SOEP The Socio-Economic Panel

743

SOEP Survey Papers

Series G – General Issues and Teaching Materials

SOEP – The German Socio-Economic Panel at DIW Berlin 2019

SOEPcompanion (v34), V.2

Selin Kara, Stefan Zimmermann, and SOEP Group

DIW SOEP

Running since 1984, the German Socio-Economic Panel study (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing.

The SOEP Survey Papers is comprised of the following series:
Series A – Survey Instruments (Erhebungsinstrumente)
Series B – Survey Reports (Methodenberichte)
Series C – Data Documentation (Datendokumentationen)
Series D – Variable Descriptions and Coding
Series E – SOEPmonitors
Series F – SOEP Newsletters
Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at **http://www.diw.de/soepsurveypapers**

Please cite this paper as follows:

Selin Kara, Stefan Zimmermann, and SOEP Group. 2019. SOEPcompanion (v34), v.2. SOEP Survey Papers 743: Series G. Berlin: DIW/SOEP.

ISSN: 2193-5580 (online)

DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany

soeppapers@diw.de

SOEPcompanion (v34), Release 2019, v.2

Selin Kara, Stefan Zimmermann, and SOEP Group

CONTENTS

CHAPTER

ONE

PREFACE

SOEP-Core is the centerpiece of the Socio-Economic Panel, a wide-ranging representative longitudinal study of private households in Germany, based at the German Institute for Economic Research, DIW Berlin. SOEP-Core was started in 1984, and in 1990—shortly after German reunification—it was enlarged to include a representative sample from East Germany. This feature makes the SOEP unique among household panel surveys worldwide. Every year since 1984, around 15,000 households and about 30,000 individuals have been surveyed by the SOEP's fieldwork organization, Kantar Public Germany. The data provide information on every member of every household taking part in the survey. Respondents include Germans living in both the former East and West Germany, foreign citizens residing in Germany, recent immigrants, and a new sample of refugees added in 2016. Some of the many topics include household composition, education, occupational biographies, employment, earnings, health, and satisfaction indicators.

The SOEPcompanion describes the current version of the SOEP-Core data (v34) and introduces users to the different SOEP-Core data structures. It also provides applications in Stata as well as instructions on how to use our various documentation services. We plan to revise the information in the SOEPcompanion annually to continue providing users a comprehensive, up-to-date introductory understanding of the SOEP.

We know that starting to use any new dataset is difficult, and this is especially true of panel data given their complexity. We hope that this introduction will help. We always welcome any feedback or tips on how to improve our documentation.

- Recommendation of our most recent version of a general short description of SOEP study: The German Socio-Economic Panel Study (SOEP)
- To the information system for efficient working with complex datasets: paneldata.org

CHAPTER

TWO

TOPICS OF SOEP-CORE

The topics of the SOEP questionnaires and the various modules they contain can be grouped into 11 areas. Some of the modules deal with aspects of life that tend to change from one year to the next, and are therefore repeated annually, while other modules are repeated at intervals of several years. How often a module is repeated is stated in the "replication" column of our topic tables. Some SOEP modules are also adapted in different ways to the different questionnaires. The questions in the "Big Five" personality traits module, for instance, are formulated differently in the mother-child questionnaires than they are in the individual questionnaire.

Note: The tables list the modules in a questionnaire, not the question items. The years specified in the "replication" column may therefore not apply to every variable, since some items may have been added later or removed. The modules listed do not represent all of the variables in the SOEP, nor do they refer to all questionnaires. Specific information can also be found in our generated datasets.

SOEP Topics

Overview of Modules in Different SOEP Questionnaires

	Individual	*Youth*	*Mother-Child A*	*Mother-Child B*	*Mother-Child C*	*Parents D*	*Mother-Child E*
Affective well-being	x	x					
Big Five personality traits	x	x		x	x		x
Birth history		x	x	x	x	x	x
Childcare			x	x	x	x	x
Educational aspirations		x				x	x
Health of child			x	x	x		x
Height and weight of child			x	x	x		
Height and weight	x	x					
Language ability German / native language	x	x					
Leisure and activities (with child)				x	x		
Life satisfaction	x	x					
Language use				x			x
Locus of control	x	x					
Family background	x	x					
Parental interest in school performance		x					x
Allowance		x					x
Political orientation	x	x					
Risk aversion	x	x					
State of health	x	x					
Strengths and difficulties					x		x
Temperament			x	x			

The SOEP Scales Manual briefly describes the theoretical background and development of all of the scales used in the Socio-Economic Panel (SOEP) study. It also provides the relevant citations as well as the items belonging to the scales and the answer format, including the verbal anchors.

2.1 Demography and Population

The demography and population topic provides information including the birth date and sex of each household member (including children) and of each interviewer; the place and history of births in the household; household size; and relationships among household members.

Questionnaire	Module	Replication	No. Vars.	Variables
Mother and Child Instruments				
	Birth history	annually	2	birthm & birthy
Individual Questionnaire				
	Country of Origin	annually	7	plj0014_v1 - plj0014_v3 , plj0022 - plj0024, plj0175
Youth Questionnaire				
	Birth history	annually	7	jl0233 , jl0234 , jl0235_v1, jl0235_v2, jl0235_h, jl0238 , jl0239
	Country of Origin	annually	8	jl0240- jl0246 , jl0419, jl0445

2.2 Work and Employment

The work and employment modules provide information on diverse employment-related topics including the respondent's first job, further training, changes in working conditions following parenthood, part-time work, and unemployment. Modules cover not only objective information such as working hours but also subjective perceptions of working conditions and feelings about work.

Questionnaire	Module	Replication	No. Vars.	Variables
Individual Questionnaire				
	Employment / education calendar	annually	12	pab0001- pab0008, pab010- pab0013
	Care period (Pflegezeit)	annually	1	plb0020
	Change of job	annually	19	plb0031_v1 , plb0031_v2 , plb0031_h, plb0032 , plb0033_v1- plb0033_v7 , plb0034 , plb0478- plb0480 , plb0284_v1, plb0284_v2, plb0284_h, plb0295
	Commuting to work	1991-2013, 2015, 2017	8	plb0589- plb0592, plb0158, plb0159_v1, plb0159_v2, plb0159_h

Continued on next page

Table 1 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Work contract (Werkvertrag)	2013, 2015	1	plb0482
	Current job	annually	24	plb0035 , plb0036_v1 , plb0036_v2 , plb0036_h, plb0037_v1 - plb0037_v3 , plb0037_h , plb0040 , plb0041 , plb0049_v1 - plb0049_v5, plb0058 , plb0063_v1, plb0063_v2 , plb0064_v1, plb0064_v2, plb0065 , plb0568 , plb0570 , plb0586
	Gross / net income, October 2014	2015	2	plb0584 , plb0585
	Employees' council (Betriebsrat)	2001, 2006, 2011, 2016	1	plb0050
	Employment, October 2014	2015	1	plb0574
	Employment status	annually	10	plb0022_v1- plb0022_v9, plb0022_h
	Evening and weekend work	2005, 2007, 2009, 2011, 2013, 2015, 2017	10	plb0205_v1- plb0205_v4, plb0206_v1- plb0206_v4, plb0218- plb0219
	Work in black economy	2015, 2016	2	plb0571 , plb0572
	Changes in workplace tools / technologies	2015, 2016, 2017	6	plb0595- plb0600
	Labor intensity	2015, 2016, 2017	2	plb0593, plb0594
	Job search	annually	15	plb0362 , plb0358_v1- plb0358_v13, plb0358_h
	Job search, motives	1994 - 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2017	1	plb0111
	Job search, preferences	1994, 1996 - 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017	1	plb0426
	Hourly wages	2017		
	Supervisory position	2007, 2009, 2011, 2013, 2015, 2017	3	plb0067- plb0069
	Maternity / parental leave	annually	1	plb0019_v1, plb0019_v2
	Occupational expectations, non-employed	1999, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015	3	plb0427- plb0429
	Use of professional skills in job	1985-2007, 2009	1	plb0357

Continued on next page

Table 1 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Working overtime	annually	16	plb0193 - plb0194 , plb0195_v1- plb0195_v3, plb0195_h, plb0196_v1- plb0196_v3, plb0196_h, plb0197 - plb0198 , plb0483 , plb0484 , plb0220 , plb0605
	Overtime, October 2014	2015	2	plb0582 , plb0583
	Financial compensation for overtime	1984-2014	3	plb0195_v1 - plb0195_v3 , plb0195_h
	Performance evaluation by superior	2004, 2008, 2011, 2016	5	plb0098- plb0102
	Workplace zip code	2016		
	Industry sector, occupational classification	annually	12	plb0072 , plb0073_v1- plb0073_v8, plb0073_h, p_nace , p_isco08
	Professional expectations, long	1985, 1987, 1989, 1991, 1993, 1994, 1996, 1998, 2000, 2005, 2009, 2013	22	plb0432_v1 , plb0432_v2 , plb0433_v1 , plb0433_v2, plb0434_v1, plb0434_v2, plb0435_v1 , plb0435_v2, plb0436_v1 , plb0436_v2, plb0437_v1, plb0437_v2, plb0438_v1, plb0438_v2, plb0439_v1, plb0439_v2, plb0440_v1, plb0440_v2, plb0441_v1, plb0441_v2, plb0442_v1 , plb0442_v2
	Professional expectations, short	2015	7	plb0433_v1, plb0433_v2, plb0437_v1, plb0437_v2, plb0440_v1, plb0440_v2, plb0588
	Leaving a job	annually	24	plb0282 , plb0298- , plb0304_h, plb0304_v1 - plb0304_v15, plb0305_v1, plb0305_v2, plc0040, plc0041_h, plc0041_v1, plc0041_v2
	Registered unemployed	annually	1	plb0021

Continued on next page

Table 1 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Side jobs	annually	25	plb0382_v1- plb0382_v5, plb0382_h, plb0392 - plb0398 , plb0573 , plc0062_v1, plc0062_v2, plc0062_h, p_isco08_sidejob, p_isco08_sidejob1, p_isco08_sidejob2, p_isco08_sidejob3, p_isco88_sidejob, p_isco88_sidejob1, p_isco88_sidejob2, p_isco88_sidejob3
	Self-employment, reasons	2010, 2015	6	plb0333- plb0338
	Short-time compensation (Kurzarbeitergeld)	1984-2001, 2003-2005, 2010, 2011	5	plc0057_v1, plc0057_v2, plc0057_h, plc0058_v1, plc0058_v2
	Standby duty	2011, 2014-2017	4	plb0212- plb0215
	Starting a new job	annually	13	plb0417_v1, plb0417_v2, plb0418 , plb0419_v1, plb0419_v2, plb0419_h, plb0420_v1, plb0420_v2, plb0421 - plb0423 , plb0424_v1, plb0424_v2, plb0240
	Start of working hours	2002, 2004, 2006, 2008, 2012, 2015, 2017	3	plb0180- plb0182
	Vacation entitlement	2000, 2005, 2010	8	plb0269- plb0276
	Entitlement to paid breaks	2015, 2016, 2017	3	plb0601- plb0603
	Paid breaks, October 2014	2015	4	plb0575 - plb0578
	Work from home	1997, 1999, 2002, 2009, 2014	3	plb0095- plb0097
	Work time regulations	2003, 2005, 2007, 2009, 2011, 2014-2017	1	plb0211
	Work, last 7 days	annually	1	plb0018
	Contractual working hours	annually	9	plb0185 , plb0186_v1, plb0186_v2, plb0186_h, plb0241_v1, plb0241_v2, plb0241_h, plb0209 , plb0210
	Working hours, October 2014	2015	4	plb0579, plb0579_h, plb0581, plb0581_h, plb0580
	Workload (effort-reward imbalance)	2001, 2006, 2011, 2016	26	plb0112- plb0137
Youth Questionnaire				

Continued on next page

Table 1 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Jobs and money	annually	11	jl0013 , jl0014, jl0017 - jl0019 , jl0023 , jl0024_h , jl0024_v1 , jl0024_v2 , jl0025 , jl0385 - jl0387

2.3 Income, Taxes, and Social Security

The income, taxes, and social security modules collect wide-ranging financial information from earnings and spending to public benefits, pensions, inheritances, taxes, and debts. They also cover assets such as real estate and other property.

Questionnaire	Module	Replication	No. Vars.	Variables
Household Questionnaire	Financial support to relatives	2010	4	hlc0091 , hlc0092, hld0004-hld0005
	Loan repayment	annually	1	hlc0115
	Expenditures on food	1998, 2000, 2001, 2003, 2005, 2007, 2009, 2011, 2016	6	hlf0435_v1 , hlf0435_v2 , hlf0435_h , hlf0436_v1 , hlf0436_v2 , hlf0436_h
	Good / adequate / bad household income	1992, 1997, 2007	18	hlc0018_v1 , hlc0018_v2 , hlc0018_h , hlc0019_v1 , hlc0019_v2 , hlc0019_h , hlc0020_v1 , hlc0020_v2 , hlc0020_h , hlc0021_v1 , hlc0021_v2 , hlc0021_h , hlc0022_v1 , hlc0022_v2 , hlc0022_h , hlc0023_v1 , hlc0023_v2 , hlc0023_h
	Income and expenses from renting / leasing	annually	10	hlc0007 , hlc0008_v1 , hlc0008_v2 , hlc0008_h , hlc0009 , hlc0111_h , hlc0111_v1 , hlc0111_v2 , hlc0112_h , hlc0112_v1 , hlc0112_v2 , hlc0176 , hlc0177

Continued on next page

Table 2 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Household income / expenses	annually	80	hlc0005_v1 , hlc0005_v2 , hlc0005_h , hlc0006_v1 - hlc0006_v3 , hlc0039_v1 - hlc0039_v3 , hlc0039_h , hlc0040 , hlc0041 , hlc0042_v1 , hlc0042_v2 , hlc0042_h , hlc0043 , hlc0044_v1 , hlc0044_h , hlc0045_v1 , hlc0045_v2 , hlc0045_h , hlc0046 , hlc0047 , hlc0049 - hlc0054 , hlc0055_v1 - hlc0055_v3 , hlc0055_h , hlc0056_v1 - hlc0056_v3 , hlc0056_h , hlc0057 , hlc0058 , hlc0059_v1 , hlc0059_v2 , hlc0059_h , hlc0060_v1 , hlc0060_v2 , hlc0060_h , hlc0061_h , hlc0061_v1 , hlc0061_v2 , hlc0062 - hlc0065 , hlc0066_v1 , hlc0066_v2 , hlc0066_h , hlc0067_v1 , hlc0067_v2 , hlc0067_h , hlc0068_v1 , hlc0068_v2 , hlc0068_h , hlc0070 , hlc0071 , hlc0077 , hlc0078 , hlc0079_v1 , hlc0079_v2 , hlc0079_h , hlc0080_v1 , hlc0080_v2 , hlc0080_h , hlc0081, hlc0082_v1 , hlc0082_v2 , hlc0082_h , hlc0083_v1 , hlc0083_v2 , hlc0083_h , hlc0084_v1 , hlc0084_v2 , hlc0084_h , hlc0085_h , hlc0085_v1 - hlc0085_v6 , hlc0090_v1 , hlc0090_v2 , hlc0090_h , hlc0121 - hlc0125
	Inheritances, gifts, lottery winnings	annually	6	hlc0178 - hlc0183
	Investments	annually	14	hlc0104 - hlc0108 , hlc0093 - hlc0098 , hlc0013_v1 , hlc0013_v2 , hlc0014 , hlc0184
	Repayment of loans	annually	4	hlc0113 , hlc0114_v1 , hlc0114_v2 , hlc0114_h
	Savings	annually	3	hlc0172 - hlc0174

Continued on next page

Table 2 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Ratio between income and expenditures	2010-2013,2015-2017 annually	13	hlc0024_v1 , hlc0024_v2 , hlc0025_v1 , hlc0025_v2 , hlc0026_v1 , hlc0026_v2 , hlc0027_v1 , hlc0027_v2 , hlc0028_v1 , hlc0028_v2 , hlc0029_v1 , hlc0029_v2 , hlc0030
Individual Questionnaire				
	Additional questions for employed people	annually	25	plc0042 , plc0043_v1 , plc0043_v2 , plc0043_h - plc0044 , plc0045_v1 , plc0045_v2 , plc0045_h , plc0046 , plc0047_v1 , plc0047_v2 , plc0047_h , plc0048 , plc0049_v1 , plc0049_v2 , plc0049_h , plc0050 , plc0051_v1 , plc0051_v2 , plc0051_h , plc0052 , plc0053_v1 , plc0053_v2 , plc0053_h , plc0054
	Additional questions for retirees / pensioners	annually	48	plc0223_v1 , plc0223_v2 , plc0223_h , plc0236_v1 , plc0236_v2 , plc0236_h , plc0238_v1 , plc0238_v2 , plc0238_h , plc0240_v1 , plc0240_v2 , plc0240_h , plc0242 , plc0243_v1 , plc0243_v2 , plc0243_h , plc0245_v1 , plc0245_v2 , plc0245_h , plc0247_v1 , plc0247_v2 , plc0247_h , plc0249_v1 , plc0249_v2 , plc0249_h , plc0251 , plc0278_v1 , plc0278_v2 , plc0278_h , plc0279_v1 , plc0279_v2 , plc0279_h , plc0281_v1 , plc0281_v2 , plc0281_h , plc0283_v1 , plc0283_v2 , plc0283_h , plc0285 , plc0286_v1 , plc0286_v2 , plc0286_h , plc0288_v1 , plc0288_v2 , plc0288_h , plc0290 , plc0516 , plc0517

Continued on next page

Table 2 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Asset balance	1988, 2002, 2007, 2012, 2017	69	plc0315 , plc0316, plc0317_v1 , plc0317_v2 , plc0318_v1 , plc0318_v2 , plc0319 , plc0328 - plc0374 , plc0411 - plc0425
	Benefits and bonuses from employer	2008, 2010, 2012, 2014-2017	14	plc0026 - plc0039
	Financial advantages from use of company car	2016,2017	1	plc0532
	Gross / net income, collective wage agreements	annually	10	plc0013_v1 , plc0013_v2 , plc0013_h , plc0014_v1 , plc0014_v2 , plc0014_h , plc0506 - plc0509
	Pension entitlements, company	2013	5	plc0441 - plc0445
	Pension entitlements, statutory	2013	4	plc0432- plc0435

Continued on next page

Table 2 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Income	annually	72	plc0015_v1 , plc0015_v2 , plc0015_h , plc0016 , plc0017_h , plc0017_v1 , plc0017_v2 , plc0064 , plc0065 , plc0073 , plc0074 , plc0075_v1 , plc0075_v2 , plc0075_h , plc0116 , plc0117 , plc0126 , plc0130 , plc0131_v1 , plc0131_v2 , plc0131_h , plc0132_v1 , plc0132_v2 , plc0132_h , plc0135 , plc0136 , plc0137_v1 , plc0137_v2 , plc0137_h , plc0138 , plc0139 , plc0152 , plc0153_v1 , plc0153_v2 , plc0153_h , plc0154 plc0155_h , plc0155_v1 , plc0155_v2 , plc0168_v1 , plc0168_v2 , plc0168_h , plc0169 , plc0170 , plc0171_h , plc0171_v1 , plc0171_v2 , plc0177 , plc0178 , plc0181 - plc0184 , plc0188_v1 , plc0188_v2 , plc0189_v1 , plc0189_v2 , plc0190_v1 , plc0190_v2 , plc0198_v1 , plc0198_v2 , plc0198_h , plc0202 , plc0203_v1 , plc0203_v2 , plc0203_h , plc0204 , plc0205 , plc0232 , plc0233_v1 , plc0233_v2 , plc0233_h , plc0234 , plc0235 , plc0273 , plc0274_v1 , plc0274_v2 , plc0274_h , plc0275 , plc0276 , plc0488 - plc0492 , plc0513 , plc0514, plc0515 , plb0471_v1 , plb0471_v2 , plb0471_h , plb0474_v1 , plb0474_v2 , plb0474_h , plb0477_v1 , plb0477_v2 , plb0477_h

Continued on next page

Table 2 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Inheritances	2001,2017	63	plc0375_v1 , plc0375_v2 , plc0376_v1 , plc0376_v2 , plc0377_v1 , plc0377_v2 , plc0378_v1 , plc0378_v2 , plc0379_v1 , plc0379_v2 , plc0380_v1 , plc0380_v2 , plc0381_v1 , plc0381_v2 , plc0382_v1 , plc0382_v2 , plc0383_v1 , plc0383_v2 , plc0383_h1 , plc0384_v1 , plc0384_v2 , plc0385 , plc0386_v1 , plc0386_v2 , plc0387_v1 , plc0387_v2 , plc0388_v1 , plc0388_v2 , plc0389_v1 , plc0389_v2 , plc0390_v1 , plc0390_v2 , plc0391_v1 , plc0391_v2 , plc0392_v1 , plc0392_v2 , plc0393_v1 , plc0393_v2 , plc0393_h1 , plc0394_v1 , plc0394_v2 , plc0395 , plc0396_v1 , plc0396_v2 , plc0397_v1 , plc0397_v2 , plc0398_v1 , plc0398_v2 , plc0399_v1 , plc0399_v2 , plc0400_v1 , plc0400_v2 , plc0401_v1 , plc0401_v2 , plc0402_v1 , plc0402_v2 , plc0403_v1 , plc0403_v2 , plc0403_h1 , plc0404_v1 , plc0404_v2 , plc0405 - plc0407
	Riester / Ruerup pension plans	2004, 2006, 2007, 2010, 2012, 2013, 2015, 2017	3	plc0430 , plc0431 , plc0313
	Pension payments	2013	3	plc0437 - plc0439
	Social security	1987, 1992, 1997, 2007, 2012, 2017	7	plc0008, plc0009, plc0111-plc0115
	Financial support to relatives or others	annually	31	plj0131 , plj0132_v1 , plj0132_v2 , plj0132_h , plj0133 - plj0135 , plj0136_v1 , plj0136_v2 , plj0136_h , plj0137 - plj0139, plj0140_v1 , plj0140_v2 , plj0140_h , plj0141 - plj0143 , plj0144_v1 , plj0144_v2 , plj0144_h , plj0145 - plj0147 , plj0148_v1 , plj0148_v2 , plj0148_h , plj0149 - plj0151

Continued on next page

Table 2 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Financial support received	2009, 2010, 2011	21	plj0152- plj0172
	Wage tax classification	1991, 1993, 2004, 2016	10	plc0091_v1 - plc0091_v9 , plc0091_h

2.4 Family and Social Networks

As a household study, the SOEP offers rich information on family and social relationships and how these connections change in different stages of life. The modules dealing with family and social networks cover the entire life cycle beginning with pregnancy and childbirth and continuing through parenthood, family formation, friendships, marriage, divorce, and death, and also provide a wealth of additional information on important life events.

Questionnaire	Module	Replication	No. Vars.	Variables
Individual Questionnaire				
	Family changes	annually	42	pld0012 - pld0014 , pld0038 - pld0040 , pld0134 - pld0156 , pld0158 - pld0171
	Family network	1991, 1996, 2001, 2006, 2011, 2016	59	pld0020 - pld0036 , pld0107 - pld0118 , plj0117_v1- plj0117_v3, plj0118_v1 - plj0118_v3, plj0119_v1 - plj0119_v3, plj0120 , plj0121 , plj0122_v1 - plj0122_v3, plj0123_v1 - plj0123_v3, plj0124 , plj0125_v1 - plj0125_v3 plj0126 , plj0127_v1 - plj0127_v3, plj0128 , plj0129 , plj0130_v1 - plj0130_v3
	Marital / partnership status	annually	5	pld0131 , pld0132_v1, pld0132_v2, pld0133, plk0001

Continued on next page

Table 3 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Friends	2003, 2008, 2011, 2013, 2015, 2017, 2018	1	pld0047
	Gender attitudes	2018	8	
	Sexual orientation	2016	1	pld0298
	Circle of friends, sociodemographics	2006, 2011, 2016	18	pld0089 - pld0106
	Circle of friends, trusted friends	1991, 1996, 2001, 2006, 2011, 2016	50	pld0062_v1 , pld0062_v2 , pld0063_v1 , pld0063_v2, pld0064_v1, pld0064_v2, pld0065_v1, pld0065_v2, pld0066_v1, pld0066_v2, pld0067 , pld0068_v1, pld0068_v2, pld0069_v1, pld0069_v2, pld0070_v1, pld0070_v2, pld0071_v1, pld0071_v2, pld0072_v1, pld0072_v2, pld0073 - pld0082 , pld0083_v1, pld0083_v2, pld0084_v1, pld0084_v2, pld0085_v1, pld0085_v2, pld0086_v1, pld0086_v2 ‘pld0086_v2, pld0087_v1, pld0087_v2, pld0088 , plf0049_v1, plf0049_v3, plf0049_h, plf0050_v1, plf0050_v3, plf0050_h
Youth Questionnaire				
	Childhood, parental home	annually	121	jl0273 , jl0279 - jl0304, jl0307 - jl0312, jl0313_v1, jl0313_v2, jl0314_v1, jl0314_v2, jl0315 , jl0316 , jl0327_v1, jl0327_v2, jl0327_h, jl0328_v1, jl0328_v2, jl0328_h, jl0446 , jl0447, jl1406 - jl1411 , jl0454 - jl0495 , jl0506 - jl0523 , j_kldb2010_jobfather, j_kldb92_jobfather, j_isco08_jobfather, j_isco88_jobfather, j_kldb2010_jobmother, j_kldb92_jobmother, j_isco08_jobmother, j_isco88_jobmother
	Parental interest in child’s performance in school	annually	8	jl0167 - jl0174

Continued on next page

Table 3 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Allowance	annually	7	jl0020 , jl0021_v1, jl0021_v2, jl0021_h, jl0022_v1, jl0022_v2, jl0022_h
	Relationship to family members	annually	28	jl0026 - jl0041 , jl0043 - jl0055 , jl1043
Mother and Child Questionnaire (Newborns)				
	Attitude towards maternal role	annually	9	health , change1 - change8
	Childcare	annually	10	suppartn , maincare , care1h , care3h - care6h , care8h , care12h , care19
	Breastfeeding	annually	3	breastf , breastfm , breastfc
	Pregnancy	annually	4	delivpl , birthpw , nchild , pregplan
	Relationship to other parent or children	annually	1	biochild
Mother and Child Questionnaire (2-3-year-olds)				
	Childcare	annually	10	care1h - care8h , care12h , care19
	Leisure and activities (with child)	annually	11	activ1 - activ9 , tvyn , tvhrs
	Language use	annually	1	language
	Breastfeeding	annually	3	breastf , breastfm , breastfc
Mother and Child Questionnaire (5-6-year-olds)				
	Childcare	annually	10	care1h - care7h , care10h , care12h , care19
	Leisure and activities (with child)	annually	13	activ1 - activ4 , activ6 - activ8 , activ10 - activ13 , tvyn , tvhrs
Parents and Child Questionnaire (7-8-year-olds)				
	Childcare	annually	13	maincare , care1h - care7h , care9h - care12h , care19
	Parenting goals	annually	18	edgoal1 - edgoal18
	Attitude toward parental role	annually	9	bepar1 - bepar6 , bepar8 - bepar10
	Parenting style	annually	18	edbeh1 - edbeh18

Continued on next page

Table 3 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Relationship to other parent or children	annually		
Mother and Child Questionnaire (9-10-year-olds)				
	Childcare	annually	12	maincare , care1h - care5h , care7h , care9h - care12h , care19
	Eating behavior	annually	10	eatweek1 - eatweek3 , eatsat1 - eatsat3 , eatson1 - eatson3
	Frequency of leisure and other activities	annually	20	freqact1 - freqact20
	Friends	annually	2	frndchld , frndadlt
	Language use	annually		language
	Parental interest in child's performance in school	annually	7	conscho1 - conscho7
	Allowance	annually	3	allow , allowpw , allowpm

2.5 Health and Care

The modules on health and care cover doctor visits, sports and fitness, alcohol consumption, health insurance, health status, and grip strength, both on respondents themselves and on other individuals in the household, such as children and deceased household members.

Questionnaire	Module	Replication	No. Vars.	Variables
Household Questionnaire				
	Satisfaction with availability of care	2002	1	hlf0318
Individual Questionnaire				

Continued on next page

Table 4 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Additional private insurance	2011-2014, 2016, 2018	11	ple0127 , ple0128_v1 , ple0128_v2 , ple0128_h , ple0129 - ple0134
	Alcohol consumption	2006, 2008, 2010, 2016	2	ple0090 - ple0093 , ple0177 , ple0178
	Qualification for additional benefits	1999-2011	1	ple0121
	Disabilities in everyday life (SF-12)	1997-2002, 2004-2018 (every two years)	2	ple0004 , ple0005
	Health insurance	annually	17	ple0097 , ple0098 , ple0099_v1 - ple0099_v5 , ple0099_h , ple0104_v1 - ple0104_v7 , ple0104_h , ple0160
	Health insurance debts	2017	1	plc0567
	Health restrictions	2011, 2012, 2013, 2015, 2017	2	ple0009 , ple0162
	Height and weight	2002 - 2018 (every two years)	2	ple0006 , ple0007
	Hospital stays	annually	3	ple0053 , ple0055 , ple0056
	Illness	2009, 2011, 2013, 2015, 2017	14	ple0011 - ple0024
	Individual health services	2016,2018	1	ple0186
	Nutritional awareness	2004-2016 (every two years)	4	ple0179 - ple0182
	Private supplementary care insurance	2016,2018	3	ple0183 - ple0185
	Sickness notifications to employer	annually	15	ple0044_v1 , ple0044_v2 , ple0044_h , ple0046 , ple0048 - ple0052 , ple0174 , ple0175 , plb0024_v1 - plb0024_v3 , plb0024_h
	Smoking	1998, 1999, 2001, 2002-2018 (every two years)	14	ple0080_v1 - ple0080_v3 , ple0081 - ple0085 , ple0086_v1 - ple0086_v4 , ple0089 , ple0176
	State of health	annually	1	ple0008
	Stress and exhaustion (SF-12)	2002 - 2018 (every two years)	10	ple0026 – ple0036
	Reduced ability to work	2001, 2002, 2004, 2006, 2008, 2010, 2015	2	ple0040 – ple0041
	Visits to the doctor	annually	2	ple0072 , ple0073
Youth Questionnaire				
	Height and weight	annually	2	jl0219 , jl0220
	State of health	annually	1	jl0218

Continued on next page

Table 4 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
Mother and Child Questionnaire (Newborns)				
	Health of child	annually	12	hospital3mb , medaid3mb , lstmedex , disord , disord1 - disord9
	Height and weight of child	annually	3	weightb , height , weight
	Physical and mental health of mother	annually	4	feeling1 - feeling4
Mother and Child Questionnaire (2-3-year-olds)				
	Health of child	annually	14	hospital12m , ill2 , ill4 - ill9 , ill11 - ill14 , ill31 , illno
	Height and weight of child	annually	3	weightb , height , weight
Mother and Child Questionnaire (5-6-year-olds)				
	Health of child	annually	11	medaid3m , hospital12m , ill0 , ill2 , ill4 , ill5 , ill7 - ill10 , illno
	Height and weight of child	annually	3	weightb , height , weight
Mother and Child Questionnaire (9-10-year-olds)				
	Health of child	annually	11	chhealth , medaid3m , hospital12m , ill4 , ill5 , ill7 - ill10 , ill32 , illno

2.6 Home, Amenities, and Contributions of Private HH

The housing, amenities, and household expenses modules provide wide-ranging information on everyday life including the type of dwelling and whether it is a rental property or owner-occupied; expenditures on personal hygiene, transportation, and vacations; and the division of household labor.

Questionnaire	Module	Replication	No. Vars.	Variables
Household Questionnaire				
	Number of books in household	2001, 2006, 2011, 2016	1	hlf0197
	Home fixtures and furnishings	annually	13	hlf0023 - hlf0037 , hlf0529 - hlf0531
	Homeowner	annually	4	hlf0013_v1 , hlf0013_v2 , hlf0013_v3 , hlf0013_h
	Home ownership / rental	annually	12	hlf0001_v1 , hlf0001_v2 , hlf0001_v3 , hlf0001_h , hlf0006 , hlf0007_v1 , hlf0007_v2 , hlf0007_v3 , hlf0007_v4 , hlf0007_h , hlf0009 , hlf0015
	Childcare situation	annually	6	ks_asc_r , kc_relaz , kc_frdn , kc_paid , kc_mindr , kc_none
	Change in residential situation	annually	5	hlf0523 , hlf0106 , hlf0107_v1 , hlf0107_v2 , hlf0107_h
	Cleaning or household help	annually	2	hlf0261 , hlf0262
	Household expenditures	2010	131	hlf0163_v1 , hlf0163_v2 , hlf0163_h , hlf0164- hlf0172, hlf0209- hlf0238 , hlf0239_v1 , hlf0239_v2 , hlf0239_v3 , hlf0239_v4 , hlf0239_h , hlf0240 , hlf0242 , hlf0241_v1 , hlf0241_v2 , hlf0241_v3 , hlf0241_v4 , hlf0241_v5 , hlf0241_h , hlf0242 - hlf0252 , hlf0159 , hlf0371- hlf0434
	Hereditary lease interest	annually	2	hlf0597 , hlf0598
	Dwelling / building type	annually	6	hlf0154 , hlf0016 , hlf0155_v1 , hlf0155_v2 , hlf0155_h , hlf0596

Continued on next page

Table 5 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Changes in home fixtures and furnishings since last year	1998, 1990, 1992, 1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010	43	hlf0163_v1 , hlf0163_v2 , hlf0163_h , hlf0164-hlf0167, hlf0209 , hlf0212, hlf0214 , hlf0215, hlf0217 , hlf0218, hlf0159 , hlf0223, hlf0228 , hlf0229, hlf0231 , hlc0116-hlc0118 , hlf0233, hlf0236 , hlf0237, hlf0169 , hlf0170, hlf0239_v1 , hlf0239_v2 , hlf0239_v3 , hlf0239_v4 , hlf0239_h , hlf0240 , hlf0241_v1 , hlf0241_v2 , hlf0241_v3 , hlf0241_v4 , hlf0241_v5 , hlf0241_h , hlf0242, hlf0244 , hlf0245, hlf0247 , hlf0248
	Leisure activities and costs, children	2006, 2008, 2010, 2012, 2014, 2016	19	ka06_spo, ka06_mus, ka06_art, ka06_oth, ka06_non, ka16_ssp, ka16_smu, ka16_sar, ka16_sth, ka16_sot, ka16_spo, ka16_mus, ka16_art, ka16_org, ka16_yth, ka16_ctr, ka16_non, kk_amtp, kk_cost
	Neighborhood amenities	1986, 1994, 1999, 2004, 2009, 2014	23	hlf0135-hlf0152, hlj0004_v1 , hlj0004_v2 , hld0001 - hld0003
	Loans, mortgages, building loan agreements	annually	6	hlf0087_v1 , hlf0087_v2 , hlf0087_h , hlf0088_v1 , hlf0088_v2 , hlf0088_h
	Lunch, childcare	1997, 2002, 2005, 2007, 2011, 2013, 2015	1	kd_lunch
	Lunch, school	1997, 2002, 2005, 2007, 2011, 2013, 2015	1	ks_lunch
	Material deprivation	2001, 2003, 2005, 2007, 2011, 2013, 2015, 2016	32	hlf0175 , hlf0177 , hlf0178_v1 , hlf0178_v2 , hlf0178_h , hlf0179-hlf0181 , hlf0183 , hlf0185-hlf0195 , hlf0613 - hlf0622 , hlf0441 , hlf0444
	Modernization costs	annually	2	hlf0599 , hlf0560
	Children in household	annually	1	hlk0044
	Neighborhood description	annually	1	hlf0153
	Financial burden of home ownership	annually	1	hlf0606

Continued on next page

Table 5 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Costs of home ownership	annually	9	hlf0601 - hlf0605 , hlf0090_v1 , hlf0090_v2 , hlf0090_h , hlf0084
	Persons in household in need of care	annually	30	hlf0291 , hlf0631 , hlf0292 , hlf0300 - hlf0304 , hlf0315_v1 , hlf0315_v2 , hlf0315_v3 , hlf0315_h , hlf0317_v1 , hlf0317_v2 , hlf0317_v3 , hlf0317_h , hlf0319 - hlf0322 , hlf0331 , hlf0332 , hlf0369 , hlf0370_v1 , hlf0370_v2 , hlf0370_h , hlf0446 - hlf0448 , hlf0595
	Pets	1996, 2006, 2011, 2016	7	hlf0254-hlf0259, hlf0196
	Photovoltaic / solar thermal system	annually	6	hlf0532 , hlf0535 - hlf0539
	Costs of comparable rental homes	1984-2014	1	hlf0094
	Reasons for moving, comparison of old and new home	1985-2013, 2015	28	hlf0108_v1 - hlf0108_v15 , hlf0108_h , hlf0109, hlf0124-hlf0132 , hlf0524 - hlf0526
	Dependence on childcare hours	2002	1	kd_rely
	Monthly rent, heating, other expenses	annually	18	hlf0069_v1 - hlf0069_v5 , hlf0069_h , hlf0074_v1 , hlf0074_v2 , hlf0074_h , hlf0078 , hlf0079 , hlf0081_v1 , hlf0081_v2 , hlf0081_h , hlf0082 , hlf0607 , hlf0608 , hlf0610
	School attendance by child	annually	2	ks_gen , ks_spe
	Second home	2011,2016	3	hlf0156-hlf0158
	Size and condition of home	annually	10	hlf0018 , hlf0019_h , hlf0019_v1 - hlf0019_v3 , hlf0071_v1 , hlf0071_v2 , hlf0071_v3 , hlf0071_h , hlf0021_v1 , hlf0021_v2 , hlf0021_v3 , hlf0021_h
	Government-subsidized housing	annually	6	hlf0011_v1 - hlf0011_v4 , hlf0011_h , hlf0073
	Childcare provider and costs	2011, 2013, 2015	7	kd_publ, kd_indep, kd_priv, kd_comp, kd_comm, kc_amtp, kc_cost

Continued on next page

Table 5 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	School provider and costs	1987, 1995, 1997, 2002, 2005, 2007, 2011, Mig 2013	7	kd_publ, kd_indep, kd_priv, kd_comp, kd_comm, ks_amtp, ks_cost
	Financial burden of home rental	annually	1	hlf0611
	Type of energy used in household		51	hlf0540-hlf0591

2.7 Education and Qualification

Education is one of the cornerstones of society today. The education, training, and qualifications modules provide extensive information on educational attainment and outcomes, the level of completed education and training, reasons for not completing education or training, educational goals, and much more, along with data on children's skill development, for instance, whether they are able to speak in full sentences or use scissors.

Questionnaire	Module	Replication	No. Vars.	Variables
Individual Questionnaire		annually		
	Completed education and training	annually	17	plg0072- plg0077 , plg0078_h, plg0078_v1, plg0078_v2, plg0079_v1 - plg0079_v4, plg0284 , plg0268, p_degree, p_field
	Further training measures	annually	3	plg0269 - plg0271
	Vocational training	annually	13	plg0012 , plg0013_v1, plg0013_v3, plg0014_h, plg0014_v1 - plg0014_v5, , plg0015_h, plg0015_v1 - plg0015_v4
	Further training, suggested / provided by employer	1989, 1993, 2000, 2004, 2008, 2014	2	plg0273 , plg0274
	Further training, reasons for not taking part	1989, 1993, 2000, 2004, 2014	5	plg0277 – plg0281

Continued on next page

Table 6 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Further training, course details and motives for participation	1989, 1993, 2000, 2004, 2008	72	plg0108_h, plg0108_v1, plg0108_v2, plg0109_h, plg0109_v1, plg0109_v2, plg0110_h, plg0110_v1, plg0110_v2, plg0111 - plg0119 , plg0120_v1- plg0120_v4, plg0121_v1 - plg0121_v4, plg0122_v1 - plg0122_v4, plg0129 - plg0149 , plg0152 , plg0154 , plg0164 , plg0165 , plg0169_h, plg0169_v2, plg0169_v1 , plg0171_h, plg0171_v1, plg0171_v2, plg0172_h, plg0172_v1, plg0172_v2 , plg0174 - plg0176 , plg0177_h, plg0177_v1 - plg0177_v6, plg0182_h, plg0182_v1, plg0182_v2, plg0183_h, plg0183_v1, plg0183_v2, plg0184 - plg0186
	Further training, course provider	2018	1	
	Further training, financing	1989, 1993, 2000, 2004, 2008, 2014, 2015, 2017	7	plg0285 - plg0291
	Lifelong learning	2014	1	plg0266
Youth Questionnaire				
	Education and career plans	annually	15	jl0177 - jl0187 , jl0438 - jl0441 , jl0197 - jl0199 , jl0201 - jl0205 , j_isco08_jobwish , j_isco88_jobwish, j_kldb2010_jobwish, j_kldb92_jobwish
	Educational aspirations	annually	12	jl0130_h , jl0130_v1, jl0130_v2, jl0131, jl0188 - jl0196 , jl0504

Continued on next page

Table 6 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	School, attendance & homework	annually	36	jl0125_h , jl0125_v1 - jl0125_v3 , jl0126 , jl0127_h , jl0127_v1 , jl0127_v2 , jl0132_h , jl0132_v1 , jl0132_v2 , jl0132_v3 , jl0133_h , jl0133_v1 , jl0133_v2 , jl0133_v3 , jl0137_h , jl0137_v1 , jl0137_v2, jl0138 - jl0157 , jl0162 - jl0166 , jl0176_h , jl0176_v1 , jl0176_v2 , jl0434 - jl0436
Parents and Child Questionnaire (7-8-year-olds)				
	Educational aspirations	annually	6	idegrad1 - idegrad3 , probgra1 - probgra3
	School enrollment	annually	3	sclenrolm , sclenroly , sclenroln
Mother and Child Questionnaire (9-10-year-olds)				
	School and homework	annually	29	sclenrolm , sclenroly , curscol1 - curscol8 , lamark , matmark , nomark , scolcon1 - scolcon7 & hwplace_h & hwsupprt
	Educational aspirations	annually	6	idegrad1 - idegrad3 , probgra1 - probgra3

2.8 Attitudes, Values, and Personality

The attitudes, values, and personality modules provide extensive information on respondents' personality traits, political orientations, concerns, satisfaction with different aspects of life, willingness to take risks, and much more.

Questionnaire	Module	Replication	No. Vars.	Variables
Individual Questionnaire				
	Affective well-being	annually	4	plh0184- plh0187
	Anomie	1992, 1993, 1995, 1996, 1997, 2008, 2013	4	plh0188-plh0191
	Attitudes towards refugees	2016	11	plj0433 – plj0443
	Big Five personality traits	2005, 2009, 2013, 2017	16	plh0212 -plh0226, plh0255
	Bundestag election	2014	1	plh0333
	Locus of control	2005, 2010, 2015	28	plh0369_v1, plh0369_v2, plh0370_v1, plh0370_v2, plh0371_v1 - plh0371_v4, plh0372_v1 - plh0372_v4, plh0373_v1, plh0373_v2, plh0374_v1, plh0374_v2, plh0375_v1, plh0375_v2, plh0376_v1, plh0376_v2, plh0377_v1, plh0377_v2, plh0378_v1, plh0378_v2, plh0379 - plh0382
	Depressive traits	2016	4	plh0339 – plh0342
	Donation of blood	2010, 2015	4	plh0131_v1, plh0131_v2, plh0132 , plh0133
	Donations	2010, 2015	2	plh0129 , plh0130
	Donations of goods	2010	8	plj0108 - plj0115
	Flourishing	annually	1	plh0334
	Goals in life (Kluckhohn)	1990, 1992, 1995, 2004, 2008, 2012, 2016	9	plh0105 – plh0112
	Impulsivity, patience	2008, 2013	3	plh0253 , plh0254
	Income justice, general	2005	12	plh0116- plh0127
	Life satisfaction	annually	1	plh0182
	Loneliness	2013, 2017	3	plh0269 - plh0271
	Lottery question	2004, 2009	1	plh0203
	Money and account balance	2016	3	plh0344 – plh0346
	Narcissism	2018		
	Optimism / pessimism	1999, 2005, 2009, 2014	1	plh0244
	Organizational and community membership	1985, 1989, 1993, 1998, 2001, 2003, 2007, 2011, 2015	7	plh0263_h , plh0263_v1, plh0263_v2, plh0264_v1, plh0264_v2, plh0264_h - plh0267
	Policy objectives (Inglehart Index)	1984, 1985, 1986, 1996, 2006, 2016	4	plh0054, plh0056, plh0058, plh0061
	Political orientation	annually	4	plh0007 , plh0011 , plh0012_v1 - plh0012_v6 , plh0012_h , plh0013
	Political leaning, left-right	2005, 2009, 2014	1	plh0004
	Reciprocity	2005, 2010, 2015	6	plh0206i01, plh0206i02, plh0206i03, plh0206i04, plh0206i05, plh0206i06

Continued on next page

Table 7 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Religious affiliation	1990, 1997, 2003, 2007, 2011, 2015-2017	12	plh0258_h, plh0258_v1 - plh0258_v11
	Risk aversion in different domains	2004, 2009, 2014	6	plh0197 - plh0202
	Risk aversion in general	annually	1	plh0204
	Satisfaction with various aspects of life	annually	11	plh0171 - plh0181
	Self-esteem	2010, 2015, 2016	1	plh0206i11
	Social responsibility	1987, 1992, 1997, 2002, 2017	11	plh0016 – plh0026
	Tendency to forgive	2010, 2015, 2016	4	plh0206i07, plh0206i08, plh0206i09, plh0206i10
	Trust, trustworthiness and fairness	2003, 2008, 2013	8	plh0192 - plh0196 , pld0043 - pld0045
	Wage justice	2005, 2007, 2009, 2011, 2013, 2015, 2017	6	plh0138- plh0141 , plh0337_v1, plh0337_v2, plh0338_v1, plh0338_v2
	Well-being aspects	1990 (only East), 1994, 1999	13	plh0091 , plh0099 , plh0100_v1, plh0100_v2, plh0101 - plh0103
	Worries	annually	13	plh0032 , plh0033 , plh0035 - plh0038 , plh0040 , plh0042 , plh0043 , plh0046 , plh0047 , plh0335 , plh0336
	10,000-euro question	2010, 2017	3	plh0134- plh0136
Youth Questionnaire				
	Affective well-being	annually	4	jl0381 - jl0384
	Attitudes and opinions	annually	4	jl0329, jl0330, jl0360 , jl0364
	Big Five personality traits	annually	17	jl0365 - jl0380
	Future	annually	11	jl0222 - jl0232
	Life satisfaction	annually	1	jl0392
	Locus of control	annually	10	jl0350 - jl0359
	Political orientation	annually	4	jl0388 - jl0391
	Risk aversion in general	annually	1	jl0349
	Sources of social inequality	annually	12	jl0337 - jl0348
	Trust	annually	3	jl0361 - jl0363
Mother and Child Questionnaire (Newborns)				
	Temperament	annually	5	temp1 - temp5
Mother and Child Questionnaire (2-3-year-olds)				

Continued on next page

Table 7 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Big Five personality traits	annually	4	char1a , char2 - char4
	Temperament	annually	7	temp1 - temp7
	Vineland adaptive behavior scales	annually	20	spch3 , spch5 - spch8 , skll1 - skll5 , mvmn1 , mvmn3 - mvmn6 , sclr2 - sclr6
Mother and Child Questionnaire (5-6-year-olds)				
	Big Five personality traits	annually	10	char1b , char2 - char10
	Strengths and difficulties questionnaire	annually	17	behav1 - behav17
Mother and Child Questionnaire (9-10-year-olds)				
	Big Five personality traits	annually	10	char1b , char2 - char10
	Strengths and difficulties questionnaire	annually	18	behav1 - behav18

2.9 Time Use and Environmental Behavior

The modules on time use and environmental behavior give information on time commitments, free time, and time planning as well as environmental awareness, for instance, the use of public transport and different energy sources, as well as what respondents think about renewable energies.

Questionnaire	Module	Replication	No. Vars.	Variables
Household Questionnaire				
	Traffic and energy	1998, 2003, 2015	73	hli0005, hli0077-hli0142
Individual Questionnaire				

Continued on next page

Table 8 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Computer usage	1997, 1999, 2000, 2001	15	pli0066_h , pli0066_v1 , pli0066_v2 , pli0067_h , pli0067_v1 , pli0067_v2 , pli0068 , pli0069 . pli0070_v1 , pli0070_v2 , pli0071_h , pli0071_v1 , pli0071_v2 , pli0072 , pli0073
	Leisure activities (long)	1990, 1995, 1998, 2003, 2008, 2013	28	pli0079 – pli0084 , pli0085_v1 , pli0085_v2 , pli0086 - pli0089 , pli0090_v1 , pli0090_v2 , pli0091 , pli0092_v1 , pli0092_v2 , pli0093 , pli0094_v1 , pli0094_v2 , pli0095_v1 , pli0095_v2 , pli0096_v1 , pli0096_v2 , pli0097_v1 , pli0097_v2 , pli0098 , pli0165 & pli0168
	Leisure activities (short)	1984-1986, 1988, 1992, 1994, 1996, 1997, 1999, 2001, 2005, 2007, 2009, 2011, 2015, 2017	15	pli0090_v1 , pli0090_v2 , pli0091 , pli0092_v1 , pli0092_v2 . pli0093 , pli0094_v1 , pli0094_v2 , pli0095_v1 , pli0096_v1 , pli0096_v2 , pli0097_v1 , pli0097_v2 , pli0098
	Use of transportation in general	1998,2003	38	pli0101 - pli0104 , pli0105_h , pli0105_v1 , pli0105_v2 , pli0106 - pli0108 , pli0109_h , pli0109_v1 , pli0109_v2 , pli0110 , pli0111 , pli0112_h , pli0112_v1 , pli0112_v2 , pli0113 , pli0114 , pli0115_h , pli0115_v1 , pli0115_v2 , pli0116 , pli0117 , pli0118_h , pli0118_v1 , pli0118_v2 , pli0119 , pli0120 , pli0121_h , pli0121_v1 , pli0121_v2 , pli0122 - pli0126
	Use of transportation to work	1998,2003	11	pli0127_v1 , pli0127_v2 , pli0128_v1 , pli0128_v2 , pli0129_v1 , pli0129_v2 , pli0130_v1 , pli0130_v2 , pli0131_v1 , pli0131_v2 , pli0132

Continued on next page

Table 8 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Use of transportation for errands	1998,2003	11	pli0133_v1 , pli0133_v2 , pli0134_v1 , pli0134_v2 , pli0135_v1 , pli0135_v2 , pli0136_v1 , pli0136_v2 , pli0137_v1 , pli0137_v2 , pli0138
	Use of transportation for excursions	1998,2003	11	pli0139_v1 , pli0139_v2 , pli0140_v1 , pli0140_v2 , pli0141_v1 , pli0141_v2 , pli0142_v1 , pli0142_v2 , pli0143_v1 , pli0143_v2 , pli0144
	Use of transportation in leisure time	1998,2003	12	pli0145_v1 , pli0145_v2 , pli0146_v1 , pli0146_v2 , pli0147_v1 , pli0147_v2 , pli0148_v1 , pli0148_v2 , pli0149_v1 , pli0149_v2 , pli0150_v1 , pli0150_v2
	Use of transportation to take children places	1998,2003	11	pli0151_v1 , pli0151_v2 , pli0152_v1 , pli0152_v2 , pli0153_v1 , pli0153_v2 , pli0154_v1 , pli0154_v2 , pli0155_v1 , pli0155_v2 , pli0156
	Trip to work	annually	26	plb0142 - plb0145 , plb0146_h , plb0146_v1 , plb0146_v2 , plb0147 - plb0155 , plb0156_v1 , plb0156_v2 , plb0157_v1 , plb0157_v2 , plb0158 , plb0159_h , plb0159_v1 , plb0159_v2 , plb0175_v1 , plb0175_v2
	Time use for different activities (weekdays)	annually	25	pli0001_v1 , pli0001_v2 , pli0038_h , pli0038_v1 - pli0038_v4 , pli0040 , pli0043_h , pli0043_v1 - pli0043_v3 , pli0044_h , pli0044_v1 - pli0044_v3 , pli0046 , pli0047 , pli0049_h , pli0049_v1 - pli0049_v3 , pli0051 , pli0052 , pli0059

Continued on next page

Table 8 – continued from previous page

Questionnaire	Module	Replication	No. Vars.	Variables
	Time use for different activities (Saturdays)	Every 2 years	29	pli0003_h , pli0003_v1 - pli0003_v3 , pli0003_v4 , pli0005 , pli0012_h , pli0012_v1 - pli0012_v3 , pli0019_h , pli0019_v1 - pli0019_v4 , pli0024_h , pli0024_v1 - pli0024_v3 , pli0031_h , pli0031_v1 - pli0031_v4 , pli0036 , pli0054 , pli0055 , pli0056 , pli0060
	Time use for different activities (Sundays)	Every 2 years	31	pli0002_v1 , pli0002_v2 , pli0007_h , pli0007_v1 - pli0007_v4 , pli0007_v5 , pli0010 , pli0011 , pli0016_h , pli0016_v1 - pli0016_v4 , pli0022_h , pli0022_v1 - pli0022_v4 , pli0028_h , pli0028_v1 - pli0028_v4 , pli0034_v1 - pli0034_v4 , pli0057 , pli0058
Youth Questionnaire				
	Leisure and hobbies	annually	27	jl0058 - jl0073 , jl0074 - jl0076 , jl0087 , jl0104 , jl0106 , jl0109 , jl0112 , jl0123 , jl0117 , jl1400

2.10 Integration, Migration, Transnationalization

Migration and forced migration are changing German society. The SOEP offers diverse migration samples and numerous specific migration questions that allow researchers to study migration-related questions in detail. The modules on integration, migration, and transnationalization provide data on migration histories, discrimination, inter-ethnic contact, education, cultural integration, transnational relations, identification with Germany, and the intention to stay in Germany.

Questionnaire	Module	Replication	No. Vars.	Variables
Individual Questionnaire				
	Circle of friends, percentage of migrants	2013,2018	1	plm0143
	Applying for German citizenship	1998-2018 (every two years)	1	plj0021
	Visiting / being visited by Germans and foreigners at home	1997-2017 (every two years)	4	plj0060 - plj0063
	Contacts abroad, thoughts about moving abroad	2009,2014	6	plj0104 , plj0105 , plj0089 - plj0092
	Disadvantage / discrimination based on ethnic origins (detailed)	2015	17	plj0048_v1 , plj0048_v2 , plj0327 - plj0339
	Disadvantage / discrimination based on ethnic origins (short)	1997-2011, 2013, 2017	1	plj0048_v1 , plj0048_v2
	Foreign language skills	2013	1	plm0135
	Integration indicators	1997, 1999, 2001, 2003, 2010, 2012, 2014, 2016, 2018	2	plj0078 , plj0080
	Intention to stay	1997-2011, 2013, 2015, 2017	4	plj0085 - plj0088
	Language use, media	2014, 2016, 2018	2	plj0226_v1 , plj0226_v2
	Language use, newspapers	1996-2012 (every two years)	1	plj0070
	Native language	2007-2011, 2013, 2015, 2017	8	plj0009 , plj0324 - plj0326
	Regional attachment	2009,2014	3	plj0043 - plj0045
	Residence status, citizenship	2018	2	
	Sense of home	1996-2014 (every two years)	2	plj0083 , plj0340
	Visited country of origin in last 2 years	1996-2018 (every two years)	2	plj0322 , plj0323
	Language ability German / native language		9	plj0071 - plj0076
Youth Questionnaire				
	Language ability German / native language		9	jl0248 , jl0251 , jl0442 - jl0444 jl1249 , jl1251

2.11 Survey Methodology

Survey methodology modules offer diverse variables on imputation, weighting, SOEP-Core fieldwork, identifiers, interview methods, survey modes, and information about the respondent's exit from the survey.

Questionnaire	Module	Replication	No. Vars	Variables
Interviewer Questionnaire				
	Identificators	annually	4	hhnr , intid , syear , wave
	Interview information	annually (youth since 2000)	4	typint , lenghtinth , lenghtintp , lenghtintj
	Demography		11	gender , birth , marital , educ , modbula , modggk , ista1 , ibstam1 , ibstav1 , imusp , irel
	Interviewer history	annually	7	startint , endint , experience , firstintm , firstintd , lastintm , lastintd
	Employment		3	iberuf , ioed , istell
	Interviewer activity		9	meancontacthh , responserate , amountinth , amountintp , amountintj , papi , capi , cawi , mail
	Patience	2006,2012	1	iged
	Health	2006, 2012, 2016	1	iges
	Risk aversion	2006,2012	1	irisk
	Life satisfaction	2006, 2012, 2016	1	izule
	Incentives	2012	2	ibbarhon , ibbeval
	Optimism	2012	1	ibopt
	Motivation & Fulfillment		21	igru01 - igru07 , ierf01 - ierf14
	Assessment of Participation		13	itebe01 - itebe13
	Interviewer Training		6	ibseval01 - ibseval04 , ibschul , ibschul02
	Big Five personality traits		22	iego01 - iego22
	Attitudes and social interaction	2006,2012	11	ibez01 - ibez05 , iverh01 - iverh06
	Plotical orientation	2006, 2012, 2016	4	ipol1 - ipol4
	Worries		16	isor01 - isor14 , isor21 - isor22
	Working hours	2012	8	ibwsist01 - ibwsist05 , ibwssol01 - ibwssol03

Continued on next page

Table 10 – continued from previous page

Questionnaire	Module	Replication	No. Vars	Variables
	Interviewer and other studies		13	ibsozer01 - ibsozer08 , ibsozerno , ibsozerso , ibef01 - ibef03
	Foreign language skills		20	ispr01 - ispr10 , ibspre01 - ibspre10
	Flags (conflicts)		6	genderconfl , birthconfl , maritalconfl , educconfl , startintconfl , ista1confl

Important documents regarding this Topic are available here

Last change: Nov 12, 2019

CHAPTER

THREE

SURVEY DESIGN

3.1 SOEP Questionnaires

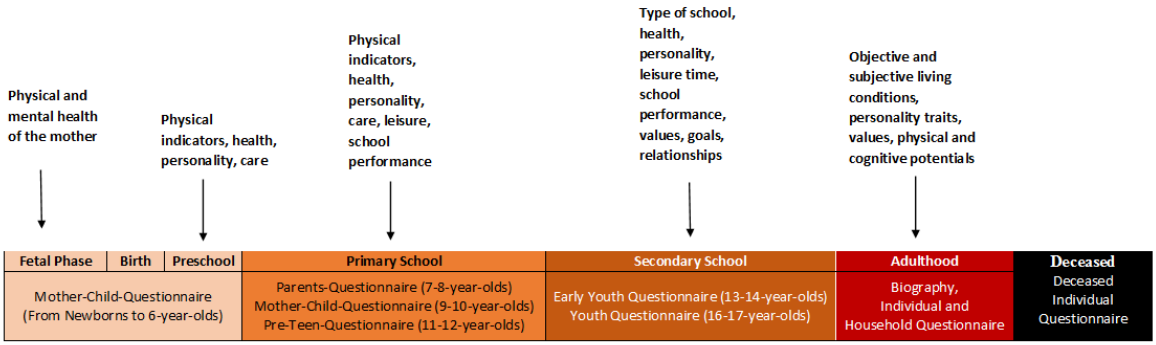
The interview methodology of the SOEP is based on a set of pre-tested questionnaires for households and individuals. Interviewers try to obtain face-to-face interviews with all members of a given survey household aged 16 and over. Thus, there are no proxy interviews for adult household members. Additionally,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

The questions in the SOEP are largely identical for all participants of the survey to ensure comparability across the participants within a given year, but of course there are differences across years. There are a few exceptions to this rule, which are due to different requirements in the target population. Up to 1996, the questionnaires for the sample of foreigners (B) and the immigrant sample (D) covered additional measures of integration or information on re-migration behavior. Between 1990 and 1992, i.e., during the first years of the German reunification process, the questionnaire for the East German sample (C) also contained some additional specific variables. From 1996 to 2012, all questionnaires were uniform and completely integrated for all of the main SOEP samples. For the IAB-SOEP Migration Sample, which was launched in 2013, specific questions were added to the SOEP questionnaires. The same is true of the IAB-BAMF-SOEP Survey of Refugees, which was launched in 2016.

Another special questionnaire is used for first-time respondents since some questions do not have to be repeated every year. Each respondent is asked to fill out a biographical questionnaire covering information on the life course up to the first SOEP interview (e.g., marital history, social background, and employment biography).

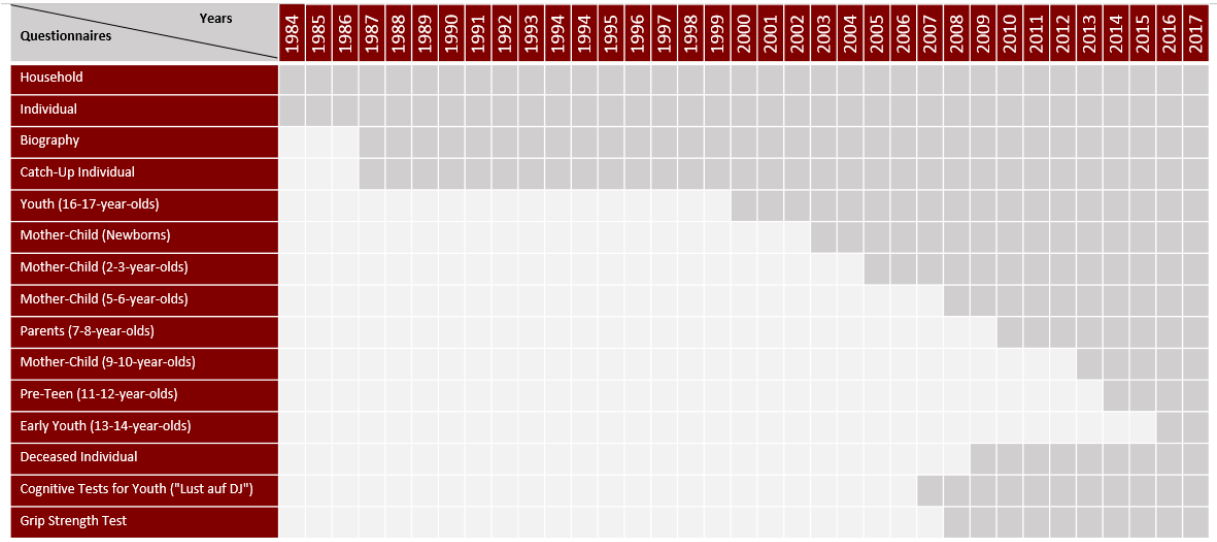
Additional information not provided directly by the respondent can be obtained from the "address logs", which are stored for every year in the $PBRUTTO and $HBRUTTO files. Every address log is filled in by the interviewer even in the case of non-response, thus providing very valuable information, e.g. for attrition analysis. For researchers interested in methodological issues, these data also contain information on the fieldwork process such as the number of contacts, reasons for drop-outs, and interview mode. For households that were contacted successfully, the address logs cover the size of the household, some regional information, survey status, etc. The individual data for all household members include the relationship to the household head, survey status of the individual, and some demographic information.

Life History



The SOEP questionnaires are designed so that people in a SOEP household can be analyzed from birth to adulthood and throughout the rest of their lives. In addition to the *Youth Questionnaire*, which was conducted for the first time in 2000/01, a series of questionnaires for specific cohorts of children living in SOEP households have been introduced since 2003. These have been completed annually since their year of introduction by mothers (in exceptional cases by fathers) with children of the appropriate age. In 2003, a questionnaire was developed for the mothers of newborn children, *Mother and Child Questionnaire (Newborns)*. The following instruments were developed in such a way that this starting cohort (born 2002/2003) can be followed up in their development and analyzed longitudinally. This was followed in 2005 by a questionnaire for mothers of 2-3-year-old children, *Mother and Child Questionnaire (2-3-year-olds)* and in 2008 by a questionnaire for 5-6-year-olds, *Mother and Child Questionnaire (5-6-year-olds)*. In 2010, the questionnaire for 7-8-year-old children, *Parents and Child Questionnaire (7-8-year-olds)*, completed by both mothers and fathers, was launched. In 2012, the questionnaire for 9-10-year-old children, *Mother and Child Questionnaire (9-10-year-olds)* was added as the last questionnaire to be answered by the mothers. This was followed by two youth instruments in which the children, aged 12, *Pre-Teen Questionnaire* and 14, *Early Youth Questionnaire*, answered questions about their own lives for the first time. These were introduced in 2014 and 2016, respectively. In 2018, the first cohort completed the entire battery of age-specific instruments and from then on, they will complete the annual questionnaires of the long-term SOEP study. Each person in a SOEP household receives the *Individual Questionnaire* as soon as they reach the age of 18, and the head of the household also receives the *Household Questionnaire*. If a respondent states in their interview that someone has died in the last year, regardless of whether the deceased person was part of a SOEP household, the *Deceased Individual Questionnaire* is given to the respondent providing the information.

Overview of the Questionnaires



Household Questionnaire

The household questionnaire in its basic form has been an important part of the SOEP surveys since 1984 and has been improved and expanded continuously. The data collected and the questionnaire itself have become so complex that the original topics are no longer sufficient. Between 1984 and 2016, the number of questions more than doubled from 46 to 97. The multitude of questions offer users many options for analysis. Each year, the number of questions varies because new innovative question modules are added or because some questions are not asked every year. An overview of the modules included at different intervals can be found in the section *Topics of SOEP-Core*. The questions provide diverse information about the respondents' households that is stored in several hundred variables. Child-specific questions asked in the household questionnaire are found in the separate dataset $kind.

Availability: Since 1984

Dataset: $h (CS), *hl* (long)

Respondent: Head of household

The following question modules are part of the core program of the Household Questionnaire:

- Change of living situation
- Neighborhood
- Building type
- Size and condition of dwelling
- Amenities
- Type of dwelling
- Loans, mortgages, building-society loans
- Hereditary lease interest
- Modernization costs
- Ownership costs

- Photovoltaic and solar thermal system
- Owner debt
- Government-subsidized housing
- Home ownership
- Rental and expenses
- Tenant debt
- Cleaning or household assistance
- Persons in need of care
- Names and birth dates of children
- Child's school attendance
- Childcare situation
- Income and expenses from renting/leasing
- Loan repayment
- Debt
- Inheritances, gifts, winnings
- Investments
- Income/expenses household
- Savings

where applicable:

+ migration-specific modules for the IAB-SOEP Migration Sample

- distinguishing repayment of loans, debt, income / expenses between Germany and foreign country

or where applicable:

+ refugee-specific modules for the IAB-BAMF-SOEP Sample of Refugees

- Information on shared accomodations
- Location preferences

Individual Questionnaire

The individual questionnaire has been a standard instrument since the beginning of the SOEP. In order to enable analysis over time, the individual questionnaire has a large number of question modules that are asked every year. There are also questions that do not have to be asked every year, as short-term changes are unlikely. In order to be able to react to current social changes, new topics are added to the individual questionnaire and repeated at intervals of more than one year.

Availability: Since 1984

Dataset: $p (CS), *pl* (long)

Respondent: Persons over 18 years in the household

The following question modules are part of the core program of the Individual Questionnaire:

- Satisfaction with various live aspects

- Satisfaction with current life situation
- Feelings
- Flourishing
- Risk aversion
- Political orientation
- Worrying
- Life satisfaction overall
- Ethnic/national origins
- Vocational training
- Completed level of education
- Higher education
- Family situation
- Family changes
- State of health
- Disability or severe disability
- Visits to the doctor
- Hospital stays
- Sick leave
- Health insurance
- Wages and collective wage agreements
- Additional questions for employees
- Additional questions for retirees/pensioners
- Government transfers
- Calendar
- Time use
- Second jobs
- Income
- Work, last 7 days
- Maternity/ parental leave
- Care period (Pflegezeit)
- Registered unemployed
- Quitting a job
- Employment status
- Start of job
- Change of job
- Job search

- Current profession
- Current job
- Working hours
- Overtime

where applicable:

+ migration specific modules for the IAB-SOEP-Migrationsample

- First Job in Germany
- Job before immigration
- Language proficiency before and since immigration
- Partnership during immigration
- Living situation since immigration
- Religion and faith of parents
- Satisfaction in various areas of life before and after immigration

or where applicable:

+ refugee specific modules for the IAB-BAMF-SOEP-Sample of Refugees

- Legal status
- Religion and faith
- Language proficiency
- Integration courses and government measures
- Special questions for interviewers concerning language
- Recognition of qualifications

Re-Interviewed

- Cultural and political participation
- Application for recognition
- Trauma screener
- OK (Judgement of different actions)
- Citizenship (inkl. connection with country of origin/ Germany)
- Disadvantages
- Location preferences
- Willingness to participate in a tandem program
- Satisfaction in various areas before and after fleeing

New respondents

- Obtaining help and knowledge about advice services
- Assessment of current situation in country of origin
- Government, democracy and woman's position

Biography Questionnaire

Availability: Since 1987

Dataset: $lela (CS), *biol* (long)

Respondent: Supplementary, one-time data from the personal questionnaire of all persons aged 18 and over in the household.

Content:

- Nationality
- Country of Origin
- Childhood
- Parents
- Life course since the age of 15
- Education
- Occupation
- Partnership/marriage
- Information on children
- Siblings

where applicable:

+ migration specific modules for the IAB-SOEP-Migrationsample

- Travel to Germany
- Stays Abroad
- Citizenship
- Language proficiency
- Work before moving to Germany
- First job in Germany
- Relationship at the time of moving to Germany

or where applicable:

+ refugee-specific modules for the IAB-BAMF-SOEP Sample of Refugees

- Travel to Germany
- Questions concerning parents of respondent
- Lodging and living situation
- Language proficiency before moving to Germany

Mother and Child Instruments

Mother and Child Questionnaire (Newborns)

Mothers of newborn children answer questions dealing primarily with pregnancy, birth, breastfeeding, and the health of the newborn child. The questionnaire also asks to what extent the mother feels that her living situation changed

after the birth of the child, how childcare is handled, and how mothers assess their baby's temperament (as a precursor to personality).

Availability: Since 2003

Dataset: $muki (CS), *bioagel* (long)

Respondent: Mother in household (child age 0-1)

Content:

- Course of pregnancy
- Childbirth
- Health screening
- Well-being
- Childcare
- Living situation

Mother and Child Questionnaire (2-3-year-olds)

Mothers of 2-3-year-old children answer questions about their child's health and how long they have been breast-feeding. The questionnaire asks again about the childcare situation and the child's temperament and includes a short scale on personality (the dimensions of agreeableness, extraversion, openness, and conscientiousness from the "Big Five"; McCrae and Costa 1987). In addition, it asks what language is spoken with the child and what activities they or the main caregiver engages in with their child (e.g., going to the playground, reading or telling stories, visiting other families with children). Mothers are asked to assess their children's adaptive behavior in the areas of communication, everyday skills, social relationships, and motor skills. This is based on a translated version of the Vineland Adpative Behavior Scale, which was reduced to 20 items for the SOEP to provide data on the child's stage of development in everyday life.

Availability: Since 2005

Dataset: $muki2 (CS), *bioagel* (long)

Respondent: Mother in household (child age 2-3)

Content:

- Personality of the child
- Well-being
- Childcare
- Language skills
- Development
- Abilities

Mother and Child Questionnaire (5-6-year-olds)

Mothers of 5-6-year-old children complete this questionnaire in the survey year when their child will turn six. It has many of the same topics as in previous years: health, childcare, a more comprehensive battery of items on the personality (from this age on, the "Big Five" dimension of neuroticism is also included) and activities that they or the main caregiver engages in with their child. In addition, the questionnaire includes a shortened version of the

Strength and Difficulties Questionnaire (SDQ), a frequently used instrument to measure the mental health of children and adolescents, reduced here to 17 items of the German SDQ.

Availability: Since 2008

Dataset: $muki3 (CS), *bioagel* (long)

Respondent: Mother in household (child age 5-6)

Content:

- Personality of the child
- Activities with children
- Well-being
- Childcare

Parents and Child Questionnaire (7-8-year-olds)

This questionnaire on 7-8-year-old children is the only age-specific instrument that is completed by both parents, as long as they live together in the same household. In this age range, questions about school attendance (date of school enrolment) and parent's aspirations for their children's level of school completion become relevant for the first time. However, the focus is on parenting goals, parenting styles, and the roles of both parents. Parenting goals range between conformity and autonomy. Parenting styles are surveyed using 18 items, which can be divided into six scales: emotional warmth, inconsistent education, monitoring, negative communication, psychological control, strict control. The items were taken from the pairfam study, as were the 10 items on the role of parents, which can be divided into three scales: autonomy, hostile attributes, and willingness to make sacrifices.

Availability: Since 2012

Dataset: $elt (CS), *bioagel* (long)

Respondent: Parents in household (child age 7-8)

Content:

- Hopes and expectations for children's educational attainment
- Parental goals
- Parental styles
- Parental role
- Childcare

Mother and Child Questionnaire (9-10-year-olds)

In addition to questions on health and child care, which are asked in almost all age groups, mothers of 9-10-year-old children are asked for more detailed information about the children's school situation. They are asked what level of schooling they would like their children to complete and what level they think is realistic, what their children's most recent grades were in their three main subjects, whether someone helps the child with homework, and whether the child likes going to school. Since friends and leisure activities are gaining in importance in this age group, some questions deal with these topics. Questions about allowance money are asked for the first time in this age group.

Availability: Since 2012

Dataset: $muki5 (CS), *bioagel* (long)

Respondent: Mother in household (child age 9-10)

Content:

- Hopes and expectations for children's educational attainment
- Education
- Parental involvement
- Leisure activities
- Family environment
- Social behavior of child
- Personality of Child
- Health of Child
- Supervision
- Allowance money

Youth Instruments

Pre-Teen Questionnaire

Young people complete a questionnaire for the first time themselves in the year they turn twelve. Here, as in the preceding questionnaires, the focus is on their school situation: what time their school day starts and ends on different days of the week, what type of school they attend, how many students are in their class, how many of their classmates are not from Germany, whether they feel discriminated against by their teacher, and what their grades were on their last report card in Math, German, and English. The questionnaire also asks how much time they spend on homework, where they do their homework, and who helps them with homework and studying. They are asked what level of schooling they would like to complete and what level they realistically expect to complete. Since friends play an important role at this age, pre-teens are asked how often they go to friends for support when they have problems. They are asked how many close friendships they have and how often their parents interfere in their choice of friends. They are asked about the educational aspirations of their three closest friends and three older siblings (if any). Several questions deal with their cultural capital and learning environment (e.g., books, musical instruments, and art in the household; whether they have a desk and a room of their own). They are asked about how they spend their free time, how much allowance money they get, and about their personality, willingness to take risks, and life satisfaction. Further questions deal with what languages are spoken with the child and who the child eats meals with.

Availability: Since 2014

Dataset: $school (CS), *biopupil* (long)

Respondent: 11-12-year-olds in the household

Content:

- Attitude
- Personality
- School (schedule, educational attainment, extra-curricular activities)
- Recreational activities
- Social and family surroundings
- Living situation

Early Youth Questionnaire

The questionnaire for early youth is designed similarly to the pre-teen questionnaire to provide important data on developmental psychology. There are fewer questions about homework and the learning environment and more questions on involvement in extra-curricular activities at school (e.g., student council, after-school clubs) since such activities build social capital. Early youth are asked about the importance of various family members and friends in their lives and about their own educational aspirations as well as those of their three best friends. They are asked how late they are allowed to stay out on school nights and weekends, and what types of activities they have taken part in without their parents (e.g., vacation, doctor visits, shopping, drinking alcohol, smoking cigarettes). They are asked how much allowance they get, and whether they have any savings. Another new topic in this age group is interest in politics and political orientations.

Availability: Since 2015

Dataset: $school2 (CS), *biopupil* (long)

Respondent: 13-14-year-olds in the household

Content:

- Self-perception
- School (schedule, educational attainment, extra-curricular activities)
- Recreational activities
- Friends
- Siblings
- Parents
- Allowance money
- Political party preferences
- Self-perceptions
- Willingness to take risks
- Life satisfaction
- Attitudes/opinions
- Future

Youth Questionnaire

In the SOEP, young people who turn 17 in the year of the survey are considered adult respondents. Like other first-time adult respondents, they receive a biography questionnaire and an individual questionnaire. Since part of the adult biography (e.g., employment history, relationships) does not yet apply to the young respondents, whereas other aspects such as relationships with parents, leisure activities, and school or vocational training play a greater role, a youth questionnaire was developed in 2000 to replace the biographical questionnaire in this age group. The content of this questionnaire corresponds in many respects to the adult biographical questionnaire so that the data can be used to supplement the information on parents (if parents do not live in the household; dataset: BIOPAREN). Health status, personality, willingness to take risks, locus of control, trust, time preferences, political preferences, knowledge of German, as well as information on the respondent's living situation, work situation, training, career plans, and educational aspirations are also covered in this questionnaire. For the period from 2000 to 2005, respondents in this age group completed the youth questionnaire and the individual questionnaire. Since 2006, they have only completed the youth questionnaire. The version used since then has been expanded to include a few additional indicators. A test was added to assess cognitive potential based on the I-S-T 2000R (Amthauer et al. 2001) using 20 subtasks each for

the components of analogies, number series, and matrices (see Solga et al. 2005). The test measures fluid cognitive abilities, a strongly biologically determined dimension of cognitive abilities that is not influenced by education and is primarily based on reasoning, processing rate, and working memory capacity (Cattell 1971; Horn 1982). Although the format of the test differs from those usually used in surveys, young people's willingness to participate has been high (Schupp and Hermann 2009).

Availability: Since 2000

Dataset: $jugend (CS), *jugendl* (long)

Respondent: 16-17-year-olds in the household

Content:

- Living
- Relationships
- Leisure and sports
- School (educational attainment, foreign languages, extra-curricular activities)
- Allowance money
- Education
- Career plans
- Future
- Background
- Childhood and Upbringing
- Attitudes/opinions
- Self-Perception
- Life satisfaction
- Political party preferences

Cognitive Tests for Youth

In 2006, a separate questionnaire with cognitive tests for adolescents was used for the first time in the SOEP. It was named "Lust auf DJ" (or "interest in DJ") as a play on disc jockey, but DJ stands for "Denksport und Jugend", or mind sports and youth. The questionnaire was created for young people between the ages of 16 and 17.

Availability: Since 2007

Dataset: *cogdj* (CS)

Respondent: 16-17-year-olds in the household as a supplement to the youth questionnaire

Content:

- Assignment of word pairs
- Complete equations
- Assign figures

Additional Instruments

Catch-Up Individual Questionnaire

The Catch-Up or "Gap" (German:Lücke) questionnaire is given to respondents who failed to respond in the previous year of the study. They are asked to provide important data about the year they missed.

Availability: Since 1987

Dataset: pluecke (CS), *plueckel* (long)

Respondent: SOEP respondents who are temporarily unavailable.

Content:

All data refer to the previous survey year

- Status of the respondent
- Occupational change
- Receipt of social benefits within the last year
- Completion of education
- Type of educational attainment
- Change of family status

Deceased Individual Questionnaire

In 2009, for the first time in SOEP-Core, information was collected on former SOEP participants who had died since the last survey in 2008. The Deceased Individual questionnaire thus completes the life history information in the SOEP. The primary aim is to obtain as much information as possible about the causes and circumstances of death of former SOEP respondents. As the questionnaire also collects information on individuals who have never participated in the SOEP survey, this can be used together with the causes and circumstances of death in socio-scientific analysis.

Availability: Since 2009

Dataset: vp (CS), *vpl* (long)

Respondent: SOEP respondents who lost a loved one.

Content:

- Relationship to the deceased
- Was the deceased a survey respondent?
- Domestic environment of the deceased
- Cause and place of death
- Last will and testament
- Health of the deceased
- Life satisfaction of the deceased
- Influence of bereavement on respondent's own life

Grip Strength Test

Availability: Since 2008

Dataset: *gripstr* (long)

Respondent: Persons over 17 years in the household

Content:

This test measures hand grip strength, which is useful in assessing respondents' physical condition.

Interviewer Questionnaire

We derive basic demographical and employment information on interviewers from personnel data of the fieldwork organization. Since 2000, Kantar Public regularly updates these information. Additionally, at irregular intervals, the SOEP interviewers complete a short version of the standard individual questionnaire themselves, which is called the interviewer questionnaire.

Availability: 2006, 2012, 2016

Dataset: *interviewer* (long)

Respondent: SOEP interviewers

Content:

- Basic Demography
- Occupational History
- Personality
- Motivation
- Interviewer Training
- Worries
- Language Skills

Last change: Nov 13, 2019

3.2 Survey Concepts and Modes

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

- Questions about a point in time (present), e.g., current employment status or current levels of satisfaction
- Retrospective questions about certain events in the past, e.g., how often have you changed jobs in the last ten years?
- Retrospective life event history since the age of 15 (in the past), e.g., employment or marital history
- Monthly calendar information on income and labor market participation (in the past), e.g., employment status January through December of last year
- Questions about a period of time (in the past), e.g., demographic changes since the last interview such as marriage or death of spouse

- Questions about the future, e.g., expected satisfaction with life five years from now, or job expectations

Survey Modes

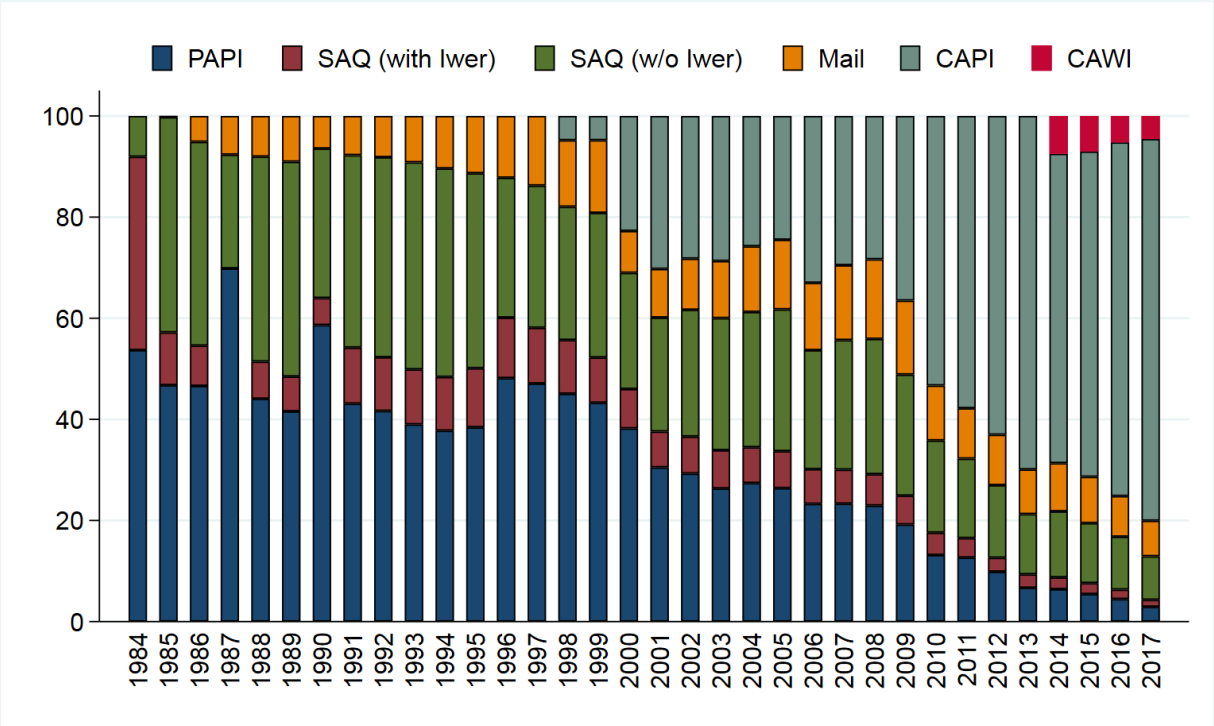
The SOEP uses several different modes to collect the data. Originally, the respondent's answers were always recorded by an interviewer who filled in the answers in a paper questionnaire, the "pen-and-paper interview" or PAPI. The personal contact between interviewer and respondent is important for the success of the survey; however, before losing a respondent due to a scheduling conflict between interviewer and respondent, the SOEP has allowed respondents to mail in the questionnaire since the second wave of subsamples A-I. This is not the same as the concept of a regular mail survey, because the interviewer still maintains personal contact with the household and schedules appointments with respondents if possible. Starting with subsample J, only "computer-assisted personal interviews" (CAPI) are allowed, and thus it is no longer possible to mail in the questionnaires.

When visiting a household, the interviewer interviews household members one at a time and can also give questionnaires to other household members to complete without the interviewer's assistance (self-administered questionnaires, SAQ). This is a time-efficient approach because it allows different household members to complete their questionnaires at the same time.

In 1998, computers were used for the first time in the SOEP for computer-assisted personal interviews (CAPI). Compared to PAPI, the CAPI mode is much more efficient in converting the data into an electronic format, which was an important asset especially with the extensions to the panel starting in the year 2000. The CAPI mode was first used parallel with PAPI, meaning that interviewers and respondents were free to chose how they wanted to do the interview. This was important for the "older" sample members (respondents as well as interviewers), who were used to the PAPI concept. Only in the most recent samples (starting in subsample J) is CAPI the sole interview mode. The figure depicts the development of modes up to 2011, showing that the CAPI mode has gained importance since its implementation.

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

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



Download STATA Code to create figure.

Last change: Nov 12, 2019

3.3 Panel Care

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

Respondents have been given gifts as tokens of appreciation since the very beginning of the study. Most of these gifts are small in-kind incentives like flowers, for which the interviewers have their own budget. In addition, the interviewers are asked to hand out a brochure with recent results from the study. Up to 2007, respondents also received a lottery ticket as a thank-you upon completion of their interview. Proceeds from the lottery benefit social projects in Germany. Since 2008, the lottery ticket has been included with the contact letter that is sent out about two weeks prior to the interview. It is thus given unconditionally, as long as the person participated in the previous wave. After a successful interview, the respondent receives a thank-you letter from survey institute along with one postage stamp as a small additional gift.

In 2009, different incentive schemes were tested in the new subsample I to increase the first-wave response rates. The basic experiment included four randomized groups of households: (1) those with the default setup of the conditional lottery ticket; (2) those with a "low" cash incentive of 5 euros per household and 5 euros per adult respondent; (3) those with a "high" cash incentive of 5 euros per household and 10 euros per adult respondent; and (4) those with a choice between a "low" cash incentive and a lottery ticket. The results showed slightly higher response rates in the cash groups, although the extra money in group (3) did not pay off.

The survey institute also does additional work to keep response rates high. Addresses are checked throughout the year to ensure that current addresses are on file. This is done, for instance, by sending out brochures about recent research based on the SOEP data and seasonal greeting cards.

Face-to-face interviews also ensure a personal relationship between interviewer and respondent, which increases the likelihood that respondents will stay in the survey. Keeping the same interviewer over time is therefore an important goal of the survey. Some SOEP respondents have in fact had the same interviewer since the beginning in 1984.

Last change: Nov 12, 2019

CHAPTER

FOUR

TARGET POPULATION AND SAMPLES

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

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

Institutionalized populations (in the true sense of the word, those living in hospitals, nursing homes, and military installations) are generally not representatively included in new samples. In 1984, for instance, only 57 institutionalized households were included. Later, however, individuals from initial survey households who have since taken up temporary or permanent residence in institutions were surveyed regularly.

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

To increase the overall sample size, SOEP started adding refresher samples in 2011. The first (in 2011) and second (in 2012) are representative of the entire population, whereas the third (2013) covers migrants. For the fourth such sample in 2014, the related study "Families in Germany" was integrated into the SOEP.

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

However, one of the ideas of the SOEP is that the users have full information available about survey methodological issues and survey design, which in this case means that you can identify the corresponding sample for each observation. In the following section, we present details on each of the samples, which unless stated otherwise are multi-stage random samples with regional clusters. The households are selected by random-walk routines.

For an extensive discussion on sampling (and weighting), see: Survey methods.

4.1 The SOEP Samples in Detail

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

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

Sample C "German Residents of the German Democratic Republic (GDR)" consists of individuals in private households in which the household head was a citizen of the German Democratic Republic (GDR). This meant that approximately 1.7% of the residential population of the GDR in June 1990 was excluded from the sample as foreigners (most of whom were living in "institutionalized" housing). In total, the sample started with 2,179 households with a sampling probability of about 0.0005.

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

Sample E "Refresher" was added in 1998, selected from the entire population of private households in Germany. The households were chosen independently of the ongoing panel and its subsamples A through D. The aim was to increase the number of observations of the general population and to preserve its representativity. The selection scheme used for sample E essentially resembles the one used in subsample A. The number of households in the first wave of subsample E was $1,060$, with a sampling probability of about 0.00005. With the 2012 data release, parts of subsample E were extracted into the SOEP Innovation Sample. It is also the first sample in which Computer-Assisted Personal Interview (CAPI) was used. At that time, interviews in Samples A-D were being conducted entirely using Paper-and-Pencil-Interviews (PAPI). To study mode effects, households from sample E were randomly allocated to either CAPI or PAPI.

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

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

Sample H "Refresher" started in 2006 as a random sample, again independently of all previous subsamples, covering all residential households in Germany. The added 1,506 households were sampled with a probability of 0.0001.

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

Sample J "Refresher Sample" started in 2011 as a random sample, independently of all previous subsamples, covering residential households in Germany. The added3,136 households were sampled with a probability of 0.0002.

Sample K "Refresher Sample" started in 2012 as a random sample, drawn independently of all previous subsamples, covering the residential households in Germany. The added 1,526 households were sampled with a probability of 0.0001.

Sample L1 "Cohort Sample" covers private households in Germany in which at least one household member was born between January 2007 and March 2010 and was therefore a child at that time. Again, migrants identified were

oversampled using an onomastic procedure. Sample L1 (as well as L2 and L3) was part of the SOEP-related study "Families in Germany" (FiD), which was integrated into the SOEP in 2014. As part of an evaluation project by the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) and the Federal Ministry of Finance (BMF), the study focused on public benefits in Germany for married people and families. Therefore, the survey instruments used in waves BA to BD differ in some respects from those used in the other samples.

Sample L2 "Family Types I" covers private households in Germany that meet at least one of the following criteria for household composition: single parents, low-income families, and large families with three or more children. Similar to Sample G, we face the problem that the eligible sub-population is relatively small and an adequate sampling frame is lacking. So again, a preceding telephone screening procedure identifies eligible households.

Sample L3 "Family Types II" covers private households in Germany that meet at least one of the following criteria for household composition: single parents or large families with three or more children. It is conducted analogously to Sample L2 to increase the number of cases in these sub-populations.

Sample M1 "Migration Sample" is a new migration sample added in 2013 with around 2,700 households drawn using register information from the German Federal Employment Agency. It includes individuals who immigrated to Germany after 1995 or second-generation immigrants.

Sample M2 "Migration Sample" was another migration sample added in 2015 with around 1,100 households drawn using register information from the German Federal Employment Agency. It includes individuals who immigrated to Germany between 2010 and 2013.

Sample M3 "Refugee Sample" was a new refugee sample added in 2016 for the IAB-BAMF-SOEP Refugee Survey in which roughly 1,769 refugee households were interviewed repeatedly. Respondents aged 18 and older who entered Germany between January 2013 and December 2016 and who had filed an asylum application by April 2016 (regardless of their current legal status) were interviewed along with the other members of their households.

Sample M4 "Refugee Family Sample": the 2016 "IAB-BAMF-SOEP Survey of Refugees" (Samples M3 and M4) is a joint project of the Institute for Employment Research (IAB), the Research Center of the Federal Office for Migration and Refugees (BAMF-FZ) and the Socio-Economic Panel (SOEP). The target population of the samples consists of 1,769 households with individuals who arrived in Germany between January 2013 and January 2016 and had applied for asylum by June 2016 or were hosted as part of specific programs of the federal states (irrespective of their asylum procedure and their current legal status). The first part of the sample (M3) was financed with funds allocated to the IAB from the research budget of the Federal Employment Agency (BA) . Sample M4 was funded by the Federal Ministry of Education and Research (BMBF) and has a focus on refugee families.

Sample M5 "Refugee Sample" M5 is the third boost sample of refugee households. The population of M5 covers adult refugees who applied for asylum in Germany between January 1, 2013, and December 31, 2016, and are currently living in Germany. The first wave of M5 was conducted in 2017. M5 added another 1,519 households of refugees who have migrated to Germany since 2013 to the SOEP framework.

Sample N "Refresher Sample (PIAAC-L)": Sample N integrated 2,314 households of former participants in the Program for the International Assessment of Adult Competencies (PIAAC and PIAAC-L) in 2017. This is the most recent addition to the SOEP-Core samples. Fieldwork in sample N was conducted between mid-March and mid-August and thus slightly later than the majority of samples A–L1.

More information about "Sample Sizes" and "Panel Attrition" can be found here

Sample-Specific Questionnaires

In SOEP it is common for special samples to receive extended, adapted, and/or integrated questionnaires in the first few years. This ensures that sample-specific questions that do not play a role in the main SOEP can also be included. In the following tables you can see which questionnaires the respective samples received, which years they ran, which raw data set they were included in, and which "long" data set they went into.

Sample Specific Instruments:

Sample A

Year	1984	1985	1986	1987	1988	1989	
Version	1	2	3	4	5	6	
Questionnaire / Wave	a	b	c	d	e	f	long Dataset
Household	$h, $kind						hl, kidlong
Household (New Respondents)		$h, $kind					hl, kidlong
Individual	$p, $pkal, biolela*	$p, $pkal					pl, pkal, biol
Individual (New Respondents)		$p, $pkal, biolela*		$p, $pkal	$p, $pkal, biolela*	$p, $pkal	pl, pkal, biol
Biography				biolela*		biolela*	biol
Catch-Up Individual (Re-Questioning employed)				$pluecke			plueckl
Catch-Up Individual (Re-Questioning unemployed)				$pluecke			plueckl
Financial Statement (Vermögensbilanz)					ev		

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

Sample B

Year	1984	1985	1986	1987	1988	1989	
Version	1	2	3	4	5	6	
Questionnaire / Wave	a	b	c	d	e	f	long Dataset
Household (Foreigners Version)			$h, $kind				hl, kidlong
Household (Foreigners Version New Respondents)			$h, $kind				hl, kidlong
Individual (Foreigners Version)	$pausl, $pkal, biolela*		$pausl, $pkal				pl, pkal, biol
Individual (Foreigners Version New Respondents)			$pausl, $pkal, biolela*	$pausl, $pkal	$pausl, $pkal, biolela*	$pausl, $pkal	pl, pkal, biol
Biography (Foreigners Version)				biolela*		biolela*	biol
Catch-Up Individual (Re-Questioning employed)				$pluecke			plueckl
Catch-Up Individual (Re-Questioning unemployed)				$pluecke			plueckl
Financial Statement (Foreigners)					ev		

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

From the start of Sample B (foreigners), respondents could complete the individual questionnaire in German or in the respective foreign language. Starting with wave 2 of the panel, there were "old" and "new" survey units (households, persons), and there were survey units with or without certain changes (e.g., households that had or had not moved; individuals who had or had not changed careers). The questionnaires took these changes into account for all subgroups. Survey procedures and tools were designed to ensure that each subgroup received the right questionnaire for them. This technique as well as the bilingual design of the foreigner questionnaires was retained for waves 3-6. In addition, retrospective information and missing information on temporary drop outs was collected. The "financial statement", which is now a survey module, was a separate questionnaire in the year 1988.

Sample Specific Instruments: Sample C

Year	1990	1991	1992	1993	1994	1995	
Version	7	8	9	10	11	12	
Questionnaire / Wave	g	h	i	j	k	l	long Dataset
Household East	ghost	$h, $kind					hl, kidlong
Individual East	$post, $pkalost, biolela*	$post, $pkalost	$p, $pkal				pl, biol, pkal
Individual East (New Respondents)		$post, $pkalost, biolela*	$p, $pkal				pl, biol, pkal
Biography East			biolela*				biol

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

SOEP researchers were determined to seize the historic opportunity of German reunification to obtain a first baseline measurement of incomes in the "old" GDR currency. The questionnaire was prepared by an East-West working group including DIW Berlin, WZB, Collaborative Research Centre 3, and the ISS at the Academy of Sciences in the GDR, with the participation of Infratest and its partner organization in the GDR. The result was a questionnaire that covered many of the same themes and questions and was structured similarly to the West SOEP questionnaire, but which focused more on the specific situation in the GDR (e.g., the housing situation).

Sample Specific Instruments:

Sample J

Year	2011	2012	2013	
Version	28	29	30	
Questionnaire / Wave	bb	bc	bd	long Dataset
Individual with Biography	$p, $lela*, $pkal			pl, biol, pkal
Household	$h, $kind			hl, kidlong
Youth	$jugend			jugendl
Individual		$p, $pkal		pl, pkal
Mother-Child (Newborns)		$muki*		bioagel
Mother-Child (2-3-year-olds)		$muki2*		bioagel
Mother-Child (5-6-year-olds)		$muki3*		bioagel
Parents (7-8-year-olds)		$elt*		bioagel
Mother-Child (9-10-year-olds)		$muki5*		bioagel
Deceased Individual		$vp		vpl
Lust auf DJ (cognitive Test)				cogdj
Grip Strength				gripstr

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

Sample K

Year	2012	2013	
Version	29	30	
Questionnaire / Wave	bc	bd	long Dataset
Individual with Biography	$p, $lela*, $pkal		pl, biol
Household	$h, $kind		hl, kidlong
Youth	$jugend		jugendl
Individual		$p, $pkal	pl, pkal
Youth			jugendl
Mother-Child (Newborns)		$muki*	bioagel
Mother-Child (2-3-year-olds)		$muki2*	bioagel
Mother-Child (5-6-year-olds)		$muki3*	bioagel
Parents (7-8-year-olds)		$elt*	bioagel
Mother-Child (9-10-year-olds)		$muki5*	bioagel
Deceased Individual		$vp	vpl
Lust auf DJ (cognitive Test)			cogdj
Grip Strength			gripstr

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

A major shift in the design of SOEP questionnaires took place with Sample J. Due to the increased panel mortality from wave 1 to wave 2 that was observed for the refresher samples F (2000- 2001), H (2006-2007), and I (2009-2010), the biographical module, with an average interview length of 17 minutes, was integrated into wave 1. If this had not been done, no biographical data would have been collected for approximately 20% of all SOEP respondents who would probably not have participated in wave 2. In comparison to the longitudinal samples, data collection in the first wave was focused on the main three questionnaires: the household, the individual, and the youth questionnaire. As the fieldwork in these refresher samples was conducted exclusively by CAPI, it was feasible to include complex modules with event-triggered question loops.

Sample Specific Instruments: Samples L1-L3

Year	2010	2011	2012	2013		
Version	27	28	29	30		
Questionnaire / Wave	ba	bb	bc	bd	Samples	long Dataset
Cohort Sample						
Household	$h, $kind				L1	hl, kidlong
Household New Respondents		$h, $kind			L1	hl, kidlong
Individual with Biography	$p, $lela*, $pkal				L1	pl, biol, pkal
Individual		$p, $pkal			L1	pl, pkal
Youth	$jugend				L1	jugendl
Catch-Up Individual (Re-Questioning)			$pluecke		L1	plueckel
Parents 1	$muki1*				L1	bioagel
Parents 2	$muki2*				L1	bioagel
Parents 3	$muki3*				L1	bioagel
Parents 4	$muki4*				L1	bioagel
Parents 5	$elt*				L1	bioagel
Parents 6	$muki5*				L1	bioagel
Screening Sample						
Household	$h, $kind				L2+L3	hl, kidlong
Household New Respondents		$h, $kind			L2+L3	hl, kidlong
Individual with Biography	$p, $lela*, $pkal				L2+L3	pl, biol, pkal
Individual		$p, $pkal			L2+L3	pl, pkal
Youth	$jugend				L2+L3	jugendl
Catch-Up Individual (Re-Questioning)			$pluecke		L2+L3	plueckel
Parents 1	$muki1*				L2+L3	bioagel
Parents 2	$muki2*				L2+L3	bioagel
Parents 3	$muki3*				L2+L3	bioagel
Parents 4	$muki4*				L2+L3	bioagel
Parents 5	$elt*				L2+L3	bioagel
Parents 6	$muki5*				L2+L3	bioagel

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

The main focus of Families in Germany (FiD) was on the families and children – the parental questionnaires (filled out by parents about their children) were about twice as long as the comparable questionnaires in SOEP-Core, and questionnaires for the 1-2-year-olds and the 9-10-year-olds were added (as of 2012, SOEP-Core had added a questionnaire for 9-10-year-olds that is partly comparable to the FiD version). In large part, FiD resembled the SOEP. Each adult was asked to answer an individual questionnaire, which, in the first two years, included retrospective questions on childhood, education, and early work experience. In addition, there were several questions designed to capture the challenges families face with regard to the return of mothers into the labor market – with respect to workplace, work schedule, overtime, daycare options, etc.

Sample Specific Instruments:

Sample M1 (IAB-SOEP-Migration Sample)

Year	2013	2014	2015	2016	
Version	30	31	32	33	
Questionnaire / Wave	bd	be	bf	bg	long Dataset
Household	$h, $kind				hl, kidlong
Individual with Biography	$p, $lela*, $pkal				pl, biol, pkal
Individual			$p, $pkal		pl, pkal
Youth		$jugendl			jugendl
Pre-Teen		$school			biopupil
Mother-Child (Newborns)		$muki*			bioagel
Mother-Child (2-3-year-olds)		$muki2*			bioagel
Mother-Child (5-6-year-olds)		$muki3*			bioagel
Parents (7-8-year-olds)		$elt*			bioagel
Mother-Child (9-10-year-olds)		$muki5*			bioagel
Deceased Individual		$vp			vpl
Lust auf DJ (Cognitive Test)					cogdj
Catch-Up Individual			$pluecke		plueckel

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

Sample M2 (IAB-SOEP-Migration Sample)

Year	2015	2016	
Version	32	33	
Questionnaire / Wave	bf	bg	long Dataset
Household	$h, $kind		hl, kidlong
Individual with Biography	$p, $lela*, $pkal		pl, biol, pkal
Youth	$jugend		jugendl
Individual		$p, $pkal	pl, pkal
Pre-Teen		$school	biopupil
Early Youth		$school2	biopupil
Mother-Child (Newborns)		$muki*	bioagel
Mother-Child (2-3-year-olds)		$muki2*	bioagel
Mother-Child (5-6-year-olds)		$muki3*	bioagel
Parents (7-8-year-olds)		$elt*	bioagel
Mother-Child (9-10-year-olds)		$muki5*	bioagel
Deceased Individual		$vp	vpl
Lust auf DJ (Cognitive Test)			cogdj
Catch-Up Individual		$pluecke	plueckel

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation
Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

Following the design shift for refresher samples since Sample J in 2011, respondents have been surveyed on their life history using the "biography questionnaire", which was integrated into the individual questionnaire from wave 1. This ensures that biographical information will be available for all target persons who provided an individual interview in participating households. Other supplementary questionnaires were not included in the survey instruments given to first-wave respondents to avoid "overburdening" respondents with an extremely lengthy first-wave interview. Questionnaires for the migration boost samples include questions that have been part of SOEP-Core for the last three decades. In addition, the survey covers each respondent's complete migration history, education, training, and employment history in Germany and abroad, and numerous aspects of cultural and living environments relevant to the social integration of migrants. The household questionnaire is identical to the questionnaire used in the SOEP-Core sample.

Sample Specific Instruments: Samples M3-M5 (IAB-BAMF-SOEP Refugee Sample)

Year	2016	2017		
Version	33	34		
Questionnaire / Wave	bg	bh	Samples	long Dataset
Household	$h		M3-M4	hl
Individual with Biography	$p, $lela*, bhpkal		M3-M4	pl, biol, pkal
Individual with Biography First Respondent		$p, $lela*	M3-M4	pl, biol
Individual with Biography non-fugitive		$p, $lela*	M3-M4	pl, biol
Youth (12-17-year-olds)		$school, $school2, $jugend	M3-M4	biopupil, jugendl
Children in household		$muki-$muki5*, $kind	M3-M4	bioagel, kidlong
Assessment of declarative knowledge and general cognitive ability in refugee children and adolescents		cog_refu	M3-M4	
Household		$h, $kind	M5	hl, kidlong
Individual with Biography		$p, $lela*	M5	pl, biol
Living Area (Wohnumfeld)		$brutt17	M5	hbrutto

* Not part of the data distribution file, only available as long-file
$: Wave abbreviation

Note that the samples are continued up to the current wave. Here only the sample specific instruments are shown with reference to the data sets.

As with every other previously established subsample of migrants in the SOEP (M1 and M2), there was a clear need for several deviations from standard SOEP-Core questionnaires to reflect the special characteristics of the target group. Several additional questions concerning migration and integration were incorporated into the individual questionnaire to better field the range of research questions and research goals of the project partners. These included topics such as ethnic background, experiences en route to Germany, language skills, integration courses in Germany, job experience, current occupation, educational background, health, attitudes, and values. The household questionnaire was much more SOEP-related than the individual questionnaire in order to establish longitudinal information on the households.

Last change: Dec 09, 2019

4.2 Eligibility and Follow-up

As mentioned, the SOEP's goal is to be representative of the residential population of Germany. All household members 16 and older are eligible for a personal interview, starting with the youth questionnaire for their age group, followed by "regular" individual questionnaires thereafter. As years go by, the children from the first wave reach age eligibility and become panel members. If they move out and start their own families, they and their new family members are also part of the survey. "New" individuals become part of the SOEP population by being born into SOEP households or as a result of residential mobility. If a person enters a SOEP household after the initial wave in which that household was surveyed, this person is asked to fill out the regular individual questionnaire if age-eligible or will

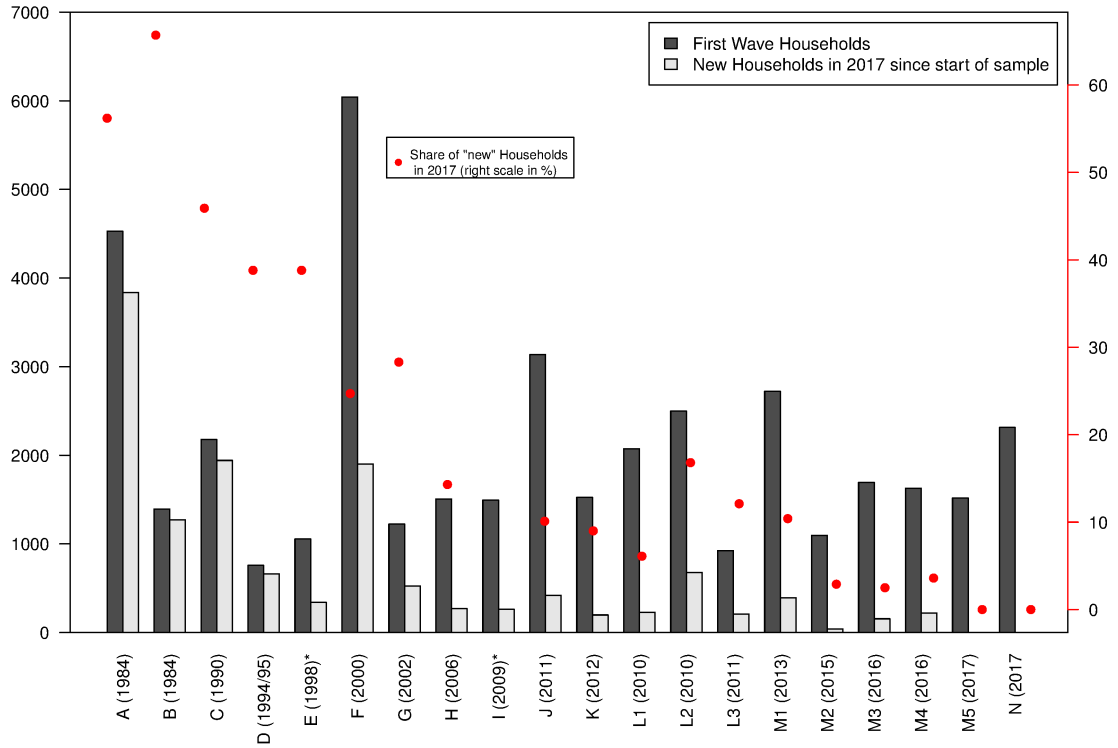
be asked to participate once old enough. In the absence of panel attrition, this would make the SOEP a self-sustaining survey.

The concept of how to follow respondents and sample members over time is important for the representativeness of the study. The basic principle for follow-up in the SOEP is that all persons participating in a wave of any subsample will be surveyed in the following years as long as they stay within the boundaries of Germany. This rule also extends to respondents who entered a SOEP household after the first wave in which it was surveyed due to residential mobility or birth. If there is a "split-off", that is, if someone moves out of the household in which they were last interviewed, the members of the new household receive a new household identifier. The table conceptualizes how new sample members and households are surveyed in the SOEP. The figure shows that as a result of the follow-up concept, several thousand "new" households had become part of the SOEP population.

Individuals or households that could not be interviewed in a given year are termed "temporary drop-outs". They are followed until there are two consecutive waves of missing interviews for all household members or until the entire household refuses to participate further. In the case of a temporary drop-out, in which a respondent participates in the survey again after not participating in the previous wave, the respondent is asked to fill out an additional short questionnaire covering key information about employment and demographics in the year of their absence.

	Existing Households	New Households
Existing Individuals	Classic case: without change of address entire household moves	Respondent moves out of existing household and forms a new household
New Individuals	Born into household, move into household	Moved into or born into HH formed after a respondent moved out of existing HH

Changes to the Sample: Old and new household in the SOEP



Download R Code to create figure

Last change: Nov 12, 2019

4.3 Development of Sample Sizes

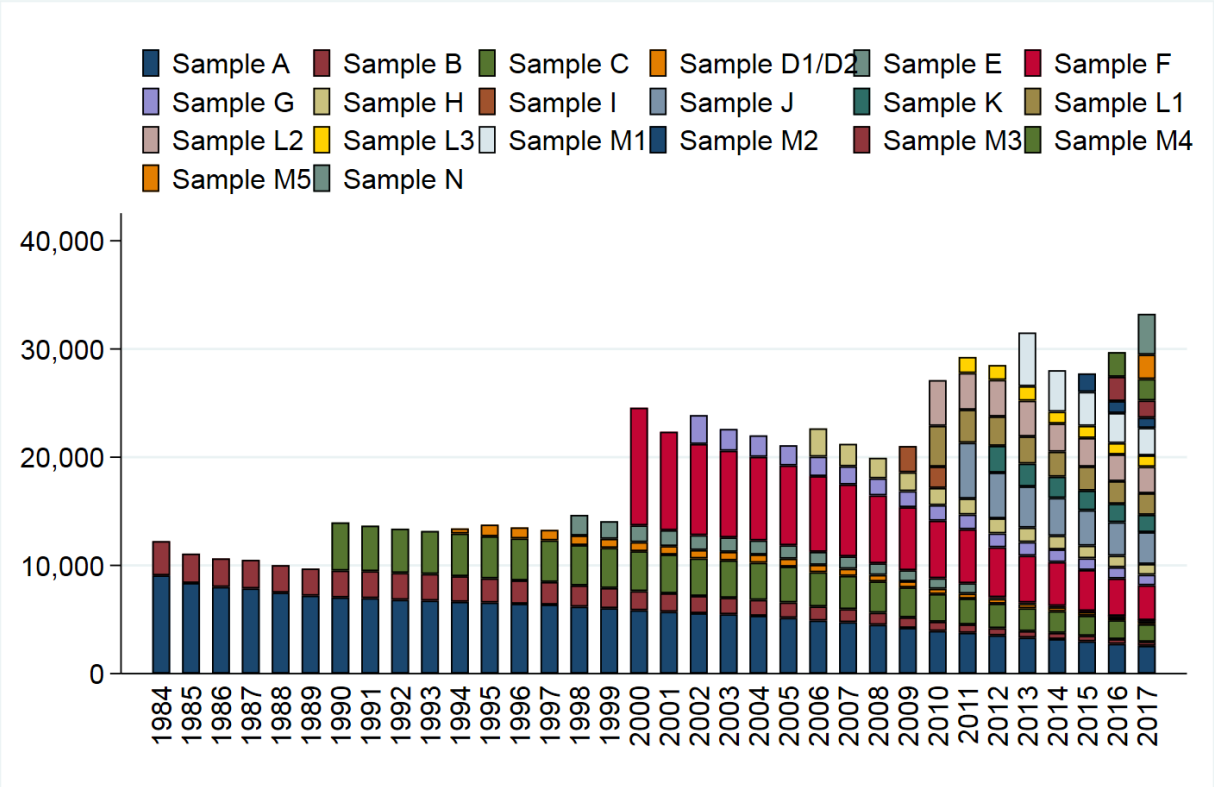
Individuals who decline to take part in the survey or are not available for an interview are kept in the so-called "gross" sample of the study as long as they continue to live in households with at least one participating respondent. If the entire household declines to participate in two consecutive waves, all individuals in the household are removed from the SOEP. The table shows the starting sample sizes of samples A through M4, the years when the samples were first collected, as well as the percentage of those persons who were eligible for an interview but declined participation ("partial unit non-response", PUNR) in the first wave. The figure illustrates the development of the number of successful person interviews since 1984. The reduction in the population size for all individual samples is mainly the result of individual-level drop-outs, refusals, moving abroad, etc. However, due to new persons moving into already existing households and children reaching the age of 16 and thereby increasing the sample size, this negative development is offset somewhat.

Starting Sample Size of the SOEP Samples

Surveyyear	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Sample / Year	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z	ba	bb	bc	bd	be	bf	bg	bh
A (Germans)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
B (Foreigners)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34
C (German Democratic Republic (GDR)							1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
D1 (Immigrants)											1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
D2 (Immigrants)												1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20	21	22
E (Refreshment Sample)															1	2	3	4	5	6	7	8	9	10	11	12	13	14	15*	16*	17	18	19	20
F (Refreshment Sample)																	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
G (High Income)																		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
H (Refreshment Sample)																							1	2	3	4	5	6	7	8	9	10	11	12
I (Incentive Sample)																										1	2	-	-	-	-	-	-	
L1 (Family Types 10)																											1	2	3	4	5	6	7	8
L2 (Family Types 10)																											1	2	3	4	5	6	7	8
L3 (Family Types 11)																											1	2	3	4	5	6	7	8
J (Refreshment Sample)																												1	2	3	4	5	6	7
K (Refreshment Sample)																													1	2	3	4	5	6
M1 (Migration Sample)																														1	2	3	4	5
M2 (Migration Sample)																																1	2	3
M3 (Refugee Sample)																																	1	2
M4 (Refugee Family Sample)																																	1	2
M5 (Refugee Sample)																																		1
N (Refreshment Sample PIAAC-L)																																		1

	Year	Households	Persons	Respondents	Children
A	1984	4528	11422	9076	2290
B	1984	1393	4830	3169	1638
C	1990	2179	6131	4453	1591
D1	1994	236	733	471	248
D1/D2	1995	541	1668	1078	517
E	1998	1057	2446	1910	466
F	2000	6043	14510	10880	2991
G	2002	1224	3538	2671	693
H	2006	1506	3407	2616	623
I	2009	1495	3428	2432	620
L1	2010	2074	7939	3770	3900
L2	2010	2500	9063	4227	4611
L3	2011	924	3645	1487	2092
J	2011	3136	6873	5161	1147
K	2012	1526	3286	2473	563
M1	2013	2723	8522	4964	2481
M2	2015	1096	3048	1689	927
M3	2016	1678	4609	2213	1744
M4	2016	1611	6737	2252	3641
M5	2017	1519	4771	2252	1847
N	2017	2314	5665	3720	1011
Total		39602	115538	72493	31911

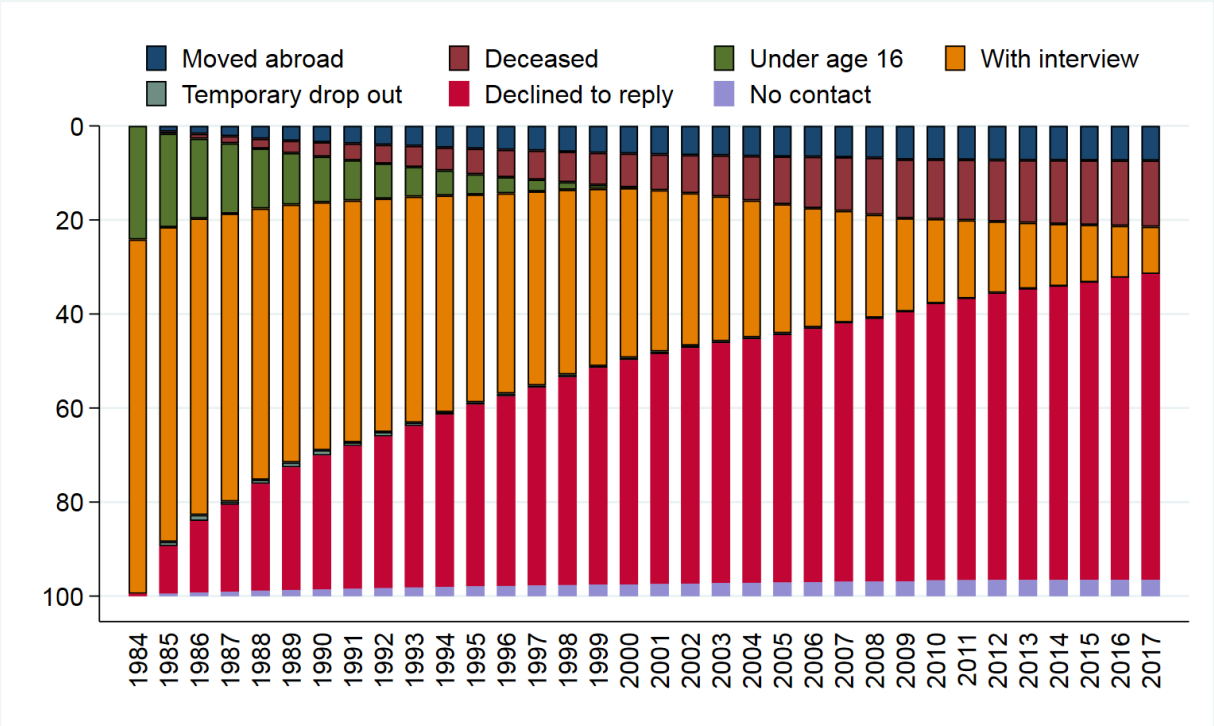
Cross-Sectional Development of Sample Size (Respondents)



Download Stata Code to create figure

This cross-sectional view is insufficient when examining the longitudinal development of the sample, which is influenced by different demographic and fieldwork-related factors. As already shown, demographic reasons for entering the panel are birth and residential mobility. Analogously, the demographic reasons for a panel exit are death and moving abroad. Fieldwork-related reasons are different, in that they relate to the interaction between the interviewer and the responding household. Respondents are either not reached for an interview (non-contact) or they decline to participate for the current year. The figure illustrates the longitudinal development of first-wave respondents in 1984, as well as their children, of samples A and B.

Longitudinal Development of the 1984 Population



Download Stata Code to create figure

Last change: Nov 12, 2019

CHAPTER

FIVE

DATA STRUCTURE OF SOEP-CORE

5.1 Principles of Data Analysis

All SOEPtutorials can be found on our YouTube Channel

The structure of panel data has three dimensions. First, the respective examination units (n) and a matrix of dependent and independent variables (y,x) are completely analogous to a cross-sectional design. Second, the dimension of time (t), whereby a distinction is made between two data formats for panel data structures - "wide" or "long" (with wide format the variable matrix is indexed with the dimension of time and with long format the respective examination units). Regardless of the selected data format, when using panel data with several survey waves, the data matrices often do not contain complete information due to the panel mortality of individual survey units or because data from new panel members are only collected at a later point in time. In both cases, the term "unbalanced panel data" is used. In contrast, the classical panel data structure, on the other hand, is "balanced", i.e., as many observations of dependent and independent variables are available for all study units as there are waves of data collection. Social science panel data often show a data structure characterized by many investigation units (large n) as well as, in relation to it, few waves and therefore measuring time (small t). When data from a panel study are available, even descriptive forms of data analysis are often of particular interest, since the identification of changes in a variable over time and the corresponding separation of interindividual and intraindividual changes can represent important social facts, particularly in the case of generalizable samples. It is of social scientific interest whether a constant 15% proportion of people whose income is below the poverty risk level is repeatedly found in the same person over time, or whether there was a even balance of increases and decreases in poverty risks and only half of the population was permanently exposed to the risk. The choice of complex analysis methods for panel data depends first and foremost on the respective measurement level of the dependent and independent variables, but also on whether they are time-constant variables (such as gender or migration background) or time-invariant variables. The statistical analysis models of panel data range from structural equation models, various regression models, event analysis, sequence data analysis, latent growth models to causal analyses using matching methods. A particular advantage of panel data is that the chronological sequence of changes can be modelled and calculated and the problem of unobserved heterogeneity, which is often encountered in the social sciences, can be significantly reduced, at least in comparison with cross-sectional data.

Cross-Sectional Data Structure (CS)

Cross-sectional data is a type of data that observes many subjects at the same point in time. Each person is assigned a row in the dataset and is only included once in such a dataset. By merging cross-sectional SOEP data across waves, you obtain a dataset in wide-format.

Row	ID	wave	sex	income
1	1	2015	m	1500
2	2	2015	m	1000
3	6	2015	f	2000
4	8	2015	m	5500

Data Structure in "Wide" Format (wide)

The SOEP data are available with different data structures. In the wide format, a respondent's repeated responses are displayed in a single row and each response in a separate column. Each column represents a variable. We provide four datasets in the wide format: ppath, phrf, hpath, hhrf.

Row	ID	sex	income2015	income2016	income2017
1	1	m	1500	1500	2000
2	2	m	1000	1200	1200
3	6	f	2000	2000	2000
4	8	m	5500	6000	6500

Data Structure in "Long" Format (long)

The long format is a condensed and user-friendly dataset structure for longitudinal section analysis. Here, each person has one line per survey year. This means that you do not have several datasets for the different waves, but one dataset in which all survey waves are represented. A person can appear more than once in such a dataset. In the long format, one line describes a person-year combination.

Row	ID	syear	sex	income
1	1	2015	m	1500
2	1	2016	m	1500
3	1	2016	m	2000
4	2	2015	m	1000
5	2	2016	m	1200
6	2	2016	m	1200
7	6	2015	f	2000
8	6	2016	f	2000
9	6	2017	f	2000

Data Structure in Spell Format (spell)

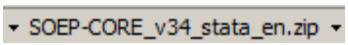
In the strict sense of the word, spell data are about time periods with a defined start and end. When handling spell data it is necessary to take potential censoring into account. Censoring denotes that the beginning (left censored) or ending (right censored) of a spell is imprecise because of missing information or the beginning or ending of a spell is outside of the period of observation. 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 are useful for looking at stays in a certain state, and transitions in and out of that state.

Row	ID	spellnr	spelltype	begin	end	censored
1	1	1	Retired	1983	2007	left and right censored
2	1	2	Housewife/husband	1983	1984	left censored
3	1	3	Housewife/husband	1994	1994	uncensored
4	1	4	Housewife/husband	1998	1998	uncensored
5	2	1	Full-Time Employment	1984	1984	left censored
6	2	2	Full-Time Employment	1985	1985	uncensored

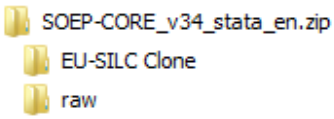
Last change: Nov 12, 2019

5.2 Data Distribution File

In the SOEP, each survey year is allocated to a data wave, which is abbreviated using the letters of the alphabet. One data wave may be released in several versions, which are displayed in SOEP with a "v" for version and the respective version number. The version number represents the survey years since the beginning of the survey. The SOEP has recently published the 34th version since the survey began in 1984. Within a data wave, updates may be made over time, such as v34.1. If updates have been made, users will be informed through various channels and be asked to order the data again. After ordering the data, the data will be sent to you in a zip file.

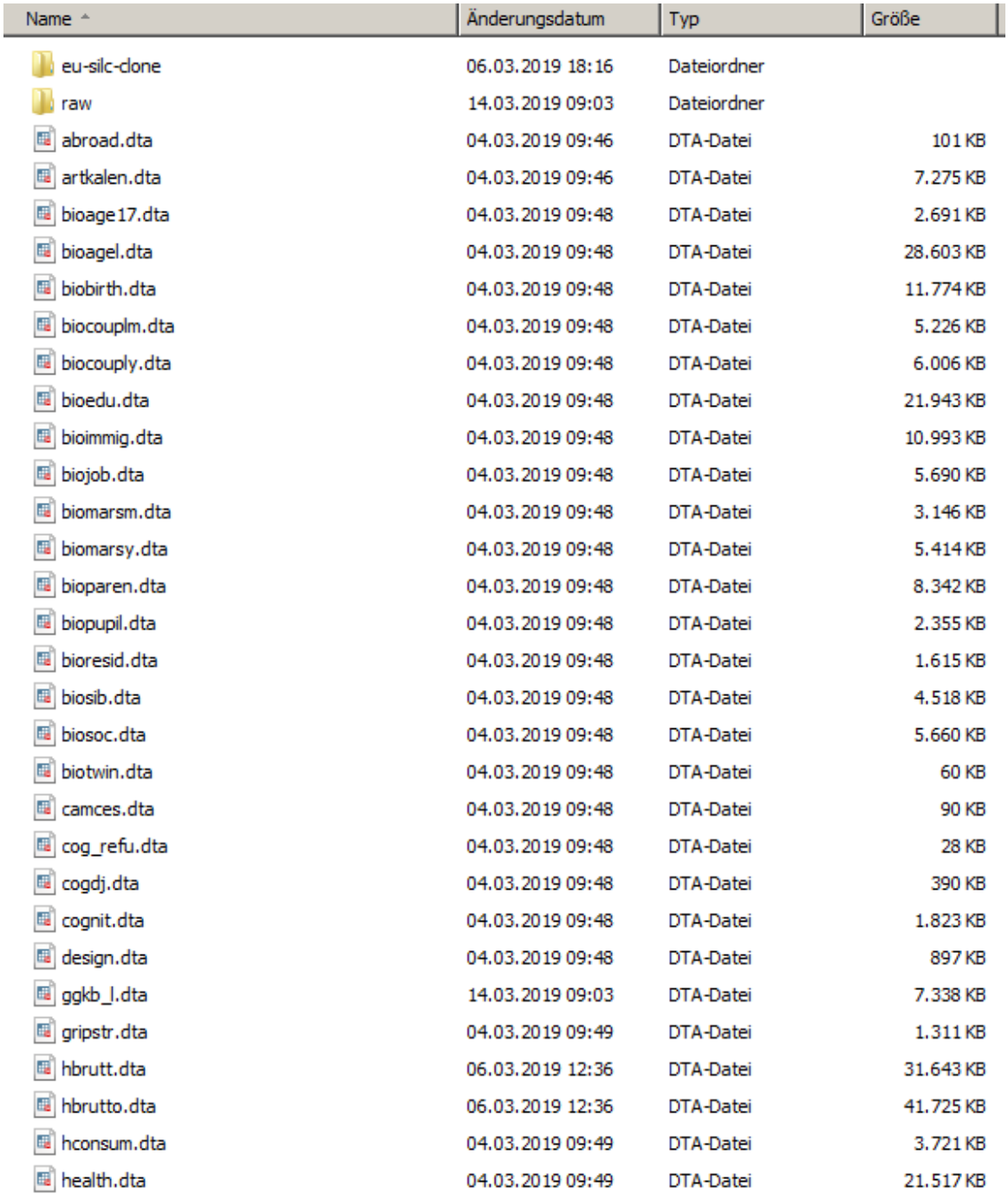


Within this zip file you will find various data sets, a "raw" subdirectory and the "EU-SILC Clone" subdirectory.



The datsets above the "RAW" subdirectory are a highly compressed and easy-to-analyze version of the SOEP data.

Note: SOEP strongly recommends that users use the data above the "raw" subdirectory.



Name	Änderungsdatum	Typ	Größe
eu-silc-clone	06.03.2019 18:16	Dateiordner	
raw	14.03.2019 09:03	Dateiordner	
abroad.dta	04.03.2019 09:46	DTA-Datei	101 KB
artkalen.dta	04.03.2019 09:46	DTA-Datei	7.275 KB
bioage17.dta	04.03.2019 09:48	DTA-Datei	2.691 KB
bioagel.dta	04.03.2019 09:48	DTA-Datei	28.603 KB
biobirth.dta	04.03.2019 09:48	DTA-Datei	11.774 KB
biocouplm.dta	04.03.2019 09:48	DTA-Datei	5.226 KB
biocouply.dta	04.03.2019 09:48	DTA-Datei	6.006 KB
bioedu.dta	04.03.2019 09:48	DTA-Datei	21.943 KB
bioimmig.dta	04.03.2019 09:48	DTA-Datei	10.993 KB
biojob.dta	04.03.2019 09:48	DTA-Datei	5.690 KB
biomarsm.dta	04.03.2019 09:48	DTA-Datei	3.146 KB
biomarsy.dta	04.03.2019 09:48	DTA-Datei	5.414 KB
bioparen.dta	04.03.2019 09:48	DTA-Datei	8.342 KB
biopupil.dta	04.03.2019 09:48	DTA-Datei	2.355 KB
bioresid.dta	04.03.2019 09:48	DTA-Datei	1.615 KB
biosib.dta	04.03.2019 09:48	DTA-Datei	4.518 KB
biosoc.dta	04.03.2019 09:48	DTA-Datei	5.660 KB
biotwin.dta	04.03.2019 09:48	DTA-Datei	60 KB
camces.dta	04.03.2019 09:48	DTA-Datei	90 KB
cog_refu.dta	04.03.2019 09:48	DTA-Datei	28 KB
cogdj.dta	04.03.2019 09:48	DTA-Datei	390 KB
cognit.dta	04.03.2019 09:48	DTA-Datei	1.823 KB
design.dta	04.03.2019 09:48	DTA-Datei	897 KB
ggkb_l.dta	14.03.2019 09:03	DTA-Datei	7.338 KB
gripstr.dta	04.03.2019 09:49	DTA-Datei	1.311 KB
hbrutt.dta	06.03.2019 12:36	DTA-Datei	31.643 KB
hbrutto.dta	06.03.2019 12:36	DTA-Datei	41.725 KB
hconsum.dta	04.03.2019 09:49	DTA-Datei	3.721 KB
health.dta	04.03.2019 09:49	DTA-Datei	21.517 KB

The data in SOEP-Core are no longer provided only as wave-specific individual files but are now pooled across all available years (in "long" format). In some cases, variables are harmonized to ensure that they are defined consistently over time. For example, the income information provided up to 2001 is given in euros, and categories are modified over time when versions of the questionnaire have been changed. The longitudinal nature of the data is one of the biggest assets of the SOEP. This is why we provide longitudinal datasets such as PL or HL. The advantage of such a dataset is that longitudinal analyses can be carried out without great effort.

If you need more information about the "long" data structure, see chapter *Data Structure in "Long" Format (long)*.

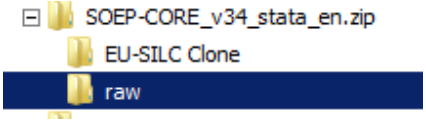
Core Datasets

The data sets above the "raw" subdirectory:

Tracking Data	Original Data	Survey Data	Generated Data	Spell Data
ppathl	pl	csamp	pgen	artkalen
hpathl	hl	design	hgen	biocouplm
pbrutto	jugendl	exit	bioage17	biocouply
hbrutto	plueckel		bioagel	biomarsm
pbr_exit	abroad		kidlong	biomarsy
	vpl		pequiv	einkalen
			biobirth	lifespell
			bioedu	migspell
			bioimmig	pbiospe
			biojob	refugspell
			bioparen	sozkalen
			biopupil	
			bioresid	
			biosib	
			biosoc	
			biotwin	
			camces	
			cogdj	
			cognit	
			cog_refu	
			gripstr	
			hconsum	
			health	
			hwealth	
			interviewer	
			mihinc	
			pflege	
			pkal	
			pwealth	
			timepref	
			trust	

Raw Datasets

In the "raw" directory, you will find all wave-specific datasets that were used to generate the long datasets on the previously presented level.



Attention: Please note that the datasets above the raw subdirectory are completely sufficient for your data analysis. The datasets used to generate the SOEP-Core data can be found in the raw subdirectory. Detailed information about the raw datasets can be found here Raw **"raw"**

Name	Änderungsdatum	Typ	Größe
ah.dta	30.01.2018 03:30	DTA-Datei	738 KB
ahbrutto.dta	30.01.2018 03:30	DTA-Datei	122 KB
ahgen.dta	30.01.2018 03:30	DTA-Datei	517 KB
akind.dta	30.01.2018 03:30	DTA-Datei	187 KB
ap.dta	30.01.2018 03:30	DTA-Datei	4.195 KB
apausl.dta	30.01.2018 03:30	DTA-Datei	205 KB
apbrutto.dta	30.01.2018 03:30	DTA-Datei	434 KB
apequiv.dta	30.01.2018 03:30	DTA-Datei	5.865 KB
apgen.dta	30.01.2018 03:30	DTA-Datei	1.952 KB
apkal.dta	30.01.2018 03:30	DTA-Datei	9.770 KB
bah.dta	30.01.2018 03:31	DTA-Datei	7.770 KB
bahbrutto.dta	30.01.2018 03:31	DTA-Datei	949 KB
bahgen.dta	30.01.2018 03:31	DTA-Datei	1.566 KB
bajugend.dta	30.01.2018 03:31	DTA-Datei	1.151 KB
bakind.dta	30.01.2018 03:31	DTA-Datei	1.315 KB
bap.dta	30.01.2018 03:31	DTA-Datei	28.594 KB
bapbrutto.dta	30.01.2018 03:31	DTA-Datei	2.697 KB
bapequiv.dta	30.01.2018 03:31	DTA-Datei	18.277 KB
bapgen.dta	30.01.2018 03:31	DTA-Datei	3.966 KB
bapkal.dta	30.01.2018 03:31	DTA-Datei	15.446 KB
bapluecke.dta	30.01.2018 03:31	DTA-Datei	117 KB
bavp.dta	30.01.2018 03:31	DTA-Datei	41 KB
bbh.dta	30.01.2018 03:31	DTA-Datei	9.127 KB
bbhbrutto.dta	30.01.2018 03:31	DTA-Datei	1.028 KB
bbhgen.dta	30.01.2018 03:31	DTA-Datei	1.706 KB
bbjugend.dta	30.01.2018 03:31	DTA-Datei	1.198 KB
bbkind.dta	30.01.2018 03:31	DTA-Datei	1.452 KB
bbp.dta	30.01.2018 03:32	DTA-Datei	34.277 KB
bbpbrutto.dta	30.01.2018 03:32	DTA-Datei	2.960 KB
bbpequiv.dta	30.01.2018 03:32	DTA-Datei	19.560 KB

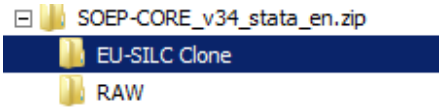
Within this "raw" directory, each wave is identified by letters of the alphabet: the first wave in 1984 is wave "A", 1985 is wave "B", and so on. To simplify the notation, the "$" sign is used when referring to all waves of one group of datasets. For example, $H refers to all household-level datasets from AH to now. For each year of SOEP data, there are single data files for households (e.g., $H) as well as for individual respondents (e.g., $P) and children (e.g., $KIND) based on interview information. These observations make up the "net" population, with each of these files containing as many records as interviews could be conducted. Additional data files with a limited number of variables based on the "address log" constitute the "gross" number of households and persons, i.e., all households and their members that were eligible for an interview in any given year. Within the "raw" directory, the datsets are stored on a wave-specific basis and are the basis for generating the majority of the long datsets described above. In addition to

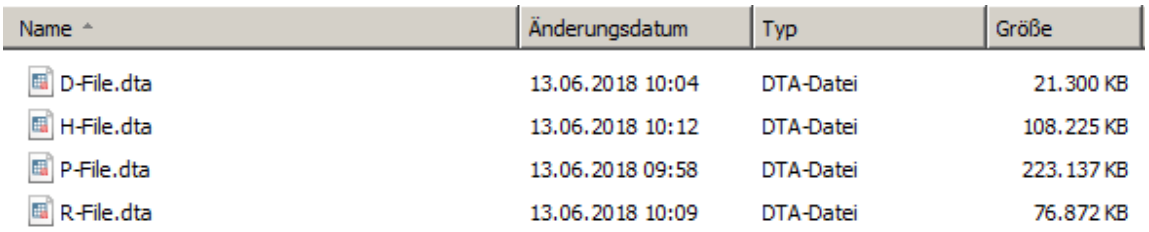
these wave-specific datsets, the "RAW" directory also contains additional datasets in cross-sectional format that have not yet been distributed in long format ($SCHOOL, $SCHOOL2, EV, EXIT, $PKALOST and PBR_HHCH).

Tracking Data	Original Data	Survey Data	Generated Data
ppfad	$p	phrf	$pgen
hpfad	$pausl	hhrf	$hgen
$pbrutto	$pluecke	pbr_hhch	$kind
$hbrutto	$h		$pequiv
	$post		$pkal
	$jugend		$pkalost
	$school		
	$school2		
	ev		
	$vp		
	biol		

EU-SILC Clone

Currently, the official German EU-SILC is provided only as a cross-sectional datset by the German Federal Statistical Office. A panel dataset will presumably be available from the year 2020 onwards (Bundesrat, 2016). As a consequence, Germany is excluded from cross-country studies exploiting the longitudinal dimension of EU-SILC. The aim of the EU-SILC clone is to provide an EU-SILC-like panel dataset for Germany from the year 2005 onwards so that Germany can be included in cross-country studies using EU-SILC panel data. The EU-SILC clone is built on the Socio-Economic Panel (SOEP) and therefore includes all EU-SILC panel variables for which the required information is recorded in the SOEP.





The EU-SILC Clone includes all of the four EU-SILC sub-datasets: The household register (D-File), the personal register (R-File), personal data (P-File), and household data (H-File). The clone datasets can be combined using the R-File, which includes both the current household and individual identifier. The identifiers in the EU-SILC Clone are unique and do not vary among the four datasets. Complete documentation on the datasets can be found here: Documentation EU-SILC.

Last change: Nov 12, 2019

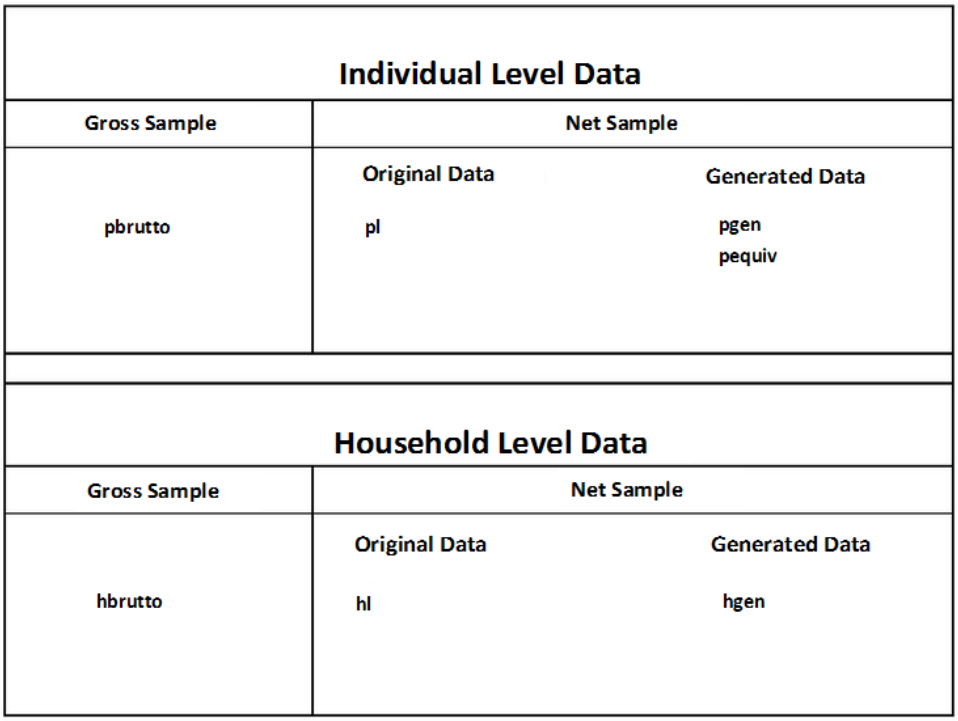
5.3 Datasets SOEP-Core

SOEP-Core contains a multitude of different datasets. To get an overview of the data, a somewhat simplified categorization helps:

There are *Tracking Data* and *Survey Data* files which describe the development of the sample, such that the user knows which individual or household was part of the interviewed sample in any given year. Then there are *Original Data* files, which contain the data from each year's questionnaires without any changes except for very basic consistency checks. To help the user with the data, there also are *Generated Data*. These contain consistently coded variables across all waves with common names, such that the users can easily use this information when combining datasets across waves. The SOEP also provides various data on the respondent's background, called biographical data. Biography data in general can conceptually be separated into biographical data which are unchanging (such as information on parent's education, or data from the Mother-Child Questionnaires) and data which may be updated through changes in a respondent's life (such as new children in the birth biography, or a job change in the job history). Some of the changing data are stored as *Spell Data*. For each spell there is a definition of the spell type, begin point, end point and the censoring status, indicating if a given employment or income spell is censored (left and/or right) or uncensored. One of the biggest assets of the SOEP data is their longitudinal nature, i.e., repeated observations of the same unit (individual or household) over time. That's why we provide longitudinal datasets, such as PL or HL. Finally, there are some files which cannot be easily categorized - some are one-time datasets, some provide information about the interviewers, some about respondents outside of Germany.

There are two datasets which should be the building block of any analysis, as they allow users to define longitudinal populations very easily: PPATHL and HPATHL. HPATHL includes all households which have been interviewed successfully at least once. Similarly, PPATHL contains all individuals who have ever lived in a household that has participated in the SOEP, i.e., that has been captured in HPATHL, including non-respondents and children. Both data files contain one record per household or individual, respectively, with wave-specific variables for each year's survey status. In addition to some time-invariant information (like gender, year of birth, migrant status), these files contain all necessary identifiers to combine other files with PPATHL and HPATHL. Although they provide essential information, PPATHL and HPATHL alone are of little use for actual analyses. The most often used sources for additional information in SOEP-Core are the cross-sectional data files provided in each survey year (or "wave") or the datasets in the "long" Format.

The SOEP datasets can be viewed based on their content classification (Tracking Data, Original Data, Survey Data, Generated Data and Spell Data), the data structure (cross-sectional (cs), wide, long, spell) and also from the respondent's perspective. From the respondent's perspective, datasets can contain gross or net information. In addition, some datasets provide information only at the household level and others provide information at the individual level.



Gross information at household or individual level is provided to users in the datasets hbrutto, hbrutt and pbrutto, pbrutt. Content information collected from household or individual questionnaires, for example, is original data and is stored in HL and PL. The SOEP team generates data from these original data, which are generated from the many SOEP questionnaires. New generated and user-friendly datasets such as pgen are created from the components of PL.

Tracking Data

Tracking data are the basis for linking your research-relevant variables. In addition to various demographic information, tracking data also provide information on how the interview was conducted. These datasets should be understood as initial data that you can use to merge your research-relevant variables via the individual and household numbers.

Dataset	Label	Format	Identifier (ID)	Additional Identifier
ppathl	Individual Tracking File	*long*	pid, syear	hid, cid, parid
hpathl	Household Tracking File	*long*	hid, syear	cid
pbrutto	Gross Individual Data	*long*	pid, syear	hid, cid, intid, hhnrold
hbrutto	Gross Household Data	*long*	hid, syear	cid, intid1, intid
pbr_exit	Cumulated Exit	*long*	pid, syear	hid, cid, hhnrold

[1] *In addition to the classic identifiers (pid, hid and cid), these datasets also have the identifiers of older data distribution versions. (pid=persnr; hid=hhnrakt; cid=hhnr).*

hpathl "Household Tracking File" (long): HPATHL consists of all waves of the raw datasets HPATH and HHRF. For all years since 1984, the HPATHL datset contains information on all households that have ever participated in the SOEP survey at any point in time. HPATHL is important for the delimitation of the unit of investigation (household), especially in longitudinal analysis. HPATHL is useful particularly for household analysis and can be used for preselection of specific households.

ppathl "Individual Tracking File" (long): PPATHL consists of all waves of the raw datasets PPATH and PHRF. For all years since 1984, the PPATHL datset contains information on all individuals who have ever lived in a SOEP household at the point in time of a survey (i.e., all respondents, but also children under 17 years of age and individuals who have never given an interview). PPATHL is important for the delimitation of the units of investigation (individuals), especially for longitudinal analysis. It contains one record for each individual and year a individual has been a member of a respondent household. It is keyed on pid and syear, the survey year identifier. It contains the Household ID, the unvarying individual characteristics, individual weights, as well as the response status for that individual in each wave.

pbrutto "Gross Individual Data" (long): PBRUTTO consists of all waves of the raw datasets $PBRUTTO. PBRUTTO covers all respondents who were either interviewed for the first time or contacted for the purpose of being interviewed again in a given wave. The datset provides gross information on all SOEP respondents' interviews as well as their positions in the panel framework.

hbrutto "Gross Household Data" (long): HBRUTTO consists of all waves of the raw datasets $HBRUTTO. HBRUTTO covers all households that were successfully interviewed for the first time in a wave or were contacted for the purpose of being interviewed again. The datasets provide gross information on all SOEP households' interviews as well as their positions in the panel framework.

pbr_exit "Cumulated Exit" (long): The dataset pbr_exit is a supplement of pbrutto for individual dropouts. Individual dropouts are removed from the original pbrutto population, so that pbrutto covers all current household members. Pbr_exit contains the corresponding register information on individual dropouts from households.

Original Data

These datsets contain respondents' direct information. The contents of these variables mirror the contents of the survey instruments. By searching the questionnaires, you can determine the exact wording of the question and obtain possible filter guidance.

Dataset	Label	Format	Identifier (ID)	Additional Identifier
pl	Individual questionnaire	*long*	pid, syear	hid, cid, intid
hl	Household questionnaire	*long*	hid, syear	cid, intid
biol	Biographical data	*long*	pid, syear	hid, cid, intid
jugendl	Youth questionnaire for first-time respondents at age 17	*long*	pid, syear	hid, cid, intid
plueckel	Follow-up questionnaire	*long*	pid, syear	hid, cid, intid
abroad[1]	Questionnaire for respondents who have moved abroad	*long*	pid, syear	hid, cid
vpl	Deceased individual	*long*	vpid, syear	hid, cid, intid

[1] *In addition to the classic identifiers (pid, hid, and cid), these datsets also have the identifiers from older data release versions. (pid=persnr; hid=hhnrakt; cid=hhnr).*

pl "Individual questionnaire" (long): The PL datset contains all waves of the $P datsets from SOEP-Core. In addition, the PL file includes all variables of all waves of the datsets $POST and $PAUSL. This means that the PL datset contains all variables from the individual questionnaire for all waves. In addition, the individual-specific data from the IAB-SOEP Migration Survey and IAB-BAMF-SOEP Refugee Survey are integrated into the PL datset.

hl "Household questionnaire" (long): HL contains all waves of the datsets $H from SOEP-Core. This means that the HL datset includes all questions of the household questionnaire. In addition, the household-specific data from the IAB-SOEP Migration Survey and IAB-BAMF-SOEP Refugee Survey are integrated into the original HL datset.

biol "Biographical data" (long): BIOL contains cumulated individual-level raw data from the biographical questionnaire and from wave-specific biographical modules of the individual questionnaire. BIOL is intended to be used in addition to the generated biographical files (by advanced users) to complete (or modify) generated biographical variables.

jugendl "Youth questionnaire for first-time respondents at age 17" (long): JUGENDL contains the waves q (2000) up to the current wave of $JUGEND in SOEP-Core. Since 2000 (wave Q), first-time respondents between the age of 16 and 17 have received a separate biographical questionnaire with additional age-group-specific questions, for instance, about their relationship to their parents or about what they do in their free time. Up to now, only some of the data collected from this survey have been processed and provided to users in dataset BIOAGE17. The complete data will be provided in individual JUGENDL dataset.

plueckel "Catch-up questionnaire" (long): The PLUECKEL datset contains all waves of the $PLUECKE datsets in SOEP-Core. Temporary drop-outs ("gaps") can cause problems for longitudinal analyses. This has especially negative consequences for the employment and income data. That is why the SOEP tries to fill in at least some of the key missing information. PLUECKEL is a small questionnaire covering information on the year previous to which the temporary drop-out occurred. It covers questions on job-related changes, employment calendar, income, education, and qualifications.

abroad "Questionnaire for respondents who have moved abroad" (CS): With the pilot study "Life outside Germany" in 2008, the longitudinal SOEP study ventured into completely uncharted methodological territory by attempting to locate the addresses of former SOEP respondents who have since moved abroad and to survey these individuals with the help of a specially developed written questionnaire on the reasons for their move. The project was discontinued due to insufficient case numbers in 2014.

vpl "Questionnaire on the deceased individual" (long): The VPL datset contains all waves of the $VP datsets of SOEP-Core. The VPL file contains information about respondents who lost a relative in the previous year. It provides information about the deceased individual and the respondent who reported the death.

Survey Data

These datsets contain information on survey methodologies used in SOEP-Core. The various datsets contain detailed exit information provided by respondents and the household weighting factors that users need for representative analysis.

Dataset	Label	Format	Identifier (ID)	Special Identifier
csamp	Sample definition	*long*	cid	
design	Survey design	*CS*	hhnr	intid
exit[1]	Cumulative drop-outs	*CS*	pid	cid, syear
pbr_hhch[1]	PBR_HHCH	*CS*	pid	hid, syear, cid, pnralt, pnrneu, hhnrold
cirdef	Randomized survey file	*long*	hhnr	

[1] *In addition to the classic identifiers (pid, hid and cid), these datsets also contain the identifiers from older data release versions. (pid=persnr; hid=hhnrakt; cid=hhnr).*

csamp "Sample definition" (long): The dataset CSAMP [SAMP] contains detailed sampling information for each of the original sample households at the case level [cid / hhnr].

design "Survey design" (CS): The dataset DESIGN provides information on the stratified sampling of the SOEP in the form of two variables. The variable STRAT identifies each of the discrete sampling groups described above. Altogether, the SOEP consists of 40 strata: one stratum in sample A, twenty-seven in sample B, one in sample C, three in sample D, one in sample E, two in sample F, four in sample G, and one in sample H. Each of these strata have unique inclusion probabilities. The variable design contains the inverse of this probability, i.e., the design weight.

exit "Follow-up study [Verbleibstudie]" (long): The dataset EXIT delivers the results from the follow-up study [Verbleibstudie] conducted by Kantar Public (formerly: TNS Infratest) in 2008/2009. This study has been used to identify reasons for (demographic) dropouts. Deceased individuals identified through the follow-up study are included in the corresponding variables in PPATH/L [todjahr, todinfo].

pbr_hhch "PBR_HHCH" (long): The dataset pbr_hhch is a subfile of pbrutto that was used from 1984 to 2009 to identify individuals from households that underwent split-offs in subsamples A-H.

cirdef "Randomized survey file" (long): This dataset includes randomized groups of original sample households [rgroup] for selection of representative shares across all subsamples with full representation of any cross-sectional and longitudinal information (variables) at all levels (case, households, individuals, spells) for the entire SOEP population across all waves.

Generated Data

The SOEP team has prepared these datsets for easy use and subjects them to additional plausibility checks and quality controls prior to data release. In most cases, they consist of several variables and different survey instruments and are described in the documentation provided. As a result, these datsets cannot be assigned 1:1 to a single survey instrument.

Dataset	Label	Format	Identifier (ID)	Additional Identifier
pgen	Generated individual data	*long*	pid, syear	hid, cid, pgpartnr
hgen	Generated household data	*long*	hid, syear	cid
bioage17[1]	Generated biographical youth information	*CS*	pid	hid, syear, cid, bymnr, byvnr, intid
bioagel[1]	Generated biographical information	*long*	pid, syear, persnre	hid, cid,
biopupil[1]	Generated biographical information	*long*	pid, syear	hid, cid
kidlong[1]	Data on children	*long*	pid, syear	hid, cid
pequiv	Cross National Equivalent File	*long*	pid, syear	hid, cid
biobirth[1]	Generated biographical information	*CS*	pid	cid, kidpnr01-kidpnr15
bioedu[1]	Generated biographical information	*CS*	pid	cid
bioimmig[1]	Generated biographical information	*long*	pid, syear	hid, cid
biojob[1]	Generated biographical information	*CS*	pid	cid
bioparen[1]	Generated biographical information	*CS*	pid	cid, fnr, mnr
bioresid[1]	Generated biographical information	*CS*	pid	hid, syear, cid, intid
biosib[1]	Generated biographical information	*CS*	pid	cid, sibpnr1-sibpnr11
biosoc[1]	Generated biographical information	*CS*	pid	hid, syear, cid, intid
biotwin[1]	Generated biographical information	*CS*	pid	cid, pnrtwin, pnrtrip, pnrquad
camces[1]	Highest educational qualification, migrants sample M1 and M2	*CS*	pid	hid, syear, cid
cogdj[1]	Data on cognitive tests (Youth)	*CS*	pid	syear, cid
cognit[1]	Data on cognitive potential	*CS*	pid	syear, cid, intid
cog_refu[1]	Data on cognitive tests (Refugees)	*CS*	pid	syear, cid, hid
gripstr[1]	Grip Strength Measures	*CS*	pid	syear, cid, intid
hconsum[1]	Household Consumption Module	*CS*	hid	syear, cid

Continued on next page

Table 2 – continued from previous page

Dataset	Label	Format	Identifier (ID)	Additional Identifier
health[1]	Data on health indicators	*CS*	pid	syear, cid
hwealth	Wealth module	*long*	hid, syear	cid
interviewer	Data on the SOEP interviewer	*long*	intid, syear	cid
mihinc	Multiple imputed data on monthly household income	*long*	hid, syear	cid
pflege	Persons needing care within the household	*long*	pid, syear	cid
pkal	Individual calendar	*long*	pid, syear	hid, cid
pwealth	Wealth module	*long*	pid, syear	hid
timepref[1]	Experiment on time preferences	*CS*	pid	hid, syear, cid
trust	Experiment on trust	*long*	pid	hid, syear, cid

[1] *In addition to the classic identifiers (pid, hid and cid), these datsets also have the identifiers of the older data release versions. (pid=persnr; hid=hhnrakt; cid=hhnr).*

pgen "Generated individual data" (long): PGEN contains all waves of the $PGEN datsets in SOEP-Core. The PGEN-file contains user-friendly data on the individual level that are consolidated from different sources. The plausibility is validated longitudinally in many respects, making the data superior to those in PL in most situations. The file contains one row for each individual (pid is unique) with a completed individual or youth questionnaire.

hgen "Generated household data" (long): HGEN contains all waves of the $HGEN datsets in SOEP-Core. In order to minimize computational effort for the user, the SOEP provides yearly status variables on the household level. The HGEN data provide a set of time-invariant variables generated from the SOEP household questionnaire. They only include households that participated in the respective year.

bioage17 "Generated biographical information" (CS): The design of the dataset BIOAGE17 is patterned after the 2001 Youth Questionnaire, which is the standard version used in subsequent years. Young people living in a panel household who reached the survey age of 17 are a special group of first-time respondents. This group of panel entrants provides more detailed information on youth and socialisation than we are able to obtain from other new sample members.

bioagel "Generated biographical information" (long): The BIOAGEL data files are generated using information collected in the "Mother & Child" and "Parent" questionnaires. BIOAGEL is now provided in one dataset.

biopupil "Generated biographical information" (long): The BIOPUPIL data files are generated using information collected in the "Pre-Teen" and "Early-Youth" questionnaires. BIOPUPIL is provided in one dataset.

kidlong "Data on children" (long): The variables stored in the KIDLONG file are based on the information collected annually and contained in the wave-specific $KIND files. The relevant information is not provided by children themselves but is obtained from answers to questions in the household questionnaire provided by the respondent within the household (usually the head of the household). This data is reaggregated at the individual level and stored as child-specific entries in the file $KIND.

pequiv "Cross-National Equivalent File" (long): PEQUIV contains all waves of the $PEQUIV datsets in SOEP-Core. The PEQUV-File is based on the Cross-National Equivalent File (CNEF) with extended income information for the SOEP. This file comprises not only the aggregated income figures from CNEF but also additional separate income components.

pkal "Individual calendar" (long): PKAL contains all waves of the $PKAL datsets in SOEP-Core. The PKAL datasets contain calendar variables from the individual questionnaire. The datset includes the individual's employment or educational status on a monthly basis as well as the individual' income status.

biobirth "Generated biographical information" (CS): The file BIOBIRTH provides information on fertility histories of adult respondents in the SOEP. Up to 2014 (version 30, wave BD), the data were stored in two separate files: BIOBIRTH containing female fertility histories, and BIOBRTHM providing male fertility histories. Fertility histories in BIOBIRTH provide information on every woman (as well as every man with panel entry since 2001) who has ever completed at least one SOEP interview.

bioedu "Generated biographical information" (CS): The SOEP contains a broad range of variables on early childhood education and care, educational participation, educational degrees, and related topics. The BIOEDU dataset is designed to provide ready-made variables on educational transitions and related topics for use in longitudinal analysis.

bioimmig "Generated biographical information" (long): The variables contained in BIOIMMIG relate to foreigners in (and migrants to) Germany. Questions deal with the 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.

biojob "Generated biographical information" (CS): The purpose of BIOJOB is to provide a file that offers the user convenient access to biographical information on past job activities. BIOJOB consists of generated variables as well as plain questionnaire information. Up to now, all but two variables in 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 respondent's biography interview.

bioparen "Generated biographical information" (CS): The dataset BIOPAREN contains biographical entries on the parents' and respondent's background. The information in BIOPAREN is obtained from two sources: from proxy entries by children on their parents in the biography questionnaire and youth questionnaire, and from direct entries by parents when the respondent lives in the same household as the parents. Please note that BIOPAREN focuses on the social parent. Biological parent identifiers can be found in BIOBIRTH.

bioresid "Generated biographical information" (CS): In 1994, questions with a focus on occupancy were introduced into the biographical questionnaire asking about the duration of residence in the current dwelling and any second residence. The information obtained from the biographical questionnaire is contained in the file BIORESID.

biosib "Generated biographical information" (CS): BIOSIB provides information on siblings living within SOEP households. The datset contains the individual identifiers of all siblings in a SOEP household. It includes information on the individual sibling's sex, year of birth, number of siblings, position in birth order, and relationship between siblings.

biosoc "Generated biographical information" (CS): Contains data on youth and socialization. Respondents of all ages describe aspects of their life at the age of 15, including their relationship with parents, grades in school, the federal state where they last attained educational qualifications, detailed information on vocational qualifications, as well as intentions to complete further education or vocational training. Questions concerning military and alternative services are also included in this datset.

biotwin "Generated biographical information" (CS): The file BIOTWIN contains all twins that were ever identified within the SOEP. To be classified as a twin, a individual is required to have exactly the same age as his or her sibling (year & month of birth), have a relationship to the head of the household that indicates that he or her and a second individuals are siblings, and have the same mother (as far as a pointer to the mother is available). Furthermore, it is not only twins that are recorded in the BIOTWIN datset, but also triplets or quadruple siblings.

camces "Highest educational qualification, migrant samples M1 and M2" (CS): The CAMCES-File provides information about computer-assisted measurement and coding of educational qualifications in surveys.

cogdj "Data from cognitive tests (Youth)" (CS): In SOEP 2006, a separate questionnaire with cognitive tests for adolescents was used for the first time: "Lust auf DJ". The acronym "DJ" stands for "Denksport und Jugend" (mind sports and youth)", but it was named for its more common association with "disc jockey". The questionnaire "Lust auf DJ" was created for all respondents aged 16-17.

cognit "Data on cognitive potential" (long): In the 2006 survey year, for the first time, short cognitive tests were carried out with a subsample of the SOEP. The goal was to employ a robust set of instruments that could be administered easily by trained interviewers within just a few minutes. COGNIT06 provides the aggregated sum scores (total values for three time packages, so-called "parcels" of 30, 60 and 90 seconds).

cog_refu "Data on cognitive tests (refugees)" (CS): The dataset contains sum scores for two competence measurements (previous school knowledge and basic cognitive skills) of youths born in 2000, 2003 and 2005 surveyed in 2017.

gripstr "Measures of grip strength (left and right hand)" (long): The data on grip strength from the survey year 2012 is now included in the GRIPSTR dataset.

hconsum "HH consumption module" (CS)": We were faced with three methodological challenges in generating the final consumption data. First, due to the design of the consumption module, inconsistent answers arose between the amounts give for monthly and annual consumption. Second, there was the common problem of missing data, here in particular item nonresponse. And third, consumption data are usually blurred by heaping. For researchers who do not want their consumption variables to include changes from all steps of data preparation, the new datset "HCONSUM" contains not only the prepared consumption variables but also flag variables providing researchers the opportunity to select individual solutions.

health "Data on health indicators" (long): Starting in 2002, the SOEP health module in the individual questionnaire has been revised and replicated at two-year intervals. In the HEALTH file, users find, for instance, the generated variables on height and weight with imputation flags and a user-friendly longitudinal checked generated variable for Body Mass Index (BMI).

hwealth "Wealth module" (long): The generated SOEP wealth data is stored in two separate data files called PWEALTH for information at the individual level and HWEALTH for correspondingly aggregated data at the household level. HWEALTH contains all information on the household level; it is purely the result of aggregating the individual-level information in PWEALTH. However, for all individuals with valid household-level information who did not respond to the individual questionnaire (partial unit non-response), imputations have been carried out and the results are included in HWEALTH.

interviewer "Data on the SOEP interviewer" (long): The SOEP aims not only to collect high-quality data on the living conditions and well-being of households, but also to provide a valuable empirical source for survey research. The INTERVIEWER file provides users with easy access to all available longitudinal information on the SOEP interviewers.

mihinc "Multiple imputed data on monthly household income" (long): The dataset MIHINC contains the complete imputation results and is available separately. To be compatible with methods for analyzing multiply imputed data, MIHINC is constructed in the "stacked" or MIM data format. It contains the following variables: HHNRAKT, SVYYEAR, MJ, MI, IHINC and IMPFLAG. Since 1995 for every survey household in all survey years, there are ten imputed values for current household income.

pflege "Persons needing care within the household" (long): Since wave B (1985), the SOEP household questionnaire includes questions on household members in need of care. In order to support individual-level analysis, this information has been restructured and is stored in the cumulative file PFLEGE.

pwealth "Wealth module" (long): For the first time in 2002, the individual questionnaire included a special module focusing on wealth. It included questions on seven different wealth components: owner-occupied property (including debt), other property (including debt), financial assets, private pensions (including life insurance and building savings contracts), business assets, tangible assets, and consumer credit. The generated SOEP wealth data are stored in two separate data files called PWEALTH for information at the individual level and HWEALTH for correspondingly aggregated data at the household level. Wealth-related variable names in the file PWEALTH consist of six digits. The first digit tells the user which wealth component is referred to, and the second to sixth digits provide more detailed information about possible filter information, the personal share, the gross amount, and the amount of any outstanding debt. In principle, a digit is coded "1" if a given variable does indeed contain this specific piece of information and "0" otherwise. The wealth information in the SOEP questionnaire is surveyed at the individual level and thus also imputed or edited at the individual level (although checked against household information for consistency).

timepref "Experiment on time preferences" (CS): Following the behavioral experiment on trust and trustworthiness carried out in the 2003, 2004, and 2005 SOEP surveys, the experiment "time preferences" was run in 2006. In this experiment on economic behavior, respondents were asked to decide how they would want to receive €200 in prize money: if they would want to receive it immediately by check or if they would want to wait and receive a larger amount later, that is, with interest.

trust "Experiment on trust" (long): The economic behavior experiment on trust and trustworthiness from survey years

2003, 2004, and 2005 served to measure trust based on an investment game, a one-off game for two players who interact anonymously. The first player receives a credit of ten points and can overwrite any number of points of the second player. Each overwritten point is doubled. The second player also receives a credit of ten points. After receiving the (doubled) points from the first player, the second player decides how much of her own credit she will transfer to the first player (zero to ten points). As with the first transfer, the recipient's points are doubled. After the decision of the second player, the game ends and the other players are paid (one point corresponds to one euro, the total is paid by check a few days later). The trust datset thus contains the information from all three waves in which the behavioral experiment was conducted.

Spell Data

Spell, duration, and event history data are used frequently in the social sciences. In the strict sense of the word, spell data are about time periods with a defined start and end. General information about the data structure of spell data can be found in the chapter *Data Structure in Spell Format (spell)*

Working with spell data:

Working with spell data (pdf):

Working with spell data (do-files):

How to generate spell data from data in wide format: Based on the migration biographies in the IAB-SOEP Migration Sample:

Generating spell data:

Dataset	Label	Format	Identifier (ID)	Additional Identifier
artkalen	Spell data from the activity calendar	*spell*	pid	cid
biocouplm	Generated biographical information	*spell*	pid	cid, coupid
biocouply	Generated biographical information	*spell*	pid	cid
biomarsm	Generated biographical information	*spell*	pid	cid
biomarsy	Generated biographical information	*spell*	pid	cid
einkalen	[deprecated] Spell data on income	*spell*	pid	cid
lifespell	Spell Information on the pre- and post-survey history of SOEP respondents	*spell*	pid	cid
migspell	Migration history	*spell*	pid	cid
pbiospe	Generated biographical information	*spell*	pid	cid
refugspell	Migration history	*spell*	pid	cid
sozkalen	[deprecated] Spell data on social benefits	*spell*	hid, cid	

artkalen "Spell data from the activity calendar" (long): 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. The information on activity status is collected on a monthly basis in the yearly individual questionnaire and stored in the file ARTKALEN.

biocouplm "Generated biographical information" (long): With the BIOCOUPLM the SOEP provides consistent and continuous partnership histories for nearly all adult respondents. BIOCOUPLM is built on the prospective information at the time of each interview. The relationsship histories are collected on a monthly basis from all adult SOEP participants since their entry into the SOEP.

biocouply "Generated biographical information" (long): With the BIOCOUPLY, the SOEP provides consistent and continuous partnership histories for nearly all adult respondents. BIOCOUPLY is built on retrospective and prospective information at the time of each interview. The relationship histories are provided on an annual basis.

biomarsm "Generated biographical information" (long): With BIOMARSM the SOEP provides consistent and continuous marital histories for nearly all adult respondents. BIOMARSM is built on the prospective information at the time of each interview. The martial histories are collected on a monthly basis from all adult SOEP participants since their entry into the SOEP.

biomarsy "Generated biographical information" (long): With BIOMARSY the SOEP provides consistent and continuous marital histories for nearly all adult respondents. BIOMARSY is built on retrospective and prospective information at the time of each interview. The marital histories are provided on an annual basis.

einkalen "[deprecated] Spell data on income" (long) The income calendar is used to gain information about sources of income throughout the year. The respondent checks off for each month all appropriate sources of income.

lifespell "Spell information on the pre- and post-survey history of SOEP respondents" The SOEP team regularly conducts follow-up studies to relocate attritors. These studies draw on official register data and allow us to determine whether a individual is still living in Germany, is deceased, or has moved abroad since the last SOEP interview. The information is combined in a spell file LIFESPELL. This dataset reports all available information on the pre- and post-survey history of all individuals who have ever been a member of a SOEP household.

migspell "Migration history" (long): MIGSPELL is derived from the migration biographies, which are collected from each new respondent of the IAB-SOEP migration samples M1 and M2. It contains data on moves by foreign-born migrants as well as on stays abroad by German-born respondents.

pbiospe "Generated biographical information" (long): The spell file PBIOSPE is based on the information on activity status over the life course, which is collected as a matrix from every respondent who completes the biographical questionnaire. The observations start at the age of 15 and end at the current age (up to age 65). To update ongoing employment information in PBIOSPE, information from the yearly individual questionnaire is also used.

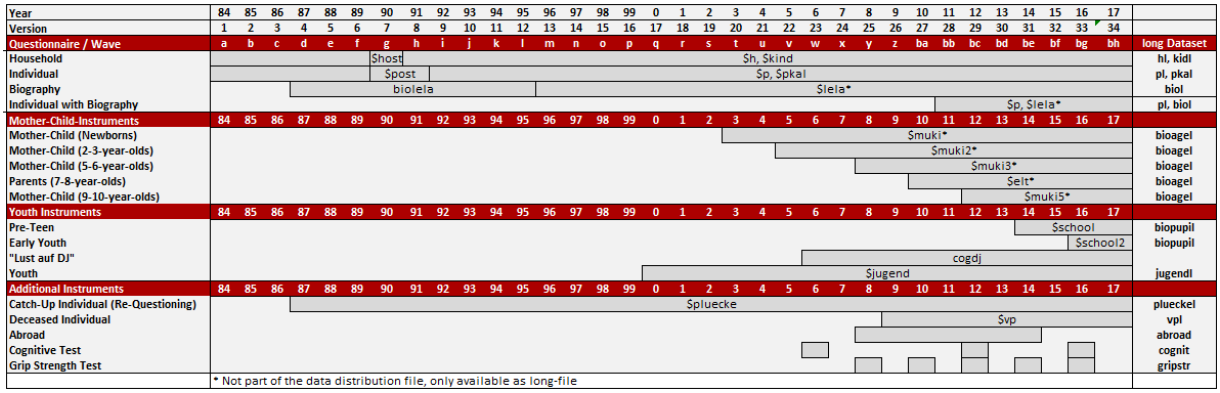
refugspell "Migration history" (long): For migration biographies in the refugee samples, we created the spell datset REFUGSPELL. The variables in MIGSPELL and REFUGSPELL are derived from different instruments and only partially overlap. The data structure allows the datset to be linked with MIGSPELL if desired.

`1992-2000` sozkalen "[deprecated] Spell data on social benefits": The file SOZKALEN provides spell data on households receiving social assistance, defining the beginning, end, and censoring status of any period of receiving 3 different types of assistance. This file is set up using information from the calendar that is collected for the previous year (between 1992-2000). Thus, it contains information on a monthly basis.

Last change: Nov 12, 2019

5.4 Data Processing

The following overview shows which datasets form the basis of each questionnaire, and the respective data processing process, from the questionnaire to the wave-specific datasets, to the prepared "long" datasets. Please note that not all datasets are based on questionnaires but that many have been prepared meticulously by our staff. The table therefore does not show the full range of datasets available.



In addition to the classic SOEP survey instruments, there are also a large number of sample-specific questionnaires whose information flows into other unlisted raw datasets (e.g. $pausl, $post, $pkalost etc.). The chapter *Sample-Specific Questionnaires* explains why such special survey instruments exist, how they become raw datasets and in which long datasets these variables can be found.

Last change: Nov 12, 2019

5.5 Dataset Identifiers

Because of the overall data structure with data on different observational levels, any analysis requires the combination of data using matching or merging procedures. These merging procedures need identifiers such that a combination of datasets becomes feasible. The central individual identifier across time is **pid**, which is fixed over time (and of course datasets). Since a person might change the household in which he or she lives at any point in time, yearly household identifiers called **hid** are necessary. The same information is also stored in **$hhnr**, facilitating matching depending on the dataset used. Finally, each individual (respondents as well as children) can be traced back to be a member of or a split-off from an original household from the very first wave. This household's ID, which is fixed no matter how often a person changes households over the course of time, is called **cid**. In addition, respondents in long data can be differentiated by survey year. The **syear** variable can be used to identify a respondent's survey year. The SOEP provides additional identifiers in the various datasets in order to identify respondents and to provide further possibilities for merging datasets. A excerpt of these additional identifiers can be found here:

Please note that these are not all identifier variables. The name of the identifier variable can change depending on the datset used.

- **parid** "Unchanging Individual identifier of Partner (PID)"
- **pgpartnr** "Individual Identifier of Partner"
- **coupid** "Couple Identifier"
- **intid** "Interviewer Identifier"
- **intid1** "Identifier of First Interviewer"
- **$hhnr** "Current Wave HH Number (=HHNRAKT/HID)"
- **hhnrold** "HH Number Previous Year With Individual Identifier"
- **vpersnr** "Individual Identifier of Deceased Indivdiual"
- **bymnr** "Individual Identifier Mother"
- **byvnr** "Individual Identifier Father"
- **mnr** "Individual Identifier Mother"

- **fnr** "Individual Identifier Father"
- **kidpnr01-kidpnr15** "PERSNR 1st Child" - "PERSNR 15th Child"
- **sibpnr1-sibpnr11** "Individual Identifier, 1st sibling" - "Individual Identifier, 11th sibling"
- **persnre** "Permanent Indivdiual Identifier Respondent" (usually mother)
- **pnrtwin** "Individual Identifier 2nd Sibling"
- **pnrtrip** "Individual Identifier 3rd Sibling"
- **pnrquad** "Individual Identifier 4th Sibling"
- **pnralt** "Old Household and Individual Identifier"
- **pnrneu** "New Household and Individual Identifier"

Last change: Nov 12, 2019

5.6 Versioning and Harmonization

In some cases, variables in long format with the same content but collected in different ways need to be harmonized to ensure that they remain consistent and comparable over time. Starting with SOEP Core v.34, SOEP offers versioning and harmonization solutions for such variables in all *Original Data* in long format. These versions and harmonizations are recognizable in the variable name. The "_v" suffix indicates possible differences in a variable. Harmonization suggestions generated by SOEP from the different versions of these variables can be recognized with the "_h" suffix. In general, particular caution is required when using variables marked "_v" or "_h":

1.) Differences in Response Options

Variables are versioned and harmonized because the response options have changed over time.

2.) Differences in Coding of Response Options

Variables are versioned and harmonized because the coding of the response options has changed over time. Since the values of certain response options can change, it is not possible to easily integrate the various wave-specific variables into a variable in long format. The variable must be appropriately harmonized to be useable.

3.) Content Differences in the Questions.

Variables are versioned and harmonized because the questions were asked differently in different years, but the content belongs together. If the content or wording of the question changes, the wave-specific variables cannot easily be integrated into a long variable.

4.) Changes in Question Type.

Variables are versioned and harmonized because the questions were asked differently in different years, for example, as a question with multiple response options and later as a question with a single response option. A possible multiple answer in certain years makes it difficult to easily integrate the wave-specific variables into a variable in long format.

5.) Euro Harmonization

Variables are versioned and harmonized because they are metric and were surveyed as DM amounts before the introduction of the euro. For the long version of the variable, metric variables based on different currencies in different years are harmonized as euro amounts.

6.) Differences in Metric Variables

Variables are versioned and harmonized if they contain a year and were provided in the wave-specific raw data with different numbers of digits. The years are standardized and presented in the harmonized version with four digits. In addition, possible problems with decimal digits in metric variables from the raw datsets are corrected for the long-format variable.

7.) Different Respondents

Variables are versioned and harmonized when different groups of respondents have received different survey instruments and the variables have not been integrated into the wave-specific raw datasets. Special samples or a specific filtering in the questionnaire can lead to certain groups of people receiving different questions that belong together in terms of content. Such different variables are harmonized in the long version of the variable.

A more detailed explanation of the versioning and harmonization concept can be found in the section *Working with harmonized Variables*

Last change: Nov 12, 2019

5.7 Missing Conventions

Survey variables might be missing, that is, lacking a valid code or value, for different reasons. In the SOEP, negative values are not valid for any variable, but are used instead to code different reasons for missing information. There are two possible origins of missing values: the respondent's answer or the survey design. In the first case, the respondent may refuse to answer or not know an answer or may report invalid values. In the second case, the interview design may exclude respondents with certain characteristics from some questions (e.g., men will never be asked if they are pregnant). The following codes are used:

Code	Label
-1	No answer / don't know
-2	Does not apply
-3	Implausible value
-4	Inadmissable multiple response
-5	Not included in this version of the questionnaire
-6	Version of questionnaire with modified filtering
-8	Question not part of the survey program this year[1]

[1] *Only applicable to datasets in long format.*

A person might decline to answer a question. This occurs mainly with sensitive questions (e.g., income-related questions) and when respondents simply do not know the answer. In such cases, the missing code is "-1" for "no answer / don't know". Note that the SOEP does not distinguish between a refusal to answer and a true "don't know". Information may be missing when a question is not asked because it is not relevant to a specific person, e.g., owner-occupiers will not be asked about the amount of rent they pay. In such cases, the question "does not apply" to this person, and the variable receives a code of "-2". Sometimes invalid answers occur when respondents fill out a PAPI interview themselves or the interviewer mistypes an answer (e.g., working hours over 168 per week). In such cases, multiple checks are carried out, and if the inconsistency remains, the variable is recoded "-3 Implausible value". Some questions contain multiple answer possibilities and respondents are asked to pick one answer. In the SOEP PAPI questionnaires, respondents sometimes ignore this request and give more than one answer (e.g., "very good" and "good" when asked about their current health status). In such cases, if the correct answer cannot be determined from the questionnaire itself, the code "-4 Invalid Multiple Answers" is assigned to this variable. With the extensions to the SOEP in recent years, entirely new samples have been added to SOEP-core. In these samples, questions are sometimes left out completely, e.g., to shorten the questionnaire or because the focus of the sample is different (as is the case with SOEP-related studies). In such cases, the variable will be set to "-5 Not included in this version of the questionnaire" for an entire subsample. With the use of CAPI, recent developments include an "integrated" individual questionnaire, i.e., the biography part and the "regular" part of the questionnaire are combined into one questionnaire. Some of the questions in the biography part are repeated in the regular part. Whereas the respondent will answer the same question twice the PAPI mode, the CAPI allows the respondent to filter around the question if it has already been asked. These cases are very rare, but if they occur, they receive a code "-6 Version of questionnaire with modified filtering".

Last change: Nov 12, 2019

CHAPTER

SIX

WORKING WITH SOEP DATA

The following exercises are taken from our SOEP Campus Workshops, a service especially for young scholars in the disciplines of sociology, economics, and psychology. Here we provide introductions to the use of the SOEP data.

6.1 Syntax Generator on paneldata.org

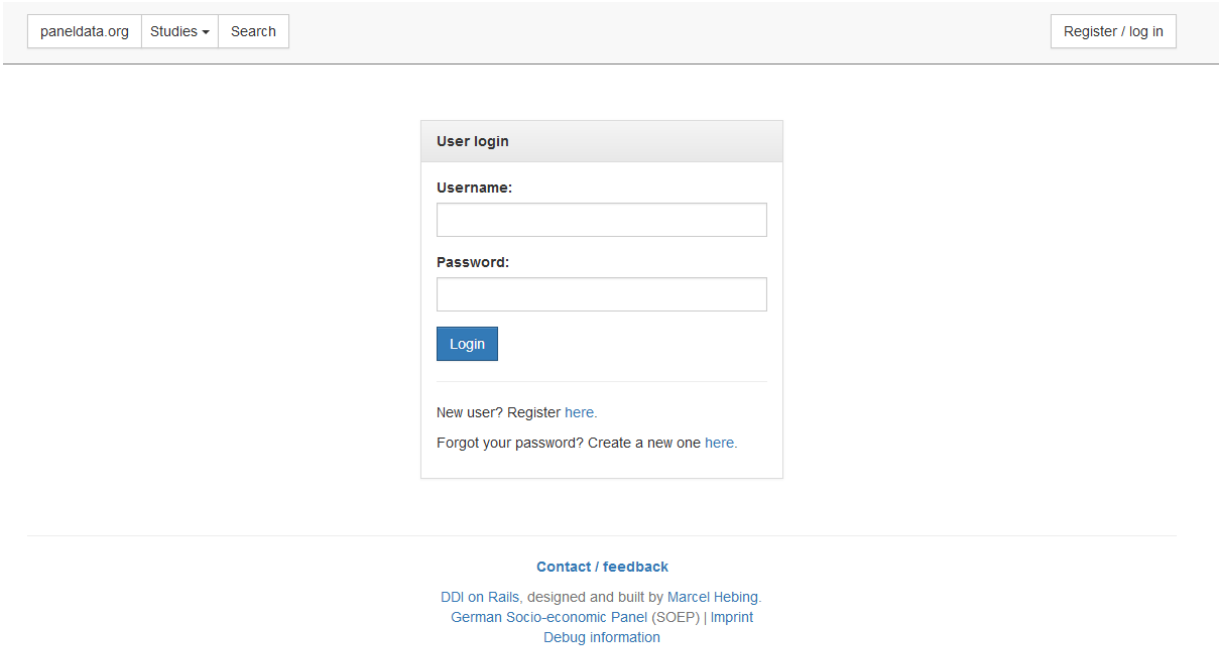
Paneldata allows registered users to collect and save variables relevant to their research in a variable basket. These variables can be simply written into a single dataset with the script generator. The script generator helps you with data management and can save valuable working time.

Open Paneldata

For our experienced users, we have temporarily equipped the old soepinfo with the current data so that the variable basket function and the script generator can also be used there.

Open soepinfo

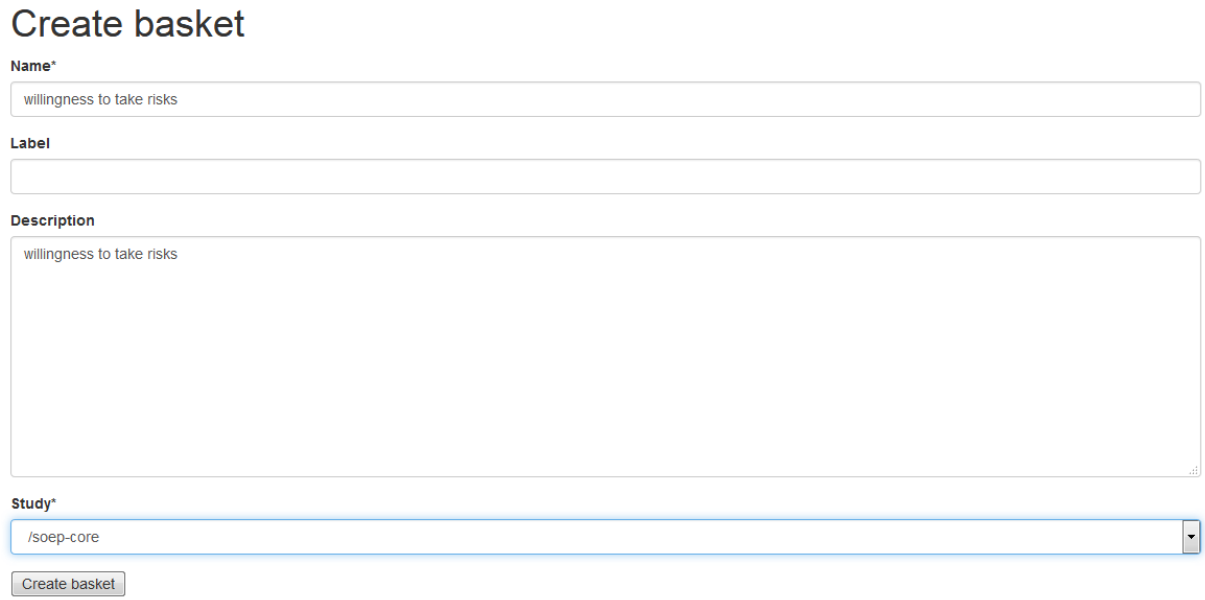
Click on “Register / login” to log in to paneldata.org.



If you have already registered, go to “user login”. As a new user, you can register under “register here”. Once you have logged in, you have access to the variable basket and the syntax generator.

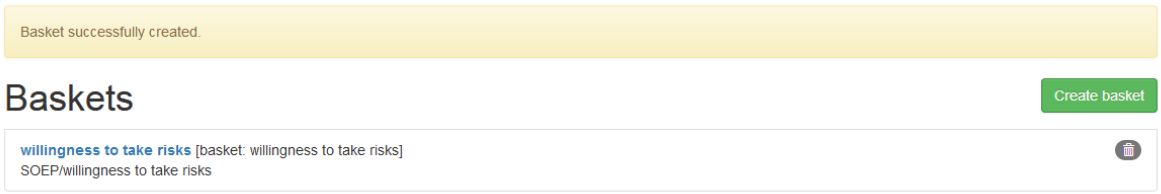
To access the activated functions, click on the button “my baskets”. You will be taken to your personal workspace on paneldata.org.

“My baskets” displays your variable baskets. If you click on “create basket”, you can create a new basket.

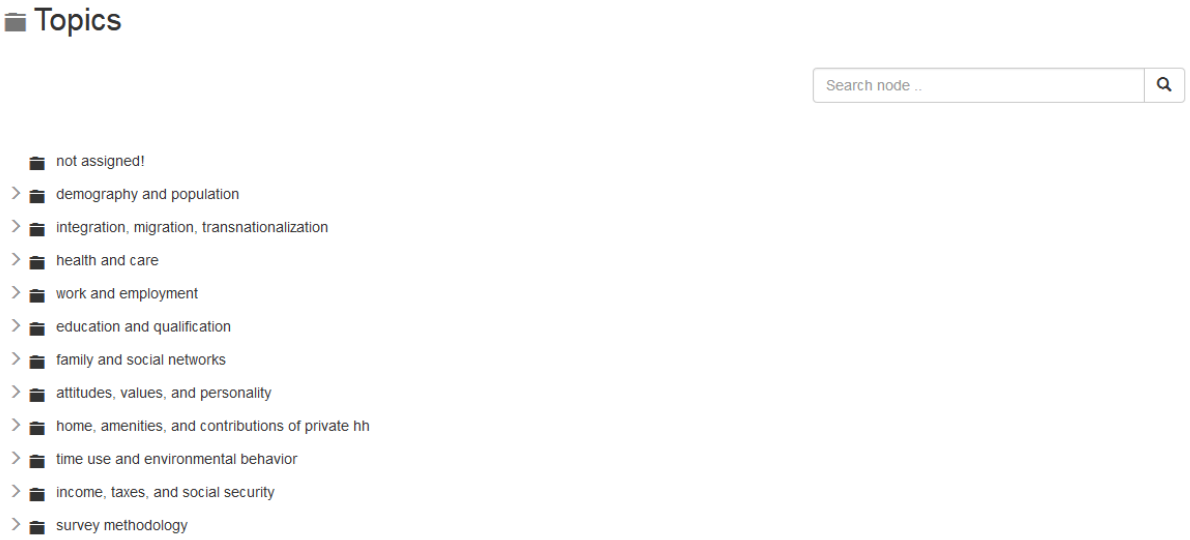


When creating the basket, first define the name of the variable basket. The name must be lower case to be accepted by

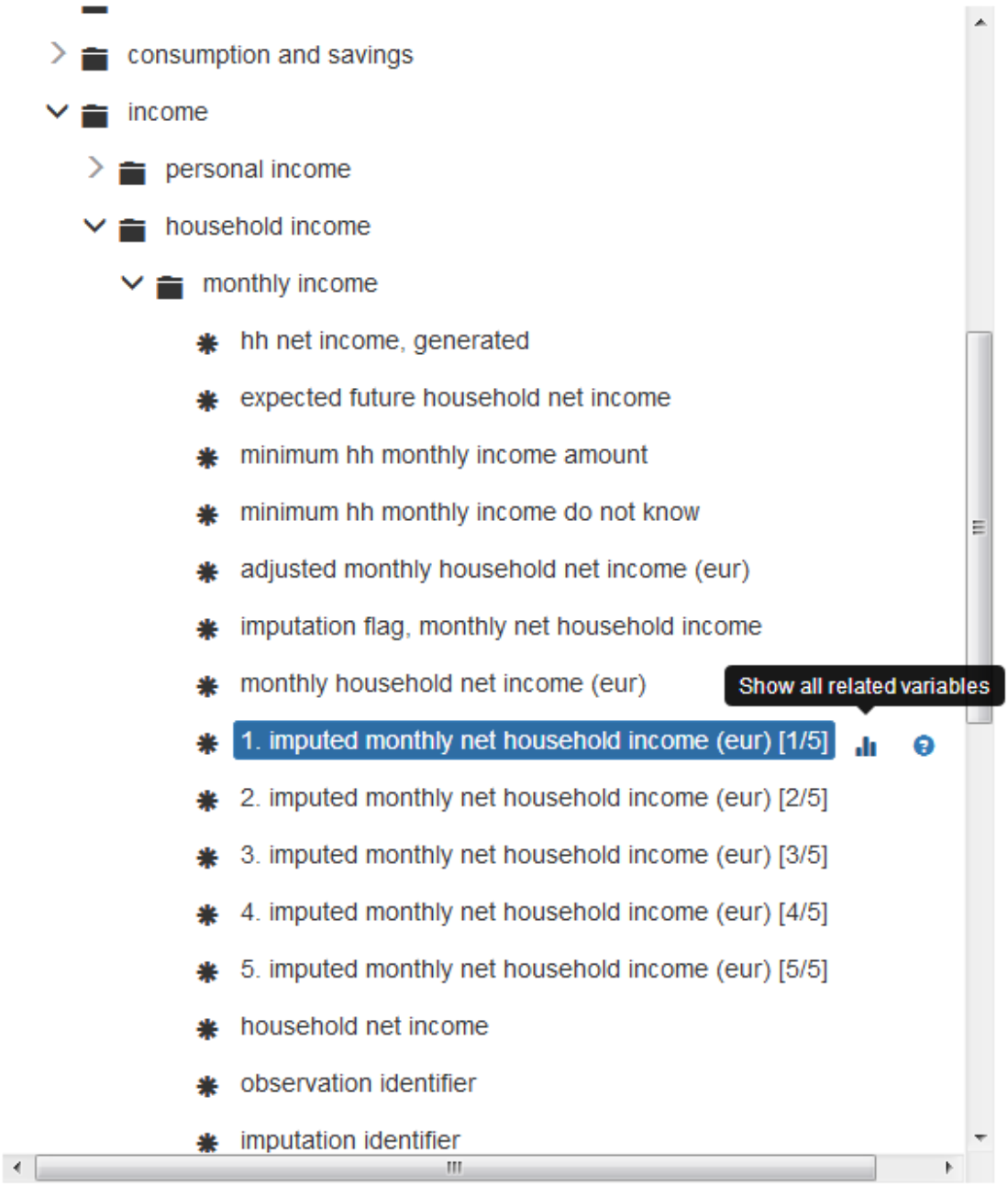
Paneldata. Optionally, you can assign a label and enter a description. Finally, you select the study that you want to use as a database for your research. Now click on “Create basket” and your newly created variable basket appears in the interface.



Now search for the relevant variables on paneldata.org and add them to your individual basket. For example, you are interested in monthly net household income. If you do not know the variable name, you can find the overarching concept using the topic search. Click on “paneldata.org” to get to the main page. Select the study SOEP-Core and click on “topics” at the top of the page.

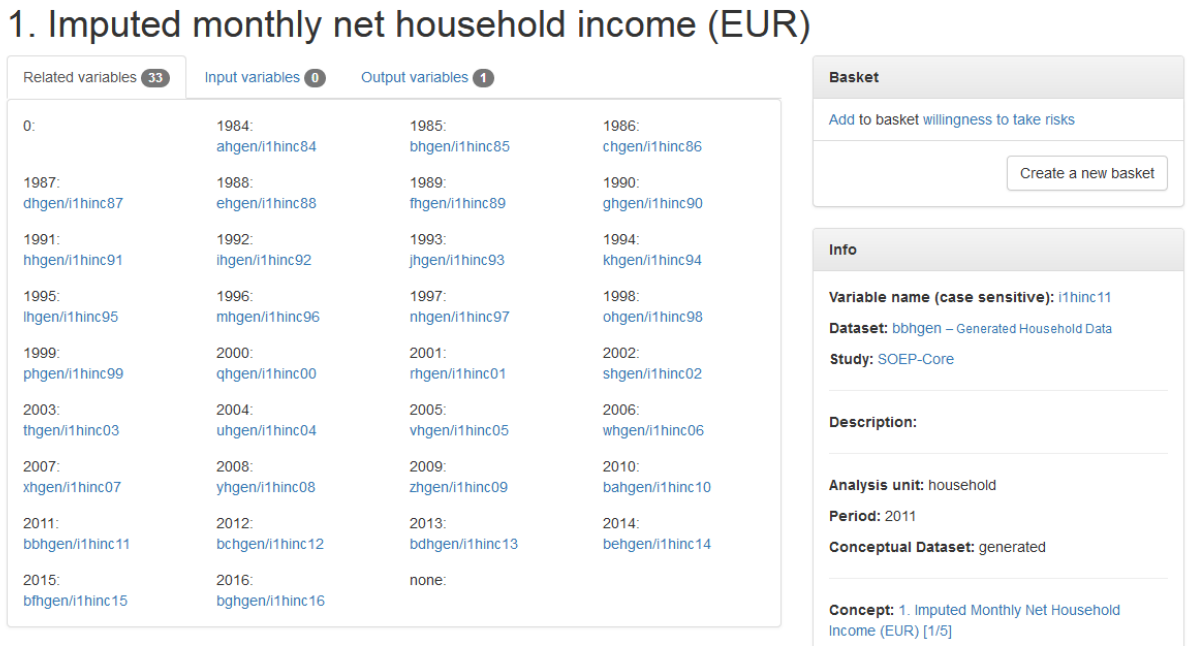


Check the different topics for income-relevant concepts and select “income, taxes, and social security”.

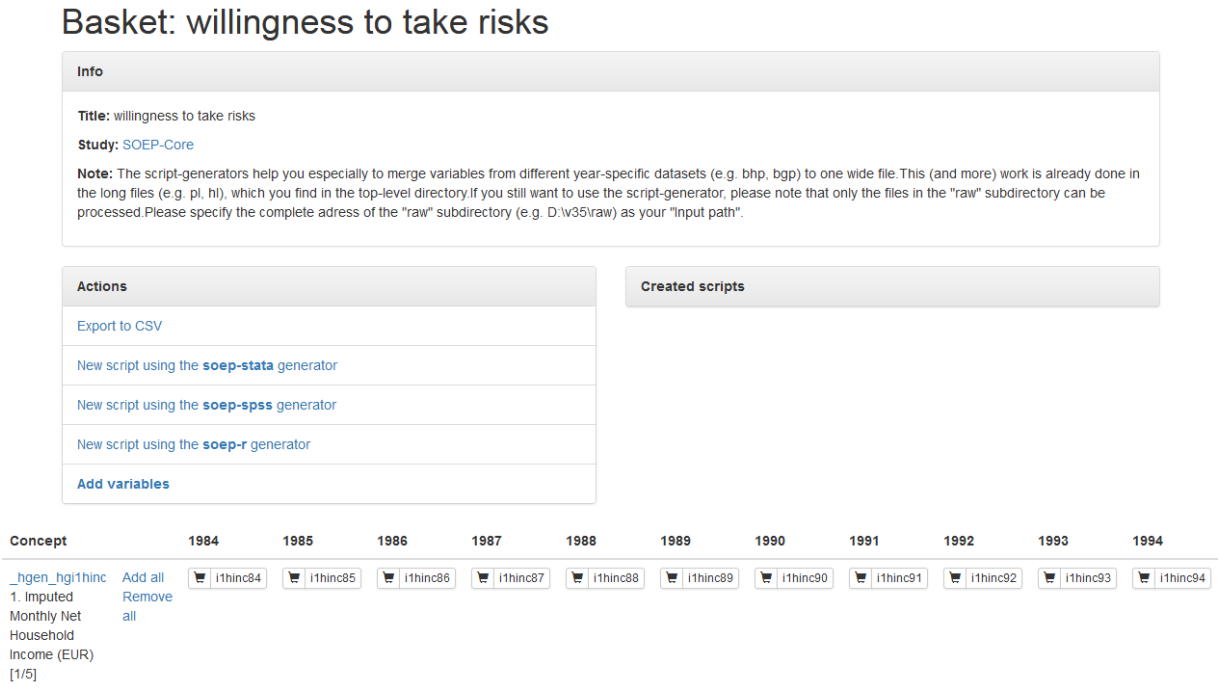


Browse the topic list and you will reach the sub-topic "income" –> "household income" –> "monthly income". There you will find the variables you are looking for. Click on "show all related variables" and you will see the history of variables.

Select the variable of your desired study SOEP-Core and you will reach the variable overview with important information about the variable. In the variable overview, you should make sure that the variable also meets your requirements.

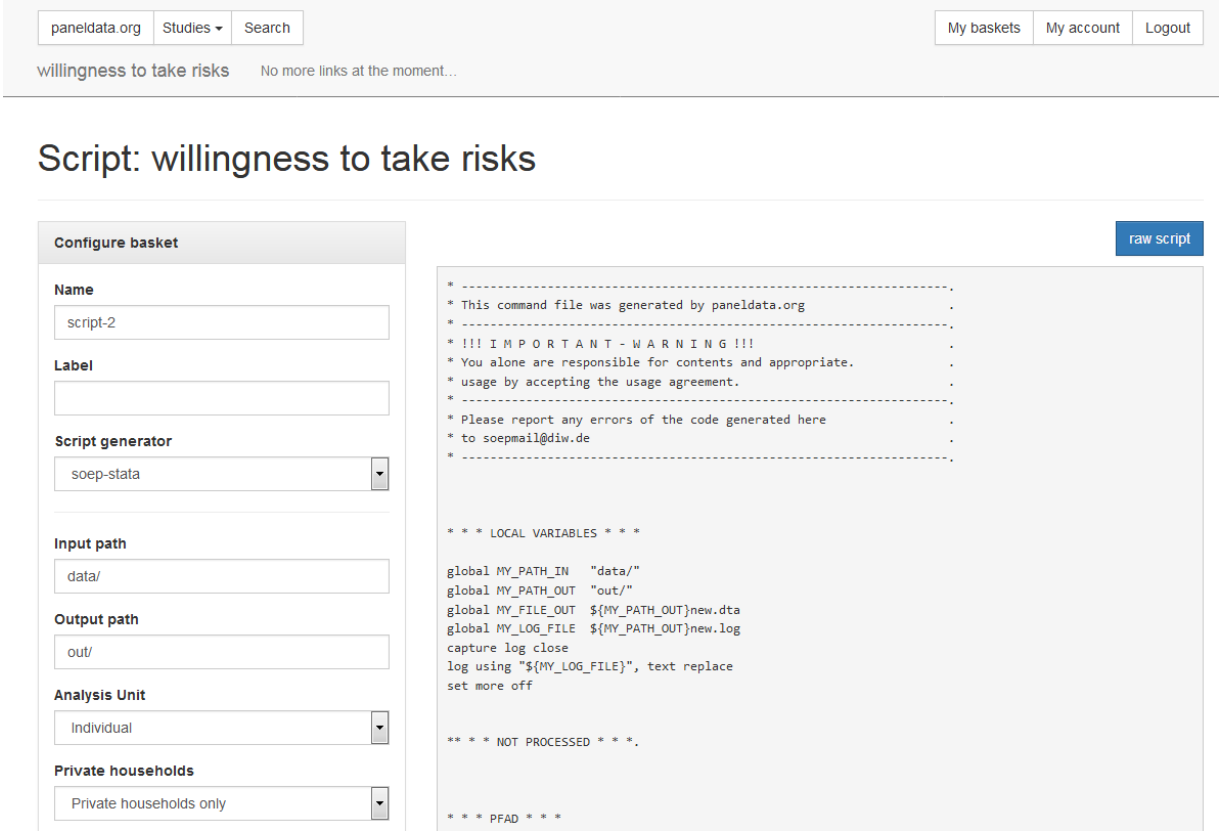


When logged in, the basket area appears in the overview of variables. Your baskets are listed there. If you want to add the variable to a basket, click on “add to basket”. If the variable is already in the basket and you want to remove it, select “remove from basket”. If you want to create a new basket within the overview of variables, click on “create a new basket” and your variable will automatically be placed in the new basket. You can access the basket overview by clicking on the name of your basket in the “basket” section. Alternatively, you can click on the button “my baskets” and you will also return to the basket overview.

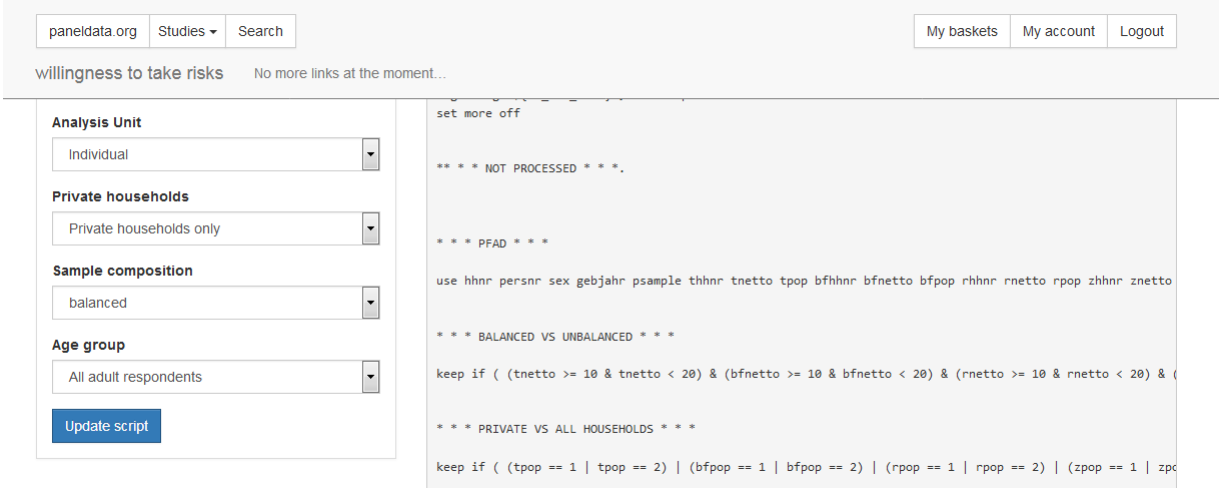


Click on the basket with your added variable and you will get an overview of all variables in your basket. With “add all”, you add the variables of all survey waves and the shopping cart is highlighted in green. If you are interested in a specific survey period, you can select the wave-specific variables by clicking on the shopping cart. Click on “remove all” to remove the variable from your basket. Furthermore, you can export your chosen variables to a CSV-File (Comma Seperated Value-File) and, for example, import them in STATA.

Once you have filled your basket and selected the desired survey waves, you can merge all variables into one dataset. To do this, click on “new script using the soep-xxx generator” in the “actions” area. You can choose between different statistical programs.



In the script generator, you can create a script that matches your preferred variables. Specify the name of your script, select the statistics program you are using. Enter the (local) addresses of the folders in which the data is to be found or the result (and temporary intermediate results) is to be written in the two path fields (“input” and “output” path).



In the “analysis unit” section, you decide whether all persons are considered individually within the household (“individual”) or whether you are interested solely in the household as a whole (“household”). With “sample composition” you can choose between “balanced” and “unbalanced”. If you select “balanced”, you will receive a dataset without missing codes. The respondents provided information on all variables. For more information about balanced and unbalanced datasets, see the section *Principles of Data Analysis*. Under “age group”, you can limit the respondents. When you are satisfied with your settings, click on “Update Script” and your script will be created.

If you click on the “raw script” button, the script is displayed in text form. Copy it to your statistical software. The result dataset has the name new(.dta, .sav). If you want to change it, you have to do it in the script. Execute the script with your statistical software and you will receive your dataset with all your chosen variables.

Last change: Nov 12, 2019

6.2 Working with Tracking Data (PPATH / PPFAD)

For all years since 1984, the PPATH dataset contains information on all persons who have ever lived in a SOEP household when a survey was conducted (i.e., all adult respondents as well as children under 17 years of age and household members who have never given an interview). PPATH is important in distinguishing research units (persons), especially for longitudinal analysis. In addition, paneldata.org uses PPATH to differentiate the study population.

Time-constant information on individuals:

- Permanent Individual ID (adults, adolescents, children)
- Original Household Number
- Gender, year of birth, month of birth, year of death if applicable
- Migration Background
- Sample Membership (psample)

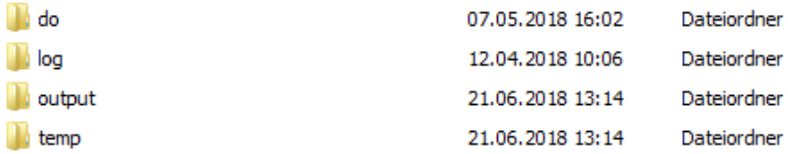
Time-varying information from individuals:

- Current Household Number: If you move to another household, the household number changes (hhnrakt or $hhnr)
- Survey Status ($netto, $netold)
- Population Membership (private household, institutional household)
- Survey Region (East or West Germany)

The dataset is explained in more detail in the following documentation:

Dokumentation PPATH:

Create an exercise path with four subfolders:



Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store your script, log files, datasets and temporary datasets. Open an empty do-file and define your paths with globals:

```
*********************************************
* Set relative paths to the working directory
*********************************************
global AVZ         "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\complete\soep-core\soep.v33.2\stata_en\"
global MY_DO_FILES "$AVZ\do\"
```

(continues on next page)

(continued from previous page)

```
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

Attention: Please note that since version 34 (v34), PPFAD has been renamed PPATH. The following ecxercises are done with version 33.2 (v33.2), where the tracking file was named PPFAD.

The global „AVZ“ defines the main path. The main paths are subdivided using the globals “MY_IN_PATH”, “MY_DO_FILES”, “MY_LOG_OUT”, “MY_OUT_DATA”, “MY_OUT_TEMP”. The global “MY_IN_PATH” contains the path to the data you ordered.

Based on the data in PPATH, answer the following questions:

1. Look at the two people with the individual IDs (variable persnr) 2102 and 19202

a) What is their gender? When were they born and when (if applicable) did they die?

Open the PPATH dataset. Search the dataset for variables that describe sex, year of birth, and year of death. Display the information from the variables for individuals 2102 and 19202.

```
use "${MY_IN_PATH}ppfad.dta", clear

* a) What is their gender? When were they born and when (if applicable) did they die?
list persnr sex gebjahr todjahr if persnr == 2102 | persnr == 19202
```

```
. * a) What gender are they? When were they born and eventually died?
. list persnr sex gebjahr todjahr if persnr == 2102 | persnr == 19202
```

	persnr	sex	gebjahr	todjahr
59.	2102	[2] Female	1927	1999
639.	19202	[1] Male	1960	-2

b) Were these people and their parents born in Germany?

In the dataset, search for a variable that describes the migration background. Display the information from the variables for indivduals 2102 and 19202.

```
* b) Were these people and their parents born in Germany?
list persnr migback if persnr == 2102 | persnr == 19202
```

```
. * b) Were these people and their parents born in Germany?
. list persnr migback if persnr == 2102 | persnr == 19202
```

	persnr	migback
59.	2102	[1] no migration background
639.	19202	[2] direct migration background

c) If they immigrated to Germany, in which year and from what country?

Search the dataset for a variable that describes the country of birth and the year of moving to Germany. Display the information from the variables on indivduals 2102 and 19202.

```
*c) If they have immigrated: In which year and from which country?
list persnr immiyear corigin if persnr  == 2102 | persnr == 19202
```

```
. *c) If they have immigrated: In which year and from which country?
. list persnr immiyear corigin if persnr  == 2102 | persnr == 19202
```

	persnr	immiyear	corigin
59.	2102	-2	[1] Germany
639.	19202	1980	[2] Turkey

d) Are these people from East or West Germany?

Search the dataset for a variable that tells whether respondents are from the East or West. Display the information from the variables for individuals 2102 and 19202.

```
*d) Are these people from East or West Germany?
list persnr loc1989 psample if persnr  == 2102 | persnr == 19202
```

```
. *d) Are these people from East or West Germany?
. list persnr loc1989 psample if persnr  == 2102 | persnr == 19202
```

	persnr	loc1989	psample
59.	2102	[2] West Germany (FRG) incl. West Berlin	[1] A 1984 Initial Sample (West)
639.	19202	[2] West Germany (FRG) incl. West Berlin	[1] A 1984 Initial Sample (West)

e) What sources provide the information on the migration background and year of death

Search the data set for variables that give you the sources of information for year of death and migration background. Display the information from the variables for individuals 2102 and 19202.

```
*e) What sources provide the information on the migration background and year of death?
list miginfo todinfo if persnr  == 2102 | persnr == 19202
```

```
. *e) From which sources does the information on the migration background and the year of death come?
. list miginfo todinfo if persnr  == 2102 | persnr == 19202
```

	miginfo	todinfo
59.	[1] direct personal w/o parental info	[5] Infratest drop-out study 2001
639.	[1] direct personal w/o parental info	[-2] Does not apply

2. How many people lived in a private household that was interviewed in 2016 and completed the individual questionnaire?

Remember that the wave-specific survey year in SOEP is abbreviated with letters. SOEP started with wave "a" in 1984 and had reached wave "bg" in 2016. For more information on this topic, please refer to the SOEPcompanion subchapter Label.

If you are interested in the 2016 survey year, the wave name indicates that you should be interested in variables with the abbreviation "bg". Search the dataset for variables with the abbreviation "bg" that describe the population. Display the characteristics of the population variables:

```
*******************************************************************************
*** Exercise 2) ***
* How many people lived in an interviewed private household in 2016 and answered the
* individual questionnaire?

*******************************************************************************

* informationen from:
* 2016 -> Wave bg
* private household -> bgpop
* Individual questionnaire -> bgnetto

tab bgpop
```

```
. tab bgpop
```

Sample Membership 2016	Freq.	Percent	Cum.
[-2] Does not apply	68,743	54.49	54.49
[1] Private HH, German HH-Head	31,696	25.13	79.62
[2] Private HH, Foreign HH-Head	13,972	11.08	90.69
[3] Institutional. HH, Collective accom	141	0.11	90.81
[4] Institutional. HH, Collective accom	3,067	2.43	93.24
[5] Not Compl. Private HH, German HH-He	5,947	4.71	97.95
[6] Not Compl. Private HH, Foreign HH-H	2,518	2.00	99.95
[7] Not Compl. Institutional. HH, Colle	31	0.02	99.97
[8] Not Compl. Institutional. HH, Colle	36	0.03	100.00
Total	126,151	100.00	

Values 1 and 2 are relevant to answer the question because they describe interviewed households. Search the dataset for variables with the abbreviation "bg" that describe the survey status. Display the characteristics of the survey status:

```
tab bgnetto
```

```
. tab bgnetto
```

Survey Status 2016	Freq.	Percent	Cum.
[-2] Does not apply	68,743	54.49	54.49
[10] Interviewee With Succesful Intervi	5,562	4.41	58.90
[12] Individual Questionnaire And Perso	8,570	6.79	65.70
[14] Individual Questionnaire And Other	30	0.02	65.72
[15] Individual Questionnaire And Exper	14,903	11.81	77.53
[17] Youth Biography First Time Surveye	535	0.42	77.96
[19] Individual Questionnaire Without H	113	0.09	78.05
[20] Children in Succesfully Interviewe	10,682	8.47	86.51
[21] Children With Mother-Child Questio	349	0.28	86.79
[22] Children With Mother-Child Questio	393	0.31	87.10
[23] Children With Mother-Child Questio	685	0.54	87.64
[24] Children age 7-8, with parental qu	746	0.59	88.24
[25] Children age 9-10, with parental q	538	0.43	88.66
[26] Students Age 11-12	559	0.44	89.11
[28] Youth questionnaire, Age 13-14	526	0.42	89.52
[29] Youth from refugee sample, age 16-	222	0.18	89.70
[30] Persons In Successfully Interviewe	12,361	9.80	99.50
[32] Successfully Completed Biography Q	1	0.00	99.50
[34] Successful Tests and Experiments	13	0.01	99.51
[90] Individual Dropouts PBR_EXIT	306	0.24	99.75
[91] Moved abroad	133	0.11	99.86
[99] Has Died	181	0.14	100.00
Total	126,151	100.00	

Respondents with survey status between 10 and 15 or survey status 19 completed the individual questionnaire. Cross-tabulate the variables bgpop and bgnetto with an appropriate restricting condition to answer the question.

```
tab bgnetto bgpop if ((bgnetto >= 10 & bgnetto <= 15) | bgnetto==19) & (bgpop==1 |␣
↪bgpop==2)
```

```
. tab bgnetto bgpop if ((bgnetto >= 10 & bgnetto <= 15) | bgnetto==19) & (bgpop==1 | bgpop==2)
```

Survey Status 2016	Sample Membership 2016 [1] Priva	[2] Priva	Total
[10] Interviewee With	5,362	173	5,535
[12] Individual Quest	1,685	5,365	7,050
[14] Individual Quest	30	0	30
[15] Individual Quest	14,055	757	14,812
Total	21,132	6,295	27,427

3. PPATH allows you to see which populations can be viewed from a longitudinal perspective:

a) How many people who answered the individual questionnaire in 2000 also took part in the survey in 2014?

Remember that the wave-specific survey year in SOEP is abbreviated with letters. SOEP started with wave "a" in 1984 and had reached wave "bg" in 2016. For more information on this subject, see the subchapter Label. The wave name shows that you are interested in the survey years 2000 and 2014. The survey years include the wave names "q"(2000) and "be"(2014). Search the dataset for variables with the abbreviations "q" and "be" that describe the survey status. Display the characteristics of the survey status under the condition that the individual questionnaire has been answered:

```
* a)How many people who answered the individual questionnaire in 2000 also took
*   part in the survey in 2014?

* informationen from:
*        2000 -> wave q
*          2014 -> wave be
*         Individual questionnaire -> $netto

tab qnetto benetto  if qnetto>=10 & qnetto<=19 & benetto>=10 & benetto<=19
*or:
//fre qnetto benetto  if qnetto>=10 & qnetto<=19 & benetto>=10 & benetto<=19
```

```
. tab qnetto benetto  if qnetto>=10 & qnetto<=19 & benetto>=10 & benetto<=19
```

Current Wave Survey Status 2000	Current Wave Survey Status 2014 [10] Inte	[12] Indi	[15] Indi	[19] Indi	Total
[10] Interviewee With	5,044	1	2,457	3	7,505
[12] Individual Quest	47	0	16	0	63
[16] Individual Quest	52	0	19	0	71
Total	5,143	1	2,492	3	7,639

A total of 7,639 respondents completed the individual questionnaire in 2000 and 2014.

b) How many people answered the individual questionnaire every year from 2000 to 2014?

The survey years include the wave designations from "q"(2000) to "be"(2014). View the relevant survey status codes to answer the question. Please consider all individuals who completed the individual questionnaire:

```
* b) How many people answered the individual questionnaire every year from 2000
*    to 2014?

/* to see all the codes */
lab list bgnetto
```

```
bgnetto:
          -6 [-6] Version of questionnaire with modified filtering
          -5 [-5] Not included in this version of the questionnaire
          -4 [-4] Inadmissible multiple response
          -3 [-3] Answer improbable
          -2 [-2] Does not apply
          -1 [-1] No Answer
          10 [10] Interviewee With Succesful Interview (_P)
          12 [12] Individual Questionnaire And Person Biography
          13 [13] Individual Questionnaire And Youth Biography
          14 [14] Individual Questionnaire And Other Questionnaires
          15 [15] Individual Questionnaire And Experiments, Test
          16 [16] Individual Questionnaire, First Time Surveyed, Age 17
          17 [17] Youth Biography First Time Surveyed, Age 17
          18 [18] Individual Questionnaire And Child under age 17
          19 [19] Individual Questionnaire Without Household Interview
          20 [20] Children in Succesfully Interviewed Households (_Kind)
          21 [21] Children With Mother-Child Questionnaire_I, Age 0-1
          22 [22] Children With Mother-Child Questionnaire_II, Age 2-3
          23 [23] Children With Mother-Child Questionnaire_III, Age 5-6
          24 [24] Children age 7-8, with parental questionnaire
          25 [25] Children age 9-10, with parental questionnaire
          26 [26] Students Age 11-12
          27 [27] Children with Mother-Child Questionnaire, Age 1-2
          28 [28] Youth questionnaire, Age 13-14
          29 [29] Youth from refugee sample, age 16-17
          30 [30] Persons In Successfully Interviewed HH Without Individual Interview
          31 [31] Successful Gap Interview (_LUECKE)
          32 [32] Successfully Completed Biography Questionnaires
          33 [33] Successful Youth Questionnaire
          34 [34] Successful Tests and Experiments
          60 [60] Only Questionnaire Without Indiv. And HH Interview
          61 [61] Gap Interview without HH reference
          62 [62] Gap Interview with drop out
          70 [70] Only Participation In Tests, Experiments, etc.
          80 [80] Individual Without Any Current Information
          81 [81] Prior Interviewee Without Any Current Information
          88 [88] Repatriate - (moved abroad before [91])
          89 [89] Repatriate - (was drop out [90])
          90 [90] Individual Dropouts PBR_EXIT
          91 [91] Moved abroad
          92 [92] Moved abroad (abroad)
          93 [93] Moved abroad (exit)
          94 [94] Person Gap with advices
          97 [97] advice to dead person (exit)
          98 [98] advice to dead person (_VP)
          99 [99] Has Died
```

Define a variable list that shows all survey statuses ($netto) from the 15 survey waves considered in total.

```
local v "netto"
local vlist "q`v' r`v' s`v' t`v' u`v' v`v' w`v' x`v' y`v' z`v' ba`v' bb`v' bc`v' bd`v
↪' be`v'"
```

(continues on next page)

(continued from previous page)

```
/* --> 15 waves */
```

Generate a variable that shows the number of waves of completed individual interviews. Note that the values 10,12,13,14,15,16,18,19 of the $netto variable mean realized interviews.

```
capture drop h1
egen h1 = anycount(`vlist'), values(10 12 13 14 15 16 18 19)
```

Display a table with its newly generated variable.

```
tab h1 if h1 == 15
```

```
. tab h1 if h1 == 15

  see notes |      Freq.     Percent        Cum.
------------+-----------------------------------
         15 |      6,665      100.00      100.00
------------+-----------------------------------
      Total |      6,665      100.00
```

A total of 6,665 people completed the individual questionnaire every year over the period 2000-2014.

c) How many people who turned 15 in 2011 and spent at least part of their childhood in a SOEP household took part in the survey in 2016?

The survey year 2011 is represented by the wave "bb" and the survey year 2016 is represented by the wave "bg". To answer the question, a variable must be generated that identifies people who were 15 years old in 2011. The age of the respondent can be determined with the year of birth, and you can limit children using the net code. Generate a variable with people who turned 15 in 2011 and had lived in a survey household as a child.

```
* c) How many people who turned 15 in 2011 and lived as children in a survey
*    household took part in the survey in 2016?

*   informationen from:
*          2011 -> wave bb
*        Age  -> 15
*        Child -> bbnetto
*        2016 -> wave bg
*         Individual Questionnaire -> bgnetto

/* People who turned 15 in 2011 and lived in a survey household as a child...*/
capture drop a15kind
gen a15kind = 1 if 2011-gebjahr == 15 & bbnetto >= 20 & bbnetto < 30

```

In order to identify all persons who were 15 years old in 2011, lived in a survey household as a child, and completed the individual questionnaire in 2016, you must use the net codes again. Create a table using the net code from 2016 to narrow down the cases appropriately.

```
// fre bgnetto if a15kind == 1 & bgnetto >= 10 & bgnetto < 20
* oder:
tab bgnetto if a15kind == 1 & bgnetto >= 10 & bgnetto < 20

```

```
. tab bgnetto if a15kind == 1 & bgnetto >= 10 & bgnetto < 20
```

Survey Status 2016	Freq.	Percent	Cum.
[10] Interviewee With Succesful Intervi	70	22.65	22.65
[12] Individual Questionnaire And Perso	2	0.65	23.30
[15] Individual Questionnaire And Exper	227	73.46	96.76
[19] Individual Questionnaire Without H	10	3.24	100.00
Total	309	100.00	

In 2016, a total of 309 people who were 15 years old at the time of the survey and had been part of a survey household as a child in 2011 completed an individual interview.

d) The person with persnr=588010 was born in 1984 in a panel household and was still part of the sample in 2009. The person changed households twice during this time. In which years?

To identify how often and when a person changed households, you must display all available household numbers in ppath for person 588010.

```
* still part of the sample in 2009. The person has changed households twice during
* this time. In which years?

* Information from:
* -> household numbers

list *hhnr if persnr == 588010
/* -> changed household
 in year d (1987)
 in year y (2008)
 no participation since bb (2011)
*/
```

```
. list *hhnr if persnr == 588010
```

	hhnr	ahhnr	bhhnr	chhnr	dhhnr	ehhnr	fhhnr	ghhnr	hhhnr	ihhnr	jhhnr	khhnr	lhhnr	mhhnr	nhhnr	ohhnr	phhnr	qhhnr	rhhnr	shhnr
25347.	58807	-2	58807	58807	73407	73407	73407	73407	73407	73407	73407	73407	73407	73407	73407	73407	73407	73407	73407	73407

thhnr	uhhnr	vhhnr	whhnr	xhhnr	yhhnr	zhhnr	bahhnr	bbhhnr	bchhnr	bdhhnr	behhnr	bfhhnr	bghhnr
73407	73407	73407	73407	73407	132608	132608	132608	-2	-2	-2	-2	-2	-2

Person 588010 has participated in the survey since the wave “b” (1985) as part of household 58807. From wave “d” (1987) to wave “x” (2007) the person was in household 73407, from wave “y” (2008) on, the person was in household 132608.

Last change: Nov 12, 2019

6.3 Generating a Cross-Sectional Dataset

This example involves generating a dataset to analyze health satisfaction determinants in 2008, and you can either use the Paneldata.org syntax generator or write your own syntax file to perform this task. You can search for the variable names in Paneldata.org (or use the variables below directly).

1. Generate a cross-sectional dataset for the year 2008, which should contain all persons with the following characteristics:

- Respondents in 2008 **"ynetto"**
- Lived in a private household in 2008 **"ypop"**

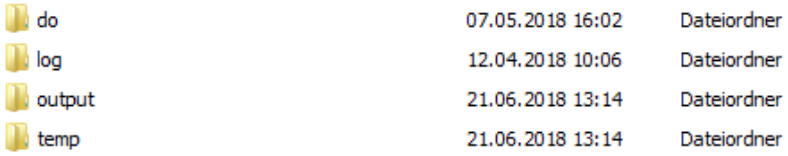
The dataset should contain the following variables of interest.

- satisfaction with health **"yp0101"**
- smoking currently yes/no **"yp10601"**
- current employment status **"emplst08"**
- monthly household net income **"hinc08"**

In addition, the dataset should contain the following additional information for a 2008 cross-sectional analysis (these variables are automatically generated by paneldata.org):

- current cross-section weighting factor **"yphrf"**
- personal number **"persnr"**
- original household number **"hhnr"**
- current household number **"yhhnr"**
- sample affiliation **"psample"**
- gender **"sex"**
- year of birth **"gebjahr"**

Create an exercise path with four subfolders:



Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store commands, log files, datasets, and temporary datasets. Open an empty do file and define your created paths with globals:

```
*************************************************
* Set relative paths to the working directory
*************************************************
global AVZ          "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\complete\soep-core\soep.v33.2\stata_en\"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global "AVZ" defines the main path. The main paths are subdivided using the globals "MY_IN_PATH", "MY_DO_FILES", "MY_LOG_OUT", "MY_OUT_DATA", "MY_OUT_TEMP". The global "MY_IN_PATH" contains the path to your data.

Use ppath as the source file together with the required variables. Keep all cases with completed interviews. In addition, your dataset should only contain respondents who can make a statement on the content of the question. For example, you can use the net code to identify and remove children from your dataset.

```
* * * PFAD * * *

use hhnr persnr sex gebjahr psample yhhnr ynetto ypop using "${MY_IN_PATH}ppfad.dta"


* * * BALANCED VS UNBALANCED * * *

keep if ( (ynetto >= 10 & ynetto < 20) )


* * * PRIATVE VS ALL HOUSEHOLDS * * *

keep if ( (ypop == 1 | ypop == 2) )


* * * SORT PFAD * * *

sort persnr
save "${MY_OUT_TEMP}ppfad.dta", replace
clear
```

Attention: Please note that since version 34 (v34), PPFAD can be found in the subdirectory "Raw" of the data distribution file. The following exercises are done with version 33.1 (v33.1), where the tracking file was named PPFAD.

Save the modified data temporarily. Now link your dataset with the weights of the SOEP and save your dataset as a master file.

```
* * * HRF * * *

use "${MY_IN_PATH}phrf.dta"
sort persnr
save "${MY_OUT_TEMP}hrf.dta", replace
clear


* * * CREATE MASTER * * *

use "${MY_OUT_TEMP}ppfad.dta"
merge 1:1 persnr using "${MY_OUT_TEMP}hrf.dta"
drop if _merge == 2
drop _merge
sort persnr
save "${MY_OUT_TEMP}master.dta", replace
clear
```

Now prepare the content variables. Search for the content variables you are looking for from the various datasets and temporarily save the datasets you have created.

```
* * * READ DATA * * *

use hinc08 yhhnr using "${MY_IN_PATH}yhgen.dta"
sort yhhnr
save "${MY_OUT_TEMP}yhgen.dta", replace
clear


use yp10601 yhhnr yp0101 persnr using "${MY_IN_PATH}yp.dta"
sort persnr
save "${MY_OUT_TEMP}yp.dta", replace
clear


use emplst08 yhhnr persnr using "${MY_IN_PATH}ypgen.dta"
sort persnr
save "${MY_OUT_TEMP}ypgen.dta", replace
clear
```

Link the datasets you have created to your master file and save for analysis.

```
* * * MERGE DATA * * *

use    "${MY_OUT_TEMP}master.dta"

sort yhhnr
merge yhhnr using "${MY_OUT_TEMP}yhgen.dta"
drop if _merge == 2
drop _merge

sort persnr
merge persnr using "${MY_OUT_TEMP}yp.dta"
drop if _merge == 2
drop _merge

sort persnr
merge persnr using "${MY_OUT_TEMP}ypgen.dta"
drop if _merge == 2
drop _merge


* * * DONE * * *

save "${MY_OUT_DATA}my_dataset.dta", replace
desc
```

You have successfully created a cross-sectional dataset for the year 2008.

2. Encode missing values into system missings (STATA)!

In SOEP, the missing codes of variables are described in detail with the values -1 to -8. To learn more about missing codes, see the section *Missing Conventions*. For content analysis, it is not always necessary to differentiate missing codes. Therefore you should be able to convert missing codes:

```
use "$MY_OUT_DATA\my_dataset.dta", clear


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

(continues on next page)

(continued from previous page)

```
*** Exercise 2) ***
* Encode missing values into missing values in system missings (STATA)!
*******************************************************************************

* mvdecode = Change missing values to numeric values and vice versa
        mvdecode _all, mv(-1=. \ -2=.t \ -3=.x \ -5=.y \ -8=.z)
```

Open the dataset for your analysis and summarize all missing codes.

3. How does average health satisfaction differ a) by gender

Satisfaction was measured on a scale of 1 to 10. To compare average satisfaction with health between women and men, you should display the mean value for both genders.

```
        *unweighted*
        tabstat yp0101, by(sex)
```

```
. *a) by sex:
.         *unweighted*
.         tabstat yp0101, by(sex)

Summary for variables: yp0101
     by categories of: sex (Sex)

             sex |      mean
-----------------+----------
        [1] Male |  6.616534
      [2] Female |  6.516729
-----------------+----------
           Total |   6.56428
-------------------------------
```

Since you have previously added the SOEP weighting factors to the dataset for your analysis, you should use the weighting for a representative analysis.

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(sex)
```

```
.          *weighted*
.          tabstat yp0101 [aw=yphrf], by(sex)

Summary for variables: yp0101
     by categories of: sex (Sex)
```

sex	mean
[1] Male	6.53008
[2] Female	6.407367
Total	6.467019

b) Employment status

Now proceed in a similar way when comparing satisfaction with health and employment status. Compare the mean values again:

```
*b) by job status:
        *unweighted*
        tabstat yp0101, by(emplst08)
```

```
. *b) by job status:
.          *unweighted*
.          tabstat yp0101, by(emplst08)

Summary for variables: yp0101
     by categories of: emplst08 (Employment Status)
```

emplst08	mean
[1] Full-Time Em	6.931818
[2] Regular Part	6.805956
[3] Vocational T	7.792453
[4] Marginal, Ir	6.739879
[5] Not Employed	6.085035
[6] Sheltered wo	5.72
Total	6.56428

Since you have previously added the SOEP weighting factors to the dataset for your analysis, you should use the weighting for a representative analysis.

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(emplst08)
```

```
.         *weighted*
.         tabstat yp0101 [aw=yphrf], by(emplst08)

Summary for variables: yp0101
     by categories of: emplst08 (Employment Status)

        emplst08 |      mean
-----------------+----------
[1] Full-Time Em |  6.847115
[2] Regular Part |  6.704637
[3] Vocational T |  7.822574
[4] Marginal, Ir |  6.615801
[5] Not Employed |  5.987851
[6] Sheltered wo |  4.937647
-----------------+----------
           Total |  6.467019
```

c) Age

Since you do not have a variable that represents age, you must generate a suitable age variable using the birth year variable. The year of birth is metric and should be categorized for analysis. Define categories for your age variable and assign suitable labels.

```
*c) by age in 2008 (<30, 30-64, 65+)

        gen age=2008-gebjahr
        gen age_3=age
        recode age_3 (17/29=1) (30/64=2) (65/120=3)
        label define age_3 1 "17-29" 2 "30-64" 3 "65+"
        label values age_3 age_3
```

Create a mean value comparison with your age variable and health satisfaction in weighted and unweighted form.

```
        *unweighted*
        tabstat yp0101, by(age_3)
```

```
.         *unweighted*
.         tabstat yp0101, by(age_3)

Summary for variables: yp0101
     by categories of: age_3

   age_3 |      mean
---------+----------
   17-29 |  7.640552
   30-64 |  6.607247
     65+ |  5.714101
---------+----------
   Total |   6.56428
```

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(age_3)
```

```
.         *weighted*
.         tabstat yp0101 [aw=yphrf], by(age_3)

Summary for variables: yp0101
     by categories of: age_3

 age_3 |      mean
-------+----------
 17-29 |  7.595288
 30-64 |  6.483365
   65+ |  5.660658
-------+----------
 Total |  6.467019
------------------
```

d) Income

As with age, generate a categorized version of income for household net income:

```
*d) by monthly houshold net income (-1.999, 2.000-3.999, 4000+ Euro)
        gen hinc08_3 = hinc08
        recode hinc08_3 (0/1999=1) (2000/3999=2) (4000/99999=3)
        label define hinc08_3 1 "<2000 Euro" 2 "2000-<4000 Euro" 3 "4000+ Euro"
        label values hinc08_3 hinc08_3
```

Display the mean values in weighted and unweighted form:

```
        *unweighted*
        tabstat yp0101, by(hinc08_3)
```

```
.         *unweighted*
.         tabstat yp0101, by(hinc08_3)

Summary for variables: yp0101
     by categories of: hinc08_3

    hinc08_3 |      mean
-------------+----------
     <2000 Euro |  6.042256
2000-<4000 Euro |   6.69125
     4000+ Euro |   7.11391
-------------+----------
          Total |  6.551677
--------------------------
```

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(hinc08_3)
```

```
.         *weighted*
.         tabstat yp0101 [aw=yphrf], by(hinc08_3)

Summary for variables: yp0101
     by categories of: hinc08_3

       hinc08_3 |      mean
----------------+----------
     <2000 Euro |  5.988714
2000-<4000 Euro |    6.6906
     4000+ Euro |  7.126235
----------------+----------
          Total |  6.446908
-----------------------------
```

e) Smoking

Since this variable is nominal, adjustments to this variable are not necessary. Display average satisfaction with health for smokers and non-smokers in weighted and unweighted form:

```
*e) by smoking yes/no

        *unweighted*
        tabstat yp0101, by(yp10601)
```

```
.         *unweighted*
.         tabstat yp0101, by(yp10601)

Summary for variables: yp0101
     by categories of: yp10601 (Currently Smoke)

       yp10601 |      mean
---------------+----------
     [1] Yes   |  6.551121
     [2] No    |  6.570124
---------------+----------
          Total |  6.564997
-----------------------------
```

```
        *weighted*
        tabstat yp0101 [aw=yphrf], by(yp10601)
```

```
.          *weighted*
.          tabstat yp0101 [aw=yphrf], by(yp10601)

Summary for variables: yp0101
     by categories of: yp10601 (Currently Smoke)

         yp10601 |      mean
-----------------+----------
         [1] Yes |  6.448555
          [2] No |  6.476664
-----------------+----------
           Total |  6.468664
-------------------------------
```

Last change: Nov 12, 2019

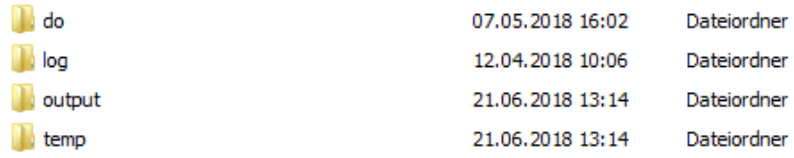
6.4 Working with Migration Data (BIOIMMIG)

With its migration and refugee samples, SOEP provides a wide range of information on people with a history of migration or forced migration.

In the BIOIMMIG dataset, you will find relevant information on the history of migration or forced migration, including refugees' and migrants' motives for leaving their country of origin, their living conditions upon arrival in Germany, as well as information in edited form on any relatives in the country of origin and the desire to return to the country of origin. For more information about this dataset and a list of the variables it contains, see: BIOIMMIG Documentation.

In the following, we will use this record and other information from the SOEP to create a status variable that you can use to distinguish whether or not people with a migration background also have a background of forced migration, that is, whether migrants are also refugees.

Create an exercise path with four subfolders:



Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store commands, log files, datasets, and temporary datasets. Open an empty do-file and define your paths with globals:

```
*************************************************
* Set relative paths to the working directory
*************************************************
global AVZ          "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\complete\soep-core\soep.v33.2\stata_en\"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global "AVZ" defines the main path. The main paths are subdivided using the globals "MY_IN_PATH", "MY_DO_FILES", "MY_LOG_OUT", "MY_OUT_DATA", "MY_OUT_TEMP". The global "MY_IN_PATH" contains the path to the data you ordered.

Task 1: Preparation of BIOIMMIG

a) Which variable contains information about the status of each person when they immigrated to Germany?

Open the record or browse the BIOIMMIG documentation and search for a variable describing the immigration status. The biimgrp variable from the BIOIMMIG data set is the appropriate variable.

```
*** Exercise 1 ********************************************************************

/*
a)        Which variable contains information about the status of each person when␣
↪they immigrated to Germany?
*/

* Immigration status is stored in the variable biimgrp.

use $MY_IN_PATH\bioimmig.dta, clear
```

b) Identify this variable in the BIOIMMIG dataset and retrieve it from the dataset together with the person number and survey year.

Open your dataset with just the required variables to maintain clarity for your analysis.

```
/*
b)        Identify this variable in the BIOIMMIG dataset and load it from the data␣
↪set, together with the individual indentifier and the survey year.
*/

use persnr syear biimgrp using $MY_IN_PATH\bioimmig.dta, clear
```

c) What are the values of this variable?

Familiarize yourself with your variable and check the coding and case numbers.

```
/*
c)        What are the values of this variable?
*/

tab biimgrp, m //Characteristics of the variable are examined.
```

```
.
. tab biimgrp, m //Characteristics of the variable are examined.
```

BI: Immigration Group	Freq.	Percent	Cum.
[-5] Not included in this version of th	5,848	3.14	3.14
[-2] Does not apply	113,969	61.23	64.37
[-1] No Answer	1,373	0.74	65.11
[1] East German	3,687	1.98	67.09
[2] Person Of German Descent From Easte	28,029	15.06	82.15
[3] German Who Lived Abroad	1,195	0.64	82.79
[4] Citizen Of EU Country (up to 2009 E	6,935	3.73	86.52
[5] Asylum seeker, refugee	9,419	5.06	91.58
[6] Other Foreigner	15,681	8.42	100.00
Total	186,136	100.00	

d) On the basis of this variable, generate the variable "escape", which only distinguishes among three groups:

- 0 = Cases where no information is available
- 1 = All persons without escape background
- 2 = Asylum seekers / refugees

After you have familiarized yourself with the variable, recode it to fit your project. Then check the case numbers of your generated variable with the source variable.

```
/*
d)        On the basis of this variable, generate the variable "Escape", which only distinguishes between three groups:
    0 = Cases where no information is available
    1 = All persons without escape background
    2 = Asylum seekers / refugees
*/

recode biimgrp (-5 -2 -1 = 0 "No Answer") (1 2 3 4 6 = 1 "no Escape") (5 = 2 "Escape"), gen(Escape)
tab biimgrp Escape, m // biimgrp and escape are compared.
```

```
. tab biimgrp Escape, m // biimgrp and escape are compared.
```

BI: Immigration Group	RECODE of biimgrp (BI: Immigration Group)			
	No Answer	no Escape	Escape	Total
[-5] Not included in	5,848	0	0	5,848
[-2] Does not apply	113,969	0	0	113,969
[-1] No Answer	1,373	0	0	1,373
[1] East German	0	3,687	0	3,687
[2] Person Of German	0	28,029	0	28,029
[3] German Who Lived	0	1,195	0	1,195
[4] Citizen Of EU Cou	0	6,935	0	6,935
[5] Asylum seeker, re	0	0	9,419	9,419
[6] Other Foreigner	0	15,681	0	15,681
Total	121,190	55,527	9,419	186,136

e) It may be that initially there is no information on the immigration status but this will change one year later. Limit the dataset to the last observation available on the respective person, since this gives you the most comprehensive information.

```
e)        It may happen that tinitially there is no information on the status of
*   immigration, but this will change in a later year. Limit the data record to
*   the last observation that is available for the respective person, since this
*   way the specification with the most information content is used.
*/

bysort persnr: egen syear_max = max(syear) //A variable is created, which shows the last existing yearly observation
keep if syear_max == syear //Annual observations which are not the last observation are deleted.
```

f) Save the generated data temporarily on your personal drive.

```
f)        Save the generated data record on your personal drive temporarily
*/

save $MY_OUT_TEMP\biimgrp.dta, replace
```

Task 2: Add basic variables from PPATH and weights

Attention: Please note that since version 34 (v34), PPFAD can be found in the subdirectory "Raw" of the data distribution file. The following exercises are done with version 33.1 (v33.1), where the tracking file was named PPFAD.

a) Load the following information from PPATH:

- Permanent Indivdiual Identifier **"persnr"**
- Household Identifier **"hhnr"** and the current household number **"bghhnr"**

- The net variable with information about the interview type **"bgnetto"**
- The gender of the person **"sex"**
- The year of birth **"gebjahr"**
- Variables on the migration background **"migback"**, **"germborn"**, **"corigin"**, **"immiyear"**
- Information about the survey status: **"psample"**

If you want to familiarize yourself with the PPATH dataset, see the section *Working with Tracking Data (PPATH / PPFAD)*.

```
/*
a)        Use the following information from PPFAD:
  - Unchanging Person ID „persnr"
  - Household number "hhnr" and the current household number "bghhnr".
  - the net variable with information about the interview type "bgnetto".
  - the gender of the person "sex"
  - the year of birth "semester"
  - Variables on the migration background "migback", "germborn" "corigin" "immiyear"
  - Information about the survey status: "bgnetto" and "psample".
*/

use persnr hhnr bghhnr bgnetto psample sex gebjahr germborn corigin immiyear migback ␣
↪using $MY_IN_PATH\ppfad.dta, clear
```

b) Merge the previously generated data using the individual identifier.

If you don't understand how to create your own cross-sectional dataset, see the chapter *Generating a Cross-Sectional Dataset*.

```
/*
b)        Merge the previously generated data set using the person number.
*/

merge 1:1 persnr using $MY_OUT_TEMP\biimgrp.dta, nogen
```

c) Add the corresponding individual extrapolation factors to the data.

```
c)        Add the corresponding data using the individual identifier.
*/

merge 1:1 persnr using $MY_IN_PATH\phrf.dta, keepus(bgphrf) nogen
```

d) Only keep respondents who completed a youth or individual questionnaire in 2016.

For example, to exclude children who have not provided immigration status information, use the net code from PPATH. Only keep individuals who completed an individual or youth interview.

```
/*
d)        Only keep respondents who completed a youth or individual questionnaire in␣
↪2016.
*/

tab bgnetto, m //Variable values are displayed

keep if inrange(bgnetto, 10, 19) // People who have a code between 10 and 19 will be␣
↪kept.
```

```
. tab bgnetto, m //Variable values are displayed
```

Survey Status 2016	Freq.	Percent	Cum.
[-2] Does not apply	68,743	54.49	54.49
[10] Interviewee With Succesful Intervi	5,562	4.41	58.90
[12] Individual Questionnaire And Perso	8,570	6.79	65.70
[14] Individual Questionnaire And Other	30	0.02	65.72
[15] Individual Questionnaire And Exper	14,903	11.81	77.53
[17] Youth Biography First Time Surveye	535	0.42	77.96
[19] Individual Questionnaire Without H	113	0.09	78.05
[20] Children in Succesfully Interviewe	10,682	8.47	86.51
[21] Children With Mother-Child Questio	349	0.28	86.79
[22] Children With Mother-Child Questio	393	0.31	87.10
[23] Children With Mother-Child Questio	685	0.54	87.64
[24] Children age 7-8, with parental qu	746	0.59	88.24
[25] Children age 9-10, with parental q	538	0.43	88.66
[26] Students Age 11-12	559	0.44	89.11
[28] Youth questionnaire, Age 13-14	526	0.42	89.52
[29] Youth from refugee sample, age 16-	222	0.18	89.70
[30] Persons In Successfully Interviewe	12,361	9.80	99.50
[32] Successfully Completed Biography Q	1	0.00	99.50
[34] Successful Tests and Experiments	13	0.01	99.51
[90] Individual Dropouts PBR_EXIT	306	0.24	99.75
[91] Moved abroad	133	0.11	99.86
[99] Has Died	181	0.14	100.00
Total	126,151	100.00	

Task 3: Generate a status variable with the following categories:.

- No migration background
- Migrant, 2nd generation
- Migrant, no information
- Migrant, not refugee
- Migrant, refugee

To generate this status variable, check the contents of the existing migration variables from PPATH (migback germborn).

```
/*
Generate a status variable with the following categories:
*/

tab migback
```

```
. tab migback
```

Migration background	Freq.	Percent	Cum.
[1] no migration background	18,099	60.91	60.91
[2] direct migration background	9,456	31.82	92.74
[3] indirect migration background	2,158	7.26	100.00
Total	29,713	100.00	

```
tab germborn
```

Born in Germany	Freq.	Percent	Cum.
[1] born in Germany or immigr.<1950	20,257	68.18	68.18
[2] not born in Germany	9,456	31.82	100.00
Total	29,713	100.00	

Use the migration variables from PPATH (migback, germborn) and link this information with your previously generated refugee variable to build the described status variable from Task 3.

```
gen Status = 0 // All persons will first receive the missing code for "no info".
replace Status = 1 if migback == 1 & germborn == 1 // "no migback"
replace Status = 2 if migback == 3                 // "2nd generation" (2nd generation migrants born by definition in Germany, therefore "& germborn == 1" here unnecessary
replace Status = 3 if germborn == 2 & Escape == 0  // "Immigrants without information"
replace Status = 4 if germborn == 2 & Escape == 1  // "Immigrants, no escape"
replace Status = 5 if germborn == 2 & Escape == 2  // "Immigrant, escape"

label def Statuslbl 0"no info" 1"no migback" 2"2. Generation" 3"Immigrants without information"  4"Immigrants, no escape" 5"Immigrant, escape"
label val Status Statuslbl // Values of the status veriable receive label
```

Task 4: Content analysis:

a) How many refugees (foreign-born with refugee/asylum status) are now in your file?

Look at your status variable previously generated in task 3 to answer the question.

```
*** Exercise 4 ************************************************************

/*
a)        How many refugees (foreign-born with refugee/asylum status) are now in your file?
*/

tab Status, m //Display Generated Status Variable
```

```
. tab Status, m //Display Generated Status Variable
```

Status	Freq.	Percent	Cum.
no info	18	0.06	0.06
no migback	18,099	60.91	60.97
2. Generation	2,158	7.26	68.24
Immigrants without information	826	2.78	71.02
Immigrants, no escape	4,098	13.79	84.81
Immigrant, escape	4,514	15.19	100.00
Total	29,713	100.00	

All 4,514 respondents who received the value 5 for the generated status variable have a direct migration background (migback==2), were not born in Germany (germborn==2), and fled their country of origin (flight==2 and biimgrp==5).

b) How many are there if you take the individual extrapolation factors into account? Interpret the results.

Look at the status variable generated in task 3 to answer the question.

```
/*
b)        How many are there if you take the individual extrapolation factors into␣
↪account? Interpret the results.
*/

tab Status [aw=bgphrf], m  //Display generated status variable weighted with analytic␣
↪weights
```

```
. tab Status [aw=bgphrf], m  //Display generated status variable weighted with analytic weights
```

Status	Freq.	Percent	Cum.
no info	17.1538018	0.06	0.06
no migback	22,182.267	75.23	75.29
2. Generation	2,161.9832	7.33	82.63
Immigrants without information	622.927131	2.11	84.74
Immigrants, no escape	3,824.1688	12.97	97.71
Immigrant, escape	675.499938	2.29	100.00
Total	29,484	100.00	

After weighting, there are approximately 675 refugees in the dataset. The weighting thus corrected the number of refugees downwards.

c) How many persons are represented in the sample, taking the extrapolation factors into account?

To use frequency weights in STATA, integer weights are required. Create an integer frequency weight from the weighting factor provided so that you can make representative statements. Then take a look at the new results.

```
/*
c)        How many persons are represented when the sample taking the extrapolation␣
↪factors into account?
*/

```

(continues on next page)

(continued from previous page)

```
gen fweight = round(bgphrf) //Frequency weights for stata require integer weight
tab Status [fw=fweight], m  //Display generated status variable weighted with␣
↪frequency weights
```

```
. tab Status [fw=fweight], m  //Display generated status variable weighted with frequency weights
```

Status	Freq.	Percent	Cum.
no info	40,818	0.06	0.06
no migback	52,781,778	75.23	75.29
2. Generation	5,144,356	7.33	82.63
Immigrants without information	1,482,236	2.11	84.74
Immigrants, no escape	9,099,488	12.97	97.71
Immigrant, escape	1,607,336	2.29	100.00
Total	70,156,012	100.00	

Around 1,600,000 people are represented.

d) What is the proportion of people over 40 years of age among the refugees?

Since the data in this exercise come from the wave "bg", we are currently in the survey year 2016; if you need a description of the wave designations, please refer to the chapter Label. To generate a suitable age variable, you can use the year of birth (year of birth). If we look at the survey year 2016, all persons born in 1976 or earlier were over 40 years old. Generate a suitable age variable and look at the proportion of refugees over 40 years of age in weighted form:

```
/*
d)        What is the proportion of people over 40 years of age among the refugees?
*/

gen ue_40 = 0
replace ue_40 = 1 if gebjahr <= 1976 // Persons receive proficiency 1 if they were␣
↪born before 1975.

tab Status ue_40 [aw=bgphrf], m row nofreq
```

```
. tab Status ue_40 [aw=bgphrf], m row nofreq
```

Status	ue_40 0	1	Total
no info	57.54	42.46	100.00
no migback	28.83	71.17	100.00
2. Generation	59.22	40.78	100.00
Immigrants without in	8.91	91.09	100.00
Immigrants, no escape	37.10	62.90	100.00
Immigrant, escape	53.04	46.96	100.00
Total	32.28	67.72	100.00

The proportion of refugees over 40 years of age is about 47%.

Last change: Nov 12, 2019

6.5 Generating a Longitudinal Dataset

This example focuses on generating a dataset to analyze determinants of health satisfaction. You can either use the syntax generator in paneldata.org or write a syntax file yourself. You can search for variable names in Paneldata.org.

In the previous examples, you created an exercise path with four subfolders as well as corresponding globals in the STATA do-file. You can use the same folders and globals for this exercise.

1.Generate an unbalanced panel dataset for the years 2006 to 2008 using paneldata.org if you wish. The dataset should contain all respondents in private households:

The data set should contain the following variables of interest:

- health satisfaction **"wp0101" "xp0101" "yp0101"**
- currently smoking yes/no **"wp9301" "yp10601"**
- current employment status **"emplst06" "emplst07" "emplst08"**
- monthly household net income **"hinc06" "hinc07" "hinc08"**

In addition, the dataset should include the following additional information for analysis from 2006 to 2008:

- cross-sectional weighting factors for all relevant years **"wphrf" "xphrf" "yphrf"**
- individual identifier **"persnr"**
- original household number **"hhnr"**
- household number for all relevant years **"whhnr" "xhhnr" "yhhnr"**
- sample membership **"psample"**
- sex **"sex"**
- year of birth **"gebjahr"**
- population membership **"wpop" "xpop" "ypop"**

If you need detailed instructions on how the script generator works in paneldata.org, you can find them in the chapter *Syntax Generator on paneldata.org*.

If you would like to assemble your dataset yourself, you can do this with the datasets you have assembled. From the previous exercise with tracking data, you may already have an idea where to get most of the variables.

Since we want to have an unbalanced panel set, the $netto variable for the years 2006 to 2008 must also be used. In addition, our analysis must limit population membership, as we are only interested in household respondents.

Tip: If a dataset is created from several variables of different datasets, it is worth sorting the individual identifier before saving the individual data sets in order to be able to merge the data sets more easily afterwards.

1.1. Create a Master File

Use ppfad as the source file together with the required variables that you may have already found in Paneldata.org or identified from the variable label in the dataset. Note that only variables from the years to be analyzed should be used.

```

use hhnr persnr sex gebjahr psample xhhnr xnetto xpop yhhnr ynetto ypop whhnr wnetto␣
↪wpop using "${MY_PATH_IN}ppfad.dta"

```

Since we want to obtain an unbalanced data set, i.e., individuals who have completed an individual questionnaire at least once within the last three years, you must restrict the variable $netto (survey status). Also, we only want to analyze private households, so we need a further restriction of the $pop (sample membership) variable.

```

keep if ( (xnetto >= 10 & xnetto < 20) | (ynetto >= 10 & ynetto < 20) | (wnetto >= 10␣
↪& wnetto < 20) )


* * * PRIVATE VS ALL HOUSEHOLDS * * *

keep if ( (xpop == 1 | xpop == 2) | (ypop == 1 | ypop == 2) | (wpop == 1 | wpop == 2)␣
↪)

```

Then we sort the persnr (individual identifier) in the datasets and save it.

```

sort persnr
save "${MY_PATH_OUT}ppfad.dta", replace
clear

```

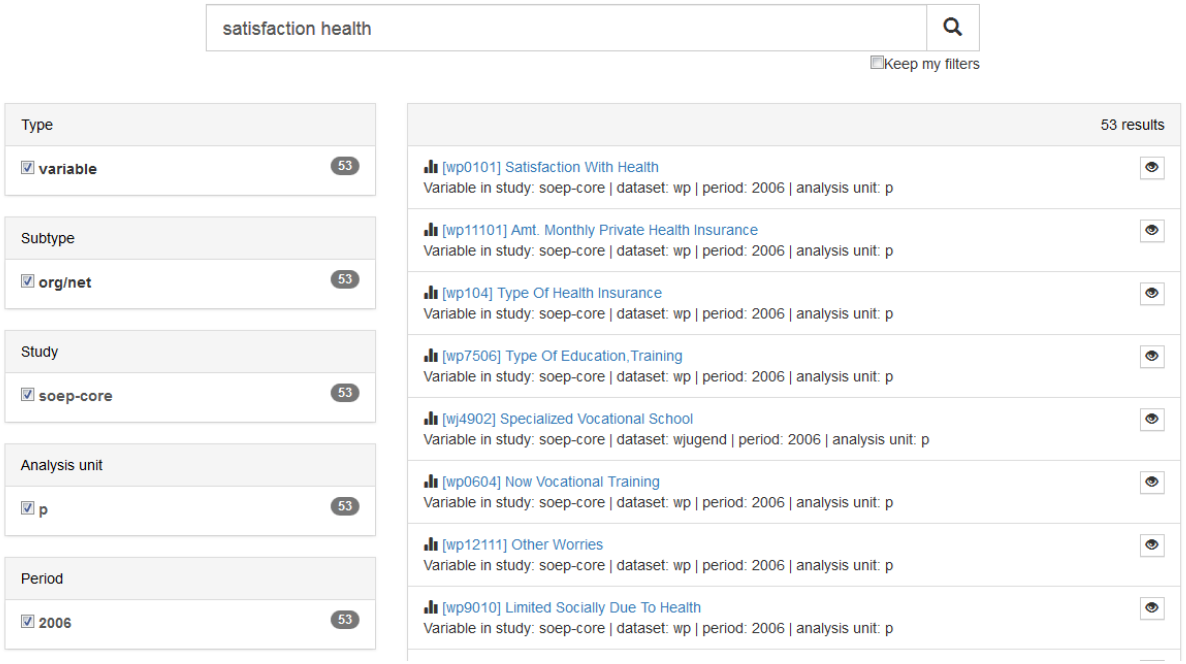
What is still missing is the cross-sectional weighting factor and the variables of interest for the analysis. To apply the weighting factors to the dataset, open the weighting dataset for the person-level phrf, sort it, and save it again.

```

use persnr wphrf xphrf yphrf using "${MY_PATH_IN}phrf.dta"
sort persnr
save "${MY_PATH_OUT}phrf.dta", replace
clear

```

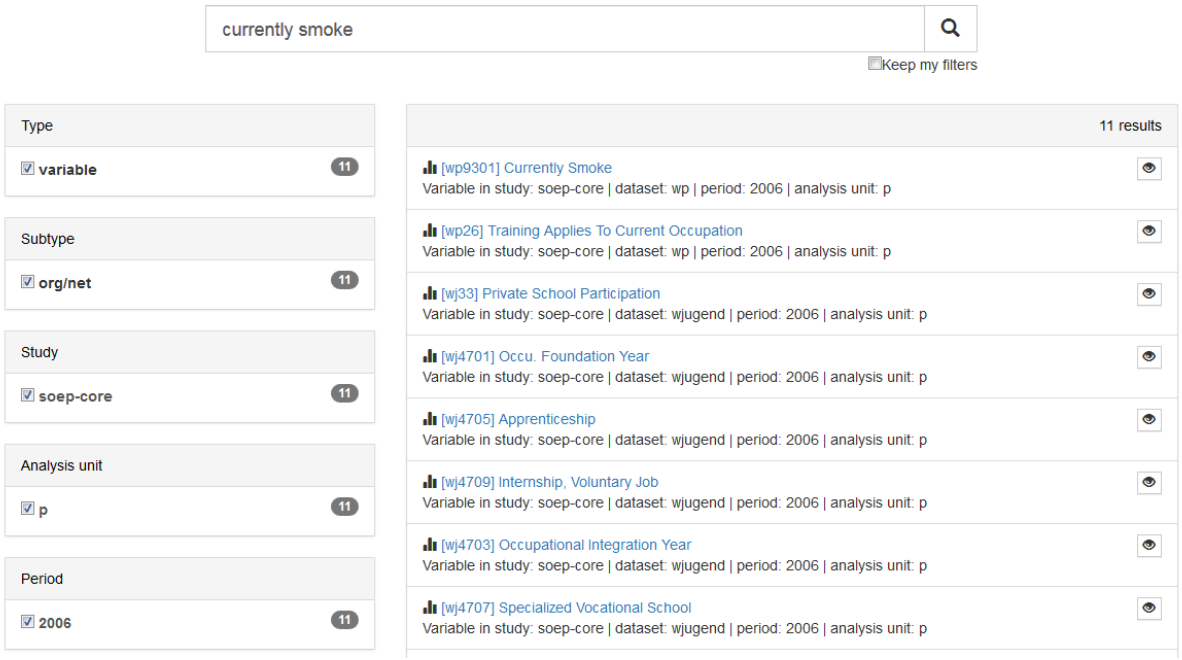
Now we come to the content variables. In order not to have to click through all of the datasets in the data release, it is recommended that the label be entered for the variable of interest from paneldata.org.

Use the filter to narrow your search. Select our main study SOEP-Core, the search type “variable”, the analysis unit “p” or “h” and the corresponding year. Once you have clicked on the year of interest, a variable history is displayed. You can use this to see which years the variable was collected and what the variable is called.

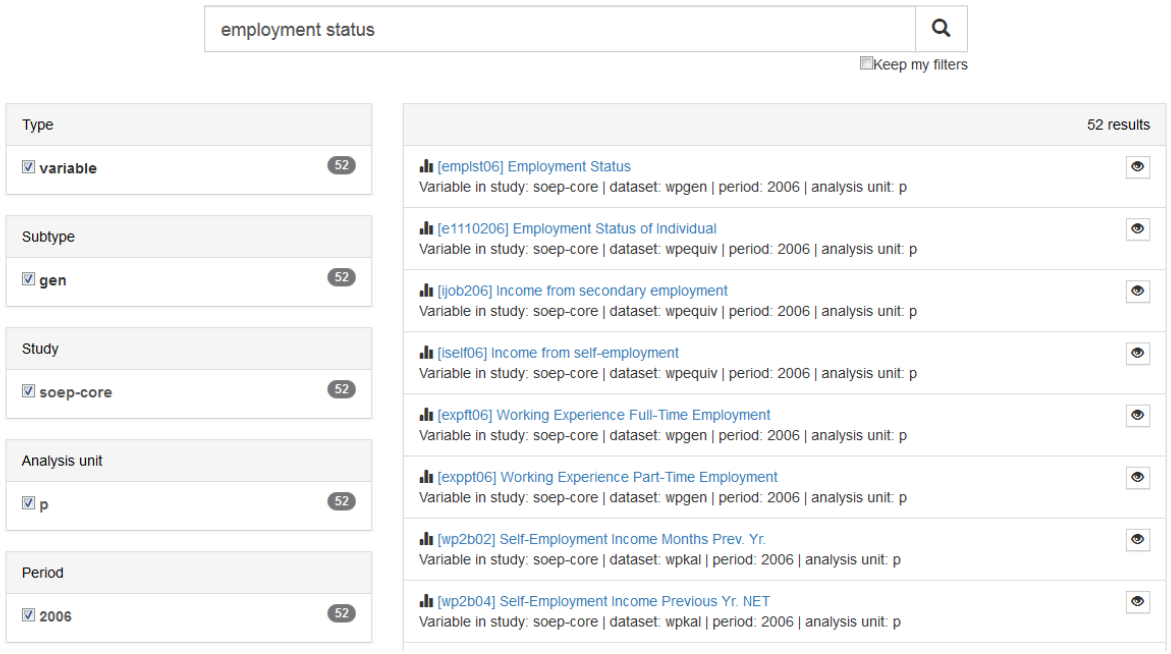
Example: Variable Label “Satisfaction Health”



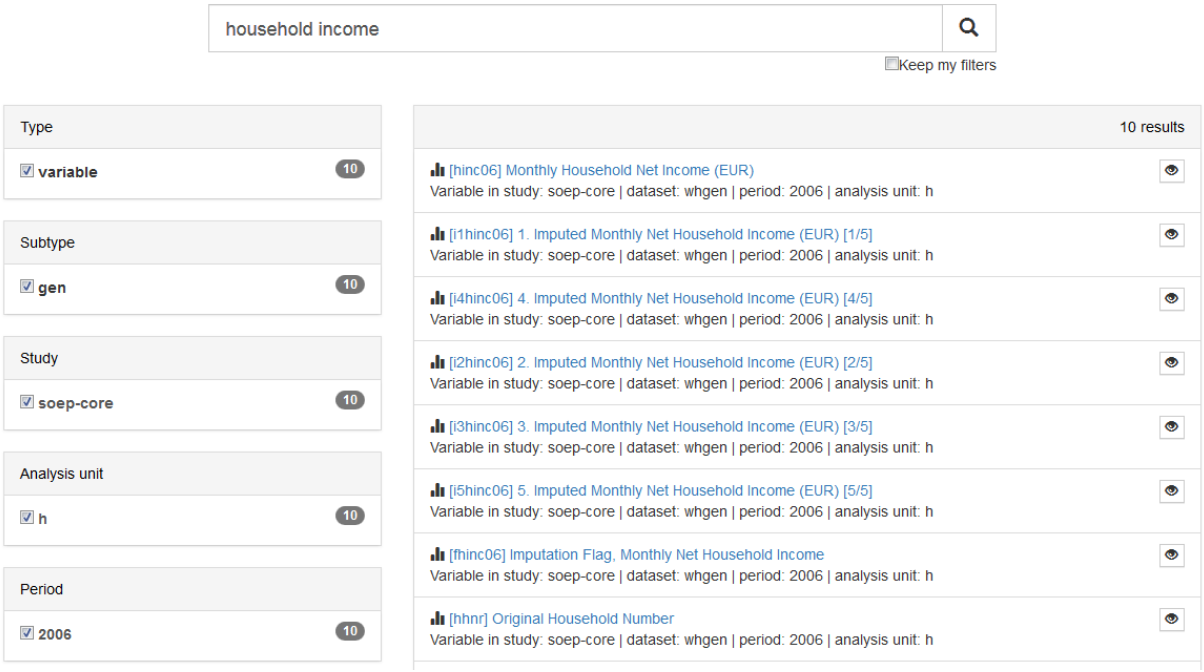
Example: Variable Label “currently smoking yes/no”



Example: Variable Label “current employment status”



Example: Variable Label “monthly net household income”



To merge the data, you can either use the script generator on paneldata.org or write the syntax manually into a do-file.

We now have all the information we need to create a master file. As already mentioned with **TIP!**, it is recommended to save the datasets sorted by the persnr (individual identifier) before merging.

```
use persnr wp0101 wp9301 using "${MY_PATH_IN}wp.dta"
sort persnr
save "${MY_PATH_OUT}wp.dta", replace
clear

* * * Persons 2007 * * *
use persnr xp0101 using "${MY_PATH_IN}xp.dta"
sort persnr
save "${MY_PATH_OUT}xp.dta", replace
clear

* * * Persons 2008 * * *
use persnr yp0101 yp10601 using "${MY_PATH_IN}yp.dta"
sort persnr
save "${MY_PATH_OUT}yp.dta", replace
clear

```

With the help of a unique identifier, which is either the household ($hhnr) or individual identifier (persnr), you can now merge all datasets or individual variables to ppfad. Which identifier to use when depends on the unit of analysis. Since we are on the individual level, our indicator is persnr (individual identifier).

We load the dataset ppfad and merge our datasets or variables to ppfad.

```

merge 1:1 persnr using "${MY_PATH_OUT}phrf.dta", keep(master match) nogen


* merge data from $p.dta
merge 1:1 persnr using "${MY_PATH_IN}/wp.dta", keepus(wp0101 wp9301)  keep(master␣
↪match) nogen // health & smoking
merge 1:1 persnr using "${MY_PATH_IN}/xp.dta", keepus(xp0101)                  ␣
↪keep(master match) nogen // health
merge 1:1 persnr using "${MY_PATH_IN}/yp.dta", keepus(yp0101 yp10601) keep(master␣
↪match) nogen // health & smoking

* merge data from $pgen.dta
local y = 6
foreach wave in w x y {
        merge 1:1 persnr using "${MY_PATH_IN}/`wave'pgen.dta", keepus(emplst0`y
↪')nogen keep(master match)
        local y = `y' + 1
}

* merge data from $hgen.dta
local y = 6
foreach wave in w x y {
        merge m:1 `wave'hhnr using "${MY_PATH_IN}/`wave'hgen.dta", keepus(hinc0`y')␣
↪nogen keep(master match)
        local y = `y' + 1
}

```

2. Encode missing values in system failings (STATA)!

After the master file has been created with all required information, the missing values, which can take between -1 to -8 in SOEP, must be recoded to missings. This step is important for converting a wide-format data set to a long format.

```
*******************************************************************************
*** Task 2) ***
* Encode missing values in systemmissings (STATA)!
*******************************************************************************

        mvdecode _all, mv(-1=. \ -2=.t \ -3=.x \ -5=.y \ -8=.z)
```

3. The data set is in "wide" format, i.e., additional years are displayed as additional variables (columns). For many analyses, it makes sense to convert datasets into the "long" format. In long format, additional years are displayed as additional lines. If the dataset covers three years, as in this example, there are three lines for each person. Convert the data set to long format using the STATA command reshape.!

Since these are cross-sectional variables, it can be assumed that each variable has at least one wave abbreviation, which makes the variable unique. Conversely, this means that the variables must be renamed before the reshape command.

Before renaming all original variables (e.g., from $P data sets) it must be checked whether the question and the answer categories were the same in all years (you can also look up the exact wording of the question in the corresponding questionnaire). If changes are made, the variables may have to be recoded.

```
*Check if original variable have changed over time
        tab1 wp0101 xp0101 yp0101
        tab1 wp9301 yp10601
        /*additionally check questionaires for exact wording*/
```

How you rename the variables is largely up to you. However, you should ensure that the name remains consistent over time and that the variable only differs according to the year (variable name + four-digit year suffix, e.g., zufr2006, zufr2007, zufr2008). You can rename the variables either manually, line by line, or for advanced users using a loop.

Example of manual renaming:

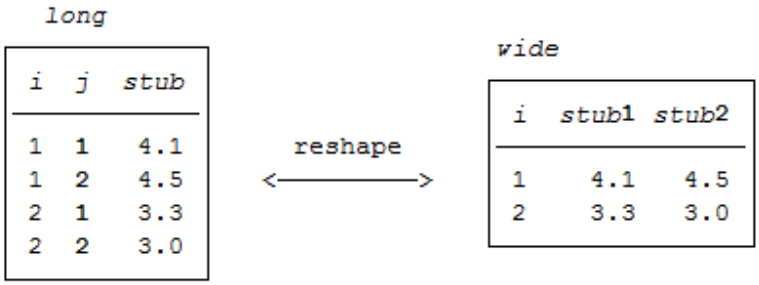
```
*rename time-variant variables
*with examples how to use loops (but can also be done "manually")
        rename wp9301 smoke2006
        rename yp10601 smoke2008
        rename wp0101 health2006
        rename xp0101 health2007
        rename yp0101 health2008
        ...
```

Example of a loop:

```
        foreach  x in 6 7 8 {
                rename hinc0`x' hinc200`x'
                rename emplst0`x' emplst200`x'
                }


        local y=2006
        foreach w in w x y {
                rename `w'hhnr hhnrakt`y'
                rename `w'netto netto`y'
                rename `w'pop pop`y'
                rename `w'phrf phrf`y'
                local y=`y'+1
                }
```

3.1. The reshape command

Now that we have made all relevant preparations, you can start to convert the dataset. If you want to convert a dataset, you can do this in both directions:



In our case, we reshape from wide to long. This means that a new variable name must be assigned for the year of the survey (j). The variable is then generated automatically. Currently, each person is assigned a line in Stata.

persnr	hhnr	wave	sex	smoke2006	smoke2008
12345	123	x	m	yes	yes
54321	211	x	m	no	no

```
*reshape dataset to long-format
        reshape long health smoke emplst hinc netto pop hhnrakt phrf, i(persnr) j(year)
        bys persnr: gen waves=_N                /*additional information: count number of waves per person*/
        tab waves
```

After the reshape command, you have one line per year for each person:

persnr	hhnr	wave	year	sex	smoke
12345	123	x	2006	m	yes
12345	123	y	2007	m	.
12345	123	z	2008	m	yes

4. Perform analyses based on the data. Try to answer the following questions:

a. Has men's and women's average satisfaction with health changed over the three years?

Satisfaction with health was measured on a scale from 1 to 10, with a value of 10 representing the highest possible level of satisfaction. To compare the average satisfaction with health between women and men, you should display the mean value for both sexes. The mean value is displayed weighted here.

```
*a) Has the average satisfaction with men's health and women changed
*   over the three years?

          mean health [pw=phrf], over(sex year)
```

```
.           mean health [pw=phrf], over(sex year)

Mean estimation                   Number of obs   =     30,765

         Over: sex year
    _subpop_1: [1] Male 2006
    _subpop_2: [1] Male 2007
    _subpop_3: [1] Male 2008
    _subpop_4: [2] Female 2006
    _subpop_5: [2] Female 2007
    _subpop_6: [2] Female 2008
```

Over	Mean	Std. Err.	[95% Conf.	Interval]
health				
_subpop_1	6.579	.0457144	6.489398	6.668602
_subpop_2	6.571889	.046199	6.481337	6.662441
_subpop_3	6.511273	.0488181	6.415588	6.606959
_subpop_4	6.475934	.0422708	6.393082	6.558787
_subpop_5	6.456594	.0429136	6.372482	6.540707
_subpop_6	6.421587	.0485101	6.326505	6.516668

The output shows the average values for men and women for all three years. The first three values show men's average satisfaction with health between 2006 and 2008, while the last three values show women's average satisfaction with health.

b. What is the proportion of people for whom health satisfaction has increased from 2006 to 2007?

To answer this question, the difference between 2006 and 2007 should be displayed. You should make sure that the analysis is conducted only within one persnr (individual identifier) and only for satisfaction in the following year.

```
*b) What is the proportion of people for whom health satisfaction has increased
*   from 2006 to 2007??
        sort persnr year
        gen diff=health-health[_n-1] if persnr==persnr[_n-1] & year==year[_n-1]+1
        tab diff if year==2007                                  /*unweighted*/
```

```
.         tab diff if year==2007                                    /*unweighted*/
```

diff	Freq.	Percent	Cum.
-10	3	0.03	0.03
-9	2	0.02	0.05
-8	14	0.14	0.19
-7	21	0.21	0.41
-6	43	0.44	0.84
-5	107	1.08	1.93
-4	202	2.05	3.97
-3	432	4.38	8.35
-2	841	8.52	16.88
-1	1,902	19.28	36.15
0	3,141	31.84	67.99
1	1,707	17.30	85.29
2	822	8.33	93.62
3	343	3.48	97.10
4	153	1.55	98.65
5	74	0.75	99.40
6	29	0.29	99.70
7	17	0.17	99.87
8	5	0.05	99.92
9	6	0.06	99.98
10	2	0.02	100.00
Total	9,866	100.00	

Since you have previously added the SOEP weighting factors to the dataset for your analysis, you should use the weighting for a representative analysis.

```
1       tab diff if year==2007 [aw=phrf]        /*weighted*/
```

```
.          tab diff if year==2007 [aw=phrf]        /*weighted*/
```

diff	Freq.	Percent	Cum.
-10	3.69881191	0.04	0.04
-9	1.514105677	0.02	0.05
-8	18.9326365	0.19	0.25
-7	17.065928	0.18	0.42
-6	37.1065342	0.38	0.80
-5	95.2821037	0.98	1.78
-4	198.375239	2.04	3.82
-3	479.45631	4.92	8.74
-2	819.914247	8.42	17.16
-1	1,853.9569	19.03	36.19
0	3,057.3252	31.39	67.58
1	1,617.6167	16.61	84.18
2	850.31852	8.73	92.91
3	358.524393	3.68	96.59
4	171.378275	1.76	98.35
5	92.2643934	0.95	99.30
6	32.9474818	0.34	99.64
7	21.31469291	0.22	99.86
8	3.08587415	0.03	99.89
9	9.23868822	0.09	99.98
10	1.68299548	0.02	100.00
Total	9,741	100.00	

The values less than 0 show a deterioration in health satisfaction. The value 0 means constant health satisfaction, and all values above 0 show a positive change in satisfaction with their health. With a value of 10, it can be assumed that these people were interviewed for the first time in 2007 or 2008.

c. In what direction and how much has satisfaction with health changed from 2006 to 2008 among people who quit smoking after 2006?

The procedure is similar to the previous question, except that the element "smoke yes/no" is added.

```
*c) In what direction and how much has satisfaction with
*   health changed from 2006 to 2008 among people who quit smoking after 2006?

        gen diff2=health-health[_n-2] if persnr==persnr[_n-2] & year==year[_n-2]+2 &␣
↪year==2008
        gen quit=.
        replace quit=0 if smoke==1 & smoke[_n-2]==1 & persnr==persnr[_n-2] &␣
↪year==year[_n-2]+2 & year==2008
        replace quit=1 if smoke==2 & smoke[_n-2]==1 & persnr==persnr[_n-2] &␣
↪year==year[_n-2]+2 & year==2008
        replace quit=2 if smoke==2 & smoke[_n-2]==2 & persnr==persnr[_n-2] &␣
↪year==year[_n-2]+2 & year==2008
        replace quit=3 if smoke==1 & smoke[_n-2]==2 & persnr==persnr[_n-2] &␣
↪year==year[_n-2]+2 & year==2008
```

(continues on next page)

(continued from previous page)

```
10          label define quit 0 "smoker" 1 "quit" 2 "non-smoker" 3 "begin"
11          label values quit quit
12          tabstat diff2, by(quit)
```

```
. tabstat diff2, by(quit)

Summary for variables: diff2
     by categories of: quit

      quit |      mean
-----------+----------
    smoker | -.1883657
      quit | -.2418953
non-smoker | -.1718027
     begin | -.0574713
-----------+----------
     Total | -.1755582
-----------------------
```

To obtain a weighted mean value, address the analysis weight after the generated variable.

```
1           tabstat diff2 [aw=phrf], by(quit)        /*weighted*/
```

```
. tabstat diff2 [aw=phrf], by(quit)       /*weighted*/

Summary for variables: diff2
     by categories of: quit

      quit |      mean
-----------+----------
    smoker | -.2351997
      quit | -.3483256
non-smoker | -.1747877
     begin | -.3205134
-----------+----------
     Total | -.2022029
-----------------------
```

This illustration shows the mean of the health variable under the condition of the quit variable that we generated beforehand. With a mean of -0.24 (weighted -0.35), the biggest change in health satisfaction is seen in people who quit smoking after 2006. For example, if a person smoked in 2006 and indicated a satisfaction value of 8, the person indicates a satisfaction value of 7.76 after he/she stopped smoking in 2008. So you can assume that when a person stops smoking, their perceived health state deteriorates. Now we have to test if the assumption is correct.

d. Does quitting smoking make your health worse? To what extent could the result of the analysis "stop smoking" be distorted?

In order to establish a connection between health satisfaction and stopping smoking, one should use the t-test or to be more specific, the one-sample t-test. It checks whether the mean value of a sample deviates significantly from a known expected value (specified in the null hypothesis).

```
*d) Does quitting smoking make your health worse? To what extent can the
*   result of the analysis "Stop smoking" be distorted?

        * Notes: So far we have not tested whether the difference is statistically␣
↪significant
                ttest diff2==0 if quit==1
```

```
. ttest diff2==0 if quit==1

One-sample t test

------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   diff2 |     401   -.2418953    .1069743    2.142158   -.4521973   -.0315932
------------------------------------------------------------------------------
    mean = mean(diff2)                                            t =  -2.2612
Ho: mean = 0                                     degrees of freedom =       400

    Ha: mean < 0                 Ha: mean != 0                 Ha: mean > 0
 Pr(T < t) = 0.0121         Pr(|T| > |t|) = 0.0243          Pr(T > t) = 0.9879
```

H0 Hypothesis: If one stops smoking, it has no effect on health.

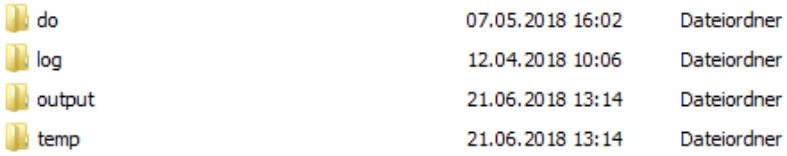
For this test we assume a 95% probability. What we want to check now is whether the H0 hypothesis can be rejected or not. If you look at the output of the test, you first see the mean value of 1 (quit smoking) of the variable quit. The last line of the output shows the significance level. If it falls below the value 0.05, one can speak of a statistically significant result. In our example, the null hypothesis can be discarded because its value is less than 0.05 percent. So quitting smoking has a significant impact on a person's perceived health.

Last change: Nov 12, 2019

6.6 Longitudinal Data Analysis

Simple cross-sectional analyses show that married people have higher life satisfaction than singles. You want to check this on the basis of longitudinal analysis with the SOEP.

Create an exercise path with four subfolders:



Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store your script, log files, datasets, and temporary datasets. Open an empty do-file and define the paths you created with globals:

```
*********************************************
* Set some useful commands
*********************************************
version 13
clear all
set more off
**increase buffer size
set scrollbufsize 2000000
**now restart stata!

*********************************************
* Set relative paths to the working directory
*********************************************
global AVZ         "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\distribution\soep-long\soep.v33.1\stata_en\"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global "AVZ" defines the main path. The main paths are subdivided using the globals "MY_IN_PATH", "MY_DO_FILES", "MY_LOG_OUT", "MY_OUT_DATA", "MY_OUT_TEMP". The global "MY_IN_PATH" contains the path to your ordered data.

Create a master file that uses the important variables from ppathl.

You should always add some variables from PPATHL to your dataset by default. Download the following information from PPATHL:

- Individual identifier **"pid"**
- Household identifier **"pid"**
- Survey year **"syear"**
- The net variable with information on the interview type **"netto"**
- The weighting variable **"phrf"**
- The gender of the person **"sex"**
- The migration background **"migback"**

```
*-------------------------------------------------------------------------------
*** Step 1) Start with basic information from PPFADL ***

use pid hid syear netto phrf migback sex using ${MY_IN_PATH}\ppfadl.dta
```

Attention: Please note that since version 34 (v34), PPFADL has been renamed PPATHL. The following exercises are done with version 33.1 (v33.1), where the tracking file was named PPFADL.

Search for matching variables and add them to your dataset

To perform your analysis, you need different SOEP variables. The SOEP offers various options for a variable search:

- Search the questionnaires for useful variables. (for more information, see the section *Variable Search with Questionnaires*)
- Find a suitable variable via the topic list of paneldata.org (for more information, see the section *Topic Search with paneldata.org*)
- Search for a suitable variable using a search term in paneldata.org (for more information, see the section *Variable Search with paneldata.org*)
- Use the documentation provided on the generated variables (for more information, see the section *Documentation on Generated Data*)

In this case, you need the variables **"pgfamstd"** (martial status) and **"plh0182"** (life satisfaction).

```
*------------------------------------------------------------------------------
*** Step 2) Add the relavant variables: here: family status and life satisfaction ***
merge 1:1 pid syear using ${MY_IN_PATH}\pgen, keepusing(pgfamstd) keep(1 3) nogen

                // merges family status from pgen
                // Documentation for PGEN can be found here
                // http://panel.gsoep.de/soep-docs/surveypapers/diw_ssp0307.pdf)


*describe using pl (directory)
                // for checking out variable names without opening the dataset

merge 1:1 pid syear using ${MY_IN_PATH}\pl, keepusing(plh0182) keep(1 3) nogen
                // merges life satisfaction from pl

save $MY_OUT_DATA\ppfad.dta, replace
```

Clean and inspect the data

Recode all missing values with commas to period decimal format.

```
*------------------------------------------------------------------------------
*** Step 3) Clean and inspect the data
mvdecode _all, mv(-8/-1)
```

Since you are interested in individual characteristics in your analysis: Delete all measurements that are not based on successful individual interviews.

```
tab netto
drop if netto>19
```

```
. tab netto
```

Current Wave Survey Status	Freq.	Percent	Cum.
[10] Interviewee With Succesful Intervi	514,447	52.79	52.79
[12] Individual Questionnaire And Perso	59,730	6.13	58.92
[13] Individual Questionnaire And Youth	318	0.03	58.95
[14] Individual Questionnaire And Other	32	0.00	58.96
[15] Individual Questionnaire And Exper	38,663	3.97	62.92
[16] Individual Questionnaire, First Ti	5,946	0.61	63.53
[17] Youth Biography First Time Surveye	4,859	0.50	64.03
[18] Individual Questionnaire And Child	8	0.00	64.03
[19] Individual Questionnaire Without H	538	0.06	64.09
[20] Children in Succesfully Interviewe	169,841	17.43	81.52
[21] Children With Mother-Child Questio	5,318	0.55	82.06
[22] Children With Mother-Child Questio	5,792	0.59	82.66
[23] Children With Mother-Child Questio	5,457	0.56	83.22
[24] Children age 7-8, with parental qu	4,875	0.50	83.72
[25] Children age 9-10, with parental q	4,097	0.42	84.14
[26] Students Age 11-12	1,759	0.18	84.32
[27] Children with Mother-Child Questio	2,186	0.22	84.54
[28] Youth questionnaire, Age 13-14	526	0.05	84.60
[29] Jugendliche 16-17 Jahre (ohne Juge	222	0.02	84.62
[30] Persons In Successfully Interviewe	128,343	13.17	97.79
[31] Successful Gap Interview (_LUECKE)	8,401	0.86	98.65
[32] Successfully Completed Biography Q	35	0.00	98.65
[33] Successful Youth Questionnaire	22	0.00	98.66
[34] Successful Tests and Experiments	122	0.01	98.67
[61] Gap Interview without HH reference	35	0.00	98.67
[62] Gap Interview with drop out	5	0.00	98.67
[80] Individual Without Any Current Inf	642	0.07	98.74
[81] Prior Interviewee Without Any Curr	359	0.04	98.78
[88] Repatriate - (moved abroad before	75	0.01	98.78
[89] Repatriate - (was drop out [90])	256	0.03	98.81
[90] Individual Dropouts PBR_EXIT	3,835	0.39	99.20
[91] Moved abroad	2,158	0.22	99.42
[92] Moved abroad (abroad)	177	0.02	99.44
[93] Moved abroad (exit)	65	0.01	99.45
[97] advice to dead person (exit)	981	0.10	99.55
[98] advice to dead person (_VP)	122	0.01	99.56
[99] Has Died	4,262	0.44	100.00
Total	974,509	100.00	

How many people contribute measurements and what is the proportion of people contributing at least 10 measurements?

Define the dataset as a panel dataset.

```
**define the dataset as panel data
```

(continues on next page)

(continued from previous page)

```
xtset pid syear
xtdes
```

```
. xtdes

     pid:  101, 102, ..., 38648901                          n =      86079
   syear:  1984, 1985, ..., 2016                            T =         33
           Delta(syear) = 1 unit
           Span(syear)  = 33 periods
           (pid*syear uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       2         4        10      25      33

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-----------------------------------
     5438      6.32    6.32 |  ................................1
     3320      3.86   10.17 |  ...........................111111
     2940      3.42   13.59 |  ..........................1111111
     2557      2.97   16.56 |  .............................1111
     2201      2.56   19.12 |  ................11111111111111111
     2049      2.38   21.50 |  ................1................
     1891      2.20   23.69 |  ..........................1......
     1774      2.06   25.76 |  ............................11111
     1740      2.02   27.78 |  .........................11......
    62169     72.22  100.00 | (other patterns)
 ---------------------------+-----------------------------------
    86079    100.00         |  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

86,079 respondents have contributed information in waves a (1984) to bg (2016) and 75% of the 86,079 respondents have provided information for at least 10 waves.

How many people took part in the survey in 2010 and contributed to continuous measurements up to 2014?

```
xtdes if syear>=2010 & syear<=2014
```

```
. xtdes if syear>=2010 & syear<=2014

     pid:  602, 901, ..., 35033302                          n =      45438
   syear:  2010, 2011, ..., 2014                            T =          5
           Delta(syear) = 1 unit
           Span(syear)  = 5 periods
           (pid*syear uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       2         3         5       5       5

     Freq.   Percent    Cum.  |  Pattern
-------------------------------+---------
    14673     32.29   32.29   |  11111
     4992     10.99   43.28   |  1....
     4342      9.56   52.83   |  .1111
     4234      9.32   62.15   |  ...11
     2669      5.87   68.03   |  11...
     2307      5.08   73.10   |  ..111
     1924      4.23   77.34   |  1111.
     1742      3.83   81.17   |  ...1.
     1548      3.41   84.58   |  111..
     7007     15.42  100.00   |  (other patterns)
-------------------------------+---------
    45438    100.00           |  XXXXX
```

14,673 respondents provided continuous information from 2010 to 2014.

Univariate inspection & analysis

How does the mean of life satisfaction change over time?

```
*------------------------------------------------------------------------------
*** Step 4) univariate inspection & analysis
table syear, content (mean plh0182)
```

```
. table syear, content (mean plh0182)
```

Survey Year	mean(plh0182)
1984	7.4257707595825195
1985	7.2370133399963379
1986	7.2855525016784668
1987	7.1372828483581543
1988	7.0825653076171875
1989	7.1014566421508789
1990	7.0492663383483887
1991	6.9480605125427246
1992	6.9156084060668945
1993	6.8846182823181152
1994	6.8577637672424316
1995	6.8879237174987793
1996	6.9003634452819824
1997	6.7927885055541992
1998	6.949559211730957
1999	6.9689054489135742
2000	7.0886578559875488
2001	7.1047582626342773
2002	7.0459656715393066
2003	6.9639754295349121
2004	6.800537109375
2005	6.9480514526367188
2006	6.9144678115844727
2007	6.9462895393371582
2008	6.9816727638244629
2009	6.9765110015869141
2010	7.2461948394775391
2011	7.1784853935241699
2012	7.1922345161437988
2013	7.3142080307006836
2014	7.2472319602966309
2015	7.3801255226135254
2016	7.3573770523071289

What proportion of people are a) married in 2014 or b) have a migration background? Compare weighted with unweighted frequency tables: Who is overrepresented in SOEP?

```
tab1 pgfamstd migback if syear==2014
tab pgfamstd [aw=phrf] if syear==2014
tab migback [aw=phrf] if syear==2014
```

```
. tab1 pgfamstd migback if syear==2014

-> tabulation of pgfamstd if syear==2014
```

Marital Status In Survey Year	Freq.	Percent	Cum.
[1] Married	16,157	57.82	57.82
[2] Married, But Separated	632	2.26	60.08
[3] Single	7,117	25.47	85.55
[4] Divorced	2,483	8.89	94.44
[5] Widowed	1,471	5.26	99.70
[6] husband/wife abroad	11	0.04	99.74
[7] Registered Same-Sex Partnership, Li	56	0.20	99.94
[8] Registered Same-Sex Partnership, Li	17	0.06	100.00
Total	27,944	100.00	

```
. tab pgfamstd [aw=phrf] if syear==2014
```

Marital Status In Survey Year	Freq.	Percent	Cum.
[1] Married	14,027.561	50.66	50.66
[2] Married, But Separated	634.611034	2.29	52.95
[3] Single	8,097.8889	29.24	82.19
[4] Divorced	2,617.4229	9.45	91.65
[5] Widowed	2,212.929	7.99	99.64
[6] husband/wife abroad	20.7802588	0.08	99.71
[7] Registered Same-Sex Partnership, Li	53.2891395	0.19	99.90
[8] Registered Same-Sex Partnership, Li	26.518149	0.10	100.00
Total	27,691	100.00	

The data show that married people are overrepresented in the SOEP and single people are underrepresented. The weighting makes it representative again for Germany.

```
-> tabulation of migback if syear==2014
```

Migration background	Freq.	Percent	Cum.
[1] no migration background	20,363	72.62	72.62
[2] direct migration background	5,190	18.51	91.12
[3] indirect migration background	2,489	8.88	100.00
Total	28,042	100.00	

```
. tab migback [aw=phrf] if syear==2014
```

Migration background	Freq.	Percent	Cum.
[1] no migration background	21,324.466	76.75	76.75
[2] direct migration background	4,464.8327	16.07	92.81
[3] indirect migration background	1,996.7017	7.19	100.00
Total	27,786	100.00	

In the SOEP sample, respondents with a direct or indirect migration background are overrepresented.

How many of those persons who reported a life satisfaction scale value of 7 in one survey year also indicated the scale value of 7 in the following survey year?

```
xttrans plh0182
```

```
. xttrans plh0182
```

Current Life Satisfaction	Current Life Satisfaction											
	0	1	2	3	4	5	6	7	8	9	10	Total
0	20.30	8.31	10.61	11.19	7.47	19.50	5.71	5.84	6.37	1.95	2.74	100.00
1	8.61	10.60	15.58	13.55	9.53	17.08	6.58	6.77	6.58	3.39	1.74	100.00
2	3.77	5.18	14.47	16.75	11.29	19.24	8.82	8.79	7.98	2.43	1.26	100.00
3	1.86	2.45	7.79	16.11	14.66	23.04	11.64	11.24	8.34	2.00	0.87	100.00
4	0.89	1.24	4.19	10.55	15.47	26.02	15.54	14.43	8.92	1.94	0.81	100.00
5	0.75	0.66	2.06	5.10	7.86	32.32	16.97	17.50	12.70	2.46	1.60	100.00
6	0.24	0.32	1.07	2.81	4.98	18.20	22.66	27.74	17.53	3.08	1.37	100.00
7	0.13	0.14	0.54	1.53	2.42	9.53	14.20	34.57	29.86	5.42	1.66	100.00
8	0.10	0.11	0.36	0.79	1.11	5.20	6.57	21.77	46.31	14.03	3.65	100.00
9	0.10	0.12	0.25	0.45	0.63	2.69	3.15	10.34	36.80	36.06	9.40	100.00
10	0.29	0.13	0.30	0.61	0.68	4.09	2.90	7.51	23.21	23.28	37.01	100.00
Total	0.44	0.43	1.23	2.58	3.48	11.67	10.92	21.61	30.36	12.03	5.25	100.00

34.57% of the respondents who reported a life satisfaction of 7 again reported a value of 7 in the following year.

Is it more likely that a highly dissatisfied person (value: 0) will be less dissatisfied the following year or that a very satisfied (value: 10) person will be less satisfied the following year?

```
xttrans plh0182
```

```
. xttrans plh0182
```

Current Life Satisfaction	Current Life Satisfaction											
	0	1	2	3	4	5	6	7	8	9	10	Total
0	20.30	8.31	10.61	11.19	7.47	19.50	5.71	5.84	6.37	1.95	2.74	100.00
1	8.61	10.60	15.58	13.55	9.53	17.08	6.58	6.77	6.58	3.39	1.74	100.00
2	3.77	5.18	14.47	16.75	11.29	19.24	8.82	8.79	7.98	2.43	1.26	100.00
3	1.86	2.45	7.79	16.11	14.66	23.04	11.64	11.24	8.34	2.00	0.87	100.00
4	0.89	1.24	4.19	10.55	15.47	26.02	15.54	14.43	8.92	1.94	0.81	100.00
5	0.75	0.66	2.06	5.10	7.86	32.32	16.97	17.50	12.70	2.46	1.60	100.00
6	0.24	0.32	1.07	2.81	4.98	18.20	22.66	27.74	17.53	3.08	1.37	100.00
7	0.13	0.14	0.54	1.53	2.42	9.53	14.20	34.57	29.86	5.42	1.66	100.00
8	0.10	0.11	0.36	0.79	1.11	5.20	6.57	21.77	46.31	14.03	3.65	100.00
9	0.10	0.12	0.25	0.45	0.63	2.69	3.15	10.34	36.80	36.06	9.40	100.00
10	0.29	0.13	0.30	0.61	0.68	4.09	2.90	7.51	23.21	23.28	37.01	100.00
Total	0.44	0.43	1.23	2.58	3.48	11.67	10.92	21.61	30.36	12.03	5.25	100.00

The rows reflect the initial values, and the columns reflect the final values. Around 20% of those who were completely dissatisfied (value: 0) in the base year remained completely dissatisfied in the following year. About 80% of these completely dissatisfied people from the base year were more satisfied in the following year. Of the completely satisfied persons (value: 10), about 37% remained just as satisfied in the following year, but 63% became less satisfied. It is more likely that a completely dissatisfied person will become more satisfied in the following year than that a completely satisfied person will become less satisfied.

Which transitions in marital status can be observed particularly frequently in the data?

```
xttrans pgfamstd
```

```
. xttrans pgfamstd
```

Marital Status In Survey Year	Marital Status In Survey Year								
	1	2	3	4	5	6	7	8	Total
1	98.49	0.90	0.00	0.10	0.50	0.01	0.00	0.00	100.00
2	4.09	74.86	0.00	18.55	1.43	1.07	0.00	0.00	100.00
3	4.09	0.15	95.63	0.02	0.00	0.06	0.04	0.01	100.00
4	4.08	0.25	0.00	95.62	0.00	0.00	0.03	0.01	100.00
5	0.36	0.07	0.00	0.00	99.57	0.00	0.00	0.00	100.00
6	12.44	25.84	0.00	0.16	0.00	61.56	0.00	0.00	100.00
7	0.00	0.00	0.00	0.32	0.00	0.00	95.82	3.86	100.00
8	0.00	0.00	0.00	3.92	1.96	0.00	5.88	88.24	100.00
Total	62.00	2.17	22.53	6.83	6.27	0.11	0.07	0.01	100.00

Survey respondents who were married but separated in the base year and reported divorce as their family status in the following year can be observed particularly frequently. (about 19%).

Simple cross sectional analyses

You now want to find the correlation between marital status and life satisfaction. Is there an effect of marriage on life satisfaction? And if so, is it a sustained effect?

First, calculate the correlation between family status and life satisfaction in from a cross-sectional perspective for 2010: Are married people happier than singles?

```
*-------------------------------------------------------------------------------
*** Step 5)simple cross-sectional analyses
table pgfamstd if syear==2010, content (mean plh0182)
```

```
. table pgfamstd if syear==2010, content (mean plh0182)
```

Marital Status In Survey Year	mean(plh0182)
[1] Married	7.394993782043457
[2] Married, But Separated	6.7182130813598633
[3] Single	7.2009811401367187
[4] Divorced	6.7114768028259277
[5] Widowed	6.7760229110717773
[6] husband/wife abroad	7.6666665077209473
[7] Registered Same-Sex Partnership, Liv	7.1500000953674316
[8] Registered Same-Sex Partnership, Liv	7

At first glance, married couples seem happier than singles.

Now generate a variable that indicates a transition from "single" to "married".

How many such transitions can you find in the data?

```
***perform longitudinal analysis
**define event: transition to marriage
generate to_mar=1 if pgfamstd==1 & l.pgfamstd==3
tab to_mar
```

```
. tab to_mar
```

to_mar	Freq.	Percent	Cum.
1	4,834	100.00	100.00
Total	4,834	100.00	

A total of 4,834 people can be observed changing status from single to married.

What is the average level of life satisfaction immediately after the transition to marriage (i.e., in the first survey in which the transition can be observed) and how high is life satisfaction immediately before the transition to marriage?

```
**standard way of life-event analysis
sum plh0182 if to_mar==1
sum l.plh0182 if to_mar==1

**alternative way
generate dif_sat= plh0182- l.plh0182
mean dif_sat if to_mar==1
```

```
. sum plh0182 if to_mar==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
plh0182	4,824	7.650498	1.522432	0	10

```
. sum l.plh0182 if to_mar==1
```

Variable	Obs	Mean	Std. Dev.	Min	Max
plh0182 L1.	4,804	7.543922	1.544923	0	10

```
. mean dif_sat if to_mar==1

Mean estimation                 Number of obs    =       4,794
```

	Mean	Std. Err.	[95% Conf.	Interval]
dif_sat	.1072174	.0227754	.0625672	.1518675

Before the transition to marriage, the average life satisfaction of the respondents is 7.54. In the following year, that is, after the transition to marriage, the average life satisfaction of the respondents is 7.65. It can be seen that with the transition to marriage, average life satisfaction rises slightly by 0.11.

Map the complete satisfaction history around the "marriage entry" event [3 years before; 3 years after].

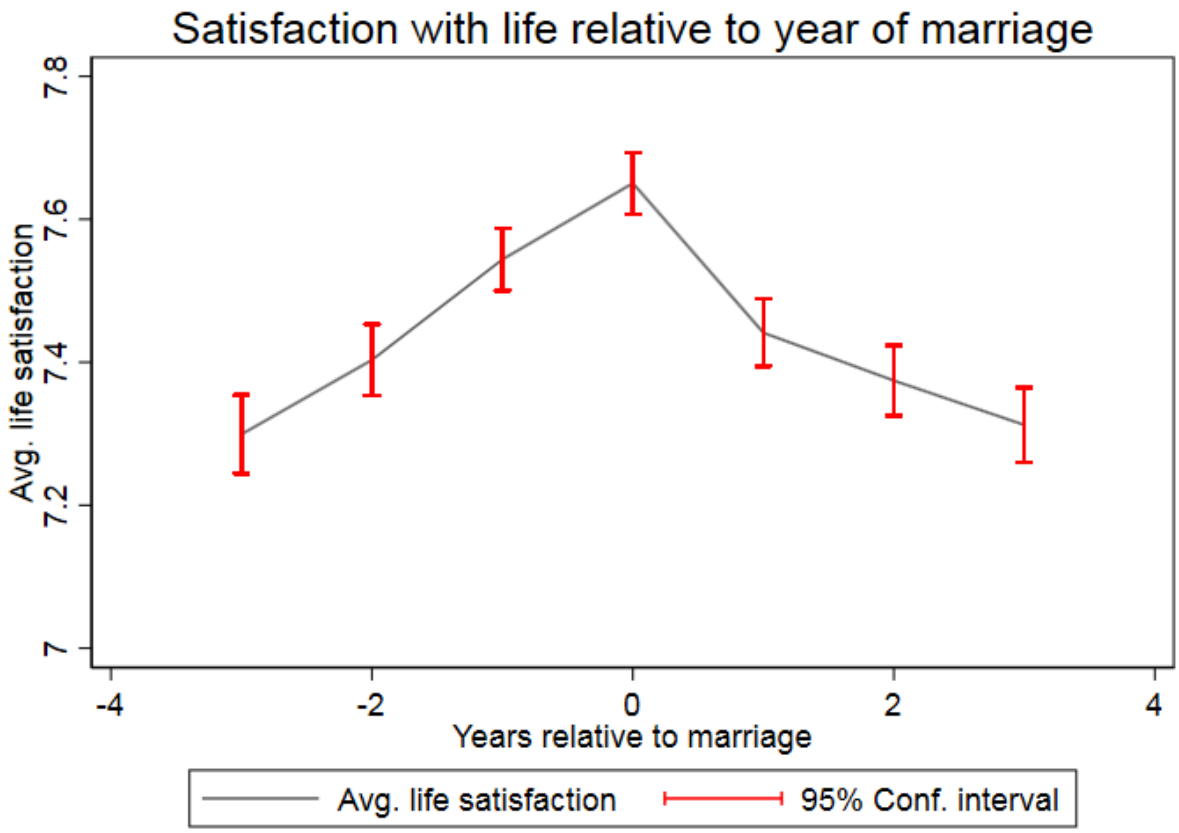
```
**preparing illustration of trajectory
generate t=0 if to_mar==1 & l.to_mar~=1 &l2.to_mar~=1 & l3.to_mar~=1 & l4.to_mar~=1 &
↪l5.to_mar~=1 & l6.to_mar~=1 & l7.to_mar~=1 & l8.to_mar~=1 & l9.to_mar~=1 & l10.to_
↪mar~=1 & l11.to_mar~=1 & l12.to_mar~=1 & l13.to_mar~=1 & l14.to_mar~=1
replace t=1 if l.t==0
replace t=2 if l2.t==0
replace t=3 if l3.t==0
replace t=-1 if f.t==0
replace t=-2 if f2.t==0
replace t=-3 if f3.t==0

table t, content (mean plh0182 n plh0182)
```

```
. table t, content (mean plh0182 n plh0182)
```

t	mean(plh0182)	N(plh0182)
-3	7.2992987632751465	3,281
-2	7.4033288955688477	3,905
-1	7.543921947479248	4,804
0	7.6504974365234375	4,824
1	7.4413299560546875	4,210
2	7.374445915222168	3,835
3	7.3124275207519531	3,444

Choose a suitable presentation for your results and let Stata create a graphic.

```
** Preparing graph of event analysis
sort t
cap drop meanplh0182
by t: egen meanplh0182 = mean(plh0182)

cap drop upper
gen upper = .
forval i = -3/3{
        su plh0182 if t == `i'
        replace upper = r(mean) + 1.96 * r(sd)/sqrt(r(N)) if t == `i'
}

cap drop lower
gen lower = .
forval i = -3/3{
        su plh0182 if t == `i'
        replace lower = r(mean) - 1.96 * r(sd)/sqrt(r(N)) if t == `i'
}

twoway (line meanplh0182 t) (rcap upper lower t, lcolor("red")) , title("Satisfaction␣
↪with life relative to year of marriage") legend(label(1 "Avg. life satisfaction")␣
↪label(2 "95% Conf. interval")) scheme(s1mono) xtitle("Years relative to marriage")␣
↪ytitle("Avg. life satisfaction")
```



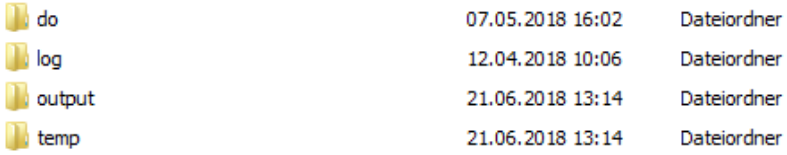
The graph shows that a positive effect on life satisfaction can be observed when family status changes from single to married. In the following years of the existing marriage, life satisfaction decreases again and approaches the initial satisfaction before the marriage.

Last change: Nov 12, 2019

6.7 Fixed Effects Estimation

Let's say you want to find out whether certain variables relevant to the labor market, such as work experience or time in education, influence a person's hourly wage. Other variables such as gender or marital status should also be taken into account. You decide to use the SOEP data to set up a fixed effects estimation model.

Create a path with four subfolders:



Example:

- H:/material/exercises/do

- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store your script, log files, datasets, and temporary datasets. Open an empty do-file and define your paths with globals:

```
*************************************************
* Set relative paths to the working directory
*************************************************
global AVZ         "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\distribution\soep-long\soep.v33.1\stata_en\"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global "AVZ" defines the main path. The main paths are subdivided using the globals "MY_IN_PATH", "MY_DO_FILES", "MY_LOG_OUT", "MY_OUT_DATA", "MY_OUT_TEMP". The global "MY_IN_PATH" contains the path to your data.

a) Generate your own SOEPwage.dta dataset. The dataset should contain information on gross monthly wage, marital status, and other personal characteristics.

To perform your analysis, you need different SOEP variables. The SOEP offers various options for a variable search:

- Search the questionnaires for useful variables. (For more information, see the section *Variable Search with Questionnaires*)
- Find a suitable variable in the topic list at paneldata.org (for more information, see the section *Topic Search with paneldata.org*)
- Search for a suitable variable using a search term in paneldata.org (for more information, see the section *Variable Search with paneldata.org*)
- Use the documentation provided for the generated variables (for more information, seethe section *Documentation on Generated Data*)

Use the various important variables of the ppfadl.dta dataset as your start file. Your source file should contain the following variables:

- Individual identifier **"pid"**
- Survey year **"syear"**
- Birth Year **"gebjahr"**
- The net variable with information on the interview type **"netto"**
- The weighting variable **"phrf"**
- The gender of the person **"sex"**
- Sample membership **"pop"**

```
use pid syear sex gebjahr netto pop phrf using "${MY_IN_PATH}/ppfadl.dta", clear
```

Attention: Please note that since version 34 (v34), PPFADL has been renamed PPATHL. The following ecxercises are done with version 33.1 (v33.1), where the tracking file was named PPFADL.

Apply the necessary content variables to your starting dataset. You need the following variables for your analysis:

- Employment status plb0022_h
- Current gross income in euros **"pglabgro"**
- Actual weekly working hours **"pgtatzeit"**
- Full-time work experience **"pgexpft"**
- Years of education or training **"pgbilzeit"**
- Marital status in survey year **"pgfamstd"**

```
merge 1:1 pid syear using "${MY_IN_PATH}/pl.dta", keepus(plb0022) keep(master match)␣
↪nogen
merge 1:1 pid syear using "${MY_IN_PATH}/pgen.dta", keepus(pglabgro pgtatzeit pgexpft␣
↪pgbilzeit pgfamstd) keep(master match) nogen
```

Only keep people who have completed an interview and who live in a private household.

```
* Only select people with completed interviews
keep if inrange(netto, 10, 19)

* Only private households
keep if pop==1 | pop==2
```

Since you are only interested in the period from 2012 to 2016, remove all survey information that does not fall within this period. To finish, save your dataset.

```
* Period from 2012 to 2016
keep if syear>=2012 & syear<=2016
```

Exercise 1: Prepare your dataset

a) Load your created SOEPWage.dta dataset. It contains information on gross monthly wage, marital status, and other personal characteristics.

```
*** Exercise 1: Prepare your dataset
* a) Load data set
use "${MY_OUT_DATA}/SOEPWage.dta", clear
```

b) Recode all missing values in systemmissings (.)

```
* b) Recode Missings
mvdecode _all, mv(-8/-1 = .)
```

For more information about the missing codes for SOEP data, see the chapter *Missing Conventions*

c) Generate the variables "hourly wage" (gross monthly wage/4.33*working time) for persons who have earned at least 1 euro and have worked at least one hour, "Married vs. Unmarried" and age.

```
* c) Generate Variables
gen wage = pglabgro/(4.33*pgtatzeit) if pglabgro>=1 & pgtatzeit>=1

gen married = 1 if pgfamstd==1 | pgfamstd==6 | pgfamstd==7 | pgfamstd==8
replace married = 0 if inrange(pgfamstd, 2, 5)

gen age = syear - gebjahr
```

d) Adjust the variable "hourly wage" from outlier values by setting values smaller than the first percentile to the same value. Set values greater than 3 times the 99th percentile to 3*99th percentile. Then generate the variable lwage = log(wage).

```
* d) Adjust wage variable
sum wage, detail
replace wage = 1/3*r(p1) if wage<1/3*r(p1)
replace wage = 3*r(p99) if wage>3*r(p99) & wage<.

gen lwage = log(wage)
label variable lwage "Log hourly wage"

save "${MY_OUT_DATA}/SOEPWage_temp.dta", replace
```

Exercise 2: Descriptive statistics

a) Define the dataset as a panel dataset.

```
*** Exercise 2: Descriptive statistics
* a)
xtset pid syear // Declaring data as panel data
```

b) What percentage of people participated in all five waves (xtdescribe)

```
* b)
xtdescribe, patterns(16) // -> unbalanced panel
```

```
. * b)
. xtdescribe, patterns(16) // -> unbalanciertes Panel

      pid:  602, 901, ..., 38647702                          n =      42808
    syear:  2012, 2013, ..., 2016                            T =          5
            Delta(syear) = 1 unit
            Span(syear)  = 5 periods
            (pid*syear uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       2         4         5       5       5

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
    17069     39.87   39.87 |  11111
     3941      9.21   49.08 |  ....1
     3044      7.11   56.19 |  1....
     2810      6.56   62.75 |  .1111
     2581      6.03   68.78 |  11...
     2040      4.77   73.55 |  1111.
     1895      4.43   77.98 |  111..
     1695      3.96   81.94 |  ...11
     1688      3.94   85.88 |  .1...
      925      2.16   88.04 |  .11..
      923      2.16   90.20 |  ...1.
      678      1.58   91.78 |  ..111
      671      1.57   93.35 |  .111.
      425      0.99   94.34 |  11.11
      402      0.94   95.28 |  111.1
      289      0.68   95.95 |  1.111
     1732      4.05  100.00 | (other patterns)
 ---------------------------+---------
    42808    100.00         |  XXXXX
```

42808 respondents have contributed information within waves bc (2012) - bg (2016) and about 40% (17069) of the 42808 respondents have provided information for all waves.

c) Describe the variable "Married" with xttab and xttrans. Take a look at some individual wage (pid=30320901, pid=30932501, pid==3101602, pid==3101801) developments with xtline.

```
* c)
* Stability of the relationship status
xttab married
```

```
. xttab married
```

married	Overall Freq.	Overall Percent	Between Freq.	Between Percent	Within Percent
0	58906	41.37	19717	46.23	94.69
1	83474	58.63	25014	58.65	95.88
Total	142380	100.00	44731	104.87	95.35
			(n = 42652)		

You can observe 41.37 percent of person-year observations with "married==no". Within the period from 2012 to 2016, 19717 people responded at least once that they were not married. During the same period, 25014 persons reported at least once that they were married. Those who were not married for at least one year responded with "married==no" in 94.69% of the observations, whereas those who were married at least once responded in 95.88 percent of the observations with "married==yes". This indicates very stable response behavior.

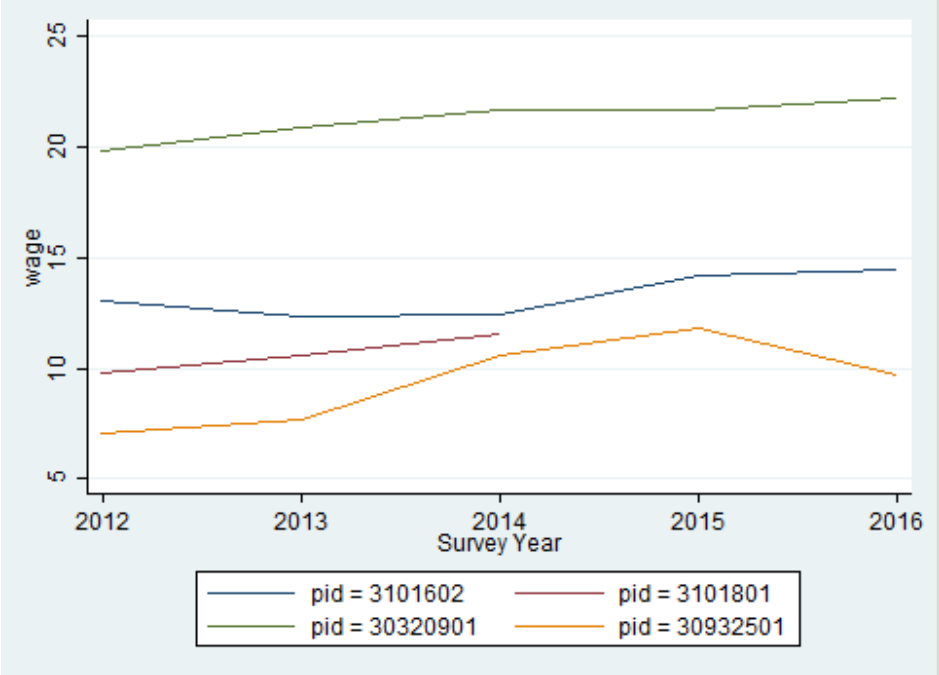
```
* Transition probabilities
xttrans married, freq
```

```
. xttrans married, freq
```

married	married 0	married 1	Total
0	39,112	1,264	40,376
	96.87	3.13	100.00
1	881	58,428	59,309
	1.49	98.51	100.00
Total	39,993	59,692	99,685
	40.12	59.88	100.00

96.87 percent of the person-year observations with "married==no" are still not married in the next period. 98.51 percent of the persons who are married indicate that they will also be married in the following period. This is evidence of stable response behavior.

```
* Individual sequences of "wage"
xtline wage if pid==30320901 | pid==30932501 | pid==3101602 | pid==3101801, overlay
```



The graphic shows a comparison of the hourly wage for four different respondents.

Exercise 3: Pooled OLS Regression

a) Execute a pooled OLS regression with "log hourly wage" as dependent variable and "married", "gender", "work experience" and "training time" as independent variables. Interpret the coefficients for "married", "gender" and "length of training". Why are these not causal effects?

```
*** Exercise 3: Pooled OLS Regression
* a) Pooled OLS
reg lwage married sex pgexpft pgbilzeit
```

```
. reg lwage married sex pgexpft pgbilzeit
```

Source	SS	df	MS	Number of obs =	78234
Model	9531.59732	4	2382.89933	F(4, 78229) =	8027.72
Residual	23221.0303	78229	.296834042	Prob > F =	0.0000
				R-squared =	0.2910
				Adj R-squared =	0.2910
Total	32752.6276	78233	.418654885	Root MSE =	.54482

lwage	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
married	.1443034	.0041241	34.99	0.000	.1362203	.1523865
sex	-.1203015	.0041704	-28.85	0.000	-.1284754	-.1121276
pgexpft	.0143396	.0001791	80.08	0.000	.0139886	.0146906
pgbilzeit	.0988842	.0007078	139.71	0.000	.0974969	.1002714
_cons	1.19645	.0121292	98.64	0.000	1.172677	1.220224

The variables married, sex, and pgbilzeit most likely correlate with other disregarded/unobserved variables that have an effect on the wage. For example, women more often work in occupations with lower wages.

b) Run the regression again with the option "vce(cluster persnr)" to get clustered standard errors. How do the standard errors of the coefficients change?

```
* b) Pooled OLS with cluster standard errors
reg lwage married sex pgexpft pgbilzeit, vce(cluster pid)
```

```
. reg lwage married sex pgexpft pgbilzeit, vce(cluster pid)

Linear regression                               Number of obs =     78234
                                                F(  4, 25133) =   2415.06
                                                Prob > F      =    0.0000
                                                R-squared     =    0.2910
                                                Root MSE      =    .54482

                                   (Std. Err. adjusted for 25134 clusters in pid)
```

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
married	.1443034	.0066788	21.61	0.000	.1312126	.1573941
sex	-.1203015	.0070382	-17.09	0.000	-.1340967	-.1065063
pgexpft	.0143396	.0003257	44.03	0.000	.0137013	.014978
pgbilzeit	.0988842	.0012169	81.26	0.000	.096499	.1012693
_cons	1.19645	.0211759	56.50	0.000	1.154944	1.237956

The standard errors are getting bigger.

Exercise 4: Fixed Effects

a) Subtract the person-specific mean value from each variable of the model. Use the "egen" function. Ideally you should also use a loop.

```
*** Exercise 4: Fixed Effects
* a) Subtract person-specific averages

gen sample = 1
foreach var in lwage married sex pgexpft pgbilzeit {

        bysort pid: egen `var'Mean = mean(`var')
        replace `var'Mean = . if `var'==.
        gen `var'Demeaned =  `var' - `var'Mean
        replace sample = 0 if `var'==.
}
bysort pid (sample): replace sample = sample[1]
```

b) Estimate the fixed effects model with the previously generated variables. Why isn't a coefficient estimated for "gender"? How do the coefficients change compared to the pooled OLS estimate? Is the effect of "married" now causally interpretable?

```
reg lwageDemeaned marriedDemeaned sexDemeaned pgexpftDemeaned pgbilzeitDemeaned, vce(cluster pid) nocons
```

```
. * b) Fixed Effects Modell
. reg lwageDemeaned marriedDemeaned sexDemeaned pgexpftDemeaned pgbilzeitDemeaned, vce(cluster pid) nocons
note: sexDemeaned omitted because of collinearity

Linear regression                               Number of obs =   78234
                                                F(  3, 25133) =  645.95
                                                Prob > F      =  0.0000
                                                R-squared     =  0.0369
                                                Root MSE      =  .24298

                                  (Std. Err. adjusted for 25134 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
lwageDemeaned|      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
marriedDemeaned|   .0197547   .0098598     2.00   0.045     .0004289    .0390805
   sexDemeaned|          0  (omitted)
pgexpftDemeaned|   .0435521   .0010848    40.15   0.000     .0414259    .0456783
pgbilzeitDemeaned|   .0660986   .0042643    15.50   0.000     .0577404    .0744568
------------------------------------------------------------------------------
```

No coefficient was estimated for gender because gender was stable over time for all observations. The coefficient of married is now significant at the 5% level!

c) Now estimate the fixed effects model using the command "xtreg lwage married sex pgexpft pgbilzeit, fe". What do you notice about the coefficients compared to task 4 b)? And with the standard errors?

```
* c) xtreg, fe
xtreg lwage married pgexpft pgbilzeit, fe vce(cluster pid)
```

```
. xtreg lwage married pgexpft pgbilzeit, fe vce(cluster pid)

Fixed-effects (within) regression               Number of obs      =     78234
Group variable: pid                             Number of groups   =     25134

R-sq:  within  = 0.0394                         Obs per group: min =         1
       between = 0.2228                                        avg =       3.1
       overall = 0.1957                                        max =         5

                                                F(3,25133)         =    643.92
corr(u_i, Xb)  = -0.4631                        Prob > F           =    0.0000

                                    (Std. Err. adjusted for 25134 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |   .0224308   .0108103     2.07   0.038      .001242    .0436196
     pgexpft |   .0443073    .001109    39.95   0.000     .0421335    .0464811
   pgbilzeit |   .0765046   .0049426    15.48   0.000     .0668168    .0861924
       _cons |   .9253963   .0644086    14.37   0.000     .7991517    1.051641
-------------+----------------------------------------------------------------
     sigma_u |  .62923787
     sigma_e |  .29340975
         rho |   .8214025   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

The coefficients are not identical to 4 b) and the standard errors become larger because model b) does not take into account the estimation of mean values in the standard errors.

d) Now add dummy variables for the years (i.syear). What happens to the effect of "labor market experience"?

```
* d) xtreg with dummy
xtreg lwage married pgexpft pgbilzeit i.syear, fe vce(cluster pid)
```

```
. * d) xtreg mit Jahres-Dummys
. xtreg lwage married pgexpft pgbilzeit i.syear, fe vce(cluster pid)

Fixed-effects (within) regression               Number of obs      =     78234
Group variable: pid                             Number of groups   =     25134

R-sq:  within  = 0.0599                         Obs per group: min =         1
       between = 0.0065                                        avg =       3.1
       overall = 0.0152                                        max =         5

                                                F(7,25133)         =    344.67
corr(u_i, Xb)  = -0.2578                        Prob > F           =    0.0000

                                  (Std. Err. adjusted for 25134 clusters in pid)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     married |    .021538   .0106165     2.03   0.042     .0007292    .0423469
     pgexpft |  -.0124634   .0024322    -5.12   0.000    -.0172306   -.0076961
   pgbilzeit |   .0606128   .0048847    12.41   0.000     .0510384    .0701872
             |
       syear |
        2013 |   .0552667   .0036671    15.07   0.000     .0480789    .0624545
        2014 |   .0980733   .0047304    20.73   0.000     .0888014    .1073451
        2015 |   .1545752   .0063392    24.38   0.000       .14215    .1670005
        2016 |   .2026541   .0080508    25.17   0.000     .1868742    .2184341
             |
       _cons |   1.882517   .0712664    26.42   0.000     1.742831    2.022203
-------------+----------------------------------------------------------------
     sigma_u |  .66907886
     sigma_e |  .29027579
         rho |   .8415946   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

Effects on the variables remain significant. The model could possibly be specified on a case-by-case basis. The Mincer equation is based on (potential) labor market experience squared.

e) Now you can also square labor market experience into the model. To what extent does the effect of labor market experience change compared to task 5d)?

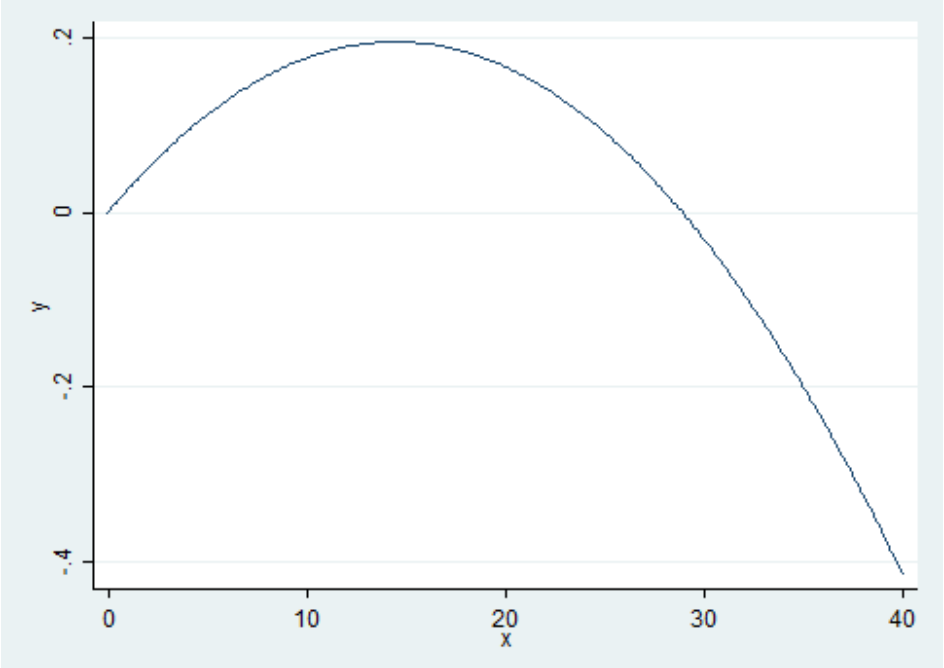
```
* e) expft squared
xtreg lwage married c.pgexpft##c.pgexpft pgbilzeit i.syear, fe vce(cluster pid)
```

```
. * e) expft auch als Quadrat
. xtreg lwage married c.pgexpft##c.pgexpft pgbilzeit i.syear, fe vce(cluster pid)

Fixed-effects (within) regression               Number of obs      =     78234
Group variable: pid                             Number of groups   =     25134

R-sq:  within  = 0.0648                         Obs per group: min =         1
       between = 0.0776                                        avg =       3.1
       overall = 0.0811                                        max =         5

                                                F(8,25133)         =    321.03
corr(u_i, Xb)  = -0.1012                        Prob > F           =    0.0000

                                          (Std. Err. adjusted for 25134 clusters in pid)
```

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
married	.0117953	.0106245	1.11	0.267	-.0090293	.0326199
pgexpft	.027049	.0035366	7.65	0.000	.0201171	.0339809
c.pgexpft#c.pgexpft	-.0009356	.0000582	-16.07	0.000	-.0010497	-.0008215
pgbilzeit	.0564758	.004831	11.69	0.000	.0470068	.0659449
syear						
2013	.0543771	.0036633	14.84	0.000	.0471967	.0615575
2014	.0971777	.0047248	20.57	0.000	.0879167	.1064386
2015	.1519717	.0063321	24.00	0.000	.1395605	.1643829
2016	.1980514	.0080426	24.63	0.000	.1822874	.2138155
_cons	1.692927	.0723071	23.41	0.000	1.551201	1.834653
sigma_u	.62325551					
sigma_e	.28951511					
rho	.82251756	(fraction of variance due to u_i)				

The coefficients of pgexpft and pgexpft^2 remain significant, whereas the coefficient for married is no longer significant.

```
graph twoway (func y = _b[pgexpft]*x + _b[c.pgexpft#c.pgexpft]*x*x, range(0 40))
```



The graph shows that the effects of the labor market experience decrease after approximately 15 years of professional experience.

f) Now estimate the model from task 5e) with longitudinal section weights. Why is the number of cases now significantly smaller? Why could the coefficient of "pgbilzeit" have changed?

Tip: Create your own longitudinal person weights, 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

---

Since you are looking at the period 2012-2016, you must create a suitable longitudinal weight. To do this, use the phrf dataset from the RAW subdirectory. Apply the required variables to your analysis dataset and generate your period-related longitudinal section weight. To understand the structure of the data distribution file and the location of the different datasets, visit the section *Data Distribution File*. For more information about the weighting datasets and other survey datasets, see the section *Survey Data*.

```
* f) Fixed Effects weighted
global MY_IN_PATH2 "\\hume\rdc-prod\complete\soep-core\soep.v33.2\stata_en\"
rename pid persnr
merge m:1 persnr using "${MY_IN_PATH2}/phrf.dta", nogen keep(master match)␣
↪keepus(bcphrf bdpbleib bepbleib bfpbleib bgpbleib)
gen wlong = bcphrf*bdpbleib*bepbleib*bfpbleib*bgpbleib
label variable wlong "Weighting BC-BG"
rename persnr pid
```

Now estimate the model from 5e) and use the created weight.

```
xtreg lwage married c.pgexpft##c.pgexpft pgbilzeit i.syear [pw=wlong], fe vce(cluster␣
↪pid)
```

```
. xtreg lwage married c.pgexpft##c.pgexpft pgbilzeit i.syear [pw=wlong], fe vce(cluster pid)

Fixed-effects (within) regression               Number of obs      =     48949
Group variable: pid                             Number of groups   =     11790

R-sq:  within  = 0.0880                         Obs per group: min =         1
       between = 0.1275                                        avg =       4.2
       overall = 0.1290                                        max =         5

                                                F(8,11789)         =     96.01
corr(u_i, Xb)  = -0.3604                        Prob > F           =    0.0000

                                    (Std. Err. adjusted for 11790 clusters in pid)
```

lwage	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
married	.0050783	.0180717	0.28	0.779	-.0303453	.0405018
pgexpft	.0237458	.0067916	3.50	0.000	.0104331	.0370584
c.pgexpft#c.pgexpft	-.0008416	.0000986	-8.54	0.000	-.0010348	-.0006484
pgbilzeit	.1392754	.0176388	7.90	0.000	.1047005	.1738503
syear						
2013	.0471116	.0076671	6.14	0.000	.0320828	.0621404
2014	.0962616	.0098515	9.77	0.000	.0769511	.1155721
2015	.1490648	.013773	10.82	0.000	.1220674	.1760623
2016	.1960915	.0171793	11.41	0.000	.1624172	.2297658
_cons	.6993781	.2279552	3.07	0.002	.2525483	1.146208
sigma_u	.63332729					
sigma_e	.29092777					
rho	.8257534	(fraction of variance due to u_i)				

The number of observations is now much smaller. The effect of pgbilzeit is greater than before. Pgbilzeit has a lower effect in the wlong==0 group, where the return is different for each additional educational year. People in the wlong===0 group may not get the returns on additional education they expected on the local labor market and may therefore move -> higher dropout probability.

Last change: Nov 12, 2019

6.8 Working with harmonized Variables

This exercise shows you how to work effectively with versioned and harmonized SOEP variables. Please note that the new SOEP versioning and harmonizing concept has only been available since SOEP-Core v34 and only applies to the original SOEP-Core data in long format.

Create an exercise path with four subfolders:

do	07.05.2018 16:02	Dateiordner
log	12.04.2018 10:06	Dateiordner
output	21.06.2018 13:14	Dateiordner
temp	21.06.2018 13:14	Dateiordner

Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store your script, log files, datasets, and temporary datasets. Open an empty do-file and define your paths with globals:

```
*********************************************
* Set relative paths to the working directory
*********************************************
global AVZ          "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-gen\consolidated\soep-long\soep.v34"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global "AVZ" defines the main path. The main paths are subdivided using the globals "MY_IN_PATH", "MY_DO_FILES", "MY_LOG_OUT", "MY_OUT_DATA", "MY_OUT_TEMP". The global "MY_IN_PATH" contains the path to your ordered data.

1.) Differences in Response Options

Variables are versioned and harmonized because the response options have changed over time.

Is this a job creation scheme (ABM) or structural adjustment measure (SAM)?

Yes ..... ☐ No ..... ☐

42. Is it an "ABM" Job (created through the government employment program) or a "1 Euro Job" (for non-profit work)?

Yes, an ABM job (government employment program) .... ☐
Yes, a 1 Euro job (non-profit work) ..... ☐
No ..... ☐

The variable plb0038_v1 was obtained from a simple yes/no question between 1992 and 2004. Since 2005, new response options have been added. The individual questionnaires from 2004 and 2005 show these differences. Through the versioning of the variable plb0038, this difference is recognizable to the data user when tabulating the variable. The variable label also shows the beginning and end of the period in which the question was asked differently.

```
use "$MY_IN_PATH\pl.dta"
tab plb0038_v1
tab plb0038_v2
```

```
. tab plb0038_v1
```

Job Creation Measure Job (1992-2004)	Freq.	Percent	Cum.
[-8] Question this year not part of the	440,893	67.61	67.61
[-2] Does not apply	191,409	29.35	96.96
[-1] No answer / don't know	927	0.14	97.10
[1] Yes	1,060	0.16	97.26
[2] No	17,852	2.74	100.00
Total	652,141	100.00	

Job Creation Measure Job (2005-2014)	Freq.	Percent	Cum.
[-8] Question this year not part of the	405,647	62.18	62.18
[-6] Version of questionnaire with modi	462	0.07	62.26
[-2] Does not apply	210,244	32.23	94.48
[-1] No answer / don't know	1,211	0.19	94.67
[1] Yes, Job Creation Measure	276	0.04	94.71
[2] Yes, Community Service	368	0.06	94.77
[3] No	34,121	5.23	100.00
Total	652,329	100.00	

The variable plb0038_v1 is recoded during the harmonization process and written into a new variable, plb0038_h, together with plb0038_v2. The harmonized version of the variable should cover the survey period from 1992 to 2014 and should be usable.

```
tab plb0038_h
tabstat plb0038_v1 plb0038_v2 plb0038_h, by(syear)
```

Job Creation Measure Job (harmonized)	Freq.	Percent	Cum.
[-8] Question this year not part of the	194,399	29.80	29.80
[-6] Version of questionnaire with modi	462	0.07	29.87
[-2] Does not apply	401,653	61.57	91.44
[-1] No answer / don't know	2,138	0.33	91.77
[1] Yes, Job Creation Measure	1,336	0.20	91.98
[2] Yes, Community Service	368	0.06	92.03
[3] No	51,973	7.97	100.00
Total	652,329	100.00	

```
. tabstat plb0038_v1 plb0038_v2 plb0038_h, by(syear)

Summary statistics: mean
  by categories of: syear (Erhebungsjahr (SurveyYear)
```

syear	plb003..	plb003..	plb~38_h
1984	-8	-8	-8
1985	-8	-8	-8
1986	-8	-8	-8
1987	-8	-8	-8
1988	-8	-8	-8
1989	-8	-8	-8
1990	-8	-8	-8
1991	-8	-8	-8
1992	-1.333284	-8	-1.17489
1993	-1.326277	-8	-1.163973
1994	-1.38004	-8	-1.231945
1995	-1.419015	-8	-1.280506
1996	-8	-8	-8
1997	-1.73387	-8	-1.67417
1998	-1.741104	-8	-1.681663
1999	-1.713454	-8	-1.64984
2000	-1.731038	-8	-1.667521
2001	-1.732585	-8	-1.669634
2002	-1.744852	-8	-1.684078
2003	-1.748795	-8	-1.688559
2004	-1.743176	-8	-1.681684
2005	-8	-1.277754	-1.277754
2006	-8	-1.255703	-1.255703
2007	-8	-1.238581	-1.238581
2008	-8	-1.245834	-1.245834
2009	-8	-1.238217	-1.238217
2010	-8	-1.224139	-1.224139
2011	-8	-1.325479	-1.325479
2012	-8	-1.264446	-1.264446
2013	-8	-1.433777	-1.433777
2014	-8	-1.432951	-1.432951
2015	-8	-8	-8
2016	-8	-8	-8
2017	-8	-8	-8
Total	-5.941218	-5.466984	-3.380835

2.) Differences in Coding of Response Options

Variables are versioned and harmonized because the coding of the response options has changed over time. Since the

values of certain response options can change, the various wave-specific variables cannot be integrated easily into a variable in long format. The variable must be appropriately harmonized to be useable.

21. What type of an employment change was that?

☞ *In the case that you have changed positions several times, please pick the appropriate reason for the most recent change.*

I have entered employment for the first time in my life ..... ☐■ *Skip to question **24!***

I have started up with paid employment again after not having been employed for a while ..... ☐

I have started a new position with a different employer (for temporary workers this includes working in an temporary workplace) ..... ☐

I have become self-employed ..... ☐

I have been taken on by the company in which I did my apprenticeship / worked as part of a state employment program / was employed on a free-lance basis ..... ☐

I have changed positions within the same company ..... ☐

26. What type of an employment change was that?

☞ *In the case that you have changed positions several times, please pick the appropriate reason for the most recent change.*

I have entered employment for the first time in my life ..... ☐➡ *Skip to question **29!***

I have started up with paid employment again after not having been employed for a while ..... ☐

I have started a new position with a different employer (for temporary workers this includes working in an temporary workplace) ..... ☐

I have been taken on by the company in which I did my apprenticeship / worked as part of a state employment program / was employed on a free-lance basis ..... ☐

I have changed positions within the same company ..... ☐

I have become self-employed ..... ☐⇨ **Did you receive funds from any government programs to start your own business?**

From 1994 to 2004, the question about “job change” was asked in the individual questionnaire as a category question with six response options. The order of the response options changed in 2005.

```
tab plb0284_v1
tab plb0284_v2
```

```
. tab plb0284_v1
```

Type Of Job Change (1994-2004)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	453,958	69.61	69.61
[-2] Does not apply	176,060	27.00	96.61
[-1] no answer	262	0.04	96.65
[1] First Time Employed	3,036	0.47	97.11
[2] Job After Break	6,571	1.01	98.12
[3] Job With New Employer	8,852	1.36	99.48
[4] New Job-Self Employed	1,319	0.20	99.68
[5] Company Taken Over	451	0.07	99.75
[6] Changed Job, Same Firm	1,632	0.25	100.00
Total	652,141	100.00	

```
. tab plb0284_v2
```

Type Of Job Change (2005-2017)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	316,675	48.56	48.56
[-5] Not Included In Questionnaire Vers	15,139	2.32	50.88
[-2] Does not apply	281,276	43.13	94.01
[-1] no answer	115	0.02	94.03
[1] First Job	4,342	0.67	94.70
[2] Returned to Past Employer After Bre	6,539	1.00	95.70
[3] New Position Different Employer	21,628	3.32	99.01
[4] Taken On Ba Company	1,519	0.23	99.25
[5] Changed Position Within Company	2,664	0.41	99.66
[6] New Job Self-Employed	2,244	0.34	100.00
Total	652,141	100.00	

In addition to the different order of the response options, the coding order also changed. The data are stored in the wave-specific "raw" datasets with different coding and are contained in the variables plb0284_v1 and plb0284_v2. To use the variable for all survey years, it is necessary to harmonize the different versions. The variable plb0284_v1 is recoded (recode (1=1)(2=2)(3=3)(4=6)(5=4)(6=5)) and then written together with plb0284_v2 as plb0284_h. The new variable plb0284_h is created by the harmonization process.

```
tab plb0284_h
tabstat plb0284_v1 plb0284_v2 plb0284_h, by(syear)
```

```
. tab plb0284_h
```

Type Of Job Change (harmonized)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	118,492	18.17	18.17
[-5] Not Included In Questionnaire Vers	15,139	2.32	20.49
[-2] Does not apply	457,336	70.13	90.62
[-1] no answer	377	0.06	90.68
[1] First Job	7,378	1.13	91.81
[2] Returned to Past Employer After Bre	13,110	2.01	93.82
[3] New Position Different Employer	30,480	4.67	98.49
[4] Taken On Ba Company	1,970	0.30	98.79
[5] Changed Position Within Company	4,296	0.66	99.45
[6] New Job Self-Employed	3,563	0.55	100.00
Total	652,141	100.00	

```
. tabstat plb0284_v1 plb0284_v2 plb0284_h, by(syear)

Summary statistics: mean
  by categories of: syear (Erhebungsjahr (SurveyYear))
```

syear	p~284_v1	p~284_v2	plb028~h
1984	-8	-8	-8
1985	-8	-8	-8
1986	-8	-8	-8
1987	-8	-8	-8
1988	-8	-8	-8
1989	-8	-8	-8
1990	-8	-8	-8
1991	-8	-8	-8
1992	-8	-8	-8
1993	-8	-8	-8
1994	-1.499068	-8	-1.496609
1995	-1.430709	-8	-1.426859
1996	-1.503738	-8	-1.500259
1997	-1.474215	-8	-1.471806
1998	-1.471506	-8	-1.471097
1999	-1.396592	-8	-1.397586
2000	-1.432495	-8	-1.428792
2001	-1.416805	-8	-1.413583
2002	-1.483425	-8	-1.480119
2003	-1.533944	-8	-1.531953
2004	-1.560425	-8	-1.555384
2005	-8	-1.564937	-1.564937
2006	-8	-1.548081	-1.548081
2007	-8	-1.492387	-1.492387
2008	-8	-1.480898	-1.480898
2009	-8	-1.473259	-1.473259
2010	-8	-1.394461	-1.394461
2011	-8	-1.482337	-1.482337
2012	-8	-1.418576	-1.418576
2013	-8	-1.937847	-1.937847
2014	-8	-1.311269	-1.311269
2015	-8	-1.540338	-1.540338
2016	-8	-1.865138	-1.865138
2017	-8	-1.635675	-1.635675
Total	-6.017658	-4.686554	-2.703361

3.) Content Differences in the Questions.

Variables are versioned when questions were asked differently in different years but the content belongs together. If

the content or wording of the question changes, the wave-specific variables cannot easily be integrated into a long variable.

108. Have you yourself ever inherited something or received a gift of great value?
We are referring to gifts or inheritance of house or land, securities, investments, other forms of wealth or large amounts of money.

Yes ..... ☐ ⇩ No ..... ☐➡ *Skip to question **109!***

155. Have you personally received an inheritance or larger endowment in the last 15 years?
We are referring mainly to transfers of home or property ownership, securities, participating interests, and other assets or larger sums of money.

Yes..... ☐ ⇩ No ..... ☐➡ *Question **157!***

In the 2001 individual questionnaire, respondents were asked whether they had ever received an inheritance. In 2017, this question was worded differently: respondents were asked whether they had received an inheritance in the last 15 years. The questions are similar but cover different time periods. Therefore, the variable is not harmonized but made available as versioned variables. Data users have to decide whether or not to use the variables in the same way.

```
tab plc0375_v1
tab plc0375_v2
```

```
. tab plc0375_v1
```

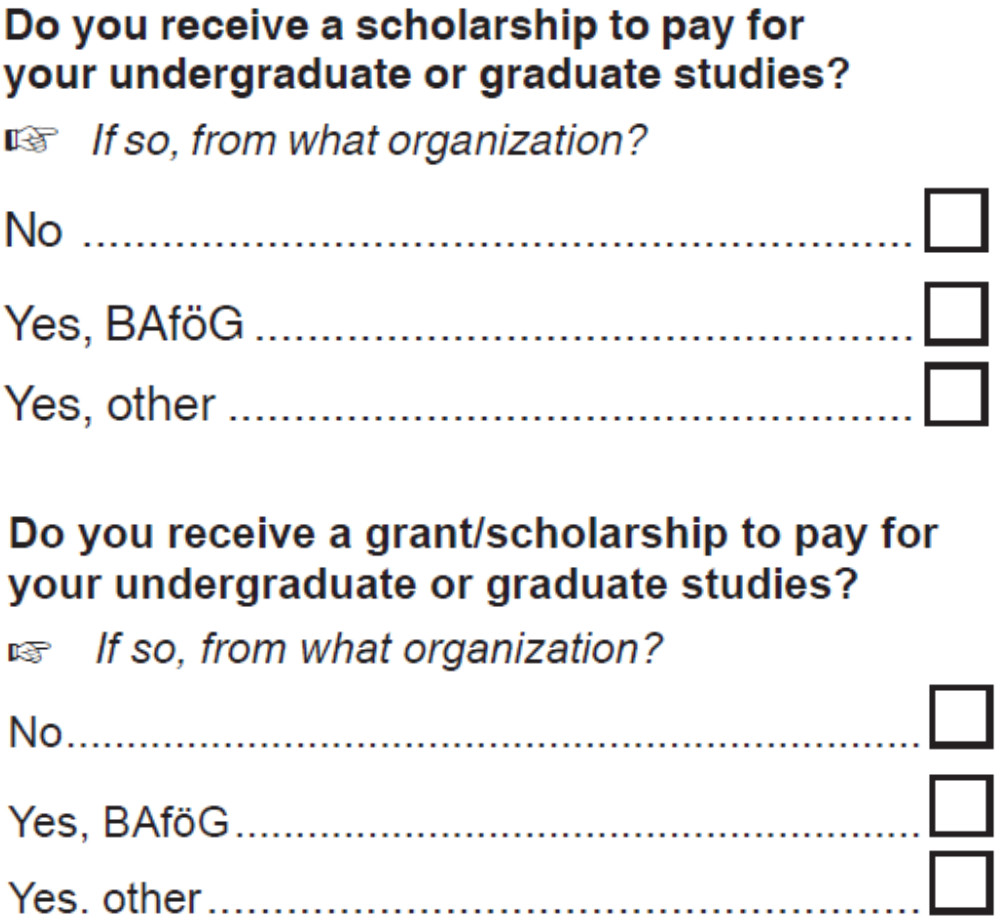
Erbschaft (jemals) (2001)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	629,790	96.57	96.57
[-1] no answer	140	0.02	96.59
[1] Yes	3,307	0.51	97.10
[2] No	18,904	2.90	100.00
Total	652,141	100.00	

```
. tab plc0375_v2
```

Erbschaft (letzte 15 Jahre) (2017)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	619,656	95.02	95.02
[-5] Not Included In Questionnaire Vers	5,703	0.87	95.89
[-1] no answer	187	0.03	95.92
[1] Yes	2,683	0.41	96.33
[2] No	23,912	3.67	100.00
Total	652,141	100.00	

4.) Change of Question Type.

Variables are versioned and harmonized when questions were asked differently in different years, for example, first as a question with multiple response options and later as a question with a single response option. The possibility to provide multiple answers in certain years makes it difficult to integrate the wave-specific variables into a variable in long format.



When comparing the question on scholarships in the individual questionnaires from 2011 and 2012, it appears that there should be no differences in the variables. Nevertheless, the two questions seem to have been asked differently and stored differently in the raw datasets. This results in several versioned variables.

```
tab plg0015_v1
tab plg0015_v2
tab plg0015_v3
tab plg0015_v4
```

Studium: Stipendium - Einfachnnenung (2007-2011)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	535,326	82.09	82.09
[-2] Does not apply	112,989	17.33	99.41
[-1] no answer	352	0.05	99.47
[1] No student aid, stipend	2,670	0.41	99.88
[2] Yes, student aid	651	0.10	99.98
[3] Yes, other	153	0.02	100.00
Total	652,141	100.00	

```
. tab plg0015_v2
```

Studium: Stipendium - Kein Stipendium (2012-2017)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	476,953	73.14	73.14
[-5] Not Included In Questionnaire Vers	11,825	1.81	74.95
[-2] Does not apply	158,864	24.36	99.31
[-1] no answer	337	0.05	99.36
[1] No student aid, stipend	4,162	0.64	100.00
Total	652,141	100.00	

```
. tab plg0015_v3
```

Studium: Stipendium - Ja, BAfoeG (2012-2017)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	476,953	73.14	73.14
[-5] Not Included In Questionnaire Vers	11,825	1.81	74.95
[-2] Does not apply	162,090	24.86	99.80
[1] Yes, student aid	1,273	0.20	100.00
Total	652,141	100.00	

```
. tab plg0015_v4
```

Studium: Stipendium - Ja, sonstiges Stipendium (2012-2017)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	476,953	73.14	73.14
[-5] Not Included In Questionnaire Vers	11,825	1.81	74.95
[-2] Does not apply	163,052	25.00	99.95
[-1] no answer	2	0.00	99.95
[1] Yes, other	309	0.05	100.00
Total	652,141	100.00	

As you can see, the variable was asked from 2007 to 2011 as a category question with three response options. As a result, respondents could only give one answer. Since 2012, the question has used binary items. It is quite possible that a respondent gave more than one answer. The harmonized version of the variable integrates the binary items from plg0015_v2, plg0015_v3, and plg0015_v4 into the harmonized version plg0015_h. The coding of the variable plg0015_v1 is used as the generation framework. In addition, the harmonization proposal takes into account the problematic multiple answers with the value four.

```
1 plg0015_h
2 tabstat plg0015_v1 plg0015_v2 plg0015_v3 plg0015_v4 plg0015_h, by(syear)
```

```
. tab plg0015_h
```

University: Scholarship (harmonized)	Freq.	Percent	Cum.
[-8] Question this year not part of sur	360,138	55.22	55.22
[-2] Does not apply	282,456	43.31	98.54
[-1] no answer	352	0.05	98.59
[1] No student aid, stipend	6,832	1.05	99.64
[2] Yes, student aid	1,901	0.29	99.93
[3] Yes, other	439	0.07	100.00
[4] Multiple Answers	23	0.00	100.00
Total	652,141	100.00	

```
. tabstat plg0015_v1 plg0015_v2 plg0015_v3 plg0015_v4 plg0015_h, by(syear)

Summary statistics: mean
  by categories of: syear (Erhebungsjahr (SurveyYear))
```

syear	plg001..	plg001..	pl~15_v3	pl~15_v4	plg~15_h
1984	-8	-8	-8	-8	-8
1985	-8	-8	-8	-8	-8
1986	-8	-8	-8	-8	-8
1987	-8	-8	-8	-8	-8
1988	-8	-8	-8	-8	-8
1989	-8	-8	-8	-8	-8
1990	-8	-8	-8	-8	-8
1991	-8	-8	-8	-8	-8
1992	-8	-8	-8	-8	-8
1993	-8	-8	-8	-8	-8
1994	-8	-8	-8	-8	-8
1995	-8	-8	-8	-8	-8
1996	-8	-8	-8	-8	-8
1997	-8	-8	-8	-8	-8
1998	-8	-8	-8	-8	-8
1999	-8	-8	-8	-8	-8
2000	-8	-8	-8	-8	-8
2001	-8	-8	-8	-8	-8
2002	-8	-8	-8	-8	-8
2003	-8	-8	-8	-8	-8
2004	-8	-8	-8	-8	-8
2005	-8	-8	-8	-8	-8
2006	-8	-8	-8	-8	-8
2007	-1.893852	-8	-8	-8	-1.893852
2008	-1.904288	-8	-8	-8	-1.904288
2009	-1.897845	-8	-8	-8	-1.897845
2010	-1.900711	-8	-8	-8	-1.900711
2011	-1.900707	-8	-8	-8	-1.900707
2012	-8	-1.934174	-1.978987	-1.994318	-1.898939
2013	-8	-1.925927	-1.975288	-1.994218	-1.885709
2014	-8	-1.920044	-1.975969	-1.994539	-1.881413
2015	-8	-2.122834	-2.175919	-2.19159	-1.889821
2016	-8	-2.391366	-2.439758	-2.45542	-1.89875
2017	-8	-2.438633	-2.49509	-2.509958	-1.892627
Total	-6.907259	-6.422924	-6.436731	-6.441162	-5.266231

5.) Euro harmonisation

Variables are versioned and harmonized because they are metric and were asked as DM amounts before the introduc-

tion of the euro. For the long version of the variable, metric variables based on different currencies in different years are harmonized as euro amounts.

Most of the variables harmonized in the long datasets are amounts of money. Before the introduction of the euro, such information was collected in DM.

57. How high was your income from employment last month?

☞ *If you received extra income such as vacation pay or back pay, please do **not** include this. Please do include overtime pay.*

☞ *Please do **not** include "Kindergeld", even if this is paid by the employer.*

Please fill in both:

- **gross income, which means wages or salary before deduction of taxes and social security**
- **net income, which means the sum after deduction of taxes, social security, and unemployment and health insurance.**

My income was: gross ☐ DM

58. How high was your income from employment last month?

☞ *If you received extra income such as vacation pay or back pay, please do **not** include this. Please do include overtime pay.*

☞ *Please do **not** include "Kindergeld", even if this is paid by the employer.*

☞ *If you are self-employed: Please estimate your monthly income before and after tax.*

Please fill in both:

- **gross income, which means wages or salary before deduction of taxes and social security**
- **net income, which means the sum after deduction of taxes, social security, and unemployment and health insurance.**

My income was: gross ☐ EURO

Euro harmonisation involves DM amounts being multiplied by the exchange rate so that the harmonized version of the variable represents euro amounts.

```
list pid syear plc0013_v1 plc0013_h if pid==7006001 & syear==2001
tabstat plc0013_v1 plc0013_v2 plc0013_h, by(syear)
```

```
. list pid syear plc0013_v1 plc0013_h if pid==7006001 & syear==2001

        |     pid   syear   plc001..   plc~13_h |
        |-------------------------------------- |
478114. | 7006001    2001       4200       2147 |
```

```
. tabstat plc0013_v1 plc0013_v2 plc0013_h, by(syear)

Summary statistics: mean
  by categories of: syear (Erhebungsjahr (SurveyYear))
```

syear	plc001..	pl~13_v2	plc~13_h
1984	1307.255	-8	667.9481
1985	1345.458	-8	687.482
1986	1418.1	-8	724.6339
1987	1467.424	-8	749.8613
1988	1521.479	-8	777.4971
1989	1620.861	-8	828.3211
1990	1143.434	-8	583.5691
1991	1513.081	-8	773.2402
1992	1651.645	-8	844.0551
1993	1807.099	-8	923.5413
1994	1872.348	-8	956.8878
1995	1938.037	-8	990.4832
1996	2003.138	-8	1023.763
1997	1990.75	-8	1017.419
1998	1952.576	-8	997.8835
1999	2046.094	-8	1045.72
2000	2002.836	-8	1023.575
2001	2055.956	-8	1050.741
2002	-8	1307.774	1307.774
2003	-8	1251.525	1251.525
2004	-8	1256.811	1256.811
2005	-8	1224.242	1224.242
2006	-8	1208.506	1208.506
2007	-8	1253.102	1253.102
2008	-8	1265.268	1265.268
2009	-8	1248.049	1248.049
2010	-8	1248.695	1248.695
2011	-8	1366.912	1366.912
2012	-8	1399.632	1399.632
2013	-8	1402.34	1402.34
2014	-8	1456.721	1456.721
2015	-8	1457.494	1457.494
2016	-8	1289.785	1289.785
2017	-8	1346.437	1346.437
Total	659.0953	814.6312	1157.022

Last change: Nov 12, 2019

6.9 Working with SOEP Regional Data

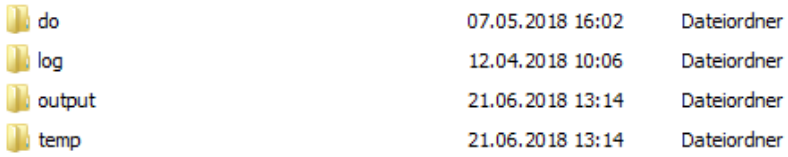
SOEP offers diverse possibilities for regional and spatial analysis. With the anonymized regional information on SOEP respondents' (households' and individuals') place of residence, it is possible to link numerous regional indicators on the levels of the federal states (Bundesländer), spatial planning regions, districts, and postal codes with the data on the SOEP households. However, specific security provisions must be made due to the sensitivity of the data under data protection law. Accordingly, data users are not allowed to give any information in their analyses that could indicate, for instance, the city or district in which respondents reside. The data nevertheless provide valuable background information for regional analysis.

Level	Available since	Data Access	Data Protection
States (Bundesländer)	1984	Standard SOEP datasets (scientific use file)	Data distribution contract
Municipal size classes (e. g., Boustedt)	1984	Data file *ggkbou*	Expanded data distribution contract covering the use of municipal size classes and data protection concept
Spatial planning regions (geocodes)	1985	Data file *Geocodes*	Expanded data distribution contract covering the use of municipal size classes and expanded data protection concept
Official county codes (KKZ) Landkreise, kreisfreie Städte	1985	SOEPremote (online access to county level regional data)	Expanded data distribution contract covering the use of SOEPremote and SOEP remote access form
Official municipality keys postal codes Microm neighbor-hood data	2000 1993 2000	Use of data only at the SOEP Research Data Center at DIW Berlin	Only by personal arrangement in the framework of our "SOEP in Residence" program

For more information and to access the data, see **Regional Data**

Assume that for your research project, you want to measure current (2016) urban-rural differences in the population. You are particularly interested in the differences in interest in politics and the different satisfaction variables provided by the SOEP. You also want to take into account demographic differences in gender and age. To be able to evaluate the potential of the data for your project, you first need an overview. For regional analysis, for example, the municipal size classes from the regional data are suitable.

Create an exercise path with four subfolders:



Example:

- H:/material/exercises/do
- H:/material/exercises/output
- H:/material/exercises/temp
- H:/material/exercises/log

These are used to store your script, log files, datasets, and temporary datasets. Open an empty do-file and define your paths with globals:

```
*********************************************
* Set relative paths to the working directory
*********************************************
global AVZ          "H:\material\exercises"
global MY_IN_PATH "\\hume\rdc-prod\complete\soep-core\soep.v33.2\stata_en\"
global region "\\hume\soep-region\DATA\soep33_de\"
global MY_DO_FILES "$AVZ\do\"
global MY_LOG_OUT "$AVZ\log\"
global MY_OUT_DATA "$AVZ\output\"
global MY_OUT_TEMP "$AVZ\temp\"
```

The global “AVZ” defines the main path. The main paths are subdivided using the globals “MY_IN_PATH”, “MY_DO_FILES”, “MY_LOG_OUT”, “MY_OUT_DATA”, “MY_OUT_TEMP”. The global “MY_IN_PATH” contains the path to the data you ordered.

a) Prepare a dataset for cross-sectional analysis covering the survey year 2016 (wave bg).

To perform your analysis, you need different SOEP variables. The SOEP offers various options for a variable search:

- Search the questionnaires for useful variables (for more information, see the section *Variable Search with Questionnaires*)
- Find a suitable variable in the topic list on paneldata.org (for more information, see the section *Topic Search with paneldata.org*)
- Search for a suitable variable using a search term in paneldata.org (for more information, see the section *Variable Search with paneldata.org*)
- Use the documentation provided by the generated variables (for more information, see the section *Documentation on Generated Data*)

Your source file should contain the following variables:

- Permanent Individual ID **"persnr"**
- Original Household Number **"hhnr"**
- Current Wave Household Number **"bghhnr"**
- The Sex of the Person **"sex"**
- Year of Birth **"gebjahr"**
- Survey Status 2016 **"bgnetto"**
- Sample Membership 2016 **"bgpop"**
- Weighting Factor 2016 **"bgphrf"**
- Satisfaction With Health **"bgp0101"**
- Satisfaction With Sleep **"bgp0102"**
- Satisfaction With Work **"bgp0103"**
- Satisfaction With Housework **"bgp0104"**
- Satisfaction With Household Income **"bgp0105"**
- Satisfaction With Personal Income **"bgp0106"**
- Satisfaction With Dwelling **"bgp0107"**
- Satisfaction With Amount Of Leisure Time **"bgp0108"**
- Satisfaction With Child Care **"bgp0109"**
- Satisfaction With Family Life **"bgp0110"**
- Satisfaction With Social Life **"bgp0111"**
- Satisfaction with Democracy **"bgp0112"**
- Political Interest **"bgp143"**
- Current Sample Region **"bgsampreg"**
- Federal State **"bgbula"**
- Spatial Category by BBSR **"bgregtyp"**
- Municipal Class Sizes "ggk"

Use the key variables from the ppath.dta dataset as your starting file.

```
use hhnr persnr bghhnr sex gebjahr bgnetto bgpop using ${MY_IN_PATH}\ppfad.dta, clear
```

Attention: Please note that since version 34 (v34), PPFAD can be found in the subdirectory "Raw" of the data distribution file. The following exercises are done with version 33.1 (v33.1), where the tracking file was named PPFAD.

Keep people who completed a questionnaire in 2016 and lived in a private household.

```
* Keep people who completed a questionnaire in 2016 and live in a private household
keep if bghhnr>0 & inrange(bgnetto, 10, 29) & inlist(bgpop, 1, 2)
keep hhnr persnr bghhnr sex gebjahr bgnetto bgpop
merge 1:1  persnr using ${MY_IN_PATH}\phrf.dta, keep(match master) keepusing (bgphrf)␣
↪nogenerate
```

(continues on next page)

(continued from previous page)

```
tempfile ppfad
save `ppfad'
```

Prepare the different datasets bgp, bghbrutto, regionl

```
* Prepare dataset bgp
use ${MY_IN_PATH}\bgp.dta, replace
keep persnr hhnr bghhnr bgp01* bgp143
tempfile bgp
save `bgp'

* Prepare dataset bghbrutto
use ${MY_IN_PATH}\bghbrutto.dta, replace
keep hhnr bghhnr bgsampreg bgbula bgregtyp
tempfile bghbrutto
save `bghbrutto'

* Prepare dataset regionl
use ${region}\regionl_v33.dta, replace
keep if syear==2016
keep syear hhnr hhnrakt ggk
rename hhnrakt bghhnr
tempfile regionl
save `regionl'
```

Merge all datasets.

```
* Merge all datasets
use `ppfad'
merge 1:1 persnr using `bgp', keep(match master) nogenerate
merge m:1 bghhnr hhnr using `regionl', keep(match master) nogenerate
merge m:1 bghhnr hhnr using `bghbrutto', keep(match master) nogenerate
```

Recode negative values as missings.

```
* Recode negative values into missings
mvdecode sex gebjahr bgp01* bgp143,mv(-5/-1)
```

Categorize the municipal class sizes from the SOEP regional dataset.

```
* Categorize community class size
gen ggk_cat=.
replace ggk_cat=-1 if ggk==-1
replace ggk_cat=1 if ggk==1 | ggk==2
replace ggk_cat=2 if ggk==3
replace ggk_cat=3 if ggk==4 | ggk==5
replace ggk_cat=4 if ggk>5 & ggk<=7

lab var ggk_cat "Community Size categorised"
lab def ggk_cat -1 "No information" 1 "<=5000" 2 "5001 - 20000" 3 "20001 - 100000" ///
4 ">100000"
lab val ggk_cat ggk_cat
```

Generate an age variable.

```
* Generate age variable
gen alter= 2016-gebjahr if gebjahr > 0
```

(continues on next page)

(continued from previous page)

```
gen alter_cat=1 if alter<=20
replace alter_cat=2 if alter>20 & alter<=30
replace alter_cat=3 if alter>30 & alter<=65
replace alter_cat=4 if alter>65 & alter<=120

lab var alter "age"
lab var alter_cat "age categorized"
lab def alter_cat 1 "<=20" 2 "21-30" 3 "31-65" 4 ">65"
lab val alter_cat alter_cat
```

Categorize a federal states variable.

```
* Categorize federal states
gen bgbula_cat=.
* Schleswig-Holstein + Hamburg
replace bgbula_cat=1 if bgbula==1 | bgbula==2
* Lower Saxony + Bremen
replace bgbula_cat=2 if bgbula==3 | bgbula==4
* Mecklenburg Western Pomerania + Brandenburg
replace bgbula_cat=3 if bgbula==13 | bgbula==12
* Saarland + Rhineland Palatinate
replace bgbula_cat=4 if bgbula==7 | bgbula==10
* Northrhine-Westphalia
replace bgbula_cat=5 if bgbula==5
* Hesse
replace bgbula_cat=6 if bgbula==6
* Baden-Württemberg
replace bgbula_cat=7 if bgbula==8
* Bavaria
replace bgbula_cat=8 if bgbula==9
* Berlin
replace bgbula_cat=9 if bgbula==11
* Saxony
replace bgbula_cat=10 if bgbula==14
* Saxony-Anhalt
replace bgbula_cat=11 if bgbula==15
* Thuringia
replace bgbula_cat=12 if bgbula==16

lab var bgbula_cat "Federal states categorized"
lab def bgbula_cat 1 "Schleswig-Holstein/Hamburg" 2 "Lower Saxony/Bremen" 3
↪"Mecklenburg Western Pomerania/Brandenburg" ///
4 "Saarland/Rhineland Palatinate" 5 "Northrhine-Westphalia" 6 "Hesse" ///
7 "Baden-Wuerttenberg" 8 "Bavaria" 9 "Berlin" 10 "Saxony" 11 "Saxony-Anhalt" 12
↪"Thuringia"
lab val bgbula_cat bgbula_cat
drop bgbula
rename bgbula_cat bgbula
```

Put the variables in your preferred order and save your dataset.

```
* Order demography and identifiers first
order persnr hhnr bghhnr syear sex gebjahr alter alter_cat bgsampreg bgbula ggk ///
ggk_cat bgregtyp

save ${MY_OUT_DATA}\zeit_online.dta, replace
```

b) You want to get an initial overview of regional differences in satisfaction with various aspects of life. Use the variable bgsampreg and cross-stabilize the variable with all satisfaction variables to identify differences between East and West Germany, display the absolute and relative frequencies.

To save the tables, save them in a log file.

```
*******************************************************************************
capture log close
log using "${MY_LOG_OUT}\satisfaction.log", replace

* Life satisfaction

local varlist bgp0101 bgp0102 bgp0103 bgp0104 bgp0105 bgp0106 bgp0107 bgp0108 ///
bgp0109 bgp0110 bgp0111 bgp0112
foreach x of local varlist {
tab bgsampreg `x' [aw= bgphrf] , row
}
```

	Satisfaction With Health										
Current Sample Region	[0] 0 Sat	[1] 1 Sat	[2] 2 Sat	[3] 3 Sat	[4] 4 Sat	[5] 5 Sat	[6] 6 Sat	[7] 7 Sat	[8] 8 Sat	[9] 9 Sat	[10] 10 S	Total
[1] West Germany	256.82471	260.7491	623.17631	1,180.878	1,226.948	2,717.234	2,324.916	4,208.661	5,384.689	2,623.874	1,546.069	22,354.019
	1.15	1.17	2.79	5.28	5.49	12.16	10.40	18.83	24.09	11.74	6.92	100.00
[2] East Germany	67.27909	65.784226	175.10943	332.81232	283.75315	686.88971	548.93801	900.241273	999.234017	454.15063	244.78919	4,758.981
	1.41	1.38	3.68	6.99	5.96	14.43	11.53	18.92	21.00	9.54	5.14	100.00
Total	324.1038	326.533325	798.28574	1,513.69	1,510.701	3,404.124	2,873.854	5,108.902	6,383.923	3,078.025	1,790.858	27,113
	1.20	1.20	2.94	5.58	5.57	12.56	10.60	18.84	23.55	11.35	6.61	100.00

	satisfaction with sleep										
Current Sample Region	[0] 0 Sat	[1] 1 Sat	[2] 2 Sat	[3] 3 Sat	[4] 4 Sat	[5] 5 Sat	[6] 6 Sat	[7] 7 Sat	[8] 8 Sat	[9] 9 Sat	[10] 10 S	Total
[1] West Germany	159.40597	235.97229	644.89468	1,096.823	1,293.988	2,220.017	2,201.258	3,256.5262	4,299.5147	2,566.6674	1,903.7127	19,878.78
	0.80	1.19	3.24	5.52	6.51	11.17	11.07	16.38	21.63	12.91	9.58	100.00
[2] East Germany	26.18853	37.661261	147.66919	280.15784	312.65151	589.20671	483.1413	627.41268	877.71191	505.79602	350.62287	4,238.22
	0.62	0.89	3.48	6.61	7.38	13.90	11.40	14.80	20.71	11.93	8.27	100.00
Total	185.5945	273.633552	792.56387	1,376.981	1,606.6397	2,809.224	2,684.399	3,883.939	5,177.227	3,072.463	2,254.336	24,117
	0.77	1.13	3.29	5.71	6.66	11.65	11.13	16.10	21.47	12.74	9.35	100.00

	Satisfaction With Work										
Current Sample Region	[0] 0 Sat	[1] 1 Sat	[2] 2 Sat	[3] 3 Sat	[4] 4 Sat	[5] 5 Sat	[6] 6 Sat	[7] 7 Sat	[8] 8 Sat	[9] 9 Sat	[10] 10 S	Total
[1] West Germany	108.18696	101.53684	226.91136	408.98276	421.367929	1,161.145	1,260.616	2,377.968	3,392.8893	1,994.521	1,091.584	12,545.71
	0.86	0.81	1.81	3.26	3.36	9.26	10.05	18.95	27.04	15.90	8.70	100.00
[2] East Germany	27.931559	21.235589	38.931778	84.358325	121.25058	286.25775	240.69159	545.04361	730.78802	333.42068	199.38207	2,629.292
	1.06	0.81	1.48	3.21	4.61	10.89	9.15	20.73	27.79	12.68	7.58	100.00
Total	136.11852	122.77242	265.84314	493.34109	542.61851	1,447.403	1,501.308	2,923.011	4,123.677	2,327.942	1,290.9661	15,175
	0.90	0.81	1.75	3.25	3.58	9.54	9.89	19.26	27.17	15.34	8.51	100.00

To view all tables, look at your generated log file.

c) Now take a closer look at satisfaction with various aspects of life with the help of SOEP regional data. Use the municipal size classes. Create a table showing satisfaction with different aspects of life and highlighting differences by sex, age, municipal size class, and federal state.

```
foreach x of local varlist {
* Tabulation of satisfaction by municipal size class and federal state
table `x' sex alter_cat, by(bgbula ggk_cat) contents(freq) column row stubwidth(20)␣
↪cellwidth(8) csepwidth(2) nomissing
* Tabulation of satisfaction by municipal size class
```

(continues on next page)

(continued from previous page)

```
table `x' sex alter_cat, by(ggk_cat) contents(freq) column row stubwidth(20) cellwidth(8) csepwidth(2) nomissing
* Tabulation of satisfaction by federal state
table `x' sex alter_cat, by(bgbula) contents(freq) column row stubwidth(20) cellwidth(8) csepwidth(2) nomissing
}
```

Federal states categorized, Community Size categorised and Satisfaction With Social Life	age categorized and Sex											
	<=20			21-30			31-65			>65		
	[1] Male	[2] Fema	Total	[1] Male	[2] Fema	Total	[1] Male	[2] Fema	Total	[1] Male	[2] Fema	Total
Schleswig-Holstein/H <=5000												
[0] Completely unsat												
[1] 1 On Scale 0-Low							1		1			
[2] 2 On Scale 0-Low												
[3] 3 On Scale 0-Low				1		1	1	1	2	1	1	2
[4] 4 On Scale 0-Low							2	2	4		1	1
[5] 5 On Scale 0-Low				1		1	5	7	12	4	1	5
[6] 6 On Scale 0-Low		1	1	2		2	5	10	15	1	2	3
[7] 7 On Scale 0-Low	3		3		2	2	17	16	33	4	4	8
[8] 8 On Scale 0-Low	1	2	3	3	4	7	21	32	53	10	8	18
[9] 9 On Scale 0-Low	1	3	4		4	4	18	22	40	9	7	16
[10] Completely sati	2	3	5	3	3	6	4	15	19	6	8	14
Total	7	9	16	10	13	23	74	105	179	35	32	67
Schleswig-Holstein/H 5001 - 20000												
[0] Completely unsat								1	1			
[1] 1 On Scale 0-Low												
[2] 2 On Scale 0-Low							3		3	1		1
[3] 3 On Scale 0-Low							3	1	4			
[4] 4 On Scale 0-Low				1		1	1	1	2	1		1
[5] 5 On Scale 0-Low							4	3	7		1	1
[6] 6 On Scale 0-Low				1	1	2	4	3	7	1	2	3
[7] 7 On Scale 0-Low		3	3		2	2	10	10	20	4	1	5
[8] 8 On Scale 0-Low	3	1	4	6	2	8	19	30	49	5	5	10
[9] 9 On Scale 0-Low	2	1	3	2	4	6	12	10	22	2	2	4
[10] Completely sati	3		3	2	1	3	4	10	14		1	1
Total	8	5	13	12	10	22	60	69	129	14	12	26
Schleswig-Holstein/H 20001 - 100000												
[0] Completely unsat								1	1			
[1] 1 On Scale 0-Low												
[2] 2 On Scale 0-Low					1	1		1	1			
[3] 3 On Scale 0-Low	1		1									
[4] 4 On Scale 0-Low				1		1		1	1			
[5] 5 On Scale 0-Low							1	7	8	3	4	7
[6] 6 On Scale 0-Low				1	1	2	4	4	8	3	2	5
[7] 7 On Scale 0-Low	2		2	3	4	7	15	13	28	1		1
[8] 8 On Scale 0-Low	1	1	2	3	4	7	22	25	47	4	4	8
[9] 9 On Scale 0-Low	1		1	6	4	10	13	23	36	3	5	8
[10] Completely sati		4	4	3	5	8	10	18	28	1	1	2
Total	5	5	10	17	19	36	65	93	158	15	16	31
Schleswig-Holstein/H >100000												
[0] Completely unsat												
[1] 1 On Scale 0-Low							1		1			
[2] 2 On Scale 0-Low							1		1			
[3] 3 On Scale 0-Low							1	3	4	1		1
[4] 4 On Scale 0-Low							5	2	7	1	1	2
[5] 5 On Scale 0-Low					2	2	6		6	7	7	14
[6] 6 On Scale 0-Low	1		1	1	1	2	13	17	30	8	7	15
[7] 7 On Scale 0-Low	3	2	5	3	10	13	25	32	57	3	9	12
[8] 8 On Scale 0-Low	2	2	4	10	8	18	44	60	104	14	20	34
[9] 9 On Scale 0-Low	8	4	12	7	10	17	25	37	62	12	15	27
[10] Completely sati	2	1	3	8	9	17	18	24	42	9	11	20
Total	16	9	25	29	40	69	139	175	314	55	70	125

To view all tables, look at your generated log file. As you can see, SOEP regional data can be used to analyze variables at the lowest regional levels.

d) Create a table that shows political interest differentiated by age, sex, and municipal size class in Bavaria

```
*******************************************************************************
capture log close
log using "${MY_LOG_OUT}\political_interest.log", replace

* Political interest
* Tabulation of political interest by municipal size class for Bavaria
table bgp143 sex alter_cat if bgbula==8, by(ggk_cat) contents(freq) column row␣
↪stubwidth(20) cellwidth (8) csepwidth(2) nomissing
```

```
. table bgp143 sex alter_cat if bgbula==8, by(ggk_cat) contents(freq) column row stubwidth(20) cellwidth (8) csepwidth(2) nomissing
```

Community Size categorised and Political Interests	age categorized and Sex											
	<=20			21-30			31-65			>65		
	[1] Male	[2] Fema	Total	[1] Male	[2] Fema	Total	[1] Male	[2] Fema	Total	[1] Male	[2] Fema	Total
<=5000												
[1] Very Strong				2	3	5	33	8	41	8	9	17
[2] Strong	11	3	14	9	15	24	124	97	221	33	24	57
[3] Not Much	13	17	30	31	32	63	129	202	331	35	36	71
[4] Not at All	12	17	29	15	19	34	42	78	120	1	6	7
Total	36	37	73	57	69	126	328	385	713	77	75	152
5001 - 20000												
[1] Very Strong	2	3	5	11	4	15	43	15	58	24	19	43
[2] Strong	10	7	17	26	14	40	138	128	266	72	45	117
[3] Not Much	21	17	38	28	38	66	187	281	468	55	74	129
[4] Not at All	14	17	31	18	31	49	68	120	188	6	13	19
Total	47	44	91	83	87	170	436	544	980	157	151	308
20001 - 100000												
[1] Very Strong	2		2	6		6	18	11	29	13	4	17
[2] Strong	6	3	9	11	10	21	56	48	104	30	26	56
[3] Not Much	11	7	18	25	34	59	85	127	212	22	26	48
[4] Not at All	9	6	15	16	27	43	53	69	122	2	4	6
Total	28	16	44	58	71	129	212	255	467	67	60	127
>100000												
[1] Very Strong	2		2	5	2	7	29	18	47	12	9	21
[2] Strong	1	5	6	25	22	47	101	85	186	40	29	69
[3] Not Much	6	13	19	26	31	57	85	142	227	22	26	48
[4] Not at All	1	4	5	12	20	32	37	50	87	3	12	15
Total	10	22	32	68	75	143	252	295	547	77	76	153

As you have seen here, the SOEP offers a wide range of possibilities for regional analysis. It is possible to allocate a multitude of regional indicators at the level of federal states, regional planning regions, districts, and postal codes.

Last change: Nov 12, 2019

6.10 How to Use SOEP IGEL

1. IGEL Workstation

IGEL refers to a computer terminal workstation for access to SOEP data:

- The terminal allows data to be entered and displayed.
- The IGEL is a so-called thin client, a computer with little computing power, which only provides a terminal to a server.

- This thin client at the SOEP guest workstation/FDZ is from the manufacturer Igel Technology, where IGEL stands for "**I** ntelligente **G** esamtlösung in der Mikro **el** ektronik".

2. Account

Access to SOEP data can only be provided in compliance with high security standards to protect respondents' confidentiality and maintain their trust in the survey. The data are also provided solely for scientific research purposes, that is, they are only made available to members of the scientific community. Researchers can therefore only access SOEP data after they have signed a data distribution contract with DIW Berlin. The same rules apply to the secure guest workstations at RDC SOEP and at other secure data access points. Since IGEL terminals also provide access to small-scale regional data, users have to sign additional contracts for these data.

All IGEL users must sign a data distribution contract with the DIW Berlin: Application for a Data Distribution Contract.

3. Logging in

Turn on the computer and the following screen should appear on the monitor. (see *figure 1*)

Fig. 1: Figure 1: IGEL start screen

At the bottom right, you should see the icon for an existing network connection appear: Two arrows, one pointing up and one pointing down.

Click on the arrow icon to see the terminal name and the existing network connection . See *figure 2*.

Fig. 2: Figure 2: Connection with LAN available

For each available server, two icons are displayed on the start screen at the top left, a red one and a blue one with the same name. See *figure 3*.

The following two servers are currently available:

1. HAUSER: Access to the SOEP survey data, including connection to small-scale regional indicators (WITHOUT coordinates).
2. MORAN: Access to the coordinates of SOEP households, but without survey data. **Access is only possible from RDC SOEP guest stations at DIW Berlin**

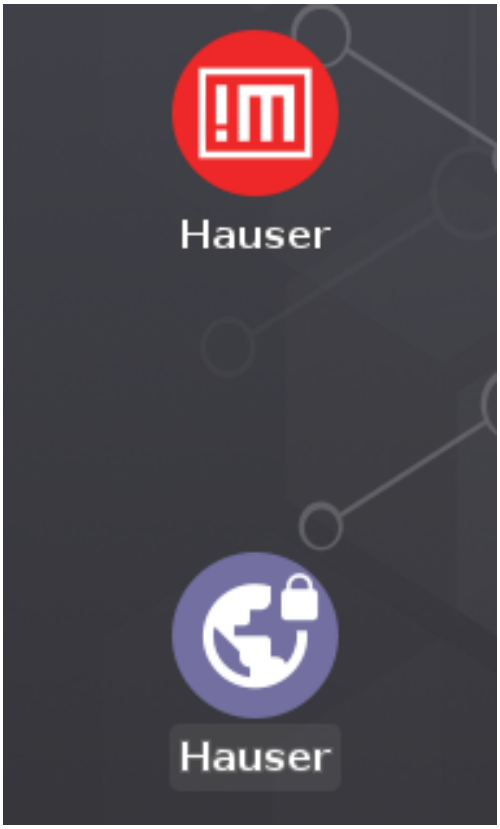


Fig. 3: Figure 3: Icons to connect with the SOEP server

Blue Icon: To connect to one of the two servers at RDC SOEP, first establish an open VPN connection by clicking on the blue icon for the server you would like to connect to. The icon in the lower right corner should then display the existing VPN connection. By clicking once on this icon, you can see the server's IP address . See *figure 4*

Red Icon: Once you have established the VPN connection to the SOEP server, click once on the red icon to start your session. The server's login window should appear, see *figure 5*. Enter the user name and password provided to you by RDC SOEP.

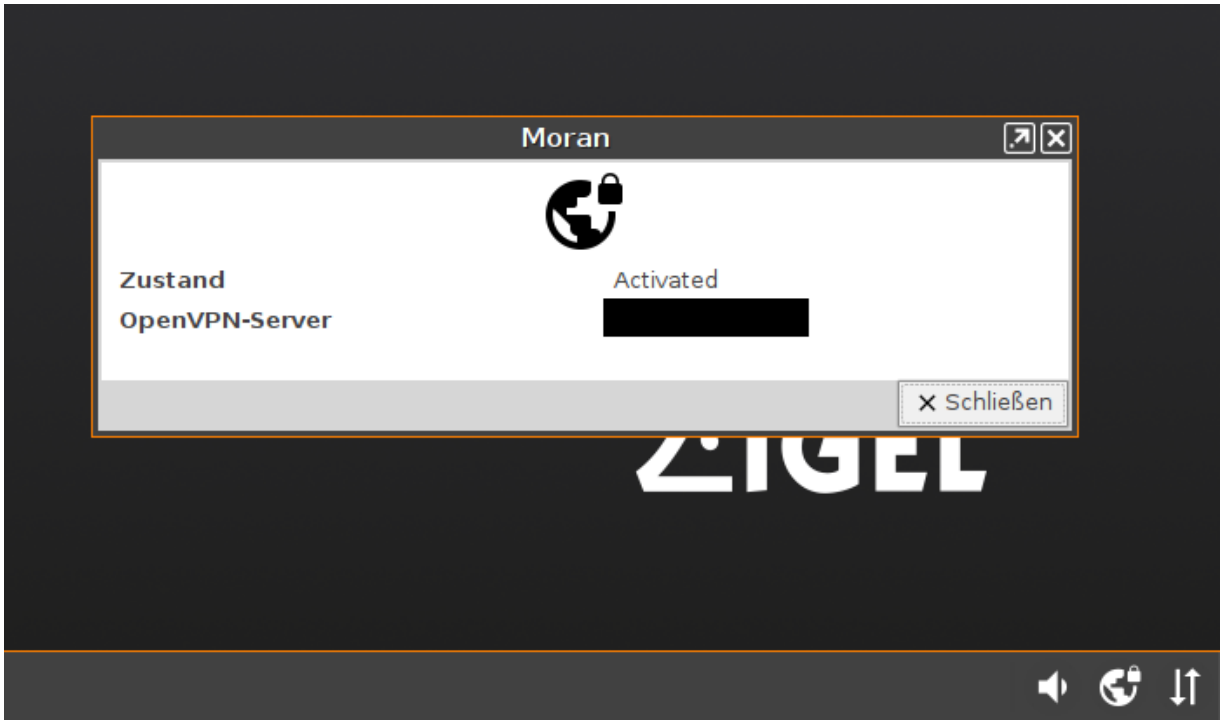


Fig. 4: Figure 4: Open VPN connection established

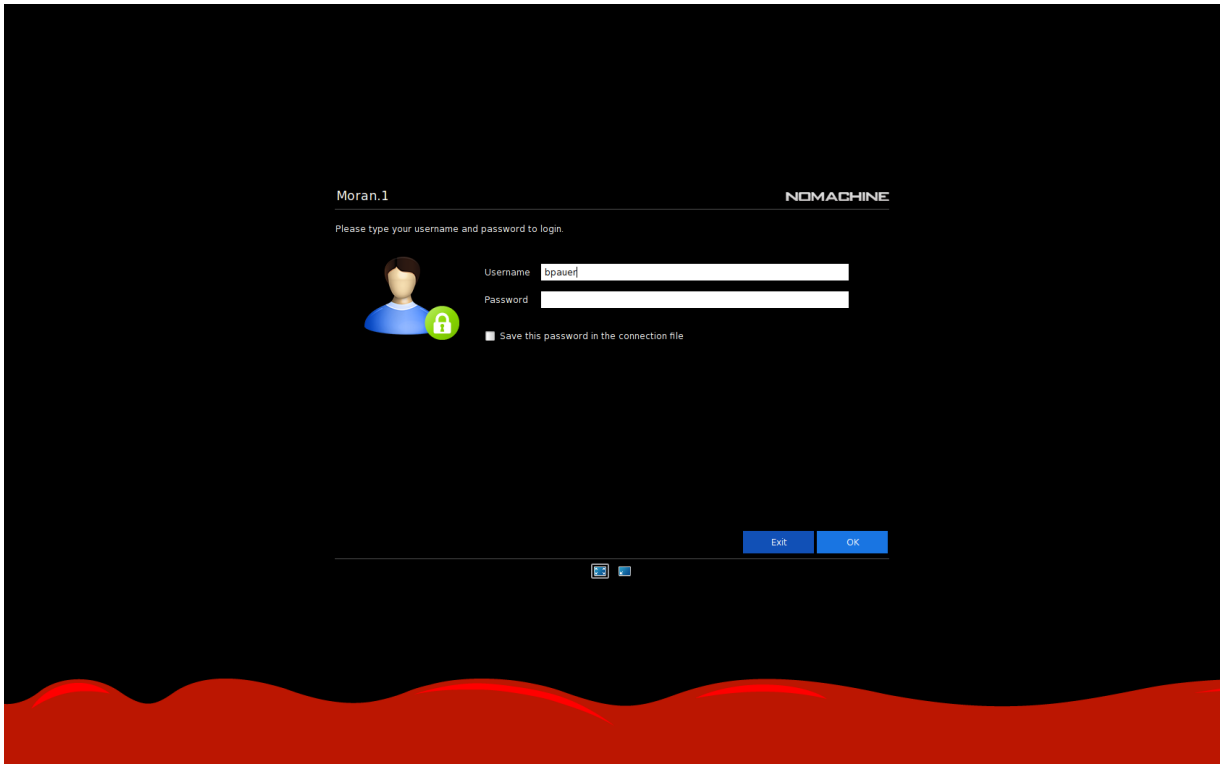


Fig. 5: Figure 5: Login to the SOEP Server

4. Working with SOEP DATA

4.1. Starting programs

- After you have logged in, a blank desktop will appear with a menu bar at the top.
- In general, programs can be started by clicking on "activities" and then either by clicking on the icon or by typing the name of the program into the search field.
- Users should inform the RDC SOEP team in advance about any additional ados in Stata or packages in R. These will be installed after checking.
- **Start Stata**: Unfortunately, there is no automatic start icon for Stata, so you have to do the following:
  1. Click on activities
  2. Enter "terminal" in the search window
  3. Start either "Terminal" or "XTerm".
  4. Enter the command "xstata-mp" into the terminal that has now appeared, and press the return key. Stata should now appear.
- The following table shows which programs are installed and available for use on each server:

Program	HAUSER	MORAN
Stata	Yes	-/-
R/RStudio	Yes	Yes
QGIS	-/-	Yes
grass	-/-	Yes
PostGis	-/-	Yes
LibreOffice	Yes	Yes
Emacs	Yes	Yes
gnome-text-editor	Yes	Yes
Nautilus (File manager)	Yes	Yes

4.2. Using SOEP data and your own data

- The latest version of the SOEP data is available at the following address directory path:

  HAUSER `~/soep-data/` or `/import/SOEP-Regio/data/`

  MORAN `~/soep-data/` or `/import/SOEP-GIS/data`
- You can store your own data and scripts in your personal home directory. `~/daten/`

4.3. Logging out

- Use the icon in the upper right corner

- click on your username and on logout:

5. Importing Scripts or External Data

- You can send these data to the RDC team before your stay. Send it to SOEP. Please use the following website:
  - cs-soep.diw.de
- As receiver for the data and scripts please use soepmail@diw.de.

- Before you come to us, please send us the data to import early enough so that we have enough time to install it.

6. Exporting Results

- Create an "export" directory in your home directory.
- In it, you can store all the results that you need for your publication.
- In addition, create a README file in your home directory (either as a text file or as a Word file using LibreOffice) describing what the individual files or subfolders contain. The description can be very brief and serves only to provide orientation in checking adherence to data protection provisions.
- Then send an email to soepmail@diw.de and ask us to export your results.

Section author: Jan Goebel <jgoebel@diw.de>

Last change: Nov 13, 2019

CHAPTER

SEVEN

WORKING WITH SOEP DOCUMENTATION

7.1 Variable Search with Questionnaires

If you come across a variable in the dataset whose variable content is unclear, you should always check whether there is a suitable questionnaire for the dataset. Under *Original Core Data* you can see whether the datasets correspond to a survey instrument. The related questionnaires can be found here:

Questionnaires

Example: Working on a research project, you come across the variable bbh5508 with the German label "Auto: Gründe" (Car: Reasons) and the English label "Reason for No Car in Household"

```
. tab bbh5508
```

Reason For No Car In HH	Freq.	Percent	Cum.
[-5] Not included in this version of th	4,529	26.93	26.93
[-2] Does not apply	9,933	59.06	85.99
[-1] No Answer	167	0.99	86.98
[1] Financial Reasons	871	5.18	92.16
[2] Other Reasons	1,319	7.84	100.00
Total	16,819	100.00	

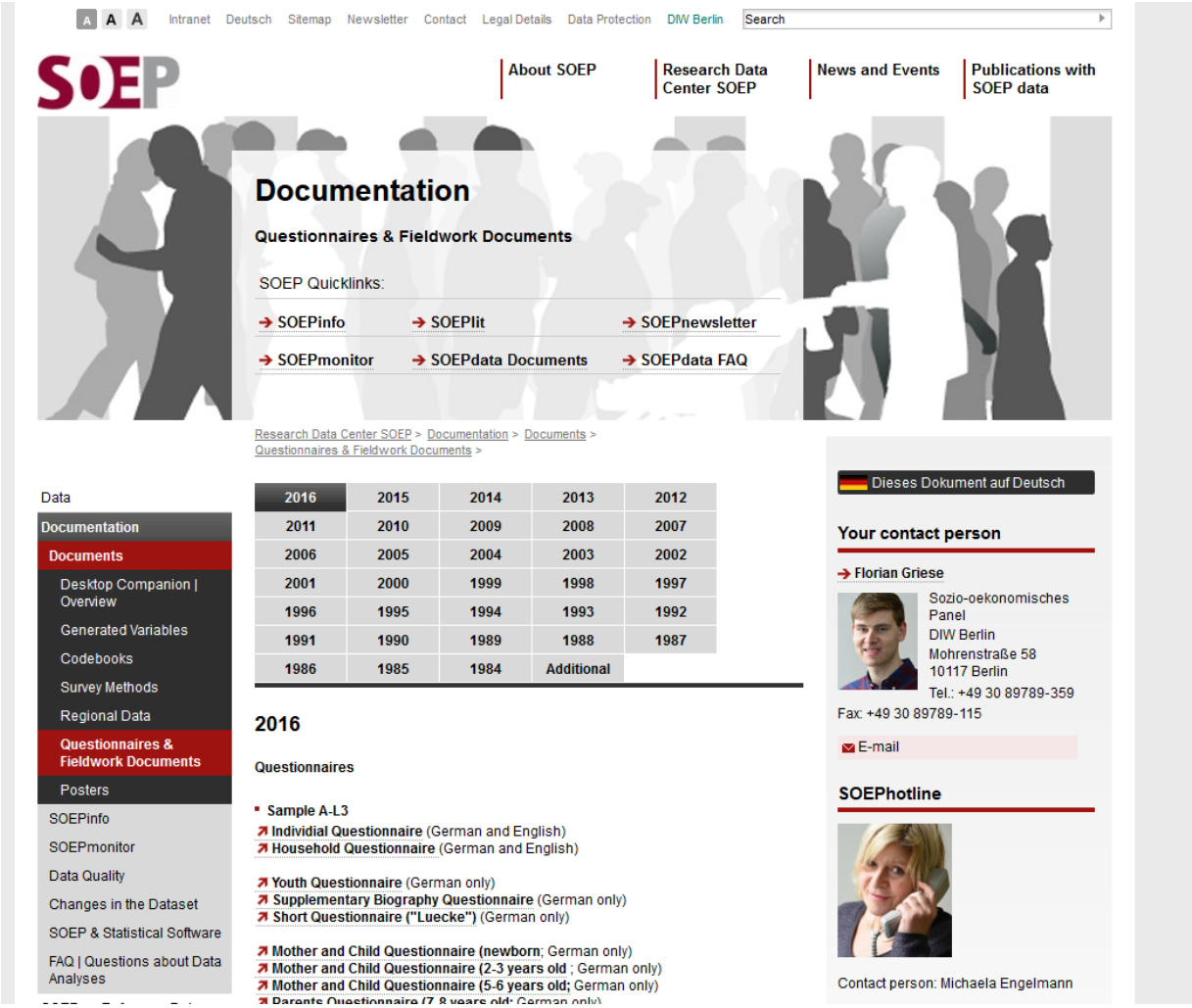
Unfortunately, it is unclear what exactly this variable represents. You should refer to the questionnaires for the complete question and possible filter instructions.

Example Variable:

bbh5508: Wave "bb" (Survey Year 2011); household questionnaire ("h"), question number 55, item 8

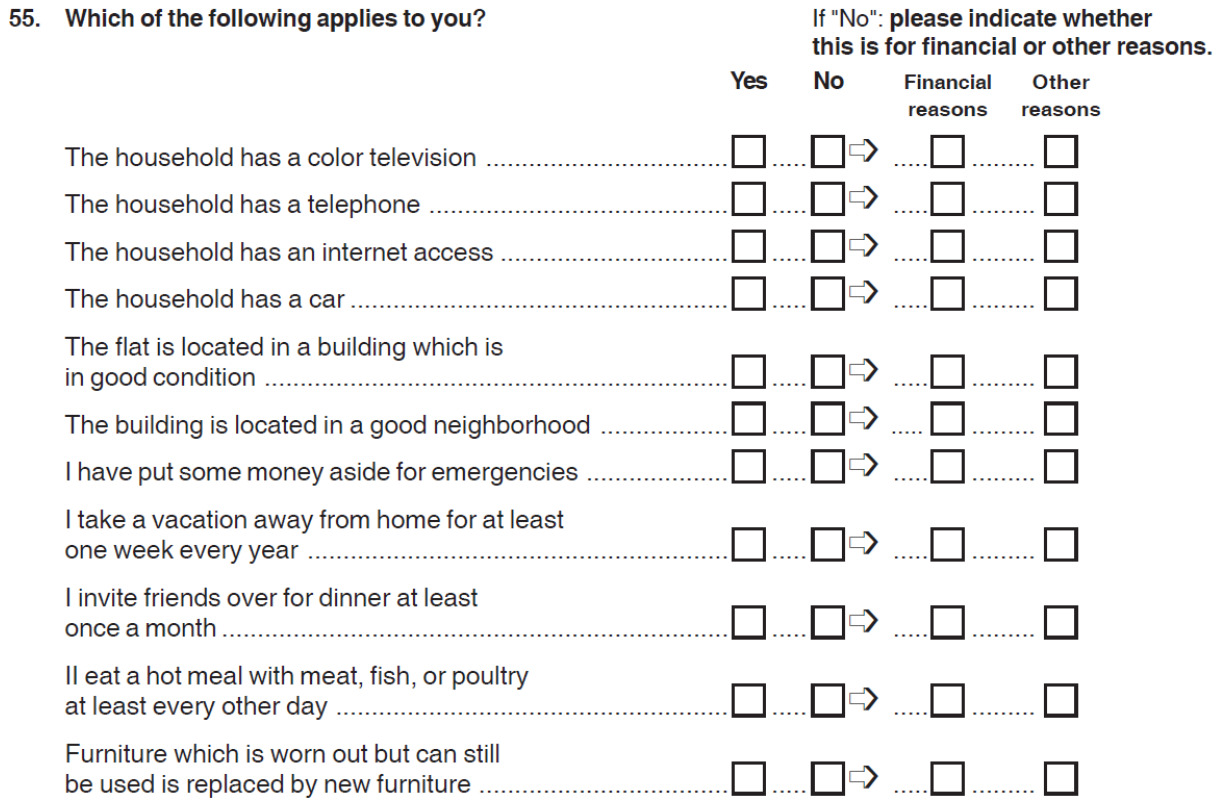
Open Questionnaires

The variable "bbh5508" can be found in the questionnaires for 2011. Select the survey year 2011 and download the household questionnaire here



Search the variable “bbh5508” in the questionnaire.

Since you are already in the correct questionnaire, you must now search for question 55.



To understand which information the variable "bbh5508" contains, you have to deal with the question. For each answer category, respondents should indicate whether or not the shown items apply to the household. If the item does not apply, respondents must answer an additional question about the reasons. Both questions should be understood as separate variables. The variable "bbh5501" indicates whether a TV is present in the household. The reasons why there is no TV in the household can be found in the variable "bbh5502". The variable "bbh5507" shows whether a car is present in the household and the variable "bbh5508" shows reasons why no car is present in the household. By looking into the questionnaire, the variable is now easier to understand. The variable "bbh5508" only contains people who do not have a car in their household and shows the reasons given.

Last change: Nov 13, 2019

7.2 Variable Search with paneldata.org

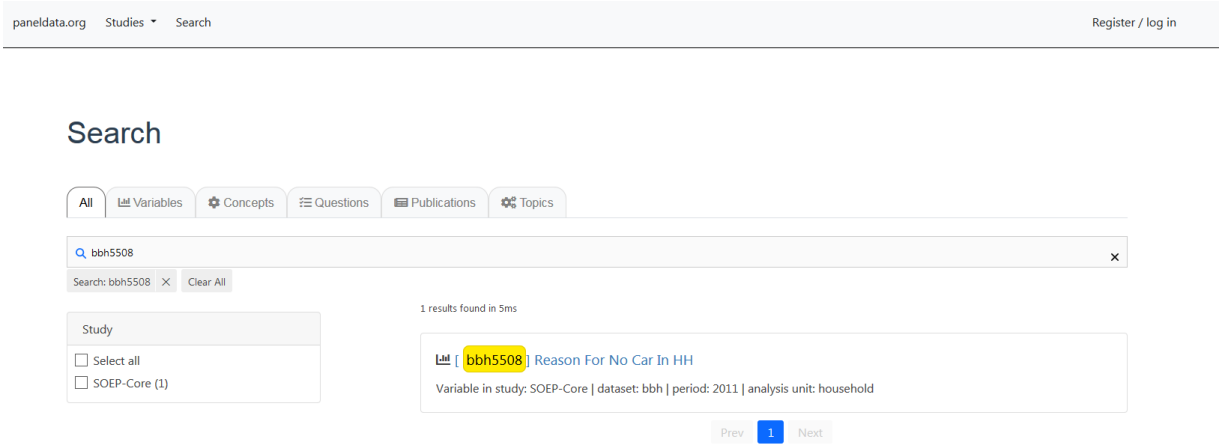
Paneldata.org also allows you to search for variables and to find more information about generated variables. It offers comprehensive frequency counts, chronologies of variables, cross-study variable linkage via concepts, a syntax generator, and a topic list for content search in the SOEP.

Example Variable:

bbh5508: Wave "bb" (Survey Year 2011); household questionnaire ("h"), question number 55, item 8

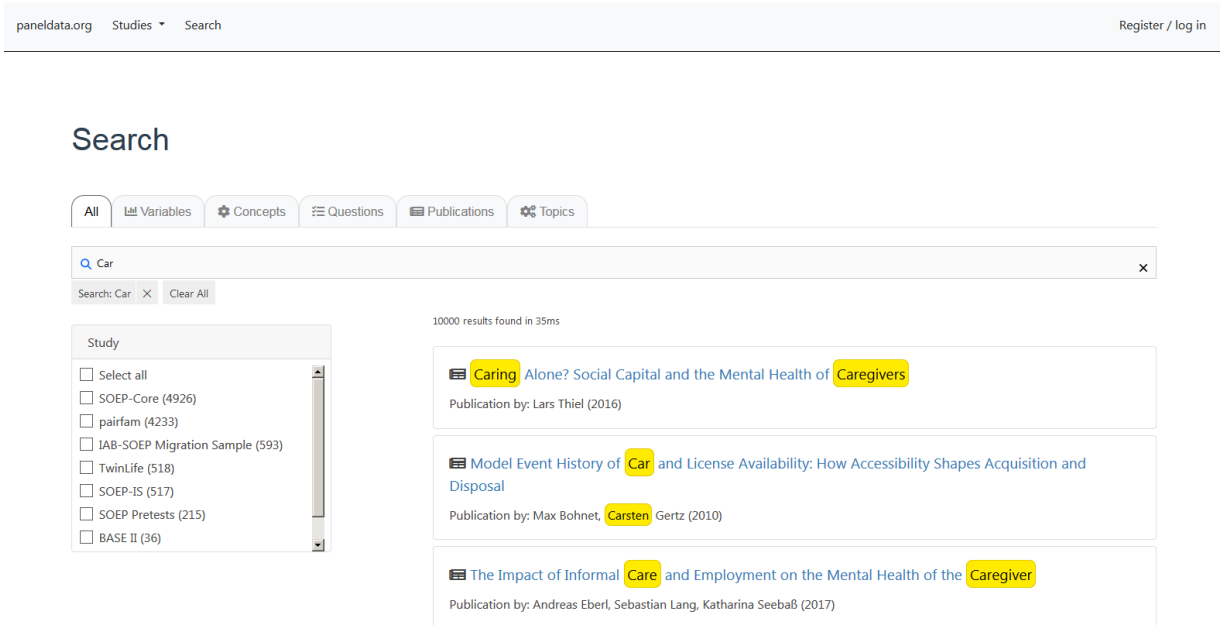
Open Paneldata

Select the study SOEP-Core. The SOEP-Core overview contains important general information about the study, e.g., data access, survey method, questionnaires, themes, terms for missing codes, all available datasets in the study and metadata-based questionnaires. To search for a variable, a dataset, or a publication, simply enter the desired search term in the search bar.

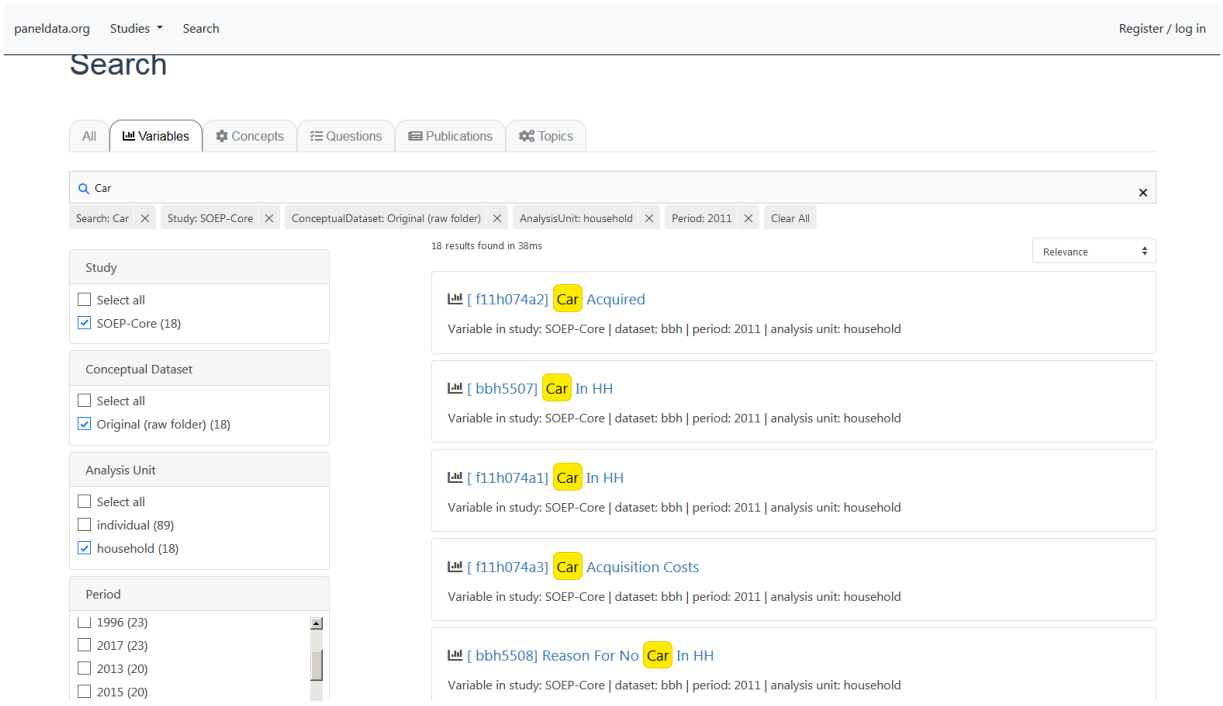


To obtain the desired results, you will need to input specific information. The results window displays all search results. You will see that the variable "bbh5508" originates from SOEP-Core data and can be found in the dataset

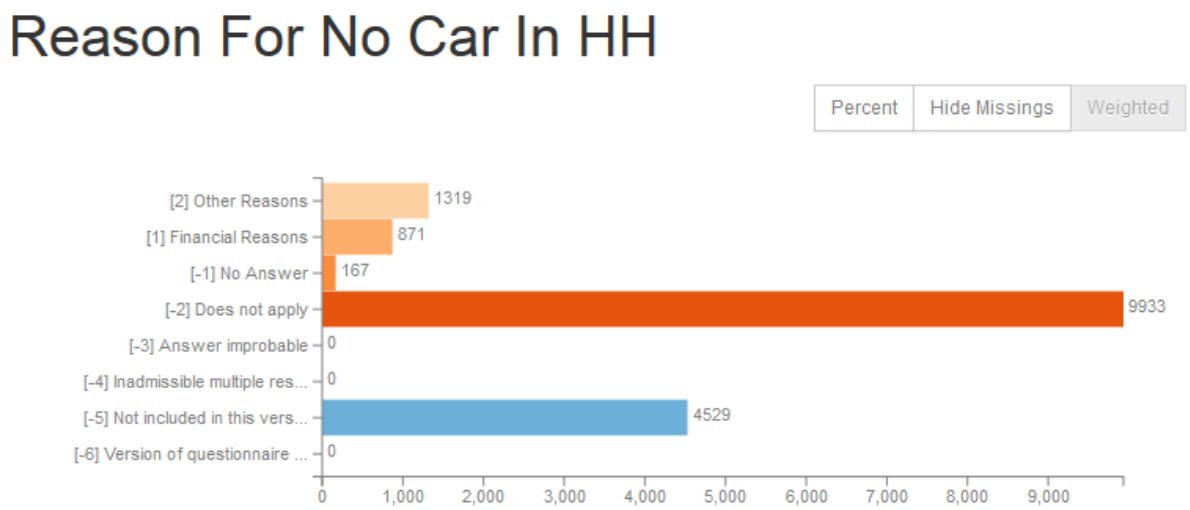
"bbh" (survey year 2011). If your search is not so specific, you can also search by keywords. We are still interested in the topic "car".



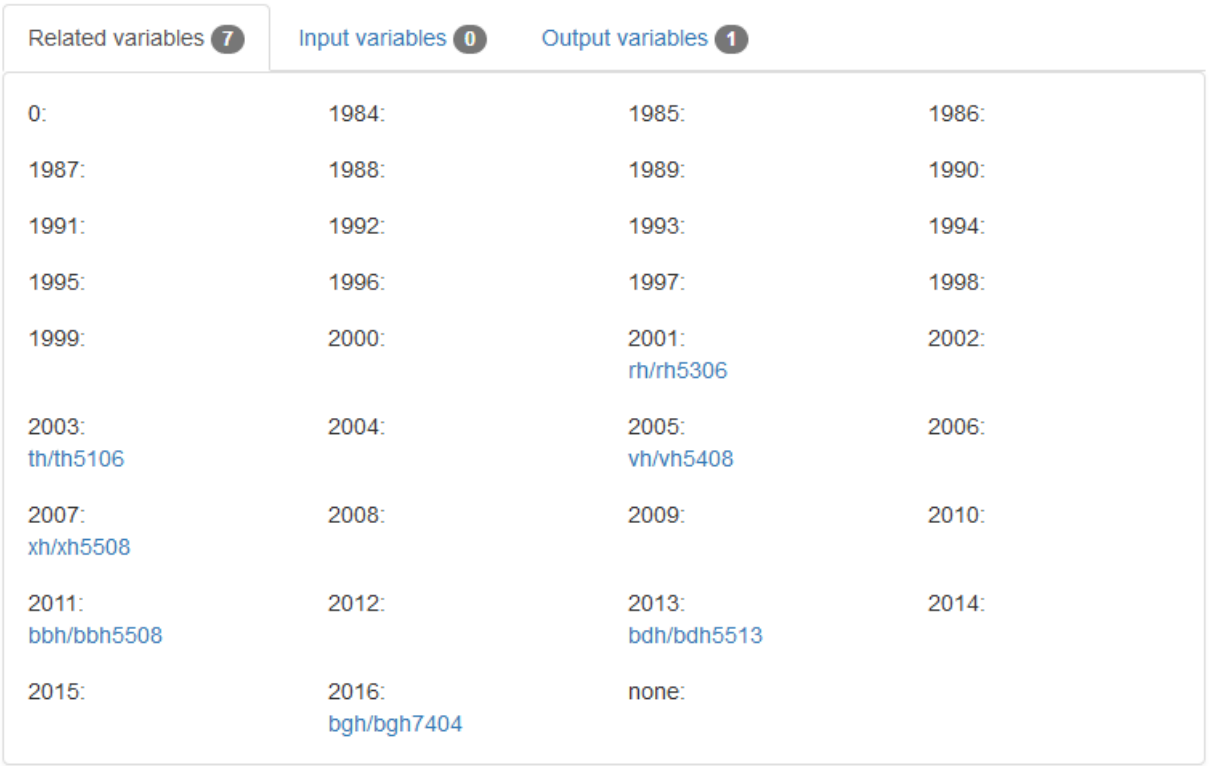
To better limit the 10000 results, the filter options on the left and on the top should be used. We are looking for variables from the "SOEP-Core" datasets. The search results should be limited with the filter options. Which survey years are of interest to me, do I want to work with original data or generated data? For more information about the different datasets in SOEP-Core, see the section *Data Distribution File*. Should the variable I am looking for be at household level or at individual level?



By filtering, the search result is limited to 18 hits, which also shows the variable we searched for. If you click on the variable “bbh5508”, you will find additional information about the variable.

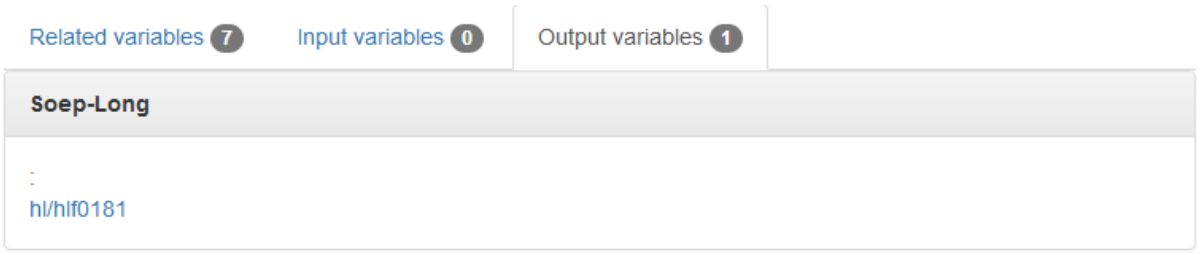


First you see the weighted absolute frequencies for the variable. It is possible to remove the missing codes from the analysis and/or to display the relative frequencies. Even without opening the dataset, paneldata obgain a good overview of the frequencies of a variable.

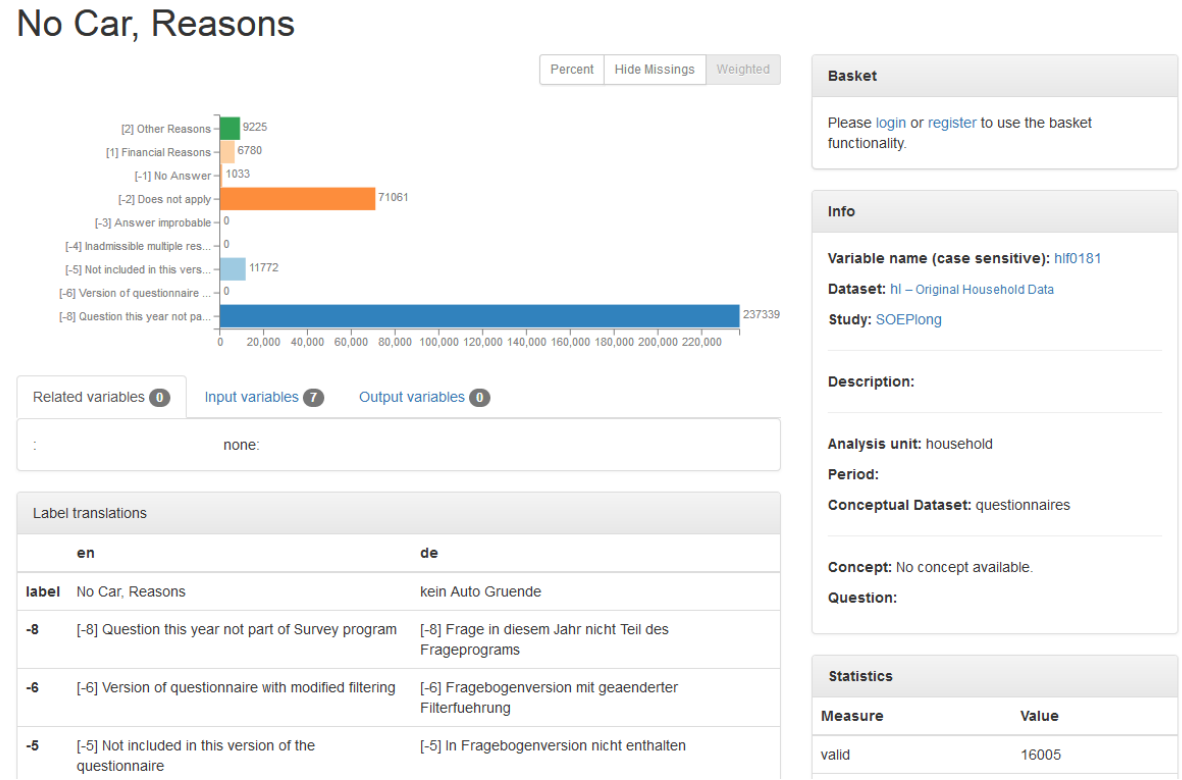


In the Related Variables section you will also find the chronology of the variable you are looking for. The sample

variable was collected in 2001, 2003, 2005, 2007, 2011, 2013. Below the survey year, the name of the variable in the respective year is displayed and can be clicked to access the respective variable page. You can see at a glance when the variable was measured, how often it was measured, and what its name is in the respective survey year.



In addition, by clicking on "Output variables", you will find a variable forwarding you to the variable in "long" format. For a more detailed understanding of the long format, read the section *Data Structure in "Long" Format (long)*.



As soon as you click on the "long" variable, you will get to the variable overview for this variable in the long format. The overview of variables does not differ. It can be seen that our example variable "bbh5508" can also be found in long format in the dataset "hl" with the variable "hlf0181".

Label translations

	en	de
label	Reason For No Car In HH	Auto: Gruende
-6	[-6] Version of questionnaire with modified filtering	[-6] Fragebogenversion mit geaenderter Filterfuehrung
-5	[-5] Not included in this version of the questionnaire	[-5] In Fragebogenversion nicht enthalten
-4	[-4] Inadmissible multiple response	[-4] Unzulaessige Mehrfachantwort
-3	[-3] Answer improbable	[-3] nicht valide
-2	[-2] Does not apply	[-2] trifft nicht zu
-1	[-1] No Answer	[-1] keine Angabe
1	[1] Financial Reasons	[1] finanzielle Gruende
2	[2] Other Reasons	[2] andere Gruende

The field “Label translations” shows the value labels of the variables in German and English. In addition, all missing codes used in SOEP are listed and explained.

Label table

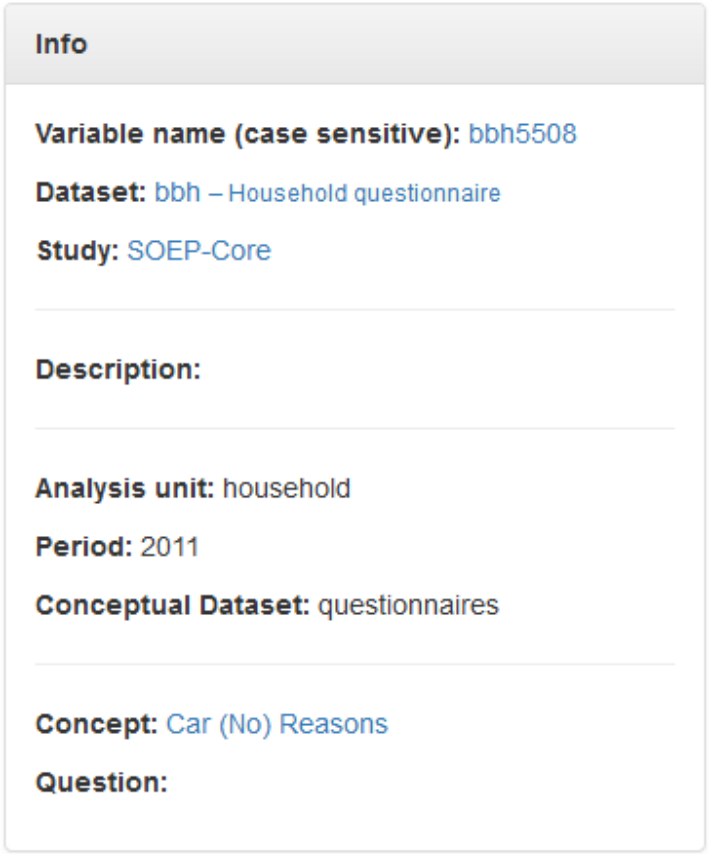
The label table provides you with an overview of label definitions across related variables to identify changes over time in longitudinal variables. The first number indicates the value code, the second number (in brackets) represents the frequency in the data. Please note that labels are simplified and values with frequency = 0 are hidden.

Variable:	rh5306	th5106	vh5408	xh5508	bbh5508	bdh5513	bgh7404
Dataset:	rh	th	vh	xh	bbh	bdh	bgh
version of questionnaire with modified filtering	-6 (0)	-6 (0)	-6 (0)	-6 (0)	-6 (0)	-6 (0)	-6 (0)
not included in this version of the questionnaire	-5 (0)	-5 (0)	-5 (0)	-5 (0)	-5 (4529)	-5 (3923)	-5 (3320)
inadmissible multiple response	-4 (0)	-4 (0)	-4 (0)	-4 (0)	-4 (0)	-4 (0)	-4 (0)
answer improbable	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)	-3 (0)
does not apply	-2 (9605)	-2 (9817)	-2 (9249)	-2 (9555)	-2 (9933)	-2 (11230)	-2 (11672)
no answer	-1 (125)	-1 (118)	-1 (211)	-1 (177)	-1 (167)	-1 (136)	-1 (99)
can not afford it	1 (827)						
financial reasons		1 (800)	1 (850)	1 (861)	1 (871)	1 (1310)	1 (1261)
other reasons	2 (1390)	2 (1326)	2 (1130)	2 (1096)	2 (1319)	2 (1494)	2 (1470)

The Label window shows you the absolute frequencies of the variable at different times of data collection. This makes it possible to identify initial trends in how response behavior has changed over a period of time. The assigned value code is output for each possible characteristic value and the absolute frequencies are displayed in parentheses.

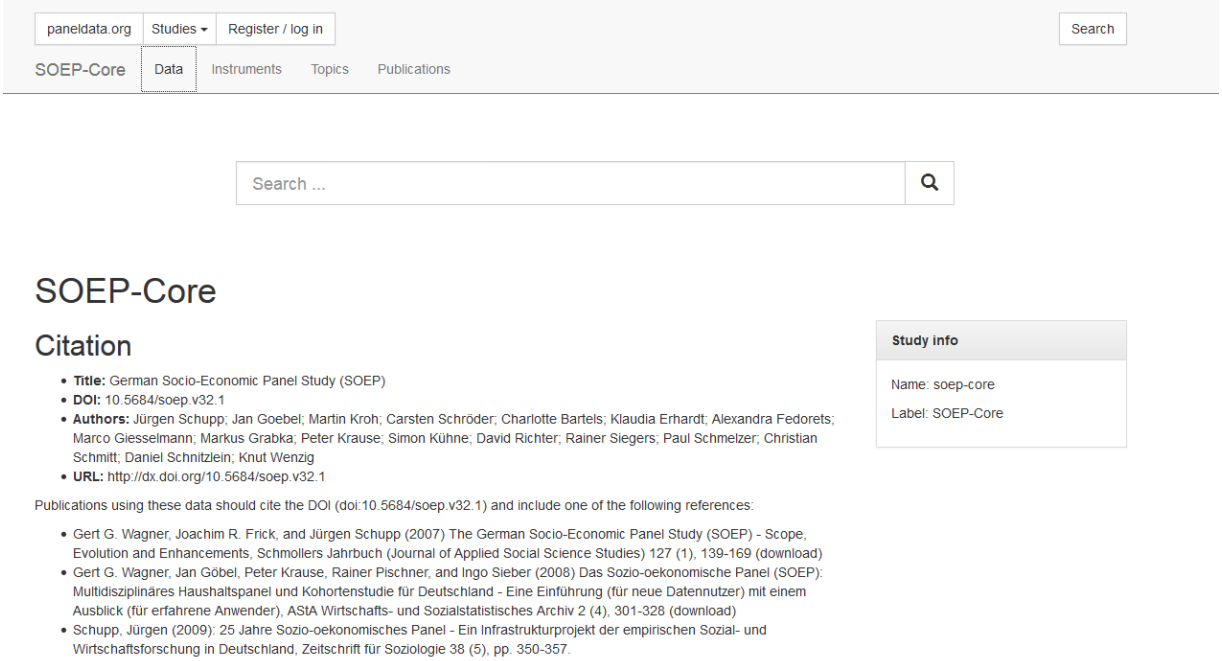
In our example, we see that for the variable “th5106” 800 respondents in the wave “t” (2003) stated “financial reasons” as the reason for not having a car in the household. For our variable “bbh5508” in survey year 2011 (wave “bb”), there were 871 respondents.

Paneldata.org is an excellent way to get an first overview of certain variables.

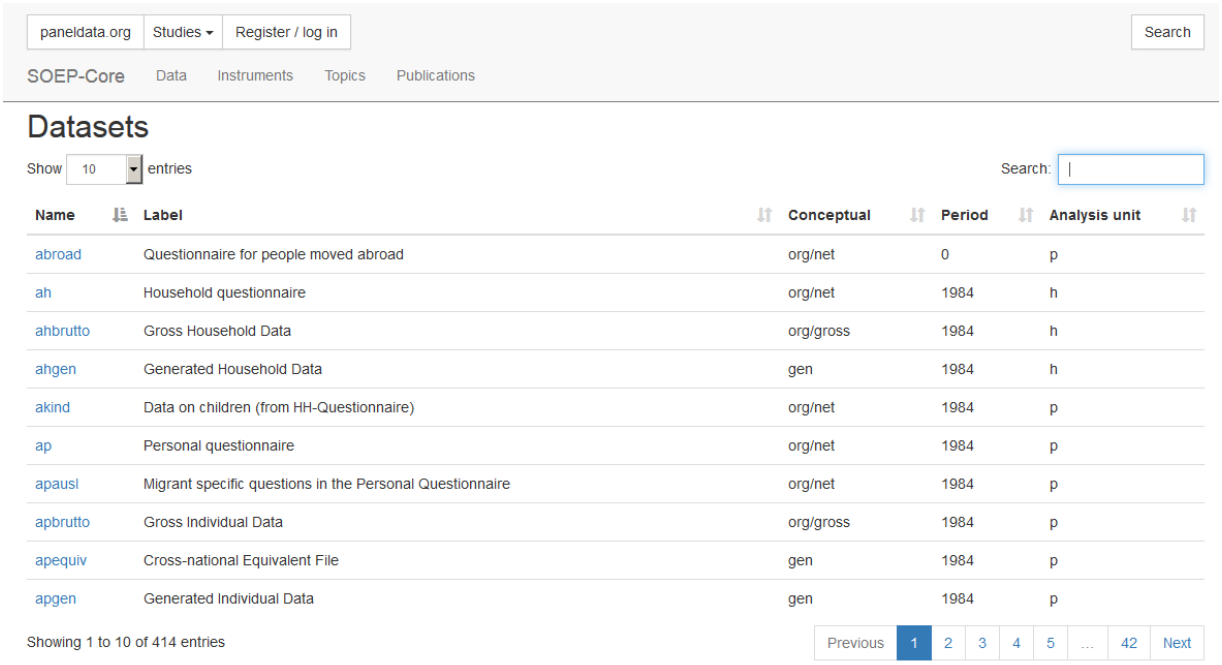


The information box on the right-hand side provides an overview of all relevant information about the variable and the dataset. In addition to basic information, you will find information about what kind of variable you are looking for under "Conceptual Dataset". In our example "bbh5508", you can see that variables with a "Conceptual Dataset: org/net" describe original variables that are assigned to a questionnaire. Generated variables are "Conceptual Dataset: gen". For an overview of the different dataset types in SOEP-Core, see the section *Data Distribution File*.

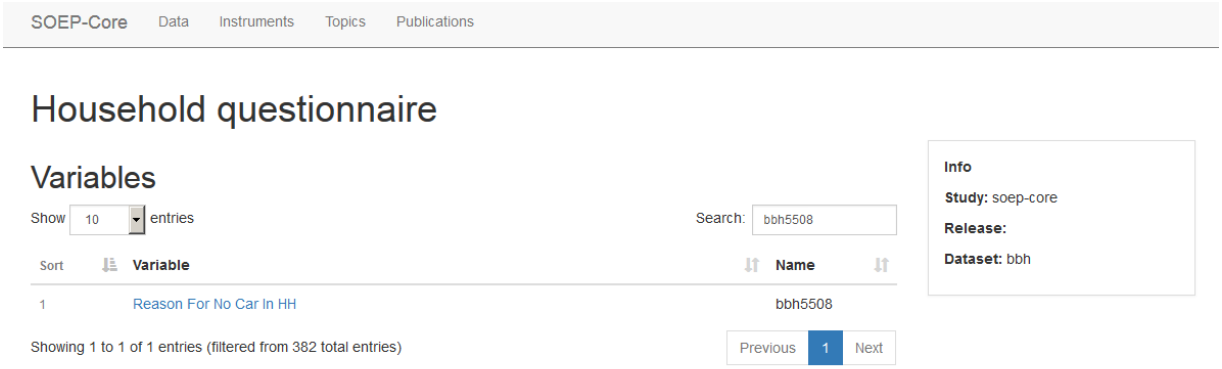
In addition to searching for keywords or using the various filter settings, you can also find what you are looking for directly in the data set search. Open paneldata.org, click on the study SOEP-Core and select the menu option "data".



Now you will be directed to an overview that shows you all datasets contained in SOEP-Core.



Enter the dataset you are looking for ("bbh") in the search bar at the top right and click on the dataset. You will be directed to an overview that shows you all variables from the "bbh" dataset.

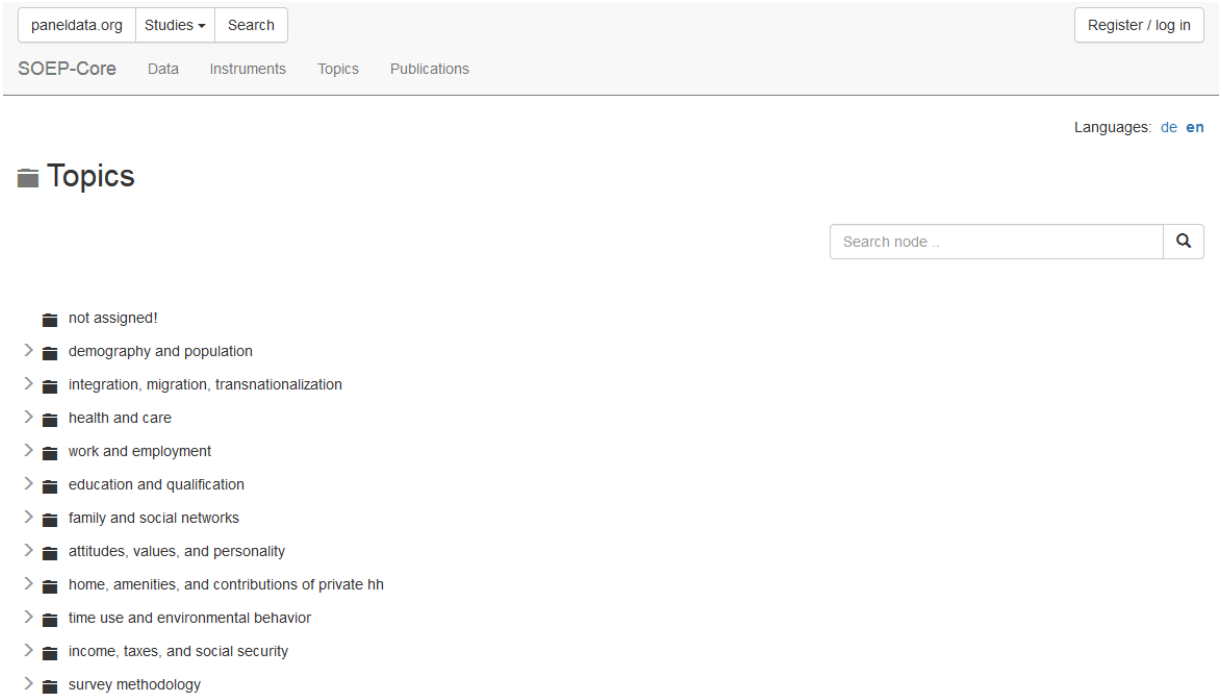


Now enter the variable you are looking for in the search bar at the top right and click on the variable of interest. You will be directed to the variable overview, where you will find detailed information on the variable. Paneldata.org offers a variety of search options to fit the user's search needs.

7.3 Topic Search with paneldata.org

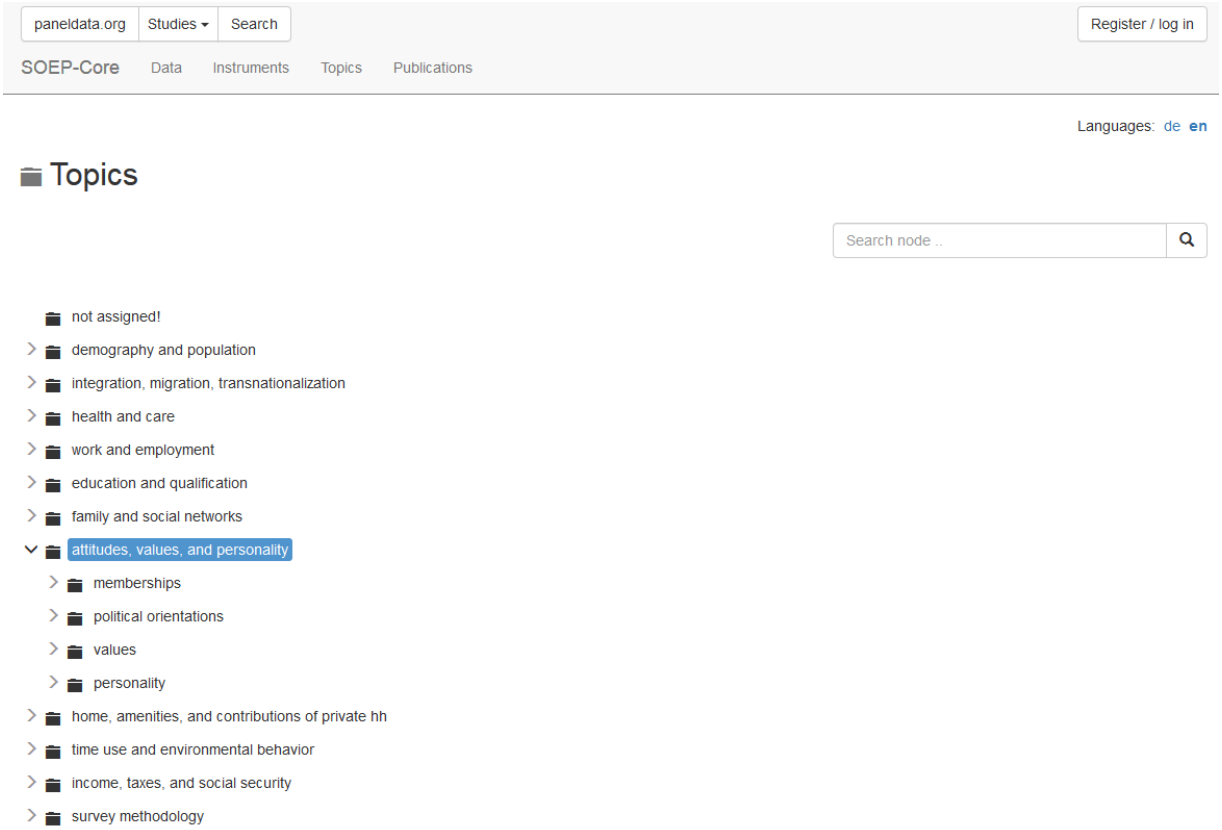
To provide an overview of the various topics in the SOEP, the variables have been grouped together on paneldata.org by topic. If you are looking for your research variables and do not want to check all datasets or questionnaires, the topic search on paneldata.org may help.

Open Paneldata and select the main study SOEP-Core. The upper navigation bar leads you to the Topics area. Click on Topics and look at the list of variables.

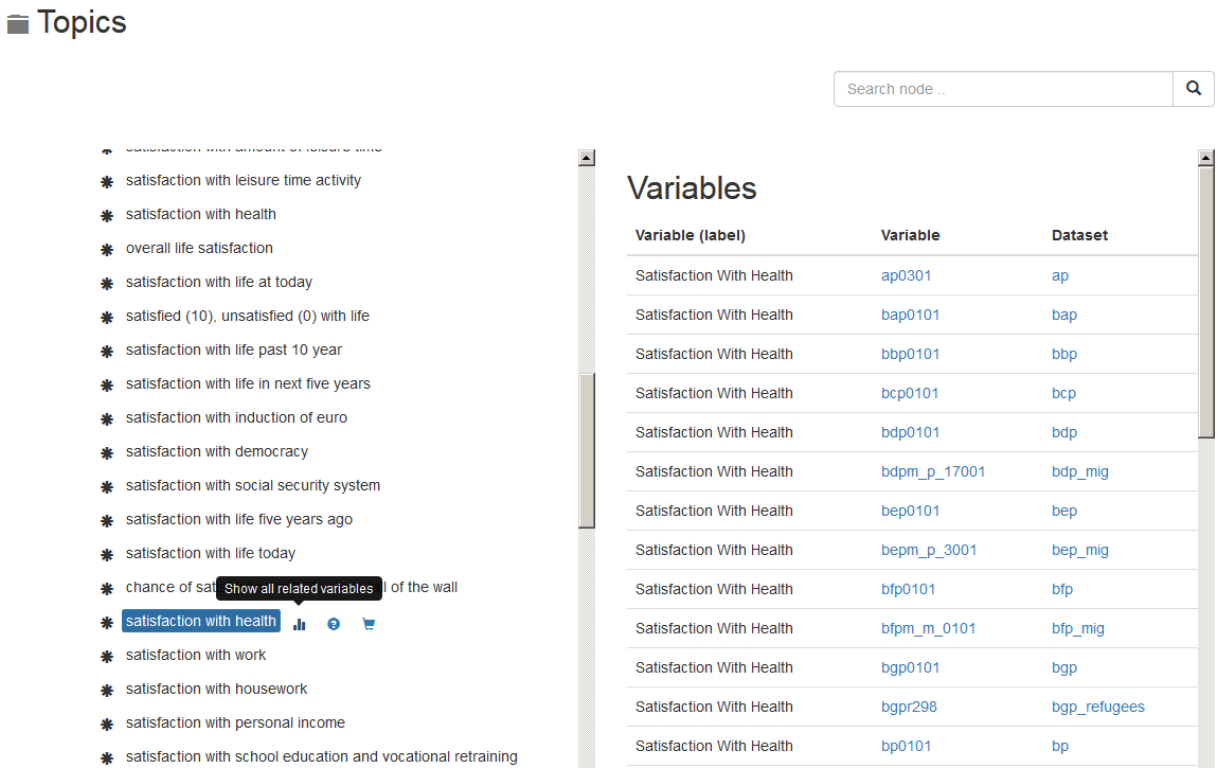


Select a topic that corresponds to your research interest, and a more detailed list of sub-topics will appear under the

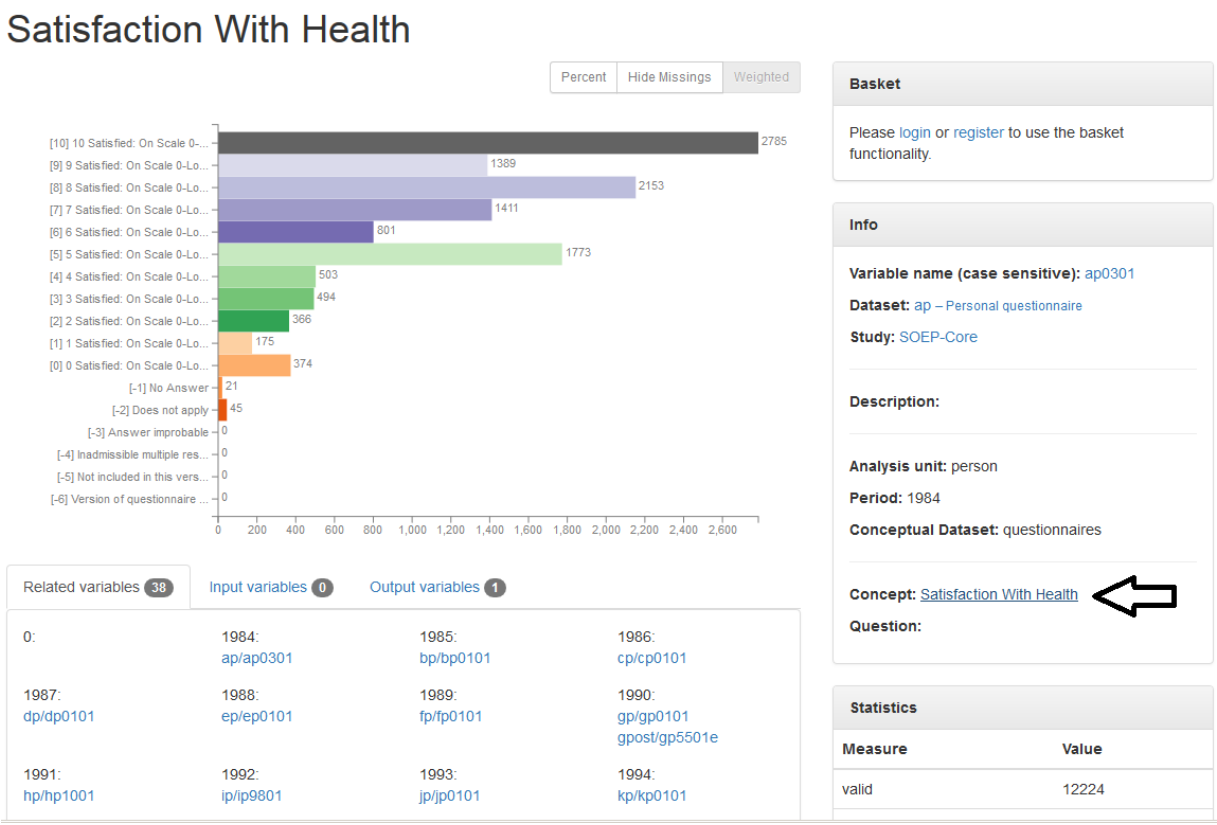
main topic heading.



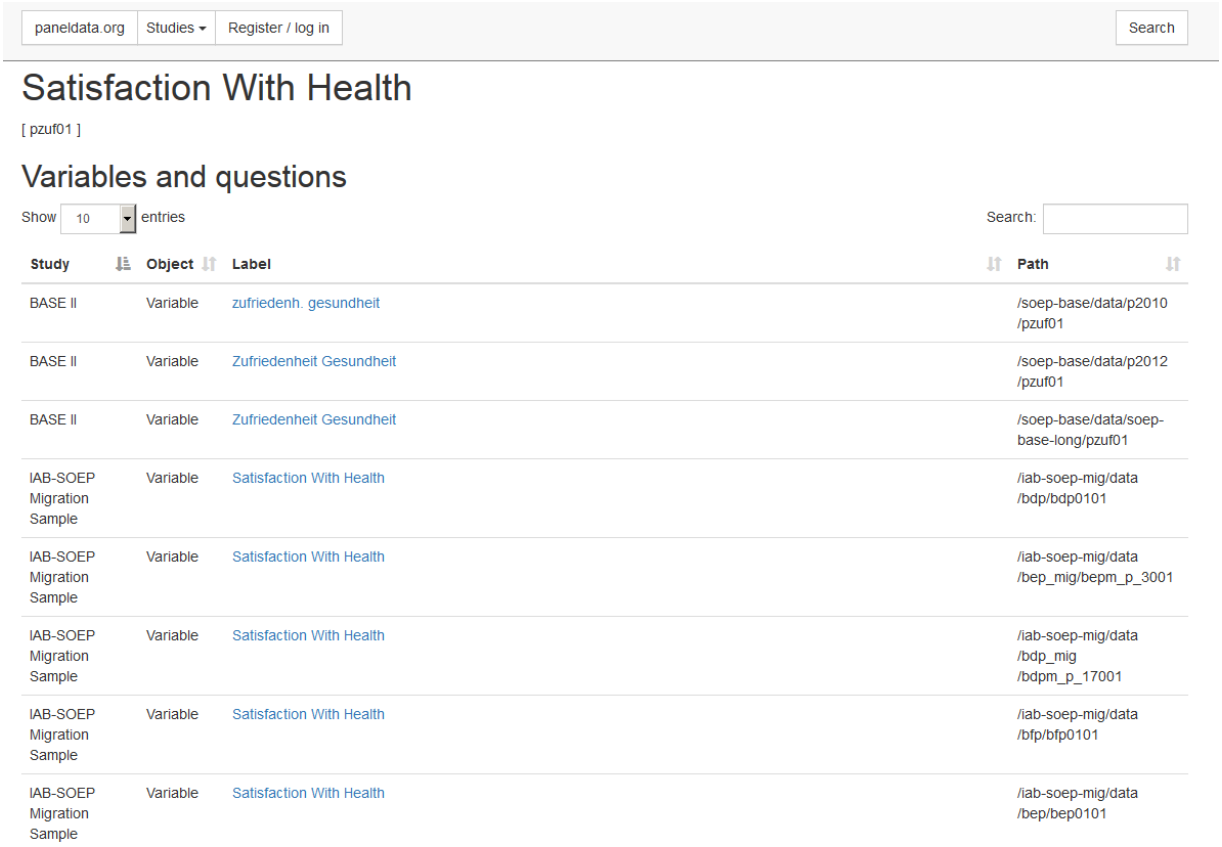
For example, if you are interested in different types of satisfaction, click on the topic “attitudes, values, and personality”. Underneath it, you will find the sub-topic “personality”. Suppose you are interested in health satisfaction. If you have found a suitable sub-topic, click on “show all the related variables”. All variables that fall under this topic will be displayed.



The paneldata topic list has three possible functions for each sub-topic. You can display all variables that belong to a sub-topic. In the future, paneldata will also display the texts of the questions from the SOEP questionnaires in which the variables in that sub-topic appear. Paneldata also allows you to keep variables from a sub-topic in a variable basket. The chapter *Syntax Generator on paneldata.org* explains in detail how to use the basket in your research and what possibilities this offers. Click on one of the variables to see the variable overview.



If you click on the concept of a variable, you will get to the concept overview. Concepts in SOEP are used to link variables with the same content. The concepts can even be used to link variables with the same content across studies.



paneldata.org | Studies | Register / log in | Search

Satisfaction With Health

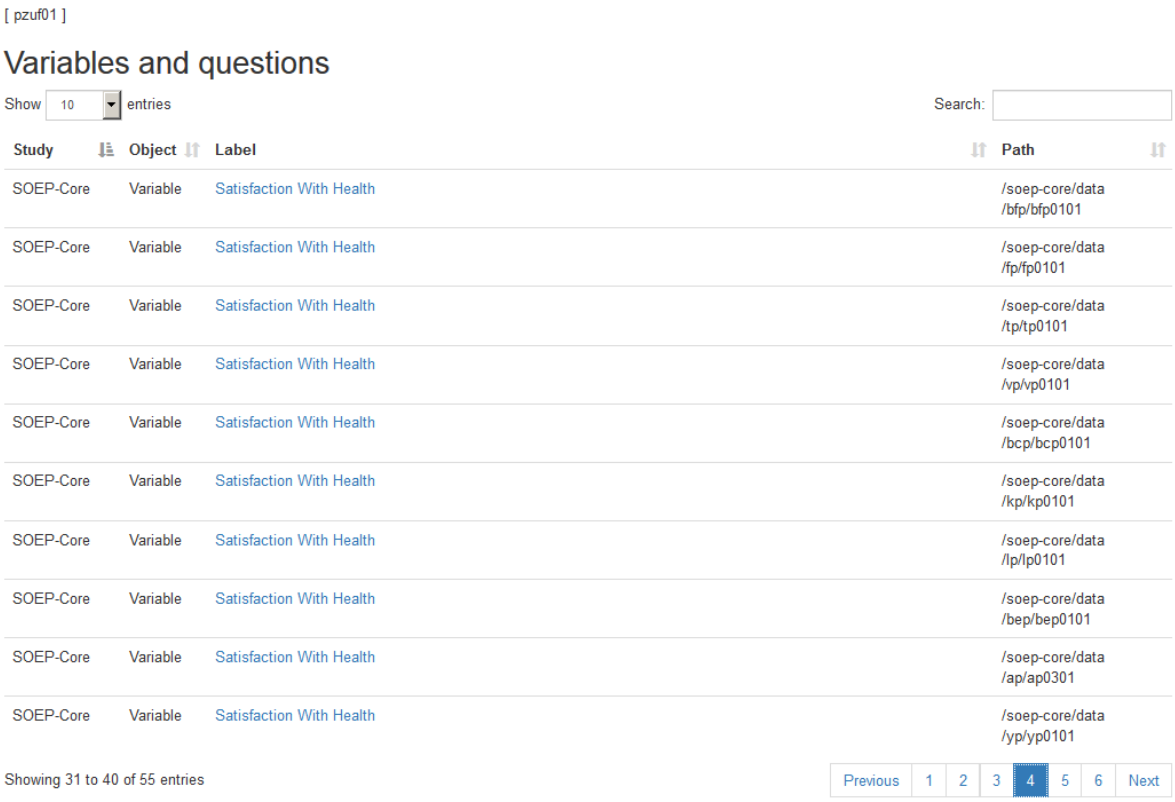
[pzuf01]

Variables and questions

Show 10 entries Search:

Study	Object	Label	Path
BASE II	Variable	zufriedenh. gesundheit	/soep-base/data/p2010 /pzuf01
BASE II	Variable	Zufriedenheit Gesundheit	/soep-base/data/p2012 /pzuf01
BASE II	Variable	Zufriedenheit Gesundheit	/soep-base/data/soep-base-long/pzuf01
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bdp/bdp0101
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bep_mig/bepm_p_3001
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bdp_mig /bdpm_p_17001
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bfp/bfp0101
IAB-SOEP Migration Sample	Variable	Satisfaction With Health	/iab-soep-mig/data /bep/bep0101

The concept overview displays the study- and wave-specific variables with this concept. The concept allows you to determine whether the variable you are looking for is also available and comparable across studies. In the column "Study" you can see which studies have the same variable linked by concept. The label of the respective variable is also displayed in the "Label" column. The column "path" shows the wave name of the variable. By clicking on the label, you will get to the overview of variables with all of the relevant information. The "Object" column in the concept overview shows you the type of information displayed.



[pzuf01]

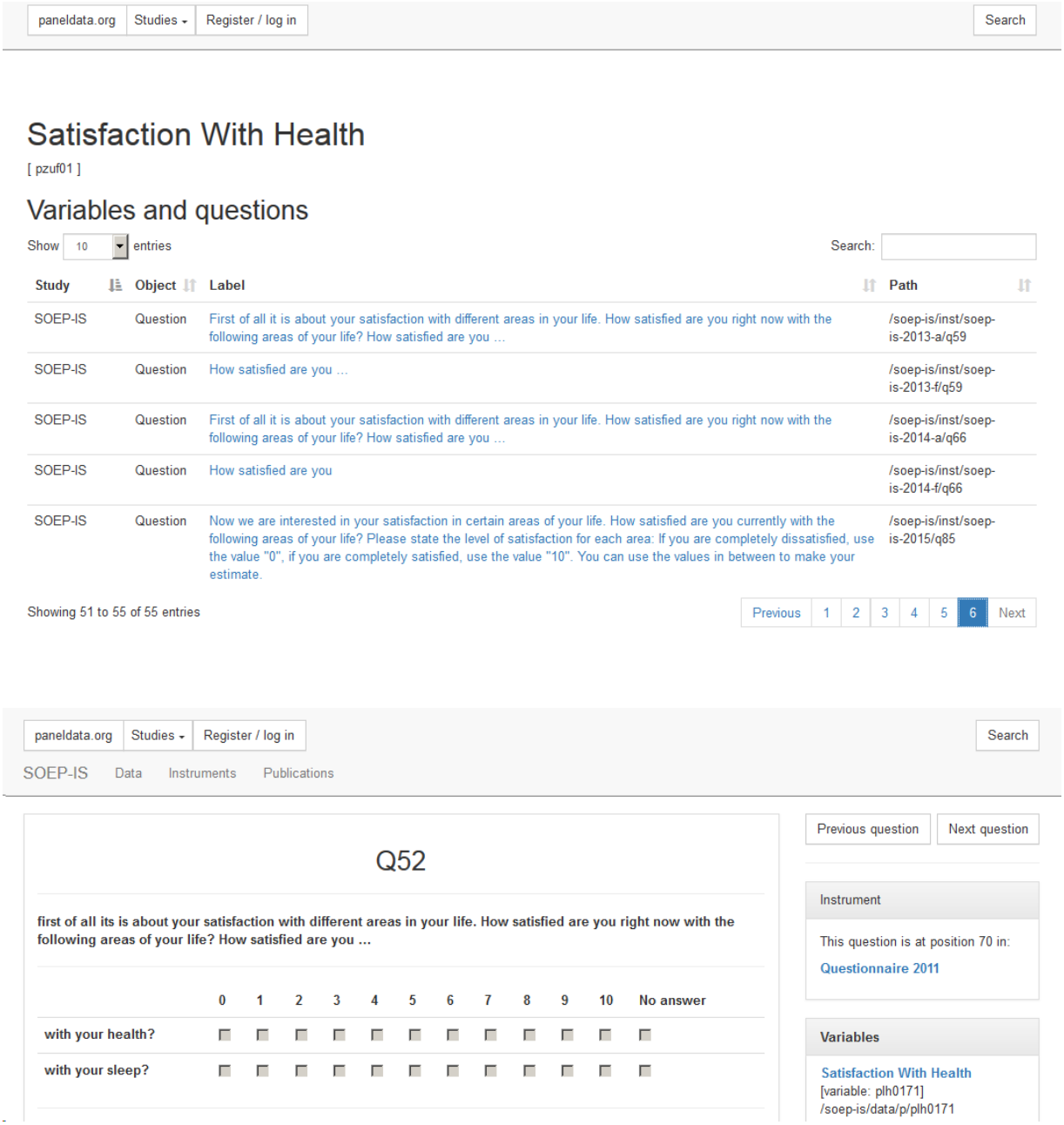
Variables and questions

Show 10 entries Search:

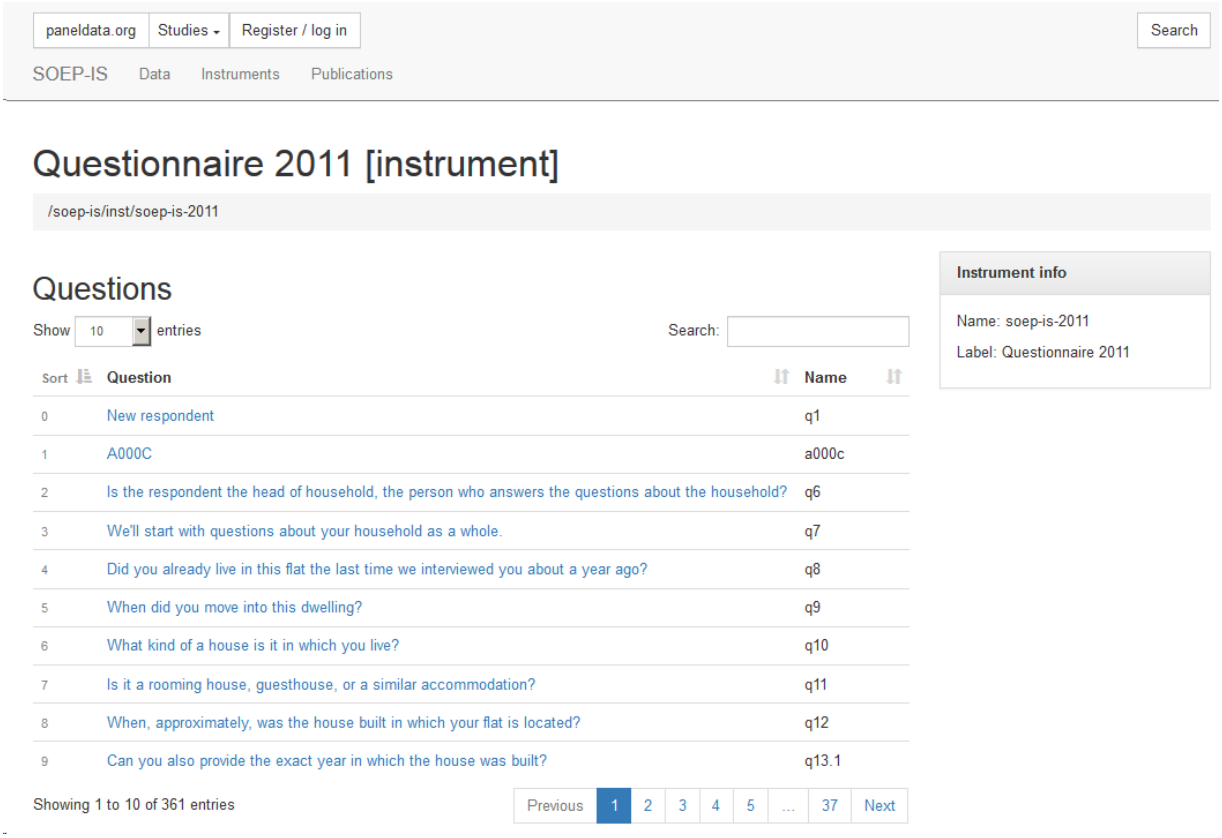
Study	Object	Label	Path
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /bfp/bfp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /fp/fp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /tp/tp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /vp/vp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /bcp/bcp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /kp/kp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /lp/lp0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /bep/bep0101
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /ap/ap0301
SOEP-Core	Variable	Satisfaction With Health	/soep-core/data /yp/yp0101

Showing 31 to 40 of 55 entries Previous 1 2 3 4 5 6 Next

In addition to the variables linked by concept, you can find the relevant questions in the concept overview. Questions are displayed in the "Object" column with question. Without having to open the questionnaire, you can read the question and identify possible differences. Click on the desired question and you will be taken to the question display.



Attention: To find out the exact wording of the question and possible filter structures, a variable search in the questionnaires is necessary. The question display in Paneldata only provides a quick overview. In the question overview, you can navigate through the questionnaire using the "next question" and "previous question" buttons. The "Instrument" section shows the position of the question in the questionnaire, the survey year, and links to the metadata-based survey instrument. Click on the survey instrument "Questionnaire 2011".



The survey instrument used in the SOEP-IS study in 2011 is now displayed. You can navigate through the questionnaire in this overview. The search bar allows you to search for research-relevant terms. Click on the question to access the question display.

Last change: Nov 13, 2019

7.4 Documentation on Generated Data

SOEP-Core contains a wide range of generated variables and datasets. To facilitate data use, we generate a large number of variables in the process of data preparation and release them with the SOEP-Core data. To make the generation process transparent to users, we provide comprehensive documentation on the numerous generated datasets and variables. For an overview, see our Documentation on Generated Data

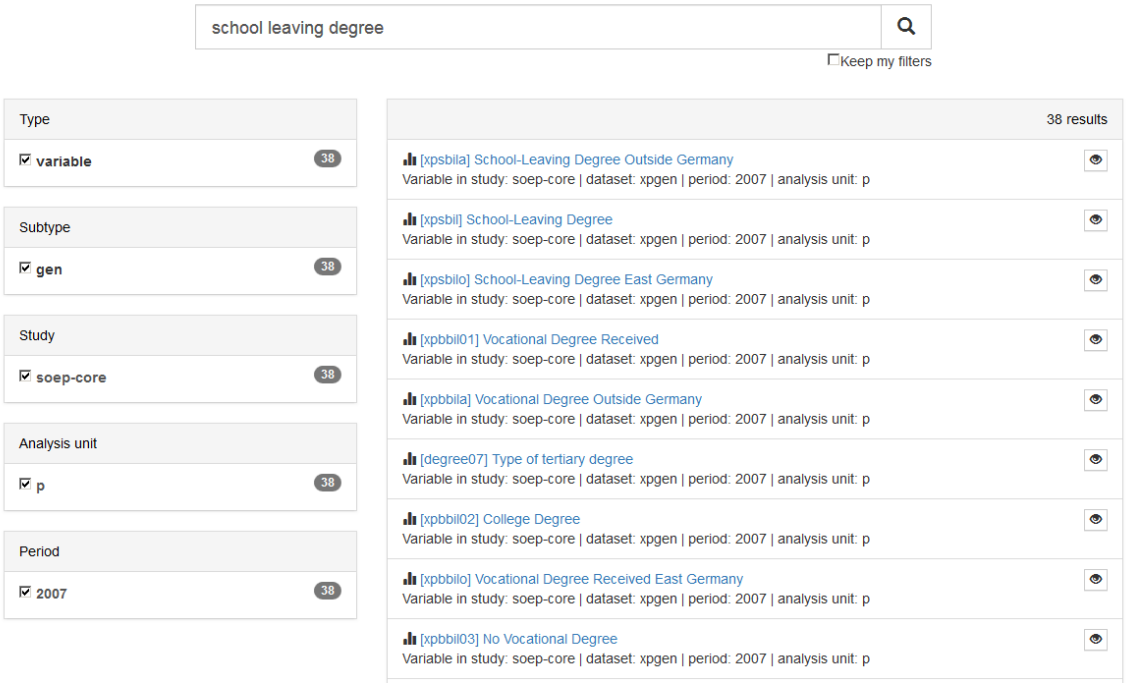
Example: A number of frequently used variables are provided in SOEP as "generated variables" (e.g., the datasets $PGEN and $HGEN). These variables are checked for consistency across waves. The documentation can be used to answer the following questions:

a) Which variable gives the highest school-leaving certificate attained by individuals surveyed in 2007?

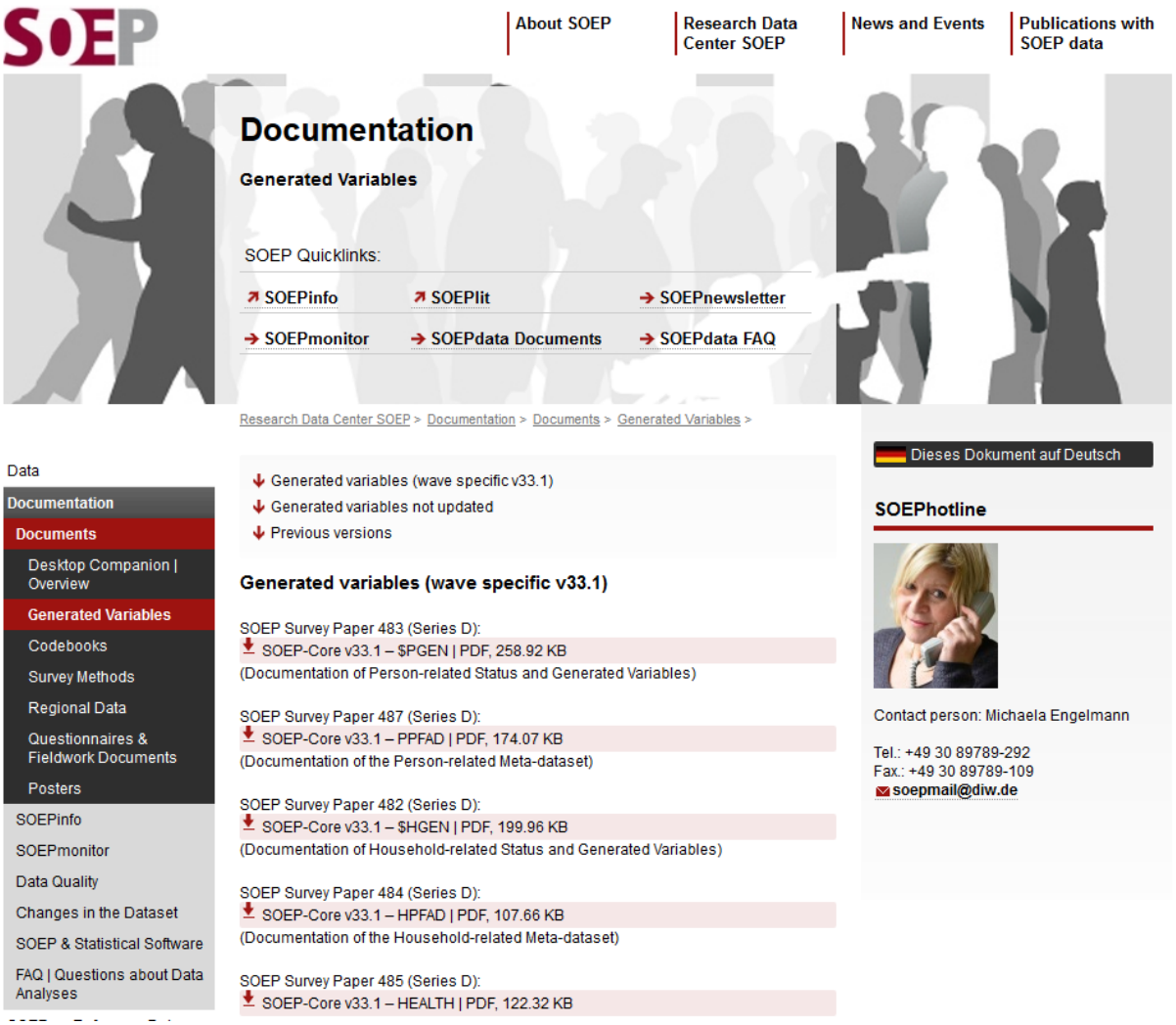
To search for the variable that provides this information, open Paneldata and enter "school-leaving certificate" in the search bar. Then specify your search by adjusting the filter settings as follows:

- type: variable
- subtype: gen
- study: soep-core

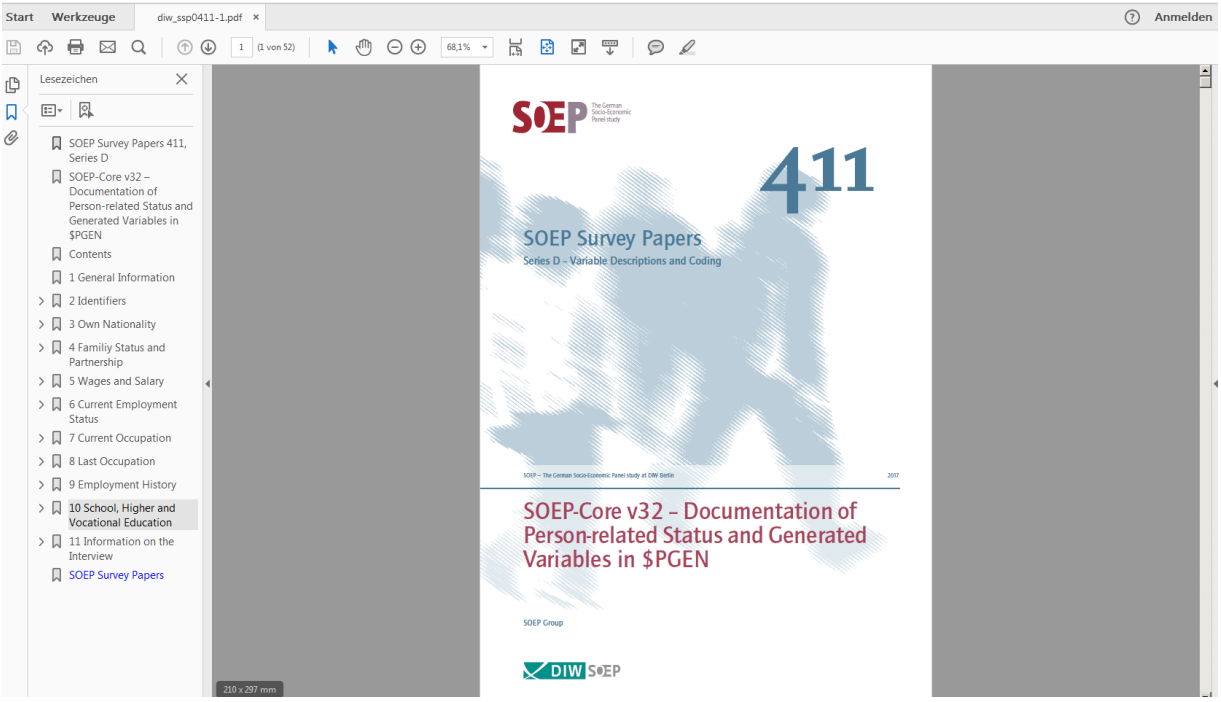
- analysis unit: p
- period: 2007



All variables could contain the information you are looking for. Since almost all variables in the search result come from the generated “xpgen” dataset, the documentation for the $pgen dataset should be used.



Now select the documentation of $pgen



The table of contents on the left gives you a classification of the dataset by topics. To find the variable you are looking for, select topic area 10.

$psbil – School-Leaving Degree [generic]

1	[1] Secondary School Degree	6411
2	[2] Intermediate School Degree	7293
3	[3] Technical School Degree	1515
4	[4] Upper Secondary Degree	5729
5	[5] Other Degree	4244
6	[6] Dropout, No School Degree	673
7	[7] Currently In School	779
-1	[-1] No Answer	1099
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0

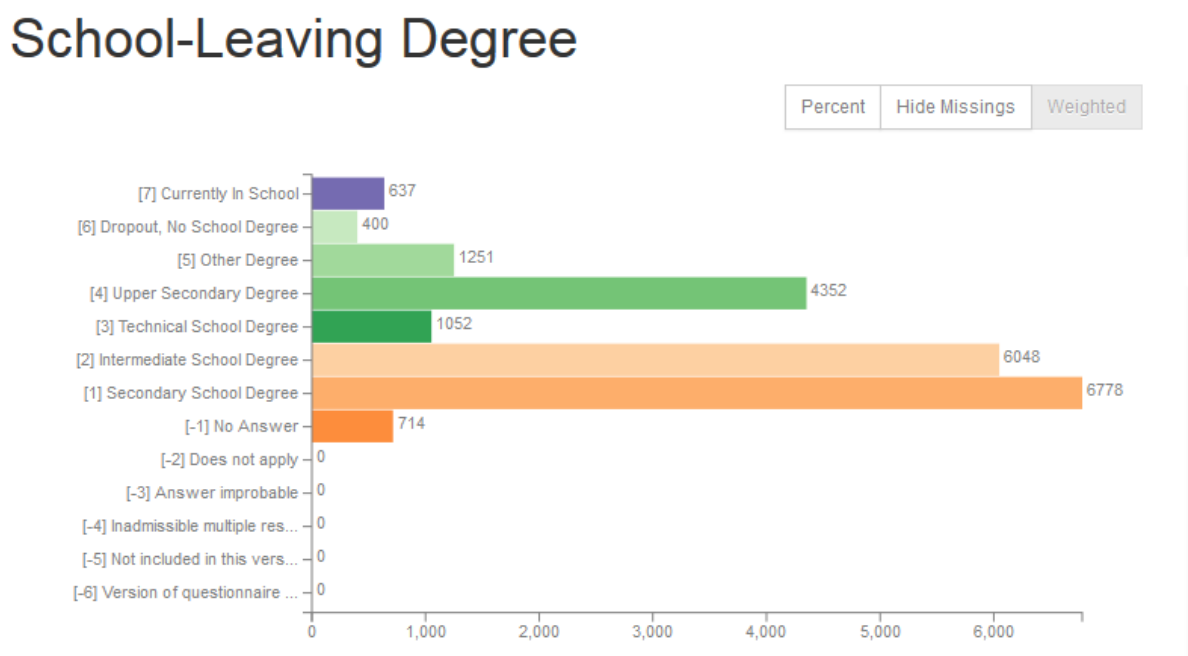
Waves: all
All respondents in all SOEP subsamples are asked about diplomas/degrees attained for completion of secondary/tertiary education (1984–1993 blue questionnaire; since 1994 biographical questionnaire) the first time they participate in SOEP. First: to generate this variable, the different diploma/degree categories provided for Subsamples B and D (see $PSBILA) as well as C (see $PSBILO) are integrated into the West German diploma/degree categories (Subsample A) and continued on in this form. Second: this data is regularly updated to take into account any changes in highest diploma/degree attained. With the survey of 2000, all educational information was collected again and is reflected in the variables. [This information can be related to a specific variable and is not necessary generic.]
For more information, contact: Peter Krause (Tel. +49-30-89789-690)

After a few searches, you will find the variable you are looking for. The documentation provides useful information about the generated variable: it comes from the biography questionnaire, which was introduced in 1994 and is administered only once per respondent. The documentation also explains the two additional variables $psbila and $psbilo in more detail: the $psbil variable is updated regularly to take into account possible changes in the respondent's highest school-leaving certificate. For this reason, the generated variable is useful in providing the most up-to-date information on completed secondary schooling.

The variable we are looking for is xpsbil and describes the highest degree in certificate attained by individuals surveyed since 2007.

b) What values do individuals with an upper secondary school-leaving certificate (Abitur) have for this variable??

Since you now know the variable you are looking for, you can use the extensive functions of paneldata.org in addition to the information from the documentation. If you search for the variable "xpsbil" in paneldata.org and click on it, the frequency counts are displayed.



In addition to the absolute and relative frequencies, you can also read the value codes of specific response categories. A translation of the answer categories can be found in the "Label translations" section:

Label translations

	en	de
label	School-Leaving Degree	Schulabschluss
-6	[-6] Version of questionnaire with modified filtering	[-6] Fragebogenversion mit geaenderter Filterfuehrung
-5	[-5] Not included in this version of the questionnaire	[-5] In Fragebogenversion nicht enthalten
-4	[-4] Inadmissible multiple response	[-4] Unzulaessige Mehrfachantwort
-3	[-3] Answer improbable	[-3] nicht valide
-2	[-2] Does not apply	[-2] trifft nicht zu
-1	[-1] No Answer	[-1] keine Angabe
1	[1] Secondary School Degree	[1] Hauptschulabschluss
2	[2] Intermediate School Degree	[2] Realschulabschluss
3	[3] Technical School Degree	[3] Fachhochschulreife
4	[4] Upper Secondary Degree	[4] Abitur
5	[5] Other Degree	[5] Anderer Abschluss
6	[6] Dropout, No School Degree	[6] Ohne Abschluss verlassen
7	[7] Currently In School	[7] Noch kein Abschluss

You can answer the question without opening the data. In the 2007 survey year, the variable “xpsbil” with the value code “4” describes the response category “upper secondary school-leaving certificate (Abitur)”.

Last change: Nov 13, 2019

7.5 Working with SOEPhelp

Attention: The following tool is available starting with Version v34 (Wave bh) and Stata Version 12.

The SOEP data contain a wide array of useful additional information. SOEPhelp is a stata.ado that displays documentation on the dataset at hand. It displays information such as variable histories directly in your Stata window.

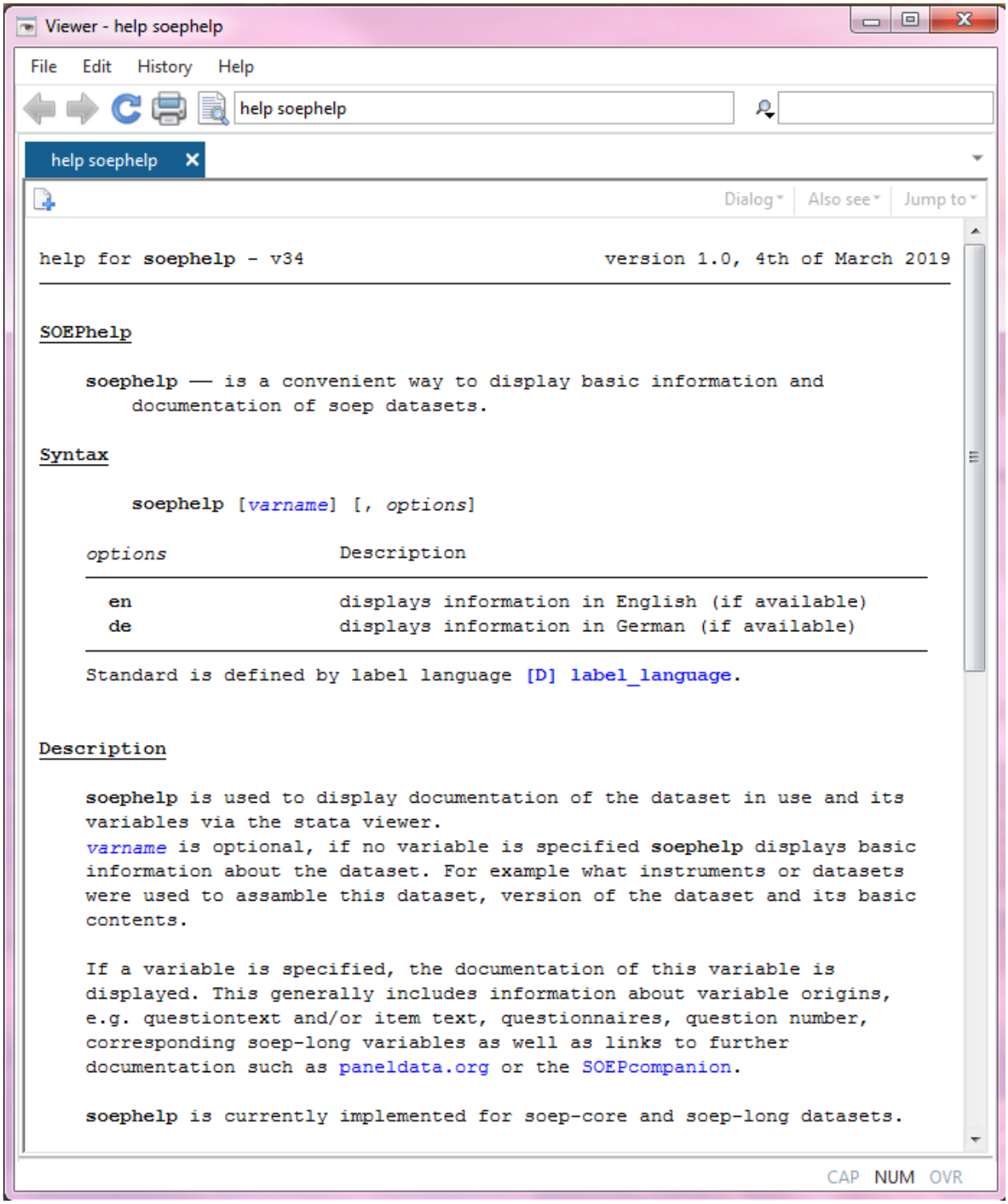
Installation

Open Stata and enter the following command:

```
net install soephelp, replace from(http://companion.soep.de/SOEPhelp/)
```

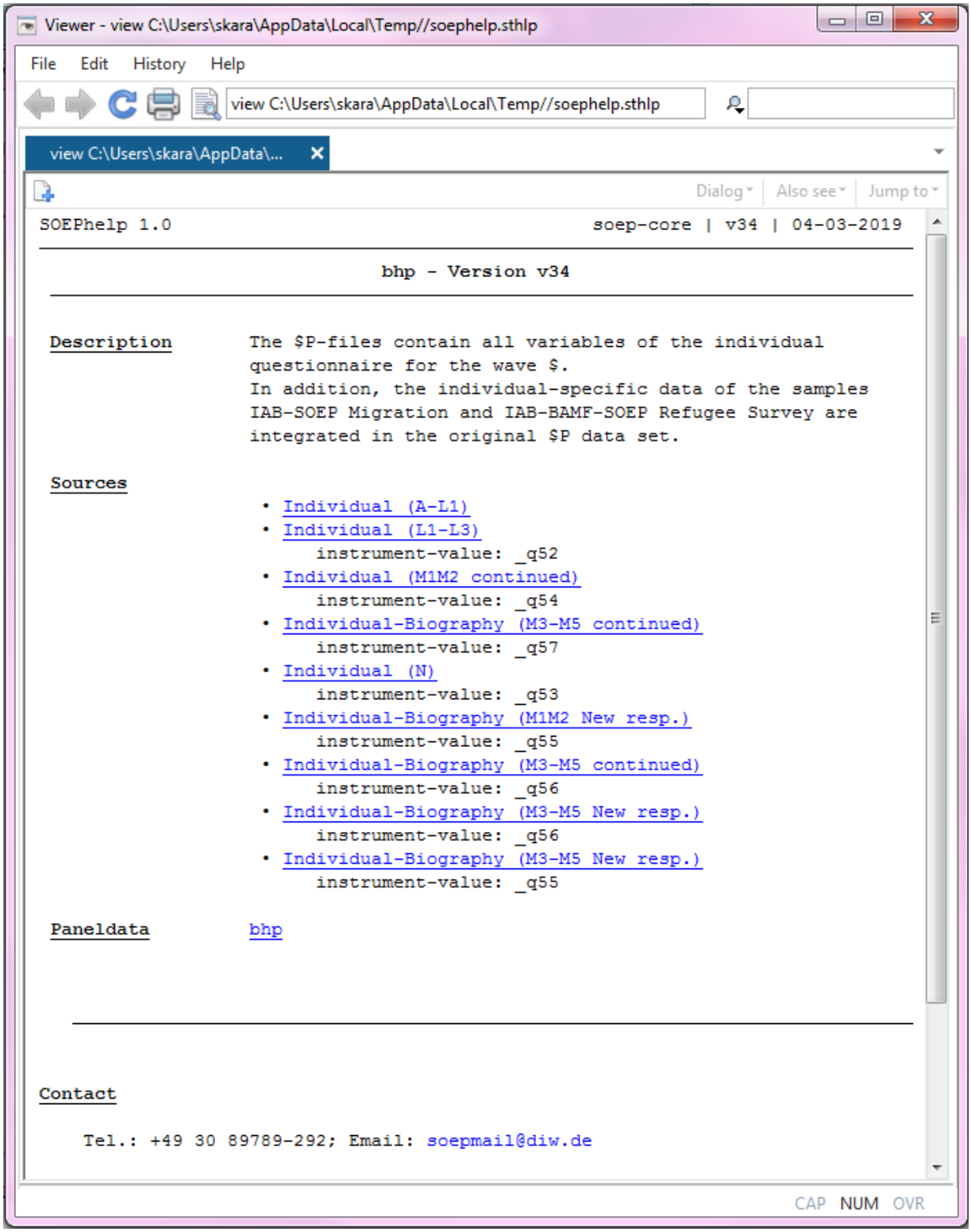
The following commands are provided by .ado:

For a general introduction to SOEPhelp, type in the command `help soephelp`. Here you will find a detailed explanation of the Stata.ado and the different ways to use it. The .ado is available in German and English.

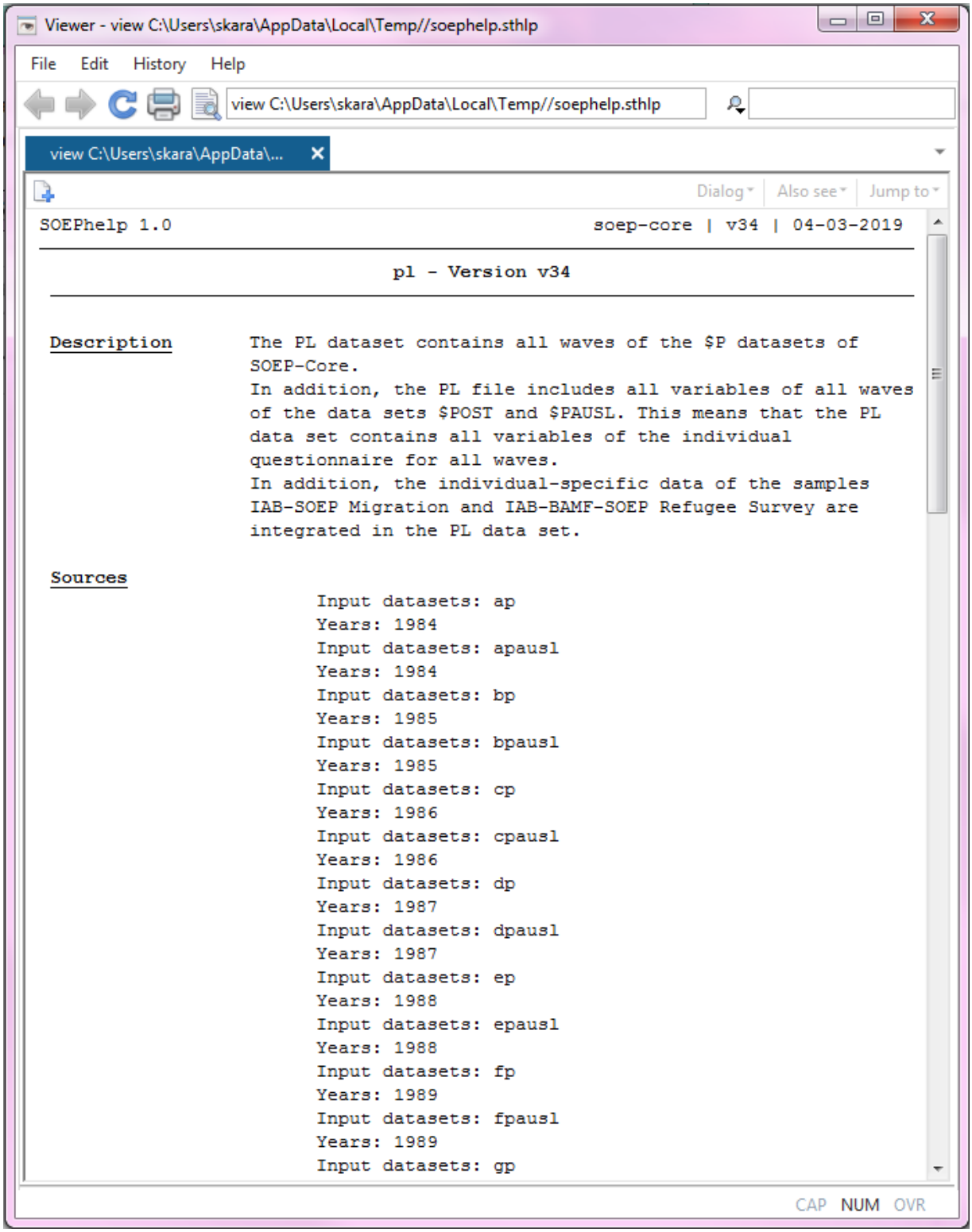


With the command `soephelp`, using wave specific datasets (subdirectory *raw*), you receive a basic description of the

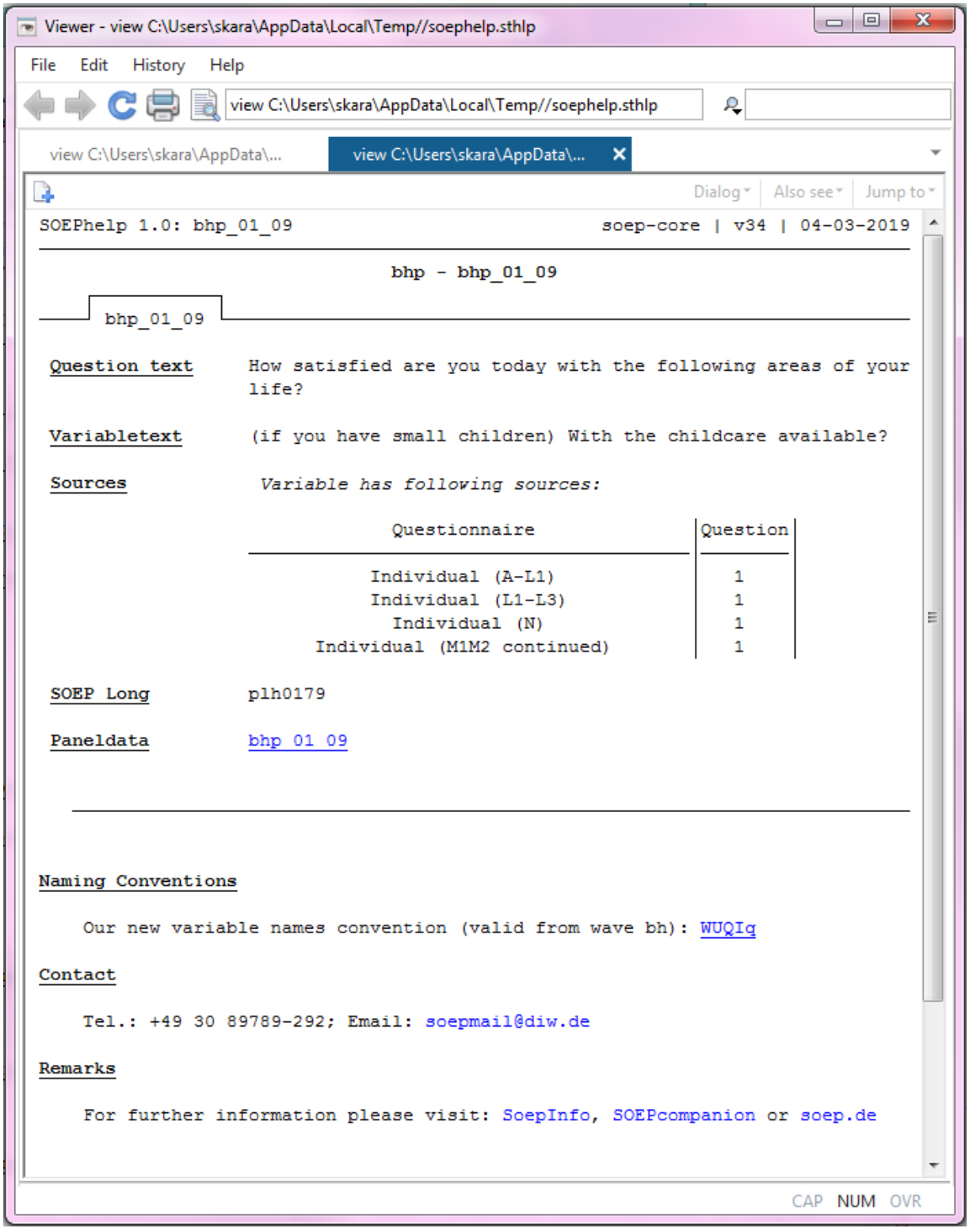
dataset as well as a list of samples contained in it, including the instruments corresponding to the sample. By clicking on the provided links, you will get to the respective questionnaires or to the dataset on paneldata.



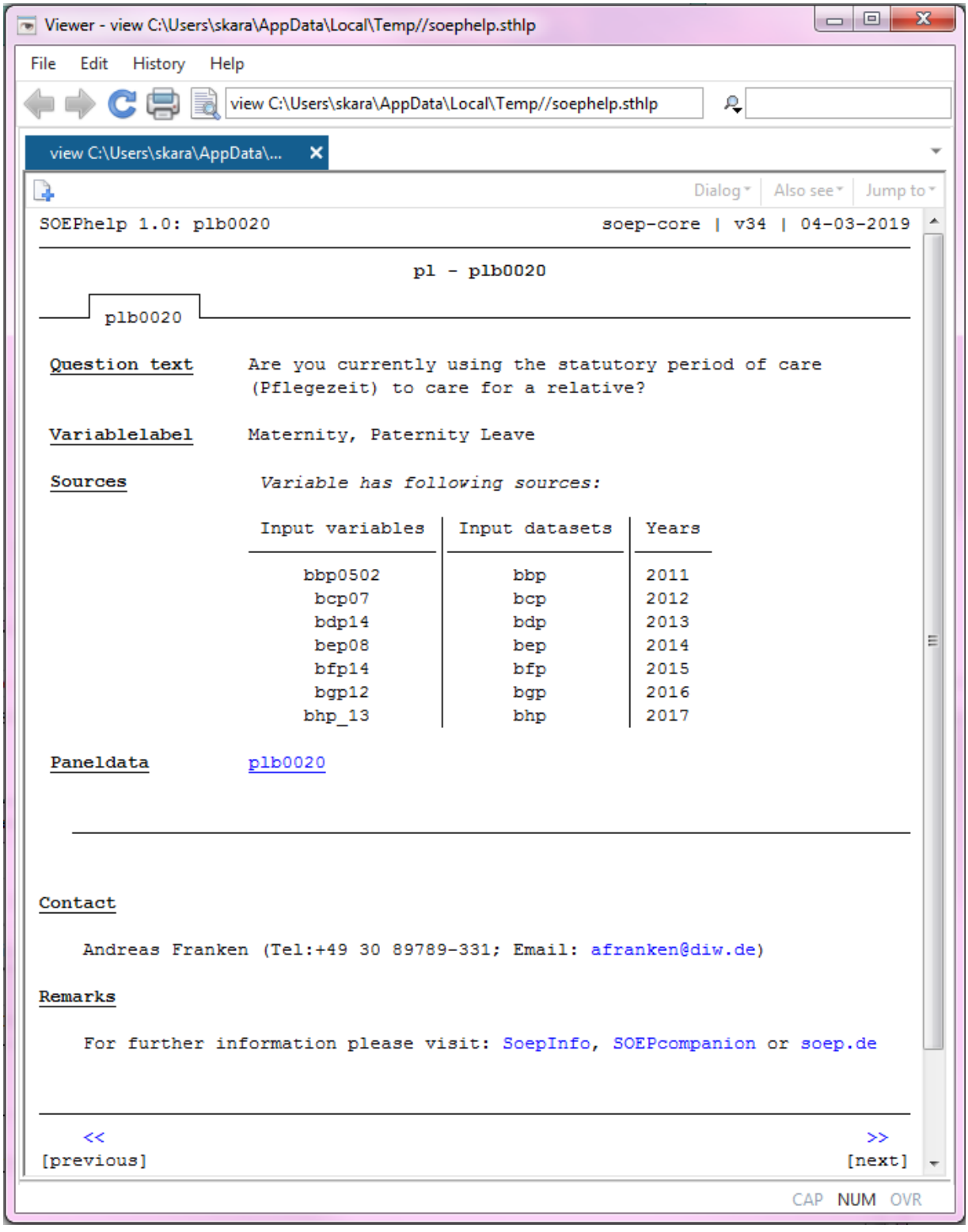
Using `soephelp` in longitudinal datasets, you also receive a basic description as well as a list of wave-specific datasets that are used to generate the longitudinal version.



If you enter the command `soephelp <variable>` in a wave-specific data set, you will get detailed information about the variable in question. The question asked in the questionnaire is displayed and in which samples and instruments the question was asked. Additionally, the command offers the corresponding long variable as well as the link of the displayed variable to the documentation at paneldata.org.



Conversely, with long data, you receive the wave-specific input variables and datasets used to generate the long-variable.

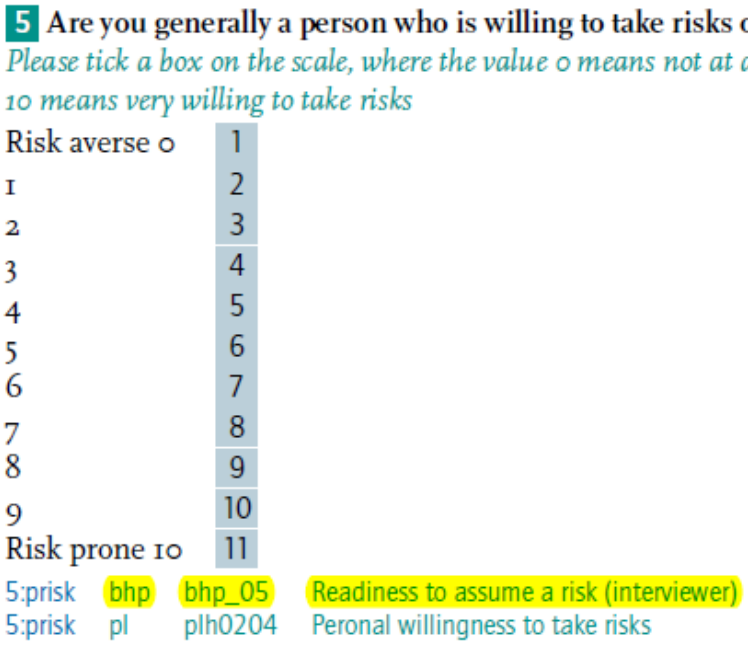


SOEPhelp is directly linked to the SOEPcompanion.

Last change: Nov 12, 2019

7.6 Working with Metadata-Based Questionnaires

Metadata-based questionnaires make it considerably easier to find the variables of interest from the perspective of the questionnaire. Each of the generated PDFs reflects a questionnaire. With the help of these documents the user learns which questions have been asked in the respective sample and in which sequence. In addition, the documents make it clear what the question variable is called and which dataset it can be found in. The example shows question 5 from the individual questionnaire of SOEP-Core, which can be found in the data set bhp under the variable name bhp_05.

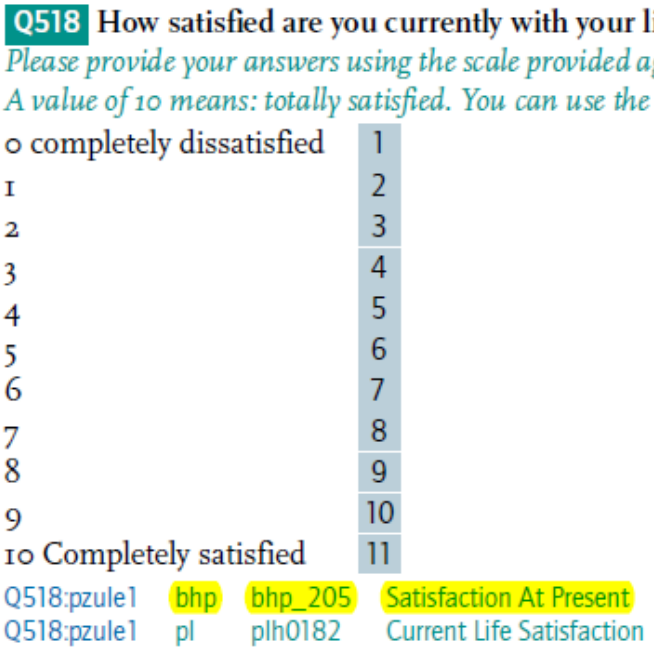

5 Are you generally a person who is willing to take risks or do you try to avoid taking risks?
Please tick a box on the scale, where the value 0 means not at all willing to take risks and the value 10 means very willing to take risks

Risk averse 0	1
1	2
2	3
3	4
4	5
5	6
6	7
7	8
8	9
9	10
Risk prone 10	11

5:prisk	bhp	bhp_05	Readiness to assume a risk (interviewer)
5:prisk	pl	plh0204	Peronal willingness to take risks

1. Example: Integrated Variable

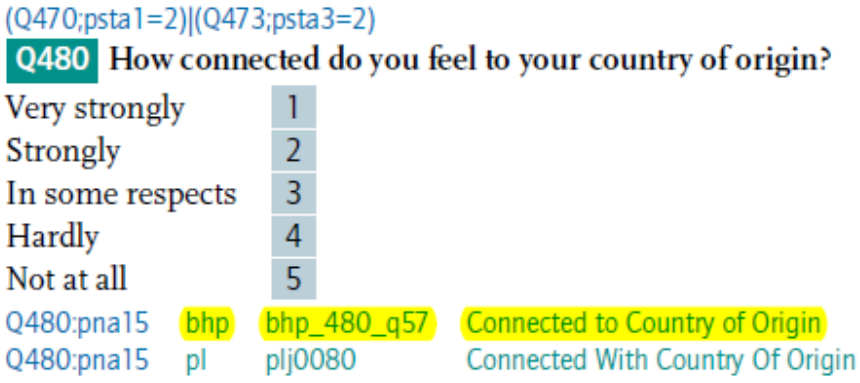
Let's say you're interested in finding out about refugees' general life satisfaction. Search the questionnaire to find which refugees were surveyed for a second time in 2017. You'll find what you're looking for under question Q518. Below the question is the information on the name of the variable and the dataset where it is found.



The general satisfaction with life can be found in the dataset bhp under the name bhp_205.

2. Example: Additional Variable

Let's say you're interested in finding out about countries or origin. You want to know specifically how connected respondents feel to their country of origin. You'll find the question in the questionniare given to refugees participating in the survey for the second time or more under question number Q480.



The information on the question is stored in the data file bhp under the name bhp_480_q57. The name indicates that the question is not in the samples A-M2 because it has the suffix _q57. This does not preclude the question from being further down the integration hierarchy in questionnaires.

Last change: Nov 12, 2019

CHAPTER

EIGHT

CONTACT INFORMATION

The first version of the SOEPcompanion (formerly Desktop Companion) was published as a PDF document by John P. Haisken-DeNew and Joachim R. Frick in September 1996. It was originally intended to give novice users a broad introduction in understanding the SOEP, its structure, depth, and research potential. The Desktop Companion was updated several times between 1996 and 2005. The first major change came in 2014, when Jan Goebel and Mathis Schröder decided to shorten the Desktop Companion to its most important content and make it web-based.

The new, completely edited version of the SOEPcompanion (formerly Desktop Companion) has a strong focus on the use of the SOEP-Core data from the perspective of a data user who has received our most recent data release from the SOEP Research Data Center. This new version is not only a web-based documentation, we also offer it as a download.

Address: SOEP, DIW Berlin, Mohrenstraße 58, 10117 Berlin, Germany

Homepage: http://www.diw.de/soep

E-Mail: soepmail@diw.de

SOEPhotline: +49 30 89789-292

Developers: Selin Kara, Stefan Zimmermann