1496 2025

SOEP Survey Papers
Series C - Data Documentations (Datendokumentationen)

# SOEP-Core – 2022: Sampling, Nonresponse, and Weighting in Sample R

Hans Walter Steinhauer, Rainer Siegers, Felix Süttmann, Reiner Gilberg, Folkert Aust, Martin Kleudgen, Sabine Zinn

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:

Hans Walter Steinhauer, Rainer Siegers, Felix Süttmann, Reiner Gilberg, Folkert Aust, Martin Kleudgen, Sabine Zinn. 2025. SOEP-Core – 2022: Sampling, Nonresponse, and Weighting in Sample R. SOEP Survey Papers 1496. 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-Core – 2022: Sampling, Nonresponse, and Weighting in Sample R

Hans Walter Steinhauer[1], Rainer Siegers[1], Felix Süttmann[1], Reiner Gilberg[2], Folkert Aust[2], Martin Kleudgen[2], and Sabine Zinn[1,3]

[1]German Institute for Economic Research (DIW Berlin)
[2]infas Institute for Applied Social Science
[3]Humboldt-Universität zu Berlin

April 14, 2025

**Abstract**
Sample R is a refreshment sample of the SOEP-Core. We apply a stratified two-stage sampling design to draw a sample of anchor persons from the population register. We oversample former lignite mining regions (brown coal) all over Germany as well as the younger part of the general population aged 18 to 45 years old. The sample adds 6,560 new households to the panel. Potential selectivities in the participating sample are corrected for using a two-step adjustment for contactability and participation. We find little evidence for selectivities and successfully reduce discrepancies between the panel sample of the SOEP and the general population in terms of age distribution and regional coverage.

**Acknowledgments**
We thank the team at infas Institute for Applied Social Science for the close collaboration in setting up the sampling design, drawing the sample, and closely monitor the field work processes.

# 1 Introduction

Panel studies provide valuable insights for social and behavioral research, allowing for the analysis of dynamic patterns and long-term trends for a given population. However, they rely on the participation of individuals who are willing to contribute their time and experiences over an extended period. Maintaining panel participation and ensuring "representativeness" for the general population pose significant challenges. Panel studies need periodic refreshment of the sample in order to provide a large enough sample. Refreshment samples introduce new households into an ongoing panel to compensate for attrition and capture changes in the target population over time. This process plays a vital role in ensuring the continued "representativeness" of the panel and safeguarding the generalizability of the findings.

The Socio-Economic Panel (SOEP), established in 1984, is one of the longest-running panel studies in the field of economic and social sciences. It provides valuable data on various socio-economic aspects of individuals and households over an extended period. The panel is designed to ensure diversity by including participants from different subgroups of the German population. The sample is refreshed periodically to account for attrition and changes in the population.

Panel attrition is a common challenge panel studies often face. It can have significant effects on the reliability of study findings. Attrition refers to the loss of participating households or individuals over time in a panel study, resulting from factors such as participant dropout, non-contactability, or noncooperation. Attrition can have several implications. If panel attrition does not occur randomly, certain factors may be associated with participants dropping out of the study. This is called selectivity and can introduce bias if the attrition is related to the variables being studied. Moreover, attrition reduces the sample size over time, leading to a loss of statistical power. Smaller sample sizes decrease the precision of the estimates and may limit the ability to detect significant effects or associations, especially for (small) subgroups. Addressing attrition is crucial in panel studies to minimize bias and ensure the robustness of the findings. However, the loss in sample size cannot always be internally compensated for, and thus refreshment samples are necessary.

To refresh the general population in the SOEP, we compared distributions for the information covered by the population register. Looking at the resulting differences we find the population aged 18 to 45 is underrepresented after survey year 2021. For this reason, we decided to oversample this age group. Further, our refreshment sample is supported by the Federal Office for Building and Regional Planning (Bundesinstitut für Bau-, Stadt- und Raumforschung; BBSR), which uses SOEP data to evaluate policy measures in former lignite mining regions. These regions are located in the federal states of Lower Saxony, Brandenburg, Saxony, Saxony-Anhalt, Thuringia, and North Rhine-Westphalia; see Figure 1. To enhance statistical power, the BBSR funded additional households in the four regions: Helmstedter Revier, Rheinisches Revier, Mitteldeutsches Revier, and Revier Lausitz. For this reason, we include additional stratification to allocate the households to these regions. We use a stratified two-stage sampling design in order to allow for oversampling the desired regions and population. At the first stage, we stratify the municipalities responsible for the population register according to federal states and lignite mining regions. At the second stage, we stratify the individuals drawn from the register according to their age. Because the stratification according to age allows households to enter the sample via individuals from different strata, we correct the design weights

accordingly. Further, contactability and participation of households in the panel might be selective. We analyze this possible selectivity using a two-step procedure, analyzing contactability first, and given a successful contact, we analyze participation in the panel. Finally, we use raking procedures in order to make sample distributions conform to those of the population and integrate the refreshment sample into the SOEP.

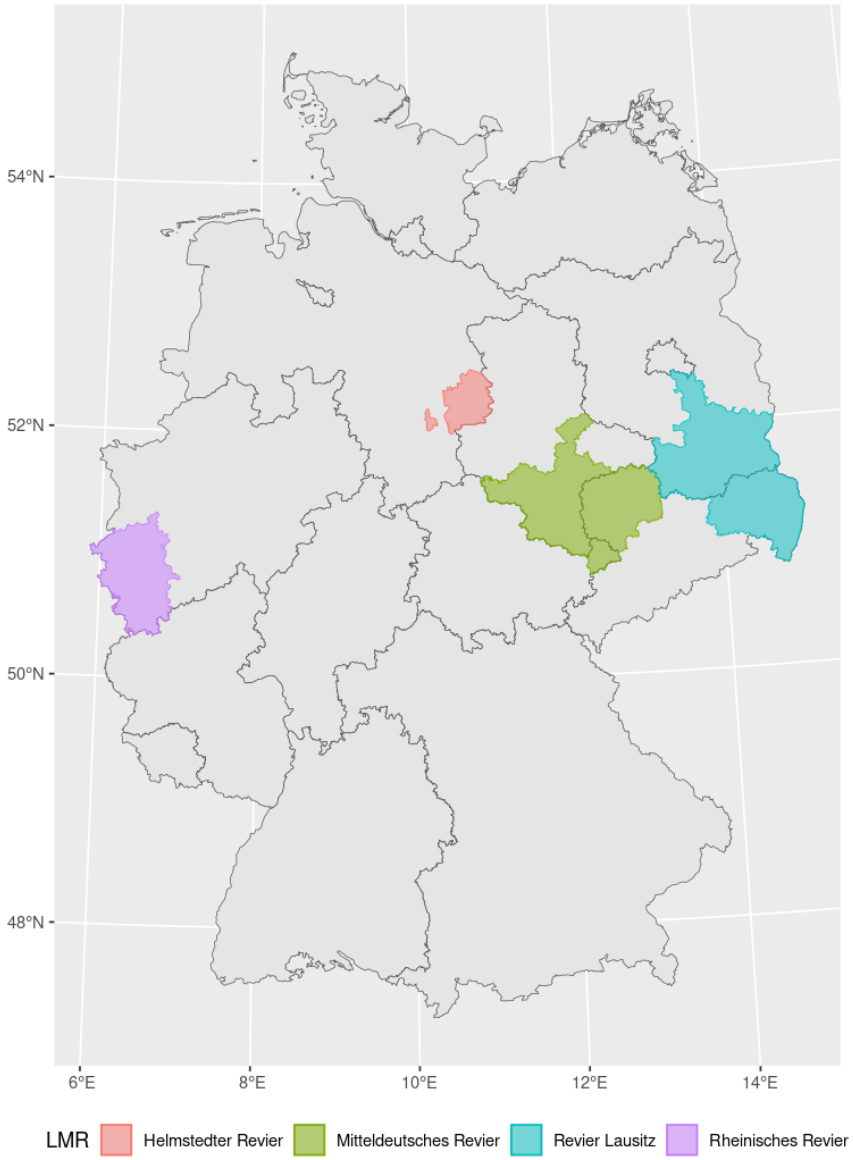


Figure 1: Lignite mining regions (LMR) in Germany.

This paper details the approach applied to refresh the SOEP-Core covering the general population. In Section 2, we provide information on the target population and the sampling frame for the refreshment sample. Details on the sampling design are given in Section 3. The fieldwork process and its results are described in Section 4. Section 5 details the different steps of weighting. Characteristics of the final weights are displayed in Section 6 and Section 7 summarizes.

# 2 Target Population and Sampling Frame

The German population register ("Einwohnermelderegister" or "Melderegister") is the administrative database that records demographic information of residents in Germany. It serves as the primary source of official population data and plays a crucial role in selecting samples from the general population in Germany. The population register is maintained by local registration offices ("Einwohnermeldeämter") in each municipality. It is a decentralized system, with each registration office responsible for keeping accurate and up-to-date records of individuals residing within the respective municipality. Residents are required by law to report any changes, such as address changes or significant life events like births, deaths, marriages, and divorces, to their local registration office. The information contained in the population register typically includes the following details:

- Personal information including full name, date of birth, gender, and nationality;
- The current residential address;
- Information on marital status, such as whether a person is single, married, divorced, or widowed; and
- Details about family relationships, including spouses, children, and other dependents residing at the same address.

Unfortunately, the population register does not contain any information on household compositions. Moreover, registration offices do not necessarily have to provide all the information listed above. Especially information on family relationships that would be useful in constructing household is not provided (see §46 BMG). Thus, sampling households directly from this register cannot be done without further work.

When drawing samples from the population register, the information listed above can be used in order to stratify the sample or focus on specific parts of the population. Further, when using the German population register as a sampling frame, a two-stage sampling design is appropriate. Here, we select the municipalities hosting the population register as primary sampling units (PSU) at the first stage and sample individuals listed in the registers within each municipality as secondary sampling units (SSU) at the second stage. Sampling individuals provides us with an address necessary to contact the household via an "Anchorperson" (anchor person).

The target population for the refreshment sample covers the general population in Germany aged 18 to 70 years living in private households. To sample this population and explicitly address the lignite mining districts we stratify the municipalities at the first stage by federal state and lignite mining districts. At the second stage, we stratify the population by age, using the information on birthdays in the register. The first stratum covers the population aged 18 to 45; and the second stratum covers the population aged 46 to 70 years. This stratification is chosen because we implement an oversampling in the stratum covering the younger part of the population. The stratification according to age at the second stage creates disjoint strata on an individual level. However, we are interested in sampling anchor persons to obtain access to the household. Here, we need to consider the different chances for one household to be sampled via different individuals in the two strata based on age. For this, we account in the weighting procedure described in Section 5.

# 3 Sampling Design

The sampling approach, applied by infas Institute for Applied Social Science, is a stratified two-stage sampling design. The stratification of the municipalities uses the federal states and the lignite mining regions. Lignite mining regions are further divided into three BBSR urban-rural district types. The oversampling of these regions is achieved by increasing the number of sampling points (a fixed number of addresses) in these regions and district types compared to a proportional draw. This factor for oversampling within the lignite mining regions varies among the three BBSR urban-rural district typologies. The measure of size for the probability proportional to size (PPS) sampling is defined as the number of individuals aged 18 to 74 in the population.[1]

The intersection of the four lignite mining districts with the BBSR urban-rural district typology and the federal state theoretically results in 30 strata. However, only 14 of these 30 strata are within the lignite mining districts and are subject to disproportionate sampling, while the remaining 16 strata are outside the lignite mining districts. For example, Hamburg, a federal city state has no lignite mining district. In contrast, Saxony has lignite mining regions, covering rural and urban districts. Sample points outside the lignite mining areas are drawn using a PPS method. Here, collapsing the strata allows for sampling in only one systematic PPS draw. This is done because there is no oversampling in these remaining strata.

In contrast, separate draws are required for each of the 14 explicit strata within the lignite mining areas. This is because the oversampling factors vary. This results in a total of 15 draws across 15 explicit strata, each with a predefined number of sample points. The number of sample points in each stratum is calculated by multiplying the number of sample points from a proportional draw by the respective oversampling factor. This relates to the BBSR district types within the lignite mining areas, with oversampling factors of 7.1 for "rural regions with various challenges", 2.9 for "other rural regions", and 1.8 for "cities". Table 1 displays the number of municipalities and sampling points in the sample as well as individuals in the population for the different federal states and lignite mining regions.

Since the lignite mining areas would also be included in a proportional draw, the total number of sample points in the 14 explicit strata of the lignite mining areas is the sum of the sample points from a proportional draw and the additional points from the oversampling. In total, 80 of the 346 sample points, or 72 of the 300 municipalities to be drawn, are in the lignite mining areas.

Within the 15 explicit strata, micro-stratification is performed as is customary for samples drawn from the population register. This is not strictly necessary but improves the distribution mapping. The implicit stratification criteria used within the 15 explicit strata include district and political municipality size class, categorized into seven levels by population size. In the stratum outside the lignite mining areas, the federal state is also used as an implicit stratification criterion. The same number of addresses is then drawn from the registration office registers for each point.

[1] At the level of the municipalities the range for age groups is not available for the range 18 to 70, but for the range 18 to 74.

Table 1: Sample stratified by federal state and lignite mining districts.

| Federal State | Lignite Mining Districts | Municipalities | Number of Points | Individuals aged 18 - 74 | Individuals aged 18 - 44 |
|---|---|---|---|---|---|
| Schleswig-Holstein | none | 10 | 10 | 575,110 | 282,715 |
| Hamburg | none | 1 | 6 | 1,352,603 | 714,477 |
| Lower Saxony | none | 25 | 25 | 1,000,748 | 498,290 |
| Lower Saxony | Helmstedter Revier | 4 | 4 | 319,908 | 157,017 |
| Bremen | none | 1 | 2 | 412,481 | 203,520 |
| North Rhine-Westphalia | none | 44 | 54 | 5,064,380 | 2,463,392 |
| North Rhine-Westphalia | Rheinisches Revier | 20 | 22 | 1,005,920 | 464,484 |
| Hesse | none | 21 | 22 | 1,206,685 | 645,142 |
| Rhineland-Palatinate | none | 14 | 14 | 426,651 | 226,643 |
| Baden-Württemberg | none | 37 | 39 | 1,954,136 | 1,039,964 |
| Bavaria | none | 40 | 46 | 2,302,377 | 1,225,197 |
| Saarland | none | 3 | 3 | 180,833 | 84,892 |
| Berlin | none | 1 | 13 | 2,684,086 | 1,408,495 |
| Brandenburg | none | 7 | 7 | 206,366 | 96,653 |
| Brandenburg | Revier Lausitz | 11 | 13 | 153,999 | 62,332 |
| Mecklenburg Western Pomerania | none | 6 | 6 | 251,317 | 126,371 |
| Saxony | none | 7 | 8 | 617,220 | 312,654 |
| Saxony | Mitteldeutsches Revier | 9 | 12 | 521,195 | 279,603 |
| Saxony | Revier Lausitz | 9 | 9 | 108,203 | 40,675 |
| Saxony-Anhalt | none | 4 | 4 | 201,805 | 95,166 |
| Saxony-Anhalt | Mitteldeutsches Revier | 17 | 18 | 356,632 | 154,594 |
| Thuringia | none | 7 | 7 | 68,270 | 26,508 |
| Thuringia | Mitteldeutsches Revier | 2 | 2 | 28,427 | 9,804 |
| Total | - | 300 | 346 | 20,999,352 | 10,618,588 |

To oversample younger age cohorts, a sufficient number of personal addresses must be drawn from the registration office registers. Doubling the number of addresses ensures enough gross addresses for the initial disproportionate sample. Additionally, the number of personal addresses per sample point must be based on the minimum expected number of participating households. It is not possible to draw a follow-up sample from registration office samples. Drawing from experiences in previous studies, we assume that a six-fold size of the sample to be realized should suffice. Together with the oversampling of the younger cohort, this yields a twelve-fold sample size.

One sample point will cover a total of 210 addresses. When sampling 346 points, this yields twelve times the required net sample size. The disproportionate initial sample is drawn from the total stock of delivered addresses after infas verifies the correctness of the draw and delivery. The initial sample is divided into tranches to maintain flexibility during the sample realization. Tranche division involves randomly assigning the addresses of the total stock per sample point and stratum (by age cohort groups), ensuring each tranche contains the same number of personal addresses per sample point and age cohort stratum.

To track the number of households in the sample that may include multiple drawn target persons, infas documents the number of individuals with identical addresses in each municipality. No personal address should be removed from the initial sample and all individuals are contacted. Whether individuals at the same address belong to the same household is determined via a specific formulation in the cover letter.

# 4 Fieldwork results and Response Rates

The addresses were validated by infas and deployed to the field. Interviews were conducted by interviewers between May 2022 and February 2023. The households sampled were provided with information sent to them via mail in advance. These letters emphasized that participation is voluntary. Table 2 details the results of the fieldwork on the household-level. In total, there were 6,560 complete or partial interviews, resulting in a response rate on the household-level, calculated according to American Association for Public Opinion Research (2023), of $RR2 = \frac{6{,}560}{22{,}673} = 0.289$. The refusal rate is $REF1 = \frac{10{,}597}{22{,}673} = 0.467$ and, thus, is similar to other samples / studies. Some addresses, although sampled and valid, were not deployed in the field because either interviewers did not contact them before the end of the field period or the numbers of desired interviews was already reached; see AAPOR code 3.11 in table 2. Other addresses are out of sample because the household has moved abroad or was screened out; see AAPOR codes 4.1 to 4.5 in Table 2.

# 5 Cross-Sectional Weighting

The derivation of weights typically involves three main steps, according to Brick and Kalton (1996). First, design weights are computed as the inverse of the inclusion probability (see Section 3). In the second step they are adjusted to account for unit nonresponse, a process referred to as sample weighting adjustment by Kalton and Kasprzyk (1986). In the final step, weights are calibrated to ensure that the estimates align with known population parameters, such as totals, ratios, or specific distributions. This step is referred to

Table 2: Fieldwork results on the household-level according to American Association for Public Opinion Research (2023).

| Final Disposition Code | Number | Percent |
|---|---|---|
| **1. Interview** | | |
| (1.1) Complete | 2,485 | 0.082 |
| (1.2) Partial | 4,078 | 0.135 |
| **2. Eligible, Non-Interview** | | |
| (2.1) Refusal | 10,597 | 0.351 |
| (2.2) Non-contact | 3,339 | 0.111 |
| (2.3) Other | 2,046 | 0.068 |
| (2.4) New address after field period | 128 | 0.004 |
| **3. Unknown eligibility, non-interview** | | |
| (3.11) Not attempted or worked | 5,281 | 0.175 |
| **4. Not Eligible** | | |
| (4.1) Out of sample / screened out | 34 | 0.001 |
| (4.2) Moved abroad | 82 | 0.003 |
| (4.4) Untraceable | 1,925 | 0.064 |
| (4.5) Non-residential building | 202 | 0.007 |
| Total | 30,197 | 1.000 |

*Note*: Totals might not add up due to rounding errors. Calculations based on `hrutt22` data set.

as population weighting adjustment. Please refer to Kroh, Siegers, and Kühne (2015) for a comprehensive understanding of the general weighting strategy employed by the SOEP and the incorporation of new samples.

Please note that the population register lists individuals without providing information about their household context. However, the SOEP is a household panel survey in which all adults are interviewed. As a result, a household with two individuals, for example, has twice the probability of selection compared to a single-person household. To determine a household's sampling probability, we must assign sampling probabilities to all members of the existing households, even though these individuals were not initially sampled as anchor persons. This requires accounting for various characteristics used in stratifying the sample, such as federal states, lignite mining regions, and age. Once the sampling probabilities for each household member are identified, we can calculate the household sampling probabilities by summing these probabilities within each household. The household weights are then derived as the inverse of these household sampling probabilities. It is important to note that the number of households in the target population is unknown and cannot be determined from the sampling frame, as the population register does not include household identifiers. For a formal documentation of this procedure see Kroh, Kühne, Jacobsen, Siegert, and Siegers (2017).

To address potential selectivity resulting from the field work and nonresponse, we employ two models that capture the success of contacting a household as well as the decision-making process of households regarding participation. The models incorporate information on

a) contact: all households that have been deployed to the field and have been approached

b) participation: all households that have been successfully contacted

The first model estimates the probability to successfully contact a household. Given a successful contact, the second model estimates the probability of the household to participate in the panel. Given the limited availability of data on households in the initial sample, we utilize area-level information regarding the residential environment provided by infas360 (see https://datenkatalog.infas360.de/). Further, we utilize design information such as stratification variables and data on processes of the field work including interviewer information and attributes of the first contact attempt.

## 5.1 Sample Weighting Adjustments

In the second step of adjusting the design weights, it is crucial to identify strong predictors of nonresponse. To achieve this, we conduct an iterative process, examining all variables included in the previously described data. We select variables that significantly influence the participation decision using bivariate regression analysis. Next, we remove variables from the set of significant predictors if their absolute correlation with one another is 0.95 or higher. This step prevents the inclusion of highly correlated variables in the analysis. The remaining variables form the basis of a preliminary nonresponse model. To obtain the final model, we apply a variable selection procedure in both forward and backward directions, using the Bayesian Information Criterion (BIC) as the selection criterion. This approach ensures a more parsimonious model by retaining only the most relevant variables. The resulting models estimating the probability to be successfully contacted as well as the response propensities used for deriving weighting adjustments, are presented in Table 3.

Examining the characteristics influencing successful contact, we found that timing in the morning is negatively related while timing in the evening is positively related, compared to a first contact midday. Contact attempts were more often successful when the first contact was established in July. Contacting the households via telephone for the first time positively impacts contactability. Interviewers whose native language is English were less likely to establish contact with a household. When the interviewer was a pensioner or in the age group between 30 and 65 years, they were more likely to contact the HH successfully. In terms of attributes of the residence and its neighborhood, we found that having no garden negatively influences the probability of successful contact. A low share of owners in a block of buildings around a given address lowers the probability of successful contact. The same is true for other types of settlement blocks, compared to purely residential, purely commercial, or mixed areas. Moreover, the probability to establish contact is lower if the dominant building type in the block is commercial, mixed, or high-rise buildings. A very high density of constructions negatively influences the probability for a success of contact. Households located in building blocks with a high share of persons with a religion other than catholic or protestant have a lower probability to be contacted successfully. In contrast, households located in building blocks with a very low share of persons with a religion other than catholic or protestant have a higher probability to be contacted successfully. Looking at attributes of the neighborhood, households located in areas with a high rate of unemployment are less likely to be contacted successfully. Households in neighborhoods with a high share of persons with Soviet or other (other than Turkish, Polish, Italian, Soviet, Yugoslavian, or Muslim) migration

background have a lower probability to be contacted successfully. Households located in a neighborhood with a high value ($9^{th}$ decile) of the index of socio-economic deprivation have a lower probability of being successfully contacted.

Table 3: Models estimating the probability to be successfully contacted as well as the response propensities used for deriving weighting adjustments.

| Variable<br>Value | Contact<br>Estimate<br>(Std. Error) | Participation<br>Estimate<br>(Std. Error) |
|---|---|---|
| Intercept | 0.456*** | −1.066*** |
| | (0.019) | (0.020) |
| **1. Attributes of first contact attempt** | | |
| Timing | −0.418*** | |
| morning | (0.033) | |
| Timing | 0.104*** | |
| evening | (0.017) | |
| Weekday | | 0.225*** |
| Sunday | | (0.067) |
| Month | 0.237*** | |
| July | (0.026) | |
| Month | | 0.578*** |
| December | | (0.040) |
| Mode | 0.137*** | 0.223*** |
| CATI | (0.023) | (0.033) |
| Mode | | 1.527*** |
| CAWI | | (0.081) |
| **2. Attributes of the interviewer** | | |
| Native language | −0.145*** | |
| English | (0.037) | |
| Occupation | 0.096*** | |
| pensioner | (0.019) | |
| Age group | 0.077*** | |
| 30-65 | (0.018) | |
| **3. Attributes of the residence (address)** | | |
| Size of garden | −0.239** | |
| no garden | (0.073) | |
| Quality of area | | 0.155*** |
| still good | | (0.034) |
| **4. Attributes of the residential area (block of buildings)** | | |
| Share of owners | −0.107*** | |
| low | (0.024) | |
| Type of settlement block | −0.381*** | |

Table 3 continued.

| Variable<br>Value | Contact<br>Estimate<br>(Std. Error) | Participation<br>Estimate<br>(Std. Error) |
|---|---|---|
| other type of settlement | (0.084) | |
| Dominant building type | $-0.221^{***}$ | |
| mostly commercial buildings | (0.043) | |
| Dominant building type | $-0.171^{***}$ | |
| mostly high rise buildings | (0.049) | |
| Dominant building type | $-0.106^{***}$ | |
| mixed buildings | (0.031) | |
| Share of inhabitants 30-45 | | $0.156^{***}$ |
| high | | (0.029) |
| Density of constructions | | $0.091^{***}$ |
| high | | (0.027) |
| Density of constructions | $-0.146^{***}$ | |
| very high | (0.024) | |
| Share of person with other religion | $-0.135^{***}$ | |
| high | (0.021) | |
| Share of person with other religion | $0.091^{***}$ | |
| very low | (0.023) | |
| **5. Attributes of the residential area (neighborhood)** | | |
| Unemployment rate | $-0.120^{***}$ | |
| high | (0.021) | |
| Share of persons with other mig. back. | $-0.098^{***}$ | |
| high | (0.020) | |
| Share of persons with Soviet mig. back. | $-0.093^{***}$ | |
| high | (0.019) | |
| Index of Socio-economic Deprivation | $0.268^{***}$ | |
| high ($9^{th}$ decile) | (0.038) | |
| Index of Socio-economic Deprivation | | $-0.195^{***}$ |
| high ($10^{th}$ decile) | | (0.049) |
| N | 24,761 | 19,167 |

Notes: Abbreviation is mig. back. for migration background. Dependent variable: Success in contacting the household (1 = yes, 0 = no), participation of the household (1 = yes, 0 = no). Significance indicated by $*** \equiv p < 0.001$, $** \equiv p < 0.01$, and $* \equiv p < 0.05$. The model is estimated using the function `glm()` with a cloglog link function in R (R Core Team, 2023).

In terms of participating in the survey, we found that participation is more likely when the households was initially contacted on a Sunday. Moreover, households have a higher propensity to participate in the survey when the first contact was in December. Additionally, when the first contact was via telephone or web, the household's probability to participate increased. Attributes of the interviewer did not play a crucial role in the household's decision to participate. When the household lives at an address with a still

good quality of the area, the participation propensity is higher. Households located in building blocks with a high share of inhabitants aged 30 to 45 have a higher participation propensity. The same holds for households in areas with a high density of constructions. Finally, households living in neighborhoods with a high value ($9^{th}$ decile) of the index of socio-economic deprivation have a lower propensity to participate in the survey.

## 5.2 Population Weighting Adjustments

In the final step of the weighting process, we use post-stratification and raking techniques to adjust the weights obtained previously. This adjustment aligns the weights with known population totals, as well as joint and marginal distributions. The specific method used depends on the available population data, with a comprehensive overview provided by Kalton and Flores-Cervantes (2003). The resulting weights from this step form the basis for deriving both cross-sectional and longitudinal weights for subsequent survey waves, beginning with wave 2. The population parameters and distributions used in these adjustments were provided by the Federal Statistical Office, based on data from the German Microcensus. Margins used in the post-stratification process are:

Number of households with at least one person of the population by

- household typology;
- household size;
- size of municipality;
- federal states and regions as shown in Table 1;
- rural-urban-classification;
- house owner; as well as
- migration background, year of immigration, and nationality

Number of persons of the population by

- sex, nationality, and age groups;
- migration background; as well as
- year of immigration

For information on the categories of the different variables see Siegers, Steinhauer, and Schütt (2022).

# 6 Characteristics of Weights

Due to stratification and disproportional allocation of households, there is some variance in the design weights. Multiplying design weights with the inverse of estimated participation probabilities increases variation in the second weighting step. The population weighting adjustments add to the variation and magnitude of weights. Resulting weights are provided in the variable `hhrf0` included in the data set `hpathl`.

After the integration step, a further post-stratification step was carried out in which the weights (previously nonresponse-adjusted, if necessary post-stratified and integrated) of

all SOEP samples were adjusted with respect to the standard marginal distributions used by SOEP that were taken from the Microcensus 2022. Using the resulting standard SOEP weighting factors (`hhrf` included in `hpathl` and `phrf` included in `ppathl`), the sample R cases can then be analyzed jointly and comparatively in combination with all other SOEP cases.

Table 4: Characteristics of weights after the steps of the weighting process (rounded to integer values).

| | | Quantiles | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Step | Min. | 10% | 25% | 50% | 75% | 90% | Max. | Mean | SD |
| DW | 39 | 260 | 637 | 887 | 1,668 | 2,028 | 4,827 | 1,177 | 859 |
| SWA | 68 | 916 | 2,141 | 3,451 | 6,156 | 8,911 | 29,431 | 4,523 | 3,622 |
| DWA | 139 | 1,264 | 2,375 | 3,962 | 6,756 | 10,686 | 23,039 | 5,154 | 4,060 |

Abbreviations: SD = standard deviation, DW = design weighting, SWA = sample weighting adjustment, PWA = population weighting adjustment.

# 7 Summary

The new Sample R is a refresher sample adding additional 6,560 households to the SOEP. The households have been seamlessly integrated into SOEP-Core. Sampling from the population register was done using a stratified two-stage sampling design. In former lignite mining regions, an oversampling was implemented to allow for detailed analyses on regions being exposed to structural changes. Through its integration into SOEP-Core, the refresher sample itself provides an additional data infrastructure for regional planning and research. Further, the refreshment sample reduces the discrepancies in the age distribution between the panel and the general population.

# References


American Association for Public Opinion Research. (2023). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (10th ed.). Retrieved from https://aapor.org/wp-content/uploads/2024/03/Standards-Definitions-10th-edition.pdf

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical methods in medical research*, *5*(3), 215–238. doi: 10.1177/096228029600500302

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of official statistics*, *19*(2), 81–97.

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey methodology*, *12*(1), 1–16.

Kroh, M., Kühne, S., Jacobsen, J., Siegert, M., & Siegers, R. (2017). *Sampling, nonresponse, and integrated weighting of the 2016 IAB-BAMF-SOEP Survey of Refugees (M3/M4)—revised version* (SOEP Survey Papers No. 477). Berlin:

DIW/SOEP. Retrieved from https://www.diw.de/documents/publikationen/73/diw_01.c.572346.de/diw_ssp0477.pdf

Kroh, M., Siegers, R., & Kühne, S. (2015). Gewichtung und integration von auffrischungsstichproben am beispiel des sozio-oekonomischen panels (soep). In J. Schupp & C. Wolf (Eds.), *Nonresponse bias: Qualitätssicherung sozialwissenschaftlicher umfragen* (pp. 409–444). Wiesbaden: Springer Fachmedien Wiesbaden. Retrieved from https://doi.org/10.1007/978-3-658-10459-7_13 doi: 10.1007/978-3-658-10459-7_13

R Core Team. (2023). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Siegers, R., Steinhauer, H. W., & Schütt, J. (2022). *SOEP-Core v37 Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2018)* (SOEP Survey Papers No. 1106). Berlin: DIW/SOEP.