SOEP The German Socio-Economic Panel study

581

SOEP Survey Papers

Series D – Variable Descriptions and Coding

SOEP — The German Socio-Economic Panel study at DIW Berlin 2018

# SOEP-Core v33.1 – Activity Biography in the Files PBIOSPE and ARTKALEN

Paul Schmelzer, Maik Hamjediers, and SOEP Group

DIW SOEP

Running since 1984, the German Socio-Economic Panel study (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:

Paul Schmelzer, Maik Hamjediers, and SOEP Group. 2018. SOEP-Core v33.1 – Activity Biography in the Files PBIOSPE and ARTKALEN. SOEP Survey Papers 581: Series D. Berlin: DIW/SOEP


ISSN: 2193-5580 (online)
DIW Berlin
German Socio-Economic Panel (SOEP)
Mohrenstr. 58
10117 Berlin
Germany
soeppapers@diw.de

# SOEP-Core v33.1 – Activity Biography in the Files PBIOSPE and ARTKALEN

**Paul Schmelzer, Maik Hamjediers, and SOEP Group**

# Activity Biography in the Files PBIOSPE and ARTKALEN

by Paul Schmelzer and Maik Hamjediers[1]

PBIOSPE and ARTKALEN encompass activities over the life course and distinguish between spells in education, employment (full- and part-time employment as well as minor employment and registered unemployment), retirement, housekeeping, parental leave and others (see Table 2 for the kind of statuses). Please note that these spells do not capture transitions between educational institutions or jobs and employers; the spells only reflect a continuous status for instance in full-time employment regardless of potential job changes. In the yearly Individual Questionnaire the respondents are also asked to report job changes between the previous and current years and this information can be added using the $PGEN data files.[2]

The spell file ARTKALEN is collected from Individual Questionnaire as a calendar-matrix of months of the previous year and respective statuses of 15 categories (for example Question 118 in 2015).[3] The information from each annual Individual Questionnaire is attached to the information of previous surveys in a way that same statuses in consecutive months were treated as one continuous spell. Thus, being for instance unemployed in the December of one year and being still unemployed in the following January is treated as one continuous spell. The generated spell file starts with the year before the entrance into the sample and ends with a respondent's last observation.

The spell file PBIOSPE is based on the information on activity status over the life course, which is collected as a matrix from every respondent answering the Biography Questionnaire (for example Question 45 in 2015).[4] The observations start at the age of 15 and end at the current age (up to age 65). This information on activity status covers only the period up to the time the biography is collected. To update the ongoing occupational career in PBIOSPE, information from the yearly Individual Questionnaire is also used by aggregating the recorded spells from ARTKALEN into yearly values.[5] In the following, the method of combining the data is described. There have been **no changes** how the data is generated since the previous version, distributed in 2016.[6] But if you have been working with older versions of the dataset (versions distributed in 2008 and earlier) you should check the section at the end of the chapter, where you will find information on previous changes.

---

[1] (based on earlier work by Rainer Pischner, Henning Lohmann, Marco Giesselmann and Mila Staneva)

[2] To add this information it is necessary to split each spell at the time point of an interview. Afterwards, reports of job changes can be merged at each respective time point.

[3] For persons who were temporarily unavailable for interviewing, it is sometimes possible to fill in the gaps in their occupational status. If these persons fill out the additional questionnaire for temporary drop-outs later on, we can use the information collected there (see files $PLUECKE).

[4] See Chapter 1 for general information on the collection of biography information.

[5] For more information, see Haisken-DeNew, John and Joachim R. Frick (2005): *DTC - Desktop Companion to the German Socio-Economic Panel Study (SOEP)*, Chapter 3.

[6] The only exception is the lack of information on short work hours (spelltyp '2' in ARTKALEN), due to omitting this question in the questionnaire of 2015.

But before we move on to the details, we provide a brief overview of the contents of ARTKALEN and PBIOSPE. Table 1 contains a list of all the variables in the datasets. The variables BEGIN and END indicate the beginning and the end of a spell. These variables are age entries in PBIOSPE and month entries in ARTKALEN starting with January 1983 as the first month. There are also variables in PBIOSPE that refer to calendar years: BEGINY and ENDY. The SPELLNR is a serial identifier of spells of each activity status of a given person. The variable SPELLTYP contains information on the activity status during the spell, e.g., employed full-time or unemployed. Note, for refugees (samples M3/M4) the category "engaged in active military combat, in captivity" was merged with category "Military/Civilian service".

**Table 1: Contents of ARTKALEN and PBIOSPE (variables)**

| Variables | Information in | |
|---|---|---|
| | **ARTKALEN** | **PBIOSPE** |
| HHNR | Original Household Number | Original Household Number |
| PERSNR | Never Changing Person ID | Never Changing Person ID |
| SPELLNR | Serial Number of the Spell per Person | Serial Number of the Spell per Person |
| SPELLTYP | Type of Spell | Type of Spell |
| BEGIN | Month Spell Begins | Age Spell Begins |
| END | Month Spell Ends | Age Spell Ends |
| BEGINY | | Year Spell Begins |
| ENDY | | Year Spell Ends |
| ZENSOR | Censor Variable | Censor Variable |
| SPELLINF | | Spell Construction Information |
| ERHEBJ | | Survey Year Biography Data |
| KALYEAR | | First Observation in Calendar Data |
| BEGINB1 | | Age Spell Begins, 1st Initial Biography Spell |
| ENDB1 | | Age Spell Ends, 1st Initial Biography Spell |
| BEGINK1 | | Age Spell Begins, 1st Initial Calendar Spell |
| ENDK1 | | Age Spell Ends, 1st Initial Calendar Spell |
| BEGINB2 | | Age Spell Begins, 2ndInitial Biography Spell |
| ENDB2 | | Age Spell Ends, 2nd Initial Biography Spell |
| BEGINK2 | | Age Spell Begins, 2nd Initial Calendar Spell |
| ENDK2 | | Age Spell Ends, 2nd Initial Calendar Spell |
| BEGINB3 | | Age Spell Begins, 3rd Initial Biography Spell |
| ENDB3 | | Age Spell Ends, 3rd Initial Biography Spell |
| BEGINK3 | | Age Spell Begins, 3rd Initial Calendar Spell |
| ENDK3 | | Age Spell Ends, 3rd Initial Calendar Spell |
| BEGINB4 | | Age Spell Begins, 4th Initial Biography Spell |
| ENDB4 | | Age Spell Ends, 4th Initial Biography Spell |
| BEGINK4 | | Age Spell Begins, 4th Initial Calendar Spell |
| ENDK4 | | Age Spell Ends, 4th Initial Calendar Spell |

As mentioned above, PBIOSPE combines information collected in the biography questionnaire and the calendar matrix of the individual questionnaire stored in ARTKALEN. The two types of information are merged into PBIOSPE following a number of rules. First of all, it is important to acknowledge that the Biography Questionnaire Matrix as well as the Individual Questionnaire Matrix allow for multiple activity statuses for a given year or month. No concept of main activity is used. A common combination is, for instance, "housewife/-husband" and

"working part-time". There are a number of other plausible combinations, but also combinations that are less plausible. However, a list of valid combinations of activity statuses defined according to legal or similar constructs would need to be based on very strong assumptions. In addition—in particular in case of the yearly matrix in the Biography Questionnaire—activities are reported that took place in a calendar year in consecutive months, which makes it impossible to exclude combinations of activities. Therefore, no data cleaning is performed at this stage. As a consequence, the data may contain information on more than one activity for a given point in time.

This also defines the rules for aggregating the monthly ARTKALEN data into yearly values. Take, for example, a person who was in full-time employment from January to November 2007, and unemployed in December 2007. The exact months are recorded in the dataset ARTKALEN. In the aggregated data, which is merged with the yearly data from the Biography Questionnaire, you find the information that the person worked full-time and was also unemployed in the year 2007. There is a second level of aggregation of ARTKALEN information as the data on type of activity, which is recorded in the variable SPELLTYP is more detailed than in PBIOSPE. The respective information is aggregated as described in Table 2.

**Table 2: Aggregation of ARTKALEN spell information into PBIOSPE**

| | PBIOSPE | ARTKALEN |
|---|---|---|
| 1 | School/University | School, College (1) |
| 2 | Apprenticeship/Training | Vocational Training (4), First Job Training, Apprenticeship (13), Continuing Education, Retraining (14) |
| 3 | Military/Civilian service | Military, Community Service (9) |
| 4 | Full-time employed | Full-Time Employment (1), Short Work Hrs (2) |
| 5 | Part-time employed | Part-Time Employment (3), Second Job (11), Mini-job (up to 400 euros) (15) |
| 6 | Unemployed | Unemployed (5) |
| 7 | House-Husband/Wife | Housewife, Husband (10) |
| 8 | Retired | Retired (6) |
| 9 | Other | Maternity Leave (7), Other (12) |
| 99 | Gap | Information on gaps in ARTKALEN is not used. Gaps are calculated on the basis of the merged dataset. |

**Table 3: Coding of the variable ZENSOR**

| Right: / Left: | not censored | censored missing | censored before gap |
|---|---|---|---|
| not censored | 1 | 2 | 3 |
| censored missing | 4 | 5 | 6 |
| censored after gap | 7 | 8 | 9 |

Note: '(99) Gap' spells are all marked as (-2)

Missing information on the beginning or end of a spell causes what is known as censoring problems. There are two types of missing data. First, data can be missing on periods outside the observation window (before the age of 15 and after the age of 65 in the case of PBIOSPE or before and after the panel participation in ARTKALEN). Second, data can be missing on years within the observation window due to item non-response in particular years or due to temporary drop-outs (the latter applies to calendar information only). In this case, we speak of "gaps." There are nine different patterns (see Table 3).

As stated above, the calendar information is used to update the biography information. However, there is also a certain overlap of the periods covered by the two types of data. This is shown in Table 4. It indicates, for persons included in PBIOSPE, the year in which the biography information was collected (variable ERHEBJ). This year is usually also the last year for which biography information is available.[7] The table also shows the first year recorded in the calendar data (variable KALYEAR). In the majority of cases (58.7 percent), the earliest calendar information is available for the year before the biography interview. This is the case for persons who answered the Biography Questionnaire in their first year as survey respondents. The calendar in the Individual Questionnaire refers to the year before the survey. There are, however, changes over time. In 1998, it was decided that first-time respondents from new samples would not be given the Biography Questionnaire in the first wave but in the second in order to reduce the entry threshold for these new respondents. Consequently, for the majority of persons in years after new samples were integrated (1999, 2001, 2003, 2007, 2010/11 – Samples E to I), the earliest calendar information is available two years before the biography information was collected. This was once again changed in 2011 and in 2013: respondents from samples J and K were given both questionnaires in their first year in SOEP and 2013 the sample M was introduced via a special Biography Questionnaire about individual migration histories without surveying the calendar data. The same applies to first-time respondents who are members of an old sample (e.g., persons who moved into a panel household) - they answer the Biography Questionnaire at the time of their first interview. The pattern is quite stable for most years before 1999. Notable exceptions is the year 1992. This is explained by the integration of East Germany into the SOEP in 1990 (Sample C). The majority of the respondents in these

[7] Please note that some biographies were collected in 2011 although they are part of Wave 27. This results from the fact that some members of Sample I were interviewed in early 2011 instead of 2010.

samples answered just the Biography Questionnaire at the entrance into the SOEP. Another exception is the year 1987. In the years 1985 to 1987, the life course matrix was not part of any of the questionnaires. Therefore the respective biography information was only available for persons who were interviewed in 1984. In 1988, biographic information was also collected for persons who became respondents in 1985, 1986, and 1987 (for all years ERHEBJ=1987).

**Table 4: Overlap between biography and calendar information**

| | First observation in ARTKALEN (compared to erhebj*) | | | | | |
|---|---|---|---|---|---|---|
| | | earlier | | | | |
| erhebj* | same year or later % | 1 year % | 2 years % | 3 years % | 4+ years % | Total n |
| 1984 | 0,1 | 99,9 | 0,0 | 0,0 | 0,0 | 11001 |
| 1987 | 0,0 | 36,4 | 33,5 | 30,1 | 0,0 | 505 |
| 1988 | 0,0 | 100,0 | 0,0 | 0,0 | 0,0 | 164 |
| 1989 | 0,5 | 99,5 | 0,0 | 0,0 | 0,0 | 193 |
| 1990 | 0,0 | 100,0 | 0,0 | 0,0 | 0,0 | 180 |
| 1991 | 0,0 | 100,0 | 0,0 | 0,0 | 0,0 | 157 |
| 1992 | 0,0 | 8,4 | 3,6 | 88,0 | 0,0 | 3930 |
| 1993 | 0,0 | 76,6 | 0,3 | 2,3 | 20,7 | 304 |
| 1994 | 0,2 | 98,3 | 0,3 | 0,2 | 1,0 | 918 |
| 1995 | 0,2 | 99,1 | 0,0 | 0,1 | 0,6 | 1037 |
| 1996 | 0,2 | 97,9 | 0,0 | 0,0 | 1,9 | 480 |
| 1997 | 0,0 | 98,5 | 0,0 | 0,0 | 1,5 | 478 |
| 1998 | 0,7 | 98,1 | 0,0 | 0,2 | 1,0 | 415 |
| 1999 | 0,1 | 26,6 | 72,8 | 0,0 | 0,5 | 1821 |
| 2000 | 0,0 | 90,2 | 0,9 | 7,7 | 1,3 | 235 |
| 2001 | 0,0 | 6,3 | 93,6 | 0,0 | 0,0 | 7529 |
| 2002 | 0,2 | 48,1 | 0,4 | 39,0 | 12,4 | 526 |
| 2003 | 0,1 | 16,9 | 81,3 | 0,1 | 1,6 | 2193 |
| 2004 | 0,0 | 68,8 | 4,2 | 20,1 | 6,9 | 432 |
| 2005 | 0,0 | 89,0 | 3,4 | 0,7 | 6,8 | 292 |
| 2006 | 0,0 | 92,2 | 4,1 | 0,0 | 3,7 | 217 |
| 2007 | 0,0 | 16,1 | 83,4 | 0,1 | 0,3 | 1858 |
| 2008 | 0,0 | 68,9 | 2,9 | 26,9 | 1,3 | 309 |
| 2009 | 0,0 | 89,5 | 2,1 | 0,5 | 7,9 | 190 |
| 2010 | 4,1 | 79,2 | 16,7 | 0,0 | 0,0 | 7887 |
| 2011 | 2,5 | 91,7 | 5,7 | 0,0 | 0,1 | 5670 |
| 2012 | 3,9 | 95,6 | 0,3 | 0,2 | 0,0 | 2205 |
| 2013 | 92,7 | 7,0 | 0,2 | 0,1 | 0,0 | 3876 |
| 2014 | 40,9 | 58,2 | 0,7 | 0,0 | 0,2 | 443 |
| 2015 | 79,9 | 15,9 | 2,4 | 1,5 | 0,2 | 1324 |
| 2016 | 0,0 | 90,0 | 7,0 | 2,2 | 0,9 | 229 |
| Total | 9,5 | 58,7 | 24,2 | 7,1 | 0,5 | 56998 |

Notes: *) Year of biography data collection (variable ERHEBJ). Source: SOEP v33 (PBIOSPE).

In addition, there are even some cases (0.5 percent) where the biography information was collected a long time after the person started to respond to the Individual Questionnaire (up to 32 years). These are respondents who failed to answer to the Biography Questionnaire at a given time and therefore the biography information was collected later. In these—albeit very rare—cases, there is substantial overlap between the periods covered by the calendar and biography information.

**Table 5: Sources of PBIOSPE spells**

| | n | % | % cum. |
|---|---|---|---|
| biography only | 225106 | 52.6 | 52.6 |
| calendar only | 136450 | 31.9 | 84.5 |
| 1 biography, 1 calendar spell | 64655 | 15.1 | 99.6 |
| 2+ biography, 1 calendar spell(s) | 584 | 0.1 | 99.7 |
| 1 biography, 2+ calendar spell(s) | 1068 | 0.2 | 100 |
| 2+ biography, 2+ calendar spell(s) | 35 | 0 | 100 |
| Total | 427898 | 100 | |

Source: SOEP v33 (PBIOSPE).

After merging the information from the Biography Questionnaire and ARTKALEN, the data is transformed into spells, whereby each spell is defined by the duration of a given status. A question that arises when merging the data is how to handle overlapping pieces of information. The basic principle is to assign a value of a given status in a given year if the status is recorded in the calendar or in the biography information or both. An example might help to illustrate this: the calendar records full-time employment for the years 2005 and 2007 while the biography records full-time employment for the period from 2000 up to 2006. The merged data from PBIOSPE contains a spell that begins in 2000 and ends in 2007. However, the initial information is restored by including additional variables, which allows for alternative ways of merging the data (see below). The variables SPELLINF, ERHEBJ, and KALYEAR contain general information on the sources of the information captured in a given spell. Table 5 shows that the majority of spells are based on biography information only (52.6 percent). Slightly less than one-third of all spells (31.9 percent) are not observed in the Biography Questionnaire but only in the calendar data. The remainder of spells contain information from biography as well as calendar data. Usually these spells combine one period observed in the Biography Questionnaire with a period observed in the calendar. Only 0.3 percent of the spells combine more than one period in any of the two sources (SPELLINF=4, 5 or 6).

The variables BEGINB1-ENDYK4 document the initial information from the two different sources and are probably not of interest to the majority of users. However, on the basis of these variables, users are able to fully separate the Biography data from the aggregated ARTKALEN

data. This is advisable if you want to use the more detailed ARTKALEN information and combine it with the yearly information from PBIOSPE for earlier years only. The variable names indicate the "source" of the original information utilized (B: Biography -Questionnaire or K: calendar information from the yearly survey). As an example, we discuss one of the spells that combines information on more than one period from any of the two sources. The spell number 4 of person 9205 starts in 1983 and ends in 1994 (SPELLTYP=4: full-time employment). As the variable SPELLINF (=5) shows, this a spell that combines one period from the biography data with two periods from the calendar data. According to the biography data, the person worked full-time from 1983 (BEGINYB1) until 1992 (ENDYB1). There is overlapping information from the calendar data available from 1986 onwards (KALYEAR). According to these data, the person worked full-time from 1986 (BEGINYK1) to 1990 (ENDYK1) and from 1993 (BEGINYK2) to 1994 (ENDYK2). During the years 1991 and 1992, no full-time employment is recorded in the calendar data, which contradicts the information from the biography data.

**Table 6: Example of combined spell**

| persnr | spellnr | spelltyp | beginy | endy | spellinf | erhebj | kalyear | beginyb1 | endyb1 | beginyk1 | endyk1 | beginyk2 | endyk2 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 9205 | 4 | 4 | 1983 | 1994 | 5 | 1998 | 1986 | 1983 | 1992 | 1986 | 1990 | 1993 | 1994 |

Source: SOEP v31 (PBIOSPE).

In PBIOSPE, no attempt is made to "resolve" such contradictions, as this would require rather strong assumptions. More important, such assumptions would differ according to the research question, which makes it even more difficult to provide a standard solution. Therefore, in such cases, we generate spells in the same manner as in less difficult cases, namely by combining the information from the calendar and the biography data. In the given example, this results in a full-time employment spell that starts in 1983 and ends in 1994. As mentioned above, there are very few spells that combine information on two or more periods (SPELLINF=4, 5, 6, less than 0.4 percent of all spells). There are even fewer such spells where the period of overlap is as long as in this example, where the biography data was collected many years after the persons joined the survey (ERHEBJ=1998, KALYEAR=1986). However, users who are interested in combining biography and calendar data in a different manner can use the variables BEGINB1-ENDYK4 to fully separate the two types of data and to recombine the data on the basis of different rules of aggregation.

***Changes in the previous version of PBIOSPE (release 2009)**:*

The description in this chapter refers to the version of PBIOSPE released in 2014 (waves 1-30). There have been no changes how the data is generated since the previous version, distributed since 2009. But users who are only familiar with older versions of PBIOSPE (releases 2008 and earlier) will observe some differences. In 2009, the data generation has been updated completely, but without changing the basic principles. Therefore, there are only a few barely discernible deviations in the main variables (due to slight changes in the consistency checks of

the data). But there are a number of visible changes in the form of additional variables or additional values in already existing variables:

- documentation of censoring:
    - gaps in the data are recorded as spells (SPELLTYP=99)
    - the variable ZENSOR is more detailed and informs about the type of censoring (end of observation window, gap due to missing data)
- documentation of set-up of single spells:
    - new variable KALYEAR: contains the first year for which calendar information is available
    - new variables BEGINB1-3, ENDB1-3, BEGINYB1-3, ENDYB1-3, BEGINK1-4, ENDK1-4, BEGINYK1-4, ENDYK1-4 (these variables replace BEGINBIO, ENDBIO, BEGINYB, ENDYB, BEGINKAL, ENDKAL, BEGINYK, ENDYK): Like the replaced variables, these variables document the original calendar and biography data. The new variables have been added to have a full documentation also for spells in which three or more initial spells are merged (spells with SPELLINF>=4). For the large majority of spells (SPELLINF<=3) only the first of each set of variables is filled. The new variables can be used to separate biography and calendar data, e.g., if you want to combine on your own biography data with data from ARTKALEN.
    - additional value in variable SPELLINF: the value 6 indicates that a spell has been constructed out of 2 or more biography and 2 or more calendar spells
- additional changes:
    - variable ERHEBJ: value -2 if no biography information for a person is available (old version: value 0)
    - The variable FEHLCODE is no longer provided, as its values appeared to be more confusing than helpful. It contained information on data problems in the biographies collected in 1984 only. Information on gaps and overlaps is now documented for all years but not in a single variable.
    - variable SPELLTYP: value 3 indicates part-time and marginal employment until 2004. In 2005 a separate category for marginal employment (also “mini-job” or “400-euro job”) is introduced (value 15) and value 3 is restricted only to part-time employment.