Introduction to the German Socio-Economic Panel (SOEP) study – Part I
Introduction to the SOEP I

1. A Short Introduction to Panel Data

2. What is the SOEP?
   a) Basic Information on SOEPcore
   b) Composition and Sample Development of SOEPcore
   c) Topics covered in SOEPcore

3. Survey Methods and Instruments

4. Data Structure

5. Documentation and Online Services
A Short Introduction to Panel Data

- Panel data: repeated survey of multiple, identical units over multiple, identical points in time

<table>
<thead>
<tr>
<th>Year</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2011</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2012</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2013</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2014</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2015</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2016</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>
Panel data: repeated survey of multiple identical units over multiple, identical points in time.

<table>
<thead>
<tr>
<th></th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2011</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2012</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2013</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2014</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2015</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2016</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Cross-section: all units, one year
Panel data: repeated survey of multiple, identical units over multiple, identical points in time.

<table>
<thead>
<tr>
<th>Year</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2011</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2012</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2013</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2014</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2015</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
<tr>
<td>2016</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
<td>✔</td>
</tr>
</tbody>
</table>

Longitudinal/time series: one unit, all years
Panel data: repeated survey of multiple, identical units over multiple, identical points in time.

<table>
<thead>
<tr>
<th>Year</th>
<th>Unit 1</th>
<th>Unit 2</th>
<th>Unit 3</th>
<th>Unit 4</th>
<th>Unit 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2010</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2011</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2012</td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013</td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2014</td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>✔</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✔</td>
</tr>
</tbody>
</table>

Pooled cross-section: several units, several years.
Complete information analysis with an unbalanced panel

<table>
<thead>
<tr>
<th>t1</th>
<th>t2</th>
<th>t3</th>
</tr>
</thead>
<tbody>
<tr>
<td>successfully interviewed in all waves</td>
<td>new subjects/households in sample</td>
<td>dropouts</td>
</tr>
</tbody>
</table>

Advantages of Panel Data

Information from different life stages

• Identification of transition phases and important life events and their (causal) effects

• Measurement of phenomena as e.g. “time in poverty”

• Control of time-constant extraneous variables (approaching causality)

• Analysis of trends
(Mis-)Predicted Subjective Well-Being Following Life Events
Advantages of Panel Data: Example 1

Figure 1: Patterns in mean actual and predicted life satisfaction around life events

Note: Sample sizes are presented in Table T. Data source: SOEP.
Potential Disadvantages of Panel Data

- Selection into sample
  → Weights

- Distortion due to panel attrition
  → Weights

- Limited generalizability due to changes in the population
  → Refreshment samples

- Trade-off: Repetition of identical questions vs. adaptation to new topics (limited survey time)
  → Rotating modules
Introduction to the SOEP - Part I

1. A Short Introduction to Panel Data

2. What is the SOEP?
   a) Basic Information on SOEPcore
   b) Composition and Sample Development of SOEPcore
   c) Topics covered in SOEPcore

3. Survey Methods and Instruments

4. Data Structure

5. Documentation and Online Services
Introduction to the SOEP - Part I

1. A Short Introduction to Panel Data

2. What is the SOEP?
   a) Basic Information on SOEPcore
   b) Composition and Sample Development of SOEPcore
   c) Topics covered in SOEPcore

3. Survey Methods and Instruments

4. Data Structure

5. Documentation and Online Services
The Socio-economic Panel Study (SOEP)

SOEPcore
- 1984 - today
- 18 subsamples, regular refreshments

SOEP IS - Innovation Sample
- From 2012 onwards
- Subsamples E and I
- Opportunity for external users to propose survey modules

SOEP RS - Related Studies
- Question+ data structure of SOEP
  - Berlin Aging Study II (BASE II)
  - Families in Germany (FiD)
  - PIAAC-L
  - BRISE
**The Socio-economic Panel Study (SOEP)**

**SOEP Core**
- 1984 - today
- 18 subsamples, regular refreshments

**SOEP IS - Innovation Sample**
- From 2012 onwards
- Subsamples E and I
- Opportunity for external users to propose survey modules

**SOEP RS - Related Studies**
- Question+ data structure of SOEP
  - Berlin Aging Study II (BASE II)
  - Families in Germany (FiD)
  - PIAAC-L
  - BRISE
SOEP IS and Related Studies

SOEP-Core (1984-)
- 1984 Germans & foreigners West (A/B)
- 1990 East-Germans (C)
- 1994/95 immigrants (D)
1998 innovation (CAPI) (E)
- 2000 addition (F)
2002 high income respondents (G)
- 2006 addition (H)
2009 innovation (incentivizing) (I)
...

SOEP-IS
- Start in 2012
- 1998 innovation (E) (CAPI)
- 2009 innovation (I) (Incentivierung)
- 2012 addition
- 2013 addition
- 2014 addition

SOEP-Related Studies
- Families in Germany (FiD) 2010-2013
- BASE II (2012-2015)
- PIAAC-L (2013-2017),
- Bonn Intervention Study (BIP) (2014-ongoing)
- BRISE (2017-2020)
Basic Information on SOEPcore

Household Panel Study: Representative annual panel survey of private households in Germany

- Since 1984
- In 2018 about 15,000 households
- Information for up to three generations due to genealogical design
- Life-Course perspective with a focus on well-being
- Individual information on all individuals living in the sampled household

- Theory-based multidisciplinary mix of subjective and objective data on households, individuals, and families.
- Provision of micro-data for research in social and behavioral sciences as well as economics for universities and independent research centers in Germany and abroad
- Possibility of regional comparison and availability local context indicators (only at SOEP Research Data Center at DIW Berlin)
Introduction to the SOEP - Part I

1. A Short Introduction to Panel Data

2. What is the SOEP?
   a) Basic Information on SOEPcore
   b) Composition and Sample Development of SOEPcore
   c) Topics covered in SOEPcore

3. Survey Methods and Instruments

4. Data Structure

5. Documentation and Online Services
Composition of SOEP Core

Original Sample
(West (1984), former East (1990))

Refreshment Samples
(Refreshments)

Special Samples
(Immigration, Family Type)

Demographic Inflows
(births, marriages, adoptions, in-movers, out-movers)

Active Panel

Attrition
(Refusals)

Deaths

Moving Abroad
Information Gathered in SOEP Households

Household Panel Study with information on individuals within the households

- Information on **the household** is gathered from the household head
- Personal interviews with all household members above the age of 18 (since 1984)
  - Since 2001: All individuals age 17 with retrospective questions on youth
  - Since 2014: Personal interviews with children living the household aged 11-12
  - Since 2016: Personal interviews with children living in the household aged 13-14
- Proxy information for younger children living in the household
- Proxy information on individuals who dropped out of the survey due to death or moving abroad
- Temporary dropouts will be re-interviewed with an additional special short questionnaire
Introduction to the SOEP - Part I

1. A Short Introduction to Panel Data

2. What is the SOEP?
   a) Basic Information on SOEPcore
   b) Composition and Sample Development of SOEPcore
   c) Topics covered in SOEPcore

3. Survey Methods and Instruments

4. Data Structure

5. Documentation and Online Services
Topics Covered in the SOEP

Multi-topic study with a multidisciplinary perspective, covering questions concerning:

- Demography and population
- Education and qualification
- Occupation and employment
- Income, taxes and social security
- Housing
- Health
- Attitudes, values and personality
- Subjective wellbeing (Life satisfaction).
- Migration and integration

Additional topics with multi-annual survey rhythms
### Topics with Multi-Annual Survey Rhythm*

*Complete list of topics is available in the SOEPCompanion (2019): Companion.soep.de*
Dimensionality of Gathered Information

Information are gathered on a temporal dimension spanning Past, Present, and Future.

Subjective & Objective
Subjective and Objective Information

Subjective

Objective

Health and Illness

138. How would you describe your current health?
- Very good
- Good
- Satisfactory
- Poor
- Bad

139. How many hours do you sleep on average on a normal day during the working week? How many hours on a normal weekend day?
- Please give only whole hours
- Normal working day: 
- Normal weekend day: 

160. Has a doctor ever diagnosed you to have one or more of the following illnesses?
- Sleep disorder
- Diabetes
- Asthma
- Cardiac disease (also cardiac insufficiency, weak heart)
- Cancer
- Stroke
- Migraine
- High blood pressure
- Depression
- Dementia
- Joint diseases (including arthritis, neuritis)
- Chronic back trouble
- Burns
- Other illness
- No illness diagnosed

161. Do you have a health problem that limits you in normal everyday life?
- Yes, severely
- Yes, somewhat
- Yes, no
- No, not at all

162. Have you had this health problem for more than a year?
- Yes
- No

First Measurement right hand

Foto: C. Kurka
Temporal Dimension of Gathered Information

Information regarding the present:
• current employment status or current life satisfaction

Information regarding the past:
• Retrospective questions about events
  • How often have you changed jobs in the recent years?
• Retrospectively assessed history of events since age of 15
  • Employment or family status
• Monthly calendar of income and employment
  • Employment status from January to December of the past year
  • Monthly changes in education or family status since the last

Information regarding the future (expectations):
• Expected life satisfaction in 5 years
• Expectations regarding (re-)employment, retirement, expectations on schooling
Information from “cradle to stretcher”
→ Ideally for all individuals in the household
Subjective and Objective Information

Extension of survey instruments besides standardized questionnaires:

- Objective measure of physical characteristics
- Instruments measuring cognitive potential
- Behavioral experiments
  - Charitable giving experiment
  - Meeting Chancellor Merkel Intervention
- Innovative modules in SOEP-IS:
  - Jelly-Bear Test
  - Saliva Collection for Epigenetics
  - Variation of the “invisible gorilla”
  - Use of Smartphone Apps for real-life data collection
Psychological Measurement Instruments

Personality Traits

- Big Five Inventory (2005, 2009, 2013)
- Impulsiveness & Patience (2008, 2013)
- Personal Networks (2006, 2011)
- Self-esteem (2010, 2015)
- Loneliness (2013)
Cognitive Abilities: Adults

- 90 seconds test
- Part of the individual questionnaire
- Only for CAPI-respondents

<table>
<thead>
<tr>
<th>Test</th>
<th>2006</th>
<th>2012</th>
<th>2016</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal Naming Task</td>
<td>x</td>
<td>x*</td>
<td>x**</td>
</tr>
<tr>
<td>Symbol-Digit-Test</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Vocabulary test</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

* only for those who answered already in 2006,
** only for those who answered already in 2012

Anger & Heineck (2009): Do Smart Parents Raise Smart Children?: The Intergenerational Transmission of Cognitive Abilities, SOEPpapers #156
Hier haben wir ein kleines Lösungsbeispiel:

Aufgabe:
Zeichen: )  ⊕  ⊖  
Zahl? ? ? ?

Lösung:
Zeichen: )  ⊕  ⊖  
Zahl? 2 1 4

Zu dem Zeichen in Feld 1 ist "2" einzugeben.
Zu dem Zeichen in Feld 2 ist "1" einzugeben.
Zu dem Zeichen in Feld 3 ist "4" einzugeben.

Welches ist das richtige Wort?
- kurehen
- krusienen
- kasseren
- kursieren
- kustieren
„Denksport Jugend“ (DJ) for 16-17 year olds

- Annually since 2006
- Incentivized intelligence-structure test (I-S-T 2000 R) (Amthauer et al. 2001)
  - ca. 30 minutes
  - 90% participation rate

**Dimensions**

- Verbal-cognitive potential (word analogies)
- Numeric-cognitive potential (numeric sequences)
- Figural-cognitive potential (matrices and pictures)
- Logic thinking (= sum index)

Beispiel:

**Wald : Bäume = Wiese : ?**

Gras... X  Heu... ☐  Futter... ☐  Grün... ☐  Weide... ☐

Hier sind Oberbegriffe und Unterbegriffe gegenüber gestellt. Für den „Wald“ sind „Bäume“, was für die „Wiese“ die „Gras“ sind. Kreuzen Sie also Gras an.

Ein weiteres Beispiel:

**dunkel : hell = nass : ?**

Regen... ☐  Tag... ☐  feucht... ☐  Wind... ☐  trocken... X

Da „dunkel“ das Gegenteil von „hell“ ist, muss zu „nass“ auch das Gegenteil gefunden werden. Hier wird also trocken gekreuzt.

Beispiel 1:  
6 + -  -  -  2 + -  -  3 = 5
Hier muss man die Zahlen 6 und 2 zusammenzählen, dann vom Zwischenergebnis 8 die Zahl 3 abziehen, um das vorgegebene Ergebnis 5 zu erhalten:

(6 + 2 = 8; 8 - 3 = 5)
Lösung:  
6 + -  -  -  2 + -  -  3 = 5

Beispiel 2:  
7 + -  -  -  2 + -  -  4 = 10
Hier sind die Zahlen 7 und 2 zu nehmen, das Zwischenergebnis ist 14, von dem muss man die Zahl 4 abziehen, um das Ergebnis 10 zu erhalten:

(7 - 2 = 14; 14 - 4 = 10)
Lösung:  
7 + -  -  -  2 + -  -  4 = 10
Introduction to the SOEP - Part I

1. A Short Introduction to Panel Data

2. What is the SOEP?
   a) Basic Information on SOEPcore
   b) Composition and Sample Development of SOEPcore
   c) Topics covered in SOEPcore

3. Survey Methods and Instruments

4. Data Structure

5. Documentation and Online Services
Development of the Survey Method

Documentation of Questionnaires

All printed questionnaires are available in quotable form as SOEP Survey Papers.

<table>
<thead>
<tr>
<th>Year</th>
<th>2017</th>
<th>2016</th>
<th>2015</th>
<th>2014</th>
<th>2013</th>
</tr>
</thead>
<tbody>
<tr>
<td>1987</td>
<td>1986</td>
<td>1985</td>
<td>1984</td>
<td>Additional</td>
<td></td>
</tr>
</tbody>
</table>
Questionnaires

Standard questionnaires
- Individual Questionnaire | PDF, 1 MB (English and German)
- Household Questionnaire | PDF, 0.9 MB (English and German)
- Youth Questionnaire | PDF, 0.64 MB (German only)
- Supplementary Biography Questionnaire | PDF, 1.13 MB (German only)
- Catch-Up Individual Questionnaire | PDF, 615.85 KB (German only)
- Pre-Teen (11-12 year olds) | PDF, 0.87 MB (German only)
- Early Youth (13-14 year olds) | PDF, 0.87 MB (German only)
- The deceased Individual | PDF, 446.06 KB (German only)
- Household Questionnaire (English and German)
- Individual with Biography (English and German)
- Individual Questionnaire (continued sample) (English and German)
- Youth Questionnaire (German only)
- Samples A-L3

Biographical questionnaires
- Supplementary Biography Questionnaire | PDF, 387.12 KB
- Household (PAPI) | PDF, 242.01 KB
- Catch-Up Individual | PDF, 173.77 KB
- Biography | PDF, 251.15 KB

Mother-child instruments
- Mother-Child-Questionnaire (Newborns) | PDF, 0.65 MB (German only)
- Mother-Child-Questionnaire (2-3 year olds) | PDF, 0.7 MB (German only)
- Mother-Child-Questionnaire (5-6 year olds) | PDF, 0.6 MB (German only)
- Parents Questionnaire (7-8 year olds) | PDF, 0.58 MB (German only)
- Mother-Child-Questionnaire (9-10 year olds) | PDF, 0.7 MB (German only)

Pupils/ early youth

Deceased person

Target-group-specific questionnaires (migrants & refugees)

For a translation of the Mother-Child instruments, please refer to the generated questionnaires with reference to variables (see below)

With reference to variables

Standard and Biographical Survey Instruments

- Standard instruments (annually)
  - Household questionnaire (household head)
    - Time-variant household information
  - Individual questionnaire (every person in household aged > 17/18)
    - Time-variant individual information

- Biographical instruments (once, during first interview)
  - Supplementary Biography Questionnaire (aged 18 and above, since 1989)
    - Retrospective questions regarding important life events and biographical data (e.g. place of birth, nationality, nationality of parents)
### Target-group-specific Instruments

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2-3 years (Muki2)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2002</td>
</tr>
<tr>
<td>5-6 years (Muki3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2002</td>
</tr>
<tr>
<td>7-8 years (FID)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2002</td>
</tr>
<tr>
<td>9-10 years (Muki5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2002</td>
</tr>
<tr>
<td>11-12 years* (pre-teen)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2002</td>
</tr>
<tr>
<td>13-14 years* (early youth)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2002</td>
</tr>
<tr>
<td>17 years** (youth)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>1983</td>
</tr>
</tbody>
</table>

Starting with birth-cohort:
- Newborns 2002/2003
- 2-3 years (Muki2) 2002
- 5-6 years (Muki3) 2002
- 7-8 years (FID) 2002
- 9-10 years (Muki5) 2002
- 11-12 years* (pre-teen) 2002
- 13-14 years* (early youth) 2002
- 17 years** (youth) 1983

* Personally interviewed, with parental consent
** personally interviewed
3 Target-group-specific Instruments

- „Luecke“ (from 2006 onwards)
  - Temporary unit non-response
- The deceased person (from 2009 onwards)
  - Family relations, cause of death, testament
  - Reasons for moving abroad, reasons to return
- IAB-SOEP-migration sample (since 2013 and 2015) (M1-M2)
- IAB-BAMF-SOEP-survey of refugees (since 2016 and 2017) (M3-M5)
Introduction to the SOEP - Part I

1. A Short Introduction to Panel Data
2. What is the SOEP?
   a) Basic Information on SOEPcore
   b) Composition and Sample Development of SOEPcore
   c) Topics covered in SOEPcore
3. Survey Methods and Instruments
4. Data Structure
5. Documentation and Online Services
When you get the data

Version number: corresponds to survey waves
We provide 4 Types of Data

- Tracking data
- Original survey data
- Generated variables
- Survey Variables
General naming conventions for datasets in SOEPcore

The SOEP provides “speaking” names for datasets, which contain hints with regards to:

- Survey unit
- Type of data
- Underlying survey Instruments
- Wave (for Wide Format in raw)
General naming conventions for datasets in SOEPcore

Subject of inquiry
- p: person
- h: household
- kid: children
- jugend: youth
- vp: deceased person

Type of data
- Original:
- Generated: *gen, bio*, *kalen, equiv, *spell, kal,
- Tracking: *brutto, *path, pbr_exit,
- Survey: *hrf, design, exit, pbr_hhch,

Underlying survey instruments
- luecke, abroad,
- cog*, gripstr, timepref, trust, …
### Biography Data Sets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>BIOAGE17</td>
<td>Information from youth questionnaires (age 17)</td>
</tr>
<tr>
<td>BIOAGEL</td>
<td>Information from mother-child questionnaires (age 0 - 10)</td>
</tr>
<tr>
<td>BIOPUPIL</td>
<td>Information from pupil and pre-teen questionnaires (age 12-14)</td>
</tr>
<tr>
<td>BIOEDU</td>
<td>Educational biographies (since 2010)</td>
</tr>
<tr>
<td>BIOIMMIG</td>
<td>Migrant information</td>
</tr>
<tr>
<td>BIOJOB</td>
<td>First job</td>
</tr>
<tr>
<td>BIOBIRTH</td>
<td>Births biographies for mothers</td>
</tr>
<tr>
<td>BIOPAREN</td>
<td>Information on mothers and fathers</td>
</tr>
<tr>
<td>BIOSOC</td>
<td>Socialization info– life course</td>
</tr>
<tr>
<td>BIORESID</td>
<td>Residence biographies</td>
</tr>
<tr>
<td>BIOTWIN</td>
<td>Information on twins, triplets and quadruplets (since 2006)</td>
</tr>
<tr>
<td>BIOSIB</td>
<td>Sibling information (since 2010)</td>
</tr>
<tr>
<td>BIOCOUPL(M/Y)</td>
<td>Relationship Biographies (month/year)</td>
</tr>
<tr>
<td>BIOMARS(M/Y)</td>
<td>Marriage Biographies (month/year)</td>
</tr>
</tbody>
</table>

Some are updated annually, others contain time-constant information.
Cumulative data sets

- **PWEALTH**  assets person (every 5 years since 2002)
- **HWEALTH**  assets households (every 5 years since 2002)
- **HEALTH**    health person (biannually since 2002)
- **PFLEGE**    in-need-of-care person (annually 1985 to present)
- **GRIPSTR**   gripping force test (biannually since 2006)

Spell data sets

- **PBIOSPE**  employment biography – 15+
- **ARTKALEN** labor status calendar
- **EINKALEN** income calendar
- **BIOMARS(Y/M)** marriage events (month)
- **BIOCOUPL(Y/M)** marriage events (year)
- **SOZKALEN** social welfare calendar (household level)
- **REFUGSPELL** information on flight history and reasons
- **MIGSPELL**  information on migration history and reasons
418 SOEP Survey Papers
Series D - Variable Descriptions and Coding

SOEP-Core v32 - Documentation on Biography and Life History Data

49 SOEPcampus http://hdl.handle.net/10419/155351
Conventions for variable names in SOEPcore

1st digit: Dataset-identifier
P = Person, H = Household,
L = Biography, J = Youth

2nd digit: Variable-identifier
L = numeric variable,
A = alphanumeric variable
C = numeric variable containing original information for recoded variables

3rd digit: Topic-identifier
a = demography and population
b = work and employment
c = income, taxes and social security
d = family and social networks
e = health and care
f = habitation, equipment and services of private households
g = education and qualification
h = attitudes, values and personality
i = time use and ecological conduct
j = integration, migration and transnationalization
k = survey methodology

4th-7th digit: run-on number

Examples:
ple0010 person numeric health & care 0010 (birthyear)
pli0044 person numeric time use 0044 (time spent on childcare weekday)
contains basic information on all individuals, who ever lived in a SOEP household including ...
- identifiers (pid, hid, cid)
- sex, year & month of birth, migback
- sample (psample)
- survey status for each year (netto)
- household type (pop)
- sample region (psamp)
- ...

we recommend using the netto variable to identify your study population

contains time consistent and time variant information on household level including
- current and root household number (hid, cid)
- sample regions (psamp)
- response status
- ...
### Documentation for PPATH (v.33.1) - SOEP Survey Papers 415

<table>
<thead>
<tr>
<th>Name</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>cid</td>
<td>Case-ID, Original Household Number</td>
</tr>
<tr>
<td>pid</td>
<td>Never Changing Person ID</td>
</tr>
<tr>
<td>sex</td>
<td>Gender</td>
</tr>
<tr>
<td>gebjahr</td>
<td>Birth Year, 4-digits</td>
</tr>
<tr>
<td>eintritt</td>
<td>Year First Contacted, Netto=10-99</td>
</tr>
<tr>
<td>erstbefr</td>
<td>Year First Surveyed, Netto=10-99</td>
</tr>
<tr>
<td>austritt</td>
<td>Year Of Last Contact, Netto=10-99</td>
</tr>
<tr>
<td>psample</td>
<td>Sample Member</td>
</tr>
<tr>
<td>letzbef</td>
<td>Year Of Last Survey, Netto=10-99</td>
</tr>
<tr>
<td>todjahr</td>
<td>Year Died, 4 Digits</td>
</tr>
<tr>
<td>todinfo</td>
<td>Year Died, Information Source</td>
</tr>
<tr>
<td>immimmigr</td>
<td>Year Moved to Germany</td>
</tr>
<tr>
<td>gebbornc</td>
<td>Born in Germany</td>
</tr>
<tr>
<td>corigin</td>
<td>County Born In</td>
</tr>
<tr>
<td>gebmonat</td>
<td>Month Of Birth</td>
</tr>
<tr>
<td>gebmonval</td>
<td>Month Of Birth, Data Source</td>
</tr>
<tr>
<td>loc1989</td>
<td>Where did you live in 1989?</td>
</tr>
<tr>
<td>migback</td>
<td>Migration background</td>
</tr>
<tr>
<td>miginfo</td>
<td>Migback: Quality of Information</td>
</tr>
<tr>
<td>gerbormig</td>
<td>Germborn: Quality of Information</td>
</tr>
<tr>
<td>coriginm</td>
<td>Corigin: Quality of information</td>
</tr>
<tr>
<td>immiyearm</td>
<td>Immimyear: Quality of Information</td>
</tr>
<tr>
<td>sexor</td>
<td>Sexual Orientation</td>
</tr>
<tr>
<td>sexorinfo</td>
<td>Sexual Orientation: Source of information</td>
</tr>
<tr>
<td>loc1989m</td>
<td>Loc1989: Source / Quality of Information</td>
</tr>
<tr>
<td>birthregm</td>
<td>Birth place: German Federal Land</td>
</tr>
<tr>
<td>arefback</td>
<td>Refugee Experience</td>
</tr>
<tr>
<td>arefinfo</td>
<td>arefback: Source of Information</td>
</tr>
</tbody>
</table>
Current Wave Survey Status (netto)

<table>
<thead>
<tr>
<th>respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>25</td>
</tr>
<tr>
<td>...</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>children</th>
</tr>
</thead>
<tbody>
<tr>
<td>89</td>
</tr>
<tr>
<td>90</td>
</tr>
<tr>
<td>91</td>
</tr>
<tr>
<td>92</td>
</tr>
<tr>
<td>93</td>
</tr>
<tr>
<td>94</td>
</tr>
</tbody>
</table>

Etc.
## Conventions for Missing Values in SOEP

<table>
<thead>
<tr>
<th>Code</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1</td>
<td>No reply</td>
</tr>
<tr>
<td>-2</td>
<td>Does not apply</td>
</tr>
<tr>
<td>-3</td>
<td>Not valid</td>
</tr>
<tr>
<td>-4</td>
<td>Invalid multiple reply</td>
</tr>
<tr>
<td>-5</td>
<td>Not available version of questionnaire</td>
</tr>
<tr>
<td>-6</td>
<td>Questionnaire with different filtering</td>
</tr>
<tr>
<td>-8</td>
<td>Question not part of questionnaire</td>
</tr>
</tbody>
</table>

Note: In SOEP no SYSTEM MISINGS („.“ in Stata)
Combining information from different datasets can be relevant for various types of research questions.

Examples:

- How does the birth of a third child affect the household income?
  - Household income from hgen or pequiv – information in births from biobirth

- What is the correlation between length of unemployment, life satisfaction and trust in democratic institutions?
  - Employment spells from artkalen – life satisfaction and trust in democratic institutions from p

If the relevant information is part of different datasets (and different levels of inquiry), they need to be merged via identifiers.
Identifiers

To combine information from different datasets and levels

- **PID**  person, invariant over time
- **HID**  current household ID, variant over time
- **CID**  root household ID, constant over time
- **SPELLNR**  spell number, in spell data
- **INTNR**  number of interviewer in data set INTVIEW
- **SYEAR**  survey year
- Regional identifiers  e.g. federal states, official district codes
For other questions, it might be necessary to combine information from different individuals.

Examples:
- How does the divorce of a child affect mental health of the parents?
- How does entering unemployment affect the grades of the children?

Information from different persons can be combined via pointers
- Certain datasets contain pointers to partners, parents or children (in addition to the identifiers for the anchor person)
How to combine the data depends on the question, e.g.:

- Analysis from the child’s perspective: Does the employment status of the mother affect the child’s education success?
  \[ \text{persnr} = \text{child} \]

- Analysis from the mother’s/parents’ perspective: Does the birth of a child affect labor supply from women?
  \[ \text{persnr} = (\text{potential}) \text{ mother} \]

- Analysis from women’s perspective: Does the partner affect the labor supply of women?
  \[ \text{persnr} = \text{woman with partner} \]

- Survey Unit which determines the **dependent variable** is the anchor person which **keeps HID and PID**

- Pointers are used to merge information on other individuals to the anchor person
## Pointers in the SOEP

<table>
<thead>
<tr>
<th>Anchor Person</th>
<th>Merged Person</th>
<th>Dataset</th>
<th>Variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
<td>Partner</td>
<td>pgen</td>
<td>pgpartz</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ppfadl</td>
<td>parid</td>
</tr>
<tr>
<td>Mother</td>
<td>Child</td>
<td>biobirth</td>
<td>kidnrXX</td>
</tr>
<tr>
<td>Child &lt; 17</td>
<td>Mother</td>
<td>kidl</td>
<td>k_pmum</td>
</tr>
<tr>
<td>Child &lt; 17</td>
<td>Partner of the mother</td>
<td>kidl</td>
<td>k_pmump</td>
</tr>
<tr>
<td>Child &gt; 17 / Adults</td>
<td>Mother</td>
<td>bioparen</td>
<td>mnr</td>
</tr>
<tr>
<td>Child &gt; 17 / Adults</td>
<td>Partner of the mother</td>
<td>bioparen</td>
<td>fnr</td>
</tr>
<tr>
<td>Person</td>
<td>Sibling</td>
<td>biosib</td>
<td>sibpnrxXX</td>
</tr>
<tr>
<td>Person</td>
<td>Twin, Triplet, Quadruplet</td>
<td>biotwin</td>
<td>pnrtwin / pnrttrip / pnrquad</td>
</tr>
</tbody>
</table>
Combining information with Pointers – An Example

How is mothers’ labor market participation affected by income changes of their partners?

1. Create a dataset of females with children containing the pointer to partners:
   
   - DepVar: Mothers’ labour market participation → PID with mothers
     
     ```
     use pid hid syear netto parid using ppathl.dta
     merge 1:1 pid hid syear using biobirth.dta, keepus(kidnr01)
     keep(3) nogen
     merge 1:1 pid hid syear using pgen.dta, keepus(pglfs pgstib)
     keep (1 3) nogen
     save $temp\mothers.dta
     ```

2. Prepare dataset for “partners” containing the relevant information and rename PID for the partner dataset

   ```
   use pid hid syear netto pglabgro using pgen.dta
   rename pid parid
   save $temp\partners.dta
   ```

3. Merge datasets keeping only merged cases

   ```
   use $temp\mothers.dta, clear
   merge 1:1 parid syear using $temp\partners.dta, keepus(pglabgro) keep (3)
   ```
In welcher Form liegen die Daten vor?

- Equivalent information but in wave-specific data sets
- $ = wave token

<table>
<thead>
<tr>
<th>wave token</th>
<th>year</th>
<th>wave nr.</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1984</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>1985</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>bg</td>
<td>2016</td>
<td>33</td>
</tr>
<tr>
<td>bh</td>
<td>2017</td>
<td>34</td>
</tr>
</tbody>
</table>
Variable names in wide – v.34 onwards

**Wave-identifier**
1-2 letters
- e.g. C = 1987, BG = 2016

**Unit-identifier**
1 letter
- e.g. H = Household, P = Person, K = child, J = Youth

**Question-number**
2-3 numbers, leading zero if < 10
- e.g. 001, 24, 157

**Item-identifier**
2 numbers with leading zeros
- e.g. 01, 04, 15

**Questionnaire-identifier**
2 numbers
Defined in instrument

**Examples:**
- \( \text{AP}_06 \) wave A (1984) person question 6
- \( \text{BEK}_78_10 \) wave BE (2014) child question 78 item 10
- \( \text{BHP}_109_01_q75 \) wave BH (2017) person question 109 item 1 questionnaire 57 *

* Personal Biography Q. (M3-M4 Re-Interviewed; CAPI)
Introduction to the SOEP - Part I

1. A Short Introduction to Panel Data

2. What is the SOEP?
   a) Basic Information on SOEPcore
   b) Composition and Sample Development of SOEPcore
   c) Topics covered in SOEPcore

3. Survey Methods and Instruments

4. Data Structure

5. Documentation and Online Services
Important sources for working with the SOEP

- SOEPcompanion (link)
- Questionnaires (link)
- SOEPhelp (link)
- Documentation of generated variables (link)
- Paneldata.org (link)
- SOEPtutorials (link)
Welcome to the SOEPcompanion!

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

The SOEPcompanion provides data users with the content of the currently released SOEP-Core data (v34) and will introduce into different versions of SOEP-Core data structures. It will also provide data users with a lot of retrievals in Stata as well as multiple instructions on how to use our various documentation services. This collection of information in the SOEPcompanion is intended to become a yearly updated main reference guide, and a practical companion in basic understanding and implementation of the SOEP.
SOEP Core Data Distribution 1984-2017 (soep.v34)

SOEP-Core version 34

- Short description
- Availability
- Online documentation and updates
- Questionnaires

Title: Socio-Economic Panel (SOEP), data from 1984-2017

DOI: 10.6884/soep.v34

Collection period: 1984-2017
Publication date: 2019-03-05
Principal investigators: Stefan Liebig, Jan Goebel, Carsten Schroder, Jurgen Schupp,
5 Documentation of Generated Variables

Generated variables (wave specific v33.1)
- Generated variables (wave specific v33.1)
- Generated variables not updated
- Previous versions

SOEP Survey Paper 483 (Series D):
- SOEP-Core v33.1 - SPGEN | PDF, 258.82 KB
  (Documentation of Person-related Status and Generated Variables)

SOEP Survey Paper 487 (Series D):
- SOEP-Core v33.1 - PFPAD | PDF, 174.07 KB
  (Documentation of the Person-related Meta-dataset)

SOEP Survey Paper 482 (Series D):
- SOEP-Core v33.1 - HPGEN | PDF, 190.96 KB
  (Documentation of the Household-related Status and Generated Variables)

SOEP Survey Paper 484 (Series D):
- SOEP-Core v33.1 - HPFAD | PDF, 107.06 KB
  (Documentation of the Household-related Meta-dataset)
SOEPhelp – a STATA ado file

net install soephelp, replace from(http://companion.soep.de/SOEPhelp/)

SOEPhelp

soephelp — is a convenient way to display basic information and documentation of soep datasets.

Syntax

soephelp [varname] [, options]

options  Description

en       displays information in English (if available)
de       displays information in German (if available)

Standard is defined by label language [D] label_language.

Description

soephelp is used to display documentation of the dataset in use and its variables via the stata viewer.
varname is optional, if no variable is specified soephelp displays basic information about the dataset. For example what instruments or datasets were used to assemble this dataset, version of the dataset and its basic contents.

If a variable is specified, the documentation of this variable is displayed. This generally includes information about variable origins, e.g. questiontext and/or item text, questionnaires, question number, corresponding soep-long variables as well as links to further documentation such as paneldata.org or the SOEPcompanion.

soephelp is currently implemented for soep-core and soep-long datasets.

Example

Open dataset documentation of current dataset.

soephelp
In our media-center and on youtube
Overview of all studies of the SOEP and other panels
All data sets and increasingly also all questionnaires
5 Working with the Basket

Basket: demo

- Title: demo
- Study: soep-core

List of scripts
- CREATE A NEW SCRIPT

Data for years 1984 to 1993:
- Variables: satisfaction with life at today
- Add all, Remove all

Data variables:
- ap8801
- bp8301
- cp9601
- cp8901
- dp89
- lp108
- gp8401
- hp10901
- ip10901
- gp109
Thank you for your attention.