

# Introduction to the German Socio-Economic Panel (SOEP)

**Joachim R. Frick**  
(DIW Berlin) <[jfrick@diw.de](mailto:jfrick@diw.de)>

February 2009  
<http://www.diw.de/gsoep>

Special thanks to John P. Haisken-DeNew (RWI-Essen) for long-term cooperation  
in developing an earlier version of this introduction.

- 1. Introduction**
- 2. Content**
- 3. Sub-samples and Survey Related Issues**
- 4. Sample Size Development**
- 5. Data Structure**
- 6. Preparing Data for Analysis**
- 7. Dealing with Variables**
- 8. Weighting**
- 9. SOEP-Online Support**
- 10. Data Distribution**
- 11. Conclusion & Prospects**

**Appendix: NEWSPELL**

### **“Division of Labor”: SOEP and official statistics**

A rather European/German perspective

- Short-running panels by Official Statistics
  - ECHP, EU-SILC
- Long-running panel surveys under academic direction
  - SOEP, BHPS/Understanding Society, HILDA, PSID

#### **Advantages**

- Complementary contents, e.g.
  - In-depth income data in official panel surveys
  - Personal traits, health, subj. indicators in academic panels
- Mutual cross-validation

### **Well-known advantages of panel data**

- ⇒ Decomposition of gross and net changes  
(e.g. poverty rate as a result of inflow, outflow and the cont. poor)
- ⇒ Causal Relationship / Evaluation  
(sequential ordering of “trigger events” and outcome)
- ⇒ Control for otherwise unobserved heterogeneity  
(RE and FE-models)

## Increasing 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
- ⇒ Observation unit “household” guarantees appropriate coverage of births and deaths → from cradle to grave
- ⇒ Linking objective outcome measures to subjective indicators (→ satisfaction as a proxy for *utility*)
- ⇒ 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 (increasing with panel duration)**

- ⇒ 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 of questions, modes, ...)
  - inter-temporal comparability of indicators vs.
  - need for adjustments to cover institutional changes

# ***1e. Introduction : SOEP Basics***



The German  
Socio-Economic  
Panel Study

## **What is "SOEP" ?**

- ⇒ The SOEP was started in 1984: Now 24 Waves available!
- ⇒ Longest-running longitudinal survey of private households and persons in the Federal Republic of Germany („Living in Germany“)
- ⇒ Started with 6,000 Households in 1984, in 2007 approx. 12,000
- ⇒ Over-sampling of Foreigners, Migrants
- ⇒ East Germans, Top-up samples, Innovations, High Income

## **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
- ⇒ Topics in Economics, Sociology, Political Science, Psychology, Geography
- ⇒ Retrospective Information on Biographical History

### **Core Questions**

- ⇒ 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

## ***2b. Content***



The German  
Socio-Economic  
Panel Study

### **Topic Modules**

- ⇒ A-1984 Employment biography since age 15 (Bio)
- ⇒ B-1985 Marriage and family biography (Bio)
- ⇒ C-1986 Social origins (Bio), first job (Bio), neighborhood
- ⇒ D-1987 Social security, early retirement, disability, child care
- ⇒ E-1988 Assets
- ⇒ F-1989 Further education or training and qualification
- ⇒ G-1990 Time use and preferences
- ⇒ H-1991 Family and social services
- ⇒ I-1992 Social security and poverty
- ⇒ J-1993 Further education or training
- ⇒ K-1994 Neighborhood, values, and expectations
- ⇒ L-1995 Time use and preferences, income questions

## ***2c. Content***



The German  
Socio-Economic  
Panel Study

### **Topic Modules Cont'd**

- ⇒ M-1996 Family and Social network
  - ⇒ N-1997 Social security and poverty
  - ⇒ O-1998 Ecology and environmental behavior, indirect taxation
  - ⇒ P-1999 Expectations, Use of time
  - ⇒ Q-2000 Further education or training, labor market
  - ⇒ R-2001 Social networks, working conditions
  - ⇒ S-2002 Social Security, Assets
  - ⇒ T-2003 Ecology and environmental behavior
  - ⇒ U-2004 Further education or training, qualification
  - ⇒ V-2005 Time use and preferences
  - ⇒ W-2006 Family and social networks
  - ⇒ X-2007 Social Security, Assets
  - ⇒ Y-2008 Further education or training, qualification
- repetition of special topical modules in (approx.) 5-year intervals

## 2d. Content

### Dimensions of Time

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

## 3a. SOEP-Subsamples ...



The German  
Socio-Economic  
Panel Study

### Subsamples: Multi-step random sampling process

- ⇒ **A** **"West-German"** residents: started in 1984, n=4,528 households  
Head is either German or other nationality than those in Sample B.
- ⇒ **B** **"Foreigners"**: started in 1984, n=1,393 households (oversampling)  
Head is either Turkish, Italian, Spanish, Greek, or Yugoslavian.
- ⇒ **C** **"East-Germans"**: started in 1990, n=2,179 households  
Head was a citizen of the GDR. (expansion of survey territory)
- ⇒ **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=1,067 households  
Random sample covering all existing subsamples (total population)
- ⇒ **F** **"Innovation sample"**: started in 2000, n=6,052 households  
Random sample covering all existing subsamples (total population)
- ⇒ **G** **"High Income sample"**: started in 2002, n=1,224 households  
Monthly net Household income > 7.500 DM (4.500 EUR in wave 2)
- ⇒ **H** **"Refreshment sample"**: started in 2006, n=1,506 households  
Random sample covering all existing subsamples (total population)

## 3b. ... Survey Related Issues



The German  
Socio-Economic  
Panel Study

### Interview Methodology (Principles)

- ⇒ 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)

Example: Interview Mode 1999/2002 (variable \$PINTA)

Oral Interview/Interviewer	43% - 29%
Self-Completed w/o. Interviewer	29% - 25%
Written (Snail-Mail)	14% - 10%
Self-Completed w. Interviewer	5% - 3%
CAPI	5% - 28%
Part Oral / Part Self-Completed	4% - 4%
Phone	.07% -.00%
Proxy	.04% -.04%

## ***3c. Survey Related Issues***



The German  
Socio-Economic  
Panel Study

### **Survey Instruments: Address Log**

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)
- ⇒ different questionnaire versions according to survey status  
*old* ("green") v.s. *new* households ("blue")

### **Survey Instruments: Questionnaires**

- ⇒ Pre-tested questions, only
- ⇒ Yearly Standard Instruments
  - Household questionnaire
  - Individual questionnaire for each HH member aged 16 and over
  - Individual questionnaire „Gap“ (Temporary Drop-outs)
  - Until 1995 specific for sub-samples <“Germans”, “Foreigners”, “East Germans”, “Immigrants”> as well as for survey status <“old” vs. “new”>
- ⇒ Biography Questionnaire „Life History“
  - Until 2000 to be answered by first time respondents aged 17+
  - Covering life-time information up until 1<sup>st</sup>/2<sup>nd</sup> interview (education and labor market, marriage and fertility, socialization and parental background, immigration)
  - Since 2001 to be answered by first time respondents aged 18 and over

### Survey Instruments: Age-specific instruments as a means to improve data for cohort and life course analysis

- ⇒ Questionnaire „Youth/Adolescence“
  - First time respondents [aged 16/17] started in 2001 with birth cohort 1984
  - Topics covered: Relationship to parents, Leisure time use, School performance, Educational intentions, Job expectations, Personality characteristics, Family expectations, Standard indicators on intergenerational mobility
  - Since 2006 further enlarged with standard person questions → respondent age: 18+
- ⇒ Questionnaire „Mother & Child I“
  - New born infants [0 to 15 months] started in 2003 with birth cohort 2002/2003
  - Topics covered: Birth related issues (weight, height, timing, etc), Problems during pregnancy, Subj. indicators (mother), Caring for baby/ support by father & third parties
  - Approximately 250 observations per year !!! To be answered by the mother.
- ⇒ Questionnaire „Your child at age ... “
  - [aged 2 to 3] started in 2005
  - [aged 5 to 6] started in 2008
- ⇒ Prospective: Questionnaire “Your child at age ... ”
  - at age 9-10 (tentative start in 2012)
  - at age 13-14 (tentative start in 2016)

### **Test / Design Stage:**

#### **Event-triggered instruments**

- Death of a relative (incl. exit interviews)
  - tested in 2007/08
  - start in 2009
- Divorce
- Moving into retirement
- Moving into institutionalized household (e.g. nursing homes)

## ***3g. Survey Related Issues***



The German  
Socio-Economic  
Panel Study

### **Developing new instruments to shed more light on unobserved heterogeneity**

#### Health (>2002)

- SF12, BMI, Smoking (2002+)
- Physical ability: grip strength (2006+; ca. 100 Interviewer)

#### Personal Traits (>2002)

- Experiment on trust (2003+; subsample in F)
- Experiment on time preference (2006; only CAPI)
- Risk Aversion (2004+; full sample)
- Big Five (2005+; full sample)
- Locus of control (2005+; full sample)

#### Cognitive ability (>2005/6)

- Symbol-digit test (90 sec speed test) (only CAPI)
- Enumerating animals ("fluency") (only CAPI)
- Teenagers: full ability measures (2005+; n=250 per year)

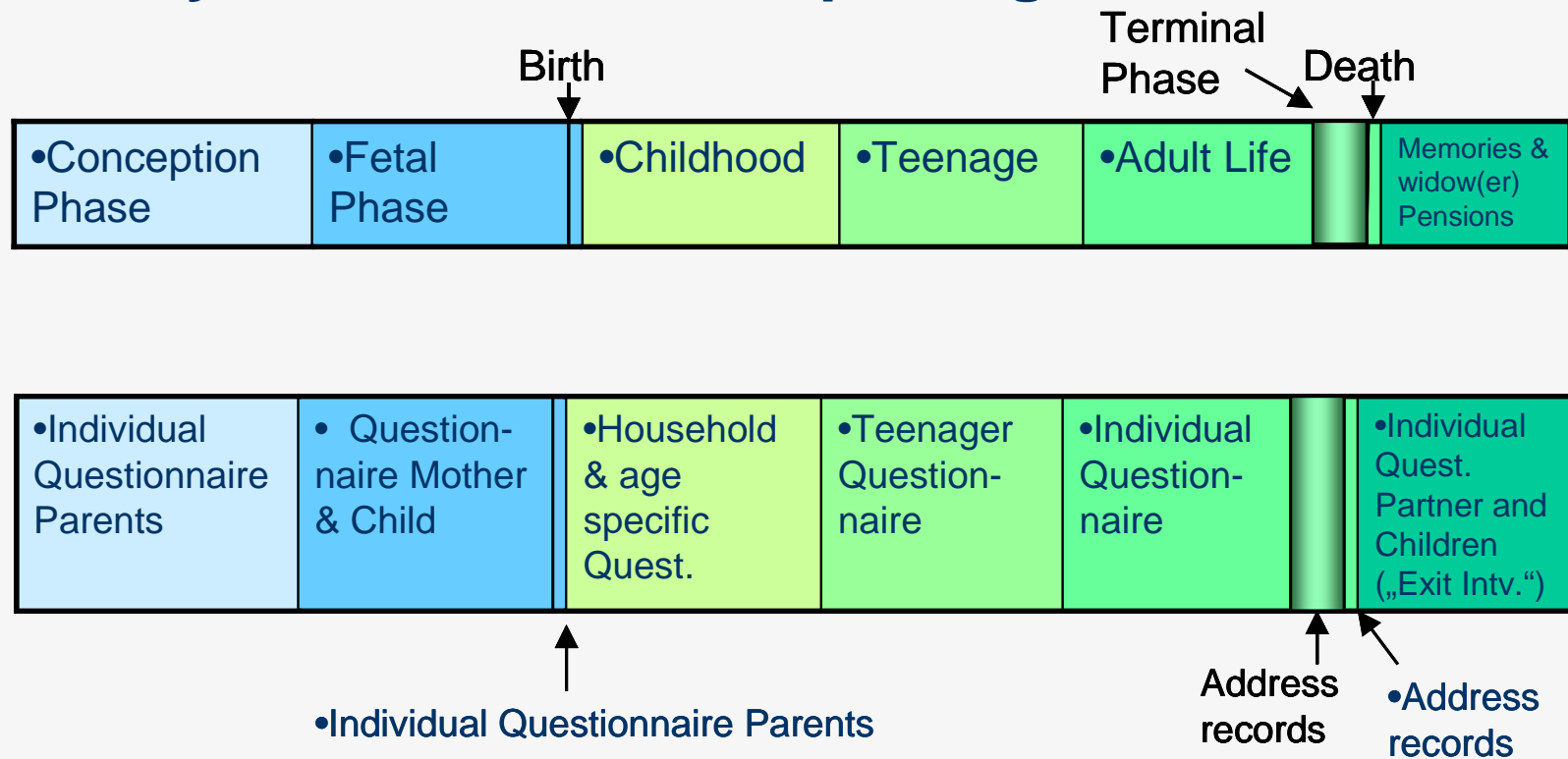
#### Biomarkers & Brain Scans (for selected respondents) ???

# 3i. ... SOEP and the life course



The German  
Socio-Economic  
Panel Study

## Survey Instruments for capturing the life course



## ***3i. Survey Related Issues***



The German  
Socio-Economic  
Panel Study

### **The Follow-Up Concept**

- ⇒ 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 persons are also followed in case of leaving the household. This had not been the case up to wave 5 (1988).

### **Temporary Drop-Outs**

- ⇒ 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).

## 3j. Survey Related Issues



The German  
Socio-Economic  
Panel Study

### The emergence of new households

		<i>Households</i>	
		<b>Old</b>	<b>New</b>
<i>Persons</i>	<b>Old</b>	<ul style="list-style-type: none"><li>• "classic case" without change of address</li><li>• entire household moves</li></ul>	<ul style="list-style-type: none"><li>• Move-out</li></ul>
	<b>New</b>	<ul style="list-style-type: none"><li>• Birth</li><li>• Move-in</li></ul>	<ul style="list-style-type: none"><li>• Birth</li><li>• Caused by splitt-offs of old persons from old households*</li></ul>

\* Remember that households *new* to the SOEP may already have existed before contacting the survey

## ***3k. Survey Related Issues***



The German  
Socio-Economic  
Panel Study

### **Starting Sample Size in Wave 1 (full 100% sample)**

Sample	Starting Year	Households	Respondents
A and B	1984	5,921	12,245
C	1990	2,179	4,453
D1/D2	1995	522	1,078
E	1998	1,067	1,923
F	2000	6,052	10,890
G	2002	1,224	2,671
H	2006	1,506	2,616

## 4a. Sample Size Development



The German  
Socio-Economic  
Panel Study

### Demographic factors

⇒ Persons exit by:

- Death
- Moving abroad

⇒ Persons enter by:

- Birth
- Moving into a SOEP household from somewhere else in Germany or from abroad
- Reaching age of 17 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

## ***4b. Sample Size Development***



The German  
Socio-Economic  
Panel Study

### **Field-work related factors (2 stages)**

- ⇒ making a successful contact to a given household
- ⇒ 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

## 4c. Sample Size Development



The German  
Socio-Economic  
Panel Study

### „Panel care“ – keep interviewees involved

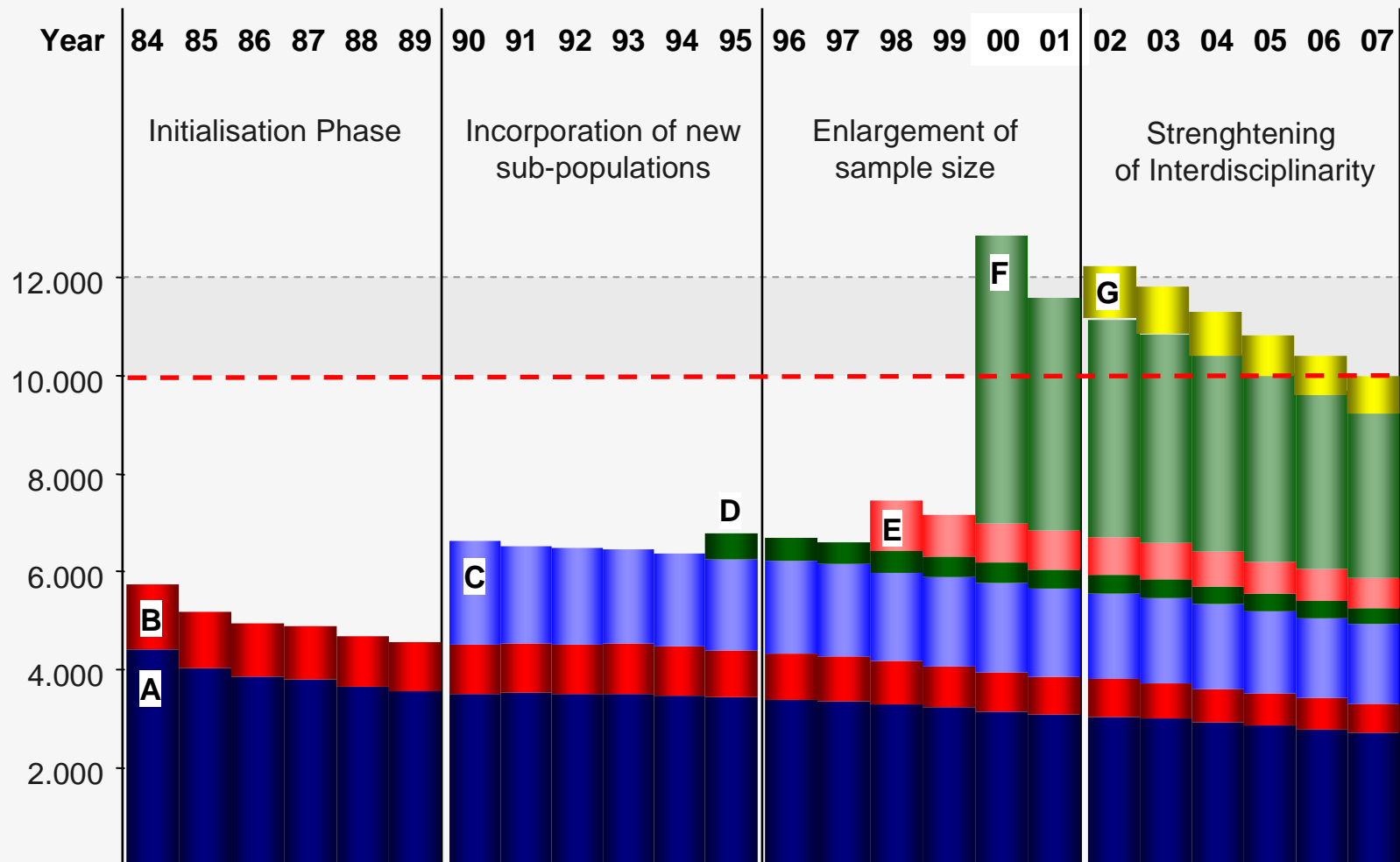
- ⇒ Announcement by mail 2 weeks prior to interview
- ⇒ For each successful interview, any respondent
  - receives a gift related to the yearly topical module (until 2007)
  - ticket for monthly nationwide lottery
    - since 2008: 5 € ex-ante incentive (until 2007: 1,50 € ex-post)
- ⇒ Addresses are 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.

# 4d. Sample Size Development



The German  
Socio-Economic  
Panel Study

## The SOEP sample 1984 – 2007 (households)



Source: TNS-Infratest Sozialforschung

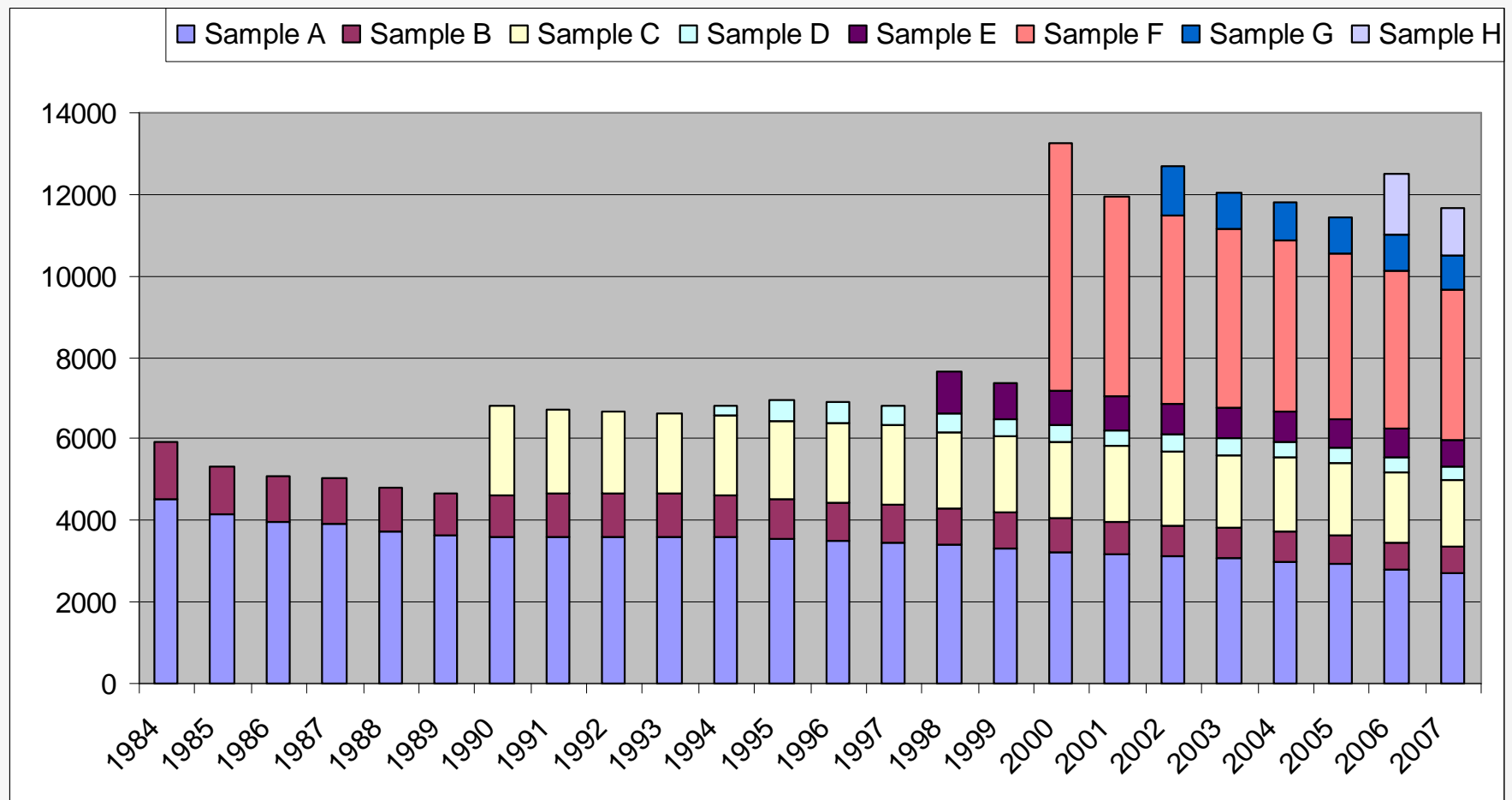
# 4d. Sample Size Development



The German  
Socio-Economic  
Panel Study

## Cross-sectional perspective

⇒ Number of successfully interviewed households by sample



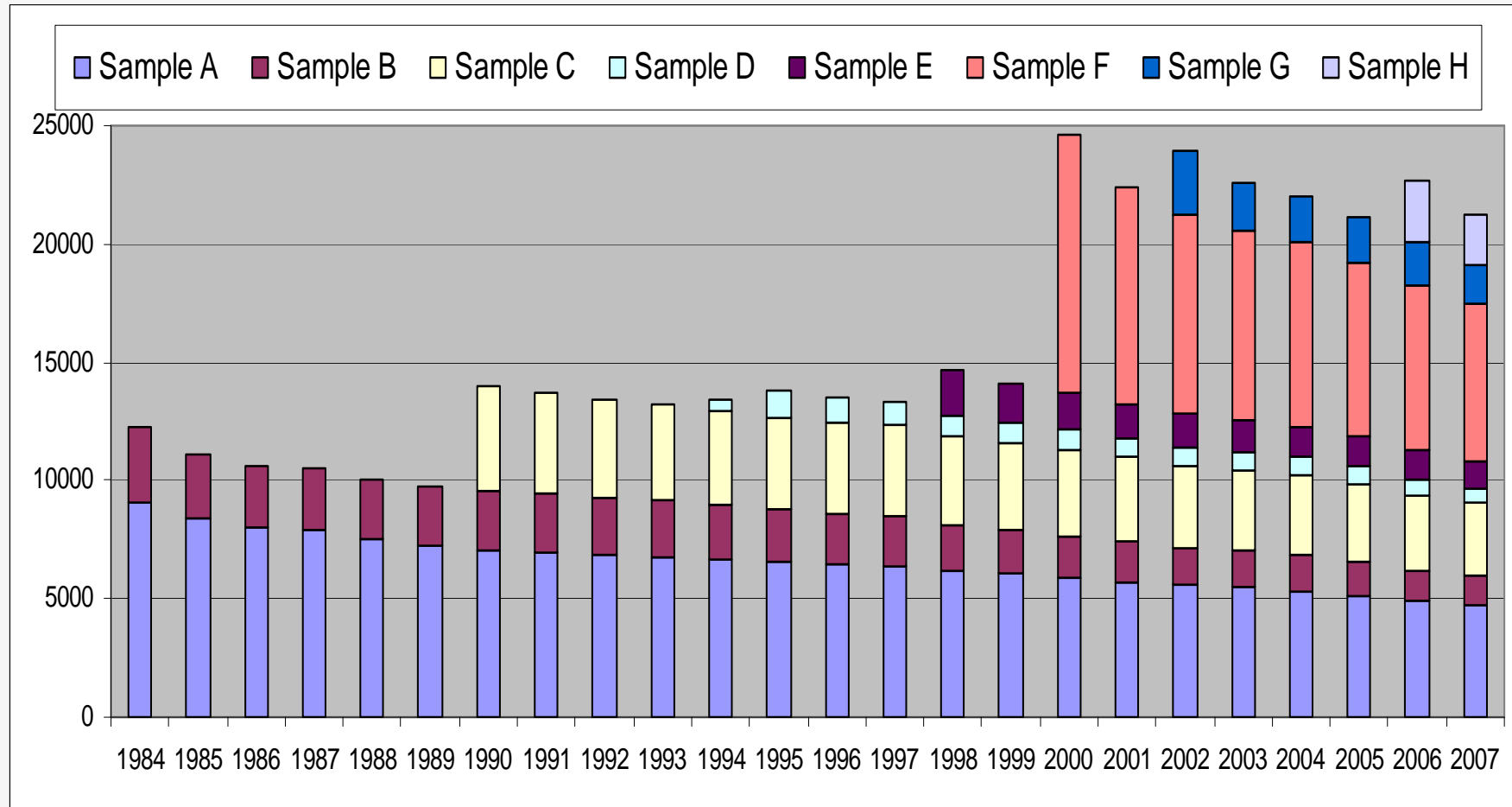
# 4e. Sample Size Development



The German  
Socio-Economic  
Panel Study

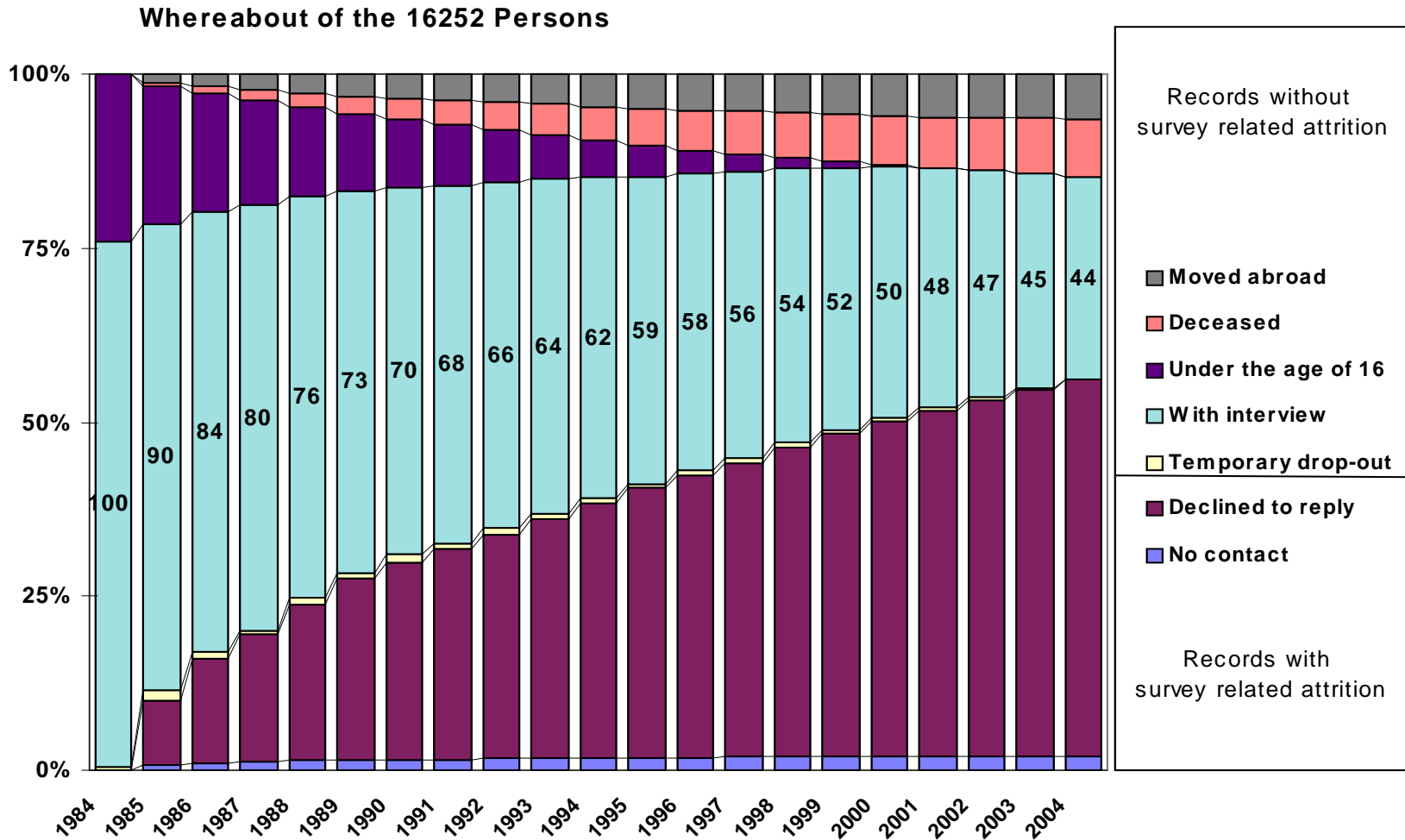
## Cross-sectional perspective

⇒ Number of successfully interviewed persons by sample



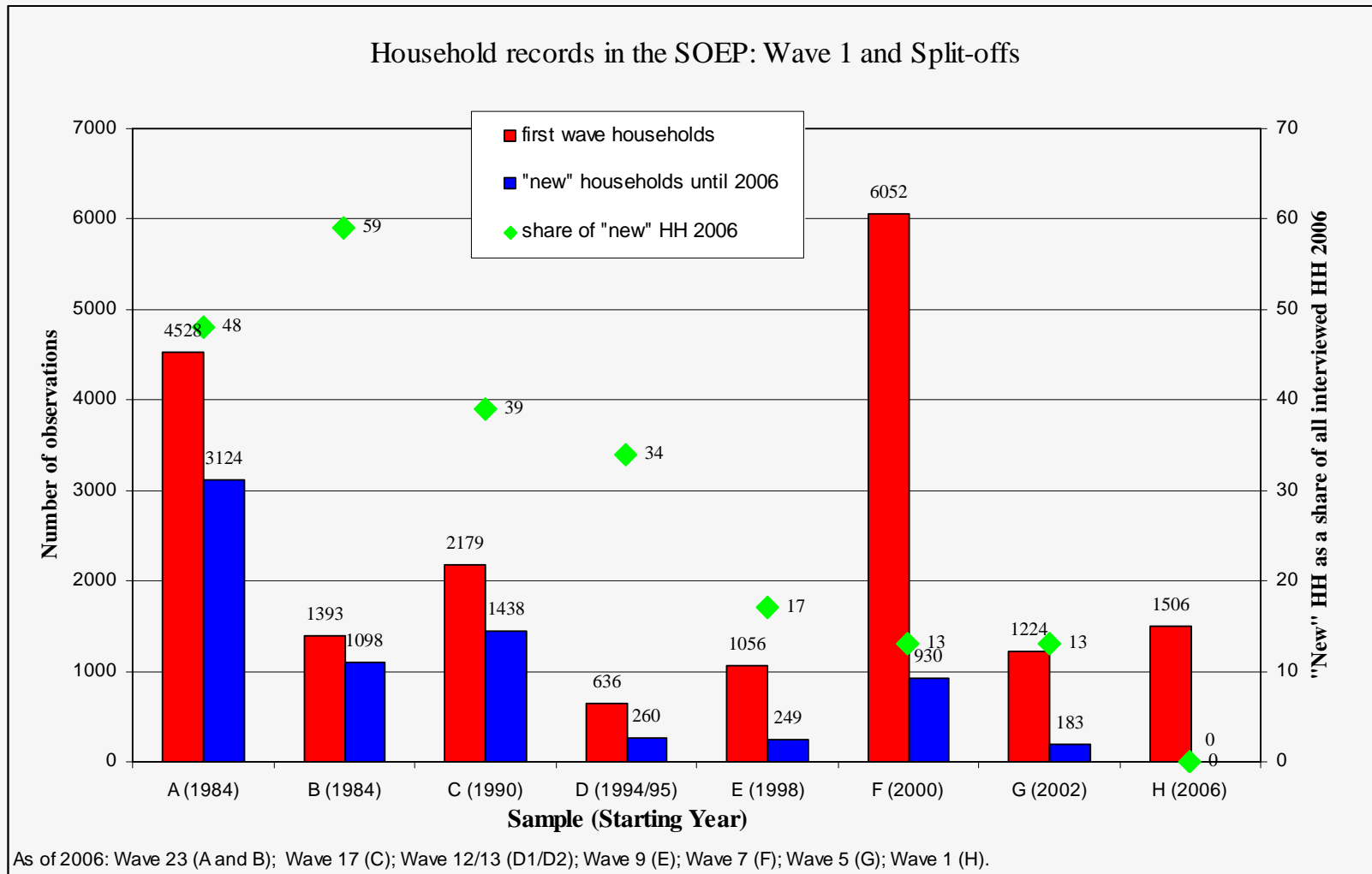
# 4f. Sample Size Development

## Longitudinal perspective: Samples A+B since 1984



# 4g. „New Households“

## Incidence of New households after Wave 1



# 5a. Preparing Data for Analysis



The German  
Socio-Economic  
Panel Study

## Cross-Sectional Data

Series of cross-sections within Panel Population

$t_0$	$t_1$	$t_2$	
			drop-outs

not yet in the sample or not yet interviewed

# 5a. Preparing Data for Analysis



The German  
Socio-Economic  
Panel Study

## Longitudinal Data

Complete case analysis with a balanced panel design

$t_0$	$t_1$	$t_2$	
			drop-outs
			successfully interviewed in all waves
			new respondents
			not yet in the sample or not yet interviewed

## 5b. Preparing Data for Analysis



The German  
Socio-Economic  
Panel Study

### Longitudinal Data

#### Downstream model

$t_0$	$t_1$	$t_2$	
			drop-outs
			successfully interviewed in all waves
			new respondents
			not yet in the sample or not yet interviewed

## 5c. Preparing Data for Analysis



The German  
Socio-Economic  
Panel Study

### Longitudinal Data

Complete information analysis with an unbalanced panel design

$t_0$	$t_1$	$t_2$	
			drop-outs
			successfully interviewed in all waves
			new respondents
			not yet in the sample or not yet interviewed

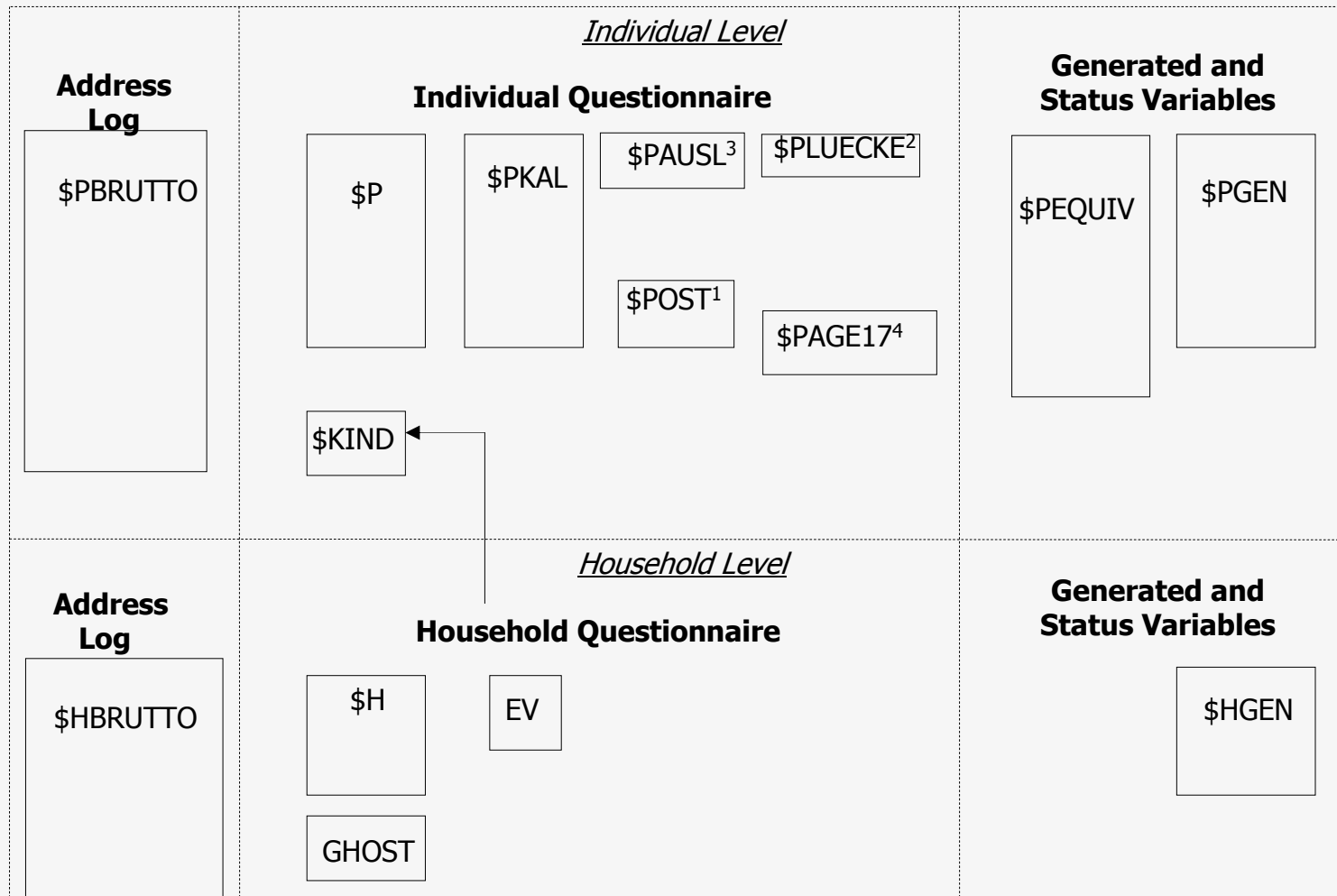
### Defining the Population and Unit of Analysis

- ⇒ Survey Data are organized cross-sectionally (in principle)
- ⇒ Additional Longitudinal “meta files” are *explicitly* designed to support panel analyses
  - Defining population of interest: Gender, Age, Region, Samples
  - Defining time period: Years, Waves
  - Balanced vs. Unbalanced panel design
  - „Long” (pooling) vs. „Wide” format
- ⇒ Spell and Event Data
  - Data stored as Person-Events *not* Person-Years

# 5e. Data Structure: Cross-Section



The German  
Socio-Economic  
Panel Study



\$: Wave specification: A, B, C... X

<sup>1</sup> Waves G and H only; <sup>2</sup> Waves B through Q only; <sup>3</sup> Waves A through L only; <sup>4</sup> Starting Wave W.

# 5f. Data Structure: Longitudinal



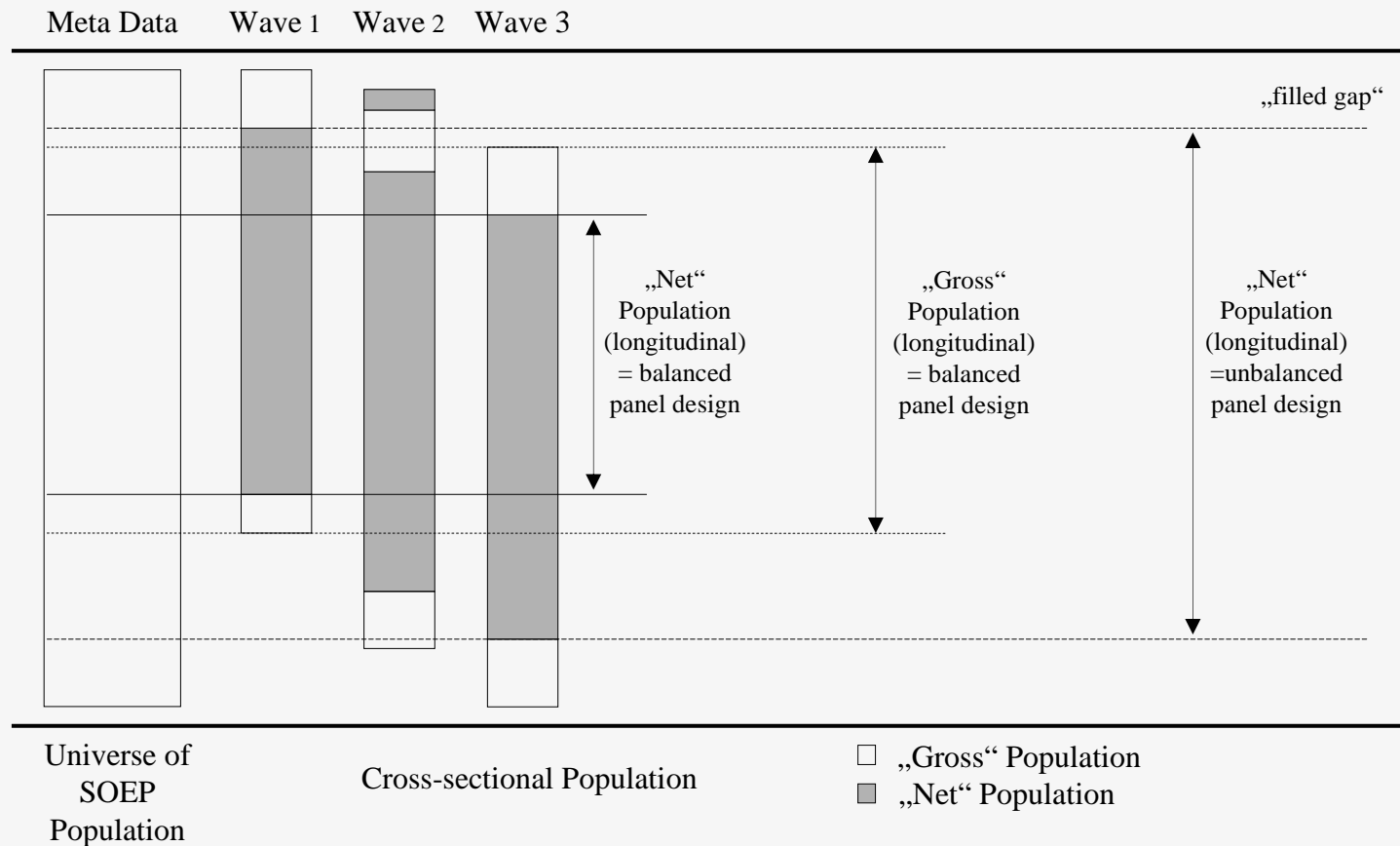
The German  
Socio-Economic  
Panel Study

<i>Individual/ Household</i>				<i>Cumulative Data</i>	<i>House hold</i>	<i>Spell</i>					<i>Individual</i>		
<i>META- DATA</i>		<i>WEIGHTING FACTORS</i>		<i>DROP- OUTS</i>	<i>SOCIAL ASSIST. (month)</i>	<i>CALENDAR (month)</i>		<i>OCCUP. BIO (year)</i>	<i>MARITAL STATUS (year) (month)</i>		<i>BIRTH (women &amp; men)</i>	<i>PARENTAL INFO</i>	
PPFAD	HPFAD	PHRF	HHRF	YPBRUTTO	SOZ- KALEN	ART- KALEN	EIN- KALEN	PBIO- SPE	BIO- MARSY	BIO- MARSM	BIO- BIRTH	BIO- PAREN	
				PERSONS NEEDING CARE							BIO- BRTHM		
				PFLEGE							BIORESID	2nd RESIDENCE	
				HEALTH							AGE SPEC. INFO		
SAMP	VARIANZ & DESIGN			HEALTH							BIOSOC	BIO- AGE17	BIOAGE01
				GRIPSTR									BIOAGE03
				WEALTH							BIOTWIN	MULTIPLES	
				PWEALTH HWEALTH							MIGRATION (migrants only)	FIRST JOB	
											BIO- IMMIG	BIO- JOB	

# 5g. Data Structure: Matching



The German  
Socio-Economic  
Panel Study



## 5h. Data Structure: Matching



The German  
Socio-Economic  
Panel Study

### Household Files

- ⇒ HPFAD multi-wave household meta information
- ⇒ HHRF household weights
  
- ⇒ \$HBRUTTO single-wave household address register
- ⇒ \$H/\$HOST single-wave household data
- ⇒ \$HGEN single-wave generated household data
  
- ⇒ EV household wealth/assets (wave 5 / 1988)
- ⇒ GGKBOU\* regional information: community size classification

ID (Matching Vars): HHNR, HHNRAKT (\$HHNR)

\* not available in Scientific Use Version for users outside the EU (Data Privacy),  
however, access possible via SOEPremote)

## 5i. Data Structure: Matching



The German  
Socio-Economic  
Panel Study

### Person Files

⇒ PPFAD multi-wave person meta information

⇒ PHRF person weights

ID (Matching Vars): HHNR, PERSNR

⇒ \$PBRUTTO single-wave person register

⇒ \$P/\$POST single-wave person data

⇒ \$PGEN single-wave generated person data

⇒ \$PKAL/\$PKALOST single-wave person monthly calendar (JAN-DEC)

⇒ \$PAUSL single-wave foreigner data (Sample B, only)

⇒ \$PLUECKE single-wave person gap data

⇒ \$KIND single-wave child data

⇒ \$PEQUIV CNEF Variables + background variables

⇒ \$PAGE17 single-wave data on 17 year old (start in Wave W)

ID (Matching Vars): HHNR, HHNRAKT (\$HHNR), PERSNR

⇒ YPBRUTTO multi-wave person register (drop-out)

ID (Matching Vars): HHNR, PERSNR, ERHEBJ

## 5j. Data Structure: Matching



The German  
Socio-Economic  
Panel Study

### Biography & Spell Data

- ⇒ BIOIMMIG Migration Biography
- ⇒ BIOJOB First Job, Entrance to Labor Market
- ⇒ BIOBIRTH/M multi-wave birth biography
- ⇒ BIOPAREN parental information
- ⇒ BIOAGE01 data on „mother & child“ (around birth)
- ⇒ BIOAGE03 data on „infants“ (2-3 year olds)
- ⇒ BIOAGE17 youth data (16-17 year olds)
- ⇒ BIOSOC data on adolescence
  
- ⇒ PBIOSPE biography calendar (spell data)
- ⇒ ARTKALEN activity calendar (spell data)
- ⇒ EINKALEN income calendar (spell data)
- ⇒ SOZKALEN HH-based social assistance calendar (spell data)
- ⇒ BIOMARSM marriage data by month (spell)
- ⇒ BIOMARSY marriage data by year (spell)

ID (Matching Vars): HHNR, (HHNRAKT/PERSNR), (SPELLNR/ERHEBJ)

## Cumulative Data

- ⇒ PWEALTH            Wealth Information Individual Level (curr. only wave S)
- ⇒ HWEALTH           Wealth Information HH Level (curr. only wave S)
  
- ⇒ HEALTH            Ind. Level Health Data (curr. 3 waves S, U, W)
- ⇒ GRIPSTR            Physical Health Measure Gripstrength (curr. only wave W)
  
- ⇒ PFLEGE            level in-need-of-care (invalidity) data (1985-2006)

ID (Matching Vars): HHNR, PERSNR/HHNRAKT, SVYYEAR/ERHEBJ

### Event and Spell Data

Although in the course of time the absolute number of observations (e.g., individuals) is almost steadily decreasing from a cross-sectional perspective, the cumulative number of events and/or spells covered by the entire data is increasing wave by wave.

#### Events

- ⇒ Re-migration: YPBRUTTO (up to 2007: >1,600 events)
- ⇒ Deaths (mortality): YPBRUTTO (up to 2007: >2,800 events)
- ⇒ Births (fertility): BIOBIRTH (up to 2007: 32,000 births with approx. 21,000 of these persons being identified within the SOEP population)

#### Spells

- ⇒ Monthly and yearly labor market status: ARTKALEN, PBIOSPE (as of 2007: e.g. 20,000 spells of unemployment, >57,000 FT employment)
- ⇒ Monthly and yearly marital status biography: BIOMARSM, BIOMARSY
- ⇒ Monthly receipt of social assistance: SOZKALEN (household)

## 6b. Preparing Data for Analysis



The German  
Socio-Economic  
Panel Study

### Status Variables

⇒ *Problem:*

- Some information is collected in the first interview only.
- Old respondents are asked for changes since last year's interview, while new respondents have to fill in the current status.

⇒ *Example:* since when with current employer

⇒ *Solution:*

- The collected information is stored in different variables.
- In order to minimize computing efforts for the user, the GSOEP provides yearly **status variables** on individual and household level, which integrate all of these information in a common variable showing the current status for all respondents.
- Thus, there is nothing else but a re-organisation of already existing data and there is almost no assumption or normative setting involved in the generating process.
- Variable \$ERWZEIT in file \$PGEN gives the number of years with the current employer for all employed respondents

Data Files: \$PGEN, \$HGEN, PPFAD, BIO\*

## 6c. Preparing Data for Analysis



The German  
Socio-Economic  
Panel Study

### Generated Variables

- ⇒ *Problem:*  
Provide a user-friendly single output variable
- ⇒ *Example:* Typically required institutional fulltime years of education
- ⇒ *Solution:*  
Variable \$BILZEIT in file \$PGEN: Based on separate variables measuring schooling attainment, completed training certificates, vocational education, university degrees, etc., a total number of years of education is generated

Data Files: \$PGEN, \$HGEN, PPFAD, BIO\*

# 7a. Dealing with Variables



The German  
Socio-Economic  
Panel Study

## Principles for Naming Survey Variables

### Digit Meaning

- 1 Wave
- 2 Unit of analysis: P=individual, H=household
- 3-4 Number of question in survey instrument (questionnaire)
- 5-6 Number of item in survey instrument (questionnaire)
- 5 or 7 Differentiation according to sub-sample:  
A=Foreigners, O=East Germans
- 2-8 Text for Variables in \$BRUTTO files  
and some occupation-related variables in \$P files

### Examples:

- AP04 Wave 1; Individual; Question 4
- BH0502 Wave 2; Household; Question 5; Item 2
- DP24G09 Wave 4; Individual; Question 24; Green version; Item 9
- AP64A Wave 1; Individual; Question 64; Sample B, Foreigners
- BIS88 Wave 2; Intl. Standard Classification of Occupation (ISCO88)

**Exceptions:** Identifiers; generated variables (\$P/HGEN, BIO\*, \$PEQUIV)

## *7b. Dealing with Variables*



The German  
Socio-Economic  
Panel Study

### **Dealing with Missing Values**

- SOEP data differentiates three kinds of missing values:

#### **Code    Meaning**

- 1    no answer / do not know / item non-response
  - 2    does not apply
  - 3    after checking for plausibility a given value was found to be implausible and finally deleted (to be interpreted like -1)
- However, there are NO system-missing values !

### **Imputation of missing income variables in case of item-non-response** (codes „-1“ or „-3“ in the original variable)

#### **Income:**

- Single imputation of missing income variables using ...
  - „Row-and-Column“ imputation based on longitudinal data (Little & Su, 1989)
  - x-sectional imputation (mostly regression-based) if no panel data available
  - Annual income (wrt previous year; stored in file \$PEQUIV)
  - Current monthly labor income (stored in file \$PGEN)
  - Imputation flags indicating imputation status
- Since release 2008: multiple imputation for current monthly HH net income („screener“); stored in file \$HGEN and MIHINC

#### **Wealth:**

- with release 2007: multiply imputed data on individual wealth (2002) stored in files PWEALTH (incl. PUNR) and HWEALTH (household level wealth aggregates)
  - wealth data for 2007, 2012, etc. will be cumulatively stacked into these files
  - wealth data release for 2007: multiple imputation using *longitudinal data*

## 8a. Weighting



The German  
Socio-Economic  
Panel Study

### Why is Weighting Necessary ?

⇒ Different design probabilities for subsamples A-G in wave 1

⇒ Unit-Non-Response

- not willing to participate in the first wave
  - 1984 subsamples A and B
  - 1990 subsample C
  - 1994 subsample D1 / 1995 subsample D2
  - 1998 subsample E
  - 2000 subsample F
  - 2002 subsample G
  - 2006 subsample H
- attrition in the subsequent waves due to
  - unsuccessful follow-up (lost contact)
  - refusal (*temporary or final*)

## 8b. Weighting

### Sampling Information Across Subsamples

#### Cross-Sectional Weighting Wave 1

	Subsample					
	A "West Germans"	B "Foreigners in West- Germany"	C "German Residents in the GDR"	D "Immigrant since 1984"	E "Refresh- ment"	F "Innovation"
Starting wave	1984	1984	1990	1994/95	1998	2000
Target Population (households)	4.500	1.500	2.000	500	1.000	6.000
Initial Response Rate in %	61	68	70	≈55-75	54	52
Sampling Probability	≈.0002	≈.0008	≈.0004	≈.0002	≈.00003	≈.00016
Average Weighting Factor	≈5,000	≈1,250	≈2,500	≈5,000	≈33,333	≈6,000

$$\text{Weight} = 1 / [ P(D_i=1) * P(R_i=1|D_i) ]$$

### External Information: Adjust Wave 1 Weights

- ⇒ **Household:** Sex of head, Age of head, Size of household, Nationality of head, Country of origin of immigrants
- ⇒ **Resident population in private households:** Sex, Age, Marital, Status, Nationality of head
- ⇒ **Resident population, total:** Sex, Age, Counties, Size of community
- ⇒ **Children up to 16 years of age:** Sex, Type of School, Nationality of head
- ⇒ **Gainfully employed persons residing in private households:** Sex, Age, Job (ISCO-1 digit), Nationality

## 8d. Weighting



The German  
Socio-Economic  
Panel Study

### Longitudinal Weights

- ⇒ **Idea:** Estimate Probability to remain in sample: [ t , t+1 ]  
Separately for Each subsample (Logit Regression)
  - ⇒ Contact Probability (CP)
  - ⇒ Response Probability (RP), given Contact
  - ⇒ Inverse Staying Probability =  $1 / [CP * RP]$
- ⇒ **Longitudinal Weight** =  $\frac{[\text{Cross-Section Wgt Starting Wave}]}{[\text{Inv Staying Prob}]}$  \*

## ***8e. Weighting***



The German  
Socio-Economic  
Panel Study

### **Significant Predictors for Contact Probability**

- ⇒ Whether Moved
- ⇒ Large City
- ⇒ Household size
- ⇒ Type of house
- ⇒ Split-off household

### **Significant Predictors for Response Probability**

- ⇒ Age/Gender of household head, Type of the household
- ⇒ Change of interviewer, Number of interviews
- ⇒ Household head Member of Wave 1
- ⇒ Person moving out, Marital status, Separation of a couple
- ⇒ Unemployment, Expected Job Loss, Occupational status
- ⇒ Welfare Recipient, Household income
- ⇒ Household income not reported, Assets, No assets reported
- ⇒ East-West Migration, Migrant, Telephone, Sub-tenant

### Ready-Made Variables Provided

- ⇒ **Cross-Section:** Person (\$PHRF) and Household (\$HHRF) level
- ⇒ **Inverse Staying Probability:** Person and Household level
- ⇒ **Sample Specific Cross-Section Weights** (Samples D, F, G)
- ⇒ **Separate weights for time series** (e.g. on income inequality) dropping wave 1 of a given sample (\$PHRF1)
- ⇒ **Support for Variance Estimation:** Random Groups, Strata

Data Files: PHRF, HHRF, DESIGN (up to release 2006: VARIANZ)

## ***9a. Online Support***



The German  
Socio-Economic  
Panel Study

### **SOEP Homepage**

- ⇒ <http://www.diw.de/gsoep>
- ⇒ Online Support Services, Documentation, FAQ, SOEPnewsletter

### **SOEP *info***

- ⇒ Information System: Frequencies (unweighted), Questionnaires
- ⇒ Getting Started Quickly: Generate Command Files !

### **Desktop Companion (DTC)**

- ⇒ In-depth Description in English
- ⇒ Basic, Extensions, Retrievals, Weighting

### **SOEP *lit***

- ⇒ Database of articles, books, papers written using SOEP
- ⇒ Easy Search Interface

### **SOEP *monitor***

- ⇒ Time series on wide range of cross-sectional and longitudinal indicators of living conditions in East and West Germany (weighted)
- ⇒ Benchmark for users !!

9b. <http://www.diw.de/gsoep>

**SOEP**

The German  
Socio-Economic  
Panel Study

## Online Information

SOEP *INFO*  
SOEP *LIT*  
DTC  
Newsletter  
Links

DIW Berlin

Deutsch Search Press Room Contact The Institute Departments **SOEP** Publications Projects Service

- SOEP-Overview
- News
- FAQ
- SOEP NEWSLETTER
- SOEPinfo
- Data Center
- Services & Documentation
- SOEPlit
- SOEP-Advisory Board
- Contact

**The German Socio-Economic Panel**

**SOEP** The German Socio-Economic Panel Study

Director: **Gert G. Wagner**, Professor of Economics

**A Representative Longitudinal Study of Private Households in the Entire Federal Republic of Germany**

**SOEP (1984-2003)**

The SOEP is a wide-ranging representative longitudinal study of private households. It provides information on all household members, consisting of Germans living in the Old and New German States, Foreigners, and recent Immigrants to Germany. The Panel was started in 1984. In 2001, there were more than 12,000 households, and more than 22,000 persons sampled.

Some of the many topics include household composition, occupational biographies, employment, earnings, health and satisfaction indicators.

The data are available to researchers in Germany and abroad in

## 9c. SOEPinfo

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

WWW-SOEPinfo (SOEP-DB 2001) - Netscape 6

http://panel.gsoep.de/soepinfo2001/

[SOEPinfo-Main-Actions] [Basic Information] [Help]

Basket:  
0 Vars

[Basket-Actions] Select all Clear Delete [Files]

### WWW-SOEPinfo

WWW-SOEPinfo is an interactive system with extensive search capabilities that provides you with detailed information on the variables in the SOEP dataset. By interactively creating lists of variables in SOEPinfo, the user can then output frequencies information, item correspondence, and even generated SPSS, SAS, and Stata command files.

**SOEPinfo should be your first step in starting any new SOEP project!**

Language Version / Spachversion:

German  English

[Variables] [Topics] [Questionnaires]

# 9d. Desktop Companion (DTC)

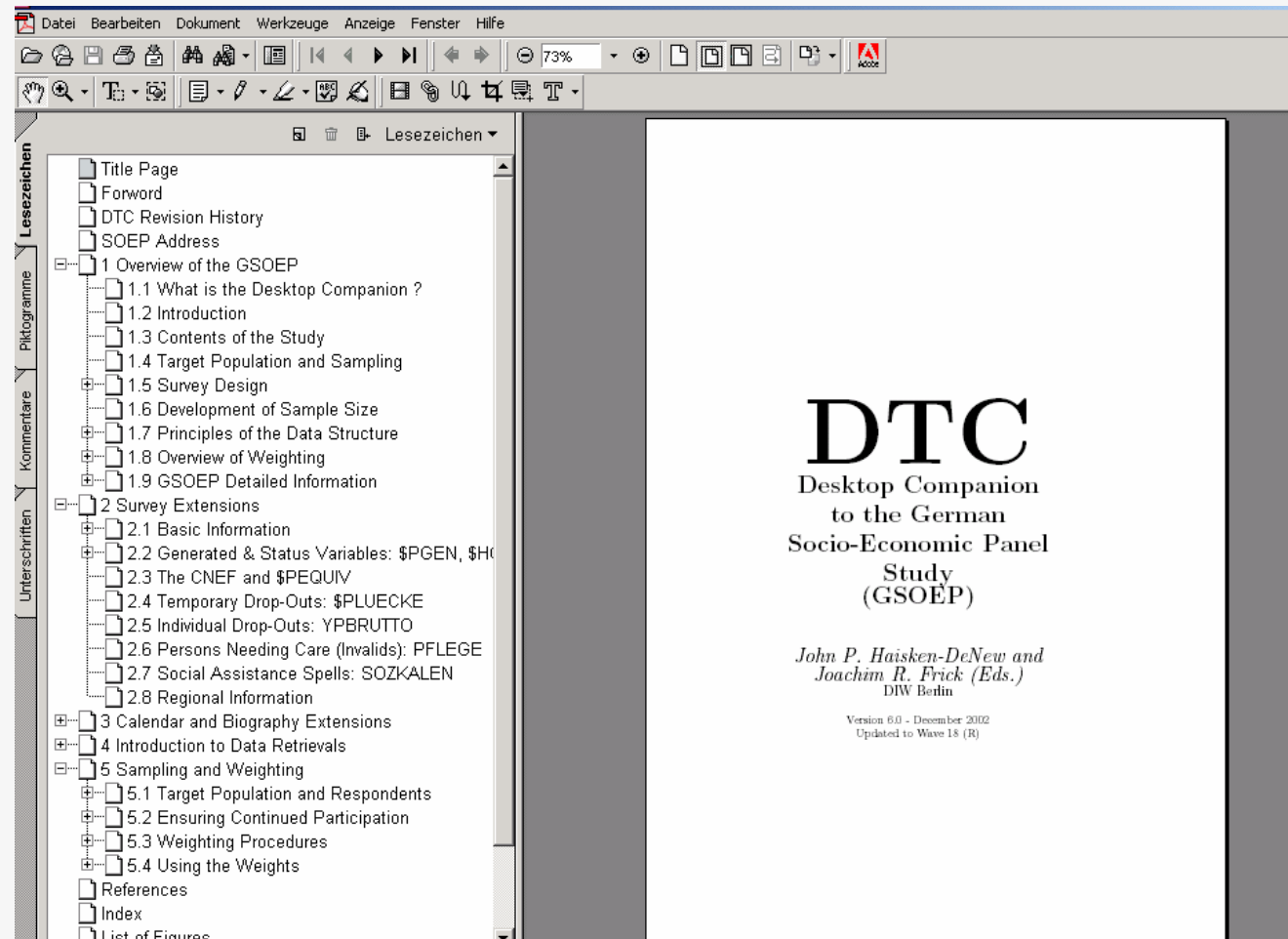
**SOEP**

The German  
Socio-Economic  
Panel Study

## Main Resource

Written in  
LaTeX,  
automatic  
Adobe PDF  
Bookmarks

Clickable



# 9e. SOEPlit - Literature

**Search for  
all DP's,  
Articles,  
Books using  
SOEP**

Input any  
Keyword, use  
Boolean  
Operators

The screenshot shows a Netscape 6 browser window with the address bar set to `http://panel.gsoep.de/soeplit/`. The page content includes the title "Das Sozio-oekonomische Panel" and the large "SOEPlit" logo. Below the logo, it states "Updated: 24. Feb 2003 -- Total Records: 2208". On the left side, there are links for "Help" and "About SOEP-lit". A search section is present with a "Search:" label, a text input field, and a dropdown menu for "Output Headline:" set to "TITLE". The "Max Hits:" is set to "50". There are "SEARCH" and "RESET" buttons. Below the search section, there is a "Problems?" link and a contact email "urahmann@diw.de". At the bottom left, it says "last update: 24.02.2003 22:38:54 CET". At the bottom right, there are links for "Home" and "Email".

## 10a. Data Distribution



The German  
Socio-Economic  
Panel Study

- ⇒ Strict data protection legislation requires ...
  - Data user contract with DIW Berlin
  - Intl. Users: Co-operation with Cornell Univ., Ithaca/NY
  
- ⇒ Data dissemination on CD-Rom / DVD 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)
    - detailed regional info not available
  
- ⇒ Low nominal data fee (30 EUR/year; 125 USD one-time)
  - yearly updates on CD-Rom / DVD
  
- ⇒ Access to sensitive information (e.g. regional data):
  - *SOEPremote* allows processing of sensitive data from outside DIW Berlin

# 10b. Data Distribution – DVD



The German  
Socio-Economic  
Panel Study

## Windows Setup Program

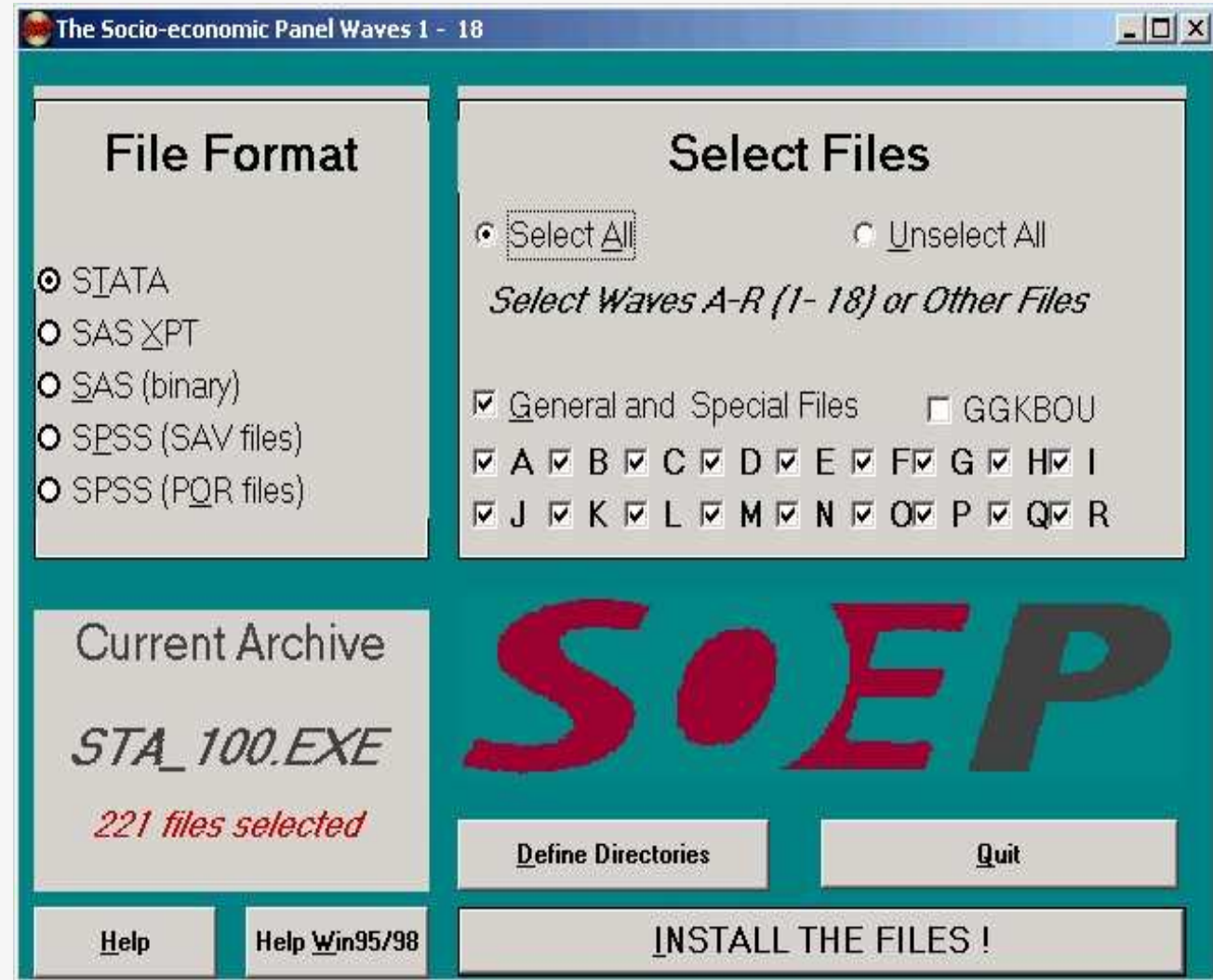
Ready-to-Go

⇒ SPSS

⇒ Stata

⇒ SAS

⇒ ASCII



### **Geo-coded data at various levels**

- county level (*Landkreis*)
- community level (*Gemeinde*)
- ZIP-code level (official geo-coded data)
- block level (commercial data)

### **Linking Neighborhood Information**

- based on geographic coordinates (latitude and longitude)

Access via **SOEPremote**



# 11a. Conclusion & Prospects



The German  
Socio-Economic  
Panel Study

## Due to recent and upcoming innovations ...

- SOEP stands for theory based, research driven data collection – not just “more and better statistics”
- Improved potential for analyses of smaller subpopulations due to increased sample size (~ 12.000 HH since 2000)
- Improved potential for intergenerational analyses based on more than 20 waves of data (PSID ~ 35<sup>th</sup> wave !)
- Improved biographical data (e.g., youth and adolescence)
  - relevant background information (RHS-variables)
  - as well as of self-contained interest
- Improved data on methodological issues  
(interview modes, imputation of missing values (income & wealth), ...)
- Improved geo-coded data (accessible via SOEPremote)
- Better controls for otherwise unobserved heterogeneity  
(behavior, health, personal traits, etc.)

## ***11b. Conclusion & Prospects***



The German  
Socio-Economic  
Panel Study

SOEP's international networking improves the cross-national infrastructure for data and analyses

- Cross-sectional databases: LIS, LWS
- Longitudinal databases:  
CNEF (1984-2005+), CHER (1990-2001), ECHP (1994-2001)
- Development of user-friendly and cross-nationally comparative micro-data (ex-post harmonization)
  - education: ISCED, CASMIN (\$PGEN)
  - labor market: ISCO88, NACE (\$PGEN)
  - regional information: NUTS (\$HGEN)
  - income: Canberra Group recommendation (\$PEQUIV)

# 11c. Conclusion & Prospects



The German  
Socio-Economic  
Panel Study

## International collaboration ...

- with other producers and analysts of panel data (e.g. BHPS, HILDA, PSID)
- and with respect to data collection (e.g. pre-testing),
- methodology (e.g. weighting, imputation),
- and substantive issues (e.g. timing of special topical modules)

→ simplifies future (ex-ante) data harmonization  
and quality of data for cross-national analyses

- Wagner, Gert G., Frick, Joachim R., Schupp, Jürgen (2007): The German Socio-Economic Panel Study (SOEP) - Evolution, Scope and Enhancements. *Schmoller's Jahrbuch - Journal of Applied Social Science Studies*. 127 (1): 139-169.
- Haisken-DeNew, John P. and Frick, Joachim R. (2005): Desktop Companion to the German Socio-Economic Panel Study (GSOEP), Version 8.0 – Update to Wave 21, DIW Berlin.

# *Appendix: NEWSPELL*



The German  
Socio-Economic  
Panel Study

by Rainer Pischner [rpischner@diw.de](mailto:rpischner@diw.de)  
(Version 2.0, April 2005)

## **Aims of NEWSPELL**

- ⇒ Creation of distinctive calendars
- ⇒ Aggregation of events
- ⇒ Disaggregation of combined events
- ⇒ Re-Definition of time range (begin and end)

## **Data supported by NEWSPELL**

### **Existing SOEP-Data (use with current SOEP-password)**

- PBIOSPE.DAT Combined activity calendars by year
- ARTKALEN.DAT Activity calendars by month
- EINKALEN.DAT Income calendar by month
- BIOMARSY.DAT Marital status by year
- BIOMARSM.DAT Marital status by month
- SOZKALEN.DAT Social assistance by month (households)

### **Self-defined Spell-Data (free use without password)**

## **Output of NEWSPELL**

- ⇒ LOG-File keeps a record of your session
- ⇒ data in spell-format with additional variables
  - previous spell-type
  - next spell-type
  - censoring status
- ⇒ data in time-series format

## **How to use NEWSPELL ?**

- ⇒ Copy all files from folder \NEWSPELL on your hard-disk
- ⇒ Write a simple command-file with any DOS-editor
- ⇒ Start the program with

*NEWSPELL command-file [password]*

## Example:

### original spell-data

```
case persnr spellnr spelltyp begin end
  1  1001      1         2      1    8
  1  1001      2         1      9   12
  1  1001      3         3      8    9
  1  1001      4         3     12   14
  1  1001      5         1     16   18
```

```
/* type 1 = full employed
```

```
/* type 2 = part time employed
```

```
/* type 3 = unemployed
```

# Appendix: NEWSPELL

## command file

```
NI=example.dat      /* a self defined dataset (input)
NS=example.spl      /* output in spell-format
NT=example.tim      /* output in time series format
NL=example.log      /* log-file of session
NR=example.res      /* output-file w/ frequencies of new vars

/* definition of new spell system
/* new type 1 = employed and unemployed
/* new type 2 = only employed
/* new type 3 = only unemployed

1=1 and 3 or 2 and 3 /* disagg. and aggregation: employed
                    and unemployed */
2=1 or 2             /* aggregation
3=3                 /* unemployed

NB=2                /* New range: Begin
NE=14               /* New range: End
```

# Appendix: NEWSPELL

## Output data files:

### time-series data:

```
case persnr begin end #2 #3 ... #14
1 1001 2 14 2 2 2 2 2 2 1 1 2 2 1 3 3
```

### spell data:

```
case persnr spellnr type begin end previous next censor
1 1001 1 2 2 7 -1 1 4
1 1001 2 1 8 9 2 2 1
1 1001 3 2 10 11 1 1 1
1 1001 4 1 12 12 2 3 1
1 1001 5 3 13 14 1 -1 2
```

sensor: (1) uncensored (2) right-censored  
(4) left-censored (5) l+r-censored