DIW Berlin / SOEP (Ed.)

Research Report

SOEP-Core v34 - PPATHL: Person-related meta-dataset

SOEP Survey Papers, No. 762

Provided in Cooperation with:
German Institute for Economic Research (DIW Berlin)

Suggested Citation: DIW Berlin / SOEP (Ed.) (2019) : SOEP-Core v34 - PPATHL: Person-related meta-dataset, SOEP Survey Papers, No. 762, Deutsches Institut für Wirtschaftsforschung (DIW), Berlin

This Version is available at:
https://hdl.handle.net/10419/203348

SOEP The Socio-Economic Panel

762

SOEP Survey Papers

Series D – Variable Descriptions and Coding

SOEP – The German Socio-Economic Panel at DIW Berlin 2019

SOEP-Core v34 – PPATHL: Person-Related Meta-Dataset

SOEP Group

DIW SOEP

Running since 1984, the German Socio-Economic Panel (SOEP) is a wide-ranging representative longitudinal study of private households, located at the German Institute for Economic Research, DIW Berlin.

The aim of the SOEP Survey Papers Series is to thoroughly document the survey's data collection and data processing.

The SOEP Survey Papers is comprised of the following series:
Series A – Survey Instruments (Erhebungsinstrumente)
Series B – Survey Reports (Methodenberichte)
Series C – Data Documentation (Datendokumentationen)
Series D – Variable Descriptions and Coding
Series E – SOEPmonitors
Series F – SOEP Newsletters
Series G – General Issues and Teaching Materials

The SOEP Survey Papers are available at http://www.diw.de/soepsurveypapers

Please cite this paper as follows:
SOEP Group, 2019. SOEP-Core v34 – PPATHL: Person-Related Meta-Dataset. SOEP Survey Papers 762: Series D – Variable Descriptions and Coding. Berlin: DIW Berlin/SOEP

ISSN: 2193-5580 (online)

DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany

soeppapers@diw.de

The German Socio Economic Panel at DIW Berlin

SOEP-Core v34 – PPATHL: Person-Related Meta-Dataset

SOEP Group

2019

The file ppathl is part of a collection, which is released with doi:10.5684/soep.v34.

Contents

1 General Information

The path datasets should be the building block of any analysis. Path Files indicate the total population at the household and individual level (over time) and provide all IDs necessary to access further files at different levels (Krause/Glass/Reher 2019a,b). Path-Files are delivered in three data formats – in long-format [H|P-PATHL] (as the most comprehensive version including weighting variables), in wide-format [H|P-PFAD] (the traditional version), and in a short-version [H|P-PATH] as a reduced population file (indicating the total of population of households and individuals) Household Level [HID|SYEAR] {Navigation File: H-PATH-L (long-format)} Individual Level [PID|SYEAR] {Navigation File: P-PATH-L (long-format)} Household Level [HID] {Population File: H-PATH} Individual Level [PID] {Population File: P-PATH} Household Level [HID (HHNRAKT)] {Path File: HPFAD (wide-format)} Individual Level [PID (PERSNR)] {Path File: PPFAD (wide-format)}. The constituting SOEP population considers three levels – cases, households, and individuals. Due to the SOEP sampling and survey process, these levels follow an implicit hierarchy. All samples refer to primary source households – indicated by the household id at the time when the survey starts – the (fixed) Case ID [CID]. New Households may emerge from these original households during the longitudinal survey process by split-offs of family members – all (current) households are therefore indicated by a (variable) Household ID [HID]. IDs for individuals living in the households are derived from the households, where they were living when they were surveyed for the first time – the (fixed) Personal ID [PID]. It is recommendet to use the (almost) time-independent (demographic) information like sample membership, sex, year and country of origin are adjusted on a wave-by-wave basis in the framework of demographic testing.

2 Primary Key, Foreign Keys and Sample Information

pid – Never Changing Person ID

A person can be uniquely identified with variable PID. Together with SYEAR primary key in this file.

syear – Survea Year

1984		16252
1985		16737
1986		15868
1987		14974
1988		14596
1989		14000
1990		19666
1991		19713
1992		19552
1993		19240
1994		19469
1995		19947
1996		19527
1997		19064
1998		21175
...	(4 rows omitted)	124175
2003		32937

2004	31540
2005	30339
2006	32747
2007	30962
2008	29005
2009	31642
2010	45977
2011	50329
2012	50120
2013	55611
2014	51684
2015	50277
2016	57287
2017	64845

Together with PID primary key in this file.

hid – ID Household

The household the person belongs to in the corresponding year (SYEAR).
For more information, contact: Peter Krause (Tel. 030-89789-690)

psample – Sample Member

1	[1] A 1984 Initial Sample (West)	270707
2	[2] B 1984 Migration (until 1983, West)	94170
3	[3] C 1990 Initial Sample (East)	125775
4	[4] D 1994/5 Migration (1984-1994, West)	25151
5	[5] E 1998 Refreshment	26652
6	[6] F 2000 Refreshment	161674
7	[7] G 2002 High Income	33947
8	[8] H 2006 Refreshment	27130
9	[9] I 2009 Innovation Sample	7130
10	[10] J 2011 Refreshment	39534
11	[11] K 2012 Refreshment	17385
12	[12] L1 2010 Birth Cohort (2007-2010)	51982
13	[13] L2 2010 Family Type (Low-Income, Single-Parent, Large Families)	56273
14	[14] L3 2011 Family Type (Single-Parent, Large Families)	21978
15	[15] M1 2013 Migration (1995-2011)	36490
16	[16] M2 2015 Migration (2009-2013)	8911
17	[17] M3 2016 Refugee (2013-2015)	9806
18	[18] M4 2016 Refugee/family (2013-2015)	14126
19	[19] M5 2017 Refugee (2013-2016)	4771
20	[20] N 2017 Refreshment (PIAAC-L)	5665
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

The sample membership never changes.
For more information, contact: Peter Krause (Tel. 030-89789-690)

cid – Case-ID, Original Household Number

Case Id - Household Source Identifier. The fixed household source id points to the first SOEP ancestor household for this person. The person does not necessarily need to have ever lived in a household with this number. This is relevant for sampling information and calculating the weights.

persnr – Never Changing Person ID

Same as PID.

3 Survey History

eintritt – Year First Contacted, Netto=10-99

1984		249816
1985		6241
1986		7546
1987		7055
1988		6992
1989		7371
1990		104540
1991		9165
1992		8237
1993		7907
1994		17142
1995		16949
1996		7252
1997		6486
1998		29902
...	(4 rows omitted)	200005
2003		7421
2004		6212
2005		5389
2006		30115
2007		4605
2008		3726
2009		10354
2010		104405
2011		62316
2012		20884
2013		38966
2014		3994
2015		11151
2016		24376
2017		12737

The year a person joined the SOEP.

For more information, contact: Peter Krause (Tel. 030-89789-690)

erstbefr – Year First Surveyed, Netto=10-99

1984		186751
1985		7700
1986		8316
1987		7617
1988		6640
1989		7231
1990		77648
1991		8640
1992		8231
1993		8034
1994		14348
1995		14656
1996		8159
1997		8015
1998		25628
...	(5 rows omitted)	179330
2004		9555
2005		9243
2006		27482
2007		8795
2008		6692
2009		10734
2010		52532
2011		44155
2012		20254
2013		28778
2014		8274
2015		11822
2016		15262
2017		12731
-2		196004

The year of a person's first interview.
For more information, contact: Peter Krause (Tel. 030-89789-690)

austritt – Year Of Last Contact, Netto=10-99

1985	2607
1986	3906
1987	3068
1988	4819
1989	5029
1990	3531
1991	4369
1992	5143

1993		6742
1994		7955
1995		8242
1996		8674
1997		8088
1998		10553
1999		11483
...	(3 rows omitted)	41613
2003		15900
2004		15652
2005		16262
2006		24635
2007		24423
2008		25256
2009		38212
2010		38508
2011		42904
2012		31643
2013		36893
2014		34484
2015		41757
2016		45461
2017		471445

The last year of a person's SOEP appearance.
For more information, contact: Peter Krause (Tel. 030-89789-690)

letztbef – Year Of Last Survey, Netto=10-99

1984		3434
1985		3169
1986		3042
1987		4404
1988		4128
1989		4065
1990		4541
1991		5358
1992		5940
1993		7159
1994		7286
1995		7103
1996		8172
1997		10413
1998		11002
...	(5 rows omitted)	69661
2004		17123
2005		21620
2006		21603
2007		23728
2008		25646

2009	27670
2010	30194
2011	38827
2012	28272
2013	32066
2014	32181
2015	34808
2016	41313
2017	309325
-2	196004

The year of a person's most recent interview.
For more information, contact: Peter Krause (Tel. 030-89789-690)

netto – Current survey status

10	[10] Interviewee With Succesful Interview (_P)	535054
12	[12] Individual Questionnaire And Person Biography	69329
13	[13] Individual Questionnaire And Youth Biography	318
14	[14] Individual Questionnaire And Other Questionnaires	32
15	[15] Individual Questionnaire And Experiments, Test	40792
16	[16] Individual Questionnaire, First Time Surveyed, Age 17	5946
17	[17] Youth Biography First Time Surveyed, Age 17	5496
18	[18] Individual Questionnaire And Child under age 17	8
19	[19] Individual Questionnaire Without Household Interview	626
20	[20] Children in Succesfully Interviewed Households (_Kind)	180175
21	[21] Children With Mother-Child Questionnaire_I, Age 0-1	6075
22	[22] Children With Mother-Child Questionnaire_II, Age 2-3	6350
23	[23] Children With Mother-Child Questionnaire_III, Age 5-6	6096
24	[24] Children age 7-8, with parental questionnaire	5479
25	[25] Children age 9-10, with parental questionnaire	5096
...	(27 rows omitted)	159442
90	[90] Individual Dropouts PBR_EXIT	4234
91	[91] Moved abroad	2447
92	[92] Moved abroad (abroad)	177
93	[93] Moved abroad (exit)	65
94	[94] Person Gap with advices	424
97	[97] advice to dead person (exit)	981
98	[98] advice to dead person (_VP)	201
99	[99] Has Died	4390
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	24
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

This variable indicates available information and files for the entire SOEP individuals. Netto-codes 10-19 (and 29) define the respondents population of PGEN, the codes 20-28 indicate

children, 30-39 unit-non-responses in partially realized households, and the codes 90-99 describe permanent (or temporary) dropouts. Further differentiations point to the survey instruments (questionnaires). The Codes 10-39 describe the population in realized (and partially realized households).
For more information, contact: Peter Krause (Tel. 030-89789-690)

nett1 – Current survey status (old 1 digit)

0	[0] Person Gap PBR_EXIT	11895
1	[1] Successful Interview _P, _JUGEND	656971
2	[2] Below Survey Age _KIND	215396
3	[3] Did Not Participate _PBRUTTO	142355
4	[4] Missing This Wave _PLUECKE	9298
5	[5] Interviewee Without Household Interview	630
-1	[-1] No Answer	0
-2	[-2] Does not apply	2649
-3	[-3] Answer improbable	63
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

Short version of NETTO-variable
For more information, contact: Peter Krause (Tel. 030-89789-690)

casemat – Case-Match, combined panel households

0	[0] CASE With HH Details	112
20605		2
20613		5
27367		14
27430		7
250996		3
272906		6
277908		1
283924		9
291102		7
292338		11
292621		7
344095		9
700495		3
701564		13
701670		3
701718		1
863637		1
3014079		2
3094650		5
3499749		2
3635481		3

-1	[-1] No Answer	0
-2	[-2] Does not apply	1039031
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

It is possible that Individuals from different original households (CID) move together in one common household. Then people with identical values for (HID) in one wave may have different values for CID. Only for those persons moving together CASEMAT contains the HID of the other household members. This information is not relevant when linking person and household data based on the current household number HID.
For more information, contact: Peter Krause (Tel. 030-89789-690)

piyear – Jahr des Interviews (Interview Year)

1984		16252
1985		16361
1986		15548
1987		14633
1988		14254
1989		13689
1990		19427
1991		19414
1992		19147
1993		18833
1994		19101
1995		19486
1996		19108
1997		18785
1998		20689
...	(6 rows omitted)	186330
2005		29908
2006		32202
2007		30460
2008		28453
2009		29655
2010		45151
2011		50254
2012		49401
2013		54819
2014		50734
2015		49465
2016		56395
2017		57897
2018		5654
-2		17752

Interview Year (indicates personal interviews realized also outside of standard SYEAR)

4 Basic Demographic Information

sex – Gender

1	[1] Male	510495
2	[2] Female	528406
-1	[-1] No Answer	356
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

Respondent's (last) sex, plausibility longitudinally validated.
For more information, contact: Peter Krause (Tel. 030-89789-690)

gebjahr – Birth Year, 4-digit

1882		2
1888		4
1892		19
1893		8
1894		13
1895		22
1896		65
1897		58
1898		54
1899		156
1900		176
1901		162
1902		304
1903		201
1904		363
...	(100 rows omitted)	962105
2005		8610
2006		7840
2007		10196
2008		9651
2009		8275
2010		8210
2011		4305
2012		3728
2013		3211
2014		2451
2015		1933
2016		1305
2017		469
2018		2
-1		5359

Respondent's year of birth, plausibility longitudinally validated.
For more information, contact: Peter Krause (Tel. 030-89789-690)

gebmonat – Month Of Birth

1	[1] January	77546
2	[2] February	70516
3	[3] March	76023
4	[4] April	69399
5	[5] May	72215
6	[6] June	67279
7	[7] July	72848
8	[8] August	70580
9	[9] September	71232
10	[10] October	68619
11	[11] November	61881
12	[12] December	65429
-1	[-1] No Answer	89822
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	13
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	105855
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

The month of birth

- was asked starting with wave W (2006) in the mother-child-questionnaire for newborns
- was asked starting with wave T (2003) in supplementary biography questionnaire, resulting in file BIOL
- was asked in wave S individual questionnaire, resulting in file SP
- is recorded for all children within the file TKIND (wave T, 2003)
- can be approximately derived for newborn children from the month of moving into the household, stored in file PBRUTTO
- can be reported by parents in the personal questionnaire (which might simultaneously establish a link to the child), stored in file PL

whereas the former information is prefered over the latter. This means the generated information (from TKIND, PBRUTTO or PL) will only be utilized if no further, questionnaire based information for the month of birth is available. The generated month of birth could only be constructed for people who were born while their parents were members of the SOEP. Several adjustments and tests of the generated data have been done which showed that – in the cases in which the generated data was also collected by PL, BIOL or $KIND – the data generation is almost always congruent with the collected data and therefore has proven to be reliable. The used source of information is stored in GEBMOVAL.
While this provides the relevant information for most of the current panel members, the information remains missing for some persons including temporary dropouts or people who exited in a previous wave. For some of them the month of birth could be reconstructed. This reconstruction remains an approximation and might differ from the true month of birth in individual cases.
For more information, contact: Christian Schmitt (Tel. 030-89789-603)

gebmoval – Month Of Birth, Data Source

1	[1] Generated from gebmonth (parents)	8708
2	[2] Ppfad, carry forward	0
3	[3] $kind, Info from mother	65265
4	[4] Info From Sp	406537
5	[5] Info From $lela	216744
6	[6] Info From bioage$$ (mother)	64888
7	[7] Info from $PAGE17	81425
-1	[-1] No Answer	89822
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	13
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	105855
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

Indicates the data source for the month of birth (GEBMONAT).
For more information, contact: Christian Schmitt (Tel. 030-89789-603)

todjahr – Year Died, 4 Digits

1984		14
1985		153
1986		298
1987		404
1988		527
1989		636
1990		592
1991		865
1992		923
1993		1156
1994		1085
1995		1663
1996		1585
1997		1486
1998		1682
...	(6 rows omitted)	13956
2005		3767
2006		3719
2007		3350
2008		4259
2009		2765
2010		2825
2011		2513
2012		2289
2013		3258
2014		2919
2015		2775
2016		3138

2017	3498
-1	357
-2	970800

The variable TODJAHR contains the four-digit year of death for persons whose death could be firmly established or a missing value code:

- (-2): persons, for whom it is unknown whether they are deceased (that is, both persons still living up to that wave, and persons whose exact whereabouts is unknown and have dropped out of SOEP)
- (-1): persons, for whom the fact of death is known, but the year of death is unknown.

todinfo – Year Died, Information Source

1	[1] From Annual Survey (pbr_exit)	53131
2	[2] survey about died person ($v)	276
3	[3] survey about parents ($lela)	16
4	[4] Infratest drop-out study 1992	17
5	[5] Infratest drop-out study 2001	4044
6	[6] Infratest drop-out study 2007	33
7	[7] Infratest drop-out study 2008/9	8173
8	[8] Modul Family changes [P]	2410
-1	[-1] No Answer	357
-2	[-2] Does not apply	970800
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

For all persons who have been identified as deceased over the course of SOEP, the variable TODINFO gives the source of this information.
53 persons were identified as deceased in the Infratest Field Organization Study (Follow-up study of drop-outs between 1984 and 1992) carried out from April – June 1992.
In the framework of the Infratest Field Organization Study (follow-up study of drop-outs) of 2001, a total of over 700 persons were identified as deceased. Among them were several with multiple entries for year of death, that is, persons who were already identified as deceased in the standard wave-to-wave follow-up procedure (stored in the file PBR_EXIT) or in the Infratest Field Organization Study of 1992. A generally very high level of correspondence was found between the information given in the standard follow-up procedure and the point of death established ex-post in the Infratest Field Organization Studies. For ten persons, the year of dropping out of SOEP was used to impute the missing year of death. In the third of those follow-up studies which has been conducted in 2007, another 21 individuals were identified as deceased between 2001 and 2005. For 18 of those persons a valid year of death could be investigated, for the remaining three observations for which the exact year of death is unknown, TODJAHR has been set to the standard missing code "-1".
When the data from the Infratest Field Organization Study contradicted the data from PBR_EXIT, the data from the Field Organization Study was used.

germborn – Born in Germany

1	[1] born in Germany or immigr.<1950	870885
2	[2] not born in Germany	168372
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

The SOEP data comprises a sizeable number of immigrants to Germany and their descendants. Several user-friendly variables identify these groups (GERMBORN, CORIGIN, IMMIYEAR, MIGBACK) and thus give information on the migration background of all persons who have ever been a part of a SOEP household (i.e., the population from PPATH). In addition, GERMBORNINFO, CORIGININFO, IMMIYEARINFO and MIGINFO indicate the quality of information given in GERMBORN, CORIGIN, IMMIYEAR and MIGBACK, respectively. Information for these variables is collected primarily from the wave-specific individual questionnaires ($P, $PAUSL) or the variations of the "biography / life history" questionnaires (in BIOL) and from the additional 16-17-year-old questionnaire in use since 2000 (JUGENDL). In addition, information from the electronic household protocol for M1 was used (answered by the household head and not included in the standard data distribution).
GERMBORN specifies whether a person was born in Germany or in another country. Persons who immigrated to Germany before 1950 are considered as being born in Germany (the Federal Republic of Germany was founded in 1949; see also IMMIYEAR). To code GERMBORN, all relevant information (see documentation on the biographical data) available on persons who have ever been a part of a SOEP household (i.e., the population from PPATH) was combined. The vast majority of persons who have ever been part of a SOEP household gave consistent information on their country of birth and GERMBORN was coded accordingly to the respondents' answers for the PPATH population. For part of the PPATH population, "no direct information" or "inconsistent information" on the person's country of birth was available (see GERMBORNINFO) and additional indicators were used to code the GERMBORN values. In this process, information on a respondent's citizenship and their parents' migration biography were additionally used. We coded the values on GERMBORN in the following order (with descending priority):

1. First, mothers' immigration history and their place of residence at the time of the respondents' birth were taken into account to determine the respondents' probable country of birth. For instance, when a respondent was born after or in the year of their mother's immigration to Germany, the respondent is considered to have been born in Germany. For the coding of a few cases, more detailed information on respondents' month of birth and mother's immigration month was available and used. When a mother's immigration year was missing, the father's immigration history was used to code a respondent's country of birth.

2. In the next step, GERMBORN was coded for the remaining "inconsistent information" cases. Respondents' information on their country of birth, their citizenship, and parental information was taken into account to identify a respondents' country of birth. While last year the latest information was considered more reliable, this year the mode

was calculated for inconsistent information on respondents' and parental country of birth. In case of varying modes, higher values were given a preference when coding, to be more sensible to foreign countries of birth. For instance, a respondent who reported being born in Germany more often than being born abroad (country of birth), who had German citizenship (citizenship), and whose parents reported more often to be born in Germany than being born abroad (parental information) was considered to have been born in Germany.

3. In a last step, GERMBORN was coded for the remaining "no information" cases. Respondents' citizenship and parental information was used to approximate their most likely country of birth. By definition, information on their country of birth was missing. In contrast to last year, the mode of parents' country of birth and citizenship was used for the coding of GERMBORN, too. For instance, respondents with German citizenship whose parents reported more often to be born in Germany than being born abroad were coded as being born in Germany.

For a few PPATH cases, the new generation procedure led to a change of the GERMBORN value. To provide the highest level of transparency possible, we include a variable for the quality of information used to create the GERMBORN variable: GERMBORNINFO. For persons who according to GERMBORN were not born in Germany, the variables CORIGIN and IMMIYEAR designate the country of origin and the year of immigration to Germany, respectively. More detailed information on this variable can be found in the documentation on the biographical data in the section on PPATH.
For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

germborninfo – Germborn: Quality of information

1	[1] consistent information	815490
2	[2] inconsistent information	32753
3	[3] no answer	191014
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

GERMBORNINFO indicates the quality of information given in GERMBORN. As in previous years, all relevant information available on persons who have ever been a part of a SOEP household (i.e., the population from PPATH) was combined into a working dataset and compared to code GERMBORN. When information in this working dataset consistently indicated that a person was born either in Germany or abroad, GERMBORNINFO was coded with a (1) for "consistent information". When inconsistent or no direct information on a SOEP person was available, GERMBORNINFO was coded with a (2) indicating "inconsistent information" or with a (3) indicating "no information". GERMBORNINFO is thus an indicator of the quality of information given in GERMBORN. More detailed information on this variable can be found in the documentation on the biographical data in the section on PPATH.
For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

birthregion – Birth place: German Federal Land

1	[1] Schleswig-Holstein	8058
2	[2] Hamburg	5454
3	[3] Lower Saxony	29099
4	[4] Bremen	2535
5	[5] North Rhine-Westphalia	60769
6	[6] Hesse	18934
7	[7] Rhineland-Palatinate	14781
8	[8] Baden-Wuerttemberg	33954
9	[9] Bavaria	44320
10	[10] Saarland	3213
11	[11] Berlin	12394
12	[12] Brandenburg	14609
13	[13] Mecklenburg-West Pomerania	9693
14	[14] Saxony	30982
15	[15] Saxony-Anhalt	17732
16	[16] Thuringia	16258
-1	[-1] No Answer	0
-2	[-2] Does not apply	716472
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

BITHREGION contains information about the German Federal State ("Bundesland") a person was born. In 2012 the SOEP asked all current respondents about the place of birth: "Where were you born? If there are other towns or cities with the same name, or if the town is very small, please state the nearest city. Please write the name of the town in the left blank and any additional information in the right blank. For example, write 'Düsseldorf', 'Frankfurt an der Oder', or 'Frankfurt am Main' in the left blank, and in the case of 'Roßdorf bei Schmalkalden', write 'Roßdorf' in the left blank and 'bei Schmalkalden' in the right blank. Since then this question has been part of the biography questionnaire and a variable BIRTHREGION is povided in dataset PPATH, which has to be updated each year for new respondents. The answers is given in clear text and coded by Kantar at the level of municipalities for german cities or villages (including the geocodes for the city center). For places outside Germany, Kantar provides only the geocodes, if possible. However, the reponses could not all be assigned to a unique municipality, therefore multiple municipality codes are provided by Kantar (up to 19 in 2012). For the variable *birthregion* in *ppfad* only those answers are used, where a unique assignment of a German Federal State ("Bundesland"), based on the possible municipality codes, was possible. For persons born in a SOEP household (the household was responding in this year) the code of the respective Federal State of this year is used.
For more information, contact: Jan Goebel (Tel. +49 30-89789-377)

corigin – Country Born In

1	[1] Germany	870885
2	[2] Turkey	27165
3	[3] Ex-Yugoslavia	4438

4	[4] Greece	8044
5	[5] Italy	11809
6	[6] Spain	5027
7	[7] Ex-GDR (only as country of origin)	0
10	[10] Austria	1931
11	[11] France	1083
12	[12] Benelux	81
13	[13] Denmark	194
14	[14] Great Britain	630
15	[15] Sweden	193
16	[16] Norway	34
17	[17] Finland	126
...	(162 rows omitted)	102489
180	[180] Bessarabia	0
181	[181] Myanmar	3
182	[182] Fiji	0
183	[183] Niger	12
222	[222] Eastern Europe	2599
333	[333] Other Unspecified Foreign Country	0
444	[444] EU-Member State (unspecif.)	0
999	[999] ethnic minority	2
-1	[-1] No Answer	2512
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

CORIGIN contains information on the country of birth for all persons who have ever been a part of a SOEP household (i.e., the population from PPATH). Respondents who were born in Germany were assigned the code (1) (see GERMBORN). Persons who were not born in Germany were assigned another country of birth than Germany depending on the information given in the wave-specific individual questionnaires ($P, $PAUSL) or the variations of the "biography / life history" questionnaires (in BIOL), and from the additional questionnaire for 16-17-year-olds in use since 2000 (JUGENDL). In addition, information from PBRUTTO or the electronic household protocol for M1 was used (both answered by the household head). The vast majority of the foreign-born population (see GERMBORN) who have ever been a part of a SOEP household (i.e., the population from PPATH) gave consistent information on their country of birth (see CORIGININFO). For around a third of the foreign-born population, no direct or inconsistent information on the person's country of birth was available. For those respondents who were not born in Germany and whose country of birth could not be determined (CORIGININFO value (2) and (3)), additional indicators were used to code their country of origin (CORIGIN). The generation process was conducted in the following order (with descending priority):

1. The respondents' country of birth which occurred most frequently, in other words the mode, was used.
2. Respondents' country of citizenship was used as their country of birth if both were not German. The citizenship variable was constructed on the basis of all information given on first, second, and previous citizenships as well as naturalizations, and includes the

countries of citizenship a respondent reported. Since citizenship information is collected annually for all persons who lived in a SOEP household, it is based on much more detailed information than the "(2) inconsistent information" collected for the country of origin. Respondents whose information on country of origin is "(2) inconsistent" answered on average three questions on their country of origin (from 2 to 5 answers).

3. Mothers' country of birth and citizenship were considered to be the respondents' most probable place of birth if the respondent was born before the mother immigrated to Germany (see also GERMBORN coding). If information on mothers' country of birth, mothers' citizenship and the respondents' citizenship was missing, fathers' country of birth and fathers' citizenship were used to code CORIGIN. In comparison to the CORIGIN coding from the last wave, in the latest version grandparents' country of birth and grandparents' citizenship were additionally used if information on mothers' and fathers' country of birth and citizenship were missing.
4. For the few cases without citizenship, (grand-)parental information and any information on their country of origin (CORIGININFO value (3)), respondents' legal status was used when it indicated that a person moved to Germany from an "Eastern European" country, resulting in the coding of a few cases to "(222) Eastern European" on CORIGIN.

If the country of birth was still missing after this procedure, CORIGIN was coded "(-1) don't know". CORIGIN includes a few more missing values than GERMBORN due to cases in which it was not possible to determine a country of birth other than Germany. To provide the highest level of transparency possible, we include a variable for the quality of information used to create the country of birth variable: CORIGININFO.
More detailed information on CORIGIN can be found in the documentation on the biographical data in the section on PPATH
For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

corigininfo – Corigin: Quality of information

1	[1] consistent information	133825
2	[2] inconsistent information	4046
3	[3] no answer	30501
4	[4] filter germborn	870885
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

CORIGININFO indicates the quality of information given in CORIGIN. In the latest wave, all relevant information available on persons who have ever been a part of a SOEP household (i.e., the population from PPATH) was compiled into a working dataset and compared to code CORIGIN. CORIGININFO indicates whether "(1) consistent", "(2) inconsistent" or "(3) no information" was available on a respondent's country of birth after these comparisons. CORIGININFO is thus an indicator for the quality of information given in CORIGIN. The filtering of CORIGIN via GERMBORN was taken into account by implementing a separate

category, "(4) Filter GERMBORN" on CORIGININFO for the persons who were considered being born in Germany on GERMBORN. More detailed information on this variable can be found in the documentation on the biographical data in the section on PPATH.
For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

immiyear – Year Moved to Germany

1950		318
1951		343
1952		136
1953		183
1954		249
1955		132
1956		336
1957		436
1958		579
1959		571
1960		967
1961		1156
1962		1412
1963		1301
1964		2050
...	(40 rows omitted)	103731
2005		1275
2006		1132
2007		1165
2008		1188
2009		1500
2010		1528
2011		1854
2012		2006
2013		3368
2014		5184
2015		13510
2016		2705
2017		299
-1		17758
-2		870885

IMMIYEAR contains information on the year of immigration to Germany for all persons who have ever been a part of a SOEP household (i.e., the population from PPATH) and who were not born in Germany (see GERMBORN). All relevant information on this variable was collected from the wave-specific individual questionnaires ($P or $PAUSL) or the variations of the "biography / life history" questionnaires (in file BIOL), and from the additional questionnaire for 16-17-year-olds in use since 2000 (JUGENDL). Since sample M, information on all of a respondent's stays in Germany has been collected (up to 15 moves between countries, see MIGSPELL and REFUGSPELL in this SOEP Survey Paper Series). For all cases in which a respondent had more than one stay in Germany, IMMIYEAR contains the respondent's last year of immigration to Germany. In addition, information from the electronic household protocol for M1 was used, which was only answered by the household head.

The vast majority of the persons who have ever been a part of a SOEP household (i.e., the population from PPATH) gave consistent information on their year of immigration (see IMMIYEARINFO). Over the course of the SOEP survey, only very few cases gave inconsistent information with regard to their year of immigration (see IMMIYEARINFO). For these cases, their latest year of immigration was used in IMMIYEAR. The respondent's year of birth was used as their year of immigration if they mentioned a year of immigration that was before their year of birth (only a few cases).
For those respondents who were not born in Germany and whose year of immigration was not available (IMMIYEARINFO value (3)), additional indicators were used to minimize the portion of missing values. These indicators were used in the following order (with descending priority):

1. When a respondent entered the SOEP for the first time because they had just moved into the household from abroad (see $PZUG from $PBRUTTO), the household entry year was considered to be the same as the immigration year.

2. Mother's year of immigration was used as a proxy for the respondent when the respondent was born before the mother immigrated to Germany. If a mother's year of immigration was missing, the father's year of immigration was used to code IMMIYEAR. If a mother's and father's year of immigration were missing, the maternal and paternal grandparents' year of immigration were used respectively.

If the year of immigration was still missing after this procedure, IMMIYEAR was coded "(-1) don't know". IMMIYEAR includes more missing values than GERMBORN and CORIGIN due to cases in which it was not possible to determine a respondent's year of immigration. However, users should be aware that the wording of questions on the year of immigration vary rather drastically over the course of the SOEP survey (see documentation on the biographical data in the section PPATH). To provide the highest level of transparency possible, we include a variable for the quality of information used to create the year of immigration variable: IMMIYEARINFO.
More detailed information on this variable can be found in the documentation on the biographical data in the section on PPATH.
For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

immiyearinfo – Immiyear: Quality of information

1	[1] consistent information	133478
2	[2] inconsistent information	543
3	[3] no answer	34351
4	[4] filter germborn	870885
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

IMMIYEARINFO indicates the quality of information given in IMMIYEAR. In the latest wave, all relevant information available on persons who have ever been a part of a SOEP

household (i.e., the population from PPATH) was compiled into a working dataset and compared to code IMMIYEAR. IMMIYEARINFO indicates whether "(1) consistent", "(2) inconsistent" or "(3) no information" was available on a respondent's country of birth after these comparisons. IMMIYEARINFO is thus an indicator for the quality of information given in IMMIYEAR. The filtering of IMMIYEAR via GERMBORN was taken into account by implementing a separate category "(4) Filter GERMBORN" on IMMIYEARINFO for individuals who were considered to have been born in Germany on GERMBORN (for more information, see GERMBORN). When information in this working dataset consistently indicated a specific year of immigration, IMMIYEARINFO was coded "(1) consistent information" and the respective year of immigration was stated in IMMIYEAR.
For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

migback – Migration background

1	[1] no migration background	755243
2	[2] direct migration background	168372
3	[3] indirect migration background	115642
4	[4] migration background, not differentiable	0
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

MIGBACK contains information on respondents' migration background for all persons who have ever been a part of a SOEP household (i.e., the population from PPATH). In comparison to GERMBORN, the variable MIGBACK is useful to identify immigrants' descendants by combining information on respondents' country of birth (see GERMBORN) and (grand-)parental information such as their country of birth and their citizenship. The information for this variable comes predominantly from PPATH (GERMBORN), auxiliary citizenship variables and the relevant biographical data sets (BIOIMMIG). The variables were also updated using information from the wave-specific individual questionnaires ($P, $PAUSL), the variations of the "biography / life history" questionnaires (in file BIOL), and the additional questionnaire for 16-17-year-olds in use since 2000 (JUGENDL).
Respondents were assigned to the MIGBACK categories based on country of birth (see GERMBORN): Being born in another country than Germany indicates, by definition, a direct migration background (2), while respondents born in Germany may have either no (1) or an indirect (3) migration background. Respondents whose parents had no migration background were assigned the code "(1) no migration background", while respondents whose father or mother had a migration background were assigned the code "(3) indirect migration background".
In comparison to the MIGBACK coding from last wave, in the latest version grandparental information were considered when an indirect migration background and parental information were missing. Please note that any updates in related variables may also lead to an update of the MIGBACK variable. For instance, a respondent who never stated his or her citizenship but later states having German citizenship will be classified as having a migration background of some form. This retrospective perspective may lead to updates of the migration background variable with every new wave. To provide the highest level of transparency possible, we include a variable for quality of information used to create the migration back-

ground variable: MIGINFO.
More detailed information on this variable can be found in the documentation on the biographical data in the section on PPATH.
For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

miginfo – Migback: Quality of information

1	[1] direct personal w/o parental info	341156
2	[2] proxy personal w/o parental info	698101
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

MIGINFO indicates the quality of information given in MIGBACK. MIGINFO provides information about the usage of (grand-)parents' migration histories in the SOEP. Overall, MIGINFO can take on two different codes: "(1) no parental information" or "(2) (grand-)parental information available". The (grand-)parental information refers to any information on the migration background of the respondents' mother, father or grandparents. This includes information on the country of birth and auxiliary citizenship variables.
Please note that the MIGINFO coding from 2015 (v32) is further differentiated between the availability of direct and proxy information on respondents. We changed the MIGINFO coding due to the introduction of the GERMBORNINFO variable in 2016 (v33). The quality of information given in MIGBACK can thus only be assessed by combining the GERMBORNINFO and MIGINFO variables. MIGBACK information is considered to be highly reliable in cases coded (2) "Parental information available" on MIGINFO and (1) "Consistent information" on GERMBORNINFO (around half of the PPATH cases). In contrast, the quality of information given on MIGBACK is considered relatively uncertain in cases where parental information ((1) "No parental information" on MIGINFO) and respondents' information was missing ((3) "No information" on GERMBORNINFO)).
In a few cases, "(1) no parental information" (see MIGINFO) was available but we were nonetheless able to identify respondents with an "(2) indirect migration background" (see MIGBACK). In these cases, respondents were born in Germany but further variables suggested that there was a migration background (e.g., ethnic Germans). MIGBACK may slightly underestimate the number of persons having an "(3) indirect migration background", since some of the respondents born in Germany with missing (grand-)parental information and for whom no further indicators were available may be the descendants of immigrants.
More detailed information on this variable can be found in the documentation on the biographical data in the section on PPATH.
For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

arefback – Refugee Experience

1	[1] without evidence of refugee experience	984251
2	[2] with evidence of direct refugee experience	34522
3	[3] with evidence of indirect refugee experience	8312

-1	[-1] No Answer	12172
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

The variable indicates asylum seekers and refugees – like for MIGBACK it is differentiated according to direct and indirect (later born children) background. (more detailed information on generation and usage can be found in Krause/Glass 2019).
For more information, contact: Peter Krause (Tel. 030-89789-690)

arefinfo – arefback: Source of Information

0	[0] without evidence of refugee experience	984251
1	[1] residence permit status (current)[current year]	8354
2	[2] residence permit status (current)[past years]	3966
3	[3] residence permit status (bioimmig)	7215
4	[4] Refugees Samples [M.] target person	1752
5	[5] Refugees Samples [M.] direct refugee experience	12082
6	[6] Partner information	432
7	[7] children[MUM], direct refugee experience	629
8	[8] children[P-MUM], direct refugee experience	9
9	[9] children[HV], direct refugee experience	4
10	[10] children[geby<=immy+5] indirect refugee experience	1986
11	[11] children[geby<=immy+10] indirect refugee experience	1458
12	[12] children[geby<=immy+10] indirect refugee experience	2547
13	[13] Refugees Samples [M.] indirect refugee experience	2321
14	[14] HH Head info [household entrance year]	28
15	[15] GER with direct refugee experience [biimgrp]	51
-1	[-1] No Answer	12172
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

The variable indicates further differentiations for asylum seekers and refugees [15 categories] – (more detailed information on generation and usage can be found in Krause/Glass 2019).
For more information, contact: Peter Krause (Tel. 030-89789-690)

loc1989 – Where did you live in 1989?

1	[1] East Germany (DDR) incl. East Berlin	175443
2	[2] West Germany (FRG) incl. West Berlin	535458
3	[3] Abroad (Ausland)	53858
-1	[-1] No Answer	50048
-2	[-2] Does not apply	224450

-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

The variable LOC1989 provides information about a person's residence prior to German reunification, distinguishing among "(1) German Democratic Republic [GDR]", "(2) Federal Republic of Germany [FRG] (including West Berlin)", and "(3) abroad". Respondents born after 1989 (GEBJAHR in PPATH) were coded as "(-2) does not apply" on LOC1989. This information has been generated for all individuals who were ever a member of a SOEP household (the population of PPATH). LOC1989 combines information from two main sources: In 2003, the individual questionnaire included information on the place of residence before German reunification (TP). Since 2004, this question has been included in the biography questionnaires (c.p. file BIOL). Along with these sources, the following indicators were used to code the variable LOC1989 (with descending priority):

1. HID in PPATHL: Place of residence in the former FRG before German reunification
2. IMMIYEAR in PPATH: Respondents who first immigrated to Germany after 1989 were coded as living "(3) abroad" in 1989
3. IMMIYEAR, CORIGIN in PPATH: Respondents who immigrated to Germany before 1990 were assumed to have been living in the "(2) Federal Republic of Germany [FRG] (including West Berlin)" in 1989
4. PSAMPLE in PPATH: Respondent's sample affiliation in 1990, differentiating between members of the former West samples (A, B) and the former East sample (C)
5. SAMPREG in PPATHL & BRMOVEIN and SYEAR in BIORESID: Respondents who moved into their current dwelling in the former FRG or GDR before 1989
6. SAMPREG in PPATHL: Respondent living in the West or East sample region in 1990

The vast majority of information given in LOC1989 is based on information from these sources. For the remaining respondents, indirect information is derived from the following proxies to code their place of residence in 1989:

1. PZUG in PBRUTTO: New entrants to the SOEP who previously lived in East Germany or abroad
2. BSSCHEND and BSSCHWO in BIOSOC: Place and year of the last school attended
3. PGRUPPE in PBRUTTO: Place of birth that was asked in 1995
4. PL: Country of origin GDR
5. PNAT_V2 in PBRUTTO: Citizens of (former) GDR
6. PL: Place of residence in 1984
7. BIOPAREN and PPATH: Parental residence in 1989 for individuals younger than 18 in 1989

For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

locinfo – Loc1989: Source / Quality of information

0	[0] Respondent born after 1989	224450
1	[1] Direct information	749643
2	[2] Indirect information	15116
-1	[-1] No Answer	50048

-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

The variable LOCINFO indicates the quality of information given in LOC1989, differentiating between direct and indirect information. LOCINFO provides information about the use of proxy information in the process of generating LOC1989 due to missing values in respondents' and their parents' residence in 1989 in the SOEP. Overall, LOCINFO can take on three different codes: either "(1) direct" or "(2) indirect information" is available on respondents or they were "(0) born after 1989".
For more information, contact: Diana Schacht (Tel. +49 30-89789-465)

sampreg – Current sample region (Berlin, West-East)

1	[1] West-Germany	828207
2	[2] East-Germany	208251
-1	[-1] No Answer	0
-2	[-2] Does not apply	2799
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

Place of residence in East- or West-Germany with regard to borders of 1990 in corresponding year (SYEAR). (For Berlin East-West-assignements are approximated by zip-codes)

pop – Sample Membership

1	[1] Private HH, German HH-Head	770923
2	[2] Private HH, Foreign HH-Head	137115
3	[3] Institutional. HH, Collective accommodation, German HH-Head	3145
4	[4] Institutional. HH, Collective accommodation, Foreign HH-Head	6685
5	[5] Not Compl. Private HH, German HH-Head	89230
6	[6] Not Compl. Private HH, Foreign HH-Head	22223
7	[7] Not Compl. Institutional. HH, Collective accommodation, German HH-Head	642
8	[8] Not Compl. Institutional. HH, Collective accommodation, Foreign HH-Head	213
-1	[-1] No Answer	0
-2	[-2] Does not apply	9081
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

POP was derived from WUM2 in HBRUTTO as well as PNAT_V* and STELL_V* (nationality and relationship to head of household in PBRUTTO). Missing values were imputed

based on the person's history. Thus, the only admissible missing value is –2, meaning not applicable. This variable is therefore particularly important, as it enters into the determination of cross-sectional weights. The variable corresponds with HPOP in HPATHL. See also the description of NETTO.

sexor – Sexual Orientation

0	[0] probably heterosexual	670030
1	[1] probably bi/homosexual	6346
2	[2] insufficient information	362881
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

The variable SEXOR combines information on the sexual orientation of respondents from various sources in the SOEP. In 2016 (wave BG), (1) a direct question about sexual orientation was introduced (self-rep). Questions on marital status in the SOEP distinguish between same-sex civil unions and different-sex marriages. This distinction has been introduced in the household questionnaire since waves 2002 (wave S), in the person questionnaire since 2011 (wave BB), and in the biographical questionnaire since 2012 (wave BC). Starting with these years respectively, we use information of (2) the head of household on marital status of all household members (civil-hh), information on the marital status (3) reported by individuals in the person questionnaire (civil-p), as well as reported (4) in the partnership biography (civil-bio). Finally, the SOEP team provides pointers to the partner of each person in the SOEP households since 1984 (see pgpartnr in pgen documentation or parid in ppathl documentation). Combining information on the gender of both partners cohabitating in the SOEP household provides (5) the final source of information on the sexual orientation of adults in the SOEP (pointer).
Self-reports on sexual orientation surveyed in 2016 distinguish between the response options heterosexual, bisexual, and homosexual. It is however impossible to clearly identify bisexual respondents from data on same-sex and different-sex partnerships even in longitudinal studies like the SOEP. This is because some bisexual respondents may be observed at periods of no-cohabitation, only same-sex, and only different-sex partnerships. Without any observed change in the partner's gender, we are unable to identify respondents as bisexual. Our approach to this problem is as follows: first, we do not seek to distinguish between homo- and bisexuals in the generated SEXOR variable. That is, we code individuals with (at least) one observation of a same-sex partnership as homo/bisexual. We code individuals with information from at least two years (arbitrary threshold) on only different-sex relationships as heterosexual. Since bisexuals in stable/multiple different-sex partnerships are misclassified as heterosexuals instead of homo/bisexuals, we add the label "probably" to our generated variable to indicate that this information is potentially erroneous. In the case of no information on partnerships or only one year of information on different-sex partnerships we consider this insufficient to make any inferences on sexual orientation in these individuals on the basis of their observed partnerships.
Finally, the sexor variable integrates both the self-reported as well as the partnership-obtained information on sexual orientation.

sexorinfo – Sexual Orientation:Source of information

0	[0] insufficient information	362881
1	[1] pointer	144834
2	[2] civil-self	1080
3	[3] pointer, civil-self	651
4	[4] civil-hh	67
5	[5] pointer, civil-hh	164626
6	[6] civil-self, civil-hh	50
7	[7] pointer, civil-self, civil-hh	97832
8	[8] BIO	593
9	[9] pointer, bio	120
10	[10] civil-self, bio	93
11	[11] pointer, civil-self, bio	23
12	[12] civil-hh, bio	0
13	[13] pointer, civil-hh, bio	670
14	[14] civil-self, civil-hh, bio	0
...	(9 rows omitted)	259001
24	[24] bio, self-rep	1130
25	[25] pointer, bio, self-rep	247
26	[26] civil-self, bio, self-rep	174
27	[27] pointer, civil-self, bio, self-rep	51
28	[28] civil-hh, bio, self-rep	0
29	[29] pointer, civil-hh, bio, self-rep	12
30	[30] civil-self, civil-hh, bio, self-rep	0
31	[31] pointer, civ-self, civ-hh, bio, self-r.	5122
-1	[-1] No Answer	0
-2	[-2] Does not apply	0
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

This integer variable indicates which sources of information coincide with the value of SEXOR for the respective respondent. Its digits in binary representation are to be interpreted as binary flags, according to the following scheme: 1=Pointer, 2=Marital status, 4=Relation to head of household, 8=Biography, 16=Self-reported. If SEXORINFO has the value 5=1x1+0x2+1x4, this means that partnership pointers and relationship to head of household variables indicate the sexual orientation which is coded in SEXOR. Similarly, a value of 16 indicates that the inference was drawn from the direct question about sexual orientation. The variable is labeled accordingly.

parid – Partner Person Number

Partner indicators have the purpose of defining couples in SOEP households and thus to make possible analyses on the dyadic level. Persons without spouse and (cohabitating) partner receive a missing code "-2" (=does not apply). Also, the variable PARTNER is coded 0, 3, 4, 5 in these cases. In couples, partner is the value of the unchanging person ID number (=PID) of the partner. The assignment of the partner ID within households is based on four sources of information: A question in the person-file, that asks (unmarried) respondents to

identify their partner in the household (bhppnr in 2017) (plk0001 in pl), the household matrix reported by the head of household at the beginning of the interview (bhstell in 2017) (stell_v1 stell_v2 stell_h in pbrutto), the partnership biography in the lifehistory calendar reported by new respondents (see also, biomars), and self-reports on marital status and life events, such as marriage, move in with partner, separation, etc. In unclear cases, due to temporal non-response for instance, we also consider longitudinal information from previous and prospective waves. Moreover, PARID is self-consistent between two individuals. For analyses of partner relationships, this information can be used to link all persons with their respective partners, and all information on both partners can also be stored in a common dataset.

partner – Status Of Partnership

0	[0] No partner	500678
1	[1] Spouse, registered partner	459182
2	[2] Partner	69946
3	[3] Probably spouse, registered partner	907
4	[4] Probably partner	1548
5	[5] not clear	4307
-1	[-1] No Answer	0
-2	[-2] Does not apply	2689
-3	[-3] Answer improbable	0
-4	[-4] Inadmissible multiple response	0
-5	[-5] Not included in this version of the questionnaire	0
-6	[-6] Version of questionnaire with modified filtering	0
-8	[-8] Question this year not part of Survey program	0

The variable PARTNER generated in the context of the partner identifier (PARID) to describe whether a person in a SOEP household has a partner in that household, and if so, the type of relationship existing between the partners. Relationships with persons outside the SOEP household are not covered by this variable. Code 0 is assigned to all single persons living in households and those with partners outside the household. Codes 1 to 4 describe relationships. To assign Codes 1 and 2, the partnership has to be definable from the perspective of both partners unanimously. If conflicting information exists between partners, the codes 3 or 4 are assigned. If it is unclear whether an individual has no partner or whether she forms a couple with one other household member, we assign the code 5. Registered partnerships (civil unions) for same-sex couples were introduced in Germany in 2001. Though, registered partnerships are legally not equal to marriage, they are listed in the same category.

5 Weighting

prgroup – Random Groups

1	128616
2	128532
3	131174
4	125335
5	130762
6	133384
7	130980
8	130474

Random Groups (Total SOEP population is randomly devided in eight groups)

pbleib – Inverse Staying Probability

inverse probability weights

phrf – Weighting factor

standard individual weights

phrf0 – Weighting factor for old samples (wave 1 of new sample)

individual weights for first wave of new samples

phrf1 – Weighting factor for new sample (wave 1)

individual weights without first wave of new samples

phrfe – Enumerated weighting factor (including non responding HH members)

standard individual enumerated weights (for all household members)

pbleibe – Inverse Staying Probability (including non responding HH members)

inverse probability enumerated weights (for all household members)

phrfe0 – Enumerated weighting factor for old samples (wave 1 of new sample)

individual enumerated weights for first wave of new samples (for all household members)

phrfe1 – Enumerated weighting factor for new sample (wave 1)

individual enumerated weights without first wave of new samples (for all household members)