SOEP The Socio-Economic Panel

827

SOEP Survey Papers

Series C – Data Documentation

SOEP – The German Socio-Economic Panel at DIW Berlin 2020

SOEP-Core – 2018: Sampling, Nonresponse, and Weighting in Sample O

Hans Walter Steinhauer, Martin Kroh, Jan Goebel

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, Martin Kroh, Jan Goebel. 2020. SOEP-Core – 2018: Sampling, Nonresponse, and Weighting in the Sample O. SOEP Survey Papers 827: 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 – 2018: Sampling, Nonresponse, and Weighting in Sample O

Hans Walter Steinhauer, Martin Kroh, Jan Goebel

March 31, 2020

Abstract

This paper provides details on the sampling design, field results and nonresponse, as well as population adjustments for the 2018 Sample O of the Socio-Economic Panel (SOEP). Sample O refreshes the SOEP-Core sample by adding the residents of households in neighborhoods that receive government funding as part of the *Soziale Stadt* (Socially Integrative City) urban development and planning program. To provide a sample of households within these neighborhoods, the SOEP implemented a novel sampling approach integrating grid level census counts on the population and georeferenced building addresses. The approach is both cost efficient and lends itself as an alternative to the often criticized random route techniques. Obtaining nearly 1,000 interviews and panel consent of households in the neighborhoods was a demanding task in the fieldwork period. Nevertheless, nonresponse on the household level was driven primarily by neighborhood characteristics such as social composition and by whether survey staff were able to contact the household.

1 Introduction

Urbanisation and gentrification are common developments across cities in Germany. These processes result in a high spatial concentration of adverse social changes, shifts in economic power, and demographic transitions that can jeopardize the stability of cities and city districts (Keim, 2011). In response to these developments, the German government has instituted a program to provide funding for neighborhoods in cities or city districts that are economically disadvantaged and deprived. This regional development initiative, called the *Soziale Stadt* (Socially Integrative City), aims at stabilizing, upgrading, and improving residential environments within these neighborhoods (Zimmermann, 2011). It does so primarily by investing in construction, infrastructure, environment, and measures to enhance the quality of life and housing. As of 2017, the *Soziale Stadt* program encompassed 891 funding programs in 513 municipalities and a total budget of 190 million euros.[1]

To analyze the impact and effectiveness of these programs, researchers need individual-level data on the people living in the neighborhoods funded through the initiative to be able to describe and monitor perceptions and how they change over time. It is crucial that these data allow for spatial referencing so they can be merged with administrative and other spatial data on the neighborhoods in question. The approach of spatial referencing allows researchers to define specific areas of interest. As a result, they are not restricted to predefined areas such as administrative regions. In addition, regional indicators on different levels can enrich the available survey data.

Being able to analyze administrative data together with panel data on individual household members enables researchers as well as public authorities to evaluate the different programs and their impacts on neighborhoods and the people living there (Goebel, Gornig, & Strauch, 2016). The SOEP is a well-known source of this kind of longitudinal data and already has provided spatial referencing of households since 2010, allowing for detailed analysis of neighborhoods (see Goebel & Pauer, 2014). These data have been used by Aehnelt, Goebel, Gornig, and Häußermann (2009) as well as Goebel, Gornig, and Häußermann (2012) to produce findings on socio-spatial polarization in Germany cities. Moreover, Wüstemann, Kalisch, and Kolbe (2017b) and Wüstemann, Kalisch, and Kolbe (2017a) used SOEP data to analyze environmental inequalities in access to urban green and blue space within Germany.

Nevertheless, an initial cooperation project between the SOEP and the Federal Institute for Research on Building, Urban Affairs and Spatial Development (BBSR) showed two limitations in the use of SOEP data for empirical analysis of research questions related to neighborhoods receiving government funding through the *Soziale Stadt* program. First, the number of households is not sufficient for the diverse research questions that need to be studied with these data, and second, the sample covers only about half of the neighborhoods receiving funding through the *Soziale Stadt* program (Goebel et al., 2016). To address these issues, the cooperation between SOEP and BBSR continues, and more households in neighborhoods receiving funding through the *Soziale Stadt* program have been added to the sample. The new Sample O refreshes SOEP-Core by increasing the number of households for general research questions, and enables researchers to address specific research questions in urban and spatial research. In so doing, Sample O enhances

[1] https://www.staedtebaufoerderung.info/StBauF/DE/Programm/SozialeStadt/soziale_stadt_node.html

the SOEP as a source of national, spatially referenced data. For more information on the SOEP in general, see Goebel et al. (2019).

Sampling via registers for our purposes is hampered by the absence of information on neighborhoods receiving funding through the *Soziale Stadt* program. If registers are not available, random route techniques are often used for random sampling of addresses. The two-step procedure involves, first, a random sampling of pre-defined areas and, second, a random selection of addresses within these areas (Arbeitskreis deutscher Markt- und Sozialforschungsinstitute, 2013). At the second sampling stage, interviewers receive a starting point and instructions which random route to take. Along this route they list a predefined number of addresses on their way. This sampling approach bears the risk that interviewers do not follow instructions and list convenient addresses instead. Moreover, Bauer (2014) and others have shown that even in the absence of interviewer effects, random route leads to unequal selection probabilities of households. Finally, for drawing a sample of households within the relevant neighborhoods, the random walk approach based on the ADM design (see Heckel & Hofmann, 2014), as used in previous samples of the SOEP (see Siegers, Belcheva, & Silbermann, 2020), could not be applied, because even if the starting point for a random walk was set within the boundaries of the relevant neighborhoods, interviewers might cross these boundaries when sampling households along the route.

In this paper we propose an alternative approach for randomly sampling addresses within a specific area. To avoid the shortcomings of the random route technique, the BBSR provided shape files restricting the areas within the cities, so that building addresses could be located within the relevant neighborhoods. For this purpose, we combine grid level census counts on the number of inhabitants within neighborhoods of the *Soziale Stadt* and georeferenced building addresses provided by the Federal Agency for Cartography and Geodesy (BKG). The grid level information from the census covers an area of 100×100 meters. Thus, the areas of the *Soziale Stadt* cover several grids of one hectare. Combining these informations enabled us to draw a random sample of addresses that definitely fall within the boundaries of the *Soziale Stadt* area. In contrast to random walk approaches, this procedure also yields positive and known inclusion probabilities for sampled addresses. In addition, directly sampling addresses reduces the workload of the fieldwork agency and the interviewer. Thus, this novel approach is, in contrast to random route sampling, cost-efficient and can be transferred to different applications, among others, to sampling of general population surveys.

This paper documents the sampling design and weighting strategy used in the 2018 Sample O of the SOEP. Section 2 describes the population and provides further information on the *Soziale Stadt* program. The novel approach used to directly sample building addresses is described in section 3. Section 4 provides detailed information on the fieldwork and its results. Weighting adjustments are presented in Section 5 and section 6 details the resulting characteristics of weights. Finally, Section 7 gives a brief summary.

2 Target Population and Sampling Frame

The target population of Sample O consists of all residents of private households in Germany within neighborhoods receiving government funding through the *Soziale Stadt* program in the year 2018. Figure 1 shows 331 municipalities (in green) that received government funding through *Soziale Stadt* program (in red) in the year 2017. Within these

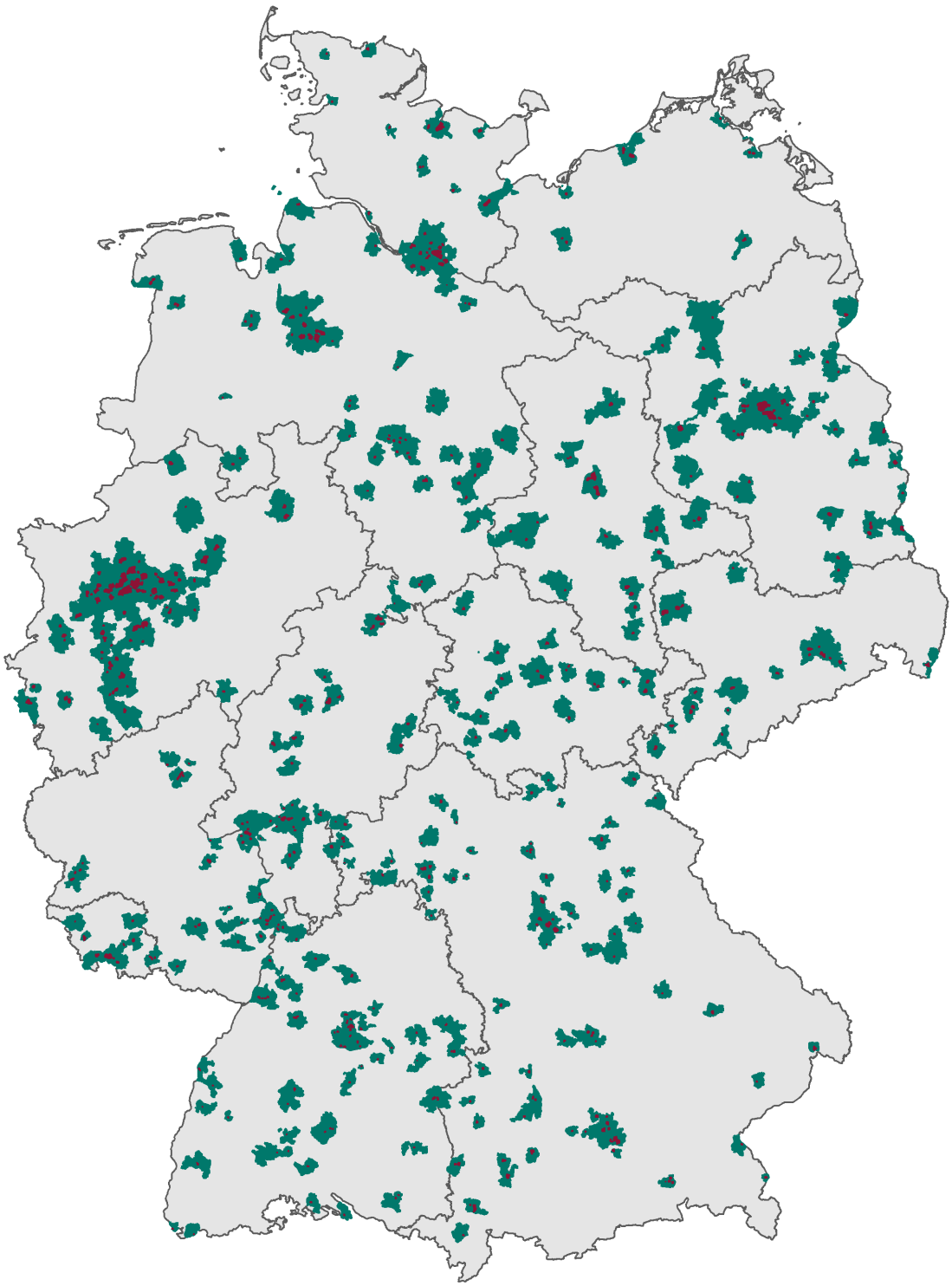

Figure 1: Municipalities (green) with neighborhoods receiving funding through the *Soziale Stadt* program (red). Source: authors' depiction based on shape files provided by the BBSR.

municipalities, there were 626 neighborhoods in the *Soziale Stadt* program in 2017. The numbers reported here differ slightly from those reported above because the BBSR did not have all neighborhoods available as spatially referenced shape files. These neighborhoods may be adjacent to one another or dispersed across the city. From the map, it can be seen that there are differences in the number of municipalities in the different federal states. Moreover, clustering of neighborhoods receiving funding through the *Soziale Stadt* program can be seen within federal city states as well as in the Ruhr area. Table 1 provides the number of municipalities and the number of funding programs by federal state. Most of the neighborhoods in the *Soziale Stadt* program are within urban areas but a few are also located in rural areas. To sample from this population, it was necessary to restrict sampling to the areas within the neighborhoods of the *Soziale Stadt*. To make this possible, the BBSR provided shape files in the first step so that the areas were known. Sampling based on the ADM design appeared imprecise because the sampling points did not precisely match the neighborhoods in the *Soziale Stadt* program, and a random walk could end outside of these areas. Also, register-based sampling did not qualify as an approach, because it would be difficult to restrict the register to addresses within the area. But spatially referenced addresses provided by the Federal Agency for Cartography and Geodesy (BKG) allowed us to restrict the addresses of buildings to the neighborhoods in the *Soziale Stadt*program. Thus, we were able to sample addresses on the basis of shape files that restricted sampling to the neighborhoods in the *Soziale Stadt* program and the addresses within that area.

Table 1: Number of municipalities and funding programs by federal state.

Federal State	Number of municipalities	Thereof with funding program	Number of funding programs
BW	1,101	47	81
BY	2,056	68	94
BE	1	1	41
BB	417	26	30
HB	2	2	14
HH	1	1	19
HE	423	22	39
MV	750	6	11
NI	945	34	53
NW	396	45	106
RP	2,304	17	36
SL	52	7	12
SN	421	15	30
ST	218	14	21
SH	1,106	10	16
TH	821	16	23
Total	11,014	331	626

Note: BW = Baden-Wuerttemberg, BY = Bavaria, BE = Berlin, BB = Brandenburg, HB = Bremen, HH = Hamburg, HE = Hessen, MV = Mecklenburg-Western Pomerania, NI = Lower Saxony, NW = North Rhine-Westphalia, RP = Rhineland-Palatinate, SL = Saarland, SN = Saxony, ST = Saxony-Anhalt, SH = Schleswig-Holstein, TH = Thuringia.

3 Sampling Design

As there is no list of private households residing within these neighborhoods, and the population register in Germany does not allow sampling on the basis of spatially referenced addresses, direct sampling approaches cannot be used. Alternative sampling schemes, for instance, a random walk approach from a given starting address, has the drawback that the interviewer may end up leaving the neighborhoods of the *Soziale Stadt* on his or her random walk. Therefore, we propose a new sampling scheme using spatially referenced geographical data. The BBSR delivered shape files defining the neighborhoods of the *Soziale Stadt*. Together with information provided by the German census from 2011 including census grids (Zensuskacheln) and the number of buildings and inhabitants in these grids, the SOEP was able to sample addresses.

Before sampling, however, and in the absence of a national, spatially referenced register, we constructed primary sampling units on the basis of census grids. These were assigned to the corresponding neighborhoods in the *Soziale Stadt* program. For sampling, primary sampling units (PSU) were defined as an aggregation of census grids within the neighborhoods of the *Soziale Stadt* program with similar numbers of inhabitants and similar spatial geographic proximity. The purpose of this aggregation was to ensure that the number of persons living within each PSU was between 1,500 and 3,000. In a second step, these PSUs were allocated to the strata defined by federal state and municipality size as follows:

$h = 1$ Baden-Wuerttemberg
$h = 2$ Bavaria, municipality size up to 100.000 inhabitants
$h = 3$ Bavaria, municipality size more than 100.000 inhabitants
$h = 4$ former western Berlin (Kreuzberg, Schöneberg, Tempelhof, Neukölln)
$h = 5$ former western Berlin (other)
$h = 6$ former eastern Berlin
$h = 7$ Brandenburg
$h = 8$ Bremen
$h = 9$ Hamburg
$h = 10$ Hessen
$h = 11$ Mecklenburg-Western Pomerania
$h = 12$ Lower Saxony
$h = 13$ North Rhine-Westphalia, municipality size more than 500.000 inhabitants
$h = 14$ North Rhine-Westphalia, municipality size 100.000 - 500.000 inhabitants
$h = 15$ North Rhine-Westphalia, municipality size up to 100.000 inhabitants
$h = 16$ Rhineland-Palatinate, Saarland
$h = 17$ Saxony
$h = 18$ Saxony-Anhalt
$h = 19$ Schleswig-Holstein
$h = 20$ Thuringia

At the first stage (indicated by the superscript ^{I}), a total of $m^I = \sum_{h=1}^{H} m_h^I = 125$ PSUs have been sampled from this stratified population. Due to a very unequal use of the *Soziale Stadt* program (see Figure 1), we reduced the inclusion probability in Berlin, Bremen, Hamburg, and North Rhine-Westphalia ($h \in \{4; 5; 6; 8; 9; 13; 14; 15\}$) by half to avoid too many sampling points in certain districts.

The PSUs were selected systematically with probability proportional to the number of

inhabitants N_{jh} estimated by the 2011 Census for PSU j in stratum h. Thus, the inclusion probability π_{jh} for PSU j in stratum h arises as

$$\pi_{jh} = m_h^I \cdot \frac{N_{jh}^*}{\sum_{j=1}^{M_h^I} N_{jh}^*}, \text{with}$$

$$N_{jh}^* = p_h \cdot N_{jh} \text{ and}$$

$$p_h = \begin{cases} \frac{1}{2}, & \text{if } h \in \{4;5;6;8;9;13;14;15\} \\ 1, & \text{otherwise.} \end{cases}$$

With m_h^I denoting the number of PSUs to be sampled in stratum h, M_h^I denoting the total number of PSUs in stratum h, and N_{jh} denoting the number of inhabitants covered by the j^{th} PSU in stratum h.

Having sampled the PSUs, buildings were selected by simple random sampling as secondary sampling units (SSUs). To sample households within these buildings, the number of households and the number of persons living in these households was simulated. To do so, in a first step, the number of households per building was simulated using an exponential distribution with $\lambda = 0.2$. Next, in a second step, the number of people N_j residing in PSU j were allocated to the buildings, proportionally to the number of households. Then, before selecting the households, a sample of buildings was drawn at the second stage (indicated by the superscript ^{II}). Here, $m^{II} \leq 40$ buildings were selected in each PSU using simple random sampling.[2] The corresponding inclusion probability for a building b is given by

$$\pi_b = \frac{m^{II}}{M^{II}},$$

with M_{jh}^{II} denoting the total number of buildings within each PSU j in stratum h. Finally, for the m^{II} selected buildings, interviewers counted the number of households. The total of $m^{III} = 80$ households to sample within a PSU was allocated proportionally to the number of households (N_b) per sampled building b. Finally, the households to be selected per building were determined and the corresponding households were selected using the Kish selection grid (*Schwedenschlüssel*, see Kish (1949) for details). Given the selected building b in PSU j in stratum h, the inclusion probability for household i is

$$\pi_i = \frac{m_b}{M_b},$$

where M_b (m_b) denotes the total number of households (to be selected) in building b.

4 Fieldwork Results and Response Rates

After sampling 2,642 addresses of buildings, the interviewers with the fieldwork agency Kantar Public visited the addresses, first, to make sure the buildings were residential

[2] To limit the workload of interviewers when listing addresses of selected buildings, we stopped selecting buildings if either a maximum of 400 estimated households or a maximum of 40 buildings was reached. The limits were chosen on the basis of the observed distributions affecting mostly outliers.

buildings and, second, to identify the households to be sampled based on the Kish selection grid (Kish, 1949).

Table 2 displays the results for the fieldwork for the 6,625 sampled households. In total, there were 935 complete or partial interviews, resulting in a response rate on the household-level, calculated according to The American Association for Public Opinion Research (2016), of RR2 = 0.153. Although the response rate at the household-level is lower than expected, the refusal rate (REF1 = 0.565) is similar to other samples / studies. Further, a substantial portion of the addresses were classified as quality-neutral drop-outs. Most of these were in buildings that did not contain any households, which is common in economically distressed neighborhoods. In addition, in some cases, the listed households in residential blocks (mostly high-rise apartment buildings) had moved out by the time of the survey. Finally, several households could not be contacted within the field period.

Table 2: Fieldwork results on the household-level.

	Number	Proportion
Interview		
Complete interview	698	0.105
Partial interview	237	0.036
No Interview		
Refusal	3,455	0.522
Non-contact	1,378	0.208
Language problems	187	0.028
Physically or mentally unable/incompetent	113	0.017
Other	37	0.006
Away/unavailable	14	0.002
Quality neutral drop-out		
Not a housing unit	506	0.076
Total	6,625	1.000

Table 3 details the number of households by federal state for the initial sample and the successfully surveyed sample. The remaining columns give the corresponding proportions. In total 6,625 were sampled, 935 of which completed a household interview.Table 3 details the number of households by federal state for the initial sample and the successfully surveyed sample. The remaining columns give the corresponding proportions. In total 6,625 were sampled, 935 of which completed a household interview.

5 Cross-Sectional Weighting

According to Brick and Kalton (1996) the computation of weights is usually performed in three steps. In the first step, design weights are calculated as the inverse of the inclusion probability, see Section 3. Second, these design weights are adjusted to correct for unit nonresponse. Kalton and Kasprzyk (1986) refer to this step as sample weighting adjustment. Finally, in a third step, weights are calibrated so that estimates conform to known population parameters, for example, totals or ratios, or to meet specific distributions.

Table 3: Number (N) and proportion (θ) of households in the initial and realised sample.

Federal State	Sampled N	Sampled θ	Realised N	Realised θ
BB	371	0.056	71	0.076
BE	742	0.112	91	0.097
BW	371	0.056	45	0.048
BY	354	0.144	129	0.138
HB	159	0.024	16	0.017
HE	424	0.064	60	0.064
HH	318	0.048	48	0.051
MV	159	0.024	36	0.039
NI	318	0.048	35	0.037
NW	1,219	0.184	190	0.203
RP	318	0.048	65	0.070
SH	212	0.032	25	0.027
SL	53	0.008	10	0.011
SN	371	0.056	36	0.039
ST	371	0.056	50	0.053
TH	265	0.040	28	0.030
Total	6,625	1.000	935	1.000

This step is referred to by Kalton and Kasprzyk (1986) as population weighting adjustment. For details on the general weighting strategy of the SOEP and the integration of new samples, see Kroh, Siegers, and Kühne (2015).

To account for possible selectivity due to nonresponse, we model the participation decision of the households using information on participating and nonparticipating households. Because there is usually very little information available on nonparticipating households, we use area-level information as well as interviewer observations on the residential environment. Information collected by the interviewer on the residential environment includes: problems speaking German, condition of the neighborhood, condition of the building, access problems due to physical barriers (e.g. locked doors, fences), access problems due to intercom system, other access problems, safety of the neighborhood, composition of the housing area, type of building (according to number of households). District-level information is obtained from INKAR online (Indikatoren und Karten zur Raum- und Stadtentwicklung; www.inkar.de). INKAR provides information on (un)employment, construction and housing, education, infrastructure, population characteristics, and other regional indicators. The time reference for the data is 2015. Detailed documentation on the variables in the data is provided by (INKAR, 2019). Lower-level information used in the nonresponse analysis is provided by Microm, mainly on the street level (www.microm.de). Microm provides information about the social structure of neighborhoods in Germany on the regional and local levels. The local level refers to different aggregations such as eight-digit postal code areas covering approximately 500 households, street-level, or household cells aggregating a few households.

5.1 Sample Weighting Adjustment

When correcting the design weights in the second step, strong predictors of nonresponse are needed. To find these, we iterate through all variables included in interviewer observations, INKAR, and Microm, and select those that significantly influence the participation decision in a bivariate regression analysis. In a second step, we omit those variables from the set of significant variables with an absolute value of correlation among each other of greater than or equal 0.95. Finally, the remaining variables enter a preparatory nonresponse model. To obtain the final model, we run variable selection in both directions using the BIC as a selection criterion. This yields a more parsimonious model. The model finally estimating the response propensities used to derive weighting adjustments is presented in Table 4.

In the final model, we find several strong predictors of nonresponse. Among the strongest are interviewer observations and regional information.

Proceeding from aggregate to building-level information, we find county-level information provided by INKAR to show positive effects only on the participation probability. Here, the strongest effect stems from the increase in the life expectancy of women, followed by the living area per resident, the decrease in the development of numbers of self-employed[3], and finally, naturalization per 1,000 foreigners.

Looking at the regional information provided by Microm, we see influences of the milieus, see Hempelmann and Flaig (2019). First, households residing in a neighborhood with a predominantly hedonist milieu have a higher likelihood of participating in sample O. Second, households residing in a neighborhood with a predominantly cosmopolitan avant-garde milieu have a lower likelihood of participating in sample O. Both of these are on the same end of the social status scale (reorientation), but the latter is in the lower / lower middle class segment, whereas the former is in the upper / upper middle class segment. Moreover, the economic status and demographics of the neighborhood affect the decision of households to participate. In neighborhoods that mainly consist of financially secure families with children, the participation propensity is higher. In contrast, neighborhoods with mainly financially weak elderly single people have a lower participation propensity. A lower participation propensity is also found in neighborhoods where Volkswagen is the dominant brand of automobile. For the Microm information, the finding that an increase in the likelihood to move or relocate results in higher participation propensities seems to be somewhat odd.

Finally, the most detailed building-level information collected by the interviewers reflects typical difficulties in the recruitment process. If the interviewer faced problems when asking the household to complete the survey in German, these problems in communication reduced the participation propensity of the household. Moreover, interviewers had higher participation propensities when the households were located in an area interviewers perceived as very safe, and when interviewers had no problems contacting the household.

The model presented above yields a pseudo-R^2 of $R^2_{final} = 0.042$ compared to the full model with $R^2_{full} = 0.026$ and the null model with $R^2_{null} < 0.001$. So for all the models considered, the model fit is quite low. Nevertheless, from the large number of variables that have been tested, only a few turn out to significantly influence the household's

[3]This decrease is because the percentages are mainly negative, thus indicating a decrease in the number of self-employed people.

Table 4: Model estimating response propensities used to derive weighting adjustments.

Variable Value	estimate (std. error)
(Intercept)	-6.057^{***}
	(0.540)
County level information (INKAR)	
Living area per resident	0.060^{***}
in m^2	(0.010)
Change in the number of self-employed people	0.056^{***}
in %	(0.012)
Naturalization	0.030^{**}
per 1,000 foreigners	(0.009)
Change in life expectancy	0.191^{***}
of female newborns	(0.038)
Regional information (MIKROM)	
Number of moves / relocations	0.069^{***}
unlikely to very likely	(0.020)
Social milieu at street level	0.011^{**}
Hedonist	(0.003)
Social milieu at the PLZ8 level	-0.565^{***}
Cosmopolitan avant-garde	(0.162)
Dominant brand of car at PLZ8 level	-0.296^{***}
Volkswagen	(0.080)
Status and phase of life at PLZ8 level	0.670^{***}
financially secure family with child	(0.154)
Status and phase of life at PLZ8 level	-0.380^{**}
financially weak elderly single person	(0.117)
Interviewer observations	
Problems speaking German	-0.574^{**}
severe	(0.181)
Access problems due to physical barriers	0.598^{***}
none	(0.073)
Safety of neighborhood	0.416^{***}
very safe	(0.088)
Log likelihood	−2,493.178
N	6,119

Notes: Dependent variable: 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, 2019).

participation decision. All in all, the findings indicate little selectivity among the households located in the neighborhoods in the *Soziale Stadt* program.

5.2 Population Weighting Adjustment

In the last step of the weighting process, we use post-stratification and raking to adjust the weights from the previous step to meet known totals as well as joint and marginal distributions. To achieve this, there are several methods that can be used depending on the data available for the population. An overview of methods is provided by Kalton and Flores-Cervantes (2003). The weights resulting from this step are the basis for cross-sectional and longitudinal weights derived for wave 2 and beyond.

The population parameters and distributions used in the population weighting adjustments were provided by the Federal Statistical Office based on the German Microcensus; see Table 6 in the Appendix. The target population has been identified by using the census grids that were used for sampling. For these areas, the parameters and distributions have been derived from the Microcensus.

At the household-level the following distributions have been used:

- Number of households by federal state
- Number of households by municipality size
- Number of households by household size
- Number of households by household type
- Number of households by year of birth and migration experience
- Number of households by earliest immigration year
- Number of households by specific nationality of the household head
- Number of households in the former territory of Germany

At the individual level the following marginal and joint distributions have been used:

- Number of persons by migration background and year of birth
- Number of persons by immigration year
- Number of persons by nationality
- Number of persons by age group and gender

6 Characteristics of Weights

The weights for the first wave of sample O have been derived in three steps (design weighting, sample weighting adjustment, and population weighting adjustment). The characteristics for the weights on the household-level resulting from each step are displayed in Table 5.

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 very little to the variation of weights and at the same time reduce the magnitude, especially of the large weights.

Table 5: Characteristics of weights after the steps of the weighting process.

		Quantiles							
Step	Min.	10%	25%	50%	75%	90%	Max.	Mean	SD
DW	148	263	307	371	526	698	1,141	433	193
SWA	544	1,037	1,537	2,253	3,779	5,410	17,008	2,846	1,960
PWA	89	552	1,071	1,918	3,530	5,540	9,440	2,559	2,032

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

7 Summary

The new Sample O is a SOEP refresher sample adding an additional 935 households to the SOEP. They have been seamlessly integrated into SOEP-Core. Sampling was done using a novel approach based on spatially referenced data. Shape files restricting the sample to residential areas in which the *Soziale Stadt*program was being implemented were provided by the BBSR. Within the neighborhoods of the *Soziale Stadt* program, buildings were sampled and within these buildings, households were selected using the Kish selection grid. This novel sampling approach turned out to work very well and was accepted by the fieldwork agency as well as by the interviewers. In addition, the novel sampling approach proved to have several advantages over the traditionally used ADM design. Through its integration into SOEP-Core, the refresher sample itself provides an additional data infrastructure for urban and regional planning and research. The data can, in future waves, be used to evaluate the *Soziale Stadt* urban development and planning program on a national level. Moreover, the data allow for analysis of households in nearby neighborhoods stemming from a random sample.

References

Aehnelt, R., Goebel, J., Gornig, M., & Häußermann, H. (2009). Soziale Ungleichheit und sozialräumliche Strukturen in deutschen Städten. *Informationen zur Raumentwicklung*, *6*, 405–413.

Arbeitskreis deutscher Markt- und Sozialforschungsinstitute. (2013). *Stichproben-Verfahren in der Umfrageforschung: eine Darstellung für die Praxis*. Wiesbaden: Springer-Verlag.

Bauer, J. J. (2014). Selection errors of random route samples. *Sociological Methods & Research*, *43*(3), 519–544. doi: 10.1177/0049124114521150

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

Goebel, J., Gornig, M., & Häußermann, H. (2012). Bestimmt die wirtschaftliche Dynamik der Städte die Intensität der Einkommenspolarisierung? *Leviathan*, *40*(3), 371–398. doi: 10.5771/0340-0425-2012-3-371

Goebel, J., Gornig, M., & Strauch, K. (2016). Nutzung von Bevölkerungsdaten für ein erweitertes Monitoring der Städtebauförderung. *Informationen zur Raumentwicklung*, *1*, 53-62.

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics*, *239*(2), 345–360. doi: 10.1515/jbnst-2018-0022

Goebel, J., & Pauer, B. (2014). Datenschutzkonzept zur Nutzung von SOEPgeo im Forschungsdatenzentrum SOEP am DIW Berlin. *Zeitschrift für amtliche Statistik Berlin Brandenburg*, *3*, 42–47.

Heckel, C., & Hofmann, O. (2014). Das ADM-Stichproben-System (F2F) ab 1997. In *Stichproben-Verfahren in der Umfrageforschung* (pp. 85–116). Springer. doi: 10.1007/978-3-531-18882-9_5

Hempelmann, H., & Flaig, B. B. (2019). *Aufbruch in die Lebenswelten*. Springer.

INKAR. (2019). *Indikatorenübersicht – Indkatoren Raum- und Zeitbezüge*. Retrieved from https://www.inkar.de/documents/Indikatoren%20Raum-%20und%20Zeitbezuege.pdf

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

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

Keim, R. (2011). Soziale Stadt und sozialräumliche Ausgrenzung: Wohnen und öffentlicher Raum. In W. Hanesch (Ed.), *Die Zukunft der "Sozialen Stadt"* (pp. 241–255). Springer. doi: 10.1007/978-3-531-92637-7_11

Kish, L. (1949). A procedure for objective respondent selection within the household. *Journal of the American Statistical Association*, *44*(247), 380-387.

Kroh, M., Siegers, R., & Kühne, S. (2015). Gewichtung und Integration von Auffrischungsstichproben am Beispiel des Sozio-oekonomischen Panels (SOEP). In *Nonresponse bias* (pp. 409–444). Springer.

R Core Team. (2019). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Siegers, R., Belcheva, V., & Silbermann, T. (2020). *SOEP-Core v35 Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2018)* (SOEP Survey Papers No. 826). Berlin: DIW/SOEP.

The American Association for Public Opinion Research. (2016). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys* (9th ed.). AAPOR.

Wüstemann, H., Kalisch, D., & Kolbe, J. (2017a). Accessibility of urban blue in german major cities. *Ecological Indicators*, *78*, 125–130. doi: 10.1016/j.ecolind.2017.02.035

Wüstemann, H., Kalisch, D., & Kolbe, J. (2017b). Access to urban green space and environmental inequalities in germany. *Landscape and Urban Planning*, *164*, 124–131. doi: 10.1016/j.landurbplan.2017.04.002

Zimmermann, K. (2011). Der Beitrag des Programms "Soziale Stadt" zur Sozialen Stadtentwicklung. In W. Hanesch (Ed.), *Die Zukunft der "Sozialen Stadt"* (pp. 181–201). Springer. doi: 10.1007/978-3-531-92637-7_8

Appendix

Table 6: Population characteristics used in post-stratification and raking procedures.

Variable	Values
Federal state	Hamburg / Schleswig-Holstein Bremen / Lower Saxony North Rhine-Westphalia Hessen Baden-Wuerttemberg Bavaria Rhineland-Palatinate / Saarland Berlin / Brandenburg Mecklenburg-Western Pomerania Saxony Saxony-Anhalt Thuringia
Municipality size	less than 20,000 residents 20,000 up to 100,000 residents 100,000 up to 500,000 residents 500,000 residents or more
Household size	single household 2 person household 3 person household 4 person household household with 5 or more persons
Household type	single two persons, no children two adults, no more than two children single parent, no more than two children single parent, at least three children more than two adults, at least three children
Indicator for year of birth and migration experience	No person in household with indirect migration experience At least one person with indirect migration experience, born 1995 or later At least one person with indirect migration experience, born 1975-1994 At least one person with indirect migration experience, born 1975-1994 and 1995 or later At least one person with indirect migration experience, born 1974 or earlier

Continued on next page

Table 6 – *Continued from previous page*

Variable	Values
Former federal territory	West Germany East Germany Berlin separated by postal code
Migration background	without direct, German nationality indirect, German nationality direct, non-German nationality indirect, non-German nationality
Year of birth	before 1965 1965 - 1974 1975 - 1984 1985 - 1994 1995 - 2004 2005 and later
Earliest immigration year	household with no persons having migration background household with at least one person having immigrated between (from-up to) 2014-2017 2010-2013 2005-2009 2000-2004 1995-1999 1990-1994 1985-1989 1980-1984 1979 or earlier
Nationality (detailed)	German Turkish Spanish Greek Italian Polish Romania Yugoslavia Russian Arabic EU-foreigner without Slovenia/Croatia EU-foreigner with Slovenia/Croatia Others

Continued on next page

Table 6 – *Continued from previous page*

Variable	Values
Age Group (from - younger than)	0 - 5 5 - 10 10 – 15 15 – 20 20 – 25 25 – 30 30 – 35 35 – 40 40 – 45 45 – 50 50 – 55 55 – 60 60 – 65 65 – 70 70 – 75 75 – 80 80 – 85 85 – 90 90 – 95 95 and older 50 – 60 60 – 70 70 – 80 80 and older 50 – 65 65 and older
Gender	Female Male