DTC Documentation

Release 1.2

SOEP

Sep 05, 2018

CONTENTS

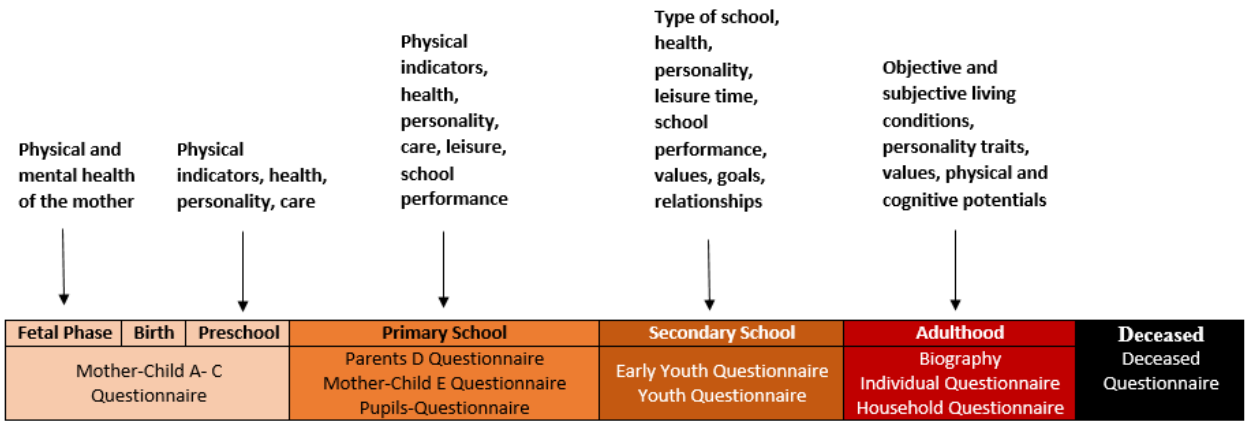
CHAPTER

ONE

CONTENTS OF SOEPCORE

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

Life History



The SOEP questionnaires are designed in such a way that people in a SOEP household can be analysed from birth to adulthood. In addition to the *Youth Questionnaire*, which was conducted for the first time in 2000/01, a series of questionnaires for certain cohorts of children living in SOEP households has been introduced since 2003. These are filled in every year 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 (0-1 years) *Mother-Child Questionnaire A (Age 0-1)*. The following instruments were developed in such a way that this starting cohort (born 2002/ 2003) can be followed up in its development and analysed longitudinally. This was followed in 2005 by a questionnaire for the mothers of 2-3-year-old children *Mother-Child Questionnaire B (Age 2-3)* and in 2008 by a questionnaire for 5-6-year-olds *Mother-Child Questionnaire C (Age 5-6)*. In 2010, the questionnaire for 7-8-year-old children *Parents Questionnaire D (Age 7-8)*, completed by both mothers and fathers, was launched. In 2012, the questionnaire for 9-10-year-old children *Mother-Child Questionnaire E (Age 9-10)* was the last questionnaire to be answered by the mothers. This was followed by two youth instruments in which the children, aged 12 *Pupils Questionnaire* and 14 *Early Youth Questionnaire* respectively, answered questions about their own life situation for the first time. These were introduced in 2014 and 2016 respectively, so that in 2018 the first cohort went through the complete battery of age-specific instruments for the first time and then, as an adult, will answer annually thematically changing topics of the long-term SOEP study. As soon as the age of 18 is reached, each person in a SOEP household

receives the *Individual Questionnaire*, the head of the household additionally receives the *Household Questionnaire*. As soon as a person dies, regardless of whether this person is part of a SOEP household, the *Deceased Persons Questionnaire* is handed over to the person providing the information.

1.1 SOEP Topics

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

1.1.1 Attitudes, Values and Personality

The character of a person offers a variety of analysis possibilities. Information about the personality of the respondents, their political orientation, concerns, satisfaction, willingness to take risks and much more can be found in the "Attitudes, Values, and Personality" section.

1.1.2 Demography and Population

In this topic you find various information about the birth dates, no matter if interviewer, children, siblings or parents. Furthermore, there is data on places and history of births in households. The household sizes and relationships between the different persons in a household are also listed, as are the sexes of all persons involved.

1.1.3 Education and Qualification

Education is one of the cornerstones of our society today, and the information that can be obtained through the SOEP is numerous. Whether school achievement, vocational training or academic success in this section is everything about the education of people. The school history, reasons for lack of further training, educational goals and so on. Furthermore, basic skills of children can be found here to, whether they are able to speak in whole sentences or use scissors, for example.

1.1.4 Family and Social Networks

As a household study, the SOEP determines rich information about family and social contacts and how these relationships change at different stages of life. The whole cycle of life with its wonderful and sad facets and a wide range of information is shown in this section: Pregnancy - birth - parenthood - kinship - circle of friends - marriage - divorce - death. And of course many more data can be found here.

1.1.5 Home, Amenities and Contributions of Private HH

In this section you will find information about the household and everything that has to do with everyday life. What kind of home do you live in? Are you an owner or a tenant? Which expenses do you have on things like personal hygiene, the car or holidays? Who's taking care of the kids? All this and much more information about living, its costs or the living environment can be seen here.

1.1.6 Health and Care

On the subject of health, numerous personal data such as the number of doctoral visits and habits like sport or alcohol consumption are recorded. There are also information on health insurance, health status and grip strength. However, health information from other people such as children or deceased persons are also displayed.

1.1.7 Integration, Migration, Transnationalization

Migration and establishment processes are changing society. With its large number of migration samples and specific migration questions, the SOEP can cover these research topics comprehensively. The area "Integration, Migration, Transnationalization" offers you analysis possibilities on migration history, discrimination, interethnic contacts, education, cultural integration, transnational relations, identification with Germany and the intention to stay.

1.1.8 Income, Taxes and Social Security

Income and finances are an essential part of our everyday life. How much money is earned and how much is spent. Child benefit, pensions, inheritance or salary, but also taxes and debts belong to this topic. No less interesting is the information on other assets such as real estate or property, plant and equipment.

1.1.9 Survey Methodology

In the "Survey Methodology" section you will find many relevant variables on imputation, weighting, field work in SOEP core, identifiers, the interviewers' working methods,survey methods and also information about our respon-

dents' exit from the survey.

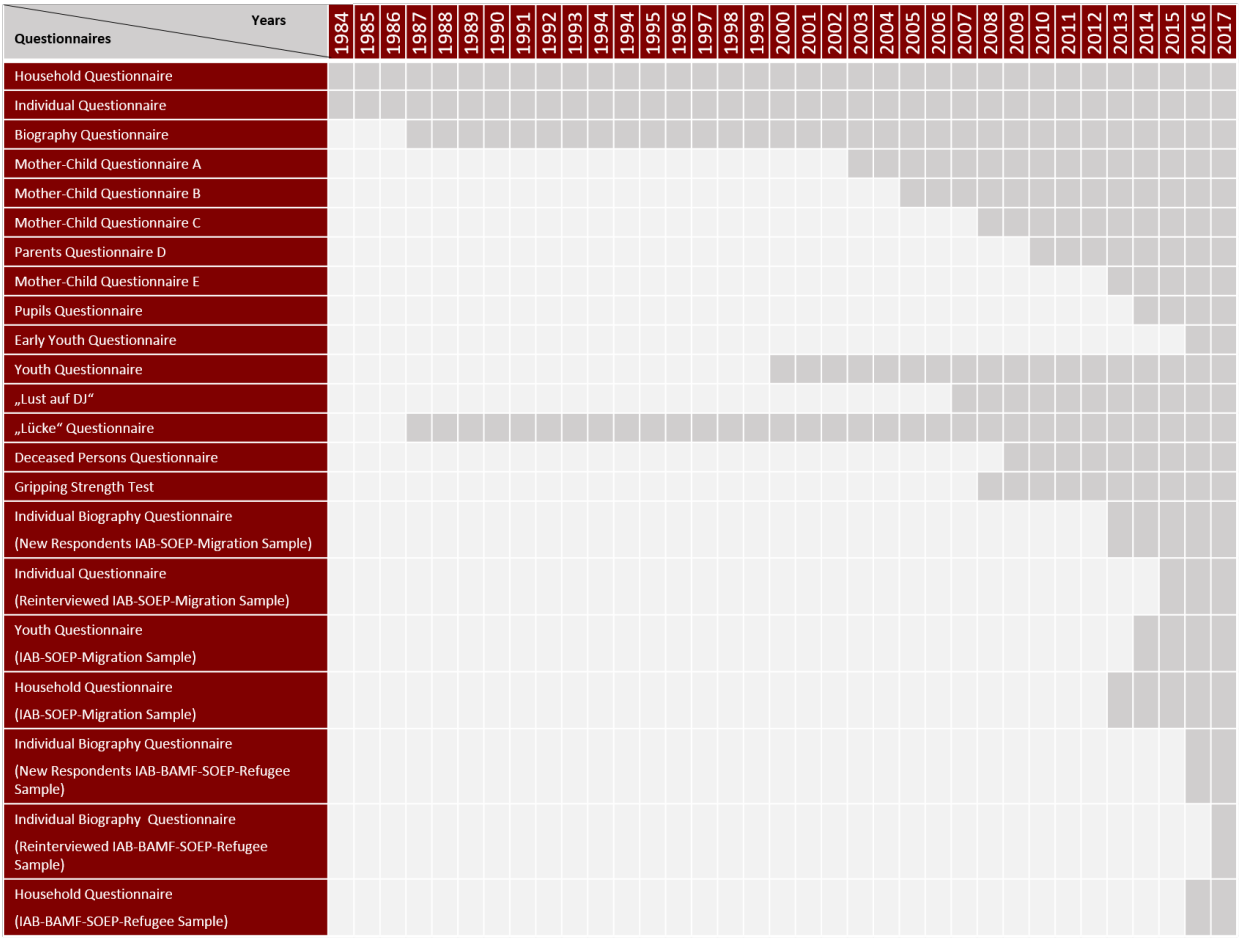
1.1.10 Time Use and Environmental Behavior

Time is a valuable resource for every human being. Information on how a person plans their time, what obligations they have at what time and how they spend their free time can be found in the "Time Use and Environmental Behavior" section. This section also provides comprehensive information on environmental awareness. Which transport infrastructure is used, which energy resources are used to what extent and what is the position on the subject of renewable energies?

1.1.11 Work and Employment

Information about the topic profession can be found in this section. From the very first job and further training, to job changes and parenthood, to part-time jobs and unemployment. However, not only objective information such as hours of work, but also perceptions of the working environment and feelings about work are shown.

1.2 SOEP Questionnaires



1.2.1 Household Questionnaire

Availability: Since 1984

Respondent: Head of household

Stable content

The household questionnaire has been a standard instrument since the beginning of the SOEP. Because the SOEP has a panel character, important questions have to be answered each year anew by the respondents. In order to enable analyses over time, the household questionnaire therefore has a large number of question modules which are asked every year. The following question modules are part of the core program of the household questionnaire.

Topics	Module	No. Vars	Variables
Household, Amenities and Contribution of private Households			
	Change of living situation	3	hlf0523 & hlf0106 & hlf0107

Continued on next page

Table 1 – continued from previous page

Topics	Module	No. Vars	Variables
	Neighbourhood	1	hlf0153
	House type	4	hlf0154 & hlf0016 & hlf0155 & hlf0596
	Size and condition of the house	4	hlf0018 & hlf0019 & hlf0071 & hcf0011
	Apartment equipment	13	hlf0023 - hlf0037 & hlf0529 - hlf0531
	Apartment status	4	hlf0001 & hlf0006 & hlf0007 & hlf0009 & hlf0015
	Loans, mortgages, building-society loans	2	hlf0087 & hlf0088
	Hereditary lease interest	2	hlf0597 & hlf0598
	Modernization costs	2	hlf0599 & hlf0600
	Owner costs	2	hlf0601 - hlf0605 & 'hlf0090 & hlf0084
	Photovoltaic and solar thermal system	6	hlf0532 & hlf0535 - hlf0539
	Owner burden	1	hlf0606
	Social housing/leased flat at a reduced rate	2	hcf0007 & hlf0073
	Apartment owner	1	hlf0013
	Rental and ancillary costs	9	hlf0069 & hlf0074 & hlf0078 & hlf0079 & hlf0081 & hlf0082 & hlf0607 & hlf0608 & hlf0610
	Tenant burden	1	hlf0611
	Cleaning or household assistance	2	hlf0261 & hlf0262
	Persons in need of care	22	hlf0291 & hlf0631 hlf0292 & hlf0300 - hlf0304 & hlf0315 & hlf0317 & hlf0319 - hlf0322 & hlf0331 & hlf0332 & hlf0369 & hlf0370 & hlf0446 - hlf0448 & hlf0595
	Name and birth of children	1	hlk0044
	School attendance for child	2	ks_gen & ks_spe
	Caring Situation for child	6	ks_asc_r & kc_relaz & kc_frdn & kc_paid & kc_mindr & kc_none
Income, Taxes and Social Security			
	Income and expenses from rental/lease	7	hlc0007 - hlc0009 & hlc0111 & hlc0112 & hlc0176 & hlc0177
	Repayments for loans	2	hlc0113 & hlc0114
	Credit burden	1	hlc0115
	Inheritance, present, lottery prize	2	hlc0178 - hlc0183
	Investments	14	hlc0104 - hlc0108 & hlc0093 - hlc0098 & hlc0013 & hlc0014 - hlc0184

Continued on next page

Table 1 – continued from previous page

Topics	Module	No. Vars	Variables
	Income/expanses household	43	hlc0005 & hlc0006 & hlc0039 & hlc0041 - hlc0047 & hlc0049 - hlc0055 - hlc0057 & hlc0059 & hlc0061 - hlc0065 & hlc0067 & hlc0068 & hlc0070 & hlc0071 & hlc0077 - hlc0085 & hlc0090 & hlc0121 - hlc0125
	Saving	3	hlc0172 - hlc0174

`Download Stable Content Household (csv)`

Replication Calendar Household Questionnaire

Besides the topics that are asked every year in the household questionnaire, there are some topics modules that are collected irregularly. Many questions 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 on the household questionnaire are added, which are not surveyed every year and are therefore not part of the standard questions of the household questionnaire. You can find a selection of irregular but recurring topics in the replication calendar:

Topics	Module	Replication	No. Vars	Variables
Home, Amenities and Contributions of private households				
	Participation (financial reasons)	2001, 2003, 2005, 2007, 2011, 2013, 2015	24	hlf0174– hlf0175 & hlf0439– hlf0444 & hlf0178– hlf195
	Material deprivation	2016	24	hlf0178-hlf0181 & hlf0186-hlf0195 & hlf0613 - hlf0622
	Household equipment since last year	1998, 1990, 1992, 1994, 1996, 1998, 2000, 2002, 2004, 2006, 2008, 2010	32	hlf0163-hlf0167, hlf0209, hlf0212, hlf0214-hlf0215, hlf0217-hlf0218, hlf0159, hlf0223, hlf0228-hlf0229, hlf0231, hlc0116-hlc0118, hlf0233, hlf0236-hlf0237, hlf0169-hlf0170, hlf0239-hlf0242, hlf0244-hlf0245, hlf0247-hlf0248
	Amount of Books in household	2001, 2006, 2011, 2016	1	hlf0197
	Pets	1996, 2006, 2011, 2016	7	hlf0254-hlf0259, hlf0196
	Reasons for Moving and Comparison	1985-2013, 2015	27	hlf0109-hlf0132, hlf0524-hlf0526
	Second Residence	2011,2016	3	hlf0156-hlf0158
	Living Environment	1986, 1994, 1999, 2004, 2009, 2014	22	hlf0135-hlf0152, hlj0004, hld0001-hld0003

Continued on next page

Table 2 – continued from previous page

Topics	Module	Replication	No. Vars	Variables
	Lunch, School	1997, 2002, 2005, 2007, 2011, 2013, 2015	1	ks_lunch
	Sponsors and Costs, School	1987, 1995, 1997, 2002, 2005, 2007, 2011, Mig 2013	7	kd_publ, kd_indep, kd_priv, kd_comp, kd_comm, ks_amtp, ks_cost
	Lunch, Childcare	1997, 2002, 2005, 2007, 2011, 2013, 2015	1	kd_lunch
	Sponsors and Costs, Childcare	2011, 2013, 2015	7	kd_publ, kd_indep, kd_priv, kd_comp, kd_comm, kc_amtp, kc_cost
	Rely on care times	2002	1	kd_rely
	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
	Independent Income of The Children	2016		
	Price Comparison Apartment for Rent	1984-2014	1	hlf0094
Health and Care				
	Satisfaction With Availability Of Care	2002	1	hlf0318
Income, Taxes and Social Security				
	Saving	2010,2016	7	hlc0024-hlc0030
	Expenditures on Food	1998, 2000, 2001, 2003, 2005, 2007, 2009, 2011, 2016	2	hlf0435-hlf0436
	Alimony	2010	4	hlc0091-hlc0092, hld0004-hld0005
	Consumption Module	2010	122	hlf0163-hlf0172, hlf0209-hlf0252, hlf0159, hlc0166-hlc0168, hlf0371-hlf0434
	Good/Low Income	1992, 1997, 2007	4	hcc0005-hcc0010
Time Use and Environmental Behaviour				
	Traffic and Energy	1998, 2003, 2015	124	hlf0540-hlf0591, hli0005, hli0077-hld0142

`Download Replication Household (csv)`

1.2.2 Individual Questionnaire

Availability: Since 1984

Respondent: Persons over 18 years in the household

Stable content

The individual questionnaire has been a standard instrument since the beginning of the SOEP. Because the SOEP has a panel character, important questions have to be answered each year anew by the respondents. In order to enable analyses over time, the individual questionnaire therefore has a large number of question modules which are asked every year. The following question modules are part of the core program of the individual questionnaire.

Topics	Module	No. Vars	Variables
Attitudes, Values and Personality			
	Satisfaction with various aspects	11	plh0171 - plh0181
	Mood	4	plh0184- plh0187
	Flourishing	1	plh0334
	Risk Aversion	1	plh0204
	Political orientation	4	plh0007 , plh0011 - plh0013
	Worries	13	plh0032 , plh0033 , plh0035 - plh0038 , plh0040 , plh0042 , plh0043 , plh0046 , plh0047 , plh0335 , plh0336
	Life satisfaction	1	plh0182
Demography and Population			
	Origin	6	plj0014 , plj0022 - plj0025, plj0175
Education and Qualification			
	Apprenticeship	1	plg0012 - plg0015 , plg0264 , plg0265
	Acquired qualification	12	plg0072- plg0079 , plg0284 , plg0268, p_degree, p_field
	Advanced training	3	plg0269 - plg0271
Family and Social Networks			
	Family situation	4	pld0131 - pld0133, plk0001
	Family changes	42	pld0012 - pld0014 , pld0038 - pld0040 , pld0134 - pld0156 , pld0158 - pld0171
Helath and Care			
	State of health	1	ple0008
	Disability or severe disability	2	ple0040- ple0041
	Visits to the doctor	2	ple0072 , ple0073
	Hospital stays	3	ple0053 , ple0055 , ple0056

Continued on next page

Table 3 – continued from previous page

Topics	Module	No. Vars	Variables
	Sickness notifications to employer	10	ple0044 , ple0046 , ple0048 - ple0052 , ple0174 , ple0175 , plb0024
	Health insurance	4	ple0097 , ple0099 , ple0104 , ple0160
Income, Taxes and Social Security			
	Employment earnings and collective wage agreements	6	plc0013 , plc0014 , plc0506 - plc0509
	Additional questions for employees	13	plc0042 - plc0054
	Additional questions for retirees/pensioners	20	plc0223 , plc0236 , plc0238 , plc0240 , plc0242 , plc0243 , plc0245 , plc0247 , plc0249 , plc0251 , plc0278 , plc0279 , plc0281 , plc0283 , plc0285 , plc0286 , plc0288 , plc0290 , plc0516 , plc0517
	Transfer payments	21	plj0131 - plj0151
Time Use and Environmental Behaviour			
	Calendar	12	pab0001- pab0008, pab010- pab0013
	Use of time	60	pli0001- pli0060
Work and Employment / Income, Taxes and Social Security			
	Secondary occupations	9	plb0392 - plb0396 , plb0573 , plc0062 , p_isco88n , p_isco08n
	Income	62	plc0015 - plc0017 , plc0064 , plc0065 , plc0073 - plc0075 , plc0116 , plc0117 , plc0126 , plc0130 - plc0132 , plc0135 - plc0139 , plc0152 - plc0155 , plc0168 - plc0171 , plc0177 , plc0178 , plc0181 - plc0184 , plc0188 - plc0190 , plc0198 , plc0202 - plc0205 , plc0232 - plc0235 , plc0273 - plc0276 , plc0488 - plc0490 , plc0494 - plc0496 , plc0513 , plc0514, plc0515 , plb0471 , plb0474, plb0477
Work and Employment			
	Work, last 7 days	1	plb0018
	Maternity/ Parental leave	1	plb0019 , plb0020
	Care period (Pflegezeit)	1	plb0020
	Registered unemployed	1	plb0021
	Quitting a profession	4	plb0282 , plb0298- plb0305 , plc0040, plc0041

Continued on next page

Table 3 – continued from previous page

Topics	Module	No. Vars	Variables
	Employment status	1	plb0022
	Start of the job	9	plb0417 - plb0424 , plb0240
	Change of job	9	plb0031 - plb0034 , plb0478- plb0480 , plb0284 , plb0295
	Job search	2	plb0362 , plb0358
	Practised profession	4	plb0072 , plb0073 , p_nace , p_isco08
	Current employment	13	plb0035 - plb0037 , plb0040 , plb0041 , plb0049 , plb0058 , plb0063- plb0065 , plb0568 , plb0570 , plb0586
	Working hours	10	plb0180 - plb0182, plb0185 - plb0188 , plb0241 , plb0209 , plb0210
	Overtime	10	plb0193 - plb0198 , plb0483 , plb0484 , plb0220 , plb0605

Download Stable Content Individual (csv)

Replication Calendar Individual Questionnaire

Besides the topics that are asked every year in the individual questionnaire, there are some topics modules that are collected irregularly. Many questions 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 on the individual questionnaire are added, which are not surveyed every year and are therefore not part of the standard questions of the individual questionnaire. You can find a selection of irregular but recurring topics in the replication calendar:

Topics	Module	Replication	No. Vars	Variables
Attitudes, Values and Personality				
	Well-being	1990 (only Ost), 1994, 1999	13	plh0091 - plh0103
	Optimism	1994, 2005, 2009, 2014	1	plh0244
	Religious Affiliation	1990, 1997, 2003, 2007, 2011, 2015	1	plh0258
	Organisational and community membership	1985, 1989, 1993, 1998, 2001, 2003, 2007, 2011, 2015	5	plh0263 -plh0267
	Personality traits (Big Five)	2005, 2009, 2013, 2017	16	plh0212 -plh0226, plh0255
	Anomy	1992,1993, 1995, 1996, 1997, 2008, 2013	4	plh0188-plh0191
	Depressive Traits	2016	4	plh0339 – plh0342
	Goals in life (Kluckhohn)	1990, 1992, 1995, 2004, 2008, 2012, 2016	9	plh0105 – plh0112, plh0343

Continued on next page

Table 4 – continued from previous page

Topics	Module	Replication	No. Vars	Variables
	Money and account balance	2016	3	plh0344 – plh0346
	Control beliefs	2005, 2010, 2015	10	plh0247 – plh0252 , plh0245 , plh0246
	Reciprocity	2005, 2010, 2015	11	plh0206 - plh0211, plh0142- plh0146
	Trust and fairness	2003, 2008, 2013	8	plh0192 - plh0196 , pld0043 - pld0045
	Narcissism	2018		
	Loneliness	2013,2017	3	plh0269 - plh0271
	Impulsivity, patience	2008,2013	3	plh0204 , plh0253 , plh0254
	Risk Aversion (long)	2004,2009	6	plh0197 - plh0202
	Lottery question	2004,2009	1	plh0203
	Policy objectives (Inglehart Index)	1984, 1985, 1986, 1996, 2006, 2016	4	plh0054, plh0056, plh0058, plh0061
	Attitudes towards refugees	2016	11	plj0433 – plj0443
	Bundestag election	2014	1	plh0333
	Political Tendency, Left-Right	2005, 2009, 2014	1	plh0004
	Social responsibility	1987, 1992, 1997, 2002, 2017	11	plh0016 – plh0026
	Donations	2010,2015	2	plh0129 , plh0130
	Donation of blood	2010,2015	3	plh0131 - plh0133
	Donations of goods	2010	8	plj0108 - plj0115
	10000 Euro Question	2010,2017	3	plh0134- plh0136
	Income justice, general	2005	12	plh0116- plh0127
Family and Social Networks				
	Family Network	1991, 1996, 2001, 2006, 2011, 2016	43	pld0020 - pld0036 & pld0107 - pld0118 & plj0117 - plj0130
	Networks, trusted person	1991, 1996, 2001, 2006, 2011, 2016	29	pld0089 - pld0088 & plf0049 - plf0050
	Networks, sociodemography	2006, 2011, 2016	18	pld0089 - pld0106
	2003, 2008, 2011, 2013, 2015, 2017, 2018		1	pld0047
	LGBT Status	2016	1	pld0298
	Gender Attitudes	2018	8	
Health and Care				
	Ilness	2009, 2011, 2013, 2015, 2017	14	ple0011 – ple0024
	Stress and exhaustion (SF-12)	2002 - 2018 (every two years)	10	ple0026 – ple0036

Continued on next page

Table 4 – continued from previous page

Topics	Module	Replication	No. Vars	Variables
	Disabilities in everyday life (SF-12)	1997-2002, 2004-2018 (every two years)	2	ple0004 – ple0005
	Height and Weight	2002 - 2018 (every two years)	2	ple0006 – ple0007
	Chronicall Illness	1984-1989, 1991, 2009, 2010, 2012, 2014, 2016, 2018	1	ple0036
	Health restrictions	2011, 2012, 2013, 2015, 2017	2	ple0009 & ple0162
	Health insurance debts	2017	1	
	Smoking	1998, 1999, 2001, 2002-2018 (every two years)	6	ple0081 & ple0086 - ple0088 & ple0176
	Alcoholic beverages	2006, 2008, 2010, 2016	2	ple0177 – ple0178
	Nutritional awareness	2004-2016 (every two years)	4	ple0179 – ple0182
	Assisted or curative care	1999-2011	1	ple0121
	Additional private insurance	2011-2014, 2016, 2018	8	ple0127 – ple0134
	Private supplementary care insurance	2016,2018	3	ple0183 – ple0185
	Insurance status	2018		
	Type of disability	2001, 2002, 2004, 2006, 2008, 2010, 2015	2	ple0140 – ple0141
	Individual health service	2016,2018	1	ple0186
Income,Taxes and Social Security				
	Labour income, hourly wage	2017		
	Earnings Work October 2014	2015	2	plb0584 , plb0585
	Balance sheet of assets	1988, 2002, 2007, 2012, 2017	67	plc0315 - plc0319 , plc0328 - plc0374 , plc0411 - plc0425
	Inheritance	2001,2017	33	plc0375 - plc0407
	Transfer payments, income	2009, 2010, 2011	21	plj0152- plj0172
	Social security	1987, 1992, 1997, 2007, 2012, 2017	7	plc0008, plc0009, plc0111-plc0115
	Entitlements, statutory	2013	4	plc0432- plc0435

Continued on next page

Table 4 – continued from previous page

Topics	Module	Replication	No. Vars	Variables
	Riester	2004, 2006, 2007, 2010, 2012, 2013, 2015, 2017	3	plc0430 , plc0431 , plc0313
	Riester payments	2013	3	plc0437 - plc0439
	Entitlements, company	2013	5	plc0441 - plc0445
Immigration, Migration, Transnationalization				
	Integration Indicators	1997, 1999, 2001, 2003, 2010, 2012, 2014, 2016, 2018	2	plj0078 & plj0080
	Native tongue	2007-2011, 2013, 2015, 2017	8	plj0009 & plj0691 - plj0693 & plj0071 - plj0073 & plj0077
	Contact, at home and abroad	1997-2017 (every two yeras)	4	plj0060 - plj0063
	Citizenship Application	1998-2018 (every two years)	1	plj0021
	Residence status, citizenship	2018	2	
	Intention to stay	1997-2011, 2013, 2015, 2017	4	plj0085 - plj0088
	Disadvantages due to origin (short)	1997-2011, 2013, 2017	1	plj0048
	Disadvantages due to origin (area)	2015	15	plj0048 & plj0327 - plj0340
	Linguistic usage, newspaper	1996-2012 (every two years)	1	plj0070
	Linguistic usage, media	2014, 2016, 2018	1	plj0226
	Visit country of origin last 2 years	1996-2018 (every two years)	2	plj0322 & plj0323
	Sense of home	1996-2014 (every two years)	2	plj0083 & plj0340
	Regional attachment	2009,2014	3	plj0343 - plj0345
	Contacts and thoughts abroad	2009,2014	6	plj0104 & plj0105 & plj0089 - plj0092
	Circle of Friends, Share of Migrants	2013,2018	1	plj0191
	Foreign language skills	2013	1	plm0135 & plj0187
Time Use and Environmental Behaviour				

Continued on next page

Table 4 – continued from previous page

Topics	Module	Replication	No. Vars	Variables
	Leisure activities (short)	1984-1986, 1988, 1992, 1994,1996, 1997, 1999, 2001, 2005, 2007, 2009, 2011, 2015, 2017	9	pli0090 – pli0098
	Leisure activities (long)	1990, 1995, 1998, 2003, 2008, 2013	19	pli0079 – pli0092 & pli0096 – pli0098 & pli0165 & pli0168
	Computer usage	1997,1999, 2000, 2001	8	pli0066 – pli0073
	Traffic behavior	1993,1998, 2003	77	pli0101 – pli0160 & plb0016 & plb0145 & plb0147 – pli0156 & plb0158 & plb0159 & plb0175 & plb0591
Work and Employment				
	Illegal Employment	2015,2016	2	plb0571 , plb0572
	Intensity of work	2015, 2016, 2017	2	plb0593, plb0594
	Work equipment	2015, 2016, 2017	6	plb0595- plb0600
	Work breaks	2015, 2016, 2017	3	plb0601- plb0603
	Employment October 2014	2015	1	plb0574
	Work breaks October 2014	2015	4	plb0575 - plb0578
	Working time October 2014	2015	3	plb0579 - plb0581
	Overtime October 2014	2015	2	plb0582 , plb0583
	Lifelong learning	2014	1	plg0266
	Continuing education, initiative	1989, 1993, 2000, 2004, 2008, 2014	2	plg0273 , plg0274
	Further training, financing	1989, 1993, 2000, 2004, 2008, 2014, 2015, 2017	7	plg0285 - plg0291
	Further education, organizer		1	
	Continuing education, reasons for failure	1989, 1993, 2000, 2004, 2014	5	plg0277 – plg0281
	Further education, course details and motives	1989, 1993, 2000, 2004, 2008	60	plg0108 - plg0122 , plg0129 - plg0149 , plg0152 , plg0154 , plg0164 , plg0165 , plg0169 , plg0171 , plg0172 , plg0174 - plg0177 , plg0182 - plg0186

Continued on next page

Table 4 – continued from previous page

Topics	Module	Replication	No. Vars	Variables
	Benefits from employer, additional benefits	2008, 2010, 2012, 2014-2017	14	plc0026 - plc0039
	Benefits from employer, company car	2016,2017	1	plc0532
	Work time regulation	2003, 2005, 2007, 2009, 2011, 2014-2017	1	plb0211
	Standby duty	2011, 2014-2017	4	plb0212- plb0215
	Work time recording			
	Overtime, compensation	1984-2014	1	plb0195
	Start of working hours	2002, 2004, 2006, 2008, 2012, 2015, 2017	3	plb0180- plb0182
	Evening - Weekend work	2005, 2007, 2009, 2011, 2013, 2015, 2017	4	plb0216- plb0219
	Contract to Provide Specific Services (Werkvertrag)	2013,2015	1	plb0482
	Wage Tax Classification	1991, 1993, 2004, 2016	1	plc0091
	Exercised profession, training	1984-2014, 2016	4	plb0076- plb0079
	Commuter Module	1991-2013, 2015, 2017	6	plb0589- plb0592, plb0158, plb0159
	Vacation claim	2000, 2005, 2010	8	plb0269- plb0276
	Work from home	1997, 1999, 2002, 2009, 2014	3	plb0095- plb0097
	Short-time allowance (Kurzarbeitergeld)	1984-2001, 2003-2005, 2010, 2011,	2	plc0057 , plc0058
	Performance evaluation in the company	2004, 2008, 2011, 2016	5	plb0098- plb0102
	Leading position	2007, 2009, 2011, 2013, 2015, 2017	3	plb0067- plb0069
	Wage justice	2005, 2007, 2009, 2011, 2013, 2015, 2017	6	plh0138- plh0141 , plh0337 , plh0338
	Workload (Effort-Reward-Inbalance)	2001, 2006, 2011, 2016	26	plb0112- plb0137
	Professional expectations, long	1985, 1987, 1989, 1991, 1993, 1994, 1996, 1998, 2000, 2005, 2009, 2013	11	plb0432- plb0442
	Professional expectations, short	2015	5	plb0433, plb0437, plb0440, plb0588

Continued on next page

Table 4 – continued from previous page

Topics	Module	Replication	No. Vars	Variables
	Employee organization (Betriebsrat)	2001,2006, 2011, 2016	1	plb0050
	Occupational qualification, use	1985-2007, 2009	1	plb0357
	Self-employment, reasons	2010,2015	6	plb0333- plb0338
	Postcode of the place of work	2016		
	Occupational expectations, non-workers	1999, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015	3	plb0427- plb0429
	Job search, preferences	1994, 1996, 1997, 1998, 1999, 2000, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2015, 2017	1	plb0426
	Job search, motives	1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2003, 2005, 2007, 2009, 2011, 2013, 2017	1	plb0111

`Download Replication Individual (csv)`

1.2.3 Biography Questionnaire

Availability: Since 1987

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

Content:

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

1.2.4 Mother-Child Instruments

Topic	Mother-Child A (Age 0-1)	Mother-Child B (Age 2-3)	Mother-Child C (Age 5-6)	Parents D (Age 7-8)	Mother-Child E (Age 9-10)
Pregnancy & Birth	x				
Nursing	x	x			
Health	x	x	x		x
Height & Weight		x	x		
Vineland Adaptive Behavior Scale		x			
Strength and Difficulties Questionnaire			x		x
Childcare	x	x	x	x	x
Linguistic Usage		x			x
Temper	x	x			
Big 5 Personality Traits		x	x		x
School & Homework				24.	x
Educational Aspirations				x	x
Parenting Goals				x	
Styles of Parenting				x	
Mother Role/ Parent Role	x			x	
Leisure and Activities (with Child)		x	x		x
Friends					x
Pocket Money					x

Mother-Child Questionnaire A (Age 0-1)

Mothers of newborn children primarily answer questions about the course of pregnancy, birth, breastfeeding and the health of the newborn child. It also asks to what extent the mother feels that her life circumstances have changed after the birth of the child, how the care of the child is regulated and how the temperament of the baby (as a precursor of the personality) is perceived by mothers.

Availability: Since 2003

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

Content:

- Course of pregnancy
- Childbirth

- Health screening
- Well-being
- Childcare
- Life circumstances

Mother-Child Questionnaire B (Age 2-3)

Mothers of 2-3-year-old children also answer some questions about their child's health and how long they have been breastfeeding. In addition, the child's care situation is asked, again the temperament as well as a short scale for recording the personality (agreeableness, extraversion, openness and conscientiousness of the Big Five; McCrae and Costa 1987). In addition, the use of language in the family and activities carried out with the children (e.g. going to the playground, reading or telling stories, visiting other families with children) are recorded. Mothers also assess their children's adaptive behaviour in the dimensions of communication, everyday skills, social relationships and motor skills. The acquisition is based on a translated version of the Vineland Adpative Behavior Scale, which was reduced to 20 items for the SOEP. This scale thus investigates the stage of development of the infant in everyday life.

Availability: Since 2005

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

Content:

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

Mother-Child Questionnaire C (Age 5-6)

The subsequent age-specific survey is carried out as soon as the children turn six years old in the survey year. Among the topics it resembles the surveys conducted in previous years: health, care situation, a more comprehensive battery of items on the personality (from this age neuroticisum is also collected) and activities that are carried out with the child. In addition, there is the Strength and Difficulties Questionnaire (SDQ), which is a shortened version of the German version of the SDQ to 17 items and is a very frequently used instrument for the mental health of children and young people.

Availability: Since 2008

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

Content:

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

Parents Questionnaire D (Age 7-8)

The questionnaire, which was developed for 7-8-year-old children, is the only age-specific instrument to be completed by both parents, as long as they live together in the same household. In this age range, questions about school attendance (time of school enrolment) and idealistic and realistic educational aspirations become relevant for the first time. However, the focus of this instrument is on the educational goals, parenting styles and the role of both parents. The educational objectives can be differentiated between conformity and autonomy. Educational styles are asked by answering 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 for recording the role of parents. The parental role can be divided into three scales (autonomy, hostile attributes, willingness to make sacrifices).

Availability: Since 2012

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

Content:

- Expectations for school achievments
- Expectations of parental educational goals
- Upbringing
- Parental role
- Childcare

Mother-Child Questionnaire E (Age 9-10)

In addition to the items on health and the care situation recorded in almost all age groups, 9-10-year-old children are asked for more detailed information on the school situation. Here, too, the idealistic and realistic educational aspirations of the mothers for their child are recorded, but also the last grades of the three main subjects, as well as the child's homework supervision and school motivation. Since friends and leisure activities are gaining in importance in this age group, questions are also asked on these topics. Whether and how much pocket money the child receives will be asked for the first time in this age group.

Availability: Since 2012

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

Content:

- Expectations (school achievments, parental educational goals)
- Education
- parental commitment
- Leisure activities for children
- Family environment
- Social behavior child
- Personality Child
- Health Child
- Supervision
- Pocket money

1.2.5 Youth Instruments

Topic	Pupils Questionnaire (Age 11-12)	Early Youth Questionnaire (Age 13-14)	Youth Questionnaire (Age16-17)
Child's State of Health		x	x
Height & Weight	x	x	x
Life Satisfaction	x	x	x
Strength and Difficulties Questionnaire	x	x	
Linguistic Usage	x	x	
Big 5 Personality Traits	x	x	x
Willingness to take Risks	x	x	x
Locus of Control		x	x
Trust			x
Time Preference			x
School Attendance & Homework	x	x	x
Parents' Interest in School Performance	x		x
Educational Aspirations	x	x	x
Cultural Capital	x		
Relationship between Family Members	x	x	x
Parentingl Behaviour			x
Importance of personal environment		x	x
Hobbies	x	x	x
Friends	x	x	x
Pocket Money	x	x	x
Saving		x	x
Political Interest		x	x
Housing Situation			x
Jobs and Money			x
Education and Career Plans			x
Future			x
Childhood and Parental Home			x
Attitudes and Opinions			x

Pupils Questionnaire

In the year in which the children turn twelve, they answer questions about their situation for the first time. Here the focus is once again on the school situation: the start and end of school are asked differentiated according to the days of the week, the type of school attended, the number of pupils in the class and how many of them do not come from Germany, whether one feels discriminated against by the teacher and the last grades in math, German and English. It also determines how much time the student spends on homework, where he or she does the homework and who helps him or her with the homework and learning. The children are asked about their idealistic and realistic graduation aspiration. Since friends play an important role as caregivers at this age, they and various family members are asked what role they play in the support and how often there are disputes. Also asked about the number of close friendships and how often the parents interfere in the choice of friends. The educational aspirations of the three best friends and

a maximum of three older siblings (if any) are asked. The cultural capital and learning environment of the pupils are assessed on the basis of various questions (e.g. availability of literature, instruments, art at home; a desk and a room for oneself). Furthermore, the type and frequency of leisure activities is again asked. The student answers whether and how much pocket money he or she receives and for the first time gives information about his or her own personality, willingness to take risks and life satisfaction. The use of the language in the family (only German or other languages) and with whom the meals are usually taken is also asked.

Availability: Since 2014

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

Content:

- Attitude
- Personality
- School (timetable, school-leaving qualification, Engagement)
- Recreational activities
- Social and family surroundings
- Life circumstances

Early Youth Questionnaire

The questionnaire for early youth is largely similar to the questionnaire for pupils in order to provide an appropriate data structure for questions relevant to developmental psychology. Fewer questions are asked about homework and the learning environment, but the question is asked whether the young person is involved in the school (e.g. as class spokesperson or in a working group) and social capital is acquired in this way. The current importance of various family members and friends is asked and, in addition to their own educational aspirations, also that of the three best friends. With regard to parents, the question is asked how long the young person is allowed to travel and stay up alone before school days and what things the 14-year-old has already done without parents (e.g. holidays, going to the doctor, exchanging something in the shop, drinking alcohol, smoking cigarettes). They ask again for the pocket money and also whether the young person has the opportunity to save money. Another new topic in this age group is the interest in politics and the inclination towards a certain party.

Availability: Since 2015

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

Content:

- self-perception
- School (timetable, school-leaving qualification, Engagement)
- Recreational activities
- Friends
- Siblings
- Parents
- Pocket money
- Party preferences
- Self-Perception
- Willingness to take risks
- Life satisfaction

- Attitudes/Opinions
- Future

Youth Questionnaire

In the SOEP, people who turn 17 in the corresponding survey year are considered adult respondents. Like other first-time adult participants, you will thus receive a CV and a individual questionnaire. Since part of the adult biography (such as the employment biography or the relationship biography) does not yet apply to the young participants and other aspects such as the relationship with parents, leisure activities, the school situation or vocational training play a greater role, a youth questionnaire was developed in 2000 which replaces the CV questionnaire in this age group and has been used since then. The content of this questionnaire corresponds in many respects to the adult CV questionnaire, so that the data can be used to supplement the information on parents (if they do not live in the household; data set: BIOPAREN). Health status, personality, willingness to take risks, locus of control, trust, time preference, political preferences, knowledge of German as well as information on the living situation, work situation, training, career plans and educational aspirations are also surveyed. For the period from 2000 to 2005, the youth questionnaire was surveyed in addition to the personal questionnaire. Since 2006, only the youth questionnaire has been recorded for 17-year-olds. Since then, it has been available in a version extended by a few indicators, and instead a test has been used to assess cognitive potential. Based on the I-S-T 2000R (Amthauer et al. 2001) the components analogies, number series and matrices with 20 subtasks each were selected for the SOEP (cf. Solga et al. 2005). With the help of these tasks, the fluid cognitive abilities are to be recorded. This is 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 the usual questionnaires in surveys, the willingness of young people to participate is high (Schupp and Hermann 2009).

Availability: Since 2000

Respondent: 16-17 year olds in the household

Content:

- Living
- Relationships
- Leisure and Sport
- School (Graduation, Foreign languages, Engagement)
- Pocket money
- Education
- Career Plans
- Future
- Origin
- Childhood and Parental Home
- Attitudes/Opinions
- Self-Perception
- Life satisfaction
- Party preferences

„Lust auf DJ“ (Denksport und Jugend) Questionnaire

In SOEP 2006, a separate questionnaire with cognitive tests for adolescents was used for the first time: “Lust auf DJ”. In this case, “DJ” stands for “Thinking Sports and Youth (Denksport und Jugend)”, but was also specifically selected to arouse the more common association of “Disc Jockey”. For all interviewees aged 16 - 17 years, the questionnaire “Lust auf DJ” was used and created.

Availability: Since 2007

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

Content:

- Assignment of word pairs
- Complete incomplete equations
- Assign figures

1.2.6 Additional Instruments

„Lücke“ Questionnaire - Re-questioning of the Individual Questionnaire (Summary)

The “Lücke” (english:gap) questionnaire relates to temporary drop outs for which significant missing data from the previous year are collected.

Availability: Since 1987

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 Persons Questionnaire

For the first time in the main wave of 2009, information should be collected on former SOEP participants who have died since the survey in 2008 or until the time of the survey in 2009. Through the questionnaire “The deceased person”, the SOEP curriculum vitae principle is thus consistently “completed”. The primary aim of the chosen concept is to obtain as much information as possible about the death circumstances of former SOEP participants. However, it also generates information about people who have never participated in the SOEP survey. The information collected in this way about otherwise “unknown” persons, however, can also be used for various analysis purposes on causes of death and the context of death can also be used in the socio-scientific analysis.

Availability: Since 2009

Respondent: SOEP respondents who lost a loved one.

Content:

- Relationship to the deceased
- Deceased part of the survey?
- Domestic environment of the deceased person
- Cause and place of death
- Legacies
- Health condition of the deceased
- Life satisfaction of the deceased
- Influence of loss on one's own life

Gripping Strength Test

Availability: Since 2008

Respondent: Persons over 17 years in the household

Content:

This test measures the strength a person can exert when gripping. This can be important for assessing the physical condition.

1.2.7 IAB-SOEP-Migrationsstichprobe

Personal Biography Questionnaire (New Respondents)

Availability: Since 2014

Respondent: Persons with a migrant background aged 18 and over in the household

Content:

- Citizenship
- Origin
- Knowledge/Skills before entering
- Migration background
- Migration biography (the way to Germany)
- Current living situation
- Childhood and Parental Home
- Life course since the age of 15
- Education/Degrees
- Family Situation
- Partnership situation before immigration
- Employment (current and past)
- Occupational Change
- Current income
- Education/further training

- Earnings
- Well-being
- Attitudes/Opinions

Individual Questionnaire (Reinterviewed)

Availability: Since 2014

Respondent: Persons with a migrant background aged 18 and over in the household

Content:

Like *Individual Questionnaire* + the following migration-specific topics:

- Training/further training at home and abroad
- Discrimination/Pursuit/War
- Employment before moving to Germany
- Amount of income in local currency
- Religious community
- Immigration parents and/or grandparents + place

Youth Questionnaire

Availability: Since 2014

Respondent: 16-17 year olds in household with a migration background

Content:

Like *Biography Questionnaire* + the following migration-specific topics:

- Circle of friends
- Degree at home or abroad
- German lessons as a foreign language
- Training at home or abroad
- Year of the parents' immigration
- Acquired degree of parents at home or abroad
- Religious community

Household Questionnaire

Availability: Since 2014

Respondent: Head of household

Content:

Like *Household Questionnaire* + the following migration-specific topics:

- Valuables in Germany or abroad

1.2.8 IAB-BAMF-SOEP-Befragung von Geflüchteten

Personal Biography Questionnaire (New Respondents)

In 2017 (second wave of surveys), the survey instruments used and the respective survey content will come closer to the SOEP standard. In addition to the personal and household questionnaire, age-specific children's instruments are used. As a rule, mothers of children of specific birth cohorts living in households (2016/2017, 2014, 2011, 2009, 2007) are asked about their educational participation in Germany, as well as information which includes the SOEP standard of the mother-child survey instruments and refugee specific additions, such as educational pathways before fleeing to Germany, language acquisition and mental illness. In addition, with the consent of the parents, specific birth cohorts (2005, 2003 and 2000) of the growing up children/young people in the participating fugitive households themselves are interviewed. The age-specific SOEP standard instruments (students, early youth and youth) also serve as a model here. In addition to specific extensions for the fugitives, the adolescents undergo a test of basic cognitive skills developed by the Institute for Quality Development in Education (IQB). The selection of questions in the personal questionnaire, which is aimed at all adult refugees, should make it possible to trace the course of integration in many areas, also in comparison to other population groups. Thus, topics specific to refugees in the first wave of the survey are updated, such as the current status of the asylum procedure or language course participation. Classic SOEP topics, such as questions about current employment, will be given more scope. Flight-specific innovations that have become established in recent immigrant samples (IAB-SOEP Migration Surveys M1 and M2 from 2013 and 2015), such as the recording of educational qualifications acquired abroad with the help of the CAMCES tool and the question module on the recognition of qualifications acquired abroad, will also be used in the personal questionnaire in 2017.

Availability: Since 2017

Respondent: Refugees over 18 years of age

Content:

- Origin and route to Germany
- Escape/Travel Reasons
- Flight/Travel expenses
- Escape/route
- Accommodation in Germany
- Reasons why Germany as a target country
- Status of the asylum procedure
- Residence permit
- Satisfaction on various aspects
- Intention to stay
- Writing and language skills (mother tongue and foreign language)
- Integration courses in Germany
- Awareness, need and use of support and consulting services
- Employment and income abroad and in Germany
- Life situation before arrival
- Schools, colleges and vocational training abroad and in Germany (+ recognition)
- Curriculum vitae from age 15
- Well-being
- Perception of life

- Attitude towards the future
- Religious community
- party preferences
- Assessment of the current situation in the country of origin
- Attitude and values
- Personality
- Social Networks
- Family situation
- Participation of children in education
- Declaration of consent for register linking

CHAPTER

TWO

TARGET POPULATION AND SAMPLES

The target population covered in the SOEP is defined as the residential population living in private households within the current boundaries of the Federal Republic of Germany (FRG). Because of changes in these boundaries (in 1990) and changes in the residential population due to migration, various adaptations have been applied to the initial sampling structure to keep the sample's representativity. In addition, certain groups have been oversampled to increase the statistical power. In 1984, the survey started with a sample covering the entire population in then West Germany (FRG), where the five biggest groups of foreigners (the so-called "guestworkers") were oversampled.

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

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

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

The different samples in the SOEP are identified by letters: sample "A" refers to the German sample drawn in 1984, "C" to the East Germans from 1990, and so on. Even though these samples are kept separate, the respondents received identical questionnaires for the most part and distinctions by sample are usually not be necessary in an analysis. However, one of the ideas of SOEP is, that the users have full information available about survey methodological issues and survey design. Which means in this case that you can of course identify the corresponding sample for each observation. In the following section, we present details on each of the samples, which - unless stated otherwise - are multi-stage random samples with regional clusters. The respondent's households are selected by random-walk routines.

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

2.1 The SOEP Samples in Detail

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

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

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

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

Sample E "Refreshment" was added in 1998, selected from the entire population of private households in Germany. The households were chosen independently from the ongoing panel and its subsamples A through D, with the targets of increasing the number of observations of the general population and preserving its representativity. The selection scheme used for sample E essentially resembles the one used in subsample A. The number of households in the first wave of subsample E was $1,060$, with a sampling probability of about 0.00005. With the data distribution of 2012, parts of subsample E have been extracted into the SOEP Innovation Sample. It is also the first sample in which the Computer Assisted Personal Interview (CAPI) was implemented. Interviews in Samples A-D at this time were completely conducted using Paperand-Pencil-lnterviews (PAPI). To study mode effects, households of sample E were randomly allocated to CAPI and PAPI mode.

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

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

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

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

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

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

Sample L1 "Cohort Sample" covers private households in Germany, in which at least one household member is a child that was born between January 2007 and March 2010. Again migrants identified by an "onomastic procedure" are oversampled. Sample L1 (as well as L2 and L3) was part of the SOEP-related study "Familien in Deutschland" (FiD), which was later integrated into the SOEP in 2014. As part of an evaluation project of 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 of waves BA to BD differ in some parts from those of the other samples.

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

Sample M1 "Migration Sample" In 2013 a new migration sample was added with around 2,700 households drawn by using register information of the German Federal Employment Agency.

Sample M2 "Migration Sample" in 2015 another migration sample was added with around 1,100 households drawn by using register information of the German Federal Employment Agency.

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

Sample M4 "Refugee Family Sample" The 2016 "IAB-BAMF-SOEP Refugee Survey" (Samples M3 and M4) is a joint project of the Institute for Employment Research (IAB), the Research Centre of the Federal Office for Migration and Refugees (BAMF-FZ) as well as 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 applied for asylum 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 from the research budget of the Federal Employment Agency (BA) allocated to the IAB. 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 acronym for the third top-up sample of refugee households. The population of M5 covers adult refugees who have applied for asylum in Germany since January 1, 2013, 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 "Refreshment Sample (PIAAC-L)" Sample N integrated 2,314 households of former participants of the Programme for the International Assessment of Adult Competencies (PIAAC and PIAAC-L) in 2017. This is the most recent addition to the SOEPCore 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

Erhebungsjahr	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
Stichproben / Jahr	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 (Deutsche)	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 (Ausländer)	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 (Deutsche Ost)							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 (Zuwanderer '94)										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 (Zuwanderer '95)											1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20	21	22	
E (Querschnitt '98)															1	2	3	4	5	6	7	8	9	10	11	12	13	14	15*	16*	17	18	19	20
F (Querschnitt '00)																1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	
G (Obere Einkommen)																		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	
H (Aufstockung '06)																						1	2	3	4	5	6	7	8	9	10	11	12	
I (Incentivierung)																									1	2	-	-	-	-	-	-		
L1 ('10)/																										1	2	3	4	5	6	7	8	
L2 ('10) /																										1	2	3	4	5	6	7	8	
L3('11)																										1	2	3	4	5	6	7	8	
J (Aufstockung '11)																											1	2	3	4	5	6	7	
K (Aufstockung '12)																												1	2	3	4	5	6	
M1 (Migranten)																													1	2	3	4	5	
M2 (Migranten)																																1	2	3
M3 (Flüchtlinge)																																	1	2
M4 (Geflüchtete Familien)																																		1

	Year	Households	Persons	Respondents	Partial Unit Non-	Children
A	1984	4528	11422	9076	0,6	2290
B	1984	1393	4830	3169	0,7	1638
C	1990	2179	6131	4453	1,9	1591
D1	1994	236	733	471	2,9	248
D1/D2	1995	541	1668	1078	6,1	517
E	1998	1057	2446	1910	3,5	466
F	2000	6043	14510	10880	5,5	2991
G	2002	1224	3538	2671	6,1	693
H	2006	1506	3407	2616	6	623
I	2009	1495	3428	2432	13,4	620
L1	2010	2074	7939	3770	6,7	3900
L2	2010	2500	9063	4227	5,1	4611
L3	2011	924	3645	1487	4,2	2092
J	2011	3136	6873	5161	9,9	1147
K	2012	1526	3286	2473	9,2	563
M1	2013	2723	8522	4964	17,8	2481
M2	2015	1096	3048	1711	19,3	927
M3	2016	1775	4823	2351	22,0	1808
M4	2016	1779	7297	2465	27,1	3915
Total		31384	105876	66894		29391

2.2 Eligibility and Follow-up

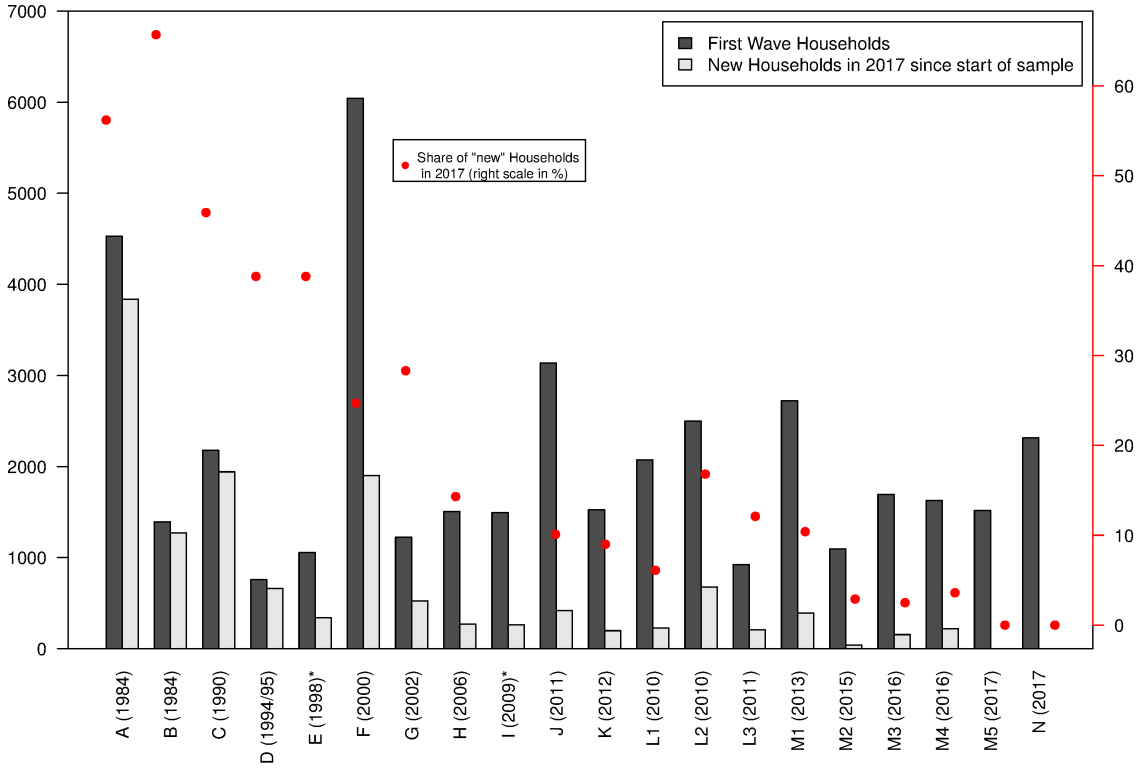
As mentioned, the SOEP's goal is to be representative of the residential population of Germany. All household members 16 and older are eligible for a personal interview, starting with the youth questionnaire at that age, followed by "regular" person questionnaires thereafter. As years go by, the children of the first wave reach age-eligibility and become panel members. If they move out and form their own families, they and their new families are still part of the survey. "New" persons become part of the SOEP population due to birth or residential mobility. In case a person enters a SOEP household after the initial wave, this person is asked to fill out the regular person questionnaire if age-eligible, or will be asked to participate once old enough. Thus in the absence of panel attrition the SOEP would be a self-sustaining survey.

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

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

	Existing Households	New Households
Existing Persons	classic case: without change of adress entire household moves	Move-out
New Persons	Birth Move-In	Move-In or birth into move-out household

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



Download R Code to create figure

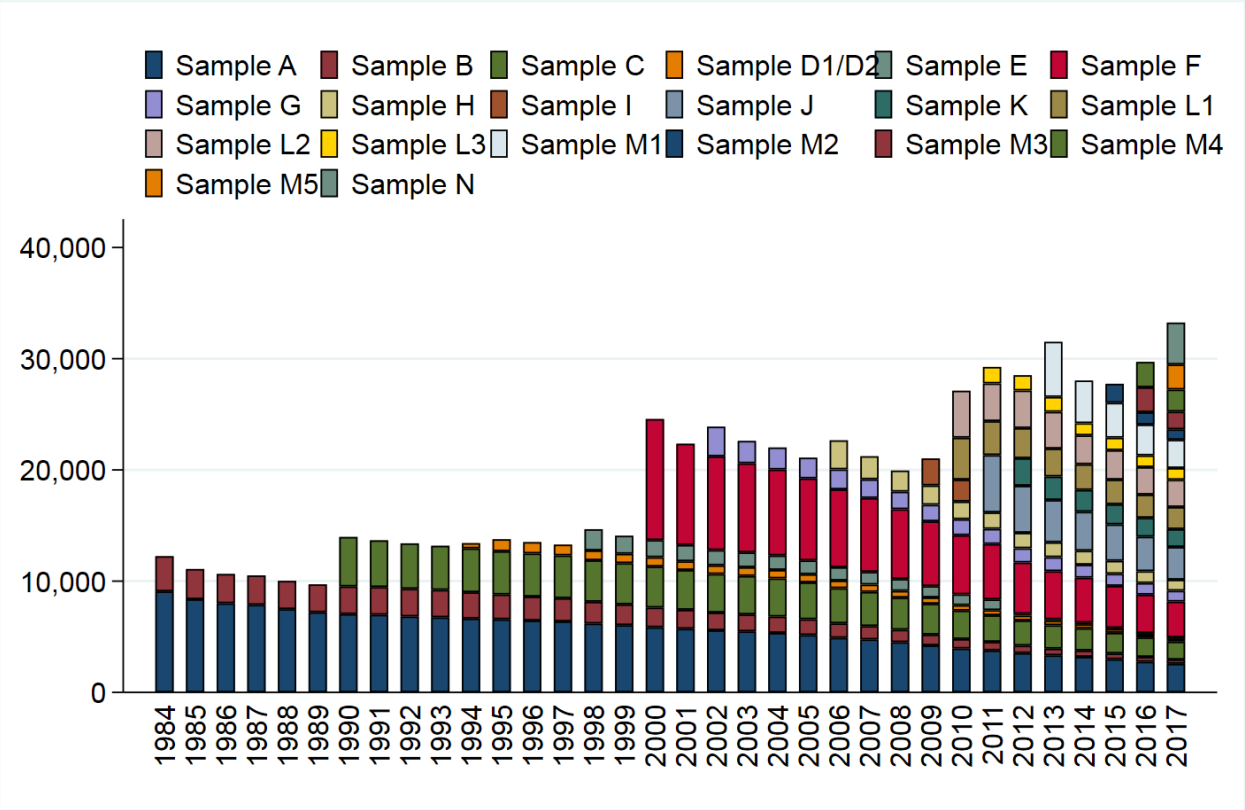
2.3 Development of Sample Sizes

Individuals who refuse participation or are not available for an interview are kept in the so-called "gross" sample of the study as long as they continue to live in households with at least one participating person. Once the entire household declines to respond in two consecutive waves of data collection, all individuals from the household are removed from the SOEP. 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 person-level drop-outs, refusals, moving abroad, etc. However, due to new persons moving into already existing households, and children reaching the minimum respondent's age of 16, and thereby increasing the sample size, this negative development is offset somewhat.

2.3.1 Starting Sample Size of the SOEP Samples

Sample	Year	Households (net)	Persons(gross)	Respondents (net)	Partial Unit Non-Response (percent)	Children (gross)
A	1984	4528	11422	9076	0.6	2290
B	1984	1393	4830	3169	0.7	1636
C	1990	2179	6131	4453	1.9	1591
D1	1994	236	733	471	2.9	248
D1/D2	1995	541	1668	1078	6.1	517
E	1998	1057	2446	1910	3.5	466
F	2000	6043	14510	10880	5.5	2991
G	2002	1224	3538	2671	6.1	693
H	2006	1506	3407	2616	6.0	623
I	2009	1495	3428	2432	13.4	620
J	2011	3136	6873	5161	9.9	1147
K	2012	1526	3286	2473	9.2	563
L1	2010	2074	7939	3770	6.7	3900
L2	2010	2500	9063	4227	5.1	4611
L3	2011	924	3645	1487	4.2	2092
M1	2013	2723	8522	4964	17.8	2481
M2	2015	1096	3048	1711	19.3	927
M3	2016	1775	4823	2351	22.0	1808
M4	2016	1779	7297	2465	27.1	3915

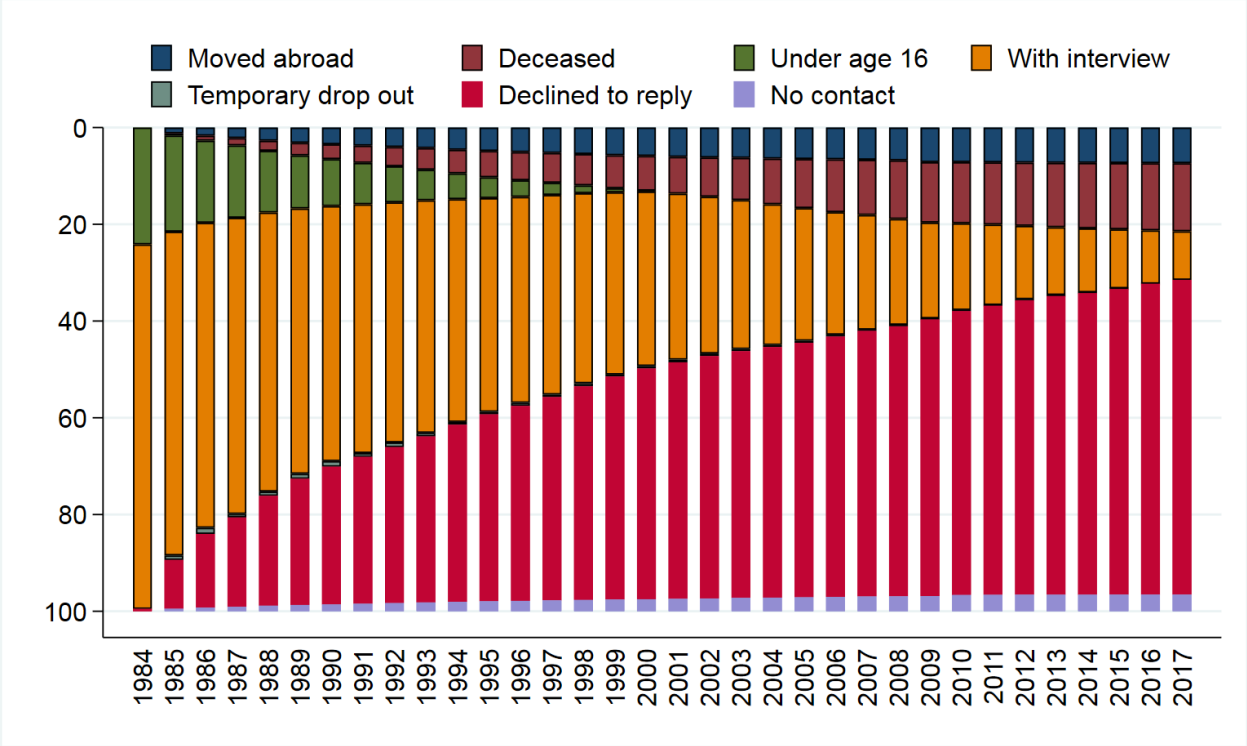
2.3.2 Cross-Sectional Development of Sample Size (Respondents) Cross-Sectional Development of



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 field-work 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.

2.3.3 Longitudinal Development of the 1984 Population



Download Stata Code to create figure

CHAPTER

THREE

SURVEY DESIGN

3.1 Survey Instruments

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

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

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

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

3.2 Survey Concepts

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

- Questions about a point in time (present) e.g. current employment status or current levels of satisfaction

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

3.3 Survey Modes

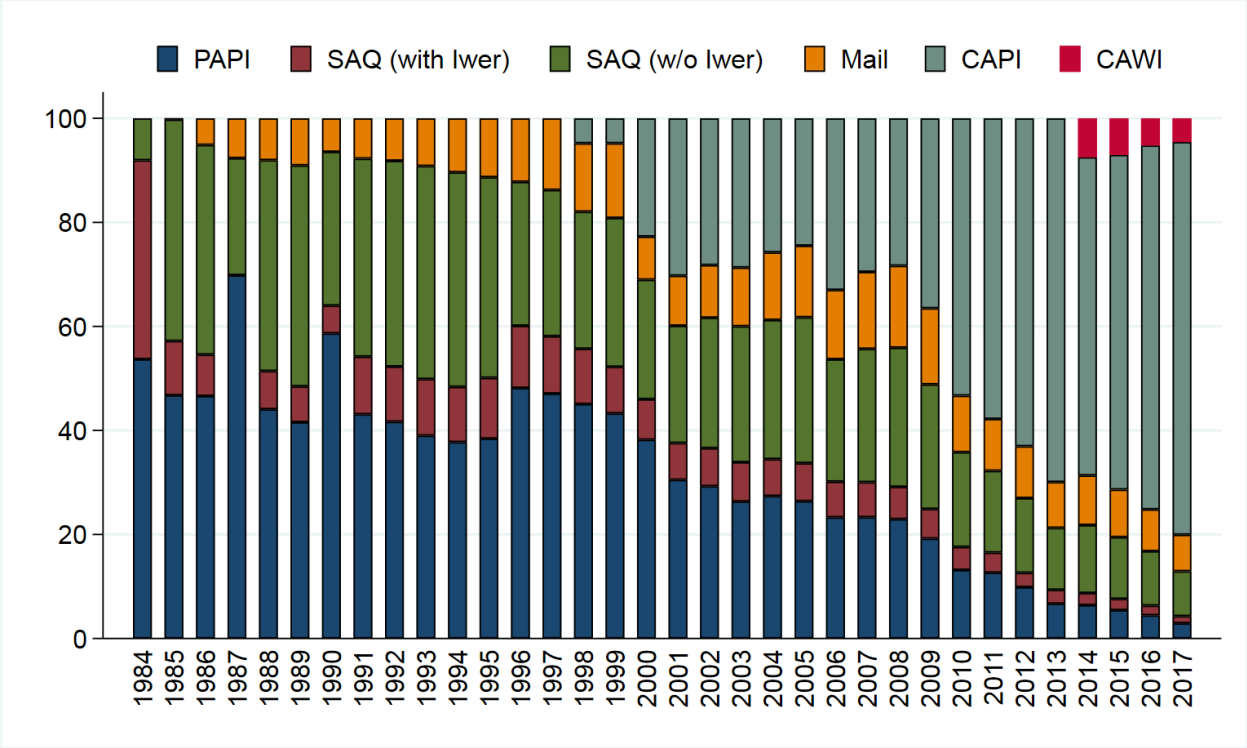
The SOEP uses several different modes to collect the data. Originally, the respondent's answers were recorded by an interviewer who filled in a paper questionnaire, the so called pen-and-paper interview or PAPI. The personal contact between interviewer and respondent is important for the success of the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

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

In 1998, interviews were conducted with computers for the first time, in computer-assisted personal interviews, or in CAPI mode. Compared to PAPI, CAPI is much more efficient in transferring the data into an electronic format, which was an important asset especially with the extensions of the panel starting in the year 2000. The CAPI mode was first conducted in parallel to the PAPI mode, meaning that interviewers and respondents were free to chose how they wanted to do the interview. This was important for the "older" sample members (respondents as well as interviewers), who were used to the PAPI concept. Only in the most recent samples (starting in subsample J), CAPI is the only mode. 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, the CAPI implementation is relatively simple compared to what would be technically feasible. For example, the SOEP basically does not use any form of dependent interviewing (i.e. referring to respondent data from previous waves), because this cannot be easily implemented in the PAPI-mode. Also, the filtering structure is very simple in the SOEP, because any respondent must be able to follow the interview path on her/his own on paper. Still, some technical features like the control of value ranges (e.g. month of birth, year of first marriage) or the randomization of scale items are implemented in the CAPI version of the questionnaire.

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



Download STATA Code to create figure.

3.4 Panel Care

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

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

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

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

CHAPTER

FOUR

PRINCIPLES OF DATA STRUCTURE

4.1 Panel Data Analysis

The data structure for panel data consists of three dimensions. At first, the respective examination units (n) and a matrix of dependent and independent variables (y,x) are completely analogous to a cross-sectional design. Another level is 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 are often not completely provided with 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. The data of social science panel data often show a data structure, which is 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 (for an overview see Andreß et al. 2013). The statistical analysis models of panel data range from structural equation models (Finkel 1995), various regression models (Giesselmann/Windzio 2012), event analysis (Blossfeld 2010), sequence data analysis (Brüderl/Scherer 2005), latent growth models (Schiedeck/Wolff 2010) to causal analyses using matching methods (Gangl 2010). 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 (Brüderl 2010).

4.2 Data Structure of 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 person 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 is stored as *Spell Data*. For each spell there is a definition of the spell type, begin, 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 (person or household) over time. That's why we provide longitudinal data sets, 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 to define longitudinal populations very easily: PPFADL and HPFADL. HPFADL includes all households which have been interviewed successfully at least once. Similarly, PPFADL contains all persons who have ever lived in a household that has participated in the SOEP, i.e. that has been captured in HPFADL, including non-respondents and children. Both data files contain one record per household or person, respectively, with wave-specific variables for each year's survey status. In addition to some time-invariant information (like gender, year of birth, migrant status), these files contain all necessary identifiers to combine other files with PPFADL and HPFADL.

Although they provide essential information, PPFADL and HPFADL 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 data sets in the long-format.

4.2.1 Cross-sectional data files (CS)

Individual Level Data		
Gross Sample	Net Sample	
	Questionnaire Data	Generated Data
$pbrutto	$p $pkal $pluecke $kind (from $h)	$pgen $pequiv

Household Level Data		
Gross Sample	Net Sample	
	Questionnaire Data	Generated Data
$hbrutto	$h	$hgen

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 all waves of one group of datasets are referred to. For example, $H refers to all household level datasets 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 which were eligible for an interview in any given year.

Data structure

Cross sectional data is a type of data, which observes many subjects at the same point of time. Each person is assigned a row in the data set and is only included once in such a data set. By merging cross-sectional SOEP data across waves (e.g. „bfp" and „bgp"), you receive a dataset in wide-format.

4.2.2 Data Structure in wide-format (wide)

The SOEP data is offered in different data structures. In 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 4 datasets in wide-format: ppfad, phrf, hpfad, hhrf

row	ID	syear	sex	income
1	1	2016	m	1500
2	2	2016	m	1000
3	6	2016	f	2000
4	8	2016	m	5500

4.2.3 Data Structure in long Format (long)

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

Row	ID	syear	sex	income
1	1	2010	f	1500
2	1	2011	f	1500
3	1	2012	f	2000
4	2	1999	m	5500
5	2	2000	m	5500

4.2.4 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 is 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

Here are some recommended literature suggestions:

Working with spell data:

Working with spell data (pdf):

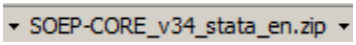
Working with spell data (do-files):

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

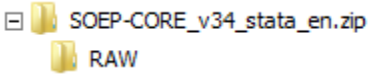
Generating spell data:

4.3 Data Sets SOEP-Core

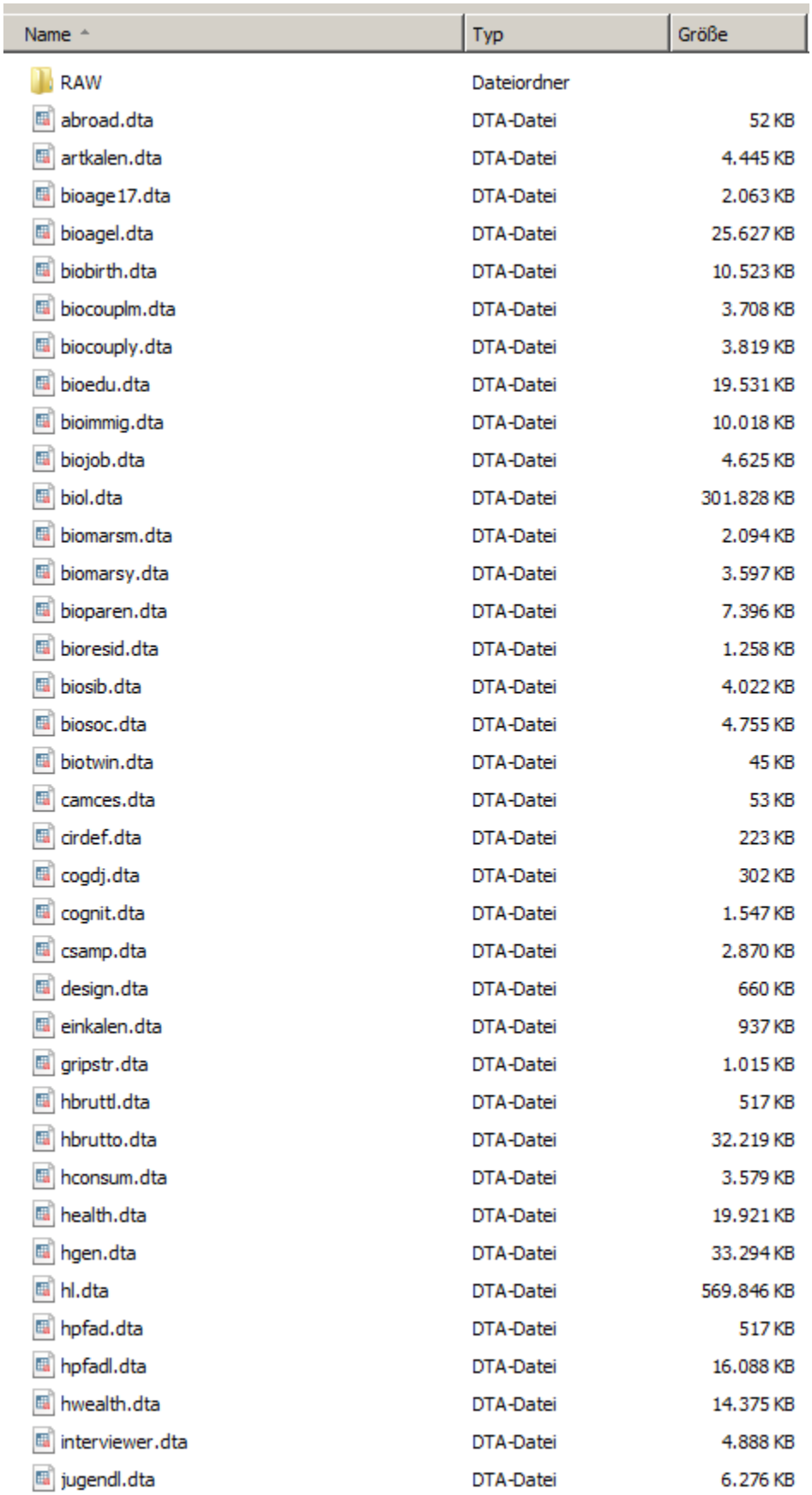
In the SOEP, each survey year is allocated to a data wave, which is abbreviated with the letters of the alphabet. The current data wave can contain several versions, which are displayed in SOEP with a "v" for version and the respective version number. The version number represents the survey years since 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 occur over time, such as v34.1. If updates have been carried out, users are informed about them via various information channels and asked to order the data again. After ordering the data, the data will be sent to you as a zip-file.



Within this zip file you will find various data sets and a "RAW" subdirectory.



The data sets above the "RAW" subdirectory are highly compressed and an easy to analyze version of the SOEP data.

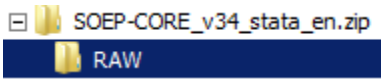


Name	Typ	Größe
RAW	Dateiordner	
abroad.dta	DTA-Datei	52 KB
artkalen.dta	DTA-Datei	4.445 KB
bioage17.dta	DTA-Datei	2.063 KB
bioagel.dta	DTA-Datei	25.627 KB
biobirth.dta	DTA-Datei	10.523 KB
biocouplm.dta	DTA-Datei	3.708 KB
biocouply.dta	DTA-Datei	3.819 KB
bioedu.dta	DTA-Datei	19.531 KB
bioimmig.dta	DTA-Datei	10.018 KB
biojob.dta	DTA-Datei	4.625 KB
biol.dta	DTA-Datei	301.828 KB
biomarsm.dta	DTA-Datei	2.094 KB
biomarsy.dta	DTA-Datei	3.597 KB
bioparen.dta	DTA-Datei	7.396 KB
bioresid.dta	DTA-Datei	1.258 KB
biosib.dta	DTA-Datei	4.022 KB
biosoc.dta	DTA-Datei	4.755 KB
biotwin.dta	DTA-Datei	45 KB
camces.dta	DTA-Datei	53 KB
cirdef.dta	DTA-Datei	223 KB
cogdj.dta	DTA-Datei	302 KB
cognit.dta	DTA-Datei	1.547 KB
csamp.dta	DTA-Datei	2.870 KB
design.dta	DTA-Datei	660 KB
einkalen.dta	DTA-Datei	937 KB
gripstr.dta	DTA-Datei	1.015 KB
hbruttl.dta	DTA-Datei	517 KB
hbrutto.dta	DTA-Datei	32.219 KB
hconsum.dta	DTA-Datei	3.579 KB
health.dta	DTA-Datei	19.921 KB
hgen.dta	DTA-Datei	33.294 KB
hl.dta	DTA-Datei	569.846 KB
hpfad.dta	DTA-Datei	517 KB
hpfadl.dta	DTA-Datei	16.088 KB
hwealth.dta	DTA-Datei	14.375 KB
interviewer.dta	DTA-Datei	4.888 KB
jugendl.dta	DTA-Datei	6.276 KB

The data in SOEP-Core are no longer only provided as wave-specific individual files but rather 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 is one of the biggest assets of the SOEP. That's why we provide longitudinal data sets, such as pl or hl. The advantage of such a data set is that longitudinal analyses can be carried out without great effort.

If you need more information about the long data structure visit the chapter *Data Structure in long Format (long)*.

In the "RAW" directory you will find all wave-specific data sets that were used to generate the long data sets on the previously presented level.



Name	Typ	Größe
ah.dta	DTA-Datei	738 KB
ahbrutto.dta	DTA-Datei	122 KB
ahgen.dta	DTA-Datei	517 KB
akind.dta	DTA-Datei	187 KB
ap.dta	DTA-Datei	4.195 KB
apausl.dta	DTA-Datei	205 KB
apbrutto.dta	DTA-Datei	434 KB
apequiv.dta	DTA-Datei	5.865 KB
apgen.dta	DTA-Datei	1.952 KB
apkal.dta	DTA-Datei	9.770 KB
bah.dta	DTA-Datei	7.770 KB
bahbrutto.dta	DTA-Datei	949 KB
bahgen.dta	DTA-Datei	1.566 KB
bajugend.dta	DTA-Datei	1.151 KB
bakind.dta	DTA-Datei	1.315 KB
bap.dta	DTA-Datei	28.594 KB
bapbrutto.dta	DTA-Datei	2.697 KB
bapequiv.dta	DTA-Datei	18.277 KB
bapgen.dta	DTA-Datei	3.966 KB
bapkal.dta	DTA-Datei	15.446 KB
bapluecke.dta	DTA-Datei	117 KB
bavp.dta	DTA-Datei	41 KB
bbh.dta	DTA-Datei	9.127 KB
bbhbrutto.dta	DTA-Datei	1.028 KB
bbhgen.dta	DTA-Datei	1.706 KB
bbjugend.dta	DTA-Datei	1.198 KB
bbkind.dta	DTA-Datei	1.452 KB
bbp.dta	DTA-Datei	34.277 KB
bbpbrutto.dta	DTA-Datei	2.960 KB
bbpequiv.dta	DTA-Datei	19.560 KB
bbpgen.dta	DTA-Datei	4.279 KB
bbpkal.dta	DTA-Datei	16.576 KB
bbpluecke.dta	DTA-Datei	216 KB
bbvp.dta	DTA-Datei	42 KB
bch.dta	DTA-Datei	8.902 KB
bcpkal.dta	DTA-Datei	16.118 KB
bcpluecke.dta	DTA-Datei	269 KB
bcvp.dta	DTA-Datei	42 KB

Within this "RAW" directory, the data sets are stored on a wave-specific basis and are the generation basis for the majority of the long data sets described above. In addition to these wave-specific data sets, the "RAW" directory also contains additional data sets in cross-sectional format that have not yet been distributed in long format ($school, $school2, ev, exit, $pkalost and pbr_hhchch).

To understand the data set and variable names, visit the *Labeling SOEP-Core* chapter.

4.3.1 Overview Data Sets

Your data distribution file contains five different types of data sets:

Tracking Data	Original Data	Survey Data	Generated Data	Spell Data
hpfad	abroad	csamp	bioage17	artkalen
hpfadl	biol	design	bioagel	biocouplm
$hbrutto	ev	exit	biobirth	biocouply
hbrutto	$h	hhrf	bioedu	biomarsm
$pbrutto	hl	pbr_hhch	bioimmig	biomarsy
pbrutto	$host	phrf	biojob	einkalen
pbr_exit	$jugend		bioparen	lifespell
ppfad	jugendl		bioresid	migspell
ppfadl	lueckel		biosib	pbiospe
	$p		biosoc	refugspell
	pl		biotwin	sozkalen
	$pausl		camces	
	$pluecke		cogdj	
	$post		cognit	
	$school		gripstr	
	$school2		hconsum	
	$vp		health	
	vpl		$hgen	
			hgen	
			hwealth	
			interviewer	
			kidl	
			$kind	
			mihinc	
			$pequiv	
			pequiv	
			pflege	
			$pgen	
			pgen	
			$pkal	
			pkal	
			$pkalost	
			pwealth	
			timepref	
			trust	

4.3.2 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 is conducted. These data sets should be understood by you as initial data. You can use the tracking data to merge your research-relevant variables via the person and household numbers.

Dataset	Label	Format	Identifier (ID)	Special Identifier
hpfad	Household Tracking File	wide	hhnrakt, $hhnr	
hpfadl	Household Tracking File	long	hid, syear, cid	
$hbrutto	Gross Household Data	wide	hhnrakt, hhnr	intid1, intid
hbrutto	Gross Household Data	long	hid, syear, cid	intid1, intid
pbr_exit	Cumulated Exit	long	pid, hid, syear, cid	hhnrold
$pbrutto	Gross Individual Data	wide	persnr, hhnrakt, hhnr	$hhnrold
pbrutto	Gross Individual Data	long	pid, hid, syear, cid	intid, hhnrold
ppfad	Individual Tracking File	wide	persnr, hhnr, $hhnr,	
ppfadl	Individual Tracking File	long	pid, hid, syear, cid	parid

hpfad „Household Tracking File" (wide): For all years since 1984, the HPFAD data set contains information on all households that have ever participated in the SOEP survey at any point in time. HPFAD is important for the delimitation of the examination unit (household), especially for longitudinal analyses. HPFAD is particularly suitable for household analyses and can be used for preselection of specific households.

hpfadl „Household Tracking File“ (long): HPFADL consists of all waves of the data sets **hpfad „Household Tracking File" (wide)** and **hhrf „Weighting and staying probabilities“ (wide)** of SOEP-Core.

$hbrutto „Gross Household Data“ (CS): $HBRUTTO covers all households, who were successfully interviewed for the first time in wave $ or were contacted for the purpose of being interviewed again in wave $. The data sets provide gross cross-sectional information on all SOEP households' interviews as well as their positions in the panel frame work.

hbrutto „Gross Household Data“ (long): HBRUTTO consists of all waves of the data sets **$hbrutto „Gross Household Data“ (CS):** of SOEP-Core.

pbr_exit„Cumulated Exit“ (long)::

$pbrutto „Gross Individual Data“ (CS) : $PBRUTTO covers all respondents, who were successfully interviewed for the first time in wave $ or were contacted for the purpose of being interviewed again in wave $. The data set provides gross cross-sectional information on all SOEP respondents' interviews as well as their positions in the panel frame work.

pbrutto „Gross Individual Data“ (long): PBRUTTO consists of all waves of the data sets **$pbrutto „Gross Individual Data“ (CS)** of SOEP-Core.

ppfad „Individual Tracking File“ (wide): For all years since 1984, the PPFAD data set contains information on all persons who have ever lived in a SOEP household at a survey time (i.e. all respondents, but also children under 17 years of age and persons who have never given an interview). PPFAD is important for the delimitation of the examination units (persons), especially for longitudinal analyses.

ppfadl „Individual Tracking File“ (long): PPFADL consists of all waves of the data sets **ppfad „Individual Tracking File“ (wide)** and **phrf „Weighting and staying probabilities“ (wide)** of SOEP-Core. It contains one record for each individual and year a person has been a member of a respondent household. It is keyed on PID, the Cross-Wave Person Identifier, and SYEAR, the survey year identifier. It contains the Household ID, and never changing individual characteristics, individual weights, as well as the response status, for that individual at each wave.

4.3.3 Original Data

These data sets contain the direct information of the respondents. The contents of these variables are 1:1 the contents of the survey instruments. By searching in the questionnaires you can determine the exact wording of the question or also possible filter guidance.

Dataset	Label	Format	Identifier (ID)	Special Identifier
abroad	Questionnaire for people moved abroad	long	persnr, hhnrakt, syear, hhnr	
biol	Biographical Data	long	pid, hid, syear, cid	intid
ev	First wealth module	wide	persnr, hhnrakt, hhnr	
$h	Household questionnaire	wide	hhnrakt, syear, hhnr	intid
hl	Household questionnaire	long	hid, syear, cid	intid
$h_refugees	Household questionnaire Refugee Sample	wide	hhnrakt, syear, hhnr	intid
ghost	East specific questions from the Household questionnaire	wide	hhnrakt, hhnr	intid
$jugend	Youth questionnaire for first time respondents at age 17	wide	persnr, hhnrakt, syear, hhnr	intid
jugendl	Youth questionnaire for first time respondents at age 18	long	pid, hid, syear, cid	intid
$p	Personal questionnaire	wide	persnr, hhnrakt, syear, hhnr	intid
pl	Personal questionnaire	long	pid, hid, syear, cid	intid
$p_mig	IAB-SOEP Migration Sample: Original Individual questionnaire	wide	pid, hid, syear, cid	intid
$p_refugees	Personal questionnaire Refugee Sample, incl. Biography	wide	persnr, hhnrakt, syear, hhnr	intid
$pausl	Migrant specific questions in the Personal Questionnaire	wide	persnr, hhnrakt, hhnr	
$pluecke	Follow-Up Questioning	wide	persnr, hhnrakt, hhnr	intid
$school	Questionnaire: Early Youth, 12-13 years old	wide	persnr, hhnrakt, syear, hhnr	intid
$school2	Questionnaire: Early Youth, 14-15 years old	wide	persnr, hhnrakt, syear, hhnr	intid
$vp	Questionnaire: the deceased person	wide	persnr, hhnrakt, syear, hhnr	vpersnr, intid

abroad „Questionnaire for people moved abroad“ (CS): With the pilot study ”Life outside Germany” in 2008, the longitudinal German Socio-Economic Panel Study (SOEP) ventured into completely uncharted methodological territory by attempting to locate the addresses of former participants in the German household panel study SOEP who have since immigrated abroad, and to survey these individuals with the help of a specially developed written questionnaire on the reasons for their international move. The project was discontinued due to insufficient case numbers in 2014.

biol "Biographical Data" (long): BIOL contains cumulated individual-level data from the biographical questionnaire.

ev „First wealth module“ (long):

$h „Household questionnaire“ (CS): The $H-files contain all questions of the household questionnaire.

hl „Household questionnaire“ (long): HL contains all waves of the data sets from SOEP-Core.

h_refugees „Household questionnaire Refugee Sample“ (CS): The $H-files contain all questions of the household refugees questionnaire.

`only 1990` **ghost „East specific questions from the Household questionnaire“ (CS):** The $host file contains east specific questions from the household questionnaire. For the year 1990 the data provides information about east specific topics about the German reunification i.e. presents from the BRD.

$jugend „Youth questionnaire for first time respondents at age 17“ (CS): Since 2000 (wave Q), first-time respondents between the ages 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 $JUGEND datasets.

jugendl „Youth questionnaire for first time respondents at age 17“ (long): JUGENDL contains the waves q (2000) up to the current wave of **$jugend „Youth questionnaire for first time respondents at age 17“ (CS)** of SOEP-Core.

$p „Individual questionnaire“ (CS): The $P-files contain all variables of the individual questionnaire for the wave $. In addition, the individual-specific data of the samples IAB-SOEP Migration and IAB-BAMF-SOEP Refugee Survey are integrated in the original $P data set.

pl „Individual questionnaire“ (long): The PL data set contains all waves of the **$p „Individual questionnaire“ (CS):** data sets of SOEP-Core. In addition, the PL file contains all variables of all waves of the data sets **$post „East specific questions from the Individual questionnaire“ (CS)** and **$pausl „Migrant specific questions in the Individual Questionnaire“ (CS)**.

`2013-2016` **$p_mig „IAB-SOEP Migration Sample: Original Individual questionnaire“ (CS):** The original data from the Sample M specific survey instrument can be found in the dataset $P_MIG, combining the individual and the biographical questionnaire. **Since the current version “v34”, the data set is not part of the SOEP-Core distribution file anymore and has to be ordered separately**. The variables are included in original or generated datasets. Variables equivalent to variables in the individual questionnaire of other samples are included in the dataset $P, Variables equivalent to variables in the biography questionnaire of other samples are included in the respective biography dataset (e.g. BIOMARSM), the comprehensively surveyed migration biography can be found in the new dataset MIGSPELL.

`only 2016` **$p_refugees „IAB–BAMF-SOEP Survey of Refugees in Germany: Original Individual questionnaire“ (CS):** The original data from the survey instruments used in Samples M3 and M4 can be found in original format in the dataset $P_REFUGEES, where the individual and the biographical questionnaires are combined. **Since the current version “v34”, the data set is not part of the SOEP-Core distribution file anymore and has to be ordered separately**. The variables are integrated in original or generated datasets. Variables equivalent to those in the individual questionnaire of other samples are included in the dataset $P. Also included in $P are all variables which will be asked more than once, but specific to the refugee questionnaire, Variables equivalent to those in the biographical questionnaires in other samples are included in the respective biographical datasets (e.g., BIOMARSM), the comprehensively surveyed migration biography can be found in the new dataset REFUGSPELL.

`1984-1995` **$pausl „Migrant specific questions in the Individual Questionnaire“ (CS)**:

$pluecke „Follow-Up Questioning“ (CS): Temporary drop-outs (“gaps”) can cause problems for longitudinal analyses. This is especially true for the employment and income data stored. That is why the SOEP tries to fill in at least some of the central missing information. $PLUECKE is a small questionnaire covering information on the year previous to which the drop-out occurred. This covers questions on job-related changes, calendar of occupation, income, education and qualification.

$post „East specific questions from the Individual questionnaire“ (CS): The $post files contain east specific questions from the individual questionnaire. For the years 1990 and 1991 the data provides information about east specific topics.

$school „Questionnaire: Early Youth, 12-13 years old“ (CS): Since 2014 the $SCHOOL-files contain all variables of the „Pre-teen (Schülerinnen und Schüler)“ questionnaire. Therefore the data sets provide variables about school, home, leisure time, health, self-perception and relationships with friends, siblings and parents.

$school2 „Questionnaire: Early Youth, 14-15 years old“ (CS): Since 2016 the $SCHOOL2-files contain all variables of the „Early Youth (Frühe Jugend)“ questionnaire. Therefore the data sets provide variables about self-perception, independence, school, leisure time or relationships with friends, siblings and parents.

$vp „Questionnaire: the deceased person“ (CS): The $VP-files contain information about respondents who lost a person in the previous year. It provides information about the deceased person and the respondent who reported the case of death.

4.3.4 Survey Data

These data sets contain surveymethodical information for SOEP core. The various data sets provide detailed exit information from respondents or household weighting factors that you need for representative analyses.

Dataset	Label	Format	Identifier (ID)	Special Identifier
csamp	Sample Definition	long	cid	
design	Survey Design	wide	hhnr	intid
exit	Cumulative drop-outs	wide	persnr, hhnr, syear	
hhrf	Weighting and staying probabilities	wide	hhnrakt, hhnr	
pbr_hhch	PBR_HHCH	wide	persnr, hhnrakt, syear, hhnr	pnralt, pnrneu, hhnrold
phrf	Weighting and staying probabilities	wide	persnr, hhnr	

csamp „Sample Definition“ (long):

design „Survey design“ (CS): The dataset DESIGN provides information on the stratified sampling of the SOEP in 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. Unique inclusion probabilities pertain to each of these strata. The variable DESIGN contains the inverse of this probability, i.e., the design weight.

exit „Cumulative drop-outs“ (CS):

hhrf „Weighting and staying probabilities“ (wide): In the SOEP database, different weighting variables for cross-sectional as well as for different kinds of longitudinal weighting are set aside for each household in the HHRF-file.

pbr_hhch „PBR_HHCH“ (CS):

phrf „Weighting and staying probabilities“ (wide): In the SOEP database, different weighting variables for cross-sectional as well as for different kinds of longitudinal weighting are set aside for each person in the PHRF-file.

4.3.5 Generated Data

The SOEP team has prepared these data sets for you in a special way. The data sets are prepared in a research-friendly manner and are subjected to additional plausibility checks and quality controls. They usually consist of several variables, of different survey instruments and are described by the documentation provided. Therefore, these data sets cannot be assigned 1:1 to a survey instrument.

Dataset	Label	Format	Identifier (ID)	Special Identifier
bioage17	Generated biographical youth information	wide	persnr, hhnrakt, syear, hhnr	bymnr, byvnr, intid
bioagel	Generated biographical information	long	persnr, hhnrakt, syear, hhnr	persnre
biobirth	Generated biographical information	wide	persnr, hhnr	kidpnr01-kidpnr15
bioedu	Generated biographical information	wide	persnr, hhnr	
bioimmig	Generated biographical information	long	persnr, hhnrakt, syear, hhnr	
biojob	Generated biographical information	wide	persnr, hhnr	
bioresid	Generated biographical information	wide	persnr, hhnrakt, syear, hhnr	intid
biosib	Generated biographical information	wide	persnr, hhnr	sibpnr1-sibpnr11
biosoc	Generated biographical information	wide	persnr, hhnrakt, syear, hhnr	intid
biotwin	Generated biographical information	wide	persnr, hhnr	pnrtwin, pnrtrip, pnrquad
camces	Highest Educational Qualification, Migrants Sample M1 and M2	wide	persnr, hhnrakt, syear, hhnr	
cogdj	Data on cognitive tests (Youth)	wide	persnr, syear, hhnr	
cognit	Data on cognitive potential	wide	persnr, syear, hhnr	intid
gripstr	Measures grip strength	wide	persnr, syear, hhnr	intid
hconsum	Hosehold Consume Module	wide	hhnrakt, syear, hhnr	
health	Data on health indicators	wide	persnr, syear, hhnr	
$hgen	Generated Household Data	wide	hhnrakt, hhnr	
hgen	Generated Household Data	long	hid syear cid	
hwealth	Wealth Module	long	hhnrakt, syear, hhnr	
interviewer	Data on the SOEP Interviewer	long	hhnr, syear	intid
kidlong	Data on children	long	persnr, hhnrakt, syear, hhnr	
$kind	Data on children (from HH-Questionnaire)	wide	persnr, hhnrakt, hhnr	
mihinc	Multiple imputed data on monthly household income	long	hhnrakt, syear, hhnr	
$pequiv	Cross-national Equivalent File	wide	persnr, hhnrakt, syear, hhnr	
pflege	Persons needing care within the household	long	persnr, syear, hhnr	
$pgen	Generated Individual Data	wide	persnr, hhnrakt, hhnr	
pgen	Generated Individual Data	long	pid, hid, syear, cid	
$pkal	Individual Calendar	wide	persnr, hhnrakt, hhnr	
pkal	Individual Calendar	long	pid, hid, syear, cid	
$pkalost	Individual Calender	wide	persnr, hhnrakt, hhnr	
pwealth	Wealth Module	long	persnr, hhnrakt, syear	

Continued on next page

Table 2 – continued from previous page

Dataset	Label	Format	Identifier (ID)	Special Identifier
timepref	Experiment on time preferences	wide	persnr, hhnrakt, syear, hhnr	
trust	Experiment on trust	long	persnr, hhnrakt, syear, hhnr	

bioage17 „Generated biographical information“ (CS): The design of the dataset BIOAGE17 is patterned after the 2001 Youth Questionnaire, which is the standard version for subsequent years. A special group of first time respondents are young persons living in a panel household, who reach the surveying age of 17 years. From this specific group of panel entrants, we are able to obtain some more detailed information on youth and socialisation than 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.

biobirth „Generated biographical information“ (CS): The file BIOBIRTH provides information on fertility histories of adult respondents in the SOEP. Until 2014 (version 30, wave BD) the data was 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 a panel entry since 2001) who has ever provided at least one successful SOEP interview.

bioedu „Generated biographical information“ (CS): The Socio-Economic Panel Study (SOEP) contains a broad range of variables which cover early child education and care, educational participation, educational degrees and other related topics. It is the aim of the BIOEDU dataset to provide ready-made variables on educational transitions and related topics in order to support analyses in a longitudinal perspective.

bioimmig „Generated biographical information“ (long): The variables contained in BIOIMMIG deal with questions related to foreigners in (and migrants to) Germany. Specifically, questions concerning desire to return to the home country, the presence of relatives in the home country, reasons for coming to Germany, and conditions upon initial arrival in Germany.

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 of BIOJOB are time-invariant. Information on occupational changes and on the age at the most recent change of occupation refer to the date of the respondent’s biography interview.

bioresid „Generated biographical information“ (CS): In 1994 questions with a focus on occupancy were introduced to the Biographical Questionnaire asking for the duration of residence in the current dwelling and any second residence. The information surveyed in the Biographical Questionnaire is stored in the file BIORESID.

biosib „Generated biographical information“ (CS): BIOSIB provides information on siblings living within the SOEP households. The data set contains the person numbers of all siblings in an observed family. It includes information on their sex, their year of birth, the number of siblings, the individual’s position within the birth order, and on the relationship between the observed siblings.

biosoc „Generated biographical information“ (CS): BIOSOC contains retrospective 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 data set.

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 person 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 persons

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 data set, but also triplets or quadruple siblings.

camces „Highest Educational Qualification, Migrants Sample M1 and M2“ (CS): The CAMCES-File provides information about Computer-Assisted Measurement and Coding of Educational Qualifications in Surveys.

cogdj „Data on cognitive tests (Youth)“ (CS): In SOEP 2006, a separate questionnaire with cognitive tests for adolescents was used for the first time: “Lust auf DJ”. In this case, “DJ” stands for “Thinking Sports and Youth (Denksport und Jugend)”, but was also specifically selected to arouse the more common association of “Disc Jockey”. For all interviewees aged 16 - 17 years, the questionnaire “Lust auf DJ” was used and created.

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. Im COGNIT06 werden den Nutzern die aggregierten Summen-Scores (jeweils Gesamtwerte für drei Zeitpakete, sog. „parcels“ von 30, 60 und 90 Sekunden) zur Verfügung gestellt.

gripstr „Measures 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 consume module“ (CS)“: We were faced with three methodological challenges in generating the final consumption data. Firstly, due to the design of the consumption module, inconsistent answers arose between the monthly and annual amounts spent for consumption. Secondly, we encountered the well-known phenomenon of missing data, here in particular item nonresponse. And thirdly, 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 data set “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 put into a two year replication period. In the HEALTH-File users find i.e. the generated variables on height and weight with imputation flags and a user-friendly longitudinal checked generated variable of the Body Mass Index (BMI).

$hgen „Generated Household Data“ (CS): In order to minimize computing efforts for the user, the SOEP provides yearly status variables on household level. The $HGEN data provides a set of time-consistent variables generated from the SOEP household questionnaire. It only includes households who participated in the respective year.

hgen „Generated Household Data“ (long): HGEN contains all waves of the **$hgen „Generated Household Data“ (CS)** data sets of SOEP-Core.

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 person-level information in PWEALTH. However for all persons with valid household level information that did refuse to 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 does not only aim at collecting high-quality data on the living conditions and well-being of households, but –as a by-product of internal quality assurance processes– it lends itself increasingly as a empirical source for survey research. The purpose of the INTERVIEWER file is to provide user convenient access to all available, longitudinal information on the SOEP interviewers.

kidlong „Data on children“ (long): The variables stored in the KIDLONG file are based on the information annually collected and stored in the wave-specific $KIND files. The relevant information is not provided by children themselves but by answers to the questions in the household questionnaire given by the respondent within the household (mostly the head of the household). This data is reaggregated at the person level and stored as child-specific entries in the file **$kind „Data on children (from HH-Questionnaire)“ (CS):**.

$kind „Data on children (from HH-Questionnaire)“ (CS): The variables from the annual $kind files are not based on answers provided by the children themselves, but by answers provided by the head of household. This data is

re-aggregated on the person level and saved as child-specific entries in the file $kind. The annual $kind datasets also contain additional information on institutional care and school attendance for children and young people.

mihinc „Multiple imputed data on monthly household income (long)“: The dataset MIHINC contains the complete imputation results and is separately available. To be compatible with methods for analysing multiply imputed data, MIHINC is constructed in the so called stacked or MIM Dataset 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 the current household income.

$pequiv „Cross-national Equivalent File“ (CS): 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 provided in the CNEF but also further single income components.

pequiv „Cross-national Equivalent File“ (long) PEQUIV contains all waves of the **$pequiv „Cross-national Equivalent File“ (CS)** data sets of SOEP-Core.

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 analyses on an individual level, this information has been restructured and stored in the cumulative file PFLEGE.

$pgen „Generated Individual Data“ (CS): The $PGEN-files contain user friendly data on the individual level which are consolidated from different sources. The plausibility is in many respects longitudinally validated, therefore the data here are in most situations superior compared to the data in $P. The file contains one row for each person (persnr is unique) with a completed individual or youth questionnaire.

pgen „Generated Individual Data“ (long): PGEN contains all waves of the **$pgen „Generated Individual Data“ (CS):** data sets of SOEP-Core.

$pkal „Individual Calendar“ (CS): The $pkal datasets contain calender variables from the Individual questionnaire. The dataset includes the activity status on a monthly basis as well as the income status of a person.

pkal „Individual Calendar“ (long) PKAL contains all waves of the **$pkal „Individual Calendar“ (CS)** data sets of SOEP-Core.

`1990-1991` **$pkalost „Individual Calendar“ (CS):**

pwealth „Wealth module“ (long): In the year 2002, the individual questionnaire included for the first time a special module focusing on wealth. This section 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 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. 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 on 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): Data set of the economic behavior experiment on trust and trustworthiness from the survey years 2003, 2004 & 2005, which serves to measure trust, based on an investment game. This is a one-off game for two actors who relate to each other 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, it decides how much of its

own credit it will transfer to the first player (zero to ten points). As with the first transfer, your points at the recipient are doubled. After the decision of the second player, the game ends and the other players are paid their income (one point corresponds to one euro, the sum is sent out as a cheque a few days later). The TRUST data set thus contains the information from all three waves in which the behavioral experiment was conducted.

4.3.6 Spell Data

General information about spell data in the SOEP can be found in the chapter *Data Structure in spell format (spell)*

Dataset	Label	Format	spellnr	Special Identifier
artkalen	Spell data from the activity calendar	wide	persnr, hhnr	
biocouplm	Generated biographical information	long	persnr, hhnr	coupid
biocouply	Generated biographical information	long	persnr, hhnr	
biomarsm	Generated biographical information	long	persnr, hhnr	
biomarsy	Generated biographical information	long	persnr, hhnr	
einkalen	[deprecated] Spell data on income	long	persnr, hhnr	
lifespell	Spell Information on the Pre- and Post-Survey History of SOEP-Respondents	long	persnr, hhnr	
migspell	Migration history	long	persnr, hhnr	
pbiospe	Generated biographical information	long	persnr, hhnr	
refugspell	Migration history	long	persnr, hhnr	
sozkalen	[deprecated] Spell data on social benefits	long	hhnrakt, hhnr	

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 are 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 build 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 build on retrospective and prospective information at the time of each interview. The relationsship 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 build 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 build 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 drop-out studies to identify the whereabouts of attritors. These studies draw on official register data and allow us to determine whether a person 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 the post-survey history of all persons 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 the moves of foreign-born migrants as well as on the stays abroad of 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 answering the Biography Questionnaire. The observations start at the age of 15 and end at the current age (up to age 65). To update the ongoing occupational career 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 data set REFUGSPELL. The variables in MIGSPELL and REFUGSPELL are derived from different instruments and only partially overlap. The data structure allows the data set to be linked with MIGSPELL if desired.

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

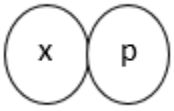
4.4 Labeling SOEP-Core

The following explanations refer to the data sets of the subdirectory “raw” in your distribution file. There is no systematic variable naming for the long files above the subdirectory “raw”.

4.4.1 Labeling Scheme of Data Sets and Variables in SOEP-Core

To distinguish the multitude of data sets and variables, the SOEP uses systematic dataset and variable names for data in cross-sectional format. These names provide a lot of information for data users. Example of a data set name:

xp



The first identifier of each data set name is the wave identifier (“x”). It can contain one or two letters. .

Each wave or survey year can be assigned using a letter in the alphabet:

1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q
2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
r	s	t	u	v	w	x	y	z	ba	bb	bc	bd	be	bf	bg	bh

As can be seen from the table, the sample data set “xp” contains survey information from the survey year 2007.

The second identifier of each data set name is the abbreviation for the respective survey instrument or, for generated data sets, the name of the content (“p”).

- h= Household
- hbrutto= Household Gross

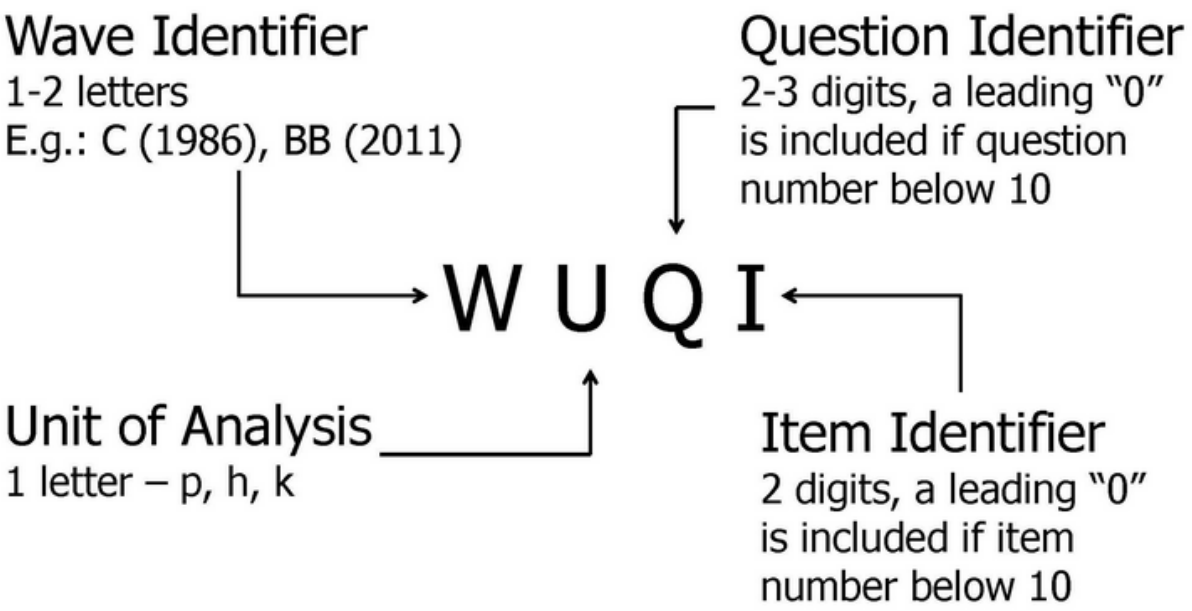
- hgen= Generated Household Data
- p= Individuals
- pbrutto= Person Gross
- p_mig= Migrants
- pgen= Generated individual data
- jugend = Youth (Ages 16-17)
- school= Pupils (Ages 11-12)
- vp= Deceased persons
- luecke= Gap Questionnaire
- hkind= Information for children from household questionnaire
- pequiv= Cross National Equivalent File
- pkal= Calendar

Further examples:

- bah = Wave „ba“ (Survey year 2010), Household data sets
- bfschool= Wave „bf“ (Survey year 2015), Pupils data sets
- zhgen = Wave „z“(Survey year 2009), Generated Household data sets

Variable names in the SOEPcore data files follow basic conventions: First, there are datasets with “speaking” variable names, where the variable name itself conveys something about the information stored in this variable. This is usally the case when the dataset is generated.

For the original datasets such as $H, $P and $KIND, the variable names are set up “around” the unit of analysis (individual - “p”, household - “h”, and child - “k”) and show before this indicator the wave in which the data was collected and after it the reference where the question can be found in the original survey instrument (see Figure 9 for an overview).



Example for a variable name: bfp0103

The first identifier of a variable name is the wave (i.e. „bf“) Every wave or rather every year can be assigned to a specific letter in the alphabet:

1984	1985	1986	1987	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q
2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
r	s	t	u	v	w	x	y	z	ba	bb	bc	bd	be	bf	bg	bh

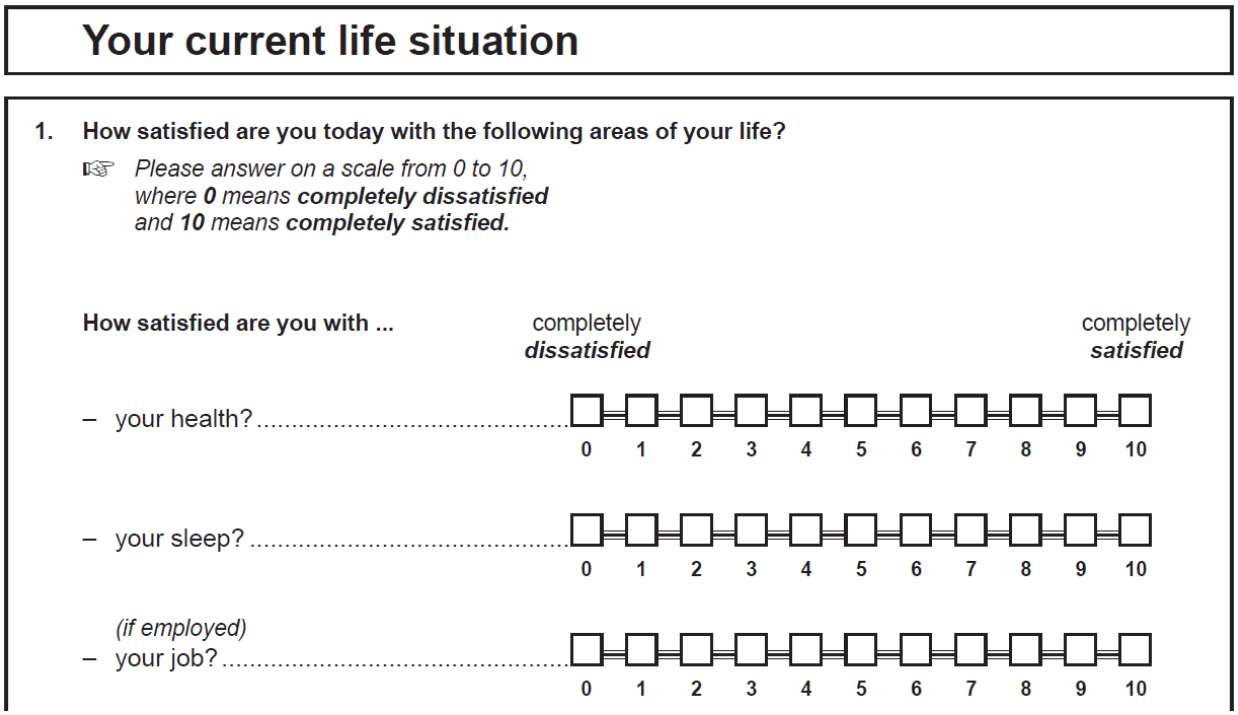
As can be seen from the table, the variable „bfp0103” contains information from the survey year 2015.

The second identifier of a variable is the abbreviation for the respective survey instrument or the type of information („p“)

- h= Household
- hbrutto= Household gross
- hgen= Generated household data
- p=Individual data
- pbrutto= Person gross
- p_mig= Person migrants (M1 und M2)
- pgen= Generated individual data

- jugend = Youth (Ages 16-17)
- school= Pupils (Ages 11-12)
- vp= Deceased people
- luecke= Gap Questionnaire
- hkind= Children information from the household questionnaire
- pequiv= Cross National Equivalent File
- pkal= Calender

The third identifier of a variable name describes the question number („01“) and a possible fourth identifier describes the position of the answer category („03“).

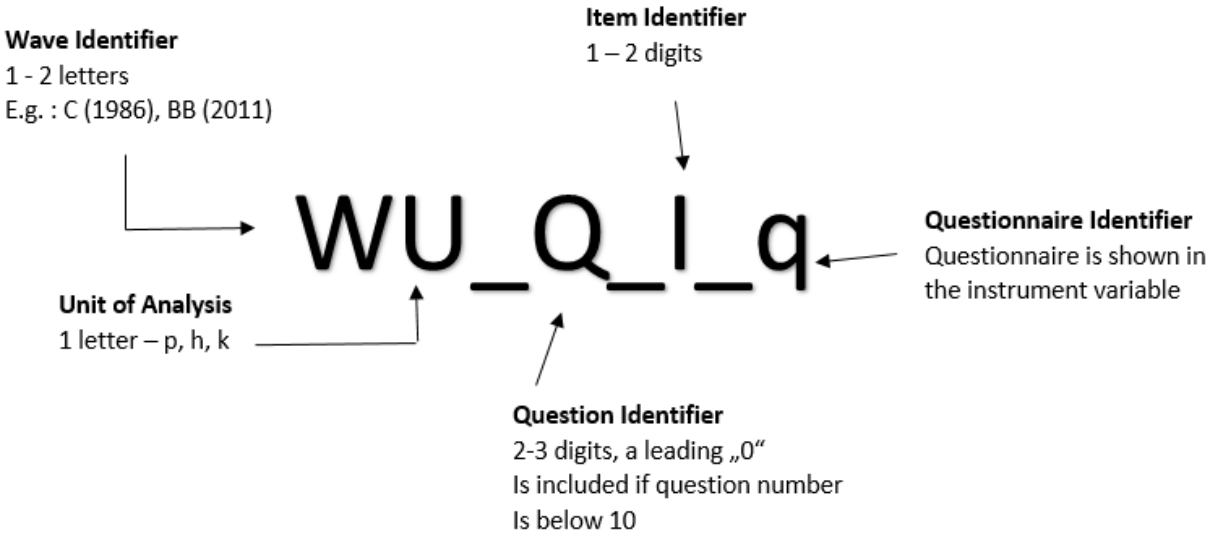


The example variable „bfp0103“ describes the „satisfaction of work“. The variable was raised in 2015 („bf“) and it can be found in the individual questionnaire („p“). In the associated individual questionnaire, the variable can be found in the first question („01“) under the third position of all answers categories („03“).

More examples: - ap06 = Wave „a“ (survey year 1984), Individual Dataset, Question 6 - th1603 = Wave „t“ (survey year 2003), Household Dataset, Question 16, Item 3 - lp10312= Wave „l“ (survey year 1995), Individual Dataset, Question 3, Item 12 - bap15604 = Wave „ba“ (survey year 2010), Individual Dataset, Question 156, Item 4

Since the data structure is getting richer every year, we extended the common variable naming convention WUQI, starting with the wave „bh“(2017). Additionally, we provide our users with an „instrument“ variable that contains all our survey instruments for each analyzing unit.

4.4.2 Extended Variable Naming Conventions



We added an underscore between question identifier and item identifier to separate question and item visually. In addition, a questionnaire identifier was introduced, which is also separated by an underscore from the item. This new version of naming variables only comes to use, if the survey instrument differs from the „original“ instrument.

Due to our different samples in the SOEP, there are some samples groups that are getting sample specific questions, like the migrant sample that started in 2013. For that specific group, we created an extended individual questionnaire, with migrant specific question and standard SOEP questions that are asked every year. For the specific questions, you can use the instrument variable to see the variables‘ source.

Let‘s take a look at the variable bhp109_01_q57

- bh= Year 2017
- P= Person questionnaire
- 109= Question 109
- _01= First Item
- _q57= ?

To know which questionnaire is the right one, you have to take a look at the instrument variable.

Value	Questionnaire
50	2017 Individual Questionnaire (A-L1 ; PAPI) [soep-core-2017-pe]
51	2017 Individual Questionnaire (A-L3 ; CAPI) [soep-core-2017-pe2]
52	2017 Individual Questionnaire (L2-L3 ; CAWI) [soep-core-2017-pe3]
53	2017 Individual Questionnaire (N; CAPI) [soep-core-2017-pe4]
54	2017 Individual Questionnaire (M1-M2 Re-Surveyed; CAPI) [soep-core-2017-p-m12]
55	2017 Questionnaire Individual-Biography (M1-M2 First-Surveyed; CAPI) [soep-core-2017-pb-m12-erst]
56	2017 Questionnaire Individual-Biography (M3-M5 First-Surveyed; CAPI) [soep-core-2017-pb-m345-erst]
57	2017 Questionnaire Individual-Biography (M3-M4 Re-Surveyed; CAPI) [soep-core-2017-pb-m34-wieder]
58	2017 Biography Questionnaire (A-L1 First-Surveyed; PAPI) [soep-core-2017-ll]
59	2017 Biography Questionnaire (A-L3; N First-Surveyed; CAPI) [soep-core-2017-ll2]

The instrument variable for identifying the exact questionnaire can be found in the respective data set. The value Q57 of the example identifies the individual biography questionnaire for re-surveyed respondents of the samples M3/M4 as the variable source. If you are now interested in the direct question in the questionnaire, open the individual biography questionnaire for refugees (Re-Surveyed), look for question number 109 and look at the first item. The variable bhp109_01_q57 was raised with the following question:

Q109: **When was the beginning of the integration course?**
- 1 Year
- 2 Month
- 99 No Details

Using the variable name and the instrument variable, you can easily identify the corresponding question in the corresponding questionnaire:

- bhp109_01_q57
- bh= Year 2017
- P= Individual questionnaire
- 109= Question 109
- _01= First Item
- _q57= 2017 Questionnaire Individual-Biography (M3-M4 Re-Surveyed; CAPI) [soep-core-2017-pb-m34-wieder]

4.4.3 Missing Conventions

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

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 for datasets in long format.*

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

CHAPTER

FIVE

WORKING WITH SOEP DATA

5.1 Working with Tracking Data (PPFAD)

For all years since 1984, the PPFAD data set contains information on all persons who have ever lived in a SOEP household at a survey time (i.e. all respondents, but also children under 17 years of age and persons who have never given an interview). PPFAD is important for the distinction of the research units (persons), especially for longitudinal analyses. In addition, paneldata.org uses PPFAD to differentiate the study population.

Time constant information of persons:

- Never changing Person ID (adults, adolescents, children)
- Original Household Number
- Gender, year of birth, month of birth, year of death if applicable
- Migrant Background
- Sample Membership (psample)

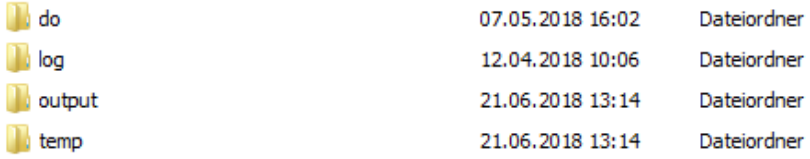
Time-varying information from people:

- 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 households)
- Survey Region (East or West Germany)

The data set is explained in more detail in a documentation:

Dokumentation PPFAD:

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 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 ordered data.

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

1. Look at the two people with the person ID (variable persnr) 2102 and 19202

a) What gender are they? When were they born and possibly died?

Open the PPFAD dataset. Search the data set for variables that describe gender, year of birth and year of death. Display the information of the variables for persons 2102 and 19202.

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

* a) What gender are they? When were they born and eventually died?
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 data set, search for a variable that describes the migration background. Display the information of the variable for persons 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 have immigrated: In which year and from which country?

Search the data set for a variable that describes the country of birth and the year of moving to Germany. Display the information of the variables for persons 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 data set for a variable that describes east-west affiliation. Display the information of the variables for persons 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) From which sources does the information on the migration background and the year of death come?

Search the data set for info variables that show you sources of information for the year of death and the migration background. Display the information of the variables for persons 2102 and 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
```

```
. *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 realised private household in 2016 and answered the individual questionnaire?

Remember that the wave-specific survey year in SOEP is abbreviated with letters. SOEP started in 1984 (wave a) and was in a survey wave "bg" in 2016. For more information on this topic, please refer to the DTC subchapter *Labeling SOEP-Core*.

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 data set for variables with the abbreviation "bg" that describe the population. Display the characteristics of the population variables:

```
*******************************************************************************
*** Exercise 2) ***
* How many people lived in a realised private household in 2016 and answered the
* personal 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 realized households. Search the data set 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-tab 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. PPFAD 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 in 1984 (wave a) and was in a survey wave "bg" in 2016. For more information on the subject, see the subchapter *Labeling SOEP-Core*. 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 data set 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 personal 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 7639 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 persons who have answered 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) of 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 person 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 6665 people completed the individual questionnaire every year over the period 2000-2014.

c) How many people who turned 15 in 2011 and lived as children in a survey 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 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 from the net code of 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 and were part of a survey household as a child in 2011, completed a 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 has changed households twice during this time. In which years?

To identify how often and when a person has changed the household, you must display all available household numbers in ppfad 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

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

5.2 Generating a cross-section Data Set

This example involves generating a data set 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-section dataset for the year 2008, which should contain all persons with the following characteristics:

- Respondents in 2008 **"ynetto"**
- Lives 2008 in private household **"ypop"**

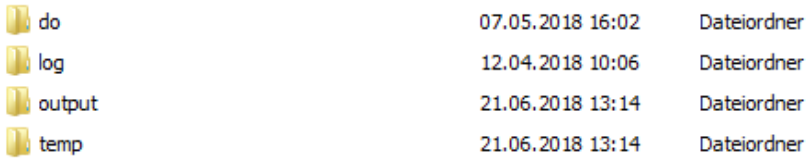
The data set should contain the following variables of interest.

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

In addition, the data set 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, data sets and temporary data sets. Open an empty do file and define your created paths with globals:

```
1 ********************************************
2 * Set relative paths to the working directory
3 ********************************************
4 global AVZ         "H:\material\exercises"
5 global MY_IN_PATH "\\hume\rdc-prod\complete\soep-core\soep.v33.2\stata_en\"
6 global MY_DO_FILES "$AVZ\do\"
7 global MY_LOG_OUT "$AVZ\log\"
8 global MY_OUT_DATA "$AVZ\output\"
9 global MY_OUT_TEMP "$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.

Use ppfad as the source file together with the required variables. Keep all cases with completed interviews. In addition, your data set 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 data set.

```
* * * 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
```

Save the modified data record temporarily. Now link your data set with the weights of the SOEP and save your data set 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 data records and temporarily save the created data records.

```
* * * READ DATA * * *

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

(continues on next page)

(continued from previous page)

```
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 your created data sets to your masterfile and save your analysis data set.

```
* * * 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 data set for the year 2008.

2. Encode missing values into missing values in system failings (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 chapter *Missing Conventions*. For content analyses 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


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

```

(continues on next page)

(continued from previous page)

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

Open your analysis data set and summarize all missing codes.

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

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

```
        *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 your analysis data set, 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 your analysis data set, 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 the 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 the income for the 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 the 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
-----------------------------
```

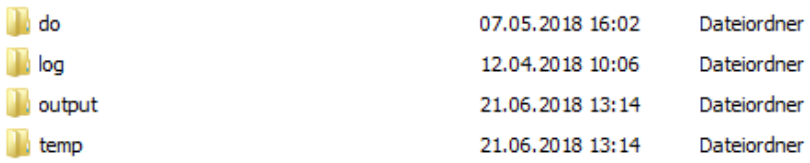
5.3 Working with Migration Data (BIOIMMIG)

With its migration and refugee samples, SOEP provides a broad spectrum of information on persons with a refugee and migration background.

In the BIOIMMIG data set you will find relevant information on the history of flight and migration, such as motives for fleeing and migration, the circumstances after arrival in Germany, but also information on relatives in the country of origin and the desire to return to the country of origin in edited form. For more information about this data set and a list of the variables it contains, see the 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 an escape background.

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, data sets and temporary data sets. 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 ordered data.

Task 1: Preparation of BIOIMMIG

a) In which variable can you find 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)        In which variable can you find 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 data set and load it from the data set, together with the person number and the survey year.

Open your data set only with the required variables to maintain clarity in your analysis data set.

```
/*
b)        Identify this variable in the BIOIMMIG data set and load it from the data set, together with the person number 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 research-relevant analysis variable and check 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 between three groups:

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

After you have familiarized yourself with the research-relevant analysis variable, recode the variable to suit 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
*/

```

(continues on next page)

(continued from previous page)

```
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 happen that initially 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.

```
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 record on your personal drive temporarily .

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

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

Aufgabe 2: Add basic variables from PPFAD and weights

a) Load the following information from PPFAD:

- Never changing Person ID **"persnr"**
- Household number **"hhnr"** and the current household number **"bghhnr"**

- The net variable with information about the interview type **"bgnetto"**
- The sex 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 PPFAD data set, visit the chapter *Working with Tracking Data (PPFAD)*.

```
/*
a)        Use the following information from PPFAD:
  - Never changing Person ID „persnr"
  - Household number "hhnr" and the current household number "bghhnr".
  - the net variable with information about the interview type "bgnetto".
  - the sex 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 record using the person number.

If you don't understand how to create your own cross-section dataset, visit the chapter *Generating a cross-section Data Set*.

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

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

c) Add the corresponding person extrapolation factors to the data record.

```
c)        Add the corresponding person extrapolation factors to the data record.
*/

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

d) Only keep respondents for whom a youth or individual questionnaire was realized in 2016.

For example, to exclude children who have not provided immigration status information, use the net code from PPFAD. Only keep persons who have conducted a completed individual or youth interview.

```
/*
d)        Only keep individuals for whom a youth or personal questionnaire was␣
↪realized 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 immigrant background
- Migration 2nd generation
- Immigration without information
- Immigration, not flight
- Immigration, Flight

To generate this status variable, check the contents of the existing migration variables from PPFAD (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 PPFAD (migback, germborn) and link this information with your previously generated escape 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 titles) are now in your record?

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 titles) are now in your record?
*/

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 home country (flight==2 and biimgrp==5).

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

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

```
/*
b)        How many are there if you take the person 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 only about 675 fugitives in the data set. The weighting thus corrected the number of fugitives downwards.

c) How many persons are represented by 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 by 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 fugitives?

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 *Labeling SOEP-Core*. 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 fugitives over 40 years of age in weighted form:

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

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		Total
	0	1	
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%.

5.4 Generating a longitudinal Data Set

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

In the previous examples you have already 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 data set should contain all respondents in private households:

The data set should contain the following variables of interest:

- Health satisfaction **"wp0101" "xp0101" "yp0101"**
- Smoking at present yes/no **"wp9301" "yp10601"**
- Current employment status **"emplst06" "emplst07" "emplst08"**
- Monthly household net income **"hinc06" "hinc07" "hinc08"**

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

- Cross-sectional weighting factors for all relevant years **"wphrf" "xphrf" "yphrf"**
- Person ID **"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 data set yourself, you can do this with the data sets you have supplied. 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 record, 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 data set is created from several variables of different data sets, it is worth sorting the person number before saving the individual data sets in order to be able to merge the data sets more easily afterwards.

1.1. Create a Master-Files

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

```
1
2 use hhnr persnr sex gebjahr psample xhhnr xnetto xpop yhhnr ynetto ypop whhnr wnetto␣
  ↪wpop using "${MY_PATH_IN}ppfad.dta"
3
```

Since we want to receive an unbalanced data set, i.e. persons who have completed a personal questionnaire at least once within the 3 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 (personal number) of the data record and save it.

```

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

```

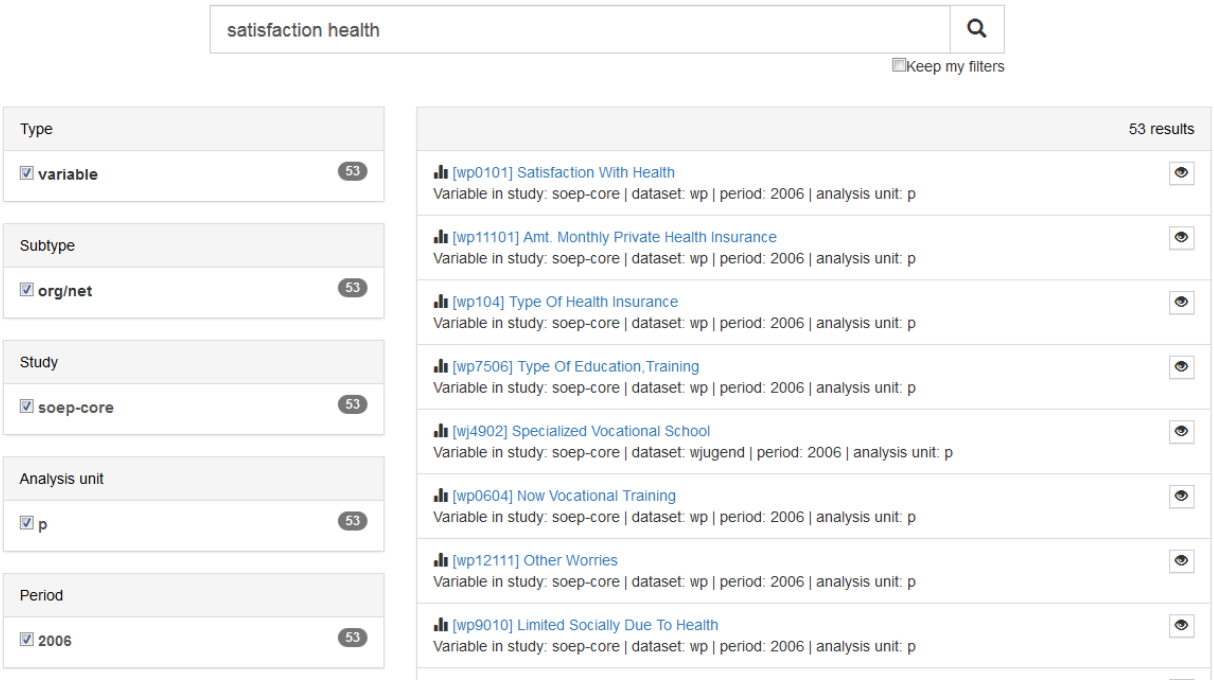
What is still missing is the cross-section weighting factor and the variables of interest in terms of content. To apply the weighting factors to the data set, open the weighting data set for the person level phrf, sort it and save it again.

```

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

```

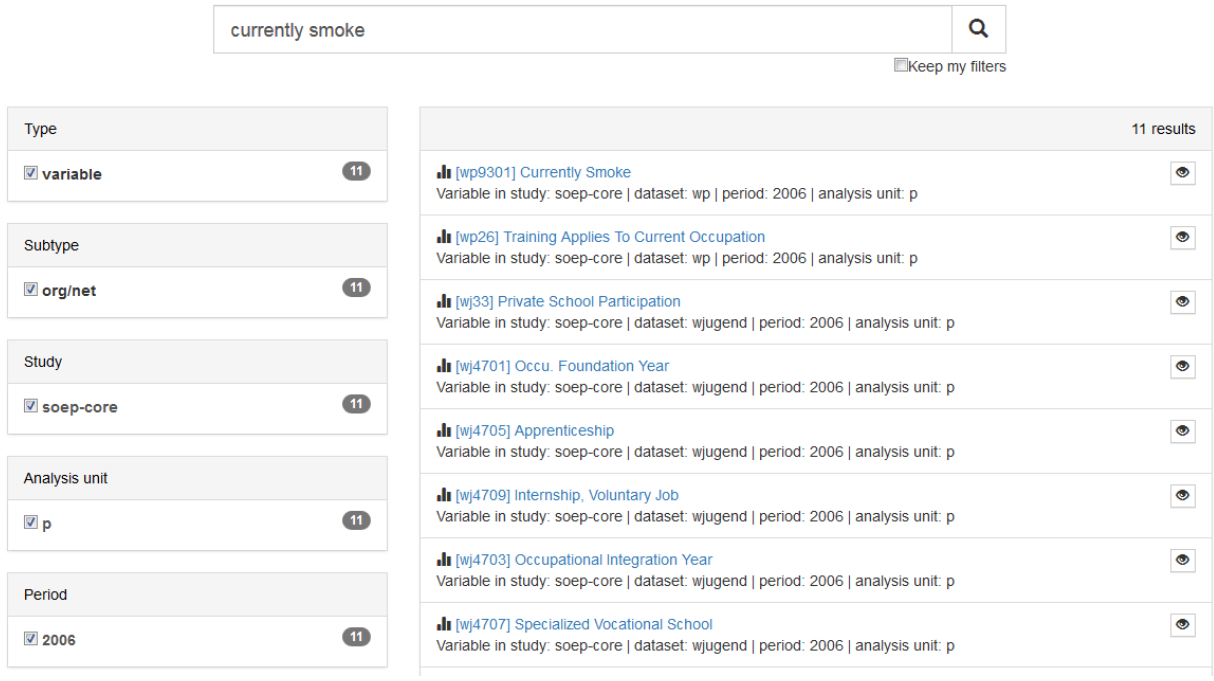
Now we come to the variables of content. In order not to have to click through all delivered data sets, it is recommended to enter the label of the variable of interest on paneldata.org.

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 in which years the variable was collected and what the variable is called.

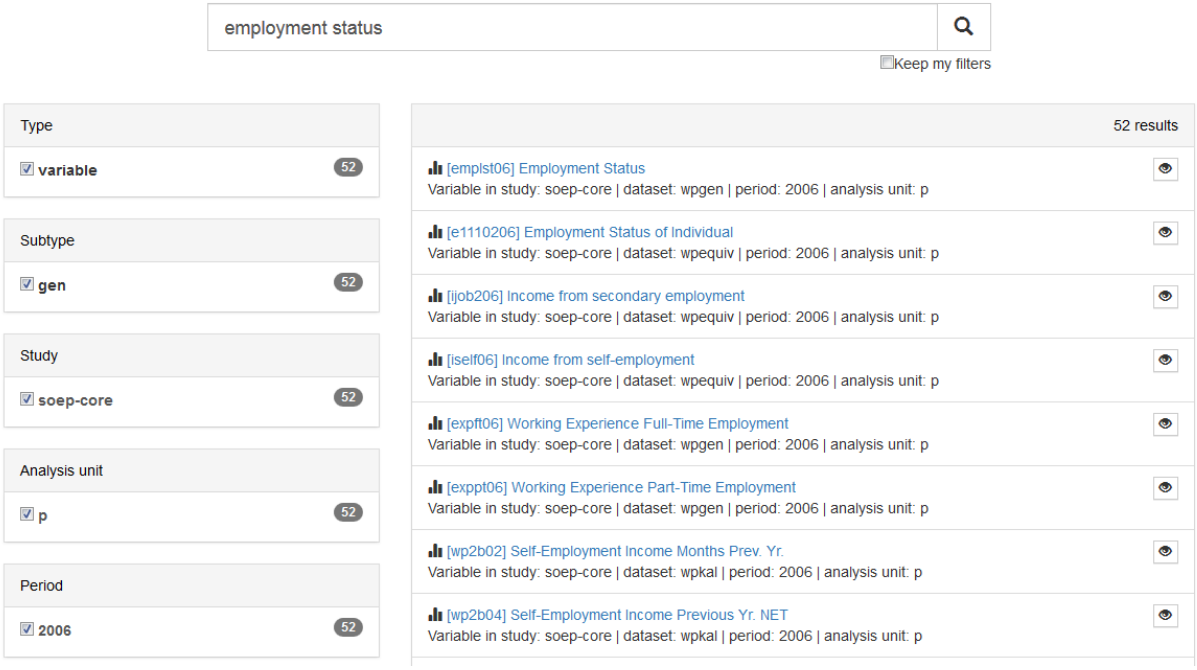
Example: Variable Label „Satisfaction Health“



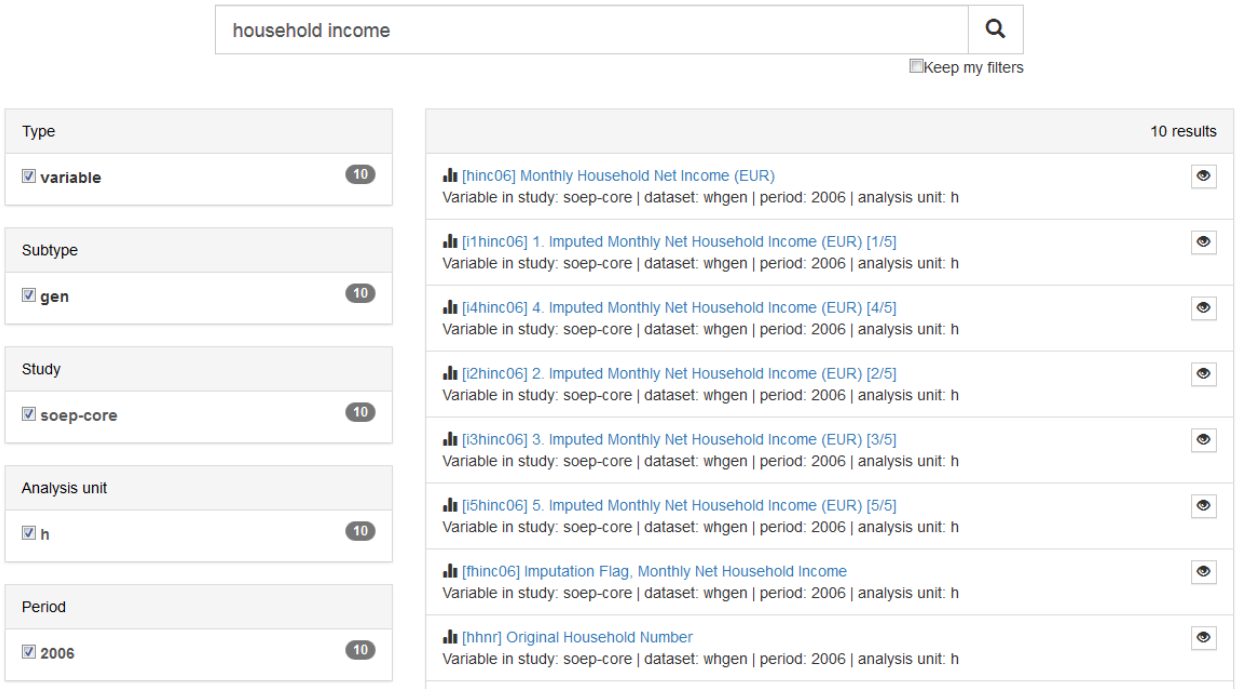
Example: Variable Label „currently smoking yes/no“



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 data records sorted by the persnr (person number) 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 indicator, which is either the household number ($hhnr) or the person number (persnr), you can now merge all data records or individual variables to ppfad. Which indicator to use and when depends on the unit of analysis. Since we are on the person level, our indicator is persnr (person ID).

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 into missings. This step is important for converting a wide-format data set to a long format.

```
*******************************************************************************
*** Task 2) ***
* Encode missing values in system failings (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 data sets into the long format. In long format, additional years are displayed as additional lines. If the data record 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-section 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 records) 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:

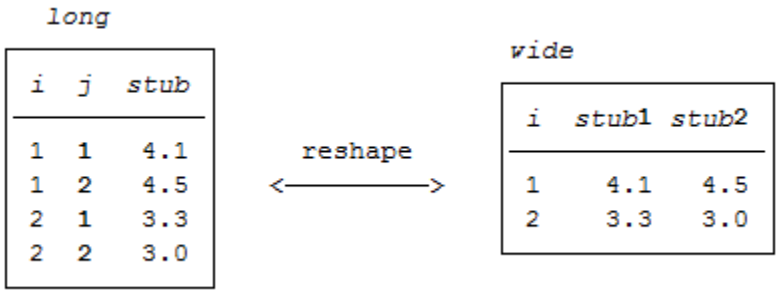
```
*rename time-variant variables
*with examples how to use loops (but can also be done "manually")
        rename wp9301 smoke2006
        rename yp10601 smoke2008
        rename wp0101 health2006
        rename xp0101 health2007
        rename yp0101 health2008
        ...
```

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 data set. If you want to convert a data set, you can do this in both directions:



In our case we reshap 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 average satisfaction with men's and women's health changed over the three years?

Satisfaction with health was measured on a scale of 10, with a value of 10 representing an extraordinarily high 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 average satisfaction with men's health between 2006 and 2008, while the last three values show average satisfaction with women's 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 only within one persnr (person ID) and the satisfaction of the following year should be analyzed.

```
*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 your analysis data set, 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 a 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 the health of people who quit smoking after 2006 changed from 2006 to 2008?

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 the health of
*   people who quit smoking after 2006 changed from 2006 to 2008?

        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)

```
label define quit 0 "smoker" 1 "quit" 2 "non-smoker" 3 "begin"
label values quit quit
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.

```
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 variable quit 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 after he/she stopped smoking in 2008 indicates a satisfaction value of 7.76. So you can assume that when a person stops smoking, the state of health that a person perceives deteriorates. Now we have to test if the assumption is correct.

d. Does quit smoking make your health worse? To what extent can 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 ttest 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
```

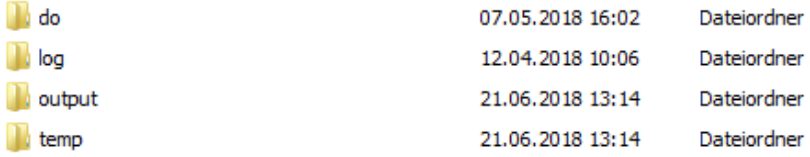
H0 Hypothesis: If one stops smoking it has no effect on health.

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 value 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.

5.5 Longitudinal Data Analysis

Simple cross section analyses show that married people have a higher life satisfaction than singles. You want to check this on the basis of longitudinal analyses 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 your created paths 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 ppfadl.

You should always add some variables from PPFADL to your data set by default. Download the following information from PPFADL:

- Person ID **"pid"**
- Household number **"pid"**
- Survey year **"syear"**
- The net variable with information on the interview type **"netto"**
- The weighting variable **"phrf"**
- The sex 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
```

Search for matching variables and add them to your data set

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 visit the chapter *Variable Search with Questionnaires*)

- Find a suitable variable via the topic list of paneldata.org (for more information visit the chapter *Topic Search with paneldata.org*)
- Search for a suitable variable using a search term in paneldata.org (for more information visit the chapter *Variable Search with paneldata.org*)
- Use the documentation provided by the generated variables (for more information visit the chapter *Documentation of 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 missings into the format of a point.

```
*-------------------------------------------------------------------------------
*** 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 personal 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 data set as a panel data set.

```
**define the data set 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  |  ..............1111111111111111
      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          |  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
```

86079 respondents have contributed information within waves a (1984) - bg (2016) and 75% of the 86079 respondents have provided information for at least 10 waves

How many people took part in the survey in 2010 and contributed to continuous measurements until 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
```

14673 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

How high is the proportion of people who will be a) married in 2014 or b) have a migration background. Compare weighted with unweighted frequency tables: Which people are 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 for Germany again.

```
-> 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 report an life satisfaction (scale value 7) in a survey year also indicate the scale value 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. People who were completely dissatisfied (value: 0) in the base year remain completely dissatisfied with around 20 % in the following year. About 80% of these dissatisfied people from the base year improve their life satisfaction in the following year. Of the completely satisfied persons (value: 10), about 37% remain just as satisfied in the following year. For 63%, however, life satisfaction worsens. It is more likely that a completely dissatisfied person (value: 0) will become more satisfied in the following year.

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 declared a divorce as family status in the following year can be observed particularly frequently. (About 19%).

Simple cross sectional analyses

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

First, calculate the correlation between family status and life satisfaction in cross section 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 4834 people can be observed changing 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, i.e. 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, the average life satisfaction rises slightly by 0.11.

Map the complete satisfaction history around the "marriage entry" event [3 years before; 3 years after].

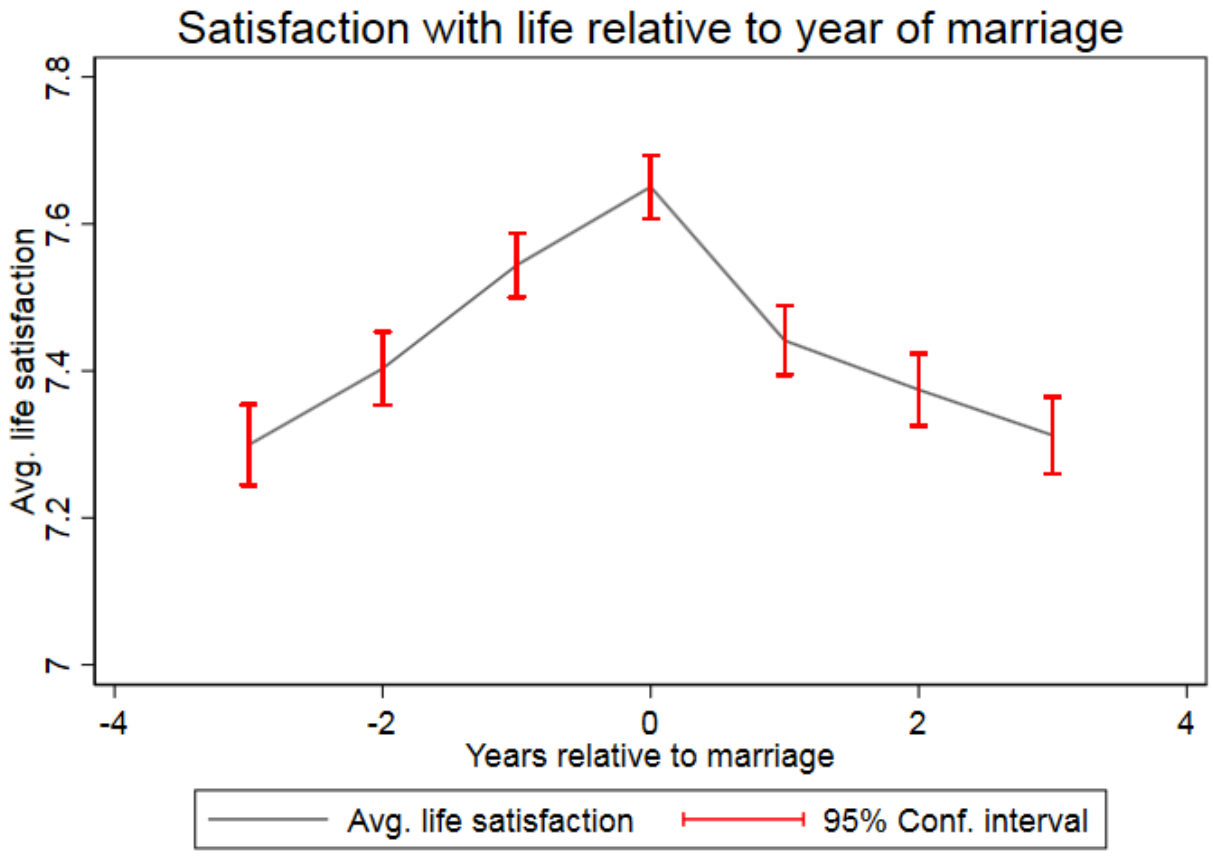
```
**preparing illustration of trajectory
generate t=0 if to_mar==1 & l.to_mar~=1 &l2.to_mar~=1 & l3.to_mar~=1 & l4.to_mar~=1 &␣
↪l5.to_mar~=1 & l6.to_mar~=1 & l7.to_mar~=1 & l8.to_mar~=1 & l9.to_mar~=1 & l10.to_
↪mar~=1 & l11.to_mar~=1 & l12.to_mar~=1 & l13.to_mar~=1 & l14.to_mar~=1
replace t=1 if 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")
```

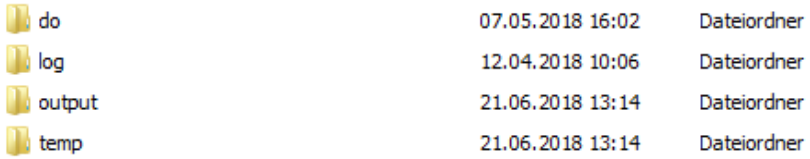


The graph shows that a positive effect on life satisfaction can be observed when the family status changes from single to married. In the following years of the existing marriage, life satisfaction decreases again and approaches the initial satisfaction before the marriage.

5.6 Fixed Effects Estimation

You want to find out whether certain variables relevant to the labour market, such as work experience or education time, influence a person's hourly wage. Other variables such as gender or marriage status should also be taken into account. You decide to use the SOEP data to set up a fixed effects estimation model.

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 created 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 ordered data.

a) Generate your own SOEPWage.dta data set. The data set 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 visit the chapter *Variable Search with Questionnaires*)
- Find a suitable variable via the topic list of paneldata.org (for more information visit the chapter *Topic Search with paneldata.org*)
- Search for a suitable variable using a search term in paneldata.org (for more information visit the chapter *Variable Search with paneldata.org*)
- Use the documentation provided by the generated variables (for more information visit the chapter *Documentation of Generated Data*)

Use the various important variables of the ppfadl.dta data set as your start file. Your source file should contain the following variables:

- Person ID **"pid"**
- Survey year **"syear"**
- Birth Year **"gebjahr"**
- The net variable with information on the interview type **"netto"**
- The weighting variable **"phrf"**
- The sex of the person **"sex"**
- Sample Membership **"pop"**

```
use pid syear sex gebjahr netto pop phrf using "${MY_IN_PATH}/ppfadl.dta", clear
```

Apply the necessary content variables to your starting data set. You need the following variables for your analysis:

- Employment Status **"plb0022"**
- Current Gross Labor Income in Euro **"pglabgro"**
- Actual Work Time Per Week **"pgtatzeit"**

- Working Experience Full-Time Employment **"pgexpft"**
- Amount Of Education Or Training In Years **"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 in your analysis, remove all survey information that does not fall within this period. To finish, save your data set.

```
* Period from 2012 to 2016
keep if syear>=2012 & syear<=2016
```

Exercise 1: Prepare your data set

a) Load your created SOEPWage.dta data set. The data set contains information on gross monthly wage, marital status and other personal characteristics.

```
*** Exercise 1: Prepare your data set
* a) Load data set
use "${MY_OUT_DATA}/SOEPWage.dta", clear
```

b) Recode all missing values in Stata Missings (.)

```
* b) Recode Missings
mvdecode _all, mv(-8/-1 = .)
```

For more information about the missing codes of SOEP data visit 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 1st 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)
```

(continues on next page)

(continued from previous page)

```
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 data set as a panel data set.

```
*** Exercise 2: Descriptive statistics
* a)
xtset pid syear // Declaring data as panel data
```

b) What percentage of people participate 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. At least once 19717 people within the period from 2012 to 2016 have stated not to have been married. 25014 persons reported to have been married at least once during this period. Those who were not married for at least one year responded with "married==no" in 94.69% of the observations. Whereas those who have been married at least once responded in 95.88 percent of the observations with"Married==Yes". A very stable response behaviour can therefore be observed.

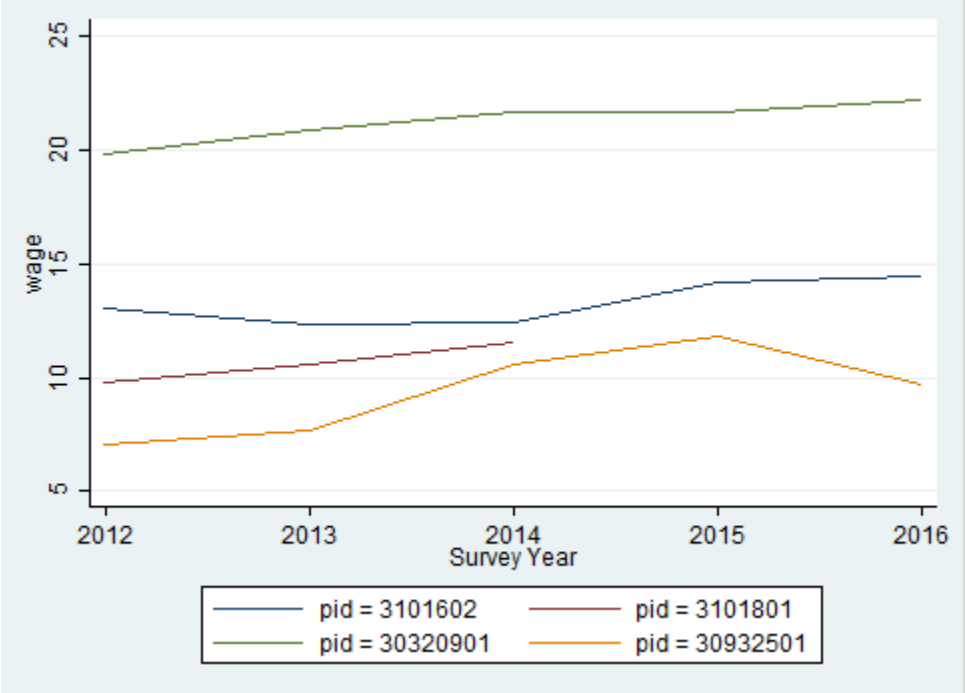
```
* Transition probabilities
xttrans married, freq
```

```
. 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 also not yet 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. A stable behaviour of the respondents can be seen.

```
* 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
				F(4, 78229) =	8027.72
Model	9531.59732	4	2382.89933	Prob > F =	0.0000
Residual	23221.0303	78229	.296834042	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 work more frequently 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 is no 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 sex because sex 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 with 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 with the effect of "labour 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) labour market experience squared.

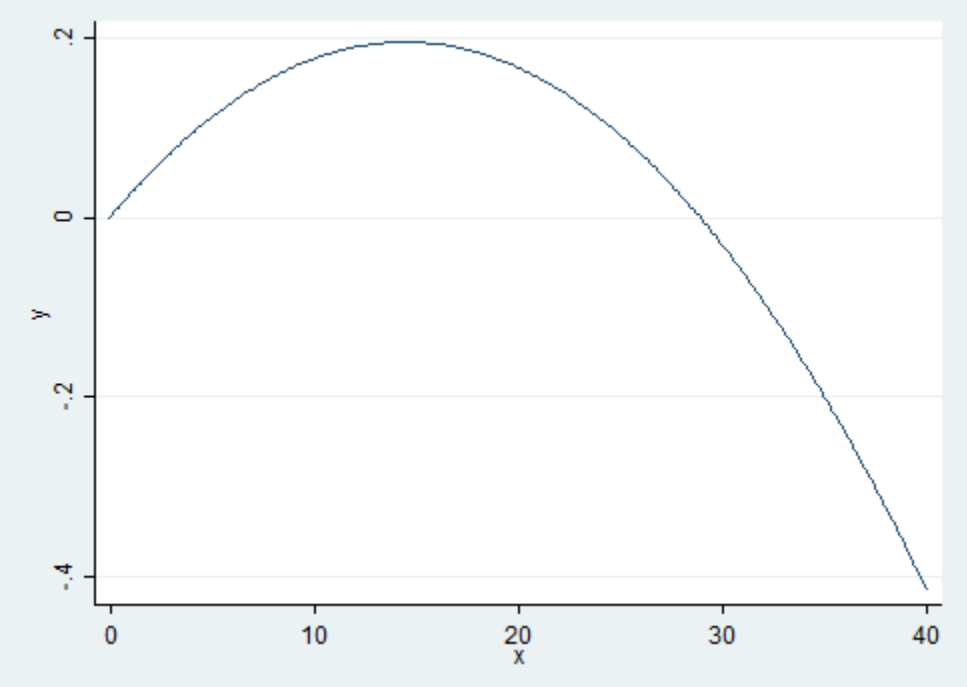
e) Now you can also square labour market experience into the model. To what extent does the effect of labour market experience change compared to task 5d)?

```
* 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)
------------------------------------------------------------------------------
             |               Robust
      lwage |      Coef.   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 labour 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 data set from the RAW subdirectory. Apply the required variables on your analysis data set and generate your period-related longitudinal section weight. To understand the structure of the data distribution file and the location of the different data sets, visit the chapter *Data Sets SOEP-Core*. For more information about the weighting data sets and other survey data sets, visit the chapter *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 stronger 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 return for the additional education they expected on the local labour market and may therefore move -> higher probability for dropout.

5.7 Working with SOEP Regional Data

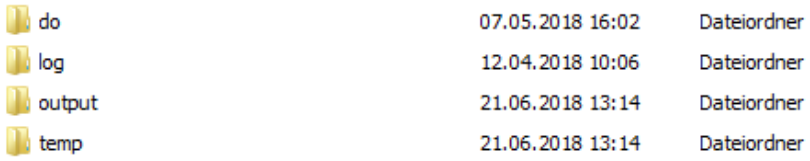
SOEP offers diverse possibilities for regional and spatial analysis. With the anonymized regional information on the residences of SOEP respondents (households and individuals), it is possible to link numerous regional indicators on the levels of the states (Bundesländer), spatial planning regions, districts, and postal codes with the SOEP data on these households. However, specific security provisions must be observed due to the sensitivity of the data under data protection law. Accordingly, you are not allowed to make statements on, e.g., place of residence or administrative district in your analyses, but the data does provide valuable background information.

Level	Available Since	Data Access	Data Protection
States (Bundesländer)	1984	Standard SOEP dataset (Scientific Use File)	Data distribution contract
Municipal size classes (e.g., Boustedt)	1984	Standard SOEP dataset with special password	Expanded data distribution contract on the use of municipal size classes & data protection concept
Spatial planning regions (geocodes)	1985	Standard SOEP dataset plus SOEP geocode disk	Expanded data distribution contract on the use of geocodes & expanded data protection concept
Official county codes (KKZ)	1985	SOEPremote (online access to county-level regional data) or at the SOEP Research Data Center at DIW Berlin	Expanded data distribution contract on the use of SOEPremote & SOEPremote access form
Official municipality key, postal codes, Microm neighborhood data	2000 1993 2000	Use of data only at the SOEP Research Data Center at DIW Berlin	Only by personal arrangements in the framework of our SOEP in residence program

For more Information and to get access visit **Regional Data**

For your research project you want to measure current (year 2016) urban-rural differences in the population. You are particularly interested in the differences in political interest and the different satisfaction variables provided by the SOEP. You also want to take into account demographic differences in gender and age. In order to be able to evaluate the research potential, you should get an overview. For regional analyses, for example, the community 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 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 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 your ordered data.

a) Prepare a cross-sectional analysis data set 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 visit the chapter *Variable Search with Questionnaires*)
- Find a suitable variable via the topic list of paneldata.org (for more information visit the chapter *Topic Search with paneldata.org*)
- Search for a suitable variable using a search term in paneldata.org (for more information visit the chapter *Variable Search with paneldata.org*)
- Use the documentation provided by the generated variables (for more information visit the chapter *Documentation of Generated Data*)

Your source file should contain the following variables:

- Never Changing Person 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"**
- Zufriedenheit mit Demokratie **"bgp0112"**
- Political Interests **"bgp143"**
- Current Sample Region **"bgsampreg"**
- Federal State **"bgbula"**
- Spatial category by BBSR **"bgregtyp"**

- Community Class Sizes “ggk”

Use the various important variables of the ppfad.dta data set as your start file.

```
use hhnr persnr bghhnr sex gebjahr bgnetto bgpop using ${MY_IN_PATH}\ppfad.dta, clear
```

Keep people who completed a questionnaire in 2016 and live 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
tempfile ppfad
save `ppfad'
```

Prepare the different data sets bgp, bghbrutto, regiobl

```
* Prepare data set bgp
use ${MY_IN_PATH}\bgp.dta, replace
keep persnr hhnr bghhnr bgp01* bgp143
tempfile bgp
save `bgp'

* Prepare data set bghbrutto
use ${MY_IN_PATH}\bghbrutto.dta, replace
keep hhnr bghhnr bgsampreg bgbula bgregtyp
tempfile bghbrutto
save `bghbrutto'

* Prepare data set 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 data sets.

```
* Merge all data sets
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 into missings.

```
* Recode negative values into missings
mvdecode sex gebjahr bgp01* bgp143,mv(-5/-1)
```

Categorize the community class sizes of the SOEP regional data set.

```
* 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
```

(continues on next page)

(continued from previous page)

```
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
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 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"
```

(continues on next page)

(continued from previous page)

```
lab val bgbula_cat bgbula_cat
drop bgbula
rename bgbula_cat bgbula
```

Put the variables in your preferred order and save your data set.

```
* 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 in Germany. 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 community size classes. Create a table showing you satisfaction with different aspects of life and revealing differences by gender, age, community size class and federal state.

```
foreach x of local varlist {
* Tabulation of satisfaction by size of community 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 size of community
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 smallest regional levels.

d) Create a table that shows you the political interest differentiated by age, gender and community size class for Bavaria

```
1  ********************************************************************************
2  capture log close
3  log using "${MY_LOG_OUT}\political_interest.log", replace
4
5  * Political interest
6  * Tabulation of political interest by size of community for Bavaria
7  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

It becomes clear that the SOEP offers a wide range of possibilities for region-related analyses. It is possible to allocate a multitude of regional indicators at the level of the federal states, the regional planning regions, the districts and the postal codes.

CHAPTER

SIX

WORKING WITH SOEP DOCUMENTATION

6.1 Variable Search with Questionnaires

If you come across a variable in the data set whose variable content is unclear, you should always check whether there is a suitable questionnaire for the data set. Under *Overview Data Sets* you can see whether the data sets correspond to a survey instrument. The related questionnaires can be found here:

Example: During your research project you come across the variable bbh5508 with the German label "Auto: Gründe" (Car: Reasons) and the Englisch 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	

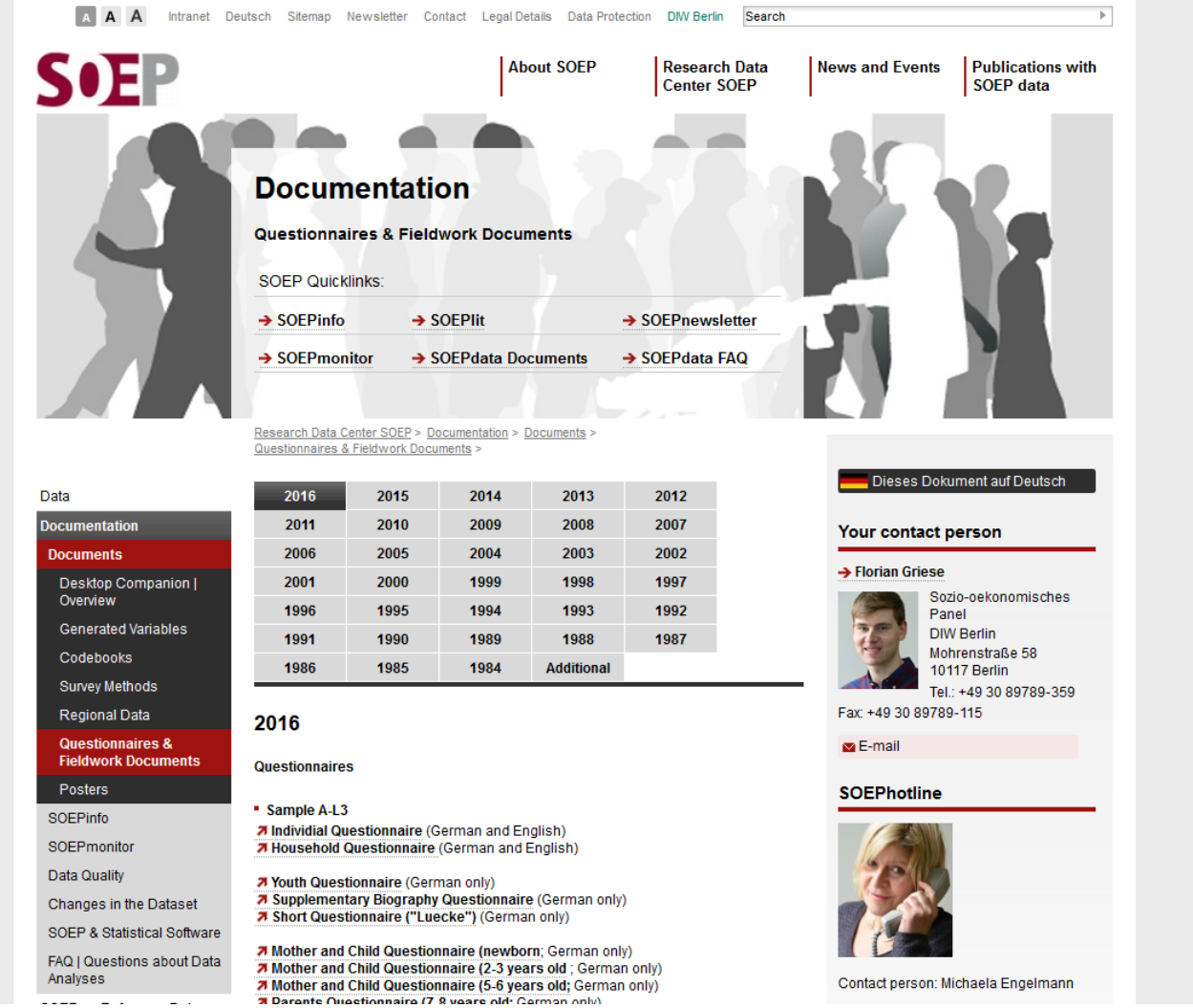
Unfortunately, it is difficult to determine the variables content from the output and also from the label designations. To understand the complete question and also possible filter instructions, you should use the questionnaires.

Example Variable:

bbh5508: Wave „bb" (Survey Year 2011); household questionnaire („h"), question number 55, item 8

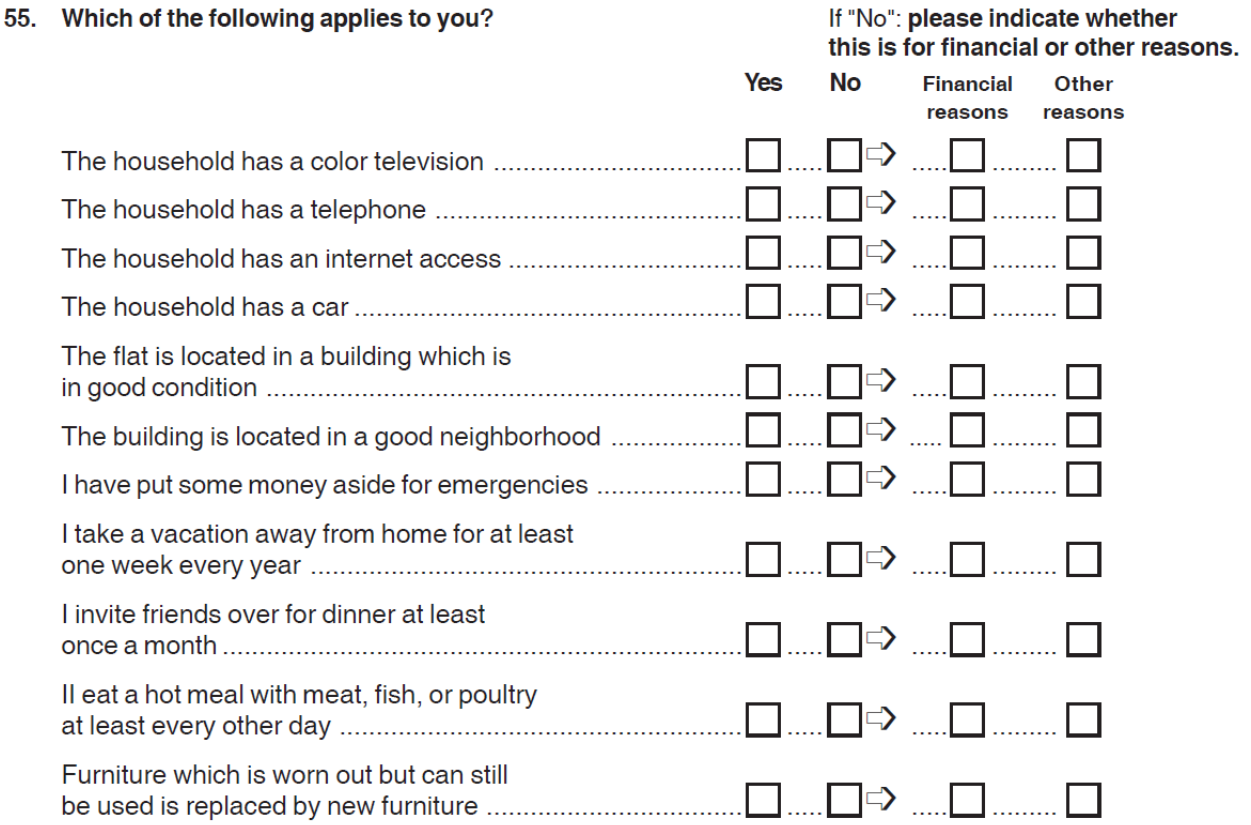
Open

The variable "bbh5508" can be found in the questionnaires for 2011. Select the survey year 2011 and download the household questionnaire.



Search the variable “bbh5508” in the

Since you are already in the correct questionnaire, you must now search for question 55.



55. **Which of the following applies to you?**

	Yes	No	If "No": please indicate whether this is for financial or other reasons. Financial reasons	Other reasons
The household has a color television	☐	☐ ⇨	☐	☐
The household has a telephone	☐	☐ ⇨	☐	☐
The household has an internet access	☐	☐ ⇨	☐	☐
The household has a car	☐	☐ ⇨	☐	☐
The flat is located in a building which is in good condition	☐	☐ ⇨	☐	☐
The building is located in a good neighborhood	☐	☐ ⇨	☐	☐
I have put some money aside for emergencies	☐	☐ ⇨	☐	☐
I take a vacation away from home for at least one week every year	☐	☐ ⇨	☐	☐
I invite friends over for dinner at least once a month	☐	☐ ⇨	☐	☐
II eat a hot meal with meat, fish, or poultry at least every other day	☐	☐ ⇨	☐	☐
Furniture which is worn out but can still be used is replaced by new furniture	☐	☐ ⇨	☐	☐

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 house 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.

6.2 Variable Search with paneldata.org

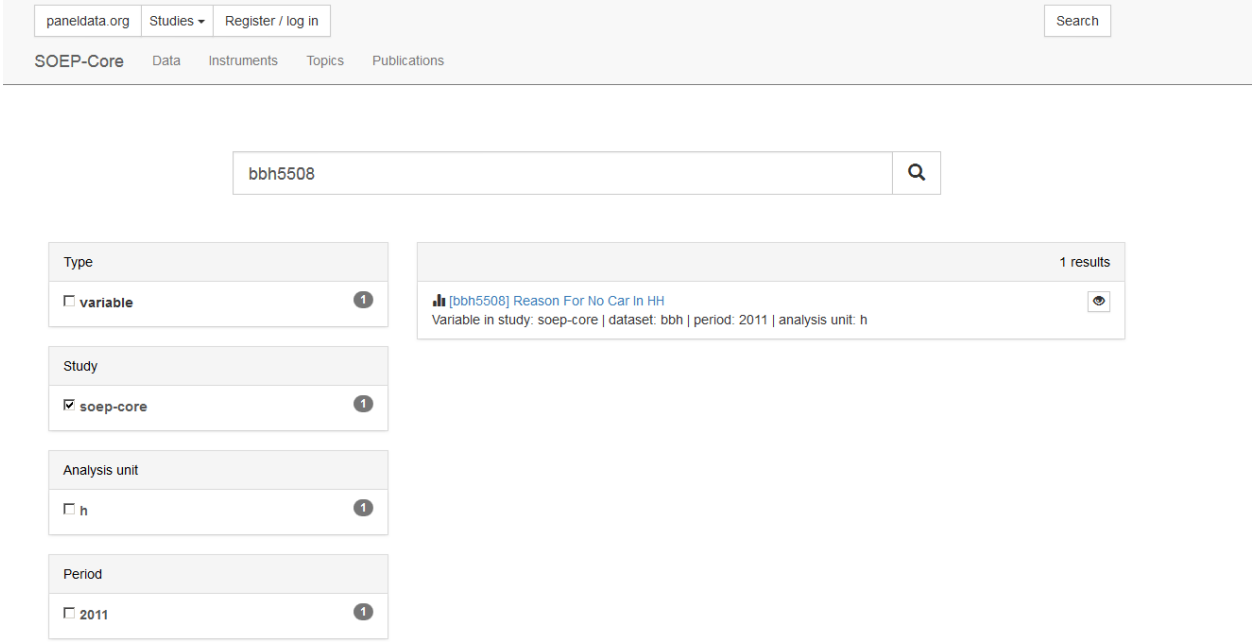
With paneldata.org it is also possible to search for variables. For example, if you want to find more information about generated variables, a search with paneldata.org is indispensable. For example, the platform offers comprehensive frequency counts, the chronology of the variables searched for, a cross-study variable linkage via concepts, a syntax generator and a topic list for content search in the SOEP.

Example Variable:

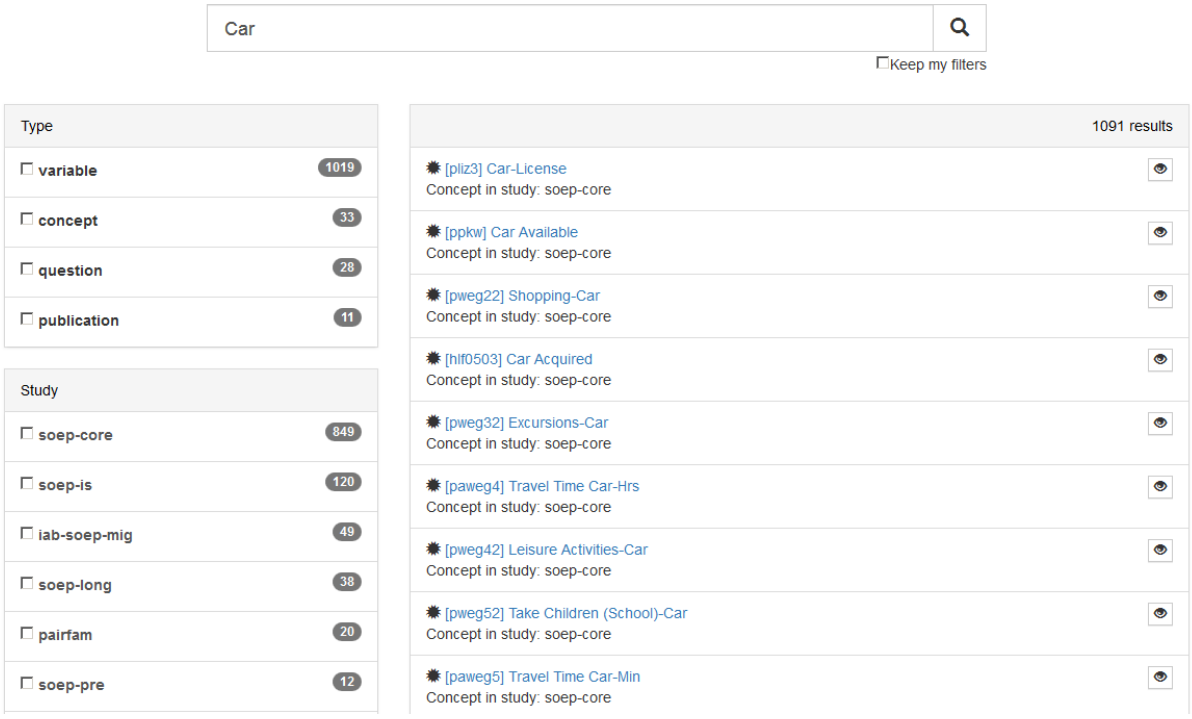
bbh5508: Wave „bb“ (Survey Year 2011); household questionnaire („h“), question number 55, item 8

Open

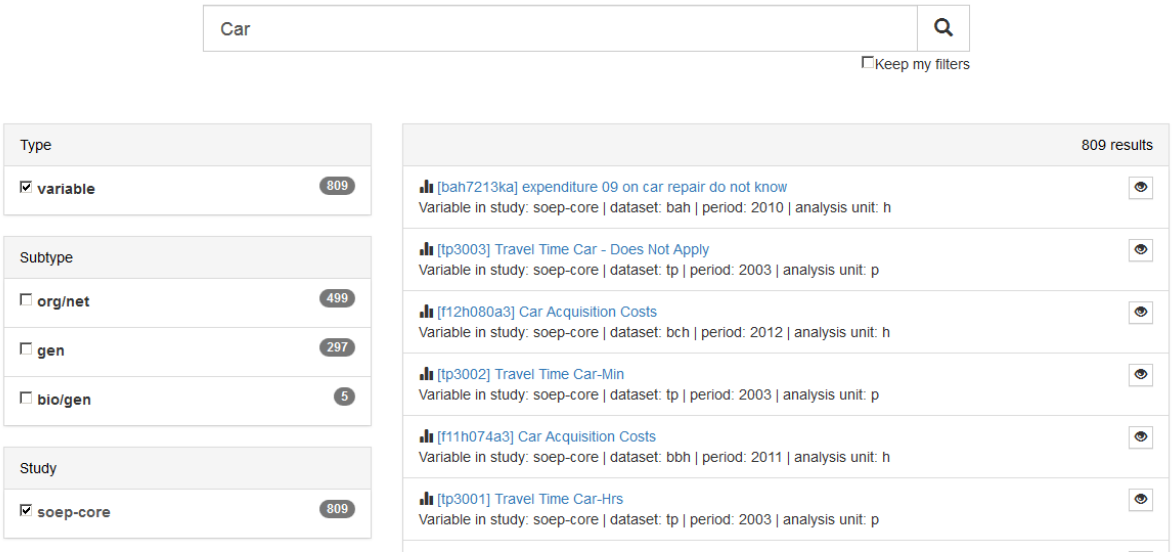
Please select the study SOEP-Core. The SOEP-Core overview contains important general information about the study, e.g. data access, survey method, questionnaires, thematic diversity, terms for missing codes, all available data sets of the study and metadata-based questionnaires. To search for a variable, a data set or a publication, simply enter the desired search term in the search field.



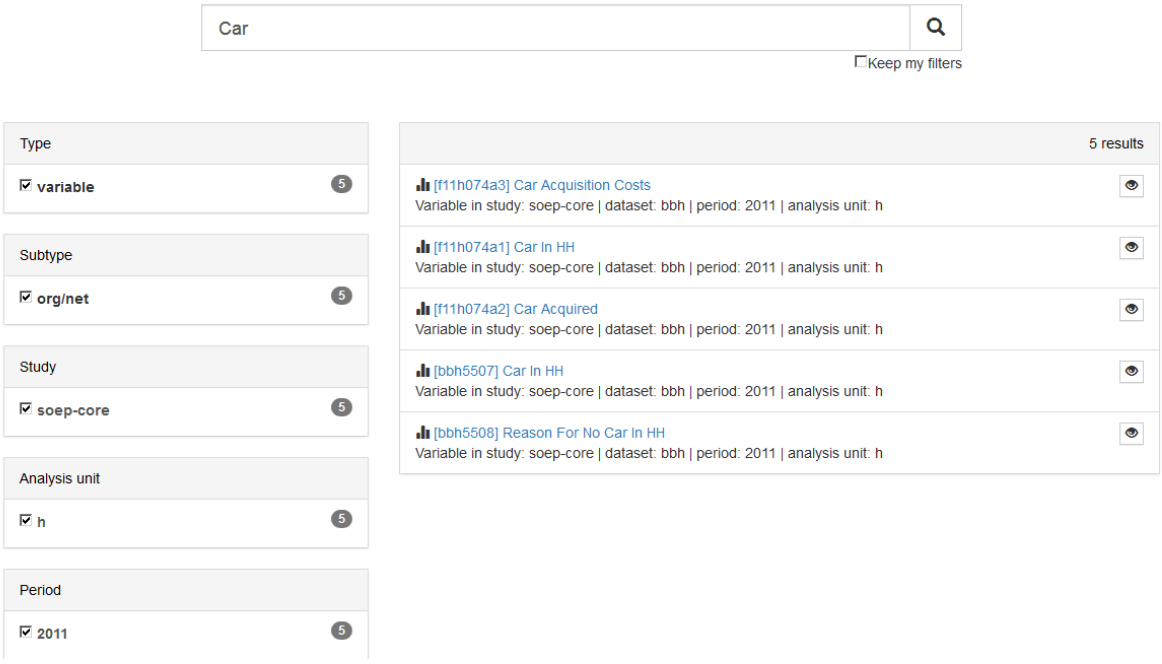
In order for the search to be successful, specific information from the user are necessary. The results window displays all results of the search. It can be seen that the variable "bbh5508" originates from the data provided by SOEP-Core and can be found in the data set "bbh" (survey year 2011). If your search is not so specific, you can also search by keywords. We are still interested in the topic "car".



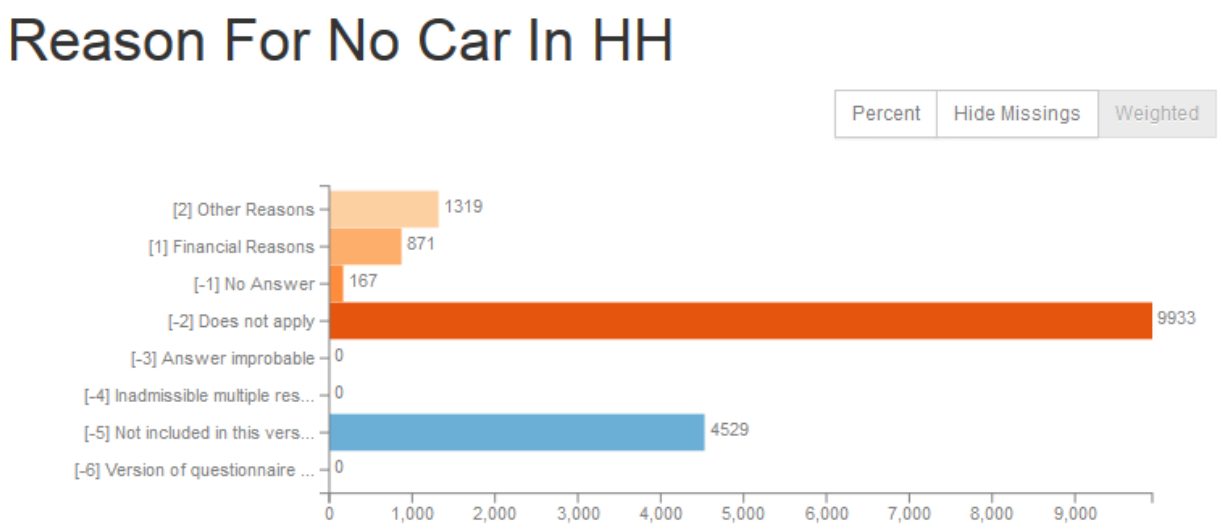
To better limit the 1091 results, the filter options on the left should be used. We are looking for variables from the ordered SOEP-Core datasets. In the windows "type" and "study" we select "variable" and "soep-core".



Now all variables are displayed, which contain the term "Car" in the SOEP-Core data. The variable search can be further limited by specifying the data set or the survey year. For more information about the different data sets in SOEP-Core visit the chapter *Data Sets SOEP-Core*. To select original data that can be assigned to a question in the questionnaire, select the subtype "org/net". The specific selection of the analyzing unit allows you to choose whether the variable should provide information on the household level("h") or on the individual level ("p"). If you are interested in household-specific variables, select "h" as the "Analysis unit". If you are explicitly interested in the survey year 2011, the variable search can be limited to five variables.



There are only five results left, which also shows our searched variable. If you click on the variable "bbh5508" you will get 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 data set, gives you a good overview of the frequencies of a variable.

Related variabels

0:	1984:	1985:	1986:
1987:	1988:	1989:	1990:
1991:	1992:	1993:	1994:
1995:	1996:	1997:	1998:
1999:	2000:	2001: rh/rh5306	2002:
2003: th/th5106	2004:	2005: vh/vh5408	2006:
2007: xh/xh5508	2008:	2009:	2010:
2011: bbh/bbh5508	2012:	2013: bdh/bdh5513	2014:
2015:	none:		

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. At one glance it is possible to see when a variable was measured, how often it was measured and what its name is in the respective survey year

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:	th5106	rh5306	xh5508	bbh5508	vh5408	bdh5513
Dataset:	th	rh	xh	bbh	vh	bdh
questionnaire version with modified filter						
other reasons	2 (1326)	2 (1390)	2 (1096)	2 (1319)	2 (1130)	2 (1494)
no answer	-1 (118)	-1 (125)	-1 (177)	-1 (167)	-1 (211)	-1 (136)
not valid						
version of questionnaire with modified filtering						
not included in this version of the questionnaire				-5 (4529)		
does not apply	-2 (9817)	-2 (9605)	-2 (9555)	-2 (9933)	-2 (9249)	-2 (11230)
can not afford it		1 (827)				
financial reasons	1 (800)		1 (861)	1 (871)	1 (850)	1 (1310)
answer improbable						
inadmissible multiple response						
not included in questionnaire version						-5 (3923)
forbidden multiple response						

The Label table window shows you the absolute frequencies of the variable at different collection times. This makes it possible to identify initial trends in how response behaviour 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 output we see that for the variable “th5106” 800 respondents in the wave “t” (2003) state “financial reasons” as the reason for the absence of a car in the household. For our example variable “bbh5508” in the survey year 2011 (wave “bb”) there are already 871 respondents.

Paneldata.org is an excellent way to get an first overview of certain variables.

Info

Variable name: bbh5508

Dataset: bbh – Household questionnaire

Study: SOEP-Core

Description:

Analysis unit: h

Period: 2011

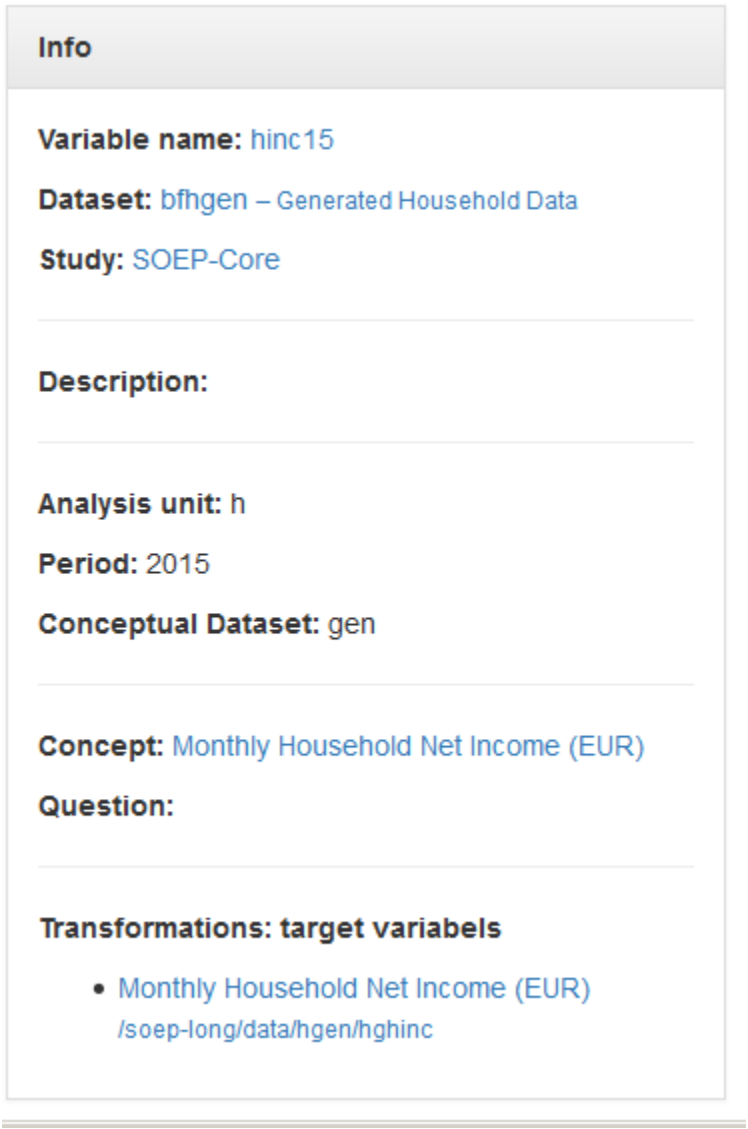
Conceptual Dataset: org/net

Concept: Car (No) Reasons

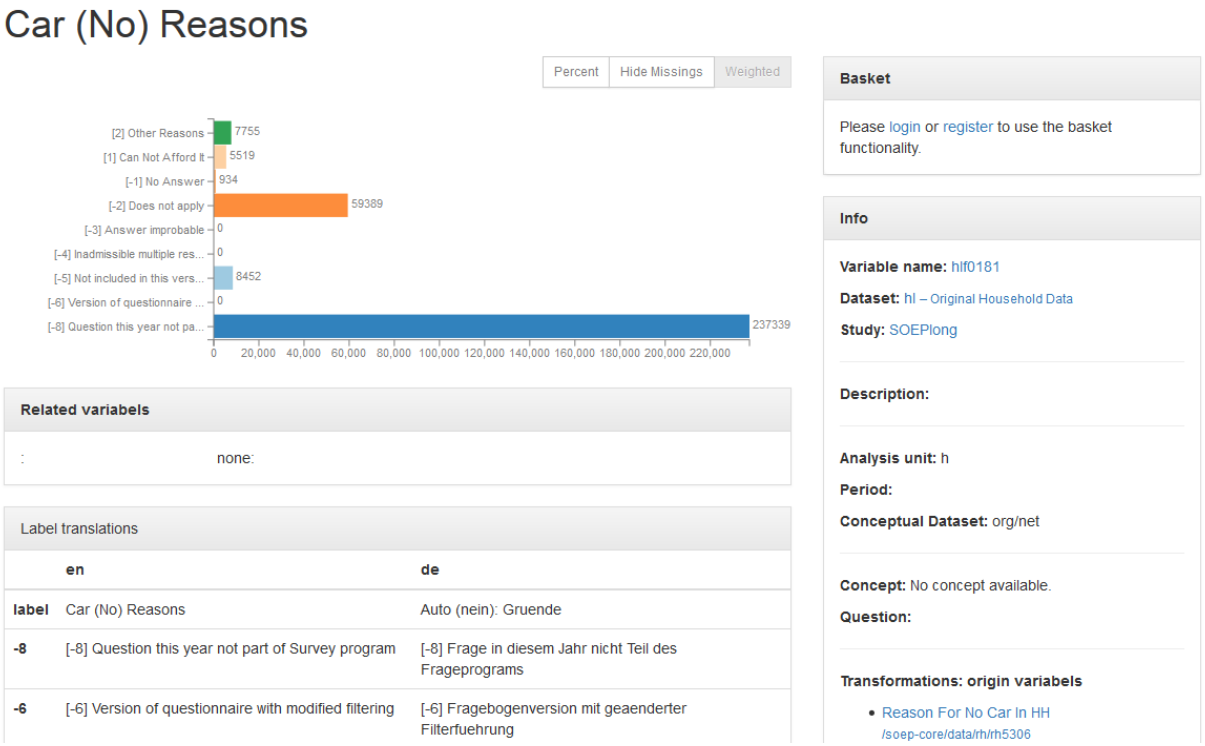
Question:

Transformations: target variabels

- Car (No) Reasons
  /soep-long/data/hl/hlf0181

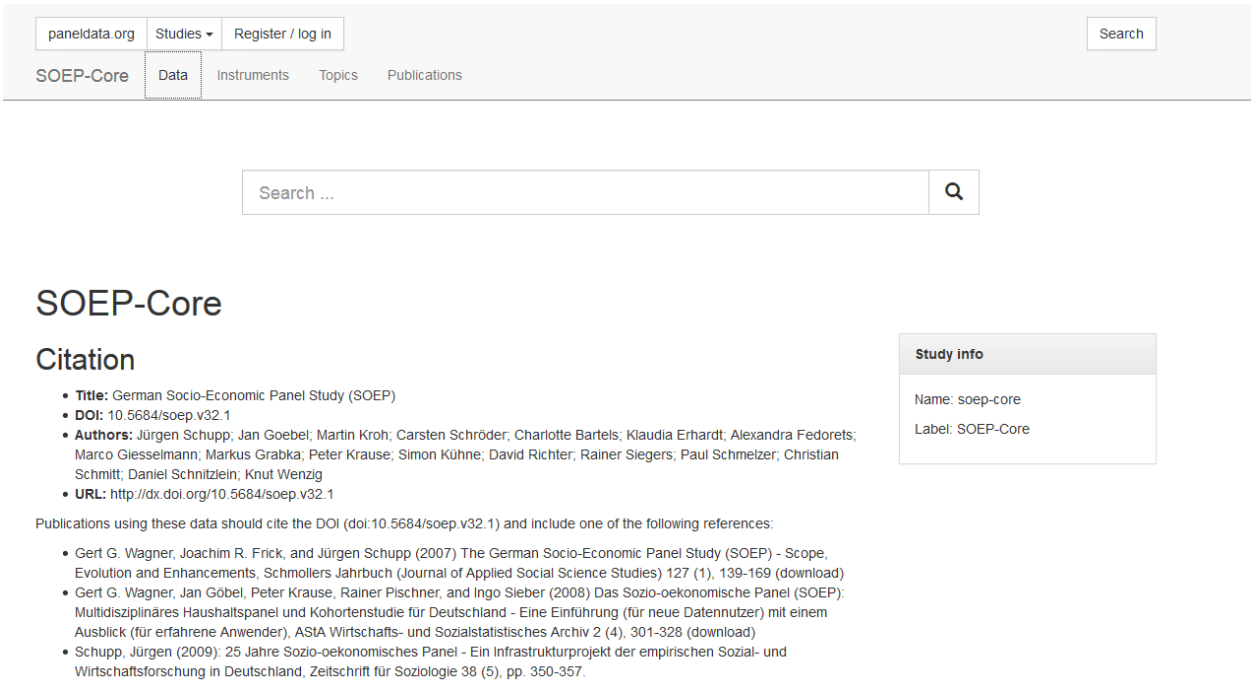


The info box on the right-hand side provides an overview of all relevant information about the variable and the data set. Beside the basic information you will find the information what kind of variable you are looking for under "Conceptual Dataset". In our example "bbh5508" you can see that variables with a "Conceptual Dataset: org/net" describe original variables that are assigned to a questionnaire. Generated variables are "Conceptual Dataset: gen". To get an overview of the different data set types of SOEP-Core, visit the chapter *Overview Data Sets*. In addition, the info box under "Transformations: target variables" provides a link or forwarding to the variable in "long" format. For a more detailed understanding of the long format, read the chapter *Data Structure in long Format (long)*.

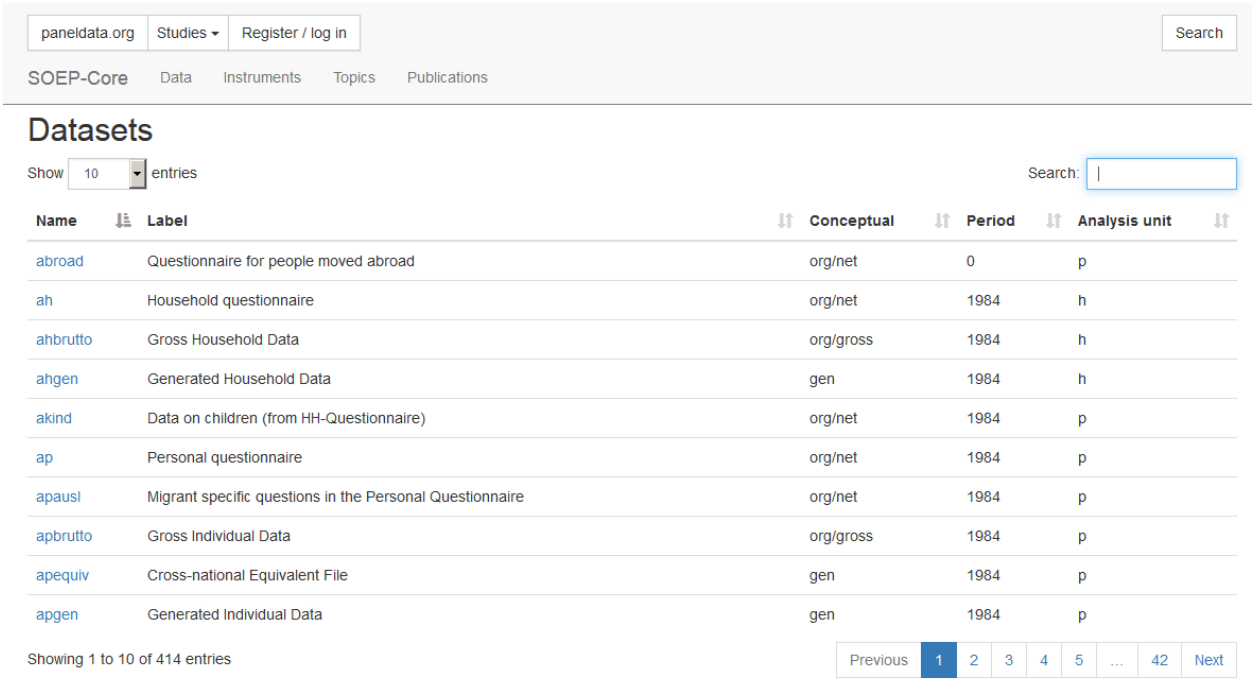


As soon as you click on the “long” variable, you will get to the variable overview for this variable in long-format. The overview of variables does not differ. It can be seen that our example variable “bbh5508” can also be found in long-format in the data set “hl” with the variable label “hlf0181”.

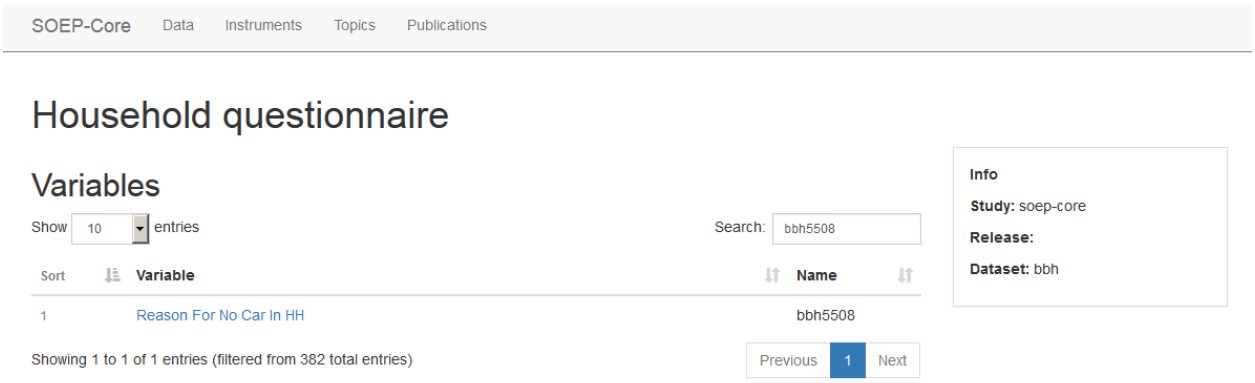
In addition to searching for keywords or using the various filter settings, you can also find what you are looking for directly in the data set search. Open paneldata.org, click on the study SOEP-Core and select the menu field “data”.



Now you get to an overview which shows you all data sets contained in SOEP Core.



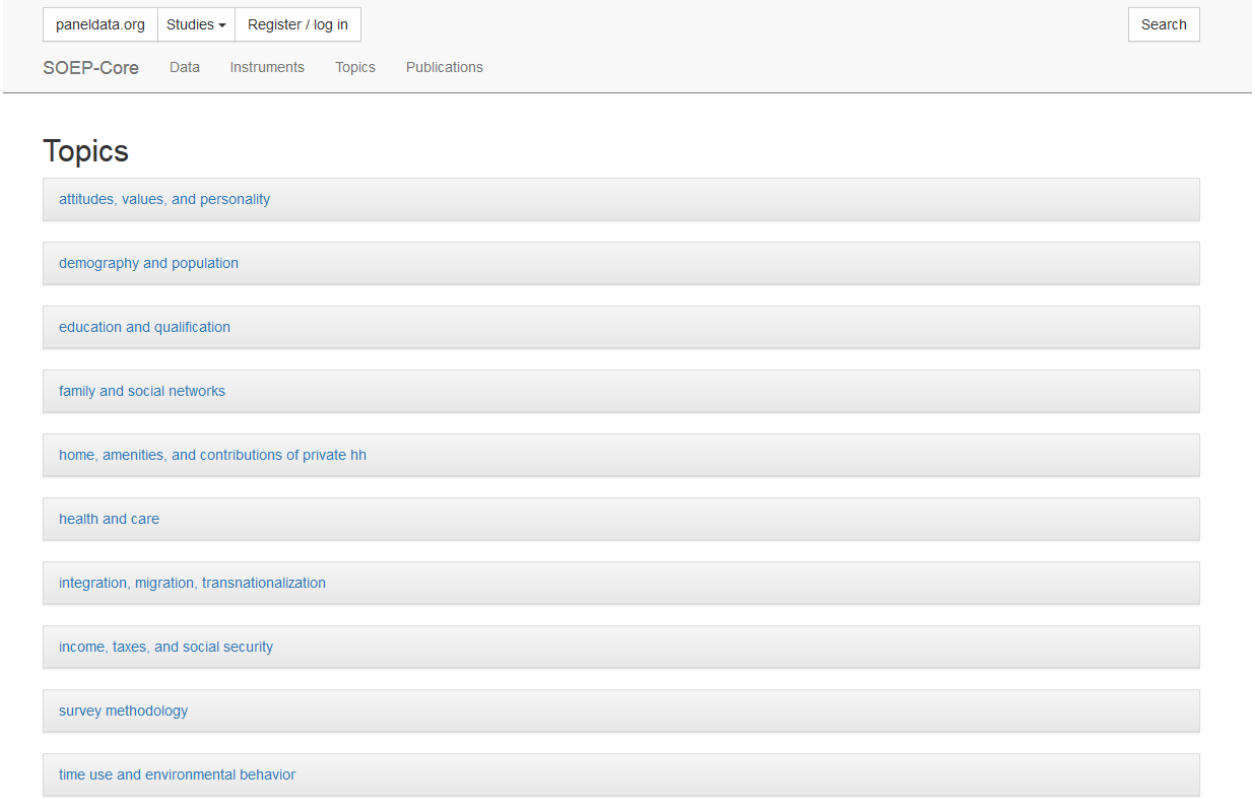
Enter the data set you are looking for ("bbh") in the search field at the top right and click on the data set. You are forwarded to an overview which shows you all variables from the "bbh" data set.



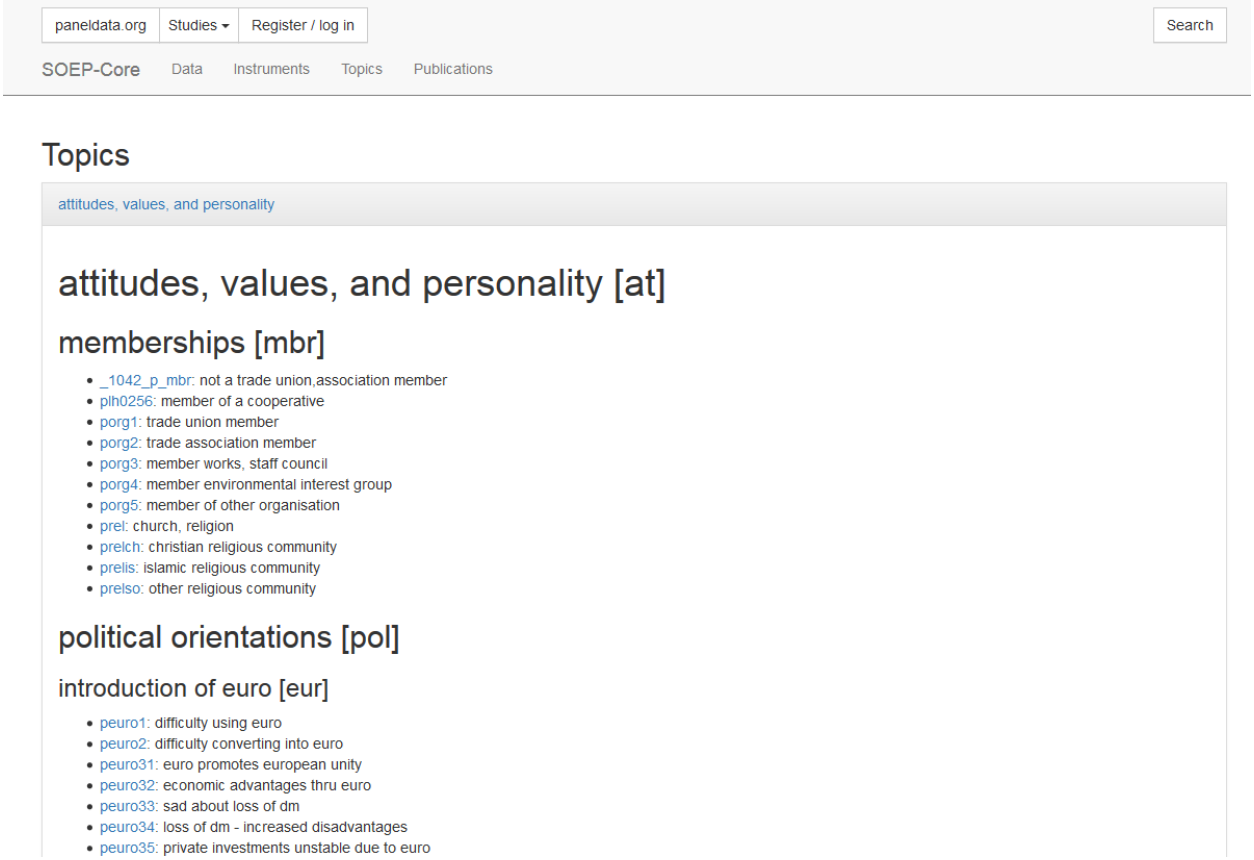
Now enter the variable you are looking for in the search field at the top right and click on the desired variable. You are then forwarded to the variable overview and receive detailed information about the variable. Paneldata.org offers the user very different search options to suit the individual search behavior of each user.

6.3 Topic Search with paneldata.org

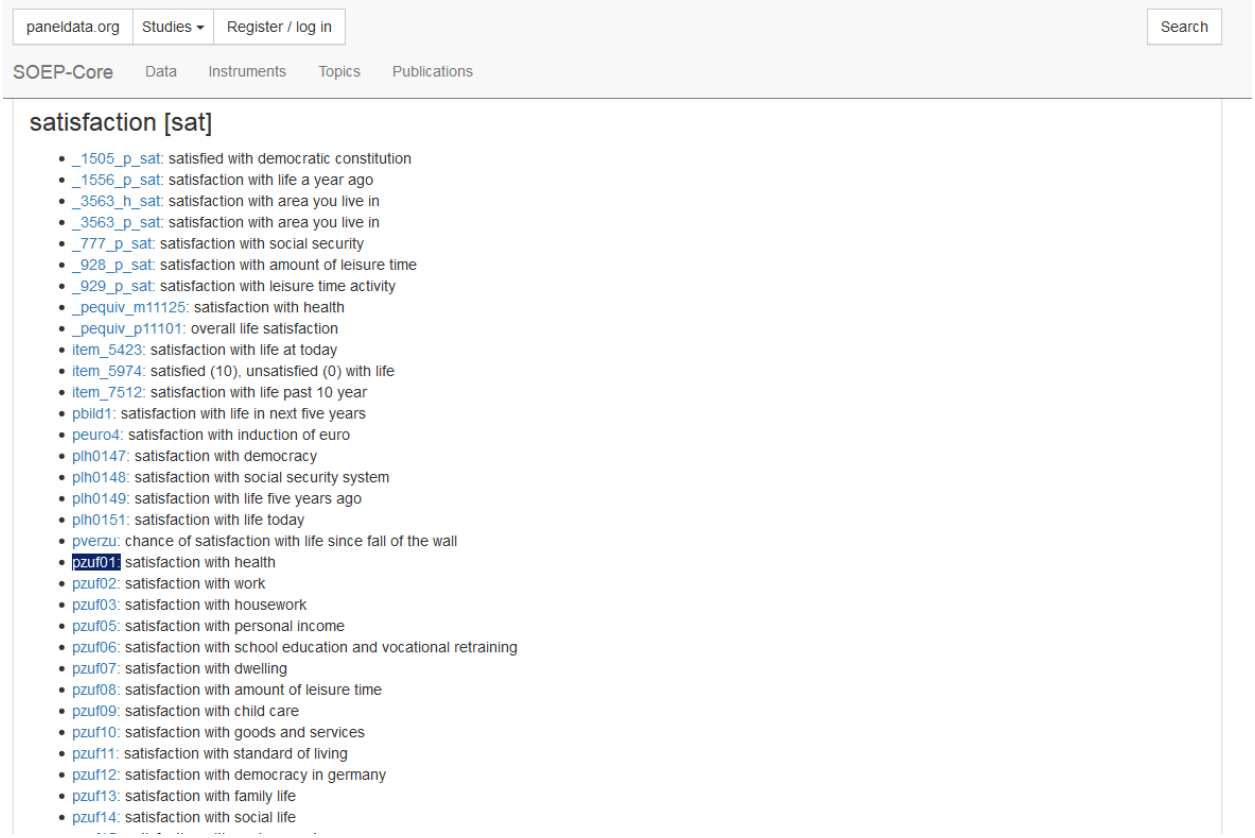
In order to obtain an overview of the content of the variables provided by the SOEP, the variables on paneldata.org were assigned to different topics. If you are looking for your research variables and do not want to check all data sets or questionnaires, the topic search on paneldata.org could help you. Open and select the main study SOEP Core. The upper navigation bar leads you to the Topics area. Click on Topics and have a look at the list of variables.



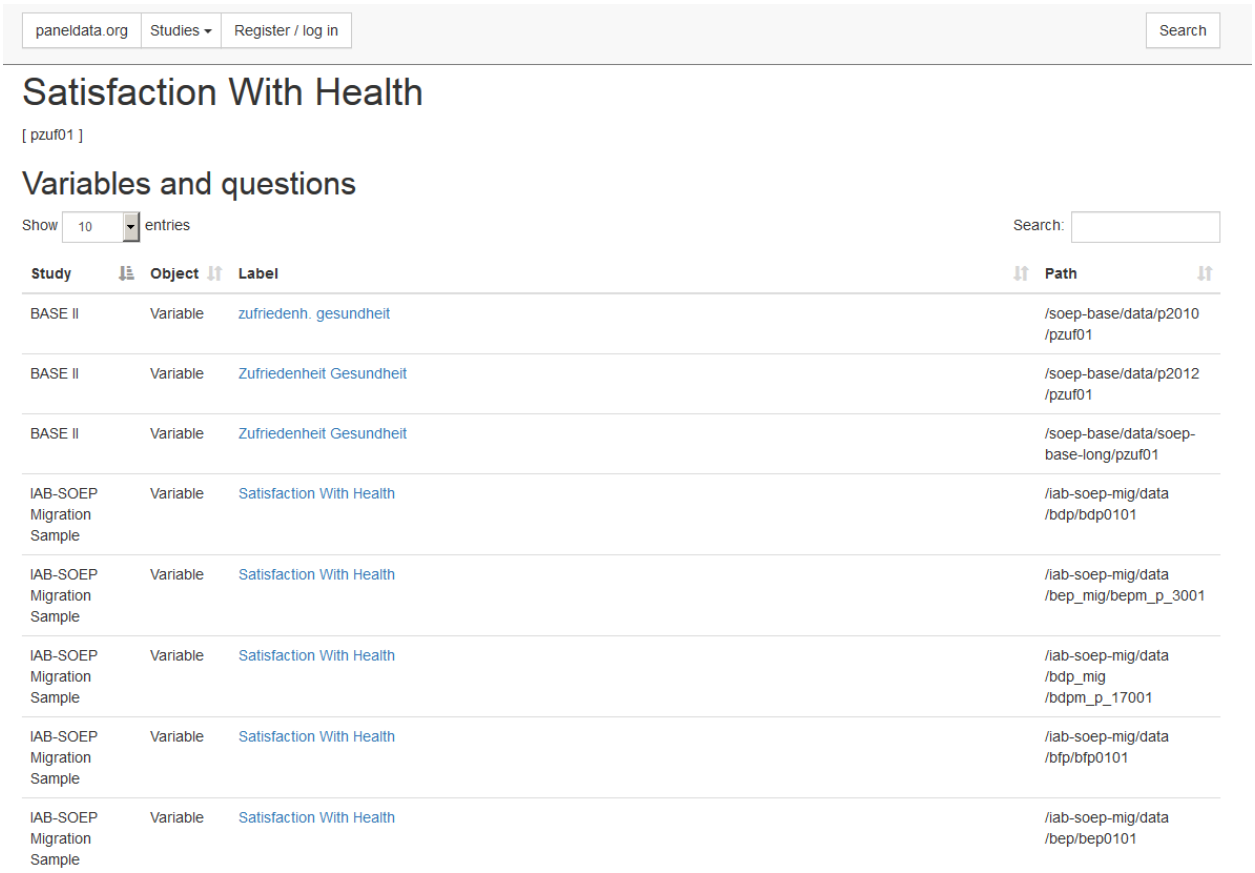
Select a topic that corresponds to your research interest and a more specific selection of topics will appear



For example, if you are interested in different types of satisfaction, select the appropriate topic “attitudes, values, and personality [at]”. With a little search you will discover the sub-topic “satisfaction[sat]”.



Suppose you are interested in health satisfaction. Based on the label, the “pzuf1” concept could be of interest to you. By clicking on the concept “pzuf1” you will get to the concept overview.



paneldata.org | Studies | Register / log in | Search

Satisfaction With Health

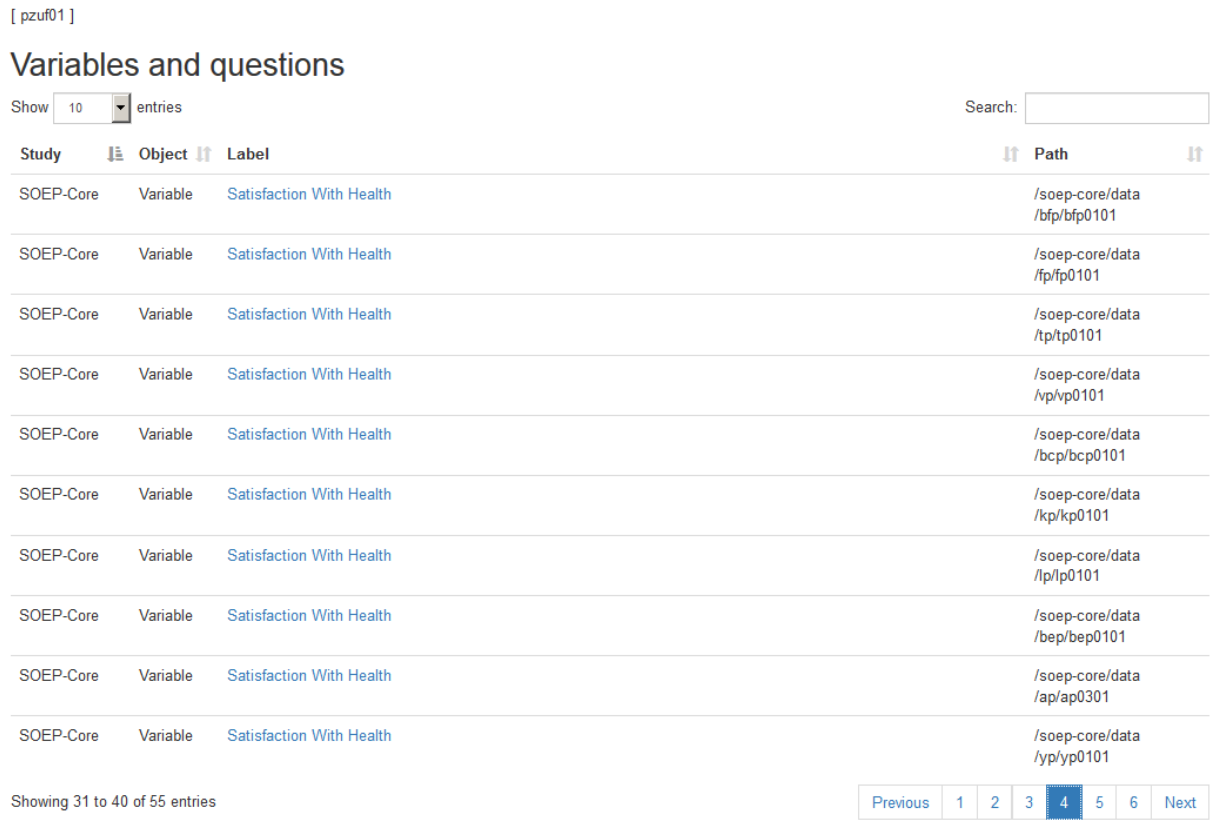
[pzuf01]

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 of the 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 in which studies the same variable is linked via the 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 known overview of variables with all relevant information. The “Object” column in the concept overview shows you the type of information which is displayed.



[pzuf01]

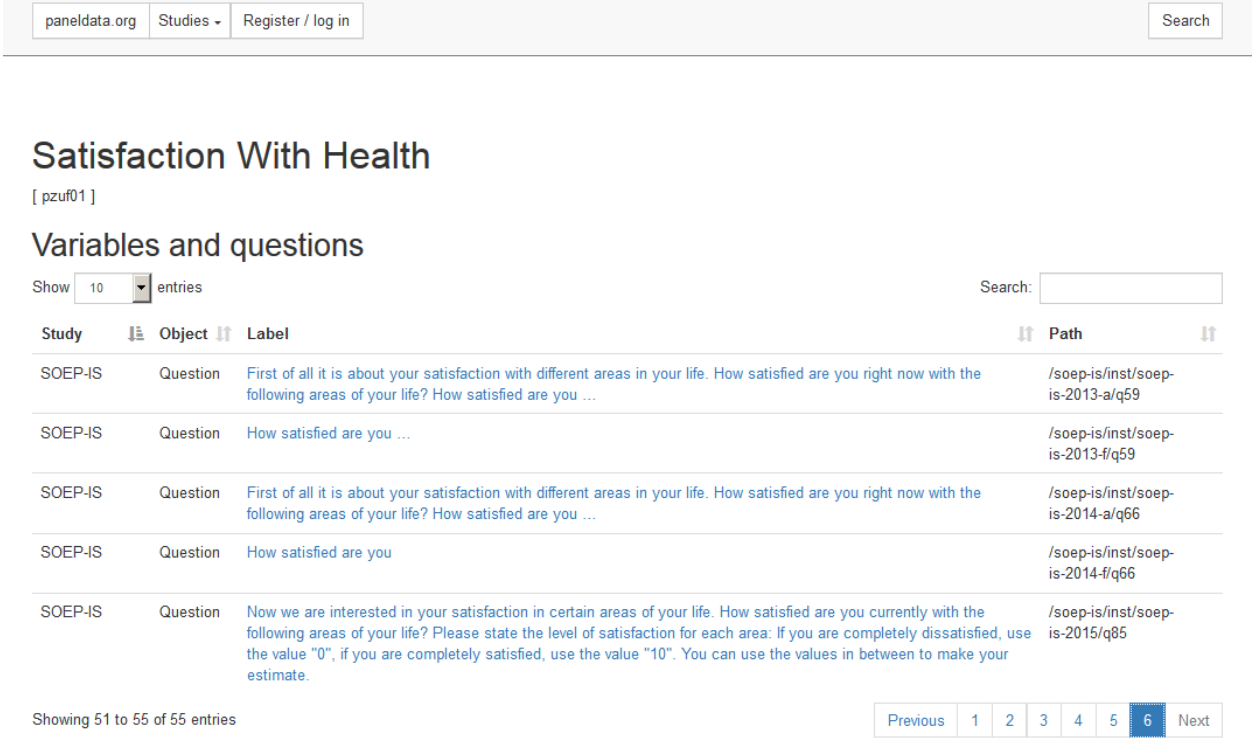
Variables and questions

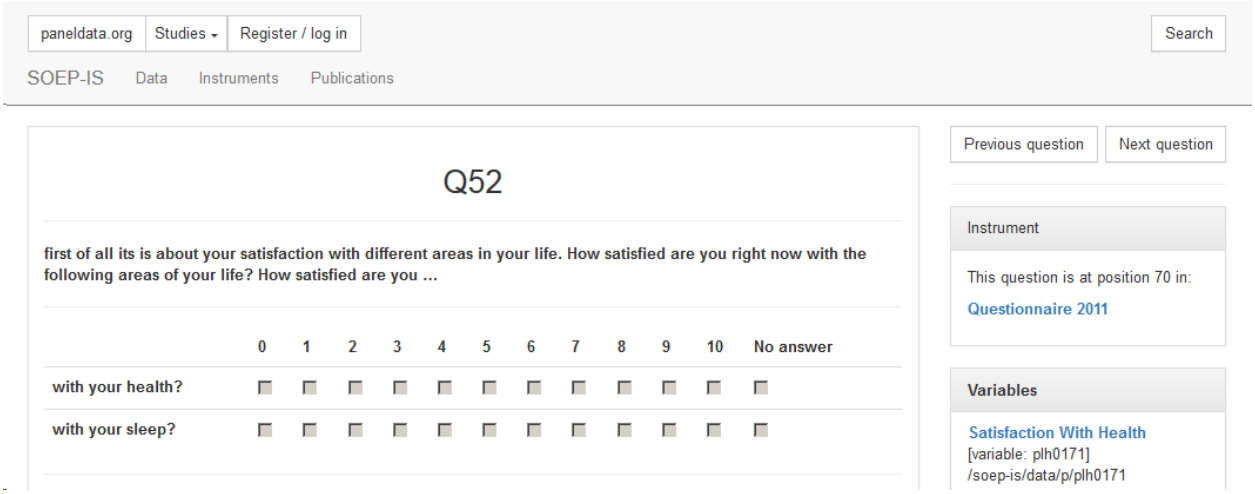
Show 10 entries Search:

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

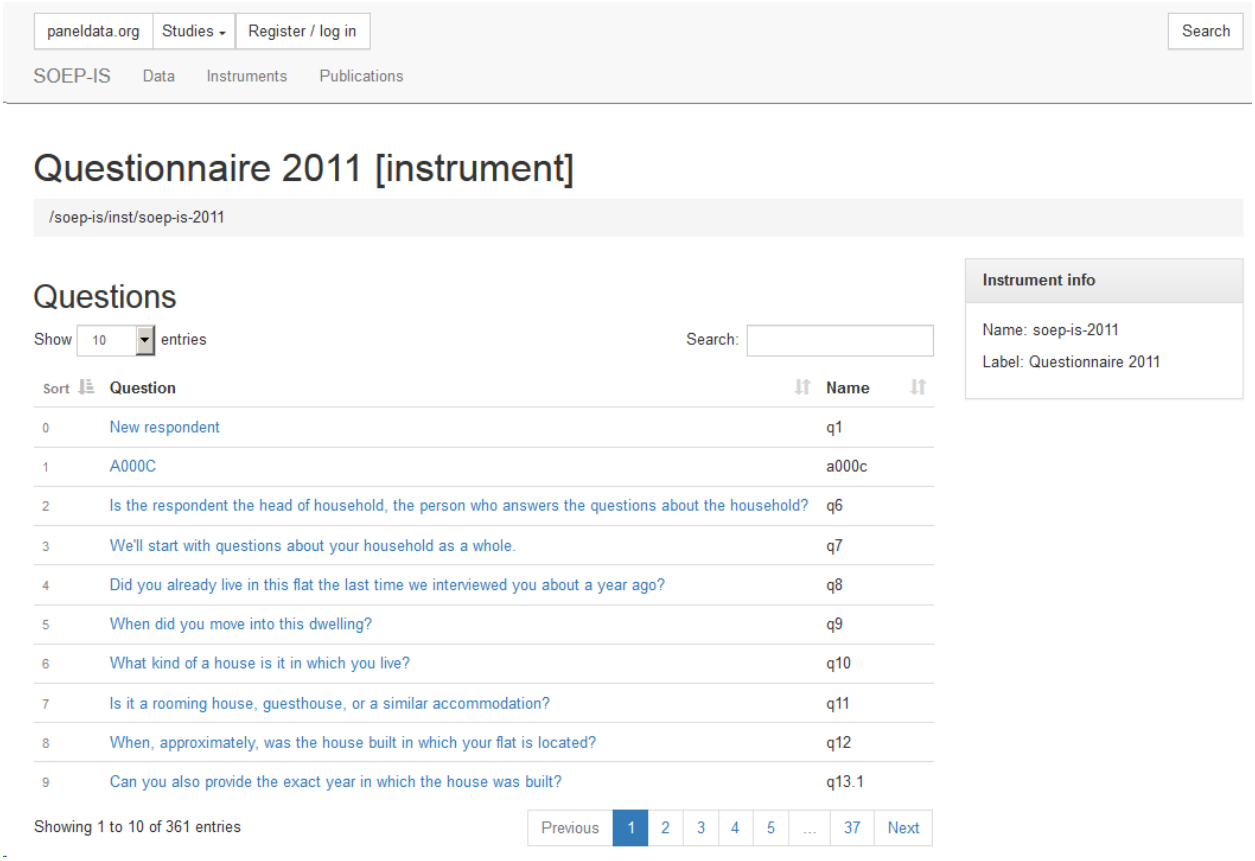
Showing 31 to 40 of 55 entries Previous 1 2 3 4 5 6 Next

In addition to the variables linked via the 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 get an overview of the question and determine possible differences. Click on the desired question and you will be taken to the question display.





Attention: The variable search via the questionnaires is unavoidable in order to find out the exact wording of the question and the possible filter structure. The question display of 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 for the survey year 2011 of the SOEP-IS study is now displayed. You can navigate through the questionnaire in this overview. The search field allows you to search for research-relevant terms. Click on the question to access the question display.

6.4 Documentation of Generated Data

The range of generated variables and data sets from SOEP-Core is very extensive. To make work easier for users, many variables are already generated for the user in the data preparation process and published with SOEP-Core. The large number of generated data sets and variables is comprehensively documented so that the generation process remains transparent for the user. Here you will find an overview of the

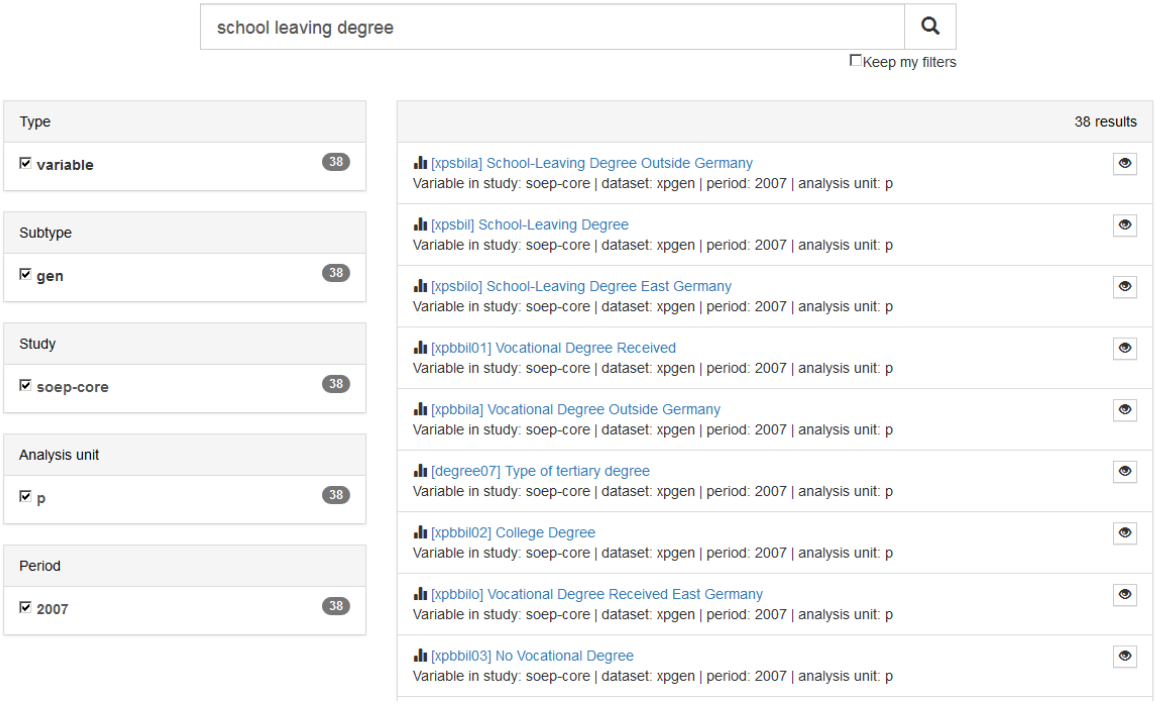
Example: A number of frequently used variables are provided in SOEP as so-called generated variables (e.g. data sets $PGEN and $HGEN). These variables are checked for consistency across waves and have a uniform name. Please use the appropriate documentation to answer the following questions:

a) In which variable is the highest school leaving degree for the persons surveyed in 2007?

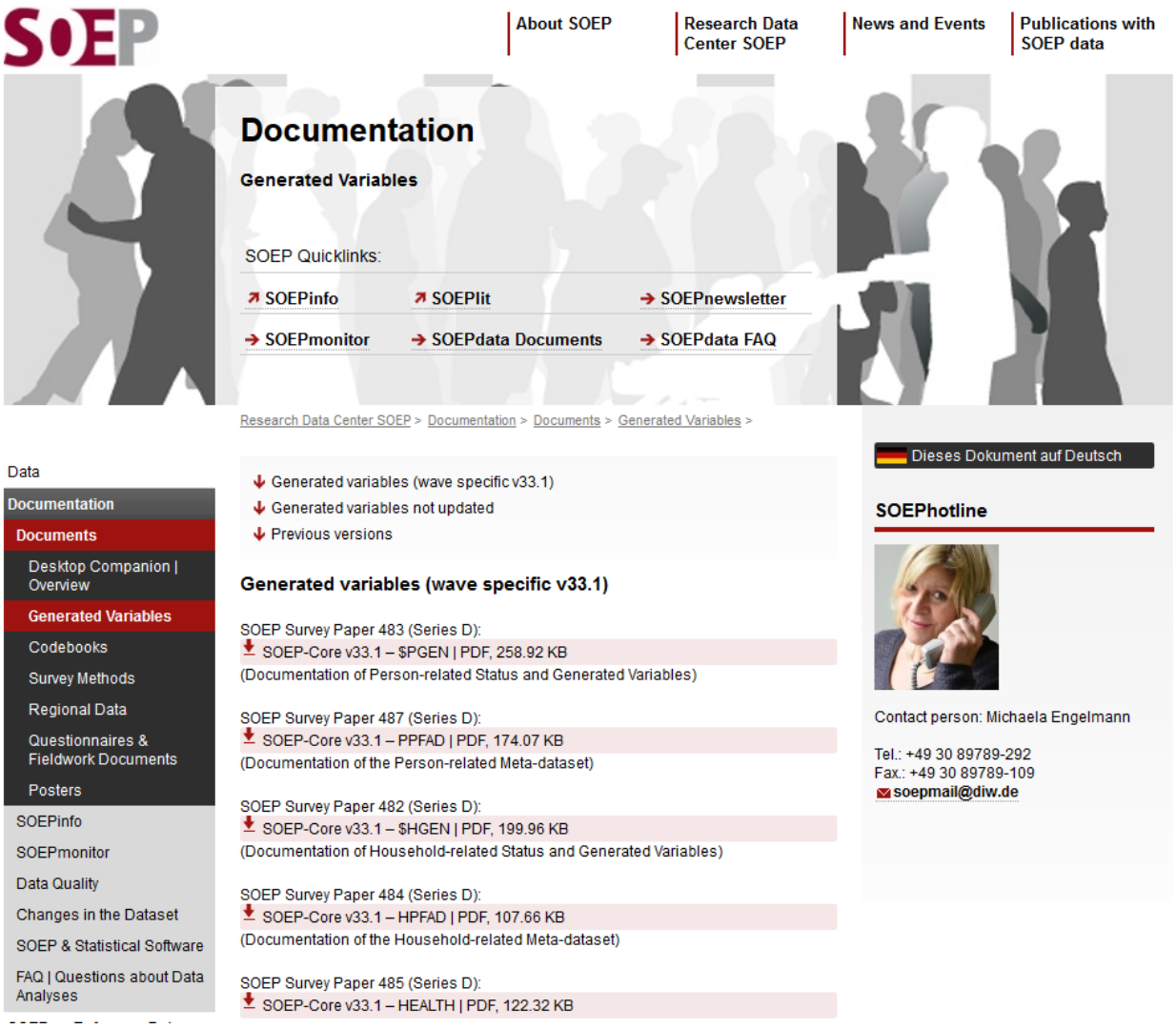
To search for the variable with the highest school leaving degree, use paneldata.org. Open and enter school leaving degree in the search field. Then specify your search by adjusting the filter settings as follows:

- type: variable
- subtype: gen
- study: soep-core
- analysis unit: p

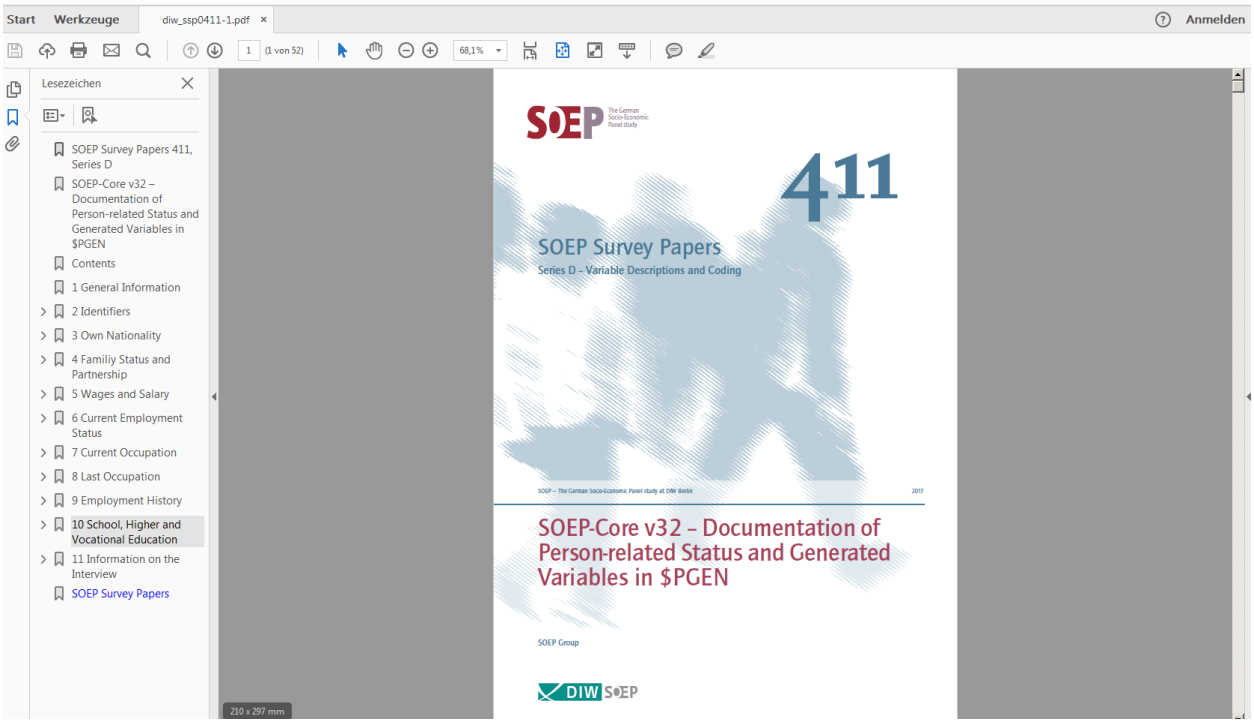
- period: 2007



All variables could contain the information you are looking for. Since almost all variables in the search result come from the generated "xpgen" data set, the documentation for the $pgen data set should be used. Open the



Now select the documentation of



The table of contents on the left shows you a thematic classification of the data set. 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.]

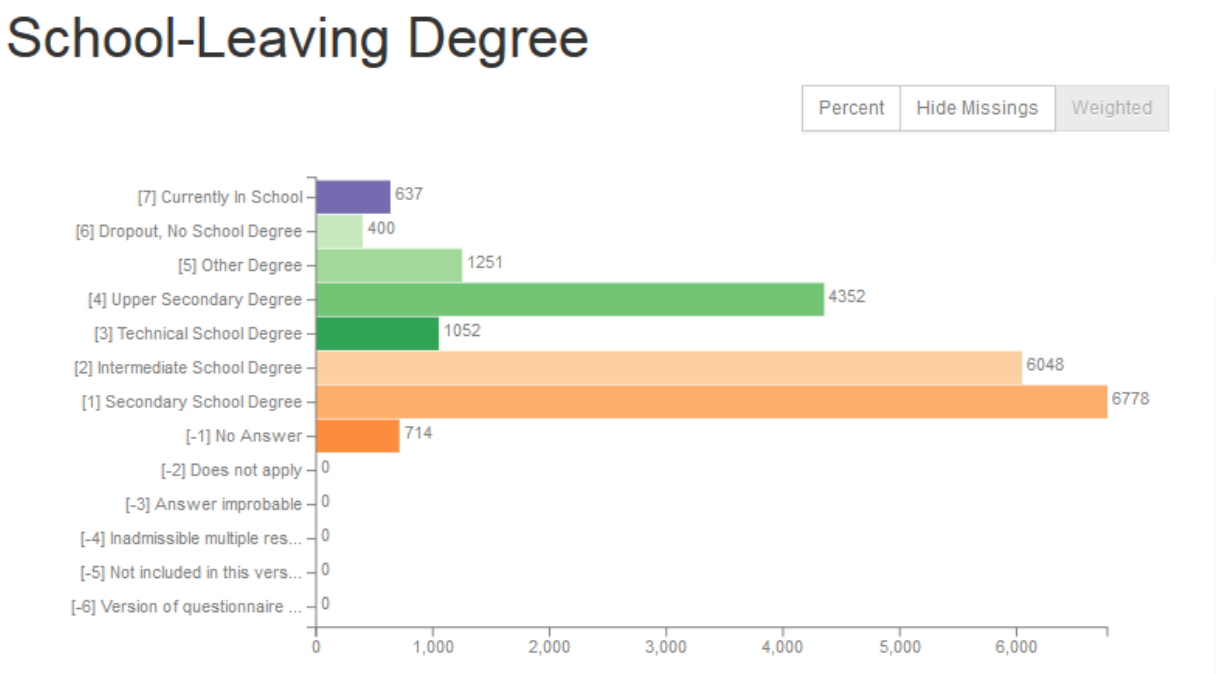
For more information, contact: Peter Krause (Tel. +49-30-89789-690)

After a few searches you will find the variable you are looking for. Some interesting information can be derived from the documentation. It can be seen that the information from the generated variable has been taken from the CV questionnaire since 1994 and is surveyed once. In addition, the two additional variables $psbila and $psbilo are explained in more detail. The documentation describes that the $psbil variable is updated regularly and also takes into account possible changes in the highest level of education. This is precisely why it is worth using the generated variable to represent the most recent highest school leaving degree of those surveyed.

The variable we are looking for is xpsbil and describes the highest school leaving degree of the persons surveyed from the survey year 2007.

b) Which values are given to persons with Upper Secondary Degree (Abitur) in 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

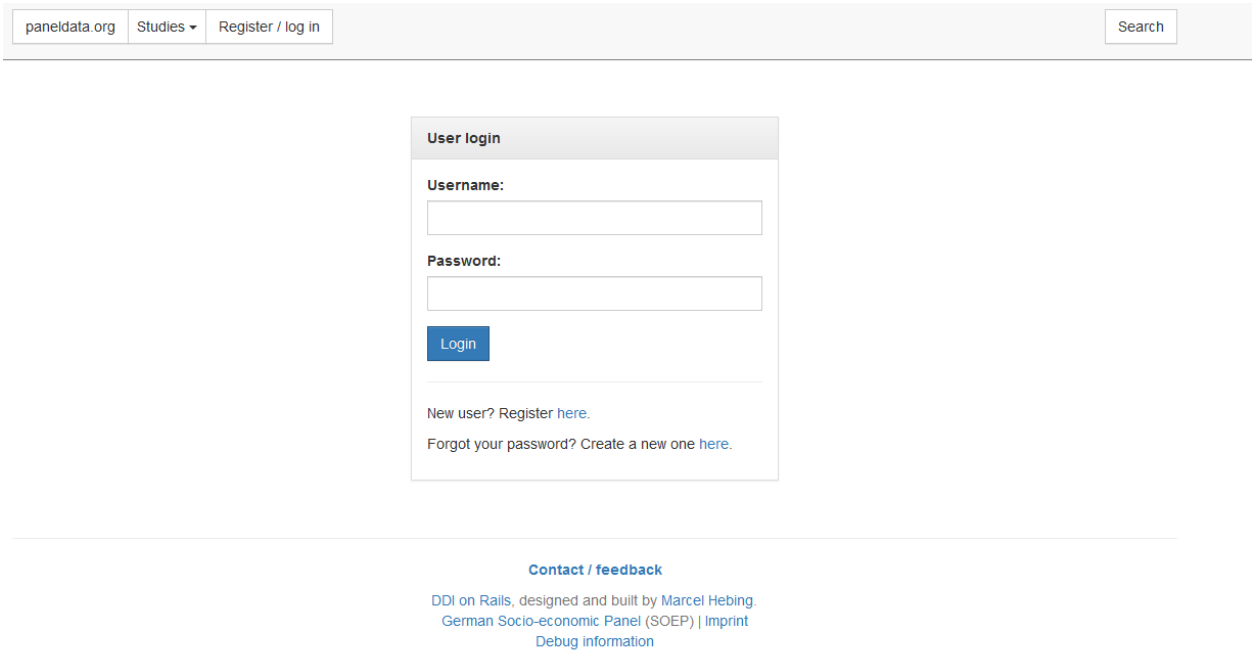
You can answer the question without opening the data. In the 2007 survey year, the variable “xpsbil” with the value code “4” describes the answer category “Upper Secondary Degree (Abitur)”.

6.5 Syntax Generator on paneldata.org

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

Open

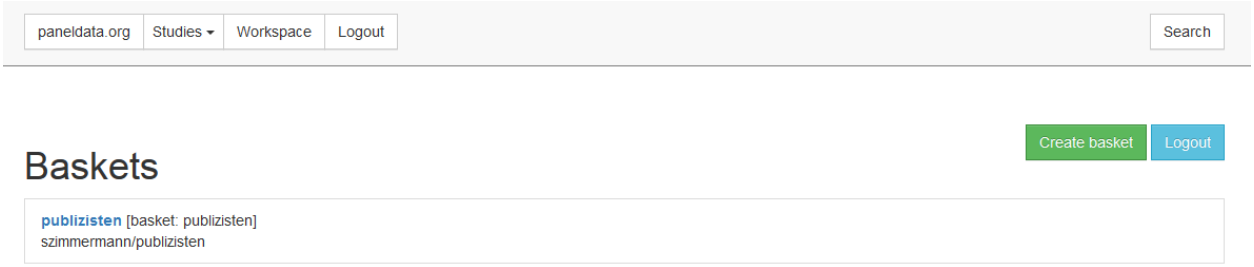
Click on the "Register/ log" to log in to paneldata.org.



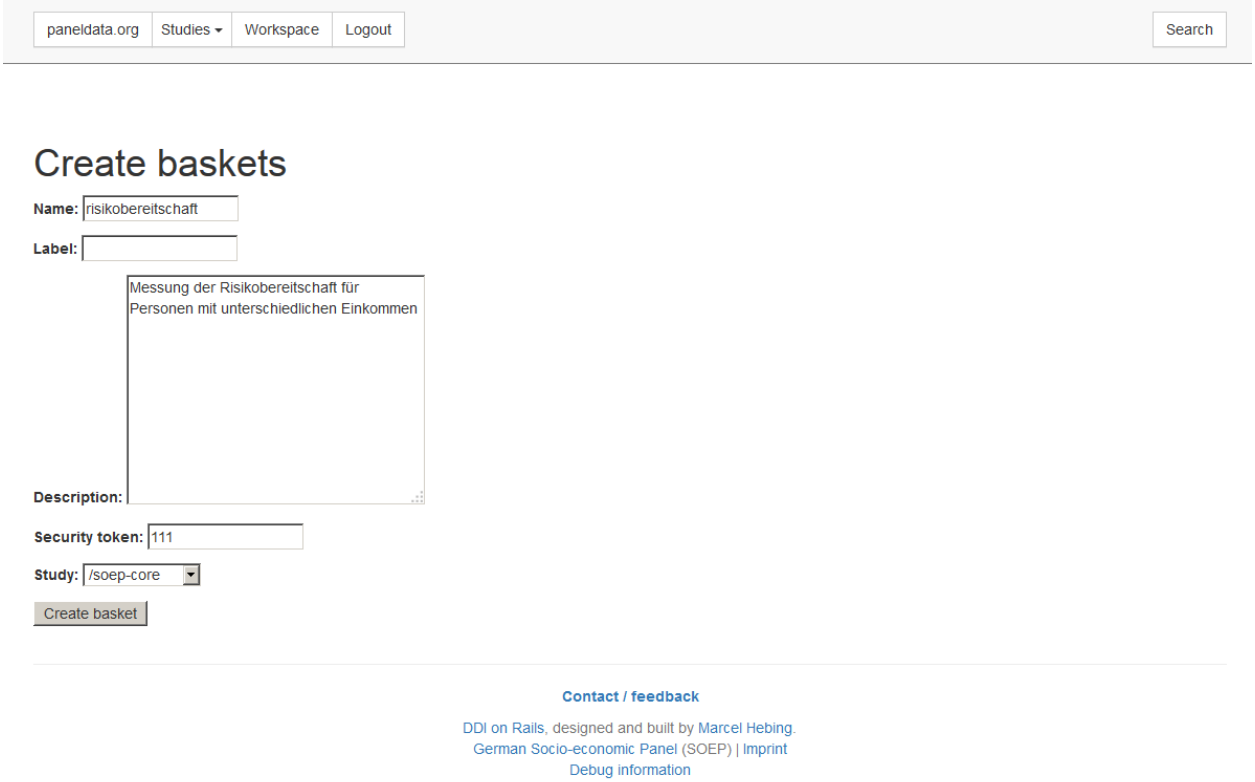
If you have already registered, you can login in the "User login" area. As a new user you can register at "Register

here". Once you have logged in successfully, you have access to the variable basket and the syntax generator.

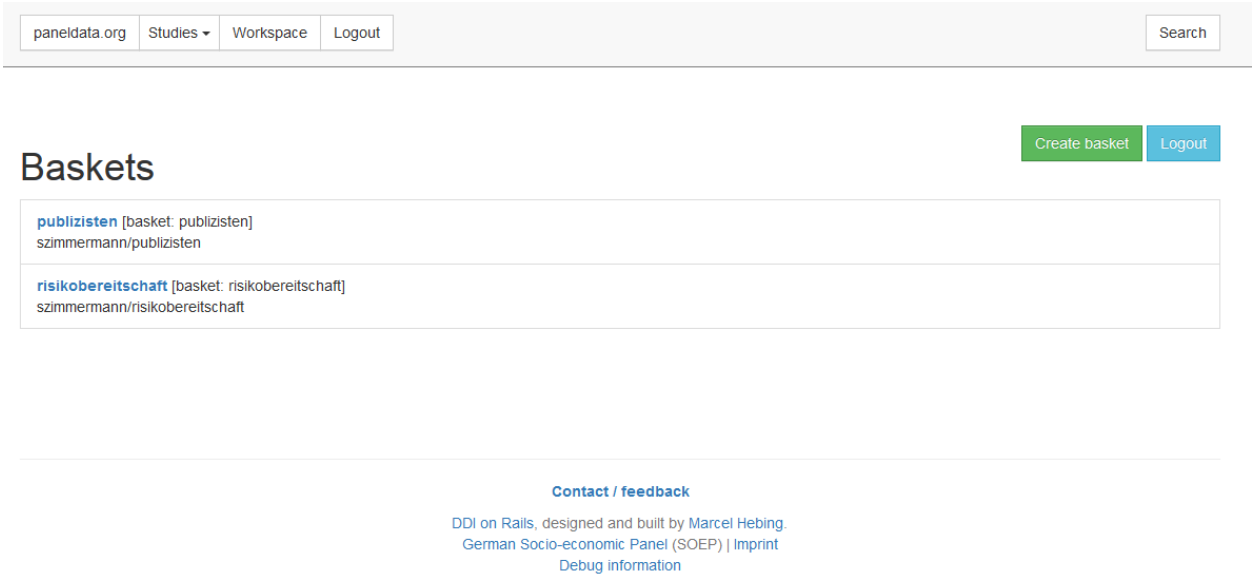
To access the activated functions, click on the navigation field "Workspace". You will be taken to your personal workspace on paneldata.org.



The "Workspace" displays your created variable baskets. If you click on "Create basket", you can create a new basket.

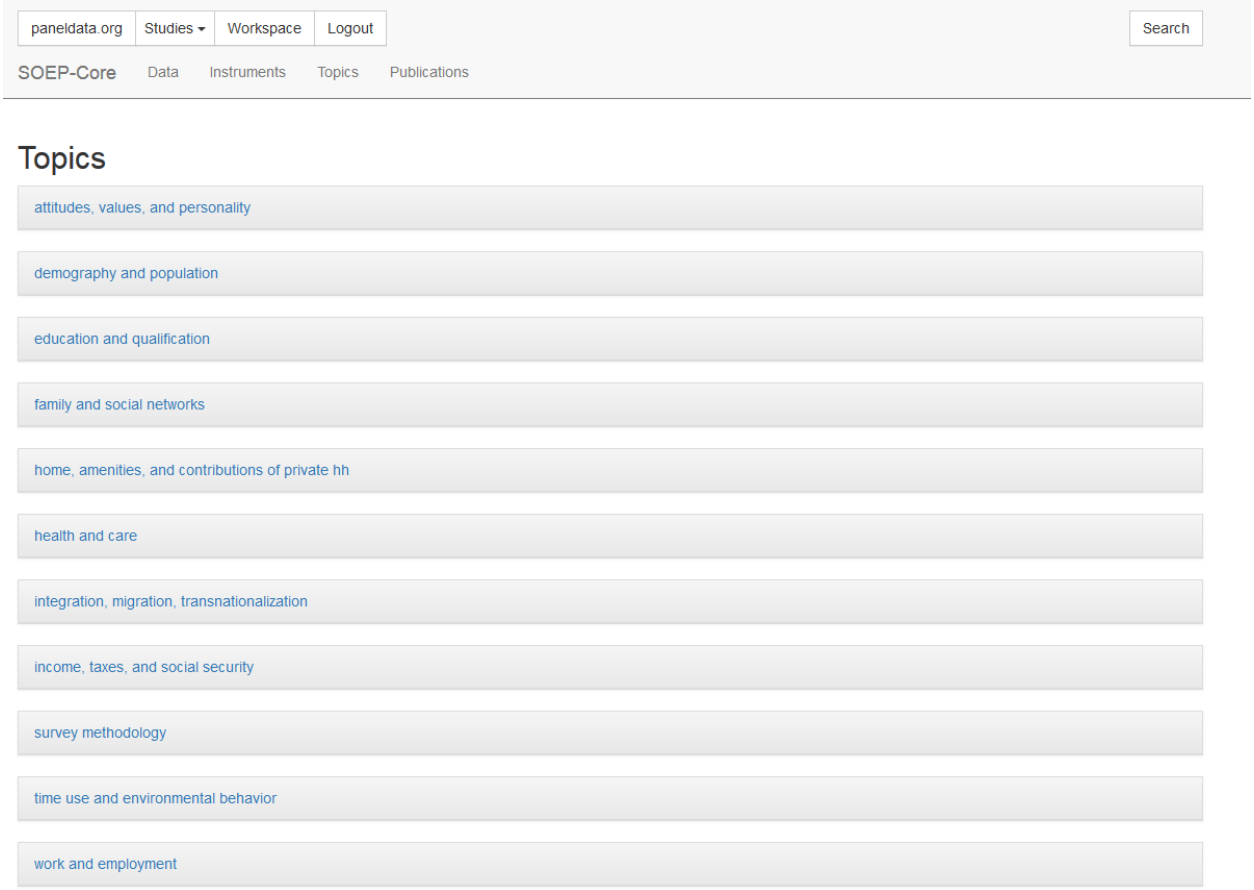


When creating the basket, first define the name of the variable basket. The name must be lower case to be accepted by Paneldata. Optionally, you can assign a label and enter a description. You can create a security key via "Security token". Finally, you select the study that you want to use as a database for your research. Now click on "Create basket" and your newly created variable basket appears in the "Workspace" interface.

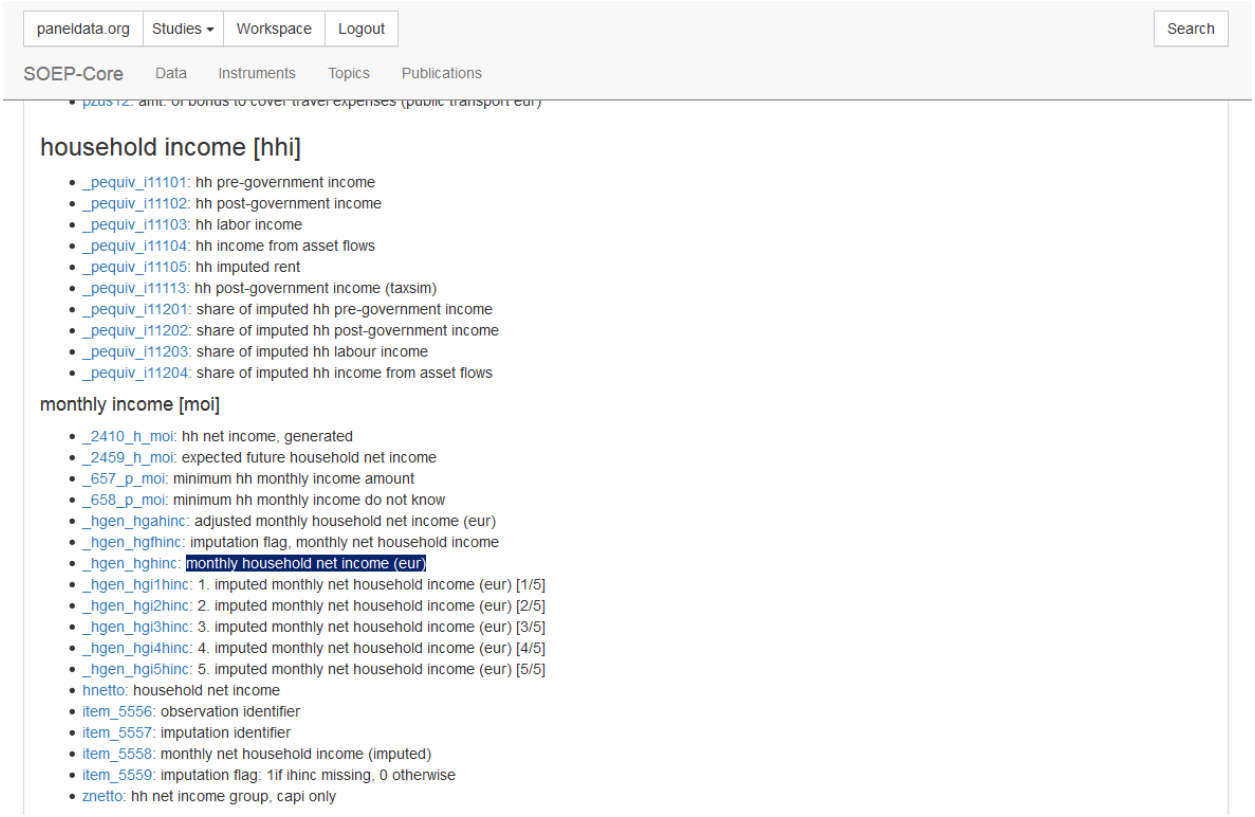


Now search for your relevant variables on paneldata.org and add them to your individual basket. For example, you are interested in the monthly net household income. If you do not know the variable name, you can find the superordinate

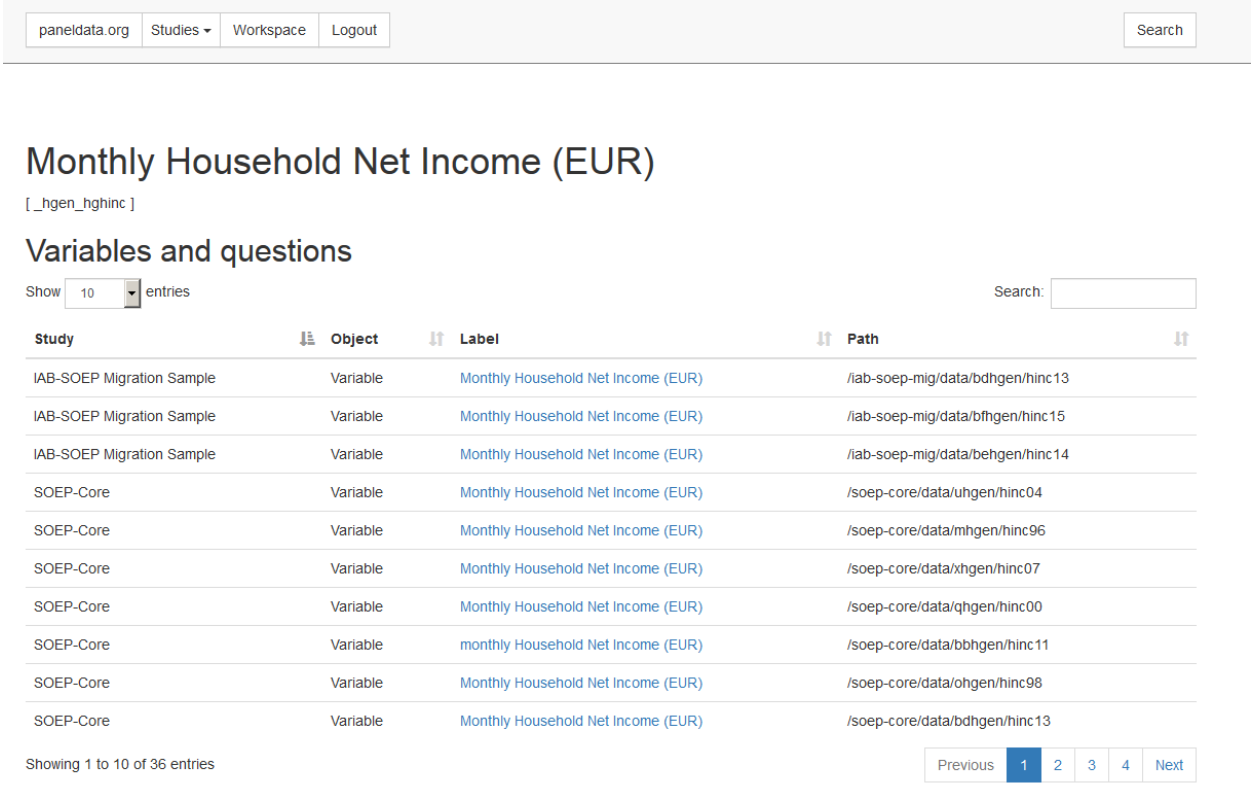
concept using the topic search. Click on the navigation field “paneldata.org” to get to the main page. Select the study SOEP-Core and click on the navigation field “Topics”.



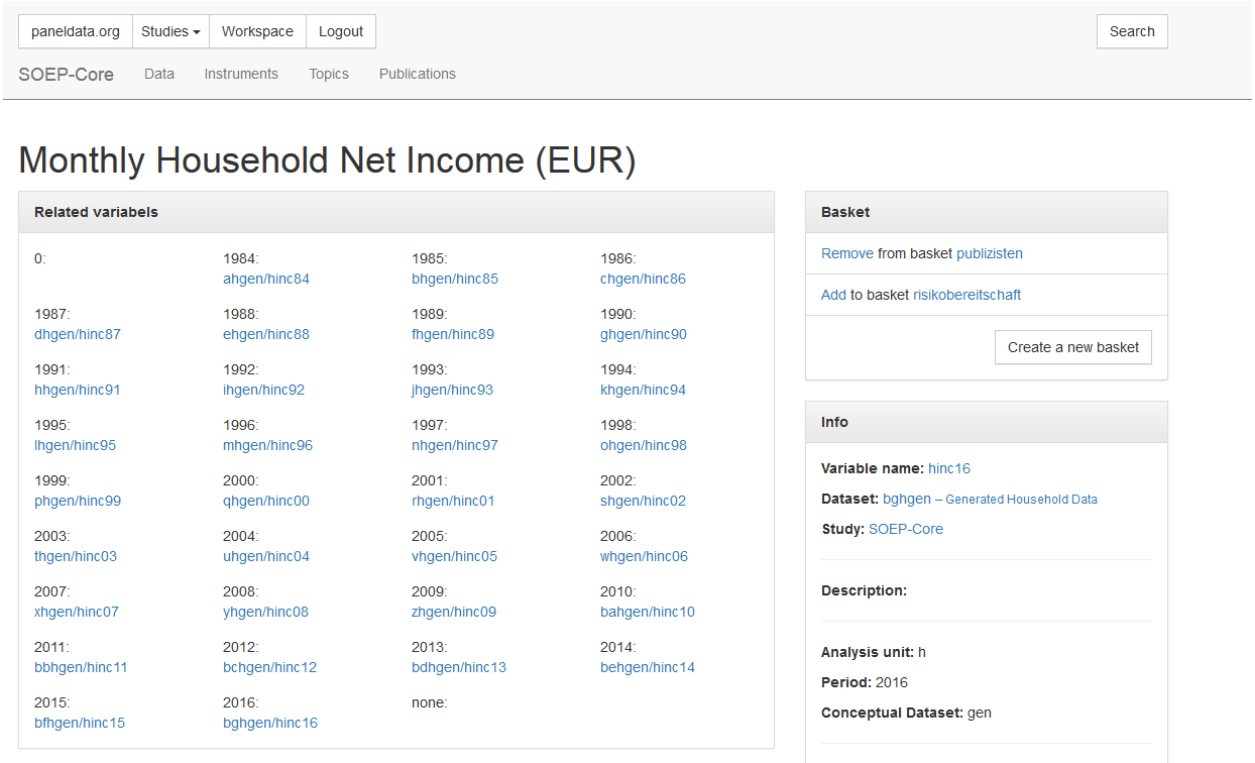
Check the different topics for income-relevant concepts and select “income, taxes, and social security”.



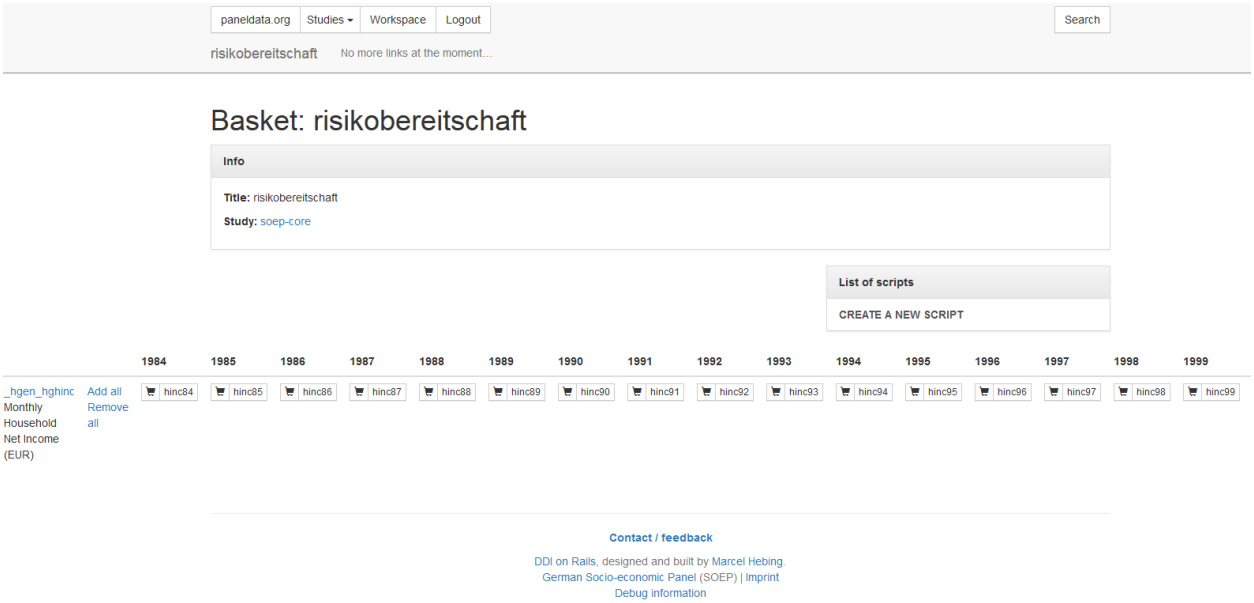
Browse the topic list and you will reach the sub-topic “household income hhi”. There you will find the concept you are looking for under “monthly income moi”. Click on the concept and you will see the history of variables, possible links to other studies and perhaps the question in metadata-based form.



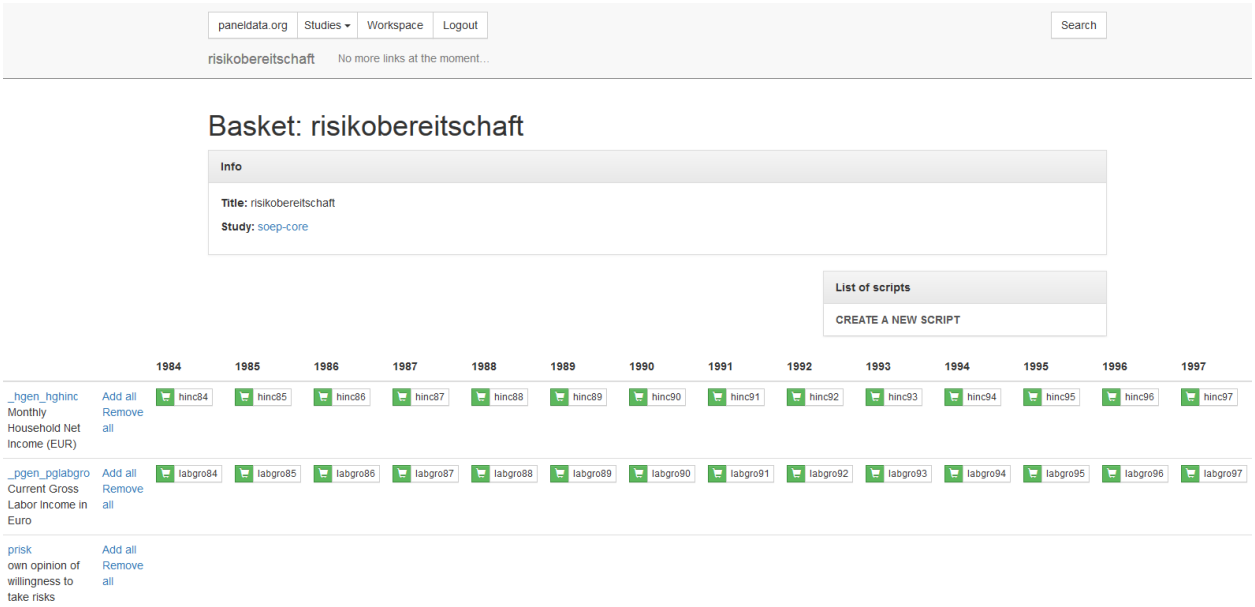
Select the variable of your desired study SOEP-Core and you will reach the variable overview with important information about the variable. In the variable overview, you should make sure that the variable also meets your requirements.



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” to go to basket creation and its variable is automatically 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 navigation field “Workspace” and you will also return to the basket overview.

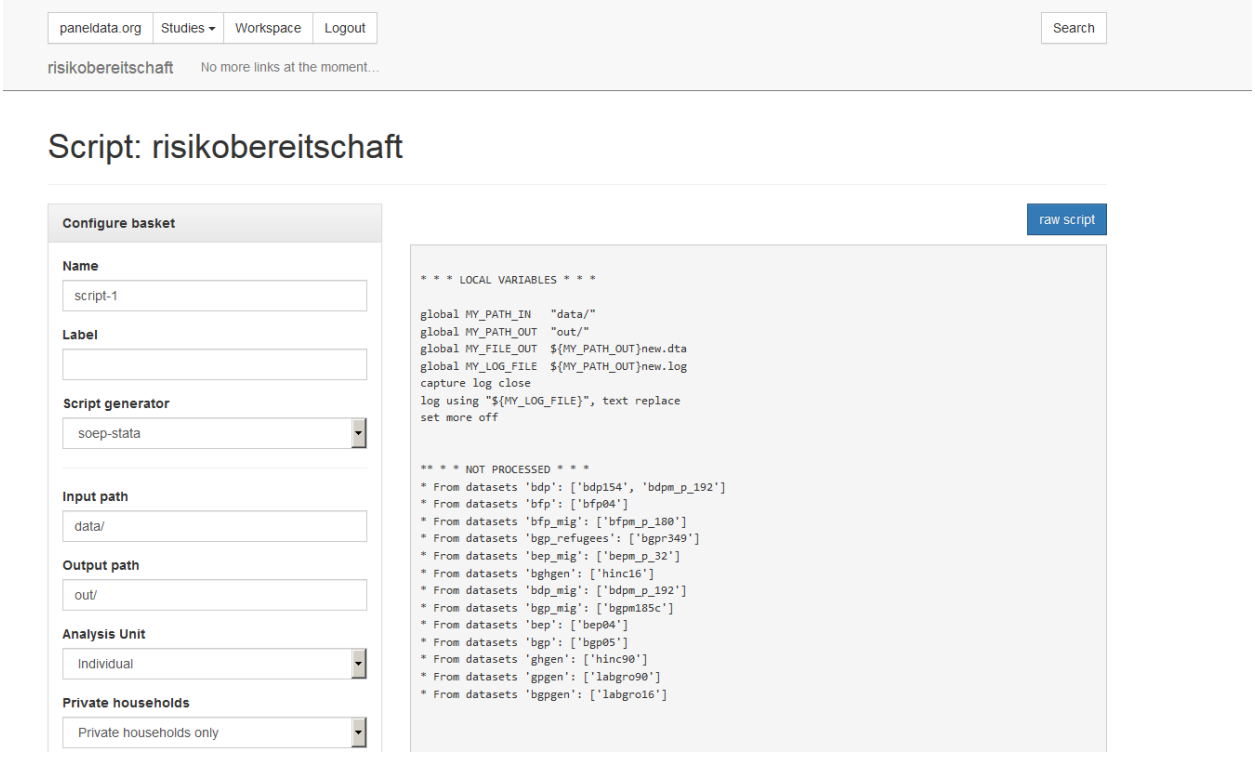


Click on the basket with your added variable and you will get an overview of all variables in your basket. With “Add all” you add the variables of all survey waves and the shopping cart is highlighted in green. If you are interested in a specific survey period, you can select the wave-specific variables by clicking on the shopping cart. Click on “Remove all” to remove the variable from your basket.



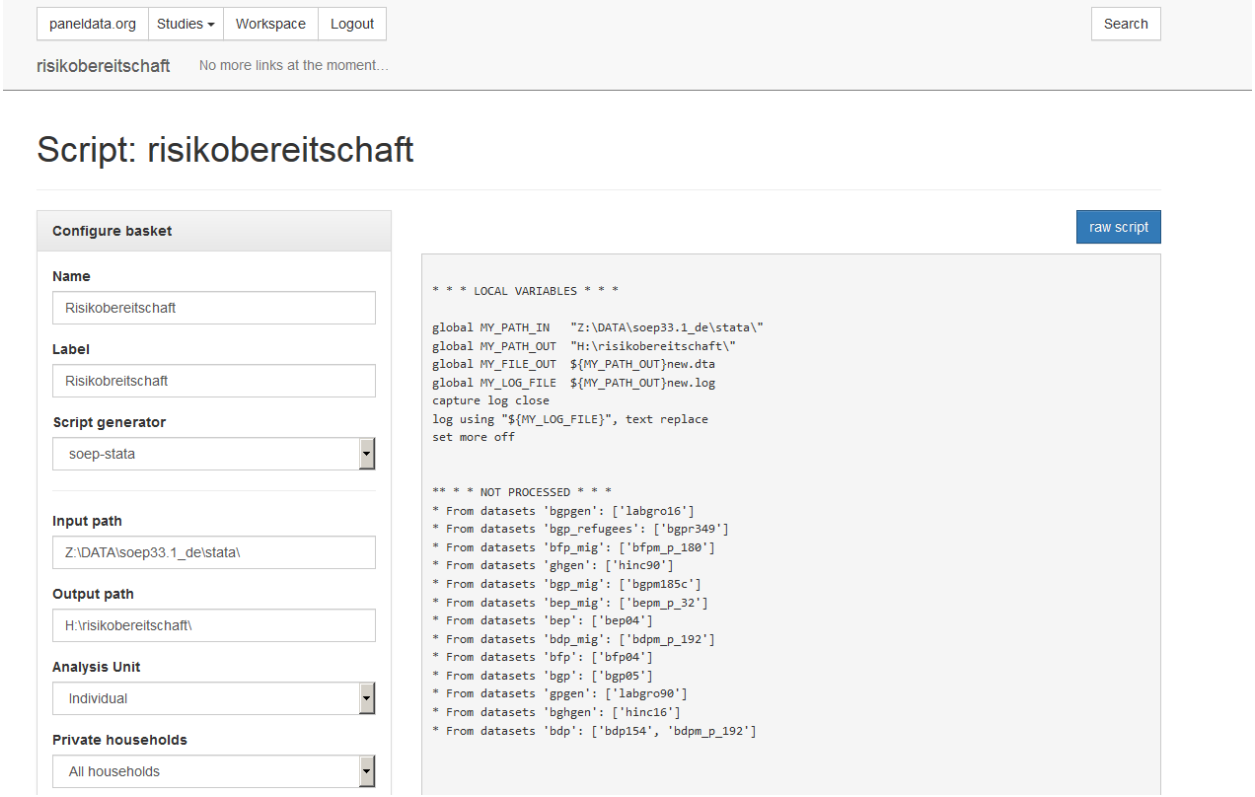
Once you have filled your basket and selected the desired survey waves, you can merge all variables into one data set.

To do this, click on “CREATE A NEW SCRIPT” in the “List of scripts” area.



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. Then enter the path where you have stored your data records in the “Input path”. In the “Output path” you write your desired output path for the created data set.

In the “Analysis Unit” section, you decide whether all persons are considered individually within the household (“Individual”) or whether you are only interested 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 data set without missing codes. The respondents provided information on all variables. For more information about balanced and unbalanced datasets visit the chapter *Panel 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 statistics software. To name the data set correctly, you should change the name of the data set in the script. Execute the script with your statistics software and you will receive your data set with all your choosen variables.