

The SOEP - A Short Manual

Christian Kagerl*

November 6, 2019

Contents

1	Introduction	1
2	Getting Access	1
3	Getting SOEP Data Ready for Analysis	3
3.1	Data Structure	3
3.2	Generated Variables	4
3.3	Searching for Variables	5
3.4	Drawing, Merging and Cleaning	6
4	Migration Data	8

*University of Bayreuth, Faculty of Law and Economics, Chair of Labour Economics – Prof. Jahn;
<https://www.vwla.uni-bayreuth.de/en/index.html>

1 Introduction

The German Socio-Economic Panel (SOEP), produced and maintained by the German Institute for Economic Research (DIW)¹, is a longitudinal data set of households and persons, spanning three decades and thousands of German respondents. Its panel structure, representative design and the wide array of survey questions make it a valuable tool for empirical research in many fields, but especially in the field of labour economics.

This manual is intended for first-time users of SOEP (e.g. for a bachelor or master thesis), giving an introduction in how to get access (via the chair of Labour Economics at the University of Bayreuth), draw, merge and clean the data (using STATA) for one's research question as well as providing helpful references.

SOEP data are proprietary and can only be accessed via university computers. Every year, the SOEP updates its data with a new wave². The data come in two formats, either already in long format or as wave-specific cross-sections.

2 Getting Access

First of all, if you are interested in using the data, you should register for a thesis at the chair of Labour Economics. After coordinating with your supervisor about the topic and the potential use of SOEP data, you [have to sign the 'Antrag auf Vertragsergänzung' as well as the 'Verpflichtung zur Wahrung des Datenschutzes'](#) which amends the university's contract with the DIW and mandates that you abide by the data protection standards. Hand in *two* copies of this at the office of Sandra Hörath (sandra.hoerath@uni-bayreuth.de, RW II, Room 1.74). The document is subsequently checked by the DIW, which can take up to two weeks, so plan your thesis accordingly.

After your request has been approved, you can access the SOEP data via university computers (preferably, use the PC pools as STATA is available there). Sandra Hörath will unlock your access and give you a username and a password, which allows you to open *one* of the [five activation files on the chair's website](#) (see Figure 1, you will be instructed as to which activation file you can use; and use Firefox!).

¹<https://www.diw.de/en/soep>

²Each (survey) year is assigned a wave. The most recent wave in 2017 was the thirty-fourth (alphabetical identifier 'bh'), the first wave took place in 1984 (identifier 'a').

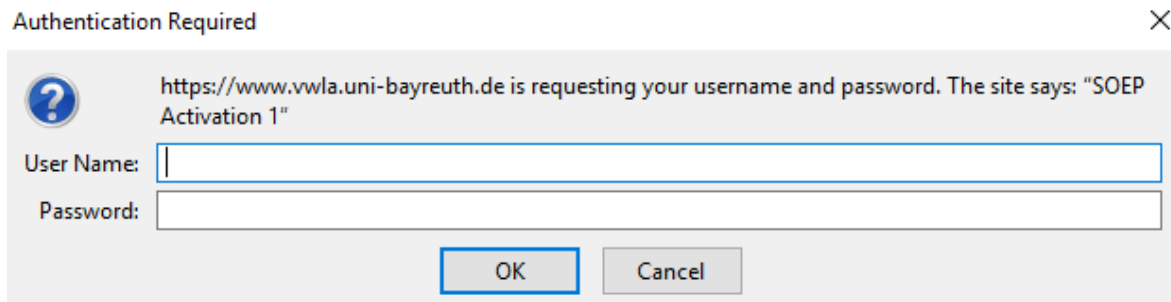


Figure 1: Authentication

After signing in with the given credentials, you are able to download a *.zip*-file that you should extract into your personal drive (the Y:\ drive), so that it is not deleted when you sign out of the university computer. It contains a single *.cmd*-file, with "SOEP_Daten_Zugang" in its name. Executing it opens a black command console, which reminds you to strictly adhere to the data protection standards (that means, for example, that you must lock the computer when you leave the PC lab). Type 'j' and then 'enter' to accept (see Figure 2).

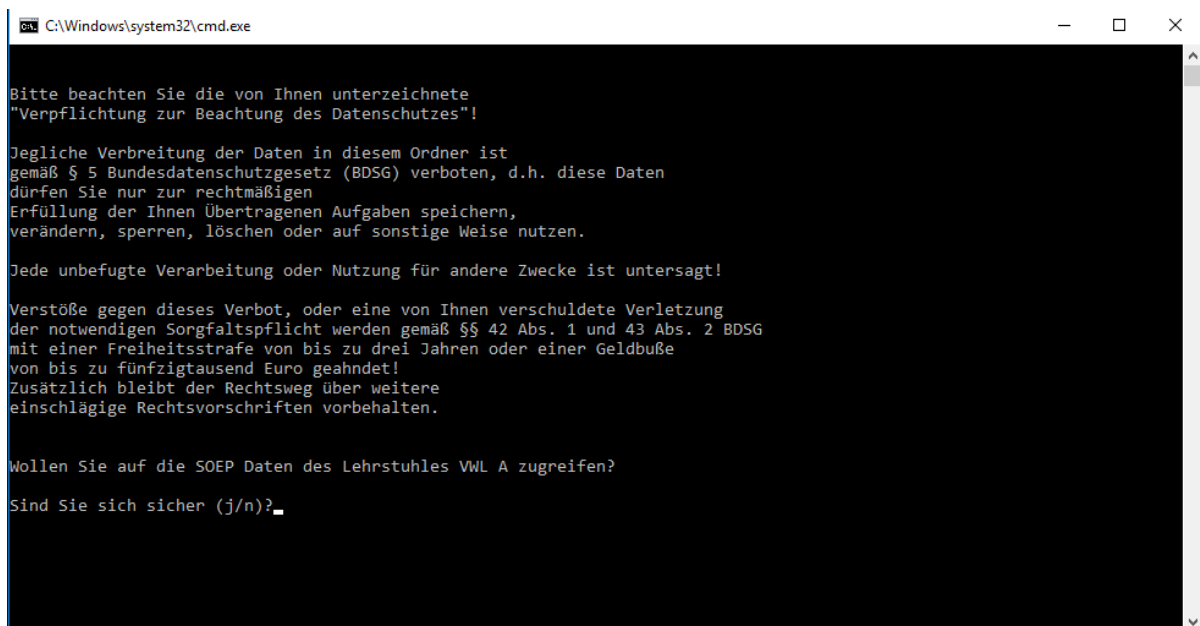
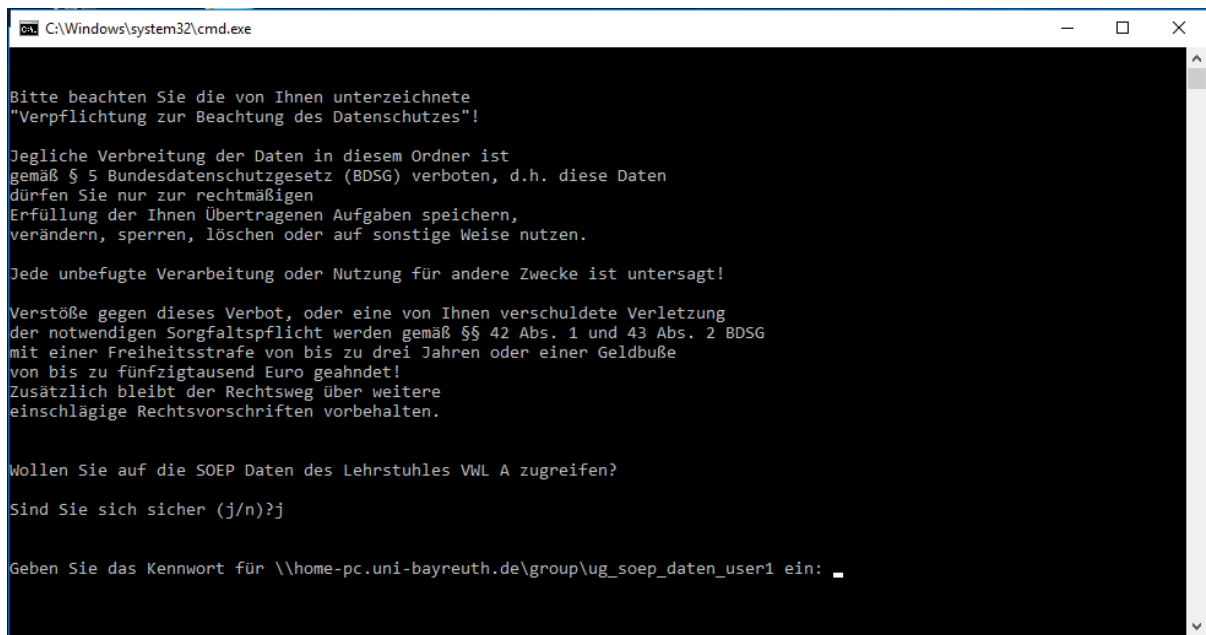


Figure 2: The Command Console 1/2

Thereafter, you're prompted to type your university password (see Figure 3), i.e. the password you use to log into the university's services like e-Learning. Do just that and press 'enter' (note that there is *no* visual confirmation while entering the password). In the end, press any button (like it says in the console).



```
C:\Windows\system32\cmd.exe

Bitte beachten Sie die von Ihnen unterzeichnete
"Verpflichtung zur Beachtung des Datenschutzes"!

Jegliche Verbreitung der Daten in diesem Ordner ist
gemäß § 5 Bundesdatenschutzgesetz (BDSG) verboten, d.h. diese Daten
dürfen Sie nur zur rechtmäßigen
Erfüllung der Ihnen Übertragenen Aufgaben speichern,
verändern, sperren, löschen oder auf sonstige Weise nutzen.

Jede unbefugte Verarbeitung oder Nutzung für andere Zwecke ist untersagt!

Verstöße gegen dieses Verbot, oder eine von Ihnen verschuldete Verletzung
der notwendigen Sorgfaltspflicht werden gemäß §§ 42 Abs. 1 und 43 Abs. 2 BDSG
mit einer Freiheitsstrafe von bis zu drei Jahren oder einer Geldbuße
von bis zu fünfzigtausend Euro geahndet!
Zusätzlich bleibt der Rechtsweg über weitere
einschlägige Rechtsvorschriften vorbehalten.

Wollen Sie auf die SOEP Daten des Lehrstuhles VWL A zugreifen?
Sind Sie sich sicher (j/n)?j

Geben Sie das Kennwort für \\home-pc.uni-bayreuth.de\group\ug_soep_daten_user1 ein: _
```

Figure 3: The Command Console 2/2

Now, the new drive Z:\ is unlocked for you, where the SOEP data are located in a folder. On that Z:\ drive, you should create a separate folder named after you (e.g. Max Mustermann). This folder serves as your repository for the created data sets (do not copy them anywhere else!) and scripts. Do not delete it after submitting your thesis as it is part of the thesis's evaluation. The folder's contents are saved after you sign out of the university computer and remain there for the next session. Crucially, for getting data access each time you do a session in the PC lab, you do not have to download the .cmd-file every time, *but you need* to execute it and sign in with your credentials each time.

3 Getting SOEP Data Ready for Analysis

3.1 Data Structure

If you are familiar with handling the SOEP, you can skip the remainder of this document. There is an additional Stata do-file available that performs an exemplary analysis with respect to a Mincer-Wage-Regression. SOEP data come in two types of formats, a top level directory that contains the data already in long format (i.e. a panel) and a sub-directory that has all the data as yearly cross-sections ('raw'). Both have identical data,

but the first directory is much easier to handle and therefore the focus of what follows. Further, the SOEP data are split into different files according to the type of data therein and according to the level of the observation, either person ('p') or household ('h'). There are tracking files, files that have the gross or original data from the surveys and files that contain generated data (indicated by 'gen' or 'equiv'). In total, there are 56 different data sets, [here is an overview](#) and [here more detailed information](#) on the types of data. In the data sets themselves, missing data points are coded as negative integers that can identify the reason for the missing value (e.g. the question asked is wave/year-specific). The SOEP's website has additional and more detailed information: On the [survey design](#), the [data structure](#), on the [samples](#) and on the [survey methodology](#).

3.2 Generated Variables

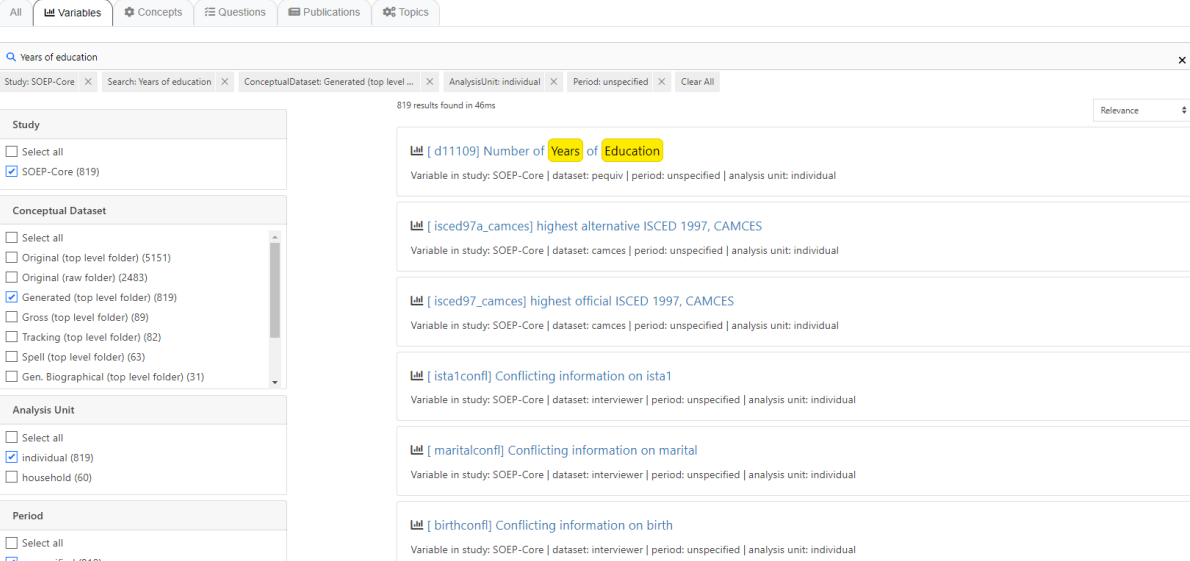
Generated variables within the SOEP are variables that do not necessarily represent any survey question specifically, but are variables that are created by the DIW based on the raw survey data. They [have extra documentation](#) and are very useful as they are already cleaned and, for example, have some missing values imputed by statistical methods (e.g. when an individual reports wage data for 2007, 2008, 2010 and 2011 but not for 2009, then the missing value is imputed). Another conceptual example concerns the education of individuals: Classically, in labor economics, such a variable measures the years of schooling/education in general an individual has attained. However, within questionnaires, such a question is typically not asked, instead the interviewee is asked what kind of degree he or she has accomplished. Generating the years of education from this is possible, but changes to the kind of degrees that exist, changes in the length of education (for instance, the length of the "gymnasium" education being 8 or 9 years in Germany) throughout the years and other issues make it a problematic, potentially fraught and time-consuming process to harmonize and aggregate. Yet, the SOEP generates some variables that are often used to make work easier for the user, those are collected in datasets that are designated as 'generated'. In this case, the SOEP includes the generated variable [d11109](#), which assigns each individual (for every year) the years of education he or she has received, based on the answers of that person throughout the surveys. Hence, the generated datasets represent a valuable source of pre-screened indicators that one can utilize.

3.3 Searching for Variables

After getting access, you need to find and select the variables that are of interest to you and your project. The [SOEP-Website provides helpful links](#) and information on potential variables can be gleaned directly from the freely available [questionnaires](#) whose contents are elaborated on [here](#)! These could be your first place to go to search for variables. Look through them and see which questions could be relevant for your project (working experience, part-time employment, job satisfaction etc.).

Another way to find variables is the site [paneldata.org](#), where you can navigate to the [SOEP-Core](#)-section to find information on the main SOEP data. In addition, it can be easier to register on the site in order to use the basket function for collecting variables of interest. Perhaps the most important thing is to use the search function [here](#). Another good starting point here is the [search for topics](#), where the variables are sorted according to themes (also [explained here](#)). Besides that, you can search for concepts (which SOEP sample), variables (crucial) and survey questions (for the last couple of years). Figure 4 presents an exemplary search query for an education variable.

Search



The screenshot displays the search interface on paneldata.org. At the top, there are navigation tabs for 'All', 'Variables', 'Concepts', 'Questions', 'Publications', and 'Topics'. The search bar contains the query 'Years of education'. Below the search bar, there are filters for 'Study' (SOEP-Core selected), 'Conceptual Dataset' (Generated selected), 'Analysis Unit' (individual selected), and 'Period' (unspecified selected). The search results show 819 results found in 46ms. The first result is '[d11109] Number of Years of Education', which is highlighted in yellow. Other results include 'highest alternative ISCED 1997, CAMCES', 'highest official ISCED 1997, CAMCES', 'Conflicting information on ista1', 'Conflicting information on marital', and 'Conflicting information on birth'.

Figure 4: The Search Function on paneldata.org

On the left hand side, you can filter the results for your search query. The ‘study’ selected here is the main SOEP study, but if, e.g., you are interested in the IAB-SOEP migration sample (see Section 4 for further details), you can select this one with that

option. Next, the filtering category ‘conceptual dataset’ allows for distinguishing the different types of datasets, be it ‘original’, ‘generated’ (see Section 3.2) or ‘tracking’ data. Behind each option, in brackets, it says ‘raw’ or ‘top level’ folder. This nomenclature differentiates the yearly cross-sectional variables (those are in the ‘raw’ folders) from the recommended variables which are already in the long-format, indicated by ‘top level’. If you select top level datasets, the ‘period’ filter no longer displays years as the data are already in a panel format, a fact denoted by an ‘unspecified’ period. Moreover, whether your unit of observation is a household or an individual, you can tailor your results accordingly with the ‘analysis unit’ option. On the right, the results are displayed with the variable name in square brackets and the variable’s data set noted. Clicking on any result leads to a new site where, in the top right, you can add the variable to an existing basket for improved overview or create a basket. Below that, the information on how the variable is named and in which data set it is to be found is listed again, also indicating its concept. For example, via this procedure, you could create a basket of variables that you need for your analysis, giving you a synopsis of what you need to draw. Scrolling down further gives you – for some variables – basic summary statistics of your chosen variable.

3.4 Drawing, Merging and Cleaning

Given that you know what variables you are looking for, you can extract the relevant data with your own code; see the accompanying File `SOEP_example.do` and the comments therein for a basic example performing a simple Mincer-Wage-Regression. In general, whether you are working on the household or individual level, you always need identifying information on your subjects for merging data sets, `pid` in the case of persons!

Some brief notes:

- Your STATA code could look like this when extracting data (the three forward slashes indicate to STATA that the lines form one logical line, but for reasons of tractability it’s easier to split the line up):

```
use var1 var2 var3 var4 ///
    var5 var6 var7 var8 ///
    using "xyz.dta"
```

```
sort identifier syear
save "your-data-file".dta, replace
```

- For merging data sets, use the unique identifier of the household or individual and the survey year (e.g. `merge 1:1 pid syear using "xyz.dta"` for persons)!
- To achieve this in the case of persons, it is advisable to always extract data on the individuals, the [sample](#) each person is part of, whether he or she resides in a [private household or not](#) and what the [current wave status](#) is; all those are part of the data set named `ppath1`.
- Depending on the nature of your analysis, you can clean the data according to your needs, for example restricting the sample to working-age adults by utilizing the `keep` command.
- Recall that missing values are coded as negative integers for all questions/variables, hence you should tell Stata that those are missing values after you have compiled your data set:

```
mvdecode _all, mv(-8/-1)
```

- Using the process of drawing as enunciated here, the panel you obtain will be unbalanced. For most empirical methods, this does not present a problem. If, however, you want to have a balanced panel, you can do this *ex post* with STATA, where x is e.g. the largest number obtained through the `tab` command:

```
xtset identifier syear
xtdescribe
bysort identifier: gen number_years=[_N]
tab number_years
keep if number_years == x
```

- To facilitate an easier interpretation of your results, you can recode variables to dummies. For example, if you have a gender indicator variable named `sex` that takes the value "1" for males and "2" for females; then


```
gen female = (sex==2)
replace female=. if sex==.
```

creates a dummy variable for women. In a similar fashion, you can use these simple commands and others to tailor any variable from the extensive questionnaires to suit your analysis (also see the exemplary *.do*-file).

The contents of your analysis are of course dependent on the topic, but the DIW has some short examples on [longitudinal analysis](#) and on [fixed effects](#). Moreover, there is a special Stata command that can help you, [see here](#).

Alternatively, it's possible to start with the cross-sectional data from the lower directory, an approach that is *not* recommended by the DIW. In that case, the basket function allows you to automatically create a script that draws data. Yet, this leaves you with a wide data set, which you would have to laboriously convert into long format. If you are interested, the process is nicely detailed in the [SOEP Companion](#), which also contains all relevant information on how to [generate scripts](#) as well as the contents, structure and design of the data.

4 Migration Data

Since migration is one of the key issues in politics and because interesting and policy-relevant questions about immigrants abound (e.g. the labour market effects of immigration and integration or how migrants' job outcomes are influenced by their pre-migration performance), this section shortly considers the extraction and handling of the [SOEP migration sample](#) that is done in collaboration with the Institute for Employment Research (IAB). Compared to the core data, the sample with migration-specific questions has a relatively short time frame as it only started in 2013. It contains roughly 5,000 persons from 2,700 households, each household being home to at least one first- or second-generation immigrant (find variables of interest via the [individual questionnaires](#) for the years 2013 through 2016 or via the website as above).

While the migration sample contains migration-specific questions (like whether immigrants feel disadvantaged at government agencies or which language is spoken within the household), the sample is also administered the most part of the regular questionnaires of

the SOEP. Therefore, the migration sample constitutes a subsample of long format SOEP data. The individuals therein can be identified by their sample affiliation (stored in the tracking file under the name `psample`); the relevant samples (their names starting with M) can be gleaned from [here](#). For finding variables within the IAB-SOEP migration sample, you can select the migration sample under the filtering option ‘study’ in the [search function](#) explained above.

Note, however, that this *does not* capture all migrants of the SOEP, only those from the recent subsamples, migrants have been part of the interviewees since the SOEP’s inception. The recent addition only serves to better understand the social and economic aspects of immigrants’ lives in Germany. If you are interested in comparisons or in older (and thus ‘longer’) samples and information, identifying information on those can be extracted from the ‘bioimmig’ dataset, via a variable named `biimgrp`. In addition – as explained [here as well](#) – in the individual tracking file ‘ppath’ you can find more information on the migration background (e.g. variables `corigin`, `migback`, `germborn` and `immiyear`) of individuals.