the interviewer's notation of house numbers along the pre-determined route. With some addresses it does not become clear until contact is made at a later date that the household does not belong to the target population (because the household members belong to sample B). This was the case in 2.8% of the addresses on the lists. With other addresses it was discovered upon closer inspection that they were business addresses (0.5%) or vacant dwellings (1.9%). And then 1.0% were either false addresses or could not be found.

Sample B (the foreigner sample) gives a different picture because addresses were supplied by the registration offices. The extent of the quality-neutral attrition due to false or no-longer-current addresses is greater here than in sample A. The average rate of attrition for the five immigrant samples is 22% of the utilized addresses.

The sample response rate is varied by stratum. In sample A, 4,554 addresses could be taken into the net sample after concluding all of the field phases and the processing work. A sample response rate of 60.6% is implied. It is common that initial responses in panel surveys are significantly lower than the response rates in subsequent waves. See Duncan and Kalton (1987) p. 109 and Duncan and Hill (1989).

Sample B yields better results. The response rates range from 64.7% for the Italians to 70.0% for the Turks.

In sample A, the main cause of attrition was refusal. Due to a long period of field work, non-contacted households could be reduced to 3.2% (a percentage not attained in normal cross-sectional surveys). Only 0.2% of households could not be surveyed due to linguistic difficulties. Lastly there is the 0.8% of the addresses for which no survey information exists. As a rule, these are reserve addresses which the interviewer didn't realize were supposed to be contacted.

The refusal rates for the foreign households from sample B are visibly lower than for the German households. However, the share of non-contacted households is higher.

From an edited gross sample with 3,114 GDR household addresses it was possible to recruit a total of 2,179 households[6]. Thus with sample C in the GDR, a response rate of 70% of the "edited gross addresses" was attained within a field time of six weeks which ended right before economic reunification

[6]Three percent of the households could not be completely surveyed until the first week of July. Since the interview date is recorded in the data set, it is possible to determine whether the data was collected before or after the currency union of July 1, 1990. This is of importance since any information on cash money as of June 1990 is in East German Marks, as compared to West German Marks (DM) thereafter.

of Germany[7]. This is a result that is practically impossible to achieve with similar studies in the FRG.

The quality of a sample (cross-sectional representativeness) can be inferred from the conformity of a distribution of characteristics in the sample with the distribution in the population. The distribution in the target population is estimated using external statistics, posing the problem that they themselves could be biased as well (particularly the income and consumer samples). It should be pointed out that the official statistics also contain institutionalized residents who are not included in the SOEP sample.

With regard to the regional distribution characteristics and the household structure, sample A reflects familiar shortcomings of survey research. The population in the metropolitain areas, and here particularly in the central zones, is more difficult to recruit for survey participation than the population in the medium and small-sized towns and communities. Elderly persons are under-represented.

The socio-demographic structure of sample B, household sizes as well as the age and sex of the household head, appears satisfactory. In regard to the regional distribution, the drop-out structure shows the same result as for sample A, namely intensified attrition in the core zones of the metropolitain areas.

A validation of the socio-economic structures of sample C is even more difficult, since the data basis of the official GDR statistics in this field is even scantier concerning projections.

The quality of sample C can be evaluated only by a few external statistics. Although in the community-size classes there are some deviations from the target population in the completed sample, they do not show the general under-representation of metropolitan population as is normally observed in the West. However, the persons/households in East Berlin were slightly under-

[7]A special problem affecting also other survey efforts in the GDR in 1990, was "total attrition". Total attrition means that a cluster of addresses ("sample point") assigned to one interviewer remained unprocessed. For one reason or another the interviewer responsible for these addresses could not or would not carry out this assignment and a replacement could not be deployed in any case in time (it could even be that the field organization was not alerted in time to the fact that the addresses remained uncontacted). In the SOEP basis survey 29 out of 360 sample points were total drop-outs. These, however, are distributed randomly throughout the entire GDR and thus have no noticeable effect on the sample. The neutrality of these drop-outs is also verified by the regional validation of the sample. The total attrition is, on the one hand, a consequence of the inadequate telecommunication facilities in GDR, and secondly an indication of the problems involved in setting up or restructuring an interview staff. The survey institute Infratest had opted for taking over an already existing interviewer network from the former GDR and reorganizing it to meet with the new standards. An analysis of interviewer effects in the survey data was undertaken. As in West Germany, these proved to be inconsequential. There is, in particular, no effect of the entry-date of interviewers into the staff.

represented. As in sample A, the elderly age groups, especially the 70-years-and-older group, are clearly under-represented in the net sample[8].

5.1.4 Attrition in the Course of Time (Wave 2 and After)

For details of the level and the structure of attrition in the SOEP see Spiess and Pannenberg (2003). Death and moving abroad are natural causes for dropping out and are not a problem for analysis. But dropping out due to refusal of respondents and in some cases due to problems of finding a household again ("unsuccessful follow-up") may cause problems when the dropouts are not random. The following characteristics have found to be of significant importance in the SOEP:

Unsuccessful follow-up

- Household moved
- Split-off
- Large City
- Single household

Refusal of Respondents

- Resident of East Berlin
- Age of head of household
- Female head
- Household moved
- Splitoff
- Separation/divorce of partner
- Change of interviewer
- Number of interviews with the same interviewer
- Low household income
- Item non-response on income
- Expected loss of job
- Migration from East to West Germany

[8]The comparative figures from the population statistics are somewhat biased, too, because the institutional residents, who are not surveyed in the SOEP, could not be excluded.

5.2 Ensuring Continued Participation

For a panel survey the tracking rules for following the sample members and ensuring their continued participation are central to the study (see Duncan and Kalton (1987), pp. 105). The procedure developed for this purpose will be explained will also be seen that procedural changes were made at several points in the course of time.

5.2.1 The Tracking Concept

For a panel survey which includes all household members the criterion for following persons must be set so as to maintain the representativeness of the selected target population. Except for immigration that results in the establishment of new households, it is possible to reflect the natural population dynamic of the target population in the panel sample. It is necessary first of all that young household members maturing into respondent age (meaning completion of their 16th year) must be interviewed. Secondly, persons who leave an "initial household" must be followed according to a specific pattern. Persons who moved in while the panel was in progress and moved out again afterwards do not have to be followed for the sake of cross-sectional representativeness. However, children who have moved directly from a foreign country into a household must be followed since they belong to the target population of the 1st wave. This is true for children who move into initial households while the panel is in progress (see Galler (1987). These tracking rules for initial persons was first chosen for the SOEP in wave 2 to wave 6.

For longitudinal analyses, especially of demographic events, it is more practical to follow not only the initial persons but all persons who have ever come in contact with the SOEP. This procedure increases the number of events which are of particular research interest. This however leads to a snowball-like growth in the sample. Hypothetically, with enough mobility and a lengthy panel duration the entire target population will take root in the sample in the end. However, this is only theoretically the case, because in practice the respondents' successive refusal to participate leads instead to a reduction in the sample. Especially little willingness to participate is shown by persons who move into an established panel household. While following non-initial persons will therefore not create problems of size in the sample, it does affect representativeness. This must be taken into account in the weighting procedures. However the goal of tracking all respondents is somewhat different. The number of analyzable cases sinks less quickly when in-moves are also followed, than when initial-persons are followed only. Since 1990 (West-wave 7) all non-initial persons have been followed because of the special attraction of following new persons, since these persons generally constitute unusually

mobile groups, which are interesting for event-oriented analyses.

In cross-sectional evaluations, the weights for households with non-initial persons are somewhat lower than, for instance, other households. The simplified follow-up rules do not affect the validity of the initial person concept for longitudinal studies. Here, as opposed to cross sections, only the initial persons have a positive longitudinal-weighting factor.

The new tracking rule is still not able to overcome a major problem which arises with the immigrants who establish new households in the FRG. Because of the fact that at the beginning of the 1980s, the immigration of migrant workers or their family members constituted the bulk of all immigration. Therefore the decision was made not to sample successively new households of immigrants.

A dramatic increase in the immigration rate in the old Federal Republic of Germany since 1988 (Eastern-Bloc citizens of German descent, Germans from the GDR and refugees necessitates supplementing the sample if information on the residential population is to be continued to be given, because this immigrant population has become very important. According to our estimates, this population in 1991 constitutes 3% of the private households in the FRG.

Re-Surveying Information

In order to sample as many respondents as possible for a long time in the SOEP a distinction is made between "final" and "temporary" drop-outs. A drop-out is considered final in the case that a household/person can no longer be found despite intensive effort to do so or if an explicit refusal is given by the respondent himself to further participation in the survey. In all other cases the attrition is assessed at first as "temporary". This means that in the succeeding wave a renewed attempt will be made to contact the household and to persuade them to participate further in the survey. If this in turn isn't successful the "temporary" becomes a "final" drop-out - this household will no longer be included in the following panel waves. If, however, the household can be motivated again to take part, a gap in the data caused by the attrition in the preceding wave will remain.

In addition to these household-related gaps, other individual-person gaps appear when a household could not be completely surveyed in one wave but in the following waves all household members are interviewed. There were, for example, 123 households with incomplete interviews in the second West-wave, 53 of which participated fully in the succeeding third wave. Accordingly, there are 250 persons to be reckoned with, who have event-histories with gaps. Although this case number is small in relation to the whole sample, it should be taken into consideration that the gaps in each wave appear repeatedly for other households or persons. A procedure was therefore developed whereby

at least the most important longitudinal characteristics for the missing year are able to be reconstructed.

In the third year of SOEP the decision was made to try to close these gaps by special field work. The results were unexpectedly good. Some 192 persons from 134 households were to be queried. Longitudinal information could indeed be reconstructed for 176 persons. In the majority of cases this was done by telephone. Owing to the good results the decision was then made to reconstruct the gaps in wave 2 as well. The long period of time between the confirmation of the gaps in wave 2 (summer 1986) and the reconstruction (Autumn 1987) proved to be a handicap. Not more than roughly 75% of the persons concerned could be included in the field work because a relatively high number of target persons had dropped out of the panel in the meantime. Since 1987 (West-wave 4) re-surveying is done routinely.

Reconstruction of Biographical Information

The central thematic content of the first topical modules of the three waves for samples A and B (West) was retrospective biographical information. Job history in yearly stages of the individual life course, beginning with the age of 15 (wave 1); marital history, information on childhood and the move away from the parental home, information from women about their children (wave 2) as well as social background, occupational starts (wave 3). This information only needs to be recorded once in the course of the panel. However, biographical information is missing completely or partially for persons who weren't included in the panel and for persons who were temporary drop-outs in the second or third wave. The retrospective questions were therefore put into a supplementary biographical questionnaire. In the fall of 1987 the new persons who had entered the SOEP in the second, third or fourth wave were sent the questionnaire along with a small gift, a pocket calendar. The reconstruction of the biographical information on the temporary drop-outs from wave 2 was done primarily by telephone - along with the reconstruction of the longitudinal characteristics in uncompleted operations.

5.2.2 The Interview Mode

The SOEP survey instruments include: (i) cover sheets (address protocol), (ii) household questionnaires and (iii) individual questionnaires. The cover sheets (the so called address protocol) records a vast amount of information on the composition of and change within the household. That information is essential for attrition analyses, the weighting of the sample and for longitudinal analyses. The questionnaires are lavishly designed in comparison to commercial questionnaires because they have to be flexibly used by the

interviewer:

- In all subsamples the questionnaire may be administered as a personal interview or as a booklet that the respondent completes himself and returns.
- In sample B the interviewer respectively the respondent has the option of conducting the interview in German, in the native language of the respondent, or in a mixed mode.

The SOEP field procedures are as follows:

- Personally conducted oral interviews are conducted whenever possible.
- The respondent, however, is permitted to fill out the questionnaire, which is handed to and explained to him by the interviewer.
- In the event of a refusal to participate or non-appearance of target persons a new interview date will be agreed upon in writing or by telephone.
- If the respondent wishes, the (new) interview date can be cancelled and, as an exception, the interview will be conducted in writing (i.e. by mail) or by telephone assistance.

These rules are obviously soft. Only one rule is strictly implemented: information on a respondent can only be obtained from the respondent him/herself. Proxy interviews which are common, for instance, in the American SIPP study and are necessary in the PSID for all household members other than the head of the household, are not allowed. There are only a handful of cases where exceptions are made, for example when an immigrant household member gives permission to another household member to fill out his personal questionnaire. With this multi-method approach the potential amount of persons who can be contacted and are willing to do an interview can be hold on a high level.

The information on the interview method applied in individual cases is available in the data. So here too systematic analyses of method-related influences can be made. To date there has been no indication that the interview method has a strong influence on the results.

5.2.3 Maintenance of Motivation of Panel Respondents

Instruments

The following methods of motivating respondents are employed (see Pol 1989, 42-45).

- Giving the study a catchy name. All sample respondents know the SOEP by the name of “Life in Germany”.
- Respondents are given an illustrated informative brochure on the aims of the study (in sample B the brochure is translated into the respondents’ native language).
- Providing an information sheet on data privacy

- Following each interview with a letter of thanks after completion of the field work for each wave.
- Providing each respondent with a ticket for a well-known TV lottery.
- Since 1987 (4th West-wave) all panel households receive a small gift ("loyalty bonus") worth 5 to 10 DM (this is less than the hourly wage rate in West Germany).

Motivation of the interviewer is certainly an important influential factor for the respondents' willingness to participate. Good training, sufficient information about the project, a clear structuring of the survey instruments and information on research results furnish a foundation for successful interviewing. For years now, all of the interviewers involved in the surveys receive a thank-you card from the client (the DIW) at the end of the year in order to underscore the relevance of their engagement. In some years the interviewers get a book with description results of the SOEP on request.

Household - Interviewer Continuity

A panel survey represents a panel not only for the respondents but for the interviewers themselves.

The SOEP had two interviewer-deployment strategies to choose from:

- Assigning the survey work to as many interviewers as possible, meaning that the extreme case number of interviewers deployed would correspond to the number of sample points (these are the clusters which are administered by one interviewer).
- Concentrating the survey work on a minimum number of highly-qualified interviewers.

A maximum deployment of interviewers resulted in few clusters of interviewer effects as possible. This strategy was chosen starting in 1984. It has become noticeable in the field work in several waves, however, that some interviewers are much better-suited than others to realize the high response rate that the SOEP requires. Moreover, a change of interviewers was determined to decrease the probability of respondents' refusal. It had to be kept in mind that the loss of a single interviewer, who interviewed a lot of households very effectively, increases the danger that a great many households refuse to participate. Thus it has proved necessary to seek a balance between as many or as few interviewers as possible.

5.2.4 Updating the Address Register

A "panel masterfile" which was built up within the survey institute Infratest is very important for the success of the field work. This file contains the addresses, telephone numbers, interview method, and so on for every household.

During the whole year the whereabouts of each surveyed household and person no longer at the same address as the previous year is checked. Information on this follow-up work gets stored in the panel files.

In the first years of SOEP annually approximately 7% of all the households from sample A are found at a new address. This share is declining since 1988 after an initial increase. Persons who leave households and establish new ones at a new address constitute each year about 4% of all households from sample A. The address research for the survey institute can be given a very good rating because it's had over 98% success since the beginning of the project. In far more than half of the cases the source for the new addresses is the interviewers themselves, who relay the new addresses to the field office. Half of the remaining addresses are obtained at the postal department; and the remainder are standard at the municipal residents' registration office.

The field-work sequence was altered after wave 3. The cases that are considered difficult to contact are scheduled first in order to ensure enough time for getting the job done without jeopardizing the field schedule.

The SOEP field work takes 8 months to complete because a more than 90% response rate has to be attained in the follow ups. The first wave of the eastern sample was an exception. The field work lasted 6 weeks only because it was an important goal of the study to finish the first eastern wave before the economic unification of Germany. In order to make reporting-date-based evaluations, the date of the interview is retained in the analyzable data record, too.

Becketti, Gould, Lillard, and Welch (1988) also recommend this for the PSID sample. In adhering to the original SOEP concept it came to light anyway that the concept of following initial persons was too complicated for the actual field work, i.e. persons were also interviewed who shouldn't have been according to the rules of the initial-person concept. Moreover, the expanded concept greatly simplifies the weighting procedure, since the sampling probabilities are now easier to calculate.

In 1990 and after 25 waves an externally funded immigrant sample of so-called Hispanics was included in the PSID for the first time. The second, equally important immigrant population, the Asians, are still missing in the PSID.

Not included in the supplement questionnaire, though, were the persons who were new to the survey because they had just turned 16, since most of the questions didn't apply to them or the information on the parental home was already available.

We believe with the first wave in the then GDR, it was an advantage for the SOEP in regard to the respondents that the project was titled "Life in Germany". In form letters and in the accompanying GDR survey brochure, it was pointed out that in the old FRG the survey had been running since

1984, and its ambitious intentions of documenting life in Germany could now be fully realized. Unlike the first SOEP wave in 1984 in the FRG, the planned interview date for the households could not be pre-announced by mail. Due to the badly-functioning East German telephone system, meetings could seldomly be arranged by phone in the routine way this has always been done by the SOEP. This did not lead to problems worth mentioning, as in the GDR, unannounced visits were a part of daily life !

For legal reasons the data privacy information is of special importance for recurring surveys but on the whole is probably not a means of winning people's trust because, as test examinations show, the impression is made that the questions are of a more sensitive nature than is in fact the case.

It could also be shown that prior length of participation in the panel - with the exception of the newly-arrived persons - has no statistically significant influence on the willingness to participate (see Rendtel (1990)).

This last kind of address research is not possible in the USA, because there is no registration of inhabitants.

5.3 Weighting Procedures

The goal of any sample is to draw conclusions from the sample and apply them to the "recorded" target population. A projection of the sample cases is required in order to be able to infer the case numbers of the target population. For SOEP we must distinguish three steps of weighting:

1. Cross-sectional weighting of wave 1
2. Weighting of longitudinal populations
3. Cross-sectional weighting of waves 2 and thereafter

5.3.1 Methodology for the Construction of Weights

The goal of population estimates is to estimate the total number of occurrences of a certain characteristic in the population from the survey sample. In the sense of statistical inference, we want to infer the unknown population parameter

$$Y = \sum_{i=1}^{N} Y_i \tag{5.1}$$

where the indicator variable Y_i shows whether the i-th element has the interesting characteristic (Y_i is one) or not (Y_i is zero). The construction of weights by inverse selection probabilities as in Horvitz and Thompson (1952)

is motivated by the randomization assumptions. The characteristics Y_i of the individual elements of the population are considered fixed quantities. The individuals i are selected at random. The stochastic indicator C_i determines whether the individual i belongs to the sample (C_i is one) or not (C_i is zero). If Y is estimated by linear estimation of the form:

$$\hat{Y} = \sum_{i=1}^{N} \alpha_i C_i Y_i, \tag{5.2}$$

then the following expected value of $\hat{Y}$ results[9]:

$$E(\hat{Y}) = \sum_{i=1}^{N} \alpha_i E(C_i) Y_i, \tag{5.3}$$

Because $E(C_i) = P(C_i = 1)$, then:

$$\alpha_i = \frac{1}{P(C_i = 1)} \tag{5.4}$$

This results in:

$$\begin{array}{rcl} \hat{Y} & = & \sum_{i=1}^{N} \frac{1}{P(C_i=1)} C_i Y_i \\ & = & \sum_{i=1}^{n} \frac{1}{P(C_i=1)} Y_i. \end{array} \tag{5.5}$$

The longitudinal weighting procedure for SOEP is recommended in Galler (1987). The selection process for a panel survey can be described as a multi-step process. The choice of a longitudinal sample over T panel waves is seen as a selection process with $2T$ steps, which are characterized as follows (to simplify the notation, index i has been omitted from the sample elements):

First step: Design selection (initialize the sample) $P(D = 1)$

Second step: Response in the first wave $P(R_1 = 1|D)$

Third step: Successfully make contact in the second wave $P(K_2 = 1|D, R1)$

Fourth step: Response in the second wave $P(R_2 = 1|D, R_1, K_2)$

2T step: Response in the $T-th$ wave $P(R_T = 1|D, R_1, ..., R_{T-1}, K_2, ..., K_T)$

[9]Only the chance of the selection process is considered when the expected values are calculated. In accordance with the randomization approach, the Yi are treated as constants.

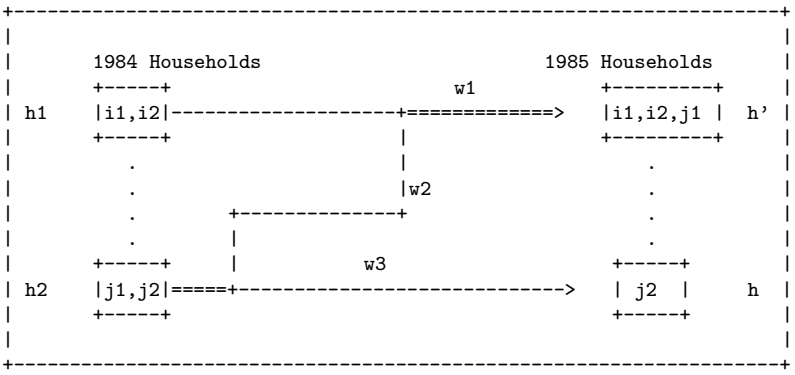
The probability of selection $P(C = 1)$ for the entire selection process over all single steps is given by the product of the single probabilities:

$$\begin{aligned} P(C=1) &= P(D=1, R_1=1, K_2=1, ...R_t=1) \\ &= P(D=1) \\ &\quad *P(R_1=1|D)* \\ &\quad P(K_2=1|D,R_1)* \\ &\quad P(R_1=1|D,R_1,K_2)* \\ &\quad ...* \\ &\quad P(R_T=1|D,R_1,...R_{T-1},K_2,...,K_T) \end{aligned} \tag{5.6}$$

Thus the problem of weighting longitudinal samples of households or persons for a certain period of time is reduced to calculating the initial probabilities of the first panel wave and determining the retention probabilities, i.e. the conditional probabilities of remaining in the panel after each selection step.

The calculation of the selection probabilities for cross-sections of wave 2 and the following waves is more complicated because of the follow-up strategies used in the panel. Table 5.1 provides an example of the selection process for households in the second wave (1985).

Table 5.1: The follow-up paths of 1984 and 1985 households.



Households in the 1984 and 1985 statistical populations are first put in separate boxes, for illustrative purposes. Some 1984 and 1985 households are related by having a person in common. The follow-up of the person in common for all possible paths is shown with arrows. In 1984 persons i_1 and i_2 live in household h_1 and persons j_1 and j_2 live in household h_2. Between 1984 and 1985 the person j_1 moves into the household h' of i_1 and i_2. This leaves person j_2 in a single member household h. Therefore household h' has

two possibilities to become part of the 1985 sample: Household h_1 is selected in the first panel wave and path w_1 is followed and!or household h_2 is selected and person j_1 is followed on path w_2.

In the first panel wave the selection of households is independent of each other. The probability that household h' is in the 1985 panel sample is therefore the sum of the two individual probabilities. That is to say selecting household h_1 (or h_2) and following path w_1 (or w_2).

As a rule a path is followed in order to reach a panel household. This is shown in Table 5.1 where a person j moves from household h_2 to an already existing panel household h'. The h_2 household characteristics which may greatly influence its selection probability in the first wave are, however not collected in the SOEP.

For this reason, it is necessary to calculate the selection probability of household h_2. For the person j who moved in, only the person characteristics PX_j are available. It is assumed that the selection probability P_j in the first wave can be described by a logit model. For the person j who moved in, both the known person characteristics PX_j and the unknown household variables HX_j of person j's former household are as follows:

$$ln\frac{P_j}{1-P_j} = PX_j\beta_1 + HX_j\beta_2 \tag{5.7}$$

The unknown household part plays the part of the error term in the specification:

$$Y_j = ln\frac{P_j}{1-P_j} = PX_j\beta_1 + \varepsilon_j \tag{5.8}$$

This relationship applies to the persons surveyed in the first panel wave and to those who moved there in the second wave. Since both the person characteristics and the logits Y_i are known for the persons surveyed in the first wave, then b_1 can be estimated from this group. Based on the estimates b_1, we can calculate the inclusion probability in wave 2. This procedure applies for the subsequent waves too.

5.3.2 Weighting of SOEP

Cross-sectional Weighting of Wave 1

As described in the previous section, the selection probabilities (and thus the weighting factors) for the first wave of a panel are of special importance, because these values are used as the starting point for deriving all other weighting factors.

The survey design of the first wave of the German Socio-Economic Panel contained a two stage selection procedure. The primary units (sample points)

are polling districts in sample A and counties in sample B. The secondary units (households of the first wave) are drawn from the polling districts using a random route procedure, and in sample B the persons drawn from that county's foreigner register.

The primary units have been stratified. For sample A, 148 regional cells, i.e. strata, were formed from the characteristics state, county, and local district. The strata sizes n_h were chosen to be proportional to the number of households N_h in those regional cells, i.e. n_h is proportional to N_h, where $h = 1, ..., 148$.

Then the primary units were chosen in each regional cell using systematic sampling proportional to size. The size proportions are related to the number of households of the primary units. The ordering of the primary units is important when using systematic sampling with random start and fixed intervals. The sequencing of primary units was carried out with the characteristics district, community, city section, and polling district number.

Within each foreigner group in sample B the districts, i.e. primary units, were sequenced by state and county. Again the primary units for each foreigner group were selected by systematic sampling proportional to size. The size of primary units was the number of foreigners of that nationalities within that area.

The second stage of the sample's selection was the performance of a random route procedure of sample points (sample A) and for sample B the selection of person from the foreigner registers. The selection of a person from the registers used again a systematic selection with random start number and fixed intervals. In some ways the random route procedure for the sample points can also be interpreted as such a systematic selection. There the procedure's inspection rules, provide a sequencing of the households. Every seventh household was chosen. Through the sequencing of the primary units and the systematic selection procedure, a kind of net is generated which covers the survey area uniformly.

One must recognize that the proportional selection probabilities at the primary unit level are such that in each primary unit (the objective was n_s equals eight households) a fixed number n_s of secondary units are drawn. If π_k is the selection probability for the k-th primary unit and $\pi_{i/k}$ the selection probability for the i-th unit in the k-th primary unit with size N_k, then the selection probability $P(C_i = 1)$ for the i-th unit is π_i:

$$\begin{aligned} \pi_i &= \pi_{i/k} p_k \\ &= n_s/N_k * \text{const} * N_k \\ &= \text{const} * n_s \end{aligned} \tag{5.9}$$

The designs of selection of households in sample A and persons in sample B are approximately identical. However, in sample B households in which several

persons are of the same nationality, the selection probabilities of households equal to the sum of selection probabilities of all household members. Hence the selection probabilities of households in sample B are proportional to the number of its household members who are 16 years and older.

Because all household members[10] were included in the survey, the selection of the households and the persons living in them is identical. Therefore the selection probabilities of a given household and of its members are the same, except for persons who have secondary residences. These persons have doubled the selection probability, i.e. an implicit assumption was made that these persons have the same selection probability at both residences.

On the basis of household or foreigner aggregates in the primary units the design probabilities $P(D_i = 1)$ can be determined for the start wave[11].

Because respondents were told at the beginning of the survey that this would be an annual survey, the respondents knew that more time and effort would be required than for a usual cross-sectional survey. Hence a sampling rate of about 65% (average of sample A and B) in the first wave can be considered a success.

The response probabilities $P(R_{1i} = 1|D_i = 1)$ for the first wave were estimated in two steps: The first step used only regional characteristics of all households (participating and non-participating). The second step compares sample information of participating households with the corresponding information from other surveys (micro-census and European Communities labor survey).

The regional characteristics of all households (participating as well as non-participating) are known. The response probabilities can be estimated in each of the 148 regional cells by using the ratio of participating households to the total number of attempted households. The reciprocal of the products $P(D_i = 1) * P(R_{1i}|D_i = 1)$ is available for the user as design and regional weights in the SOEP-database under the label `AHDESREG` (in the file *HHRF* which contains the weighting variables for households) or `APDESREG` (in the file *PHRF* with personal weighting variables).

These calculations of the selection probabilities can be enhanced by linking the estimated number of household and person characteristics with those from other corresponding surveys. This is done because we assume:

- The population estimate in the other surveys are more exact than those estimated from the panel.

[10] All household members 16 years and older were questioned directly. Data on younger household members were obtained indirectly from the household questionnaires.

[11] The accuracy of these design weights is dependent on the accuracy of the estimated number of households and foreigners. The use of inaccurate household numbers can lead to a "design defect", i.e. even during error free execution of the survey designs the calculated selection probabilities do not correspond to the actual selection probabilities.

- This adaptation also makes the calculation of other characteristics more precise.

The higher precision of the Micro Census is derived from its 50 times larger sample size and the respondents' obligation to provide information.

The second argument is more difficult to substantiate. The modified selection probabilities are not uniquely defined in that the estimation results agree with the $J < n$ restrictions from the Micro Census. A specific solution is achieved if the modified weights and the original weights have minimal information distance, see Ireland and Kullback (1968). This specific solution is characterized by

$$\tilde{P}(R_{1i} = 1|D_1 = 1) = P(R_{1i} = 1|D_1 = 1) * exp(\sum_{j=1}^{J} \lambda_j m_j(i)) \qquad (5.10)$$

for the modified response probabilities $\tilde{P}(R_{1i} = |D_i = 1)$. See (Rendtel 1987).

When the i-th unit has the j-th characteristic, then $m_j(i)$ takes on the value one, which is controlled by the distribution of the micro-census. The unknown l coefficients can be determined by iterative application of the "should/is" adaptation, i.e. the "iterative proportional fitting" algorithm. Alternatively the l's can be ascertained by a direct minimization of the information distance with the Newton-Raphson algorithm (see Merz (1983)).

The adaptation of estimation results to certain marginal distributions is therefore equivalent to the assumption that the response probabilities are in accordance with a main effects model, where the main effects are generated by the variables for the marginal distribution. This supports the original assertion that the adaptation of additional data increases the estimation precision of other characteristics. Under the main effects model the uncorrected estimation of population totals is biased. However the validity of this model cannot be checked empirically. To do this it would be necessary to know the model variables for missing households, which is generally not the case[12]. The relationship of the response probabilities may be even more complex. For instance, if significant interaction effects appear alongside the main effects, then the adaptation procedure can lead to increased distortions of population estimates.

The three-step-weighting process of wave 1 of samples A and B takes into account all information which are available for the calculations of sampling probabilities. Thus the final step of adjusting the marginal distributions of the sample and external statistics is changing the weighting factors slightly only. But the three-step-procedure make the variance of the weighting factors

[12]If possible, missing households were surveyed at least on household size and sex, birth year, and occupation of the head of household (see: field organization 1985a, p. 72).

bigger than a one-step-procedure might do. This argument was used to do the weighting of the first wave of sample C a little bit different.

The weighting of wave 1 of sample C is a one-step-procedure only by adjusting some marginal distributions of the sample to external statistics.

Chosen for the marginal adjustment of the 1st wave samples A and B were the characteristic combinations shown in the left column of Table 5.2, which affix a total of 316 restrictions to the projection results. For the frame adjustment for sample C, person-related and regional restrictions could be given only, because for 1990 no valid household-structural data were available from the GDR. With the projection of the 115 characteristic combinations shown in the right column in Table 5.2.

Table 5.2: Marginal distributions used for first wave adjustment

Sample A + B	Sample C
Households • Sex of the head of household (sample A only) • Age of the head of household • Size of the household • Nationality of the head of household (sample B only)	
Resident population in private households • Sex • Age • Marital status • Nationality of the head of household (sample B only)	Resident population • Sex • Age
Foreign institutionalized population of sample B: • Sex	
Pupils total: • Sex • Type of school • Nationality of the head of household (sample B only)	
Gainfully employed persons residing in private households: • Sex • Age • Job (main ISCO groups) • Nationality (sample B only)	
	Resident population • Counties • Size of Community

In 1998, a new sample was selected from the population of private households in Germany. The new sample, also denoted as subsample E, was selected independently from the ongoing panel (subsamples A through D). The selec-

tion scheme used for sample E essentially resembles the scheme also used in selecting subsample A. Therefore, the design weights are calculated in the same way as the wave one design weights for subsample A elements. The number of observed and valid private households in subsample E in 1998 was 1067, covering a total of 1932 successfully interviewed persons aged 16 and older. The number of children within these households in 1998 was 468.

Subsample F was again selected independently from all other subsamples from the same population of private households. With one exception, the selection scheme was essentially the same as for selecting subsample A and E (Spiess, 2001). Within each PSU, 24 households were selected according to the same schema as in subsamples A and E. However, "German" households (all adults aged $\geq$ 16 having "German" nationality) were selected mainly using the first 12 addresses within each PSU (although a few were selected from the second 12 addresses as well). The "Non-German" households (at least one adult not having German nationality) were selected using all the 24 addresses. The number of observed and valid private households in subsample F in 2000 was $6,052$ households. These $6,052$ households covered $2,993$ kids (age < 16) and $11,532$ adult persons, valid interviews are available for $10,890$ of the adult persons. The sampling probabilities are approximately 0.00028 for "German" households and 0.0005 for "Non-German" households.

For estimation purposes it is recommended to use all subsamples, i.e. subsample A through F. If design-based estimators are calculated, then the standard cross-sectional weights `$HHRF` and `$PHRF` can be used as described below. Since subsample D is a sample from a subpopulation of that population considered to be represented by subsamples A–C in 1995, E and F are selected from the same populations considered to be represented by subsamples A–D and A–E in 1998 and 2000, respectively, standard weights had to be modified accordingly. Therefore, if for waves L through N (1995 — 1997) cross-sectional analyses are to be carried out and only subsamples A–D are used, then the standard cross-sectional weights can be used. However, if only subsamples A–C are used then the standard weights have to be multiplied by 1.053. If from wave O (1998) to wave P (1999), cross-sectional analyses are carried out using all the subsamples A–E, then the standard cross-sectional (individual as well as on household level) weights can be used. However, if only subsamples A–D are used then the weights have to be multiplied by 1.25. Correspondingly, if only subsample E is used, then the weights should be multiplied by 5. Consequently, if for that time period, only subsamples A–C are used, then the weights have to be multiplied by $1.316 \approx 1.25 \times 1.053$. In the same way, if from 2000 on (wave Q) all subsamples, i.e. A–F are used, then the standard weights can be used. If, however, only subsamples A–E are used, then the corresponding cross-sectional weights have to be multiplied by 1.82. See `QPHRFAE` for persons and `QHHRFAE` for households. If only subsam-

ple F is used, the corresponding cross-sectional weights have to be multiplied by 2.22. See `QPHRFF` for persons and `QHHRFF` for households. Accordingly, if only subsamples A–D are used then the weights should be multiplied by $2.27 \approx 1.82 \times 1.25$, if only subsamples A–C are used, the weights should be multiplied by $2.393 \approx 1.82 \times 1.25 \times 1.053$. The derivation of these factors are given in Spiess and Rendtel (2000) and Spiess (2001b).

Starting with data released in 2002, there are some notable conceptual changes in the cross-sectional weighting scheme:

- For the initial weighting factors of sample F, more detailed design information is used, taking into account the sample-point specific realization rate of potential addresses.
- For each sample, A-F, the initial weights are "top trimmed" at a value of 10 times the respective median. This reduces the variation of weighting factors in general and the impact of outliers.
- Finally when adjusting these initial weights to marginal distributions of external statistics, i.e. from the German Mikrozensus, we now use more detailed age categories. This adjustment is made separately for East and West Germany, as well as for Samples A-E and Sample F.
- Since the selection scheme of subsample G is different from those of the other subsamples and selection probabilities are not available, subsample G is not integrated into the standard weighting scheme. Rather G-specific weights have been developed to be used if only this population of high income households is of interest. Such weights are denoted as `$HHRFG` at the household level (in the file *HHRF*) and `$PHRFG` at the individual level (in the file *PHRF*), respectively. Basically, the G-specific weights of the first wave are derived from the corresponding weights of high-income households in subsamples AF assuming similar socio-economic structures (see Pischner (2005) for details). For analyses based on the integrated SOEP-sample covering all subsamples A through G, the weights `$HHRFAG` and `$PHRFAG` may be used. These are arrived at by reducing subsample G specific weights relative to those of subsamples A through F, ie. the overall share of "rich" households remains constant. The proportion of high income households coming from subsamples A through F versus those coming from subsample G is determined by the respective number of observations as described in Pischner (2005). The analyses of attrition and the derivation of staying probabilities and longitudinal weights (see variables `$HBLEIB` and `$PBLEIB` for households and individuals, respectively) is done in analogy to samples A through F.

Longitudinal Weighting

The weighting of longitudinal populations is straightforward. To get an estimate for time $t + x$ it is only necessary to know for certain subgroups in the population how big the drop-out rate is. The inverses of the drop-out rate give the weighting factor.

The calculation of dropout rates can be done by cross-tabulations or - much better - by LOGIT-regression analysis. For details see the background paper "Documentation of Sample Sizes and Panel Attrition in the SOEP" by Spiess and Pannenberg (2003). To determine the reasons for attrition in wave $t+1$ the characteristics of a household in wave $t = 0$ can be used. Additionally

characteristics of the fieldwork for wave $t + 1$ can be used, for example the information that a household moved or the interviewer changed.

This analysis is done for each wave. The longitudinal weighting factors adjust from one wave (beginning with wave 2) to another wave. To arrive at the correct weight for longitudinal analyses in the course of multiple waves, the longitudinal factors need only be multiplied by each other.

Cross-sectional Weighting of Wave 2 and thereafter

For cross-sectional weighting not only "old friends" must be weighted but new members of the sample too. The inclusion of the non-initial persons is no problem as long as the sample probabilities for households in the year in question are known or can be estimated. Thus it is not necessary, as the PSID is doing, to assign zero-weights to persons who join old households.

But because the selection probabilities of households are arrived at solely by the selection probabilities of its members at the start of the panel as well as the follow-up rules, households with new arrivals have higher chances of selection than households without (because there were at least two paths by which they could be reached). As a consequence, households with new arrivals have to be assigned a lower weight. If one applies the household weight to all household members (i.e. non-initial persons too), this lower weight compensates for the increase in case numbers caused by the new arrivals.

5.3.3 Conclusion from a user's point of view

If one regards the entire weighting scheme with the practical eye of someone who wishes to analyze SOEP data, the previous sections mean that per wave for each household in the data set a cross-sectional and a longitudinal weighting factor are made available in *PHRF* and *HHRF*. The cross-sectional factor furnishes for the entire sample the valid values for the survey year in the target population.

5.4 Using the Weights

5.4.1 Technicalities

Person-related weighting variables are stored in file *PHRF*. Household related weighting variables are stored in file *HHRF*.

The file *PHRF* contains one record for each person who was ever listed in the panel database. The records are ordered by the person identification number (`PERSNR`). The file *HHRF* contains one record for each household that was ever listed in the panel database for a total of more than 21,000

households after wave 17, ordered by the *original* household identifier HHNR and the *current* household identification number (HHNRAKT).

All weighting variables are initialized to zero. Therefore the values of the variables are valid for all persons and households even if they did not participate in a particular panel wave or stopped participating altogether[13].

The following variables may be used to weight cross-sections of waves 1 (wave A) through 21 (wave U):

- In *PHRF*:
  APHRF, BPHRF,CPHRF, ... UPHRF
- In *HHRF*:
  AHHRF, BHHRF, CHHRF, ... UHHRF.

There is a flexible way of weighting longitudinal samples. The accompanying weighting factors can be easily determined by the use of the calculated "staying" probabilities, i.e. the probability that a person or household participates in the named wave and also participated in the previous wave. The reciprocal of that probability is stored in the following variables:

- In *PHRF*:
  BPBLEIB, CPBLEIB, DPBLEIB, ... UPBLEIB
- In *HHRF*:
  BHBLEIB, CHBLEIB, DHBLEIB, ... UHBLEIB.

For example with the help of these variables a longitudinal sample from wave 5 (wave E), to wave 21 (wave U) can be constructed. The weighting factor referring to persons can be labeled EUPHRF (but is not stored in file *PHRF*). The variable EUPHRF can be calculated as follows:

- EUPHRF = EPHRF * FPBLEIB * GPBLEIB * HPBLEIB * ... * UPBLEIB.

The first letter "E" signifies the start and the second letter "U" the end of the longitudinal sample. In general the weighting factor for a longitudinal sample can be calculated as the product of the weighting factor of the start wave and all the "staying" factors to the end of the longitudinal sample.

The new immigrant sample (sample D) of the SOEP provides information concerning immigrants to Germany. For details, see Rendtel, Pannenberg, and Daschke (1997). It covers households, containing at least one member who immigrated to the western states of Germany between 1984 and 1993. Because few people immigrated to the eastern states of Germany over this period, they were not included in this sample. People living in public institutions (nursing homes, mental institutions, etc.) or provisional housing for asylum seekers

[13]Their participation in the individual panel waves is recorded in files *PPFAD* or *HPFAD*. The number of entries in *PHRF* and *PPFAD*, or *HHRF* and *HPFAD* is the same, respectively.

are not included unless they were in the sample before moving into such an institution. From a user's point of view, the weighting procedure of sample D has to fulfill some requirements:

- Representative structural and longitudinal analysis of the new immigrant population (sample D).
- When combining sample D with the other three SOEP samples, the weighted entire SOEP population has to be representative for the private households in Germany in 1995.
- Concerning the "old" SOEP samples (A, B, C), the computed weights provided in the database have to guarantee a lasting representative analysis of the population of interest.

However, we have to point out that sample D includes some households, who have a positive selection probability with respect to sample A and B. These are so-called "mixed" households (households, living in West- Germany) with new immigrants. These households make up about 13% of the entire sample D and hence are not ignorable. Moreover, due to some inconsistencies during the field work, there are also some East German households, who migrated to West Germany after the starting date of sample C (so-called *Spät-Übersiedler*), in the D-sample, though they are not part of the target population. To get rid of these overlapping problems, we decided to assign zero weighting factors to all *Spät-Übersiedler* households (N=48; 8% of the sample). Facing the problem of positive selection probabilities for mixed households with respect of the "old" SOEP, it was necessary to create two different weighting variables to integrate sample D into the SOEP weighting scheme. The first variable is to be used with the full SOEP and the second is to be used when working with sample D, only. When the entire sample (A,B,C and D) is used, the "mixed" households of sample D are assigned a zero weight.

Beginning with wave 12 (1995), the following sample weights are included in the SOEP data files:

- `$PHRF` Person level weight for cross-sectional analysis for wave $ (199$), if all samples are used. This weight assigns zeros to all "mixed" households in sample D.
- `$HHRF` Household level weight for cross-sectional analysis for wave $ (199$), if all samples are used. This weight assigns zeros to all "mixed" households in sample D.
- `$PHRFD` Person level weight for cross-sectional analysis for wave $ (199$), if only sample D is used.
- `$HHRFD` Household level weight for cross-sectional analysis for wave $ (199$), if only sample D is used.

If only sample A and B are used in a weighted analysis, it has to be taken into account that beginning with wave 12 (1995), the A and B weights

are reduced to reflect the fact that immigrants are contained now in sample D. Therefore, in this special case, the sample weights of A and B must be multiplied by 1.053 (for details, see Rendtel, Pannenberg, and Daschke (1997)). This value represents the weight of immigrants in the calculation of the weights.

5.4.2 Additional Comments

Institutionalized Population in Samples A, and B

Due to practical problems the sampling rules used in the first wave of the panel almost entirely excluded the institutionalized part of the population of sample A from being surveyed. The institutionalized population of sample B could be surveyed by using the registration registers. This covers mostly persons who live in special residences for workers (*Arbeiterwohnheim*). A person who moves to an institution after the first wave can remain in the sample because of the panel's follow-up rules. Theoretically it is possible that eventually such moves will produce a representative sample of institutionalized respondents in sample A. But thus far there has been a very slow increase of such respondents in institutions. In sample A, the number of person interviews with respondents in institutions increased to 30 in wave five. The number of respondents in the foreigner sample fell by half from 47 in wave one to 22 in wave five. Experience shows that certain institutions (e.g. institutions that are closed to the public, such as jails) can be reached only intermittently. Finally the rate of response of respondents who move to old age homes is lower than average.

Hence the number of institutionalized respondents in sample A is too small to draw statistical inferences about the approximately one million persons in institutions[14]. This is the case for sample B too[15].

In order not to exclude them a-priori for cross-sectional analyses, a positive weighting factor is assigned to institutionalized respondents that corresponds to the reciprocal of their selection probability.

The household level also supports institutional households. This includes private households in institutions and household-like living in institutions. Even though there are no independent economic units in the institutionalized population by definition, most household questionnaire questions can be answered meaningfully[16].

[14] If one were to apply the average selection rate of one to 5000 for the institutional population of sample A, then 250 respondents would be required.

[15] The selection rate for sample B is about one to 1000. According to the 1985 microcensus sample B encompasses 40,000 persons in institutions.

[16] For example questions on dwelling conditions and rent.

These households are also assigned a positive weighting factor, being the reciprocal of the selection probability. But only "private households" are meaningful when estimating absolute numbers. In this case the institutionalized households must be excluded from the sample.

Cross-sectional evaluations of persons can be done in different ways. In each wave the weights of all persons residing in interviewed households can be counted. This group also includes a part of the population living in institutions. If such people are excluded, then the remainder is a statistical population representing the population in the micro-census defined as "persons living in primary residences in private households". The institutional population can be measured in the current wave $ from variable `$WUM2` in file *$HBRUTTO* or alternatively by using the variables `$POP` in *PPFAD*.

In most cases the desired weight is for the non-institutionalized population aged 16 years and older[17]. By excluding the institutionalized population, comparisons with the Micro Census are possible. Since there are only very few sample-persons in institutions, they can be ignored in calculating proportions. This is especially true for the German part of the sample.

Cross-Section Sample Populations

The statistical cross-sectional population is the population residing in primary residences in private households. Excluding persons residing in institutions, the weighting factors add up in each wave to the corresponding populations in the Mikrozensus, separately for East and West Germany since 1990.

Most of the information is obtained by way of the person questionnaire. All respondents aged 16 and older are asked to complete the person questionnaire. The statistical population used for the estimation is the population 16 years and older residing in primary residences in private households.

The cross-section weighting scheme for all waves except for sample specific starting waves as in Section 5.3.2 is based on the longitudinal weighting scheme. We adjust the cross-section data to the distribution of some main indicators, according to the results given by the German Mikrozensus (current population survey). This procedure is carried out for East- and West-Germany, separately.

The adjustment concerns the following populations: (a) Size of Household (1-, 2-, 3-, 4-, 5 and more persons in a household) (b) Sex and Age of the members of a household (5 year age brackets up to 80, 81 and older) and (c) Nationality (Germans, foreigners). The adjustment of the cross-sectional weights is done separately for Samples A-E and Sample F.

The cross sectional weight of institutional and Immigration household (sample D) remained unchanged. See Pischner (2000) for more details.

[17]The computation is: Survey year minus year of birth and greater than 16.

Table 5.3: Forming Longitudinal Samples with 5 panel waves

		Ends in wave				
		A	B	C	D	E
Starts in wave	A	x				x
		x				x
		x				x
		x				x
	B		x			x
			x			x
			x			x
	C			x		x
				x		x
	D				x	x
	E					

Longitudinal Samples

For example after five panel waves there are 10 possible longitudinal samples. Table 5.3 shows these longitudinal samples.

A longitudinal sample of the statistical population from wave X to wave Y includes the following persons:

- Persons who belong to the cross-sectional statistical population in wave X
- Persons who remain in the survey area through wave Y (have not moved away or died).

This longitudinal sample of the population is the natural statistical population. It is used for studying the life-course of panel members who participated in the panel from the first wave to the last wave.

Therefore the longitudinal sample always contains only a part of the cross-section of the first wave. The difference is due to losses from migratory movements and deaths occur over the course of successive waves. The population definition for migratory movements is: The longitudinal section of the statistical population equals the cross-section of the statistical population minus the respondents who died. The population definition for respondents who died is: The longitudinal section of the statistical population equals the cross-section of the statistical population minus the losses due to migratory movements.

Because of missing reference data in the Micro Census, the (estimated!) longitudinal sections of the statistical population are based solely on the internal update of the weighting values.

Establishing a population definition for households over several waves of data is more difficult. This is because a household does not form a stable unit over time. It can split up or join with another household. The composition of a household changes when a person dies or is born or moves into or out of a household. Various household definitions could be made for a given longitudinal sample. The most restrictive version embodies those households whose household composition does not change during the time interval.

The definition used in the German Socio-Economic Panel for a longitudinal sample on the household level is as follows[18]: A household stays in the longitudinal section, despite any composition change due to a move, a birth, a death, or if the entire household moves. If a household splits, then one of the household parts stays in the longitudinal sample. If children move out of their parents' household, then the parents' household stays. In case of a household split, that part of the household stays in the longitudinal sample that remains in the old dwelling or if the old dwelling is given up, it is that part of the household to which the former head of the household belongs.

The longitudinal household weighting factors also apply to more restrictive versions of the household longitudinal sample. The user may simply assign a zero weight to households outside the longitudinal section definition. For example all households that experienced a change in household composition.

In a cross-sectional sample there is a one-to-one relationship between persons and the household in which they live. This relationship does not necessarily apply to longitudinal samples. For example persons who live in households of the longitudinal samples do not belong to the longitudinal person sample since longitudinal households also contain persons who moved into old households after wave one. Conversely there are also longitudinal persons who live in non-longitudinal households such as children who left their parents' household.

Estimating Demographic Events: Death and Out-Migration

The panel enables researchers to estimate the extent of "death" and "out-migration" during a certain time span. These events are recorded in the file *YPBRUTTO*, cumulating drop-outs in all waves since 1984. For more information, see Chapter 4.

If a person dies, then this information is valid for all subsequent panel waves. Therefore the probability that this information remains in the longitudinal sample is one. For example: to calculate the number of persons

[18]This longitudinal definition is consistent with the way the panel survey is actually carried out. The households of the longitudinal sample are the old sample households. An old household retains the same identification number (`HHNRAKT`) across all waves in the SOEP database.

that died between 1984 (wave A) and 1988 (wave E) the weighting factors are assigned as follows:

$$AEPHRFi = \begin{cases} APHRF_i & \text{if person i died after wave A} \\ ABPHRF_i & \text{if person i died after wave B} \\ ACPHRF_i & \text{if person i died after wave C} \\ ADPHRF_i & \text{if person i died after wave D} \end{cases} \tag{5.11}$$

Out-migration is done similarly.

References

Becketti, S., W. Gould, L. Lillard, and F. Welch (1988): "The Panel Study of Income Dynamics after 14 Years: An Evaluation," *Journal of Labor Economics*, 6, 472–492.

Biewen, M. (2002): "The Covariance Structure of East and West German Incomes and its Implications for the Persistence of Poverty and Inequality," Discussion Paper 292, DIW Berlin.

Bishop, J. A., V. K. Chow, and L. A. Zeager (2003): "Decomposing Lorenz and Concentration Curves," *International Economic Review*, 44(3), 965–978.

Blau, D. M., and R. T. Riphahn (1999): "Labor force transitions of older married couples in Germany," *Labour Economics*, 6, 229–251.

Böltken, F. (1996): "Neuabgrenzung von Raumordnungsregionen nach den Gebietsreformen in den neuen Bundesländern," Discussion Paper 5, Bundesforschungsanstalt für Landeskunde und Raumordnung, Bonn.

Borchard-Tuch, C. (2004): "Unzufrieden ohne Job," Die Tageszeitung (Berlin) vom 29.10.2004.

Buechel, F., and J. W. Falter (1994): "Der Einfluss von Langzeitarbeitslosigkeit auf die Parteibindung in der Bundesrepublik Deutschland," *Zeitschrift fuer Parlamentsfragen*, 25, 186–202.

Burkhauser, R. V., B. A. Butrica, M. C. Daly, and D. R. Lillard (2001): "The Cross-National Equivalent File: A Product of Cross-National Research," in *Soziale Sicherung in einer dynamischen Gesellschaft. Festschrift fuer Richard Hauser zum 65. Geburtstag*, ed. by I. Becker, N. Ott, and G. Rolf, pp. 354–376. Campus, Frankfurt/New York.

Burkhauser, R. V., M. Kreyenfeld, and G. G. Wagner (1997): "The German Socio-Economic Panel - A Representative Sample of Reunited Germany and its Parts," *DIW-Vierteljahresbericht*, 66, 7–16.

Butrica, B. A. (1997): "Imputation methods for filling in missing values in the PSID-GSOEP Equivalent File 1980-1994," Cross-National Studies in Aging. Program Project Paper, Center for Policy Research, The Maxwell School. Syracuse, NY: Syracuse University.

Clark, A. E., E. Diener, and Y. Georgellis (2000): "Lags and Leads in Life Satisfaction: A Test of the Baseline Hypothesis," University of Orleans, mimeo.

Clark, W. A. V., M. C. Deurloo, and F. M. Dieleman (1997): "Entry to Home-ownership in Germany: Some comparisons with the United States," *Urban Studies*, 3, 7–19.

Constant, A., and D. S. Massey (2002): "Return Migration by German Guestworkers: Neoclassical versus New Economic Theories,"

International Migration, 40(4), 5–38.

DiPrete, T. A., and H. Engelhardt (2004): "Estimating Causal Effects with Matching Methods in the Presence and Absence of Bias Cancellation," *Sociological Methods and Research*, 32(4), 501–528.

Duncan, G. J., and D. H. Hill (1989): "Assessing the Quality of Household Panel Data - The Care of the Panel Study of Income Dynamics," *Journal of Business and Economic Statistics*, 7(4), 441–452.

Duncan, G. J., and G. Kalton (1987): "Issues of Design and Analysis of Surveys Across Time," *International Statistical Review*, 55, 97–117.

Dustmann, C., and A. vanSoest (2002): "Language Fluency and Earnings: Estimation with Misclassified Language Indicators," *The Review of Economics and Statistics*, 83(4), 663–674.

Fabig, H. (1999): "Income Mobility and the Welfare State: An International Comparison with Panel Data," *Journal of European Social Policy*, 9(4), 331–349.

Frick, J. R., and M. Grabka (2003): "Missing Income Data in the German SOEP," DIW Berlin, SOEP.

Frick, J. R., and T. Schneider (2005): "Biography and Life History Data in the German SOEP," DIW Berlin, SOEP.

Frijters, P., J. P. Haisken-DeNew, and M. Shields (2004a): "How Well Do Individuals Predict Their Future Life Satisfaction? Evidence from Panel Data Following a Nationwide Exogenous Shock," ANU, RWI-Essen, Uni-Melbourne.

Frijters, P., J. P. Haisken-DeNew, and M. A. Shields (2004b): "Investigating the Patterns and Determinants of Life Satisfaction in Germany Following Reunification," *Journal of Human Resources*, 39(3), 649–673.

Galler, H. P. (1987): "Zur Längschnittgewichtung des Sozio-ökonomischen Panel," in *Analysen 1987*, ed. by H.-J. Krupp, and U. Hanefeld, pp. 295–317. Campus, Frankfurt-New York.

Gangl, M. (2004): "Welfare States and the Scar Effects of Unemployment: A Comparative Analysis of the United States and West Germany," *American Journal of Sociology*, 109(6), 1319–1364.

Geishecker, I., and H. Goerg (2004): "International Outsourcing and Wages: Winners and Losers," DIW Berlin, Univ. of Nottingham.

Goodin, R. E., H.-J. Dirven, B. Headey, and R. Muffels (1999): *The Real Worlds of Welfare Capitalism*. Cambridge University Press, Cambridge.

Haisken-DeNew, J. P. (2001): "A Hitchhiker's Guide to the World's Household Panel Data Sets," *The Australian Economic Review*, 34(3), 356–366.

——— (2005): "SOEP Menu: A Menu-Driven Stata/SE Interface for

Accessing the German Socio-Economic Panel," http://www.soepmenu.de, Essen, Germany.

HANEFELD, H. (1987): *Das Sozio-Ökonomische Panel*. Campus, Frankfurt.

HELBERGER, C. (1988): "Eine Überprüfung der Linearitätsannahme der Humankapitaltheorie," in *Bildung, Beruf, Arbeitsmarkt*, ed. by H.-J. Bodenhüfer, pp. 151–170. Duncker und Humblot, Berlin.

HORSTKÖTTER, D., AND R. HÜBNER (2004): "Verteilter Wohlstand," *Capital*, 20, 18–24.

HUEBLER, O., AND A. KOENIG (1999): "Betriebliche Weiterbildung, Mobilitaet und Beschaeftigungsdynamik - Empirische Untersuchungen mit Individual- und Betriebsdaten," *Jahrbuecher fuer Nationaloekonomie und Statistik*, 219(1/2), 165–193.

HUNT, J. (1999): "Has Work-Sharing Worked in Germany?," *Quarterly Journal of Economics*, 114, 1–32.

INFRATEST-SOEP-GRUPPE (1996): "Methodenbericht Zuwandererbefragung II," Infratest-Burke Sozialforschung, München, mimeo.

IRELAND, C., AND S. KULLBACK (1968): "Contingency Tables with Given Marginals," *Biometrika*, 55, 179–186.

JUERGES, H. (1998): "Beruflich bedingte Umzuege von Doppelverdienern - Eine empirische Analyse mit Daten des SOEP," *Zeitschrift für Soziologie*, 27, 358–377.

KRAFT, K. (2001): "Unemployment and the Separation of Married Couples," *Kyklos*, 54(1), 67–88.

KREYENFELD, M. (2000): "Timing of First Births in East Germany after Unification," MPI-Rostock, mimeo.

LECHNER, M. (1998): *Training the East German Labour Force*. Physica Verlag, Heidelberg.

LITTLE, R. J., AND H.-L. SU (1989): "Item Non-Response in Panel Surveys," in *Panel Surveys*, ed. by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, pp. 400–425. John Wiley, New York.

LUCAS, R. E., A. E. CLARK, Y. GEORGELLIS, AND E. DIENER (2003): "Reexamining Adaptation and the Set Point Model of Happiness: Reactions to Changes in Martial Status," *Journal of Personality and Social Psychology*, 84(3), 527–539.

MARGOLIS, D. N., V. SIMONNET, AND L. VILHUBER (2000): "Early Career Experiences and Later Career Outcomes: A Comparison of the United States, France and Germany," York University Toronto, mimeo.

MCGINNITY, F. (2002): "The Labour Force Participation of the Wives of Unemployed Men," *European Sociological Review*, 18(4), 473–488.

MERZ, J. (1983): "Die konsistente Hochrechnung von Mikrodaten nach dem Prinzip des minimalen Informationsverlusts," *Allgemeines Statistisches*

Archiv, 67, 342–366.

Oxley, H., T. T. Dang, and P. Antolin (2001): "Poverty Dynamics in Six OECD Countries," *OECD Economic Studies*, 30(1).

Pannenberg, M. (2002): "Long - Term Effects of Unpaid Overtime: Evidence for West Germany," Discussion Paper 293, DIW Berlin.

Pischner, R. (2000): "Überarbeitete Querschnittshochrechnung der Wellen G-N (1990 bis 1997) des Sozio-ökonomischen Panels (SOEP) unter Einbeziehung der Ergänzungsstichprobe E (Welle O)," DIW Berlin, mimeo.

——— (2005): "Estimate of the Integrated Cross-sectional Weighting (Samples A–G) for 2002-2004," Berlin.

Rendtel, U. (1990): "Teilnahmeentscheidung in Panelstudien: Zwischen Beeinflussung, Vertrauen und sozialer Selektion," *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 42, 280–299.

Rendtel, U., M. Pannenberg, and S. Daschke (1997): "Die Gewichtung der Zuwandererstichporbe des Sozio-ökonomischen Panels," *DIW-Vierteljahresbericht*, 66, 271–285.

Schmid, K.-P. (2002): "Was heisst schon solidarisch? Auch unter der rot-grünen Regierung bleibt die Kluft zwischen Arm und Reich in Deutschland bestehen," Newspaper article, Die Zeit, 14. November 2002, p.28.

Schraepler, J.-P. (2002): "Respondent Behavior in Panel Studies - A Case Study for Item-Nonresponse by Means of the German Socio-Economic Panel (GSOEP)," Discussion Paper 299, DIW Berlin.

Schwarze, J. (1991): "Ausbildung und Einkommen von Männern - Einkommensfunktionsschätzungen für die ehemalige DDR und die Bundesrepublik Deutschland," *Mitteilungen aus der Arbeitsmarkt- und Berufsforschung*, 24, 63–69.

Siebern, F. (2000): "Better LATE? Instrumental Variables Estimation of the Returns to Job Mobility during Transition," *German Economic Review*, 1(3), 335–362.

Siedler, T. (2004): "Is the receipt of social assistance transmitted from parents to children? Evidence from German panel data," DIW Berlin, Essex University.

Sozialforschung, I. (2002): "Lebenslage und Vermgensbildung von Haushalten im oberen Einkommensbereich, Sondererhebung im Rahmen des SOEP 2002, Methodenbericht," Mnchen.

Spiess, M. (2000): "Derivation of Design Weights: The Case of the German Socio-Economic Panel (GSOEP)," Discussion Paper 197, DIW Berlin.

Spiess, M., and M. Pannenberg (2003): "Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (1984 until 2002)," Discussion Paper 28, DIW Berlin Research Notes.

Spiess, M., and U. Rendtel (2000): "Combining an Ongoing Panel with a New Cross-Sectional Sample," Discussion Paper 198, DIW Berlin.

Szydlik, M. (2000): *Lebenslange Solidaritaet? Generationenbeziehungen zwischen erwachsenen Kindern und Eltern*. Leske und Budrich, Opladen.

VanKerm, P. (2004): "What Lies Behind Income Mobility? Reranking and Distributional Change in Belgium, Western Germany and the USA," *Economica*, 71(282), 223–239.

von Rosenbladt, B., and F. Stutz (1998): "SOEP '98. Erstbefragung der Stichprobe E. Methodenbericht," Infratest-Burke Sozialforschung, München, mimeo.

Winkelmann, L., and R. Winkelmann (1993): "Why Are the Unemployed So Unhappy? Evidence from Panel Data," *Economica*, 65, 1–15.

Index

List of Figures

List of Tables

Contract to Use the Public Use Version of the German Socio-Economic Panel

This is a contract between the

Deutsches Institut fuer Wirtschaftsforschung (DIW),
Koenigin-Luise-Strasse 5, D-14195 Berlin, Germany

and

<<User>>, described below as the data recipient.

1 The DIW gives the data recipient the right to use data files of the public use version of the German Socio-Economic Panel (SOEP).

2 The following are restrictions to this right of use:

2.1 The data recipient agrees not to transfer the data files or make them available in any form to other persons or institutions. The only exceptions to this restrictions are collaborators and assistants of the data recipient, who work in the same institution on the project outlined in section 2.3. The same restrictions apply to all researchers who work on the project.

2.2 The data can only be used for scientific research outlined in the data recipient project (section 2.3). Under the contract the use of this data for commercial purposes is strictly forbidden. Commercial purposes as stated here could include purposes for which the data recipient is not paid.

2.3 The use of the data is allowed only for the following research project:

<<Data Recipients's Official Research Project Title>>
The data cannot be used for any other purpose.

2.4 The SOEP data cannot be matched with any other data.

2.5 The material describing the data, which is attached to the data files, must be kept under lock and key and can only be given to entitled data users as described in section 2.1.

2.6 The data recipient must ensure that a password protection system is used to prevent unauthorized access to the SOEP data. The password should be changed regularly. This is true for both mainframe and PC storage of data. Access to the SOEP data from off-site locations, except by machines owned by the data recipient's institution is not permitted.

The data recipient also guarantees the SOEP data will only be used on PCs under the data recipient's supervision. Access to such PCs must be controlled as discussed in section 2.5. All data recording media containing SOEP data are to be kept under lock and key.

2.7 If the tasks are completed for the purpose for which the data were requested, the the data provided as well as any backup copies, extract files and help files are to be erased. The DIW is to be immediately notified upon completion of the project listed under section 2.3. The obligation to protect the privacy of SOEP data stays with the data recipient until the data are erased and the project is declared completed.

2.8 The data recipient agrees not to try to identify or publish single data records. The data recipient must agree to conform to the German law concerning the protection of the privacy of personal data.

2.9 The data recipient agrees to send the DIW one copy of all research papers developed from the SOEP data at no charge.

3 The DIW agrees to produce a public use SOEP without charge, but the data recipient agrees to pay for the materials used in copying the data.

4 In all disagreements concerning the interpretation of the right to use the SOEP data, the DIW reserves the right to the final decision.

5 The right to use the data ends if the data recipient leaves the principal institute he or she was associated with at the time this contract was signed, or if that institute is dissolved, taken over by new management or becomes a new institute. The data provided, as well as any backup copies are to be erased. The DIW must be notified of all such changes. Otherwise the DIW can unilaterally withdraw the right to use the SOEP data.

6 To be valid, any changes or supplements to this contract must be made in writing. It is agreed that if adjudication is necessary, it must occur in Berlin.

Berlin,
Head of Administration
German Institute for Economic Research, Berlin

<<Date>>

<<User's Signature>>

<<User>>

<<User's Position>>

<<User's Institute>>

Sworn Statement of Upholding Data Secrecy and Security

<<User>>

<<User's Institute>>

swears to uphold data secrecy and data security as defined by Article 5 of the German Data Protection Law (BDSG-Bundesdatenschutzgesetz).

The following is noted:

According to Article 5 of the German Data Protection Law (BDSG), it is illegal to use, or process individual/person information without rightful authorization. That is to say that such information may be saved, changed, transferred, locked or deleted or used in any other way, only to the extent that you are legally authorized to do so. All unauthorized data use or processing for any other purposes is strictly forbidden.

Authorized users, in the course of their authorized use of the data must, must provide the necessary data security. Known abuses or security breaches are to be reported immediately to the institute's data security officer and the DIW.

Data security abuses, according to Article 43 of the German Data Protection Law (BDSG), can be punished with fines and/or imprisonment. In most cases, data security abuses can also have serious employment consequences, such as diciplinary actions and employment termination of the offending parties.

Questions concerning data security will be happily answered by your institute's data security officer and the DIW.

Please confirm with your signature that you have read this document, and understand your obligations and responsibilities in upholding data security.

<<Date>>

<<User's Signature>>

<<User>>