SOEP*long*

An Application for the German Socio-Economic Panel Study *(SOEP)*

Arno Simons, Katja Möhring, Peter Krause

Content

- Motivation
  - – General remarks: new views on longitudinal data
- Application
  - – User‘s perspectives: working with longformat data
- Technical Implementation
  - – Using STATA for creating data-files & docu-files
- Outlook
  - – Current stage of work & further steps

1) Motivation

1) New views on longitudinal data

(1) SOEPlong improves working with SOEP data

- New data structure (longform) reduces number of files and variables substantially

(2) SOEPlong provides new perspectives on data

- The current version of SOEPinfo for cross-sectional data indicates all corresponding item lists for each variable at any (survey) year
- SOEPlong minimizes the list of corresponding variables (in principle is each cross-sectional variables only once related to a long variable).

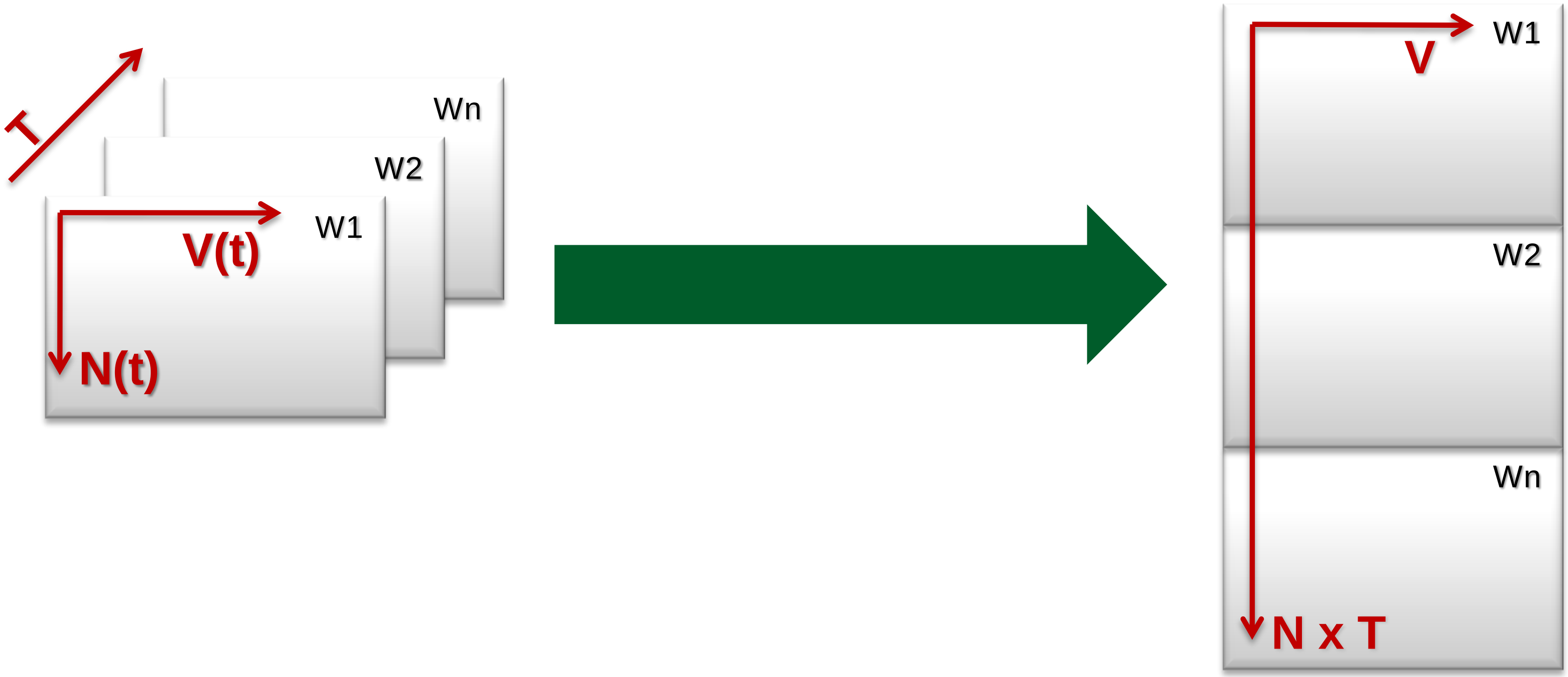
(3) SOEPlong accepts development of instruments

- SOEPlong accepts, that not only the measures (variables) but also the instruments (questionnaire, tests, etc.) may change over time
- SOEPlong provides therefore detailed information about development and consistency of variables over time

1) New data structure

Cross sectional data

Long format



1) SOEPlong – Files

SOEPlong data with fixed structure:

[p/hpfadl p/hbrutto p/hgen pkal pequiv kidl pbr-exit]

• [1984-2008]	C-Files: 154;	L-Files: 10.
• [1984-2008]	C-Vars: ~15.000 ;	L-Vars: ~730.

SOEPlong data with variable structure:

[pl hl; (... lela jugend)]

• [1984-2010]	C-Files: 69;	L-Files: 2.
• [1984-2010]	C-Vars: ~18.000 ;	L-Vars: ~3.200.

1) SOEPlong – L-Vars & IDs

L-Vars:

• Hxxx,	Pxxx,	L-Vars, (org, rec, repl., int;)
• HCxxx,	PCxxx,	L-Vars, (Org-Info in cases of recodes/replace)
• HAxxx,	PAxxx,	L-Vars, (Strings)

IDs

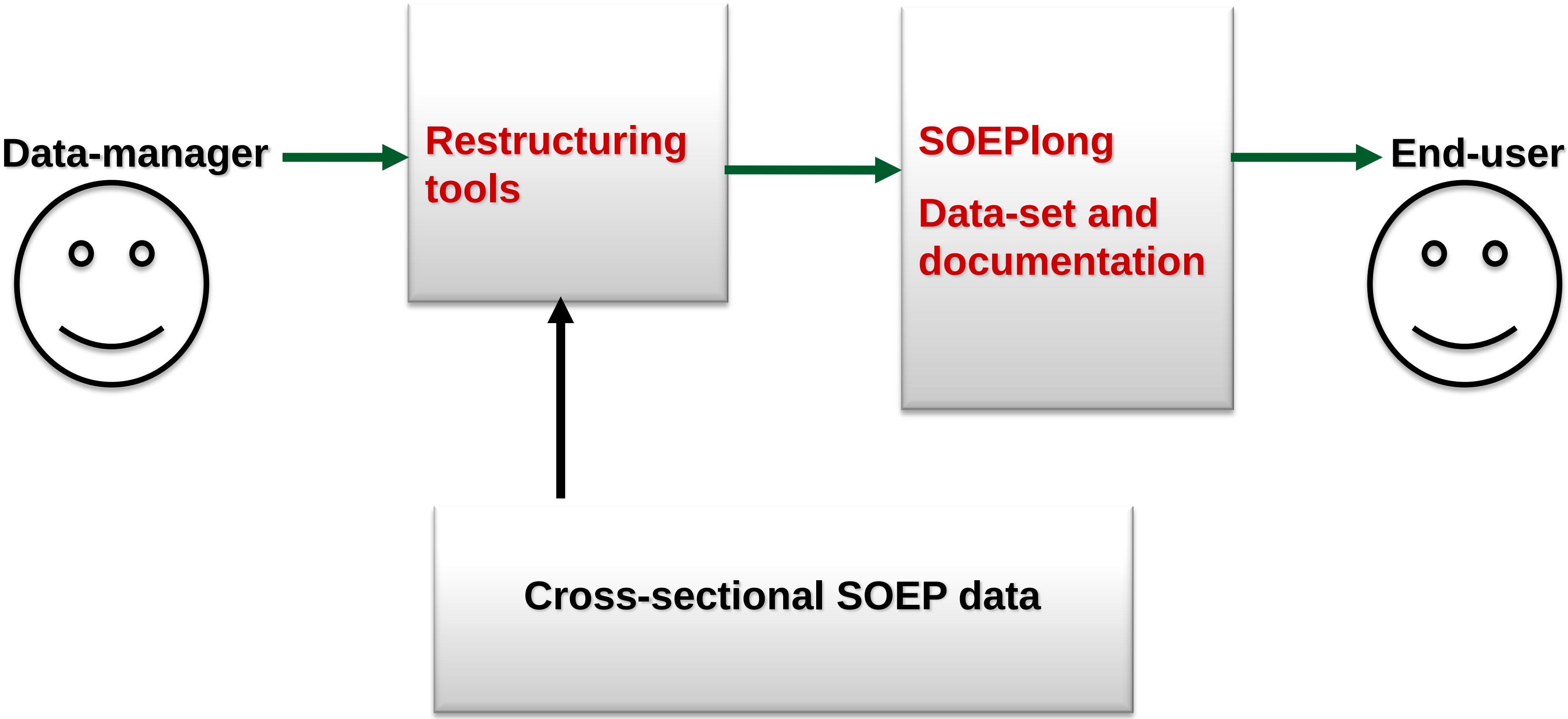
	SOEP(classic):	**SOEP*long*:**
• Erhebungsjahr:	ERHEBJ [c-var]	SVYYEAR [l-var]
• Personen-ID:	PERSNR [c-var]	PID [l-var]
• Haushalts-ID:	_HHNR [c-var]	HID [l-var]
• Ursprungs-ID:	HHNR [c-var]	CASEID [l-var]

1) New perspective

New data structure in long format …includes changes in our perspectives, regarding not only the physical structure of the files but also the analytical approach and resulting information requirements.

- Current representations of time series or panel data usually do not indicate changes in population, distribution, and instruments (questionnaire, tests) over time.

1) New data structure



Data-manager

Restructuring tools

SOEPlong

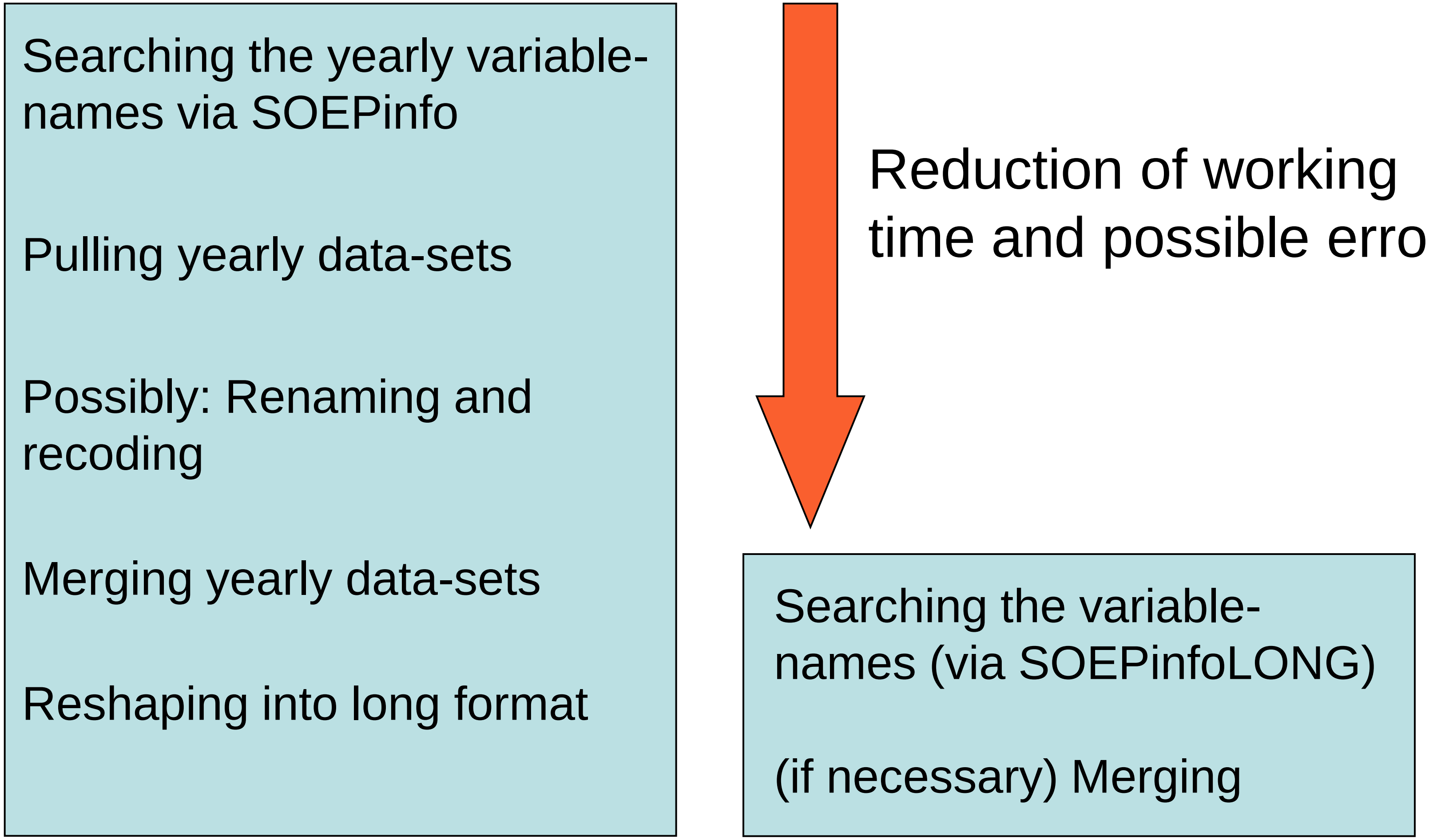
Data-set and documentation

End-user

Cross-sectional SOEP data

2) Application

2) End-user



Searching the yearly variable-names via SOEPinfo

Pulling yearly data-sets

Possibly: Renaming and recoding

Merging yearly data-sets

Reshaping into long format

Reduction of working time and possible errors

Searching the variable-names (via SOEPinfoLONG)

(if necessary) Merging

2) Requirements for usage

Longitudinal perspective: Simplicity + Flexibility

End user:

- Easily accessible data structure
- One data-set with original (but reshaped) and modified variables (*l-org-vars* and *l-vars*)
- Usage with various statistical software packages

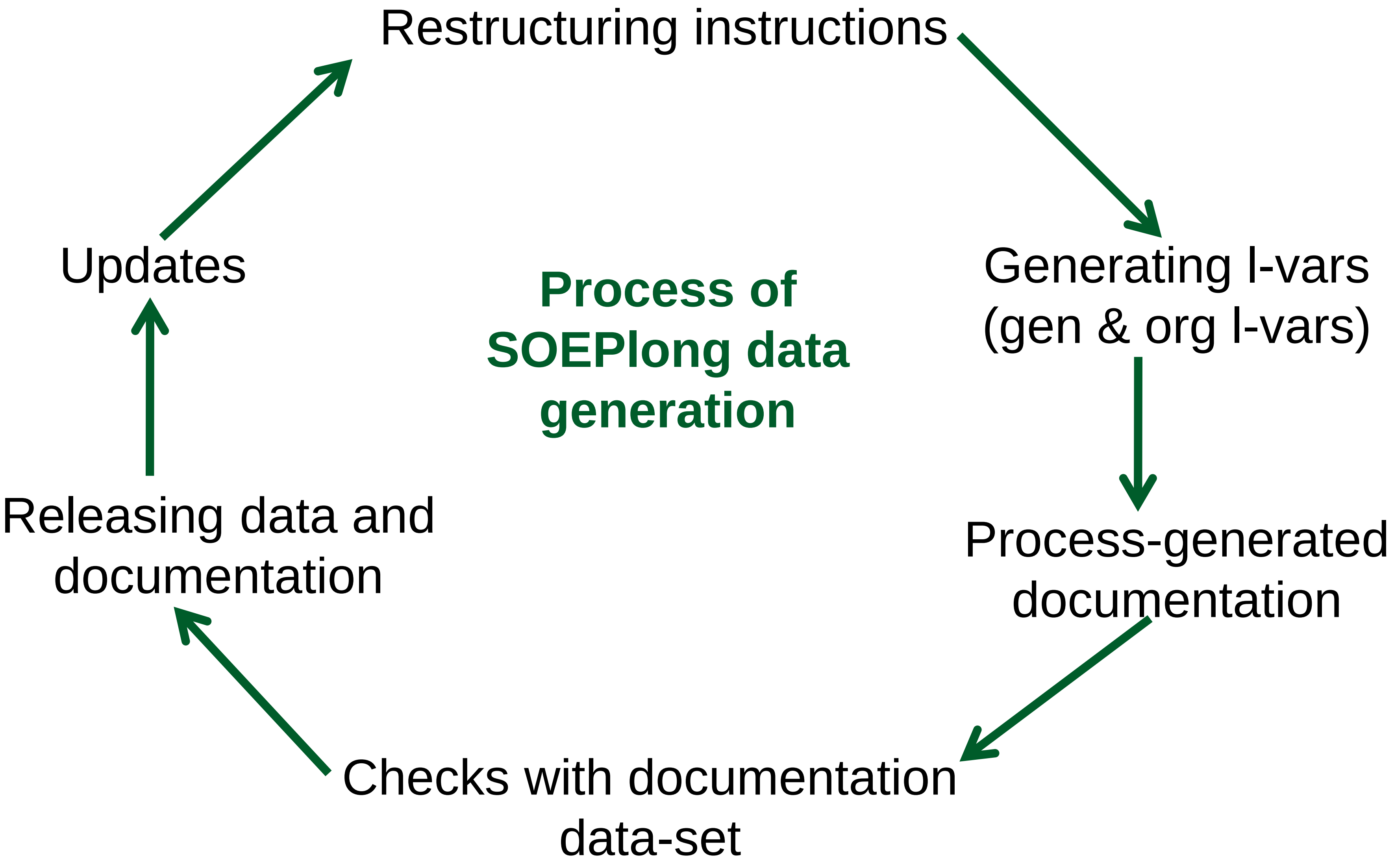
Data manager:

- One solution for data generation and documentation
- Process-generated documentation
- Dynamic updates

2) Rules for data conversion

1. **Harmonize everything** over time
2. **Keep everything** (original and modified variables in one data-set
→ full flexibility for the user)
3. **Document everything** (process-generated and documentation data-set)

2) Steps



Restructuring instructions

Generating l-vars (gen & org l-vars)

Process-generated documentation

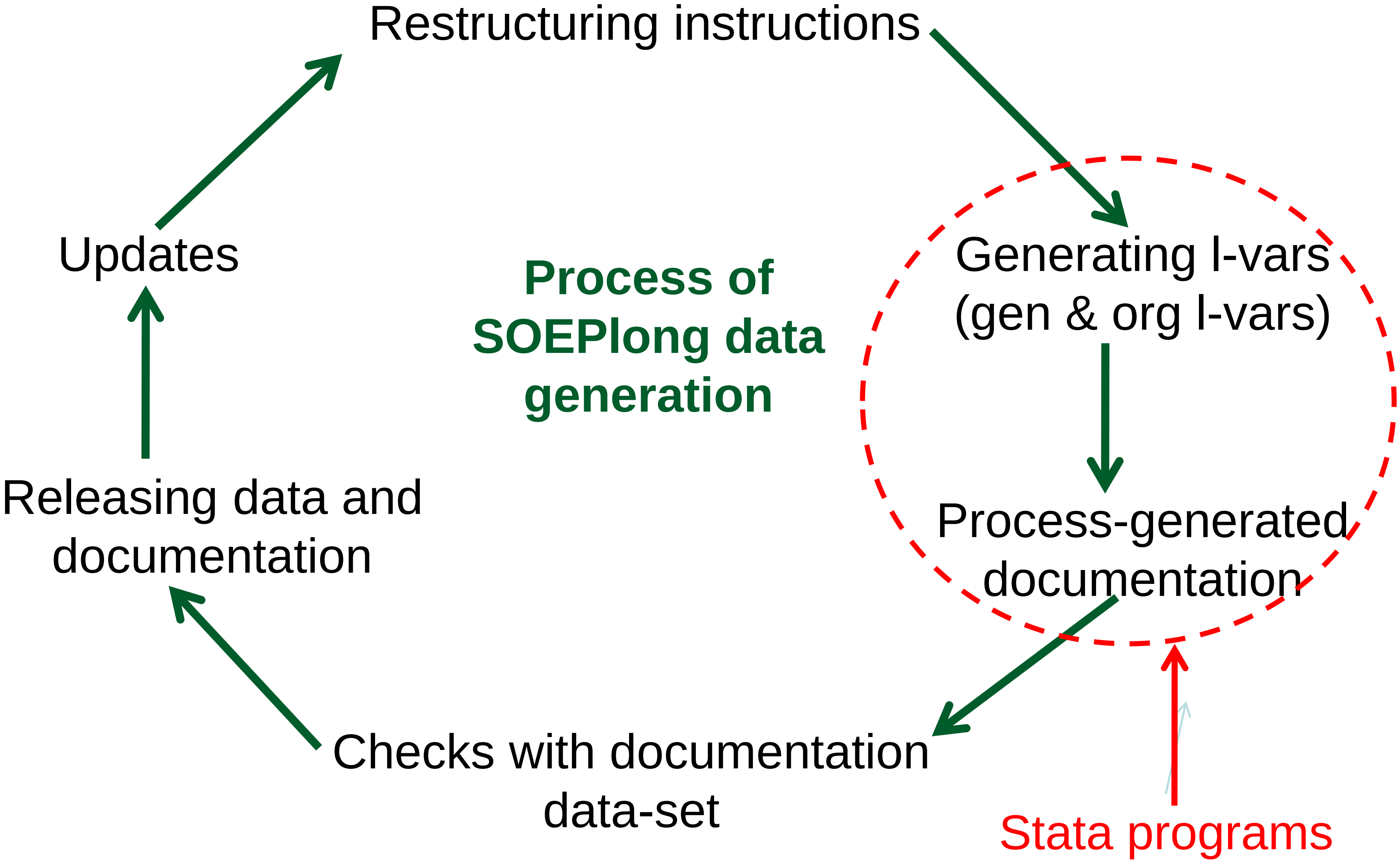
Checks with documentation data-set

Releasing data and documentation

Updates

Process of SOEPlong data generation

2) Steps



Restructuring instructions

Generating l-vars (gen & org l-vars)

Process-generated documentation

Checks with documentation data-set

Releasing data and documentation

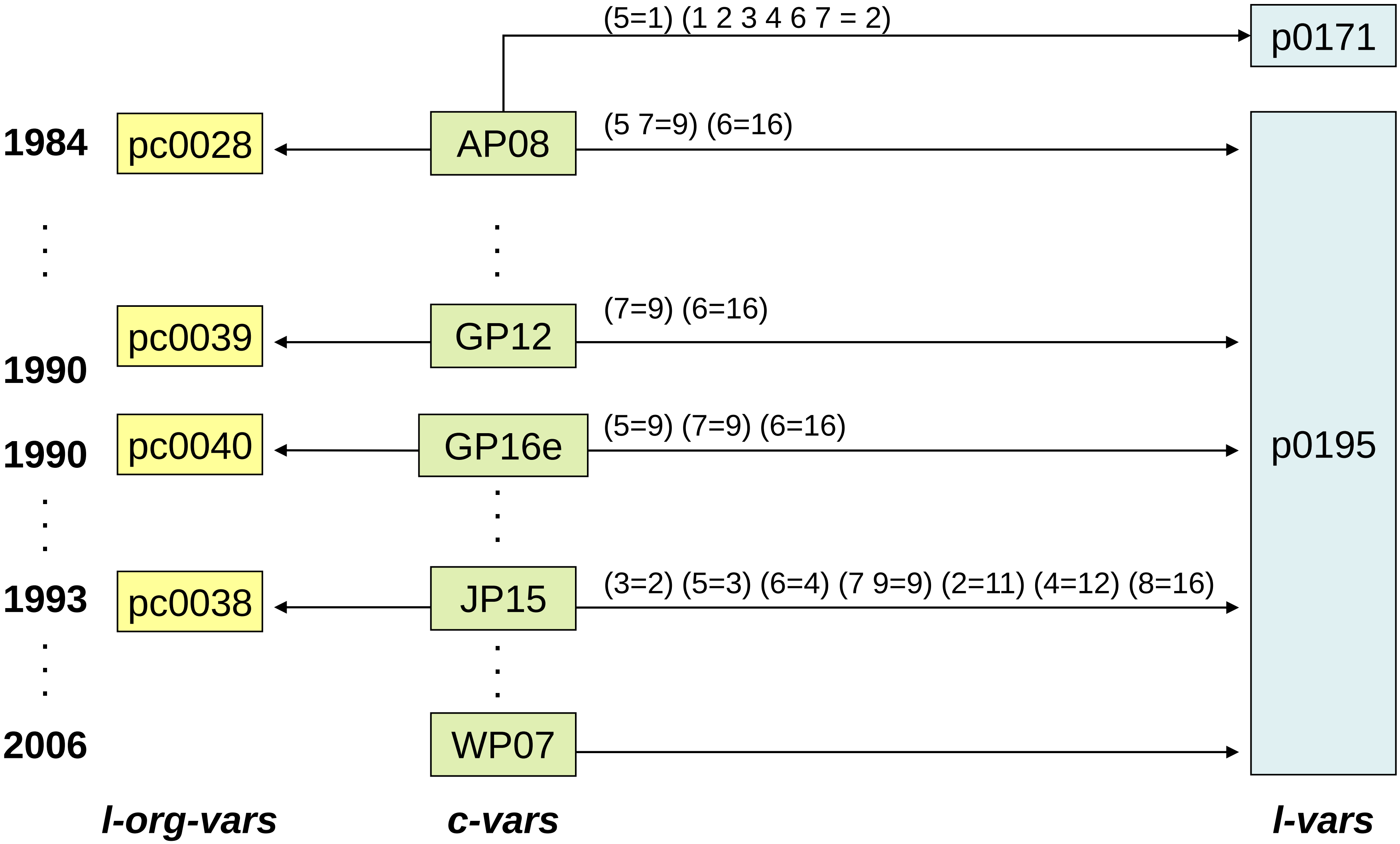
Updates

Process of SOEPlong data generation

Stata programs

3) Technical implementation

3) Vars and recodes (example)



(5=1) (1 2 3 4 6 7 = 2)
p0171

1984 pc0028 AP08 (5 7=9) (6=16)

. . .

1990 pc0039 GP12 (7=9) (6=16)

1990 pc0040 GP16e (5=9) (7=9) (6=16)

p0195

. . .

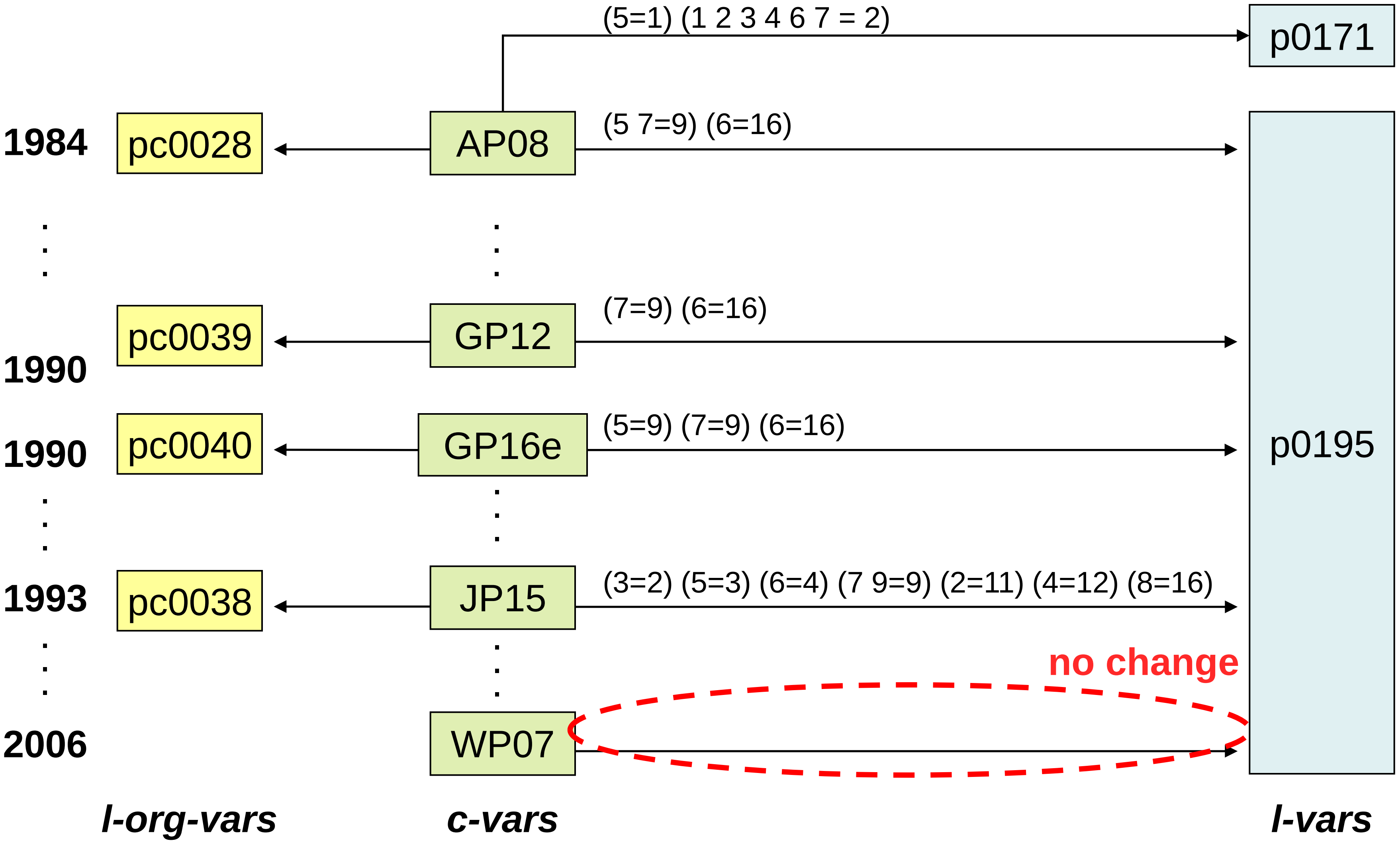
1993 pc0038 JP15 (3=2) (5=3) (6=4) (7 9=9) (2=11) (4=12) (8=16)

. . .

2006 WP07

l-org-vars *c-vars* *l-vars*

3) Vars and recodes (example)



(5=1) (1 2 3 4 6 7 = 2)

p0171

1984 pc0028 AP08 (5 7=9) (6=16)

.
.
.

1990 pc0039 GP12 (7=9) (6=16)

1990 pc0040 GP16e (5=9) (7=9) (6=16)

p0195

.
.
.

1993 pc0038 JP15 (3=2) (5=3) (6=4) (7 9=9) (2=11) (4=12) (8=16)

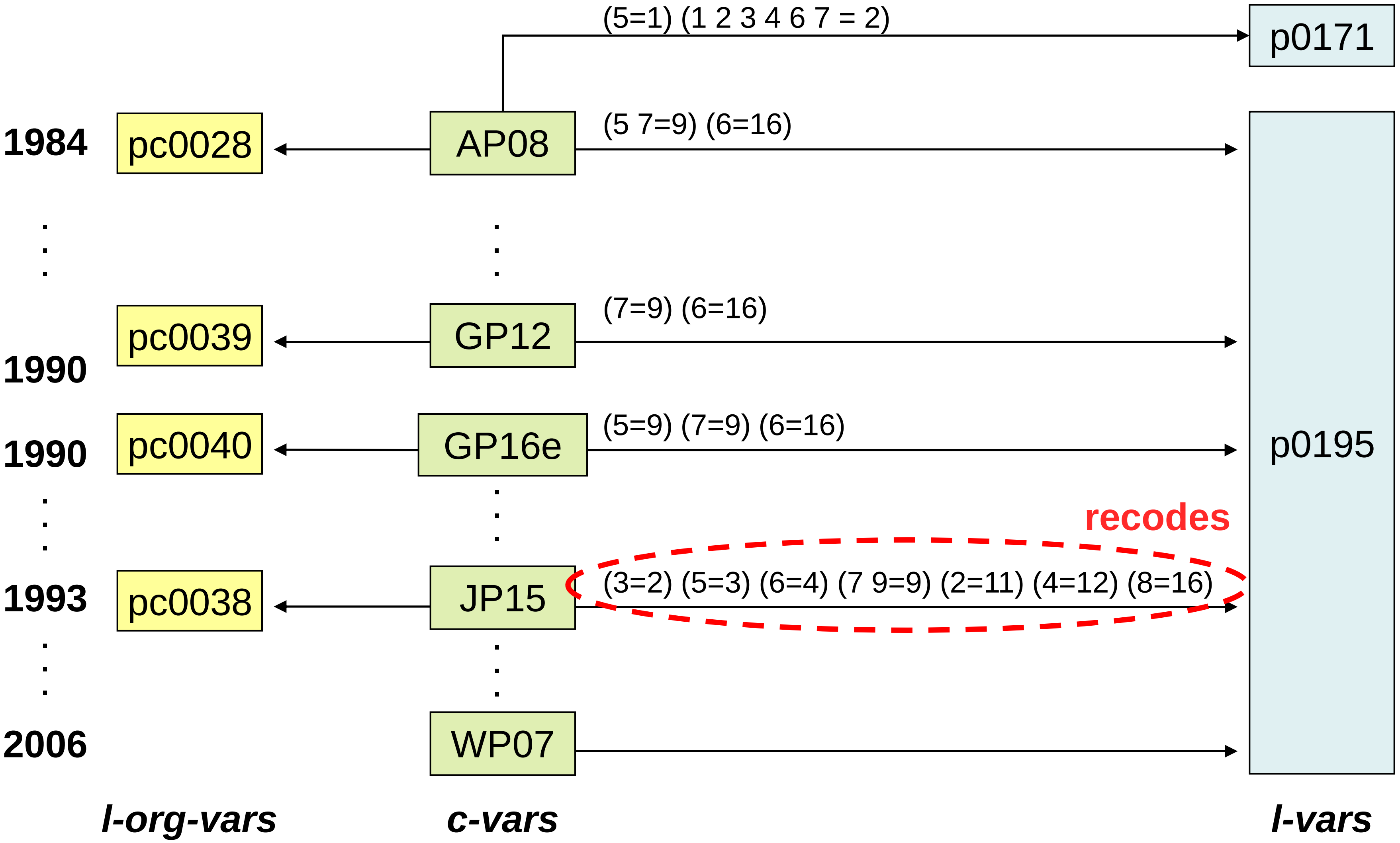
.
.
.

no change

2006 WP07

l-org-vars ***c-vars*** ***l-vars***

3) Vars and recodes (example)



(5=1) (1 2 3 4 6 7 = 2) p0171

1984 pc0028 AP08 (5 7=9) (6=16)

1990 pc0039 GP12 (7=9) (6=16)

1990 pc0040 GP16e (5=9) (7=9) (6=16)

recodes

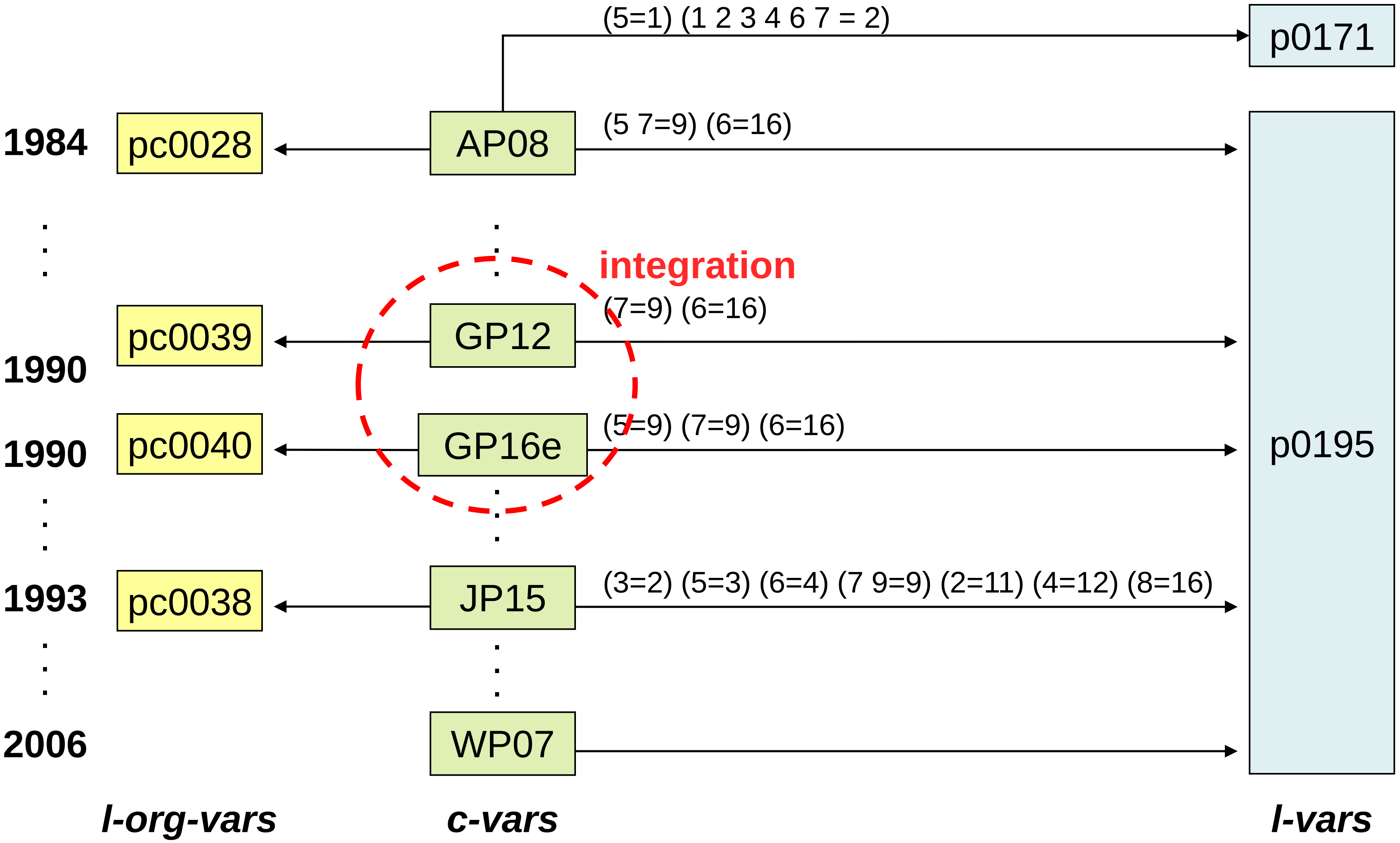
1993 pc0038 JP15 (3=2) (5=3) (6=4) (7 9=9) (2=11) (4=12) (8=16)

2006 WP07

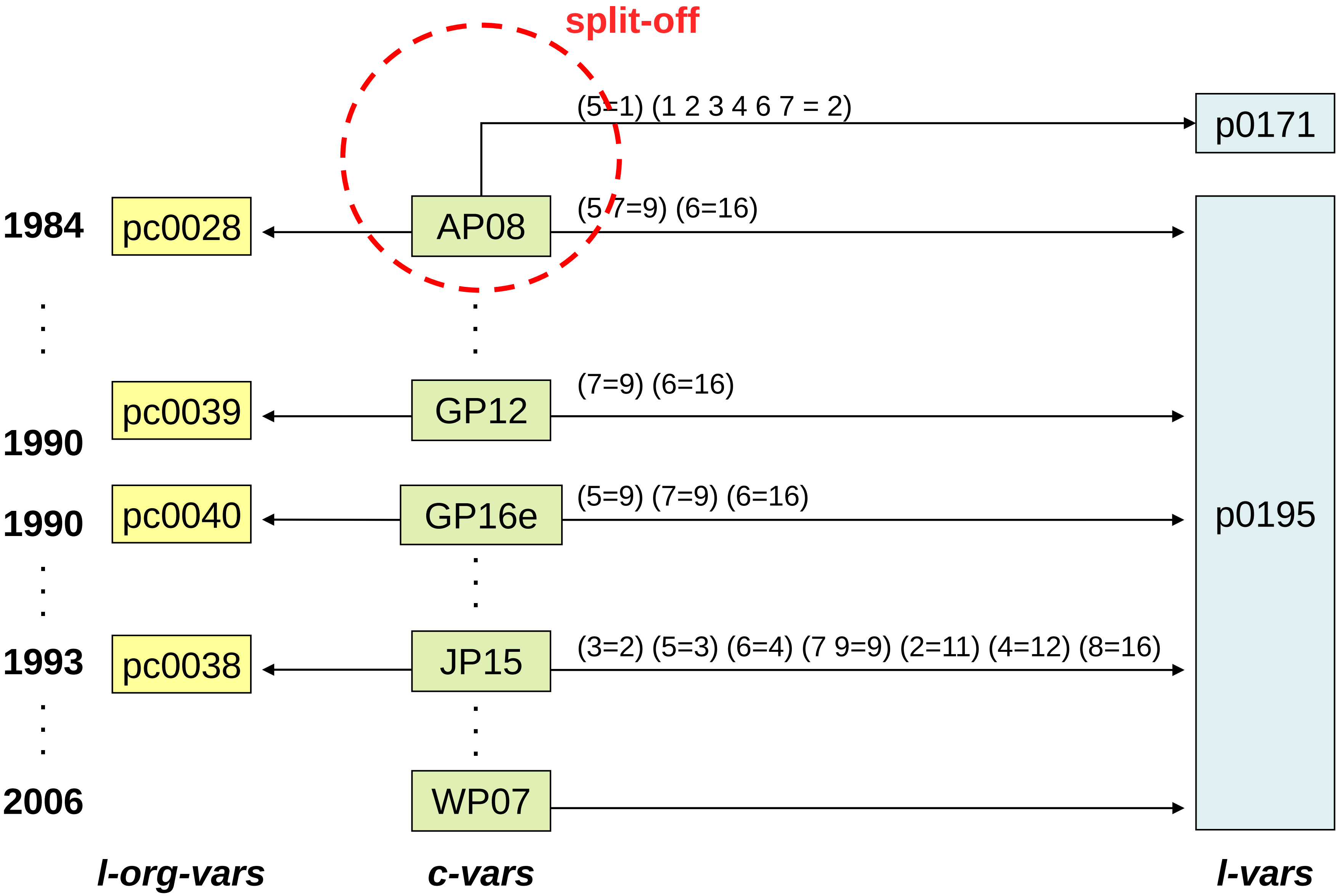
p0195

l-org-vars *c-vars* *l-vars*

3) Vars and recodes (example)



3) Vars and recodes (example)



split-off

(5=1) (1 2 3 4 6 7 = 2)

p0171

1984 pc0028 AP08 (5 7=9) (6=16)

1990 pc0039 GP12 (7=9) (6=16)

1990 pc0040 GP16e (5=9) (7=9) (6=16)

p0195

1993 pc0038 JP15 (3=2) (5=3) (6=4) (7 9=9) (2=11) (4=12) (8=16)

2006 WP07

l-org-vars *c-vars* *l-vars*

3) Program architecture

What we need:

- Original datasets
- Intable
- New command “longform”

What we get:

- Dataset in long format

3) Original datasets (example)

ap.dta	(1984)
gp.dta	(1990)
gpost.dta	(1990)
jp.dta	(1993)
wp.dta	(2006)

3) Intable (example)

(year)	dataset	cvar	lvar	recode
1984	ap	ap08	p0195	(5 7=9) (6=16)
1984	ap	ap08	p0171	(5=1) (1 2 3 4 6 7 = 2)
1984	ap	ap08	pc0028	
1990	gp	gp12	p0195	(7=9) (6=16)
1990	gp	gp12	pc0039	
1990	gpost	gp16e	p0195	(5=9) (7=9) (6=16)
1990	gpost	gp16e	pc0040	
1993	jp	jp15	p0195	(3=2) (5=3) (6=4) (7 9=9) (2=11) (4=12) (8=16)
1993	jp	jp15	pc0038	
2006	wp	wp07	p0195	

3) New command “longform”

Syntax:

longform, path(*anything*) ids(*anything*) [soep]

Requires folder “path” that includes:

- Original datasets
- Intable

3) …interactive!

4) Outlook

STATA-programs

- Improving and generalizing “longform” commands
- Release of ado file “longform”

SOEP – data dissemination

- SOEP*long* files as additional standard data release
- Full (web-based) documentation

Thank you ☺

Appendix

Documentation of SOEP*long* –
Files & Variables (Beta-Release 1984-2009)

Files in SOEP*long* (Beta release, 1984-2009)

RecType No.	SOEP*long* Files	SOEP-Files	Number of Variables	Total In Database
1	PPFADL	Long[ppfad,phrf]	37	611935
2	HPFADL	Long[hpfad,hhrf]	17	245915
3	PPFAD	[ppfad]	18	66189
4	HPFAD	[hpfad]	2	28465
5	CASEINFO	[samp,design]	3	44367
10	PBRUTTO	[a-z][pbrutto]	46	604078
11	PBR_EXIT	[pbr_exit]	43	7264
20	HBRUTTO	[a-z][hbrutto]	57	245915
30	PL	[a-z][p,pausl,post]	2425	421578
40	KIDL	[kidlong]	58	112572
60	HL	[a-z][h,host]	708	220562
80	PGEN	[a-z][pgen]	49	422734
81	PKAL	[a-z][pkal]	251	412922
82	PEQUIV	[a-z][pequiv]	207	553643
90	HGEN	[a-z][hgen]	49	220562

Docu-Files in SOEP*long* for PL and HL (Beta release, 1984-2009)

SOEPlong – DOCU_Years_All **SOEPlong – DOCU_All**

DOCU_ ALL covers all L-Vars

DOCU_Years_ALL contains all L-Vars and C-Vars for all years

The following docu-files are available for households (HL) and individuals (PL):

PL_Docu_years_all	HL_Docu_years_all
PL_Docu_all	HL_Docu_all.

SOEP*Long* – Docu_All

L_Vars	Variables Long-Format	
obsall	N	all years (total)
meanall	Mean	all years (total)
minall	Minimum	all years (total)
maxall	Maximum	all years (total)
num_yr	N of years with valid observations	
maxdiffPC	Indicator for changes in population (% of valid observations)	
maxdiffPC10	maxdiffPC (last 10 years)	
maxdiffCV	Indicator for changes in distribution (Coefficient of Variation)	
maxdiffCV10	maxdiffCV (last 10 years)	
num_recode	N of years with recodes	
num_replace	N of years with replace operations	
split	1=modified vars; 2=original vars; 3=copy of original vars in L-Vars	
label_de	Variable labels (L_Vars)	
topic	[Topics – SOEP*Info*]	

SOEP*Long* – Docu_Years_All

L_Vars	Variables [Long-Format]	
svyyear	Survey year	
obsall	N	all years (total)
meanall	Mean	all years (total)
minall	Minimum	all years (total)
maxall	Maximum	all years (total)
obsyr	N valid observations	for each svyyear
obsyrm	N total observations	for each svyyear
meanyr	Mean	for each svyyear
sdyr	Standard deviation	for each svyyear
minyr	Minimum	for each svyyear
maxyr	Maximum	for each svyyear
recode	Recodes	
replace	Replace	
ost	1=Integration of files for East-German samples	
ausl	1=Integration of files for samples of foreigners	
C_Var	Link to Cross-sectional SOEP variables	
C_OrgVar	Original C-Vars represented in L-Vars	
L_OrgVar	L_Vars with original C_Vars	
L_InputVar	L_Vars, Input-Vars (Ausgangsvariablen im long-format)	
history	Sequence of variable modifications	
split	1=modified vars; 2=original vars; 3=copy of original vars in L-Vars.	
label_de	[Variable labels – as in last C_OrgVar]	

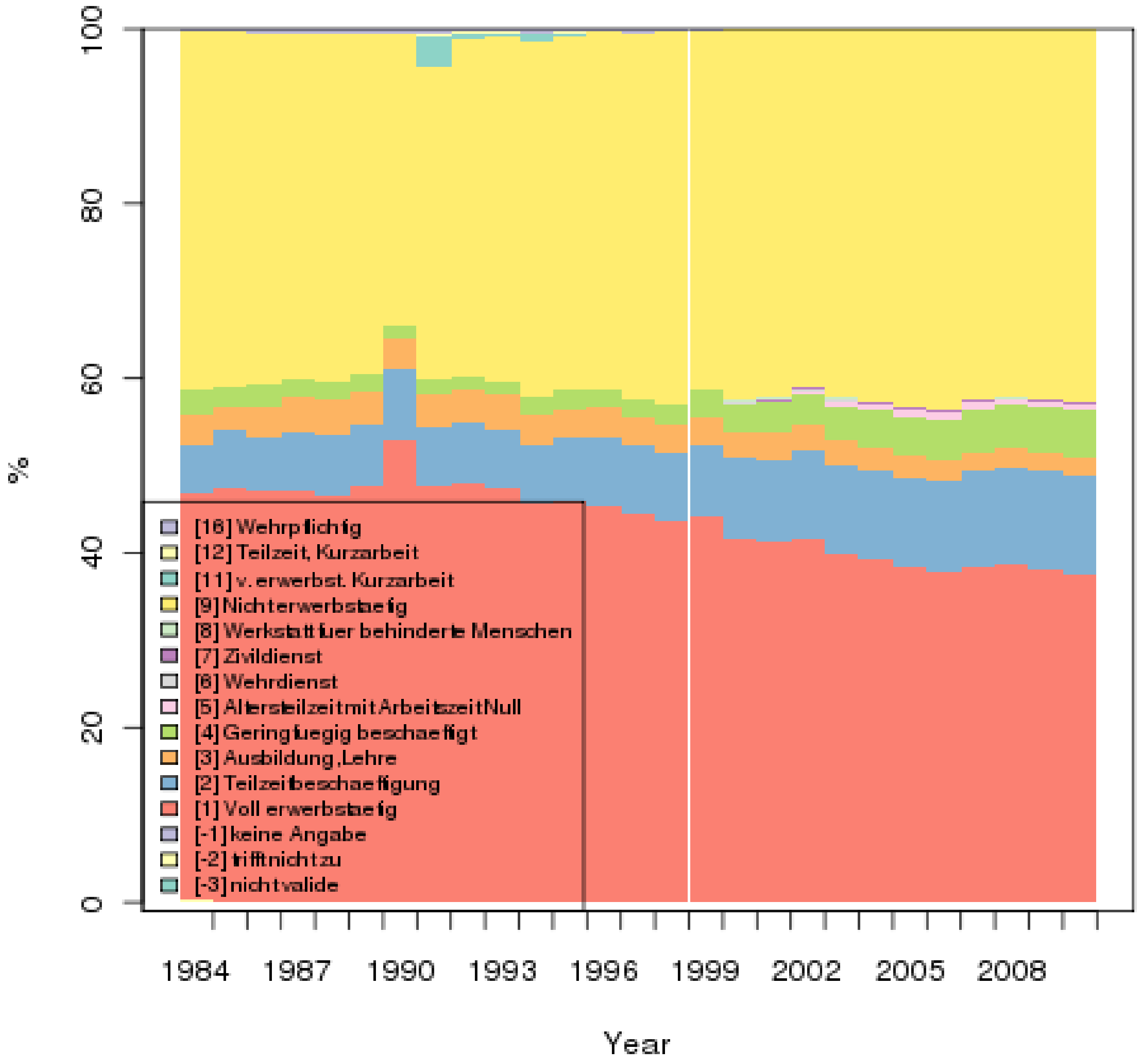
Examples for graphical presentations* of SOEP*long* Variables

*Thanks to Jan Goebel

SOEP Das Sozio-oekonomische Panel

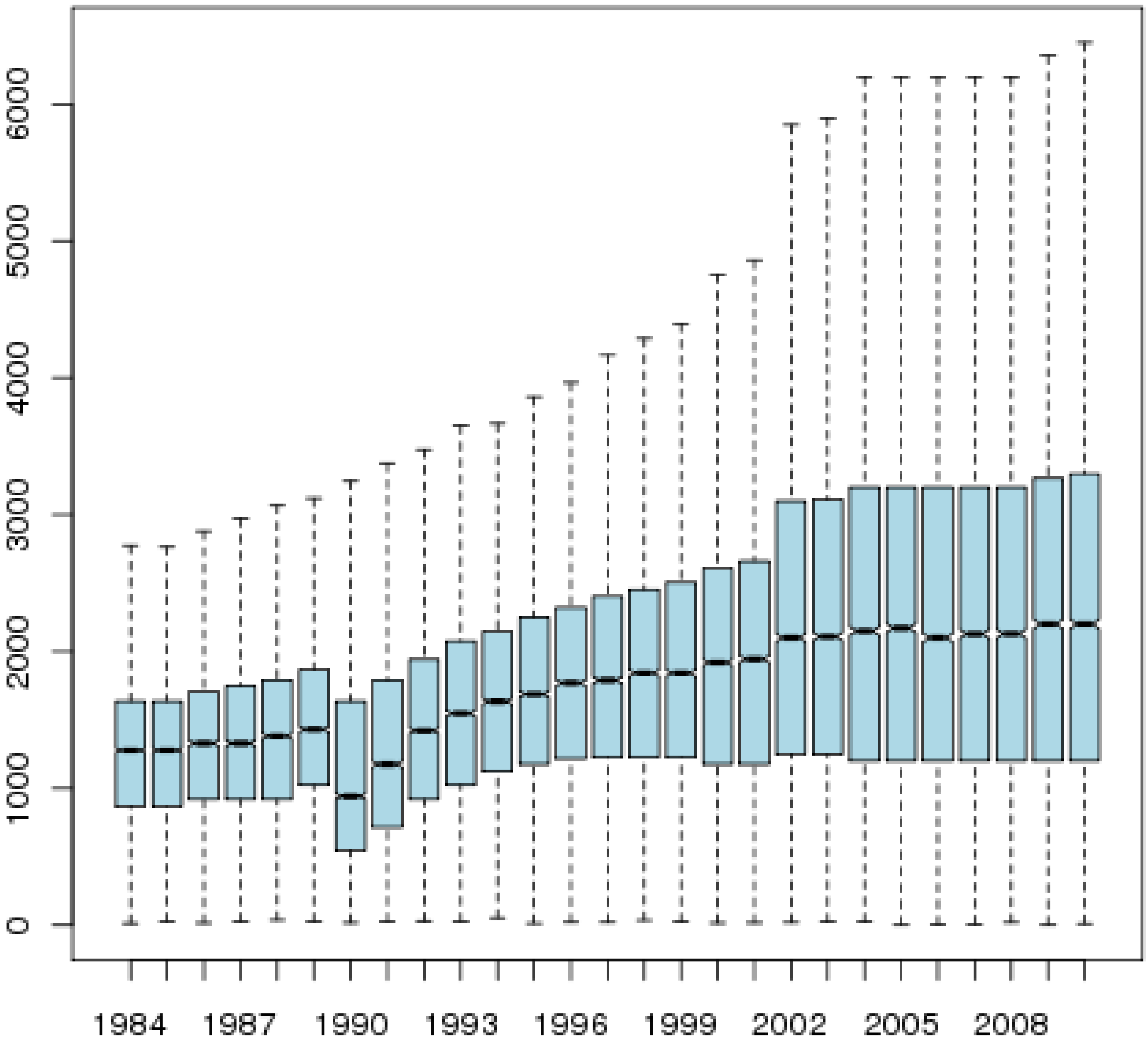
DIW BERLIN

Variable: p0195
Erwerbsstatus



Variable: p0295
Bruttoverdienst letzten Monat

Das Sozio-oekonomische Panel



6000
5000
4000
3000
2000
1000
0
1984
1987
1990
1993
1996
1999
2002
2005
2008