Synthetic data for open and reproducible methodological research in social sciences and official statistics

Burgard, Jan Pablo; Kolb, Jan-Philipp; Merkle, Hariolf; Münnich, Ralf

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Empfohlene Zitierung / Suggested Citation:
Burgard, J. P., Kolb, J.-P., Merkle, H., & Münnich, R. (2017). Synthetic data for open and reproducible methodological research in social sciences and official statistics. *AStA Wirtschafts- und sozialstatistisches Archiv : eine Zeitschrift der Deutschen Statistischen Gesellschaft*, *11*(3-4), 233-244. https://doi.org/10.1007/s11943-017-0214-8

AStA Wirtsch Sozialstat Arch (2017) 11:233–244
https://doi.org/10.1007/s11943-017-0214-8

ORIGINALVERÖFFENTLICHUNG

Synthetic data for open and reproducible methodological research in social sciences and official statistics

Jan Pablo Burgard · Jan-Philipp Kolb · Hariolf Merkle · Ralf Münnich

Received: 21 November 2017 / Accepted: 24 November 2017 / Published online: 11 December 2017

Abstract Open and reproducible research receives more and more attention in the research community. Whereas empirical research may benefit from research data centres or scientific use files that foster using data in a safe environment or with remote access, methodological research suffers from the availability of adequate data sources. In economic and social sciences, an additional drawback results from the presence of complex survey designs in the data generating process, that has to be considered when developing and applying estimators.

In the present paper, we present a synthetic but realistic dataset based on social science data, that fosters evaluating and developing estimators in social sciences. The focus is on supporting comparable and reproducible research in a realistic framework providing individual and household data. The outcome is provided as an open research data resource.

Keywords Open and reproducible research · Open data · Household data · Simulation experiments

J. P. Burgard · H. Merkle · R. Münnich (✉)
FB IV, VWL, Wirtschafts- und Sozialstatistik, Universität Trier, Universitätsring 15, 54296 Trier, Germany
E-Mail: muennich@uni-trier.de

J. P. Burgard
E-Mail: burgardj@uni-trier.de

H. Merkle
E-Mail: merkle@uni-trier.de

J.-P. Kolb
Leibniz-Institut für Sozialwissenschaften in Mannheim, GESIS, B2 1, 68159 Mannheim, Germany
E-Mail: Jan-Philipp.Kolb@gesis.org

Zusammenfassung In der Forschung nehmen Vergleichbarkeit und Reproduzierbarkeit immer mehr an Bedeutung zu. Die empirische Forschung profitiert dabei von Forschungsdatenzentren und Scientific Use Files. Für die angewandte Methodenforschung sind geeignete Datenquellen kaum verfügbar. Für angewandte Methodenforschung dagegen sind geeignete Datenquellen kaum verfügbar, obwohl gerade in den Wirtschafts- und Sozialwissenschaften komplexe Stichprobendesigns bei der Entwicklung und Anwendung von Schätzmethoden berücksichtigt werden müssen.

In dieser Arbeit wird ein synthetischer, jedoch realistischer Datensatz vorgestellt, der gerade die Evaluierung und Entwicklung von Schätzmethoden in den Sozial- und Wirtschaftswissenschaften unterstützt. Der Schwerpunkt liegt dabei auf vergleichbarer und reproduzierbarer Forschung in einer realistischen Umgebung in Bezug auf Individual- und Haushaltsdaten. Dieser Datensatz wird der Forschungsgemeinde frei zur Verfügung gestellt.

Schlüsselwörter Vergleichbare und reproduzierbare Forschung · Open Data · Haushaltsdaten · Simulationsstudien

1 Introduction

Statistical applications using individual and household data in general are split into the two areas design-based and model-based inference. Official statistics is mainly interested in parameters of a finite population like totals, means, and proportions. The adequate underlying concept of inference is design-based with respect to the underlying sampling process. Empirical researchers using household- and individual-level data are mainly interested in statistical models which are based on model inference. Kalton (2002) stresses the importance of design-based inference and points out that new model-based methods like imputation for handling missing values and small area statistics urge the needs of considering both types of inference. Design- and model-based inference are certainly using the same data which in social sciences and humanities are mainly based on complex samples. One recent and highly discussed topic is poverty measurement, which brings both *worlds* together (Pratesi 2016). In Europe, the major data source is the European Union Statistics on Income and Living Conditions (EU-SILC, http://ec.europa.eu/eurostat/web/microdata/european-union-statistics-on-income-and-living-conditions).

Developing adequate estimators for household and individual data relies on having appropriate data available. Due to disclosure limitations such data, however, are not accessible for a large part of the research community. Further, evaluating or developing estimation methods using complex survey designs is only possible when having data at universe level as well as corresponding sample files. With the ongoing increase in computational power, these evaluations and advances in methodological research for social sciences and social statistics can be achieved using Monte-Carlo simulation techniques. A huge advantage of using this kind of simulation is the possibility to compare and evaluate estimation strategies in a close-to-reality environment using complex survey sampling designs.

To foster open and reproducible research for design-based Monte-Carlo simulations, an appropriate realistic universe including samples drawn using different underlying survey designs has to be made available. The present paper presents the AMELIA dataset which provides a realistic framework for open and reproducible research based on EU-SILC data (http://www.amelia.uni-trier.de). This data source was first developed within the AMELI (Advanced Methodolgy for European Laeken Indicators) research project enabling comparative research for poverty measures (http://ameli.surveystatistics.net and Graf et al. 2011a) and further developed within the InGRID (Inclusive Growth Research Infrastructure Diffusion) research infrastructur (Merkle and Münnich 2016). Within the EU research infrastructure InGRID2 (http://www.inclusivegrowth.eu), the data resource will be turned into a data repository for individual- and household-level data. The aim is to foster methodological research in the area of poverty measurements as well as working and living conditions. The steps suggest an enhancement with complex sampling designs, the integration of longitudinal data for EU-SILC, and the integration of major variables based on the Eurosystem's Household Finance and Consumption Survey (HFCS, https://www.ecb.europa.eu/stats/ecb_surveys/hfcs/html/index.en.html).

Following the aims of the research projects and the user requests, the data resource shall support considering the following ideas:

Investigation of design-based properties of statistical methods Budget constraints or the interest in rare population often urge introducing sophisticated sampling schemes. In many countries contributing to the HFCS, oversampling routines yield so-called informative sampling designs (HFCN 2013). Using simulation studies on this dataset enables to understand possible impacts of oversampling (or other complex designs) on the inference of statistical methods, i.e. the significance of variables in a model.

Releasing realistic test data for comparative research Using unit-level data for individuals and households often suffers from disclosure limitations. Establishing new methods like small area estimation methods, however, depends on the availability of microdata with the necessary granularity. Model-based properties may also be investigated under different sampling designs. Furthermore, the realistic setting also provides a resource for investigating competing methods.

Assistance in peer-review process In many articles, new methods are applied to interesting datasets. However, since the datasets are often under subject to data provision contracts, reviewers cannot test their own routines to compare or better understand details of a study. The availability of a common similar open dataset allows to apply the methods on the open dataset to enable reviewers to compare results with their own methods.

Open and reproducible research One major task of open and reproducible research is to foster sustainable use of data.

Code benchmarking Statistical research often focuses on the implementation of better or alternative algorithms. Archived data can provide an essential base for storing methods and results that allow comparing these results under concurring algorithms, with different software packages or implementations.

AMELIA mimics real data, i.e. displays marginal distributions and basic interactions between variables of EU-SILC data. Even if the dataset is an ideal platform for methodological research, it cannot be used for displaying real data driven output such as poverty or social exclusion indicators in Europe.

Fienberg (1994) shows the conflicts arising between microdata access and needs for confidentiality (see also Hundepool et al. 2012, Chapts. 3.8 and 6). De Wolf (2015) deals with this issue in case of public files of EU-SILC data (for further reading see https://ec.europa.eu/eurostat/cros/search/custom-taxonomy/knowledge-repository-general-innovation-area/disclosure-control_en).

Additional to mimicking the original distribution to provide one *dataset*, we aim to provide a realistic dataset that is safe in terms of anonymity but supports methodological research in economic and social sciences and social statistics considering the items above. The data resource will be accompanied by a database indicating use and applications of the dataset.

2 AMELIA Platform

2.1 Methods

Within the AMELI project, the aim was to produce a synthetic but realistic dataset based on EU-SILC variables that allows investigating and further developing statistical methods for estimating poverty and social exclusion. Further, while mimicking the underlying real distributions, rare variable combinations should not be resampled or duplicated to avoid disclosure risks.

The main purpose of the AMELIA dataset is to foster comparing different survey methods and to assess reproducibility which are not only relevant for journal reviewers and editors, but also for validation and comparison within the scientific community. Synthetic data should reflect the properties of the underlying microdata, like the hierarchical structure of households within communities and regions. The process of synthetic data generation invokes the preservation of the original correlation between variables as well as the realistic heterogeneity between and within hierarchical levels. There are manifold methods to create synthetic universes. Amongst others, these techniques are described in Münnich et al. (2003), Münnich and Schürle (2003), Kolb (2012) and Alfons et al. (2011b) as well as in microsimulation literature like Lovelace and Dumont (2016) and Rahman and Harding (2016). An established method to create synthetic universes beside regression modelling is called synthetic reconstruction which is drawing from conditional distributions (Huang and Williamson 2001). These distributions can be deduced from census tables. Several statistics vary considerably for heterogeneous populations (Burgard and Münnich 2010). Therefore, it is important that this heterogeneity is also reflected in the synthetic dataset.

The use of synthetic data for statistical disclosure control is presented in Reiter and Drechsler (2007), Drechsler et al. (2008a,b), Drechsler and Reiter (2008, 2012), Templ and Alfons (2010), Drechsler (2011), Hundepool et al. (2012), and Templ (2017). An extensive overview of the methods used can be found in Alfons et al.

(2011a) and Kolb (2012). The remaining part of this section refers to these two publications describing briefly the generation process of AMELIA.

Prior to drawing household variables for AMELIA, some initial steps have to be conducted. The population size as well as the number of regions and the region sizes have to be defined. The original EU-SILC dataset is divided into four regions which reflect the main regions in Europe. Regional effects can be introduced by using a regional indicator within the models and by separately drawing households within these regions. Variable outcomes are sampled from a series of conditional distributions which are deduced from EU-SILC data preventing sparse cells to ensure avoiding disclosure limitations. Variables are generated step by step considering only relevant explanatory variables for the conditional distributions. Cross tabulations are derived for categorical variables as well as for categorized metric variables. Afterwards, the categorized metric variables are retransformed to metric variables while respecting for observed marginal distributions.

2.2 Contributions to Open and Reproducible Research

According to Peng (2015), reproducibility is important to create trust in data analysis. This also holds for the comparison and evaluation of statistical methodology. As introduced before, simple access to data especially in social sciences is seldom due to disclosure control legislation and agreements. Besides the availability of data, releasing the program code used to generate the results of a study is essential for the evaluation of scientific research by editors or within the scientific community. The authors of this paper support the use of platforms for program code sharing and consider the AMELIA platform as a relevant contribution to facilitate open and reproducible research.

Stodden (2015) gathered causes why reproducibility especially in the field of statistics might fail. Among reasons like misapplications of tests, she declares the lack of code and data as a source of issues. The AMELIA dataset and the corresponding samples are a remedy in the field of methodological research. Especially, it is a tool to assess the stability of statistical findings under a wide variety of settings. Further, the data use is free of charge. By providing a fixed set of samples under a given sampling design, the randomization of the simulations can be fully controlled and reproduced by third parties.

The generation of new variables invokes the revision of data and meta data. For this purpose, a version control system is in the AMELIA data description (Burgard et al. 2017). Users of AMELIA can provide ideas or code for enhancing the AMELIA dataset, either with variables or sampling designs. The core team will decide on the inclusion of enhancements into the official AMELIA data.

For the next versions of the AMELIA dataset, the following is planned:

- Enhancement with two-stage and unequal probability sampling designs
- Inclusion of first oversampling routines
- Inclusion of major HFCS variables
- First version of longitudinal data

In exchange for a free use of the AMELIA data, we expect that the user cites this article as well as the AMELIA homepage in her or his publication.

All scripts creating the core dataset were implemented in R. In order to support open and reproducible research, the code for new scenario variables will be made available on the AMELIA homepage. Users of the dataset are encouraged to make codes available via e. g. www.runmycode.org which is described in Hurlin et al. (2014).

The AMELIA platform addresses many of the FAIR principles described in Wilkinson et al. (2016). These principles are a guideline for data stewardship and management. The data should be findable (F), accessible (A), interoperable (I) and reusable (R). The dataset is easily accessible via the AMELIA platform (www.amelia.uni-trier.de). The access is independent of the computer software used. Interoperability is guaranteed by providing the data as RData and CSV files. CSV files can be processed by practically any analytical software. Additionally, RData files are provided for R users. Also, a data description (Burgard et al. 2017) is provided on the AMELIA platform. This data description explains the variables of AMELIA and the scheme and use of the sampling designs.

2.3 Properties of the Dataset

The synthetic dataset AMELIA consists of approximately 10 million individuals in 3.7 million households. AMELIA encompasses typical socio-economic variables like age, gender, marital status, activity status and different sources of income. The latter are an important source for poverty measurement within EU-SILC. Four regions with 11 provinces in 40 districts are implemented. Additionally, 1592 cities and communities are provided. These regional levels provide an important information for implementing realistic sampling designs. The sampling designs for collecting households and individuals mainly follow the typical European household survey designs. Additional sophisticated designs will be added, e. g. using oversampling routines. The currently implemented sampling designs are available with sampling fractions of 0.16%, 1% and 5%. For each design and each sample size, 10000 samples are drawn and split into files comprising 100 samples. The sample number, identifier of the drawn households, and probabilities of a household entering the sample are stored. More details on the sample designs is provided in the data description provided on the AMELIA platform (Burgard et al. 2017).

2.4 Community Recognition

Extensive design-based simulation studies have a long tradition. Increasing computing power, however, allowed to introduce more and more sophisticated studies incorporating several sources of error.

Within the DACSEIS project (Data Quality in Complex Surveys within the New European Information Society, Münnich and Wiegert 2001), which was supported under the fifth Framework Programme of the European Commission, variance estimation methods were investigated within several European surveys. The natural basis for this simulation study was a group of datasets containing all the relevant

information of the surveys to realize the according survey designs as well as the estimation methods. One very computer-intensive challenge was the inclusion of missing values and imputation. The output is provided in the DACSEIS recommended practice manual.

The simulation environment was later used as a prototype for the seventh Framework Programme, especially the AMELI project. AMELIA was an important component of the AMELI project. Several design-based simulation studies were conducted on this dataset. Graf et al. (2011b) dealt with the analysis of the two component Dagum distribution and its fitting on the equivalized disposable personal income. Small area estimators are compared in Lehtonen et al. (2011). Bruch et al. (2011) evaluated the performance of different variance estimators based on different sampling designs. All simulation results were presented in Hulliger et al. (2011). Kolb et al. (2011) examined indicators on poverty and social exclusion.

The AMELIA dataset was further developed and used by the research community within the InGRID (Merkle and Münnich 2016). The aim was to evaluate estimators of income poverty and inequality using income classes considering different sampling designs. The AMELIA platform is an outcome of this project and provides a first source for open and reproducible methodological research in social sciences as well as in survey and official statistics. Within the Horizon 2020 research infrastructure, the aim is to enhance the AMELIA data resource to further promote comparable methodological research using EU-SILC and related data.

3 Application Example

Estimation strategies can comprise a mix of data handling steps or processes such as data editing, imputation for missing data, sampling designs, estimation and uncertainty measurements. Whereas each single step is mathematically proved to be sound, the analysis of their interactions is mathematically ambitious and sometimes impossible. Therefore, Monte-Carlo simulations are used for this purpose. As a use case for such a simulation we will explore the estimation of regional parameters by using a small area methods (Rao and Molina 2015) method. These methods typically make use of models for the prediction of population values. Additionally to using the survey data gathered within a region of interest, a model across regions is applied to reduce the variability of predictions on regional level. This is called borrowing strength. If the model is inappropriate small biases on regional level may occur.

One of the most widely used models in this context is the Battese-Harter-Fuller (BHF) model (Battese et al. 1988). Basically, it assumes a random intercept model, where the regions form the grouping variable. For a variable y_{ij} for unit i in region j the random intercept model is specified as:

$$y_{ij} = x'_{ij}\beta + u_j + e_{ij},$$

with x_{ij} being the $p + 1$ vector of the covariates and β the $p + 1$ vector of coefficients. The random effect u_j is constant for all units i in area j. As the

areas are assumed to have a certain deviation from the overall regression model, the prediction conditional on u_j is of interest. The prediction from the model is thus:

$$\widehat{\overline{y}}_j = \overline{x}'_j \widehat{\beta} + \widehat{u}_j.$$

A popular way to evaluate the performance of new estimators is to perform a model-based simulation. That is, under the model assumptions a sample is generated R times and the model is then estimated on each samples. Then, typical measures such as the bias and the root mean squared error (RMSE) are deduced from the distribution of parameter estimates over the R repetitions.

This, however, does not reflect the situation statistical offices and researchers are in, when deciding on the estimator to choose. The sample the face is not a sample from a model-based population, but a sample drawn according to a sampling design. That is, the simulation used to choose an estimator should be design-based as well. The results from both approaches can differ significantly.

The design-based simulation analyses the properties of an estimator under the randomization of a sampling design. Fur this purpose, 10,000 samples are drawn according to a stratified sampling design with proportional allocation or with optimal allocation. As the focus of the example is to compare the model-based and design-based simulation approaches, we try to make the settings as comparable as possible. This is achieved by taking as dependent variable a prediction from a model. The new variable $y_{ij}^{\text{Des}} = x'_{ij}\beta + u_j + e_{ij}, u_j \sim \text{N}(0, \sigma_u^2), e_{i,j} \sim \text{N}(0, \sigma_e^2)$ is generated only once, and thus fixed over the simulation. This approach is denoted by an upper index *Des*. In contrast, for the model-based simulation not the whole population is needed, but only one sample. The first sample of the stratified proportional allocation design is taken and in each simulation run the variable of interest is generated anew. For the *modela* simulation, both the random effect u_j and the individual error term e_{ij} are generated in each simulation step r $y_{ij}^{\text{Moda},r} = x'_{ij}\beta + u_j^r + e_{ij}^r, u_j^r \sim \text{N}(0, \sigma_u^2), e_{i,j}^r \sim \text{N}(0, \sigma_e^2)$. This approach is denoted by an upper index *Moda*. In contrast, *modelb* takes the random intercept to be a fixed population value, and only the individual error is drawn in each simulation run $y_{ij}^{\text{Moda},r} = x'_{ij}\beta + u_j + e_{ij}^r, u_j \sim \text{N}(0, \sigma_u^2), e_{i,j}^r \sim \text{N}(0, \sigma_e^2)$.

A typical issue when dealing with survey data in social sciences is that not all relevant variables can be observed. We mimic this situation by estimating the models twice. The results named with a trailing 1 are estimated with the true model, the one that generated the data. Those, with a trailing 2, are lacking three variables in the model.

The setting *modela1* is the classical model-based simulation, where the dependent variable is generated fully under the model assumptions. As can be seen in Fig. 1, it shows to have the lowest bias over all areas and competing settings. When omitting three variables, in setting *modela2*, some areas show to have an increase in bias. When comparing these results with the fixed random intercept u_j, the effect of the violation of the model assumption $E(u_j) = 0, \forall j = 1, ...$ on the bias can be seen. Many areas show to have at least small biases, in both cases, with the full model (*modelb1*) and with the reduced one (*modelb2*). The sampling design of

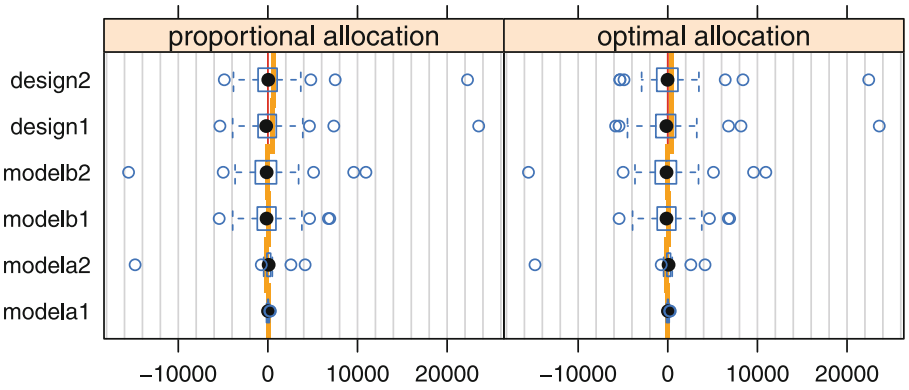

Fig. 1 Bias of the BHF estimate of EDI on district-level for the model-based and design-based simulation

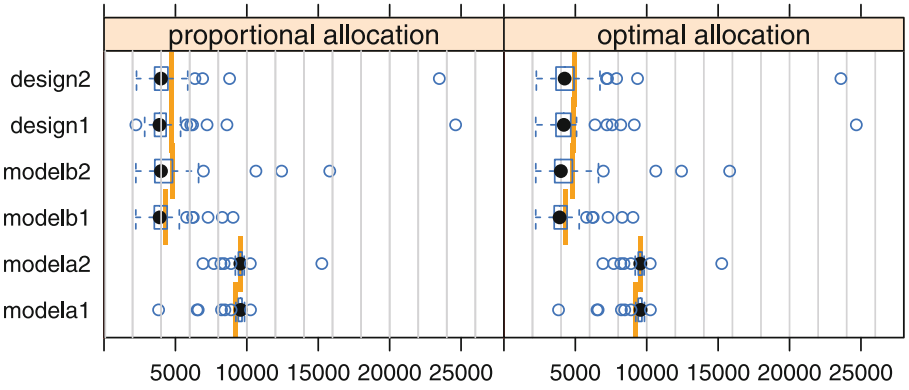

Fig. 2 Root means squared error (RMSE) of the BHF estimate of EDI on district-level for the model-based and design-based simulation

course has no effect on the model-based simulations as they do not make use of the randomization of the survey design.

When comparing the design-based simulations with the *modelb* simulation, the results are almost alike. This is due to the fact, that both use the same data-generating process. However, when introducing the optimal allocation we can observe an increase in RMSE (Fig. 2). The optimal allocation is optimal for the estimation of the population mean or total. But for the estimation of regional figures it turns out to be counter-productive. This effect is not observable when using the model-based simulation as it does not take the sampling design into account. In surveys conducted for the social sciences, often, due to budgetary constrains, complex sampling schemes are applied. In order to evaluate the impact of such sampling design applied to estimators at hand, such design-based simulations provide important information on practical use of these estimators. This is especially important in the context of official statistics. Therefore, for open, reproducible and comparable research, it is of utmost importance to have common simulation frameworks as benchmark vehicles.

4 Conclusion

The aim of the AMELIA data source is to provide a platform for comparable, open, and reproducible research in social sciences and social statistics. The data are already in use by several research projects and enable methodological advances in survey statistics using design-based methods, investigating statistical methodologies under complex survey designs including oversampling, comparing estimation strategies under different practical settings, e. g. in the presence of non-response. Further, it allows examining microsimulation methods based on single samples. Moreover, it serves as a transnational access source within the research infrastructure InGRID2 (www.inclusivegrowth.eu), and, hence, supports the research community in social sciences. In the future, the platform will be enhanced by further variables and survey designs as well as a longitudinal component.

Acknowledgements The authors thank the European Commission for financially supporting the InGRID2 research infrastructure under Horizon 2020 which allows us to further develop the AMELIA dataset fostering open and reproducible research. The first version of the AMELIA dataset was produced within the FP7-SSH-2007-1-217322-AMELI project. It was further developed within the FP7-INFRASTRUCTURES-2012-1-312691 InGRID project. Further, we thank Florian Volk for carefully editing the original AMELIA dataset, Simon Lenau for enhancing the sample drawing routines as well as Florian Ertz for his contributions the data description (Burgard et al. 2017). The current version of the dataset is based on the research in Alfons et al. (2011a), Chapt. 4, as well as Merkle and Münnich (2016), Chapt. 1. Finally, we thank the editors and two anonymous reviewers for providing helpful comments to improve this paper.

References

Alfons A, Filzmoser P, Hulliger B, Kolb J-P, Kraft S, Münnich R, Templ M (2011a) Synthetic data generation of SILC data. Research project report WP6 – D6.2, FP7-SSH-2007-217322 AMELI. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP6-D6.2-240611.pdf. Accessed December 7, 2017

Alfons A, Kraft S, Templ M, Filzmoser P (2011b) Simulation of close-to-reality population data for household surveys with application to EU-SILC. Stat Methods Appl 20(3):383–407

Battese G, Harter R, Fuller W (1988) An error-components model for prediction of county crop areas using survey and satellite data. J Am Stat Assoc 83:28–36

Bruch C, Münnich R, Zins S (2011) Variance estimation for complex surveys. Research project report WP3 – D3.1, FP7-SSH-2007-217322 AMELI. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP3-D3.1-20110514.pdf. Accessed December 7, 2017

Burgard J, Ertz F, Merkle H, Münnich R (2017) AMELIA - data description v0.2.2.1. Trier University. http://amelia.uni-trier.de/wp-content/uploads/2017/11/AMELIA_Data_Description_v0.2.2.1.pdf. Accessed December 7, 2017

Burgard J, Münnich R (2010) Modelling over- and undercounts for design-based Monte Carlo studies in small area estimation: an application to the German register-assisted census. Comput Stat Data Anal 56(10):2856–2863

Drechsler J (2011) Synthetic datasets for statistical disclosure control: theory and implementation. Lecture notes in statistics, vol. 201. Springer, New York

Drechsler J, Bender S, Rässler S (2008a) Comparing fully and partially synthetic datasets for statistical disclosure control in the german IAB establishment panel. Trans Data Priv 1(3):105–130

Drechsler J, Dundler A, Bender S, Rässler S, Zwick T (2008b) A new approach for disclosure control in the IAB establishment panel – multiple imputation for a better data access. AStA Adv Stat Anal 92(4):439–458

Drechsler J, Reiter J (2008) Disclosure risk and data utility for partially synthetic data: an empirical study using the German IAB establishment survey. Technical report. Institute for Employment Research, Regensburg

Drechsler J, Reiter JP (2012) Combining synthetic data with subsampling to create public use microdata files for large scale surveys. Surv Methodol 38:73–79

Fienberg SE (1994) Conflicts between the needs for access to statistical information and demands for confidentiality. J Off Stat 10(2):115–132

Graf M, Alfons A, Bruch C, Filzmoser P, Hulliger B, Lehtonen R, Meindl B, Münnich R, Schoch T, Templ M, Valaste M, Wenger A, Zins S (2011a) State-of-the-art of laeken indicators. Research project report WP1 – D1.1, FP7-SSH-2007-217322 AMELI. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP1-D1.1-20110418.pdf. Accessed December 7, 2017

Graf M, Nedyalkova D, Münnich R, Seger J, Zins S (2011b) Parametric estimation of income distributions and indicators of poverty and social exclusion. Research project report WP2 – D2.1, FP7-SSH-2007-217322 AMELI. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP2-D2.1-20110409.pdf. Accessed December 7, 2017

Eurosystem Household Finance and Consumption Network (2013) The eurosystem household finance and consumption survey – results from the first wave. Technical report. European Central Bank, Frankfurt am Main

Huang Z, Williamson P (2001) A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata. Population Microdata Unit, Department of Geography, University of Liverpool, Liverpool

Hulliger B, Alfons A, Bruch C, Filzmoser P, Graf M, Kolb J-P, Lehtonen R, Lussmann D, Meraner A, Münnich R, Nedyalkova D, Schoch T, Templ M, Valaste M, Veijanen A, Zins S (2011) Report on the simulation results. Research project report WP7 – D7.1, FP7-SSH-2007-217322 AMELI. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP7-D71.pdf. Accessed December 7, 2017

Hundepool A, Domingo-Ferrer J, Franconi L, Giessing S, Nordholt ES, Spicer K, De Wolf P-P (2012) Statistical disclosure control. John Wiley & Sons, Hoboken

Hurlin C, Pérignon C, Stodden V (2014) Runmycode.org: a research-reproducibility tool for computational sciences. In: Implementing reproducible research. CRC Press, Boca Raton, pp 367–381

Kalton G (2002) Models in the practice of survey sampling (revisited). J Off Stat 18(2):129–154

Kolb, J.-P. (2012). Methoden zur Erzeugung synthetischer Simulationsgesamtheiten. PhD thesis, Universität Trier.

Kolb J-P, Münnich R, Beil S, Chatziparadeisis A, Seger J (2011) Policy use of indicators on poverty and social exclusion. Research project report WP9 – D9.1, FP7-SSH-2007-217322 AMELI. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP9-D9.1-20110331.pdf. Accessed December 7, 2017

Lehtonen R, Veijanen A, Myrskylä M, Valaste M (2011) Small area estimation of indicators on poverty and social exclusion. Research project report WP2 – D2.2, FP7-SSH-2007-217322 AMELI. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Ameli_Delivrables/AMELI-WP2-D2.2-20110402.pdf. Accessed December 7, 2017

Lovelace R, Dumont M (2016) Spatial Microsimulation with R. CRC Press, Boca Raton

Merkle H, Münnich R (2016) The Amelia dataset - a synthetic universe for reproducible research. In: Berger YG, Burgard JP, Byrne A, Cernat A, Giusti C, Koksel P, Lenau S, Marchetti S, Merkle H, Münnich R, Permanyer I, Pratesi M, Salvati N, Shlomo N, Smith D, Tzavidis N (eds) InGRID deliverable 23.1: case studies, pp WP23–D23 (http://inclusivegrowth.be)

Münnich R, Schürle J (2003) On the simulation of complex universes in the case of applying the German Microcensus. DACSEIS research paper series No.4. https://publikationen.uni-tuebingen.de/xmlui/bitstream/handle/10900/47281/pdf/DRPS4.pdf?sequence=1&isAllowed=y. Accessed December 7, 2017

Münnich R, Schürle J, Bihler W, Boonstra H-J, Knotterus P, Nieuwenbroek N, Haslinger A, Laaksonen S, Wiegert R, Eckmair D, Quatember A, Wagner H, Renfer J-P, Oetliker U (2003) Monte carlo simulation study of European surveys - DACSEIS deliverables 3.1 and 3.2. Technical report, University of Tübingen. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/Dacseis_Deliverables/DACSEIS-D3-1-D3-2.pdf. Accessed December 7, 2017

Münnich R, Wiegert R (2001) The DACSEIS project. DACSEIS research paper series 1. https://www.uni-trier.de/fileadmin/fb4/projekte/SurveyStatisticsNet/DRPS1.pdf. Accessed December 7, 2017

Peng R (2015) The reproducibility crisis in science: a statistical counterattack. Significance 12(3):30–32

Pratesi M (2016) Analysis of poverty data by small area estimation. John Wiley & Sons, Hoboken

Rahman A, Harding A (2016) Small Area Estimation and Microsimulation Modeling. CRC Press, Boca Raton

Rao J, Molina I (2015) Small area estimation. Wiley, New York

Reiter JP, Drechsler J (2007) Releasing multiply-imputed synthetic data generated in two stages to protect confidentiality. IAB Discussion Paper 200720. Institut für Arbeitsmarkt und Berufsforschung (IAB), Nürnberg

Stodden V (2015) Reproducing statistical results. Annu Rev Stat Appl 2:1–19

Templ M (2017) Statistical disclosure control for microdata: methods and applications in R. Springer, Cham

Templ M, Alfons A (2010) Disclosure risk of synthetic population data with application in the case of EU-SILC. Springer, Berlin, Heidelberg, pp 174–186

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The fair guiding principles for scientific data management and stewardship. Sci Data 3:160018

de Wolf P-P (2015) Public use files of EU-SILC and EU-LFS data. Technical report, Joint UN-ECE/Eurostat work session on statistical data confidentiality in Helsinki, European Comission/Eurostat, Luxemburg. https://ec.europa.eu/eurostat/cros/system/files/d4.1presentationhelsinki.pdf. Accessed December 7, 2017