1133 2022

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

SOEP-CoV: Project and Data Documentation

Simon Kühne, Stefan Liebig, Daniel Graeber, Thomas Rieger

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:

Simon Kühne, Stefan Liebig, Daniel Graeber, Thomas Rieger. 2022. SOEP-CoV: Project and Data Documentation. SOEP Survey Papers 1133 Series C. Berlin: DIW/SOEP

ISSN: 2193-5580 (online)

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

soeppapers@diw.de

SOEP-CoV:
Project and Data Documentation

Simon Kühne
Universität Bielefeld
simon.kuehne@uni-bielefeld.de

Stefan Liebig
Sozio-oekonomisches Panel (SOEP)
sliebig@diw.de

Daniel Graeber
Sozio-oekonomisches Panel (SOEP)
dgraeber@diw.de

Thomas Rieger
Sozio-oekonomisches Panel (SOEP)
trieger@diw.de

June 3, 2022

Abstract: SOEP-CoV - "The Spread of the Coronavirus in Germany: Socio-Economic Factors and Consequences" is a joint research project of SOEP at DIW Berlin and Bielefeld University. The project was launched in April 2020 immediately after the outbreak of the virus in Germany and aimed to establish a survey database for research about the short- and long-term societal impacts of the virus in Germany. In this documentation, we provide an overview of the study design and survey methods and offer details about the SOEP-CoV data, which are openly available for scientific research.

Acknowledgments: This project was funded by the German Federal Ministry of Education and Research (BMBF) as part of its 2020 "call for proposals for research on COVID-19 in the wake of the Sars-CoV-2 outbreak" (funding reference numbers 01KI2087A and 01KI2087B). The authors would like to thank all other members of the SOEP-CoV project team: Theresa Entringer, Jan Goebel, Markus M. Grabka, Florian Griese, Martin Kroh, Hannes Kröger, Johannes Seebauer, Hans-Walter Steinhauer, Carsten Schröder, Jürgen Schupp and Sabine Zinn. We also thank Deborah Bowen for her support. For the content of questionnaires and data analysis, the project team has also worked with colleagues at the Robert Koch Institute (RKI), Charité Berlin, the Max Planck Institute for Human Development (MPIB), the Institute for Interdisciplinary Research on Conflict and Violence (IKG), and the Social Science Research Center Berlin (WZB). Moreover, in an add-on to the SOEP-CoV project, SOEP at DIW Berlin launched a joint project with the Institute for Employment Research (IAB) in Nuremberg to interview refugees to Germany about their experiences and lives during the pandemic.

1 Introduction

The spread of the coronavirus, SARS-CoV-2, poses major challenges at the individual and the societal level. Researchers are studying how individuals and societies are dealing with these challenges, and what health, psychological, social, and economic effects they have experienced. Meaningful answers to these questions can only be provided using a generalizable database. Moreover, many questions can only be studied with longitudinal data and additional contextual information (such as family or household characteristics).

The SOEP-CoV project (see Kühne et al., 2020), launched in 2020, aimed at creating a suitable database based on an ongoing household panel survey in Germany, the Socio-Economic Panel Study (SOEP, see Goebel et al., 2019). The project is a collaboration between SOEP at DIW Berlin (Co-PI Stefan Liebig) and Bielefeld University (Co-PI Simon Kühne). The SOEP-CoV project collects data to study the factors that influence the short, medium, and long-term socio-economic consequences of the coronavirus in Germany. SOEP-CoV focuses on the following topics: a) prevalence of the virus, health behavior, and health inequalities, b) labor market and employment, c) social life, networks, and mobility, d) mental health and well-being, and e) social cohesion.

For SOEP-CoV, a subset of regular SOEP respondents were interviewed by telephone between April and August 2020 and a second time between January and February 2021. SOEP-CoV data are freely available to the scientific community (see Section 3.2 for more information).

2 Survey Design & Methods

2.1 Sampling & Fieldwork

SOEP-CoV uses the Socio-Economic Panel Study (SOEP, see Goebel et al., 2019) as the basis for the survey design. The Socio-Economic Panel (SOEP) is one of the largest and longest-running multidisciplinary panel studies in the world. Currently, about 30,000 people in 20,000 households are surveyed annually. Since the same households participate in the study every year, it is possible to follow respondents across the entire life course. The SOEP is a household sample and at the same time a longitudinal and thus multi-cohort sample. The SOEP is based on random samples representing the population of all private households in Germany. This makes it an excellent tool for drawing general conclusions about the realities of people's lives in Germany – and thus also about the effects of the pandemic. For SOEP-CoV, this results in two important features:

(1) Since the same people participate in the SOEP study every year, there is already comprehensive information available on them from recent years. Researchers can use this information in combination with the SOEP-CoV data to understand how our society has changed since the beginning of the pandemic.

(2) Researchers will also be able to use these data to see how the pandemic affects life in Germany over the coming years and decades as many SOEP-CoV participants and households will remain part of the regular SOEP surveys.

SOEP-CoV consists of three waves:

1. The first wave of SOEP-CoV (excluding samples P, Q, M3-M5), which ran from March 31, 2020, to July 4, 2020. This part of the survey was conducted by the SOEP group and Bielefeld University. It has 6,694 observations.

2. The second wave of SOEP-CoV (same sample as 1. but with some panel attrition), which ran from January 18, 2021, to February 15, 2021. This part of the survey was also conducted by the SOEP group and Bielefeld University. It has 6,038 observations.

3. The third wave of SOEP-CoV (consisting only of refugee samples M3-M5), which ran from July 9, 2020, to August 21, 2020. This part of the survey was conducted by the Institute for Employment Research (IAB) and the SOEP group. It has 1,439 observations.

While the dataset provided by the SOEP Research Data Center includes all three waves, **this article focuses on the first two waves (1. and 2.)**. For more information on the third wave, which targeted refugees (3.), we refer to the documentation provided by IAB.

For the first SOEP-CoV survey (1.), the participating SOEP households were divided into nine subsamples or tranches. The subsamples were structured in such a way that household composition within each one was representative of the population of private households in Germany. Subsamples 1 to 5 were surveyed every two weeks and subsamples 5 to 9 were surveyed weekly to measure how the crisis was affecting private households over time. A total of 12,000 "anchor" respondents in households were asked to participate in the SOEP-CoV study. The (rounded) number of households in each of the nine samples was: 3000, 3000, 2,000, 1000, 600, 600, 600, 600, 600. This meant that more households were interviewed at the beginning of the survey, when Germany had stricter social distancing restrictions in place, than at the end, when restrictions had been loosened. This design reflects the assumption that the effects of the crisis are likely to have been more severe at the beginning and to have decreased over time. For the second SOEP-CoV survey (2.), the individuals who participated in the first wave were surveyed again. This time, all respondents answered the same questionnaire.

Both the first and second survey (1. and 2.) were conducted using computer-assisted telephone interviewing (CATI). The fieldwork was conducted by Kantar Public. Repeated attempts were made to reach respondents who could not initially be reached by calling repeatedly at different times of day.

2.2 Weighting

Survey weighting makes it possible to compensate for distortions in statistics resulting from selective participation behavior (due, for instance, to households having less time available to participate during the pandemic). Weights are available for the first SOEP-CoV wave 2020 (1., variable `phrf20_core`) and for the SOEP-CoV wave targeting refugees (3., variable `phrf20_ref`). For details on the creation of `phrf20_core`, see Siegers et al. (2020). Please note that we renamed *phrf_cati* to phrf20_core.

Researchers can generate weights for the 2021 SOEP-CoV wave (2.) by estimating their own response model based on the transition of wave 1 households to wave 2 households.[1] For instance, a logistic regression (1 = participation in wave 2, 0 = non-participation in wave 2) can be estimated

1—A detailed description of the construction of nonresponse weights can be found in Kroh et al. (2014).

based on wave 1 information available for both respondents and nonrespondents. After that, the inverse of the estimated participation probabilities can be multiplied by wave 1 weights to construct the final wave 2 sample weights.

Finally, in some cases, it may also be acceptable to apply wave 1 weights to the respective wave 2 households. Please note that in this case, however, there is more potential for biased estimates as some types of individuals/households may have been more likely than others to participate again in wave 2.

If researchers are using only some of the SOEP-CoV wave 1 subsamples for analysis, they should generate subsample-specific weights. Siegers et al. (2020, pp. 17–18) describe in detail how to construct these weights.

2.3 Questionnaires

Questionnaire design for the telephone interviews aimed at a large overlap with standard SOEP questions to maximize comparability with pre-COVID survey years.

Throughout the first-wave fieldwork period, a number of smaller changes were made in the composition of questions. This resulted in a total of six different questionnaires for wave 1 in 2020 (please click on the links to be redirected to the questionnaire webpages):

- Wave 1, 2020, Subsample 1 Questionnaire
- Wave 1, 2020, Subsamples 2 and 3 Questionnaire
- Wave 1, 2020, Subsample 4 Questionnaire
- Wave 1, 2020, Subsamples 5 and 6 Questionnaire
- Wave 1, 2020, Subsamples 7 and 8 Questionnaire
- Wave 1, 2020, Subsample 9 Questionnaire

For wave 2 of SOEP-CoV in 2021, we used one version of the questionnaire only:

- Wave 2, 2021 Questionnaire

All questionnaires with mappings between the items and the data will be made available on the webpage of the DIW Berlin.[2]

3 Data

3.1 Files

The data of the SOEP-CoV survey are included in `cov.dta`, `cov_brutto.dta` and `cov_contact.dta`.

- `cov.dta`: This data set consists of the individual-level responses to the SOEP-CoV questionnaires. It is keyed on `pid` (Person ID) and `syear` (Survey Year), and contains the data for all three waves described in section 2.1.

  The data are in long-format. Concepts from the SOEP-Core study, or concepts that are very close to them, are harmonized in the waves conducted by SOEP (1. and 2.) to follow

2—See https://www.diw.de/de/diw_01.c.785835.de/alle_frageboegen_der_soep-studien.html

the long-naming convention. The SOEP-CoV-specific variables in 1. and 2. are named as follows:

pcovXXXY_Z

where

- pcov is the prefix of all SOEP-CoV-specific variables in 1. and 2.
- XXX is a three-digit number indexing the position of the item (battery) in the questionnaire. The first SOEP-CoV-specific item appearing in the questionnaire is therefore called pcov001, the second pcov002 and so forth.
- If several items form a battery, Y is a lowercase letter starting with a. It indexes the items of the battery in consecutive order. For example, pcov010a-pcov010h give answers concerning the usage of various sources of information relating to COVID-19.
- _Z is a suffix providing additional information for one of the following two cases:
  - If the same item was answered for more than one person in a household, _Z is a number starting with 2 numbering the persons for whom an answer was given. For example, the information concerning the COVID-test of person 1 in a household is given in pcov001, for person 2 in pcov001_2 and so forth.
  - If the question for an item was changed slightly across different SOEP-CoV questionnaires, _Z equals _n to indicate the altered item (for example, pcov017_n). Note that this convention also applies to items from SOEP-Core that have been slightly altered in SOEP-CoV questionnaires (for example, prisk_n).

In cases where concepts from the refugee survey (3.) match the concepts from 1. and 2., the SOEP-Core harmonizations and SOEP-CoV naming conventions described above are also used in wave 3.

Since there is no dedicated path-file, the sample indicator and the weights are contained in the variables sample1 and phrf20_core as well as phrf20_ref, respectively. The weights also effectively distinguish between the SOEP-CoV study conducted by the SOEP group and the corresponding study by the IAB. That is, phrf20_core is filled for interviews conducted by the SOEP-group in 2020 (1.) and phrf20_ref for interviews conducted by IAB (3.). For interviews conducted by the SOEP-group in 2021 (2.), no weights are provided. Please see Section 2.2 for a guide on how to create these weights for 2021.

- cov_brutto.dta: This data set contains household-level information on the gross sample of the SOEP-CoV study. It also provides summary information on the fieldwork process. The data set is keyed on hid (Household ID) and syear (Survey Year). The variable tranche indicates the tranche to which the respondent belongs in the first wave of the SOEP-CoV study.

- cov_contact.dta: This data set contains the contact history of the telephone survey between the survey institute and the households. It documents which household (hid) was attempted to be contacted on which date (ContactDate) by what type of phone number (fest_mobil). Since some households could not be contacted on the first attempt, they were called several times (ContactNumber). The result of the respective contact is documented in the variables ResponseStatus and Resp_Label. The data are exclusively process-generated data.

3.2 Access

The data are provided by the SOEP Research Data Center. They will be included in the standard SOEP data release from version 37 on. For more information on data access, please contact soepmail@diw.de.

References

J. Goebel, M. Grabka, S. Liebig, M. Kroh, D. Richter, C. Schröder, and J. Schupp. The German Socio-economic panel (SOEP). *Jahrbücher für Nationalökonomie und Statistik*, 239(2):345–360, 2019. URL https://doi.org/10.1515/jbnst-2018-0022.

M. Kroh, K. Käppner, and S. Kühne. Sampling, Nonresponse, and Weighting in the 2011 and 2012 Refreshment Samples J and K of the Socio-Economic Panel. *SOEP Survey Papers*, 260(Series C), 2014. URL https://www.diw.de/documents/publikationen/73/diw_01.c.570746.de/diw_ssp0260.pdf.

S. Kühne, M. Kroh, S. Liebig, and S. Zinn. The Need for Household Panel Surveys in Times of Crisis: The Case of SOEP-CoV. *Survey Research Methods*, 14(2 SE - Research initiatives), jun 2020. doi: 10.18148/srm/2020.v14i2.7748. URL https://ojs.ub.uni-konstanz.de/srm/article/view/7748.

R. Siegers, H. W. Steinhauer, and S. Zinn. Gewichtung der SOEP-CoV-Studie 2020. *SOEP Survey Papers*, 888(Series C), 2020. URL http://hdl.handle.net/10419/224082.