Introduction to the German Socio-Economic Panel (SOEP)

Jan Goebel

IZA Red Cube Seminar
May 20, 2009
http://www.diw.de/soep
Roadmap

- Introduction
- Sub-samples and Survey Related Issues
- Sample Size Development
- Content
- Data Access
- Innovations
Household Panel Data

• The Data
  • Survey of private households and persons in Germany (official title „Living in Germany“)
  • Annual Interviews since 1984; 25th (!) wave in 2008
  • Started with 5,921 Households in 1984 (2007: 11,689) and 12,245 persons (2007: 21,234)

• Institutional Facts
  • Run by social scientists, totally research driven!
  • Located: department SOEP at DIW Berlin; institutionalized funding since 2003 (WGL institute)
  • Field-work agency: TNS Infratest Sozialforschung
What are we measuring?

- Representative micro-data on persons, **households**, families
- **Objective** (e.g. income) and **Subjective** (e.g. satisfaction) indicators
- Measure **Stability** and **Change** in Living Conditions
- **Coverage of all life-stages**
  - Observation unit “household” guarantees appropriate coverage from birth to death
- **Effects on third persons**
  - Effect of divorce not only on separated partners but also children.
- **Joint decisions**
  - Occupation of partner affects labor supply
  - Family diet and individual health.
- **Dyadic analysis**
  - Intergenerational research (mobility)
  - Sibling estimates, twin studies

Introduction
Well-known advantages of panel data

• **Heterogeneity**
  • Analysis of intra-individual change vs inter-individual difference
  • Controlling for (inter-individual) unobserved heterogeneity.

• **Causality**
  • Controlling for temporal order of events and behavior better suited for the identification of causal effects than cross-sectional data.
Figure 2
Adaptation to Good and Bad Events

Life Satisfaction

Year

-5 -4 -3 -2 -1 0 1 2 3 4 5

Year of Event

- Widowhood
- Divorce
- Unemployment
- Marriage

Diener et al. 2006, American Psychologist, 61(4), 310
Opportunities for long-running panels

- Cumulation of rare events by pooling (mortality, divorce, ...)
- Increasing coverage of changes in institutional settings and the potential impact on individual behavior
- Comparison of “intentions” and actual behavior (how relevant are expectations at the individual level?)
- Intergenerational analysis (linking parents and children)
- Increasing potential for cohort analysis
- Improved efficiency in statistical modeling due to increased number of observations per individual
Potential Caveats

- Bias due to selective attrition
- Panel effects (incl. *positive* learning effects !!!)
- Representative coverage of changes in underlying population due to immigration
- “Compensation rules” for long-term shrinking number of observations
- Continuity of questioning (phrasing, modes, …)
  - inter-temporal comparability of indicators vs.
  - need for adjustments to cover institutional changes

→ increasing with panel duration!
Introduction

Sub-samples and Survey Related Issues

Sample Size Development

Content

Data Access

Innovations
Sub-Samples

- **A "West-German" residents**
  - started in **1984**, n=4528 households
  - Head is either German or other nationality than those in Sample B.

- **B "Foreigners"**
  - started in **1984**, n=1393 households
  - Head is either Turkish, Italian, Spanish, Greek, or Yugoslavian (oversampling)

- **C "East-Germans"**
  - started in **1990**, n=2179 households.
  - Head was a citizen of the GDR. (expansion of survey territory, oversampling)

- **D "Immigrants"**
  - started in **1994/95**, n=522 households
  - At least one HH member has moved to Germany after 1984. (expansion of survey population)

- **E "Refreshment sample"**
  - started in **1998**, n=1067 households
  - Random sample covering all existing subsamples (total population).

- **F "Innovation sample"**
  - started in **2000**, n=6052 households.
  - Random sample covering all existing subsamples (total population).

- **G "High Income sample"**
  - started in **2002**, n=1224 households.
  - Monthly net Household income > 7.500 DM (oversampling)

- **H "Refreshment sample"**
  - started in **2006**, n=1506 households.
  - Random sample covering all existing subsamples (total population)
Methodology-Mix with standardized instruments

- Face-to-face individual interviews with all HH-members 16+
  - Household interview with "head of household"
  - Paper-and-pencil interviews (samples A through D, E1, F, G)
  - No proxy interviews / (almost) no phone interviews
  - Self-Completers: Data Agency resolves inconsistencies

- Since 1998 stepwise implementation of CAPI

- prospective: self-admin. Interviews via Internet (Test in 2004)
Development of the Interview-Mode

Samples and Survey Development of the Interview-Mode

PAPI □ CAPI □ SAQ (with Interviewer) □ SAQ (without Interviewer) □ SAQ (mail)
Containing general information (filled in by interviewer)

⇒ on households (e.g. size, housing area, regional information); on individuals (e.g. sex, year of birth, relation to head)

⇒ on the process of field work (e.g. number of contacts, reason for drop-outs, interview mode)
The SOEP Questionnaires

The Standard Instruments

• Household questionnaire (annual)
• Personal questionnaire (annual, all 17+)
• Personal „gap“ questionnaire (temp. drop-outs)
• Biography questionnaire (second interview only)

Special Populations

• Sub-sample-specific questionnaires for “Germans”, “Foreigners”, and “East Germans” (until 1995)
• Age-specific questionnaires: Age1 (since 2003), Age 2 (since 2005), Age 5 (since 2008), Age 12 (prospective 2013), Age 17 (since 2001)
• Interviewer questionnaire (since 2006)
• Event-triggered questionnaires planned
Survey Instruments for capturing the life course

- Conception Phase
- Fetal Phase
- Childhood
- Teenage
- Adult Life
- Memories & widow(er) Pensions

- Individual Questionnaire Parents
- Questionnaire Mother & Child
- Household & age specific Quest.
- Teenager Questionnaire
- Individual Questionnaire
- Individual Quest. Partner and Children („Exit Intv.“)

Samples and Survey
Introduction
Sub-samples and Survey Related Issues
**Sample Size Development**
Content
Data Access
Innovations
### Starting Sample Size in Wave 1

<table>
<thead>
<tr>
<th>Sample</th>
<th>Starting Year</th>
<th>Households</th>
<th>Respondents</th>
</tr>
</thead>
<tbody>
<tr>
<td>A and B</td>
<td>1984</td>
<td>5,921</td>
<td>12,245</td>
</tr>
<tr>
<td>C</td>
<td>1990</td>
<td>2,179</td>
<td>4,453</td>
</tr>
<tr>
<td>D1/D2</td>
<td>1995</td>
<td>522</td>
<td>1,078</td>
</tr>
<tr>
<td>E</td>
<td>1998</td>
<td>1,067</td>
<td>1,923</td>
</tr>
<tr>
<td>F</td>
<td>2000</td>
<td>6,052</td>
<td>10,890</td>
</tr>
<tr>
<td>G</td>
<td>2002</td>
<td>1,224</td>
<td>2,671</td>
</tr>
</tbody>
</table>
• All persons in HH are to be surveyed also the following years.

• At same address as well as after a residential move within Germany (→ regional mobility)

• Personal interviews at age of 16 (→ demographic development)

• Persons moving into an existing SOEP household.

• Since 1989 these (new) persons are also followed in case of leaving the household. This had not been the case up to wave 5 (1988).
• Principle: follow until two consecutive temporary drop-outs of all household members or a final refusal.

• Gaps: small questionnaire including questions on central information which is missing for the year of the drop-out (e.g. employment status).
The emergence of new households

<table>
<thead>
<tr>
<th>Persons</th>
<th>Old</th>
<th>New</th>
</tr>
</thead>
</table>
| Old     | • “classic case” without change of address  
       | • entire household moves                | • Move-out                               |
| New     | • Birth                                  | • Birth                                  |
|         | • Move-in                                | • Caused by split-offs of old persons from old households* |

* Remember that households *new* to the SOEP may already have existed before contacting the survey
• Persons exit by:
  • Death
  • Moving abroad

• Persons enter by:
  • Birth
  • Moving into a SOEP household from somewhere else in Germany or from abroad
  • Reaching age of 16 years (minimum respondents age is given by the calendar year, in which a person turns 17 years of age)
  • Split-offs of at least one *old* person from an *old* household
1. Making a successful contact to a given household
2. Realizing a successful interview

- Social groups typically associated with problems in respect to re-contacting and re-interviewing:
  - Single person households
  - Mobile households and persons
  - Young adults leaving parental home
“Panel care“ – keep interviewees involved

- Announcement by mail 2 weeks prior to interview
- For each successful interview, any respondent
  - receives a gift related to the yearly topical module (until 2007)
  - ticket for monthly nationwide lottery
  - since 2008: 5 € ex-ante incentive (until 2007: 1,50 € ex-post)
- Addresses are kept up to date by the field work agency
- Households receive the brochure „Living in Germany“ and information about data privacy regulations
- After interview (during summer): thank you-letter and „porto-card“
- Special treatments:
  - contact via phone
  - central case-by-case treatment for “problematic” households
  - mailing more information on request
  - website “Leben in Deutschland”
- The interview situation (face-to-face) ensures a personal
  - relationship, which makes it harder to withdraw from the survey.
  - Thus, the stability of the interviewer over time is very crucial.

Sample Size Development
The SOEP sample 1984 - 2010

Sample Size Development

Year: 84 85 86 87 88 89 90 91 92 93 94 95 96 97 98 99 00 01 02 03 04 05 06 07 08 09 10

Households

Set-up phase

Inclusion of new populations

Upgrading of sample size

Consolidation and innovations

Sample Size Development
Cross-sectional perspective
⇒ Number of successfully interviewed households by sample

Sample Size Development
Cross-sectional perspective

Number of successfully interviewed persons by sample
Longitudinal perspective: Samples A+B since 1984

Whereabout of the 16252 Persons

- Moved abroad
- Deceased
- Under the age of 16
- With interview
- Temporary drop-out
- Declined to reply
- No contact

Sample Size Development
Incidence of New households after Wave 1

Household records in the SOEP: Wave 1 and Split-offs

<table>
<thead>
<tr>
<th>Sample (Starting Year)</th>
<th>Number of observations</th>
</tr>
</thead>
<tbody>
<tr>
<td>A (1984)</td>
<td>4528</td>
</tr>
<tr>
<td>B (1984)</td>
<td>3124</td>
</tr>
<tr>
<td>C (1990)</td>
<td>1393</td>
</tr>
<tr>
<td>D (1994/95)</td>
<td>2179</td>
</tr>
<tr>
<td>E (1998)</td>
<td>1438</td>
</tr>
<tr>
<td>F (2000)</td>
<td>1098</td>
</tr>
<tr>
<td>G (2002)</td>
<td>1506</td>
</tr>
<tr>
<td>H (2006)</td>
<td>636</td>
</tr>
</tbody>
</table>

As of 2006: Wave 23 (A and B), Wave 17 (C), Wave 12/13 (D1/D2), Wave 9 (E), Wave 7 (F), Wave 5 (G), Wave 1 (H).

Sample Size Development

Sample Size Development
Introduction
Sub-samples and Survey Related Issues
Sample Size Development
Content
Data Access
Innovations
• Questions about a point of time (present)
  e.g. current employment status or current levels of satisfaction

• Retrospective questions on certain events (past)
  e.g. how often did you change your job during the last ten years?

• Retrospective life event history (past)
  e.g. employment or marital history

• Monthly calendar on income and labor market (past)
  e.g. employment status January through December last year

• Questions concerning future prospects (future)
  e.g. satisfaction with life five years from now, or job expectations
Core Questions

➤ Replication every year

- Population and demography
- Education, training, and qualification
- Labor market and occupational dynamics
- Earnings, income and social security
- Housing
- Health
- Household production
- Basic orientation (preferences, values, etc.)
- Satisfaction with life in general and various aspects
### Topic Modules

Replication every ~ 5-10 years

<table>
<thead>
<tr>
<th>Year</th>
<th>Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>1984</td>
<td>Employment biography</td>
</tr>
<tr>
<td>1985</td>
<td>Marriage and family biography</td>
</tr>
<tr>
<td>1986</td>
<td>Social origins (Bio), neighborh.</td>
</tr>
<tr>
<td>1987</td>
<td>Social security, early retirement</td>
</tr>
<tr>
<td>1988</td>
<td><strong>Assets</strong></td>
</tr>
<tr>
<td>1989</td>
<td><strong>Further education or training</strong></td>
</tr>
<tr>
<td>1990</td>
<td>Time use and preferences</td>
</tr>
<tr>
<td>1991</td>
<td>Family and social services</td>
</tr>
<tr>
<td>1992</td>
<td>Social security and poverty</td>
</tr>
<tr>
<td>1993</td>
<td><strong>Further education or training</strong></td>
</tr>
<tr>
<td>1994</td>
<td>Neighborhood, values</td>
</tr>
<tr>
<td>1995</td>
<td>Time use and preferences</td>
</tr>
<tr>
<td>1996</td>
<td>Family and social network</td>
</tr>
<tr>
<td>1997</td>
<td>Social security and poverty</td>
</tr>
<tr>
<td>1998</td>
<td>Ecology and envir. behavior</td>
</tr>
<tr>
<td>1999</td>
<td>Expectations, Use of time</td>
</tr>
<tr>
<td><strong>2000</strong></td>
<td><strong>Further education or training</strong></td>
</tr>
<tr>
<td>2001</td>
<td>Social networks, working cond.</td>
</tr>
<tr>
<td>2002</td>
<td>Social Security, assets</td>
</tr>
<tr>
<td>2003</td>
<td>Ecology and envir. behavior</td>
</tr>
<tr>
<td><strong>2004</strong></td>
<td><strong>Further education or training</strong></td>
</tr>
<tr>
<td>2005</td>
<td>Time use and preferences</td>
</tr>
<tr>
<td>2006</td>
<td>Family and social networks</td>
</tr>
<tr>
<td>2007</td>
<td>Assets</td>
</tr>
<tr>
<td><strong>2008</strong></td>
<td><strong>Further education or training</strong></td>
</tr>
</tbody>
</table>
Personality concepts

- Trust and Fairness (2003, 2008)
- Big Five Inventory (2005, 2010)
- Reciprocity (2005, 2010)
- Eagerness (2008)
- Impulsiveness (2008)
## Regional indicators

<table>
<thead>
<tr>
<th></th>
<th>Federal States</th>
<th>Regional Policy Regions</th>
<th>Counties</th>
<th>Communities</th>
<th>Zip-Codes, delivery area</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Corresponding NUTS Level</strong></td>
<td>NUTS1</td>
<td>NUTS2</td>
<td>NUTS3</td>
<td>NUTS5/LAU2</td>
<td>-/-</td>
</tr>
<tr>
<td><strong>Number of Regions (2005)</strong></td>
<td>16</td>
<td>97</td>
<td>439</td>
<td>12141</td>
<td>17201</td>
</tr>
<tr>
<td><strong>Number of Regions with SOEP respondents (2008)</strong></td>
<td>16 (100%)</td>
<td>97 (100%)</td>
<td>439 (100%)</td>
<td>2561 (21,1%)</td>
<td>2420 (14,1%)</td>
</tr>
<tr>
<td><strong>Mean number of SOEP Households (Respondents) per region in 2008</strong></td>
<td>684,9 (1246,6)</td>
<td>112,9 (205,6)</td>
<td>24,9 (45,4)</td>
<td>4,3 (7,8)</td>
<td>4,5 (8,2)</td>
</tr>
<tr>
<td><strong>Available for SOEP waves since ...</strong></td>
<td>1984</td>
<td>1985</td>
<td>1985</td>
<td>2000</td>
<td>1993</td>
</tr>
<tr>
<td><strong>Access policy</strong></td>
<td>Standard-DVD</td>
<td>Extra CD and password (Geocodes)</td>
<td>SOEPremote or at DIW Berlin</td>
<td>Only at DIW Berlin</td>
<td>Only at DIW Berlin (only one computer, in the near future via SOEPgeo at DIW Berlin)</td>
</tr>
</tbody>
</table>
Since 2000 the geographical coordinates for all (gross and net sample) households are at block level available.

New possibilities for example for mobility studies or segregation.

Pilot study in cooperation with DLR, to test the integration and possibilities of spatial information of the earth's surface with social scientific questions ("Synergies in socioeconomic and remote sensing").

Data access for external guest researchers (scheduled to end of 2009) and internal researchers possible through a special server environment at DIW Berlin only.
Cross-National Equivalent File (CNEF)

- Standardization of Various Measures
  (Focus: Income, Demographics, Employment, Health)

- USA  
  PSID - Panel Study of Income Dynamics
- Germany  
  SOEP - German Socio-Economic Panel Study
- Great Britain  
  BHPS - British Household Panel Study
- Canada  
  SLID - Survey of Labour and Income Dynamics
- Australia  
  HILDA - Household, Income and Labour Dynamics in Australia
- Switzerland  
  SHP - Swiss Household Panel

SOEP Homepage
- http://www.diw.de/soep
- Online Documentation
- All questionnaires in German and English

SOEPINFO
- Information System: Frequencies, Questionnaires
- Getting Started Quickly: Generate Command Files!

Desktop Companion (DTC)
- In-depth Description in English
- Basic, Extensions, Retrievals, Weighting
- Updated Every Year

SOEPLIT
- Database of articles, books, papers written using GSOEP
- Easy Search Interface
Search for all DP‘s, Articles, Books using SOEP!

• Search by Author, Year, Title or even over all fields.
• If available links to digital documents are provided.
SOEPinfo

- English or German
- Questionnaires
- Direct Search
- Topics
- Word Search
- Options
- Help

Content
Introduction
Sub-samples and Survey Related Issues
Sample Size Development
Content
Data Access
Innovations
• Strict data protection legislation requires …
  • Data user contract with DIW Berlin
  • Intl. Users 95% sample: Co-operation with Cornell Univ., Ithaca/NY

• Data dissemination currently on CD-Rom only
  • detailed written documentation
  • full German and English labeling (variables and values)
  • within EU: 100% sample
  • outside EU: International Scientific Use Version (95% sub-sample)
  • some variables dropped or recoded (regional info)

• Low nominal data fee (30 EUR/year; 125 USD one-time)
  • yearly updates on CD-Rom
• Access to sensitive information (counties or zip-codes, microm, geo-coordinates, plaintext answers):
  • Visiting DIW Berlin, normal workplace (counties, microm and plaintext answers)
  • Remote access via SOEPreMOTE (counties)
    - Email with stata syntax, fully automatic, 24/7
    - Visiting DIW Berlin, access to SOEPgeo Server (Zip-codes and geo-coordinates)
Introduction
Sub-samples and Survey Related Issues
Sample Size Development
Content
Data Access
Innovations
• **SOEPgeo** … allows processing of highly sensitive geo-coordinates (latitude and longitude) and data on zip-code level available on the premises of DIW Berlin. SO: with SOEPgeo, the SOEP department acts as a “Research Data Centre” (FDZ).

• … will improve current situation, as up to now data on a zip-code level can be analyzed on a single non-networked computer only (for internal as well for external users).

• … combines geo-coordinates and SOEP IDs on one server only which can not accessed by any user directly.

• It is intended that this server-based infrastructure will be available to all guest researchers in fall 2009.
Secure Shell Connection (SSH)

Server B (GIS Server):
- **Software:**
  - GRASS, R
- **Data:**
  - Geo-Coordinates (without SOEP ID)
  - External spatial data

Server C (SOEP Server):
- **Software:**
  - Stata, R
- **Data:**
  - SOEP (without Geo-Coordinates)
  - Indicators from external spatial data

Server A (Highly secured Server):
- **Software:**
  - postgres
- **Data:**
  - SOEP Household ID
  - Geo-Coordinates
  - Indicators from external spatial data

User

DIW Intranet

Input of Geo-Coordinates and prepared indicators only through pre-defined database rules

Output of SOEP Household ID and prepared indicators only through pre-defined database rules

Innovations

Data Access
Innovations

Innovation Projects 2001-2008

• Introduction of *event/age-specific questionnaires* (newborns, toddlers, preschoolers, youth, emigrants)
• Identification of twins
• Introduction of questions about *personal traits*
• Introduction of a new module on *individual wealth* holdings
• Two pretests of web-based surveying
• Measuring above and **beyond “survey questions”**
  • Evaluation of health (by taking grip strength)
  • Behavioral experiments
  • Evaluation of cognitive abilities
• Improvements of imputation techniques (MI)
• Improvement of industry (NACE) and ISCO coding – ready for ISCO2008
• Linking *neighborhood* data to SOEP households (microm)
• Linking *geographical codes* to SOEP households
• Establishment of an *interviewer survey*
• (Test) survey of attriters
Ongoing Innovation Projects

• Link to remote sensing data
• Sampling twins as a related sample (pretest study)
• Test of fully web-based SOEP questionnaires
• Test of mobile-phone-based interviewing and assessments
• Test of the event-triggered questionnaire “The Deceased Person” (in main SOEP wave of 2009)
• Test of statistical matching: SOEP and Social Security Register Data
• Pretests on educational aspirations of parents
• Pretest on factorial design (“vignettes approach”)
• Pretests evaluating the cognitive and non-cognitive skills of preschoolers and links with educational institutions (day care, school)
• Pretests of saliva sample collection

Innovations
Innovation Projects in Planning Stage

• Pretest on the longitudinal impact of incentives
• Pretest on better measurement of incomes, savings and expenditures
• Standardization of personality indicators, mood and emotions
• Questionnaire for children 8-9 years of age
Thank you for your attention!

E-mail: jgoebel@diw.de
SOEP-Hotline: soepmail@diw.de
SOEP-Homepage: www.diw.de/soep/