Small area estimation for city statistics and other functional geographies

RALF MÜNNICH, JAN PABLO BURGARD,
FLORIAN ERTZ, SIMON LENAU,
JULIA MANECKE, HARIOLF MERKLE

2019 edition

STATISTICAL
WORKING PAPERS

eurostat

Small area estimation strategies for city statistics and other functional geographies

RALF MÜNNICH, JAN PABLO BURGARD,
FLORIAN ERTZ, SIMON LENAU,
JULIA MANECKE, HARIOLF MERKLE

2019 edition

Manuscript completed in October 2019.

Printed by the Publications Office in Luxembourg.

Luxembourg: Publications Office of the European Union, 2019

Print: ISBN 978-92-76-11707-0		doi:10.2785/828660	KS-TC-19-006-EN-C
PDF: ISBN 978-92-76-11706-3	2315-0807	doi:10.2785/898627	KS-TC-19-006-EN-N

Abstract

The aim of the investigation is to examine the extent to which so-called small area estimation methods are able to increase the precision of estimations of statistical indicators at city level using common European social surveys. In this context, the focus is on the performance of different estimation strategies against the background of various sampling designs commonly used by national statistical institutes for social surveys in Europe.

As no census information on the variables of interest is available, the true distribution of said estimators of interest cannot be investigated and evaluated. Therefore, a design-based Monte Carlo simulation study using a synthetic but close-to-reality population is performed to examine the research question at hand. The repetitive drawing of samples using common sampling designs facilitates the comparison of various design-based, model-assisted, model-based and synthetic estimation approaches in a realistic environment mirroring the main properties of the population of the European Union.

In the course of the simulation study, after repetitively drawing samples according to the examined sampling designs, the estimation strategies are applied to estimate the at risk of poverty or social exclusion (AROPE) rate per area as target parameter. The model-assisted and model-based estimation approaches both incorporate auxiliary information, which is assumed to be available in practice.

The results show that, despite the introduction of a slight bias, all of the investigated estimation approaches are able to increase the efficiency of the estimation at municipality and city level. However, the improvement of the estimation quality critically depends on the underlying sampling design. Furthermore, it can be observed that the potential for improvement by using small area estimation approaches instead of the classical design-weighted estimation techniques decreases with an increasing area-specific sample size. Therefore, the choice of the estimation strategy and whether to apply small area estimation approaches depends on the underlying sampling design of the survey to be used as well as on the sample size of the areas of interest.

Keywords: Small area estimation, variance estimation, city statistics, AROPE, simulation study

Authors:

Economic and Social Statistics Department Trier University, Faculty IV, Economics

Ralf Münnich, Jan Pablo Burgard, Florian Ertz, Simon Lenau, Julia Manecke, Hariolf Merkle.

Acknowledgement:

The authors would like to thank Bernhard Stefan Zins for providing his expertise in variance estimation for the AROPE indicator and the respective estimation routines. Further, we are grateful to the contract manager Britta Gauckler, Valeriya Angelova Tosheva and Gian Luigi Mazzi for providing important input to the project. Finally, we thank Jan Weymeirsch for assisting in the finalisation of the article and the project management team at GOPA for facilitating administrative tasks.

Contents

Figures

Tables

Abbreviations

AES	Adult Education Survey
AIK	Aikake Information Criterion
AMELI	Advanced Methodology for European Laeken Indicators
AROPE	At Risk Of Poverty or social Exclusion
ARPR	At Risk Of Poverty Rate
BHF	Battese-Harter-Fuller
BLUP	Best Linear Unbiased Predictor
DOU	Degree of Urbanization
EBLUP	Empirical Best Linear Unbiased Predictor
FUAs	Functional Urban Areas
GOPA	Gesellschaft für Organisation, Planung und Ausbildung mbH
GREG	Generalised Regression
ICT	Information and Communication Technology
InGRID	Inclusive Growth Research Infrastructure Diffusion
ISCED	International Standard Classification of Education
LFS	Labour Force Survey
LWI	Low Work Intensity
NSI	National Statistics Institute
NUTS	*Nomenclature des Unités Territoriales Statistiques*
PSU	Primary Sampling Unit
REML	Restricted Maximum Likelihood
SAE	Small Area Estimation
SILC	Survey on Income and Living Conditions
SMD	Severly Materially Deprivated
SSU	Second stage Sampling Units

1 The need for information on small spatial units

The demand for reliable information on the level of small spatial units, specifically cities and functional urban areas (FUAs), has increased significantly in recent years. An overview of European cities and communities from 2016 can be drawn from Figure 9 in Appendix B (see also https://ec.europa.eu/eurostat/statistics-explained/pdfscache/72650.pdf). To this end, Eurostat has set up a city data collection containing data on various variables from registers, censuses, and sample surveys. These data were collected at the level of the cities and their respective functional urban areas. However, cities and functional urban areas are usually not incorporated in the sampling design of social surveys. This implies considerable challenges for the estimation of unknown parameters as the relevant areas (spatial units, which are the focus here) or domains (groups built by certain characteristics) might have unplanned and small sample sizes. Using only data from observations sampled from the area in question and weighting them when computing an estimate, i.e. using a so-called direct design-weighted estimation method (such as the well-known estimator by Horvitz and Thompson (1952)), will lead to unbiased estimates. However, unplanned and small sample sizes due to the disregard of cities and functional urban areas at the design stage of the sample survey might result in imprecise estimates with large standard errors. In a given application, it might even be the case that some areas of interest may not have been sampled at all. So-called small area estimation methods may be used to improve the quality of such estimates. These mostly model-based approaches incorporate additional auxiliary information from further areas by means of a previously defined model. This enables an increase in precision of the estimates and even the estimation for areas which have not been sampled at all.

The aim of the project *Small Area Estimation (SAE) for city statistics and other functional areas part II* was to investigate how small area estimation methods might be used in the context of sample designs used in common European social surveys like the *European Union Statistics on Income and Living Conditions (EU-SILC)* to estimate statistical indicators like the at risk of poverty or social exclusion rate (AROPE). As opposed to databases available at *Eurostat*, national statistical institutes (NSIs) have access to spatial identifiers, like LAU-2 codes, and a wealth of auxiliary information and could therefore employ small area estimation methods in a decentralised fashion. Here, the performance of different estimation strategies in combination with different sampling designs will be investigated.

A major hurdle for the investigation of any estimation method using sample survey data is that there is hardly ever census information on the variables of interest available. Accordingly, the true distribution of the related estimators (i.e. the distribution of the computed point estimates resulting from drawing all possible samples from the population) cannot be known and the estimator's properties not investigated as a consequence. In survey statistics, this problem is typically overcome by using a design-based Monte Carlo simulation study. The starting point is a large *synthetic* but close-to-reality population, i.e. a population of vectors of data points that share certain traits of the real population in question. To give an example, if there is a positive correlation between the number of members in a household and the overall household income found in empirical (real) data, such a positive correlation is constructed for the synthetic (non-real) data as well. Once such a population has been built, samples can be repeatedly drawn from it using sampling designs that are quite alike those really used in sample surveys like EU-SILC. Given a large enough number of drawn samples, we can assume that the distribution of point estimates is *reasonably* close to the estimator's real distribution. We are then in a position to compare the performance of different estimators (e.g. the performance of a Horvitz-Thompson type estimator to the performance of a small area estimator). For this project, we use such a design-based simulation study. The synthetic population we use here is called *AMELIA*. Together with the sampling designs under consideration and the specific estimators we compare, it will be described in Section 2.

In Section 3 we will present the main findings of our design-based simulation study. As the whole simulation setup is rather extensive, we focus on some crucial points that are of interest to the practitioner at an NSI.

Based on these major findings, we discuss the implications for a practical implementation of the methods investigated in Section 4.

The authors would like to address an important caveat right at this point. In the greater scheme of things, small area estimation methods are relatively young and a fruitful area of statistical research. A meaningful application of these complex methods necessitates a level of statistical knowledge of the user that is well above that provided by, say, basic courses in descriptive and inferential statistics in typical economics programmes at universities. Therefore, it is impossible to derive a manual including hard-and-fast rules like the following: *Given situation X and auxiliary variable Y, use estimator Z*. We will point out situations in which small area estimation methods may lead to an improvement of point estimates. However, given the very diverse survey and data situations in Europe (which are themselves subject to changes over time), statisticians at NSIs should be well-trained in order to apply such methods. The actual institutional framework prevailing in the respective member state has a considerable impact as well and should therefore be accounted for. Otherwise, there is a considerable risk of reaching less than desirable outcomes. In this light, this paper tries to shed some light on potential ways to harness small area estimation methods for the estimation of city statistics.

2 A framework for the investigation of different options

2.1 The synthetic population AMELIA

As already mentioned in Section 1, the close-to-reality synthetic population AMELIA, which was created to perform design-based Monte Carlo simulation studies in survey statistics, is used as the starting point of our investigation (Burgard et al., 2017b). AMELIA was originally created within the research project Advanced Methodology for European Laeken Indicators (AMELI) which was funded by the European Commission within its Seventh Framework Programme (Alfons et al., 2011). In its generation, mimicking the main properties of the population of the European Union was a main objective. These properties were found in EU-SILC data. After the AMELI project, the AMELIA dataset has been published on the AMELIA platform (see www.amelia.uni-trier.de) which is an outcome of the project Inclusive Growth Research Infrastructure Diffusion (InGRID), to foster open and reproducible research in survey statistics (Merkle and Münnich, 2016). Not only the synthetic population itself has been published and made freely available. The population is accompanied by already drawn samples with various underlying sampling designs. A detailed data description is provided on the AMELIA platform (Burgard et al., 2017a). The main properties of AMELIA are listed below:

- 10,012,600 individuals
- 3,781,289 households
- Regional structure
  - 4 regions (NUTS 1)
  - 11 provinces (NUTS 2)
  - 40 districts (NUTS 3)
  - 1,592 municipalities (LAU)
- Variables on household and individual level

Since AMELIA is based on EU-SILC, important poverty-related issues are already covered, i.e. variables that are necessary to calculate the at-risk-of-poverty rate (ARPR) consistent with EU-SILC. The ARPR is the share of people living in a household that has less than 60% of the national median equivalised disposable income available (c.f. Eurostat, 2018b). For the simulation study of this project, however, new variables had to be generated. These variables are necessary to calculate the at risk of poverty or social exclusion rate (AROPE, c.f. Eurostat, 2018a) which is a composite indicator comprising three subindicators and the chosen target parameter in our simulation study. A person is only counted once no matter how many of the subindicators apply. One of the subindicators is the `ARPR`, the other two cover the topics material deprivation and work intensity. All persons in a given household are considered to be severly materially deprivated (`SMD`, c.f. Eurostat, 2018c) if this household cannot afford at least four of nine items (c.f. Eurostat, 2010), which had to be generated for this project.

A person lives in a household with low work intensity (`LWI`, c.f. Eurostat, 2018d) if the household members of working age worked less than 20% of their potential within the last twelve months. All persons between 18-59 are considered as working-age persons with a few exceptions. Students between 18 and 24 are excluded from this group as well as households composed only of children, students under 25, or people aged 60 or above. These are not taken into account at all. Six new variables giving the number of months a person spent in different working situations (like full-time employment, part-time employment, unemployment, etc.) were generated in AMELIA consistent with the EU-SILC definitions (c.f. Eurostat, 2010). The generation of these variables was rather involved and included a discretisation of the original variables, latent class analysis, multinomial logistic regression models and random draws from outcomes. The reader interested in the details of this procedure is referred to Deliverable 1 of this project.

2.2 Construction of large cities in AMELIA

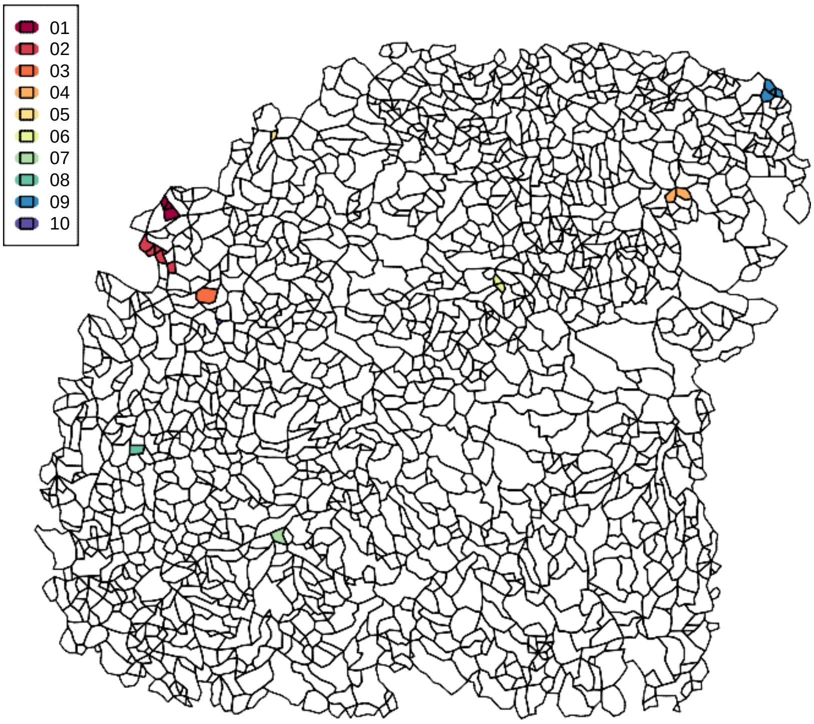
Since the specific aim of this research project was an investigation of the potential gains of employing small area estimation methods to estimate certain indicators on the level of cities and functional urban areas, the basic synthetic population dataset, as described in the previous subsection, had to be suitably amended. AMELIA consists of 1,592 municipalities (variable `CIT`) of varying degrees of urbanisation (variable `DOU`) and varying household- and individual-level population sizes (the average number of households and individuals per municipality being 2,375 and 6,289, respectively). At this point it is worthwhile to remember that the aim of a design-based Monte Carlo simulation study is not a one-to-one *reproduction* of a real population and its (spatial) structures. The key point is to *mimic* some important characteristics of the data. Therefore, the absolute size of municipalities and cities is not important but rather their relative sizes. New *synthetic* large cities had to be integrated into AMELIA.

The first starting point for this extension is the degree of urbanisation of municipalities that could either be thinly-populated, have an *intermediate population density*, or be *densely-populated*. A given municipality could only be part of a large city if it belongs to the group of densely-populated municipalities.

Within this pre-selection of municipalities (covering approximately one third of the overall household population and one fifth of the municipalities), first the ten municipalities with the largest population of individuals are chosen to be the cores of the new large cities to be constructed.

As an assumption made to facilitate the further process, the non-core areas of the large cities have to belong to the same higher-level spatial unit (i.e. one of the 40 districts, variable DIS) as their respective cores. Additionally, all municipalities forming one large city have to be connected spatially. Following this algorithm, we created ten large cities within AMELIA. These are labelled in descending order of population size and are shown in Figure 1. Further details are given in Table 1, where N_{HH} and N_{IND} are the household and individual population, respectively.

Figure 1: Large cities



Source: Ertz, F. (2020): Regression Modelling with Complex Survey Data: An Investigation Using an Extended Close-to-Reality Simulated Household Population. Ph.D. dissertation. Trier University. To be published.

2.3 Archetypical sampling designs

Different sampling designs were implemented for the simulation in order to mimic the various national sampling schemes of European social surveys. In preparation, the publicly available information for the European Statistics on Income and Living Conditions (EU-SILC)[1], EU labour force

[1] cf. https://circabc.europa.eu/w/browse/7af111b3-b700-4321-9902-695082dcb7e1

survey (LFS)[2], Adult Education Survey (AES)[3] and the Survey on information and communication technology usage (ICT)[4] were scanned. The first two surveys mentioned are the predominant data sources for the estimation of indicators of income and social exclusion. From this information, certain characteristic sampling schemes used for the surveys throughout the EU could be identified. For details on this, the reader is referred to Deliverable 1 of this project.

Table 1: Large cities

City number	CIT(s)	N_{HH}	N_{IND}	REG
1	322, 323, 326	116 218	287 882	1
2	311, 305, 306, 309, 310, 312	97 027	240 432	1
3	292	33 177	81 831	1
4	1372, 1369	6 872	20 065	4
5	1088	4 323	11 816	3
6	1250, 1255	8 072	23 266	4
7	400	4 584	11 787	2
8	189	4 660	11 693	1
9	1532, 1523, 1530, 1546	12 666	36 704	4
10	278	4 657	11 678	1

Source: Ertz, F. (2020): Regression Modelling with Complex Survey Data: An Investigation Using an Extended Close-to-Reality Simulated Household Population. Ph.D. dissertation. Trier University. To be published.

For example, a vast majority of countries apply stratification by regional, population size and/or degree of urbanization (DOU) variables for the mentioned surveys, but others are used as well. Sampling units most often include households (or related concepts), but in many cases, two-stage sampling is applied such that households are the second stage sampling units (SSUs), while larger regional aggregates are sampled as primary sampling units (PSUs) at the first stage.

Table 2: Sampling designs for the Monte Carlo simulation study

	Stage 1			Stage 2	
	PSU	Strata	fr_1 (%)	SSU	fr_2 (%)
SRS_H	HID	–	0.16	–	
SRS_P	PID	–	0.16	–	
STSI_H1	HID	PROV	0.16	–	

[2] cf. https://ec.europa.eu/eurostat/documents/7870049/8699580/KS-TF-18-002-EN-N. pdf/ce2e7a97-6b8c-44b8-8603-3a4606e5b335
[3] cf. https://ec.europa.eu/eurostat/statistics-explained/index.php/Adult_Education_Survey_(AES)_methodology#Quality_reports
[4] cf. https://circabc.europa.eu/w/browse/8b3c3278-b860-4d53-8ea3-a4f9f33f74fe and https://circabc.europa.eu/w/browse/bdcfc229-16b0-496d-8ade-c0498c28470f

	Stage 1			Stage 2	
	PSU	**Strata**	fr_1 (%)	**SSU**	fr_2 (%)
STSI_H2	HID	DOU	0.16	–	
STSI_H3	HID	PROV × INC C	0.16	–	
STSI_H4	HID	DIST × DOU	0.16	–	
STSI_P	PID	AGE C	0.16	–	
TS_H1	CIT	PROV	16.00	HID	1
TS_H2	CIT	PROV × DOU	16.00	HID	1
TS_H3	CITG	PROV	16.00	HID	1
TS_P1	CIT	PROV × DOU	16.00	PID	1
TS_P2	DIST	PROV	16.00	PID	1

AGE C: Age class **CIT:** City **CITG:** Group of cities
DIST: District **DOU:** Degree of urbanization **HID:** Household ID
INC C: Income class **PID:** Person ID **PROV:** Province
PSU: Primary sampling unit fr_1**:** Sampling fraction of PSUs
SSU: Secondary sampling unit fr_2**:** Sampling fraction of SSUs within sampled PSUs

Sources: see section 5 - References

Based on this information, typical sampling designs covering the range of realistic scenarios were constructed for the simulation study. Table 2 provides an overview of the twelve archetypical sampling designs used as the basis for the comparative simulation study.

3 Estimation strategies

In the framework of our simulation study, the application of small area estimation methods will be analysed using the share of persons living at risk of poverty or social exclusion (AROPE) per area as target parameter. This section introduces the investigated estimation strategies and their implementation as well as potential adaptations in order to estimate the target parameter given the data situation mirrored in the simulation study.

3.1 Design-based estimation

A common method of direct design-weighted estimation is the estimator by Horvitz and Thompson (1952). Let y_k be the variable of interest of unit k and let π_k be the corresponding inclusion probability. The design weight, w_k, is the inverse of the units' inclusion probability. In addition, S_d is the set of sampled units belonging to area d (while U_d is the set of all N_d units in area d. For each area with the running index $d = 1, \dots, D$, the total value $\tau_d = \sum_{k \in U_d} y_k$ is to be estimated. The Horvitz-Thompson estimator is an unbiased estimation function for τ_d and is given by

$$\hat{\tau}_d^{HT} = \sum_{k \in S_d} \frac{y_k}{\pi_k} = \sum_{k \in S_d} w_k y_k \tag{1}$$

Thus, the weighted values of the sampled units are summed up. Since this estimator only uses information from the area of interest, the estimation procedure is also referred to as direct estimation.

When the area-specific mean $\mu_d = \frac{1}{N_d} \sum_{k \in U_d} y_k$ is of interest and the area-specific size N_d is known, an unbiased estimator of μ_d is

$$\hat{\mu}_d^{HT} = \frac{\hat{\tau}_d^{HT}}{N_d} = \frac{1}{N_d} \sum_{k \in S_d} \frac{y_k}{\pi_k} \tag{2}$$

In the present application, the estimation of proportions is of interest. Proportions are a special case of the mean, if the variable of interest y_k is dichotomous. This implies that $y_k = 1 =$ if the *k*th unit has a characteristic of interest, i.e. is living at risk of poverty or social exclusion in this application, and $y_k = 0$ if the *k*th unit does not have this characteristic (Lohr, 2009, p. 30). If the estimator (2) is used for the estimation of proportions, it cannot be ruled out that $\hat{\mu}^{HT} > 1$. Respective estimates might be corrected downwards to the value 1, whereby however the estimation is no longer unbiased.

Alternatively, both τ_d and N_d might be estimated and used to estimate μ_d. The estimator is also called the weighted sample mean and is given by

$$\mu_d^{HT_w} = \frac{\hat{\tau}_d^{HT}}{\hat{N}_d} = \frac{\sum_{k \in S_d} y_k/\pi_k}{\sum_{k \in S_d} 1/\pi_k} \tag{3}$$

(SÄRNDAL et al., 1992, p. 182). Both approaches will be compared in the simulation study.

3.2 Model-assisted estimation

Population registers containing information at the level of households and even persons are an extensive source of auxiliary information. These are highly suitable for the stabilisation of estimation.

The generalised regression (GREG) estimator is a so-called model-assisted estimation approach. Its purpose is to reduce the design variance of the estimator by using a model that describes the relationship between the variable of interest y_k and the auxiliary variables x_k. The combination with a classical design-based estimator, such as the unbiased Horvitz-Thompson estimator, preserves the property of a low design bias. This asymptotic unbiasedness is given even if the model is misspecified (see Särndal et al., 1992, p. 227).

The GREG estimator for the total of the variable y_k in area d is given by

$$\hat{\tau}_d^{GREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in S_d} w_k (y_k - \hat{y}_k) \tag{4}$$

(cf. Lehtonen and Veijanen, 2009, p. 229). Here, $\hat{y}_k$ is the estimated variable of interest for each unit k. The first part of the GREG estimator shown in (4) is the sum of the variables of interest predicted from the model $\hat{y}_k$ over all units belonging to area d. Although this synthetic estimation component usually has a low variance due to the underlying model, a bias cannot be avoided. However, this bias is corrected by the so-called bias correction term, i.e. by the weighted sum of the residuals from the sample. Thus, the GREG estimator is asymptotically design-unbiased.

A further modification of the GREG estimator is implemented in the simulation study, which additionally includes the area size N_d. It has a smaller variance than estimator (4) and is given by

$$\hat{\tau}_d^{GREG\_N} = \sum_{k \in U_d} \hat{y}_k + (N_d/\hat{N}_d) \sum_{k \in S_d} w_k (y_k - \hat{y}_k) \tag{5}$$

(cf. Lehtonen and Veijanen, 2009, p. 234).

In the simulation study, the model applied to determine the relation between y_k and the auxiliary variables x_k depends on the respective sampling design. If persons are the final sampling units, the variable of interest is dichotomous, i.e. the person is either living at risk of poverty or social exclusion or not. Correspondingly a probit model is used within the GREG estimator. In case households have been sampled as final sampling units, the variable of interest is the number of persons living at risk of poverty in the respective household. In this instance, the relation between the target variable and the auxiliary variables is modelled using a poisson model.

3.3 Model-based estimation

3.3.1 Fay-Herriot estimator

The area-level estimator according to Fay and Herriot (1979) is using certain auxiliary information that have been aggregated for the area of interest. Therefore, the model is especially applied in cases where the availability of data on micro level is limited. The area-level model can be divided into two parts: the sampling model and the linking model (see Jiang and Lahiri, 2006, p. 6). The sampling model for each of the D areas of interest with index $d = 1, \ldots, D$, is given by

$$\hat{\mu}_d^{Dir} = \mu_d + e_d \tag{6}$$

with a direct estimator $\hat{\mu}_d^{Dir}$. It is assumed that the sampling errors e_d are independent and $e_d \sim N(0, \psi_d)$ with the sampling variance ψ_d. Therefore, it is supposed that $\hat{\mu}_d^{Dir}$ is a design-unbiased estimator for μ_d.

In the context of the linking model, the assumption of a linear relation between the parameter to be estimated, μ_d, and true area-specific auxiliary variables is made. Hence,

$$\mu_d = \overline{\mathrm{X}}_d^T \beta + v_d \tag{7}$$

applies with $v_d \sim N(0, \sigma_v^2)$. Here, $\overline{\mathrm{X}}_d$ designates the population average of the used auxiliary variables in area d. The random effect v_d incorporates variations between the areas that cannot be explained by the fixed effect of the regression term. The variance of the random effects σ_v^2 is also called model variance as it measures the variance between the areas, which cannot be explained by the fixed component of the model. $\overline{\mathrm{X}}_d^T \beta$ is the regression term with the vector of regression coefficients β, which measures the fixed effects over all areas. This is the relationship between the variable to be explained and the auxiliary information. In combination, the sampling model and the linking model result in the linear mixed model

$$\hat{\mu}_d^{Dir} = \overline{\mathrm{X}}_d^T \beta + v_d + e_d \tag{8}$$

with $v_d \overset{iid}{\sim} (0, \sigma_v^2)$ and $e_d \overset{ind}{\sim} (0, \psi_d)$

as a basis for the Fay-Herriot estimator. Here, the direct estimator, which has been built on the basis of a sample, forms the dependent variable. By assuming that the model variance σ_v^2 is known, the best linear unbiased predictor (BLUP) is given by

$$\hat{\mu}_d^{FH} = \overline{\mathrm{X}}_d^T \hat{\beta} + \hat{v}_d \tag{9}$$

with $\hat{v}_d = \gamma_d (\hat{\mu}_d^{Dir} - \overline{\mathrm{X}}_d^T \hat{\beta})$

and $\gamma_d = \frac{\sigma_v^2}{(\psi_d + \sigma_v^2)}$

(see Rao and Molina, 2015, p. 124). As the so-called shrinkage factor γ_d measures the relation between the model variance σ_v^2 and the total variance $\psi_d + \sigma_v^2$, it might be considered as the uncertainty of the model with respect to the estimation of the area-specific mean values $\hat{\mu}_d$. The vector of regression coefficients β is estimated by the weighted least squares method and is given by

$$\hat{\beta} = \left(\sum_{d=1}^{D} \frac{\overline{\mathrm{X}}_d \overline{\mathrm{X}}_d^T}{(\psi_d + \sigma_v^2)} \right)^{-1} \left(\sum_{d=1}^{D} \frac{\overline{\mathrm{X}}_d \hat{\mu}_d^{Dir}}{(\psi_d + \sigma_v^2)} \right) \tag{10}$$

By plugging into $\hat{v}_d = \gamma_d (\hat{\mu}_d^{Dir} - \overline{\mathrm{X}}_d^T \hat{\beta})$ into $\hat{\mu}_d^{FH} = \overline{\mathrm{X}}_d^T \hat{\beta} + \hat{v}_d$ the BLUP might be transformed as follows:

$$\hat{\mu}_d^{FH} = \gamma_d \hat{\mu}_d^{Dir} + (1 - \gamma_d)\overline{\mathrm{X}}_d^T \hat{\beta} \tag{11}$$

As a result of the transformation, it is visible the model-based estimator according to Fay and Herriot (1979) is a weighted average of the direct estimator $\hat{\mu}_d^{Dir}$ and the regression-synthetic estimator $\overline{\mathrm{X}}_d^T \hat{\beta}$. The weight of the single components hereby depends on the shrinkage factor γ_d. Hence, if the sampling variance of the direct estimators ψ_d is comparatively high in an area d, the respective γ_d tends to be comparatively low. As the direct estimator for this area is considered to be unreliable, a correspondingly large weight is placed on the regression-synthetic part of the BLUP. If, on the contrary, a low area-specific sampling variance ψ_d or a high general model variance σ_v^2 is given, the weight increases and more confidence is put in the direct estimator of the respective area.

In the practical application, σ_v^2 is unknown and has to be estimated as well. For this purpose a number of approaches exist. Within the following estimation, the variance parameter has been estimated by means of the restricted maximum likelihood (REML) method. For details with respect to this approach, we refer to Rao and Molina (2015, pp. 102-105; 127-128). By replacing the model variance σ_v^2 by the estimated variance of the random effects $\hat{\sigma}_v^2$ in (9) and (10), the empirical best linear unbiased predictor (EBLUP) is obtained.

3.3.2 Battese-Harter-Fuller estimator

In contrast to the area-level models described in the previous section, unit-level models do not use aggregate information but micro-level information instead, which enables a more efficient estimation. The standard procedure is the Battese-Harter- Fuller estimator (cf. Battese et al., 1988).

The model underlying the Battese-Harter-Fuller estimator and assumed for the population is a special form of the general mixed linear regression model and given by

$$y_{dk} = x_{dk}^T \beta + v_d + e_{dk},\ d = 1, \dots, D,\ k = 1, \dots, N_d \tag{12}$$

with $v_d \overset{iid}{\sim} (0, \sigma_v^2)$ and $e_{dk} \overset{iid}{\sim} (0, \sigma_e^2)$. The vector of the regression coefficients β measures the relationship between the variable of interest y_{dk} and the auxiliary variables x_{dk}^T over all areas and units. The term e_{dk} describes the individual sampling error of the units within the unit-level model. As in (8), the variance of the random effects σ_v^2, also referred to as model variance, measures the variance between the areas that cannot be explained by the fixed component of the model. It is also assumed that σ_v^2 and σ_e^2 are independent of each other.

Assuming that the mixed regression model (12) also applies to the sample, the mean value of the variable of interest per area is estimated by the BLUP according to Battese, Harter and Fuller (1988):

$$\hat{\mu}_d^{BHF} = \overline{X}_d^T\hat{\beta} + \hat{v}_d \tag{13}$$

with $\hat{v}_d = \gamma_d(\bar{y}_d - \bar{x}_d^T\hat{\beta})$

and $\gamma_d = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2/n_d}$

(cf. Rao and Molina, 2015, p. 174 f.), where $\bar{y}_d$ and $\bar{x}_d$ are the sample averages of the variable of interest and the auxiliary variables in area d, respectively. The auxiliary information $\overline{X}_d$, on the other hand, includes both units included and not included in the sample. The BLUP can also be transformed into a composite estimation function:

$$\hat{\mu}_d^{BHF} = \gamma_d\left(\bar{y}_d + (\overline{X}_d - \bar{x}_d)^T\hat{\beta}\right) + (1-\gamma_d)\overline{X}_d^T\hat{\beta} \tag{14}$$

Here, it has to be recognised that the Battese-Harter-Fuller estimator is a weighted average of the direct sample regression estimator $\bar{y}_d + (\overline{X}_d - \bar{x}_d)^T\hat{\beta}$ and the regression-synthetic component $\overline{X}_d^T\hat{\beta}$. The weighting factor γ_d indicates for each area the share of the model variance in relation to the total variance and determines how much weight is given to the respective components. With a high model variance of σ_v^2 or a large area-specific sample size n_d, respectively, much confidence is placed in the direct sample regression estimator. In turn, the BLUP tends to approach the synthetic component if the model variance is low or the sample size is small. Accordingly, for areas in which no unit has been sampled ($n_d = 0$, so $\gamma_d = 0$), the BLUP consists entirely of the synthetic estimator. However, this assumes that the auxiliary characteristics of the units of this area are present, so that the area-specific average value $\overline{X}_d$ can be taken into account in the estimation.

However, since the model variance σ_v^2 and the variance of the sampling error σ_e^2 are not known in practice, they have to be estimated. There are various methods for estimating the variance components. By replacing the variance components of the BLUP with the corresponding estimated values, the unit-level EBLUP is created according to Battese, Harter and Fuller (1988).

3.3.3 Measurement error model

When using model-based small area methods, it is generally assumed that the auxiliary information $\overline{X}_d$ is correct and free of errors. However, this is not always the case in practice. Especially in this application, it is mostly inevitable to use covariates from a survey which, however, tend to be subject to sampling errors. Thus, it cannot be guaranteed that the auxiliary variable averages $\overline{X}_d$ are actually the true population averages. Ybarra and Lohr (2008) show that the Fay-Herriot estimator can be even more inefficient than the simple direct design-weighted estimator when using incorrect auxiliary information $\widehat{\overline{\mathrm{X}}}_d$.

The solution proposed by Ybarra and Lohr (2008) is a conditionally unbiased estimation procedure based on a so-called measurement error model and used for erroneous covariables. First, it is assumed that $\widehat{\overline{\mathrm{X}}}_d \overset{ind}{\sim} N\left(\overline{\mathrm{X}}_d, C_d\right)$, where C_d is the known variance-covariance matrix of the estimated mean values of the register variables. Furthermore, $\widehat{\overline{\mathrm{X}}}_d$ is independent of v_d and e_d (see Rao and Molina, 2015, p. 156). Like the Fay-Herriot estimator, the measurement error estimator is also a linear combination of the direct estimator and a regression-synthetic part:

$$\hat{\mu}_d^{ME} = \gamma_d \hat{\mu}_d^{Dir} + (1-\gamma_d)\widehat{\overline{\mathrm{X}}}_d^T \beta \tag{15}$$

The weighting factor γ_d depends not only on the model variance σ_v^2 and the design variance ψ_d but also on the variability of the estimated auxiliary variables. The optimal weighting factor, which minimises the MSE of the measurement error estimator over all linear combinations, is given by

$$\gamma_d = \frac{\sigma_v^2 + \beta^T C_d \beta}{\sigma_v^2 + \beta^T C_d \beta + \psi_d} \tag{16}$$

The more inexactly $\widehat{\overline{\mathrm{X}}}_d$ is measured, the greater are C_d and the weight γ_d, which is put on the direct estimator $\hat{\mu}_d^{Dir}$. If the measurement of $\widehat{\overline{\mathrm{X}}}_d$ is made without error ($C_d = 0$), $\hat{\mu}_d^{ME}$ is reduced to the Fay-Harriot estimator by $\gamma_d = \sigma_v^2/(\sigma_v^2 + \psi_d)$. Assuming that the parameters β, σ_v^2, and ψ_d are known, the MSE of (15) is

$$MSE(\hat{\mu}_d^{ME}) = \gamma_d \psi_d \tag{17}$$

Since $0 \leq \gamma_d \leq 1$, the MSE of the measurement error estimator is at most as large as the MSE of the direct estimator ψ_d. The MSE of the Fay-Herriot estimator, on the other hand, can be greater than ψ_d if incorrect auxiliary information is taken into account (see Ybarra and Lohr, 2008, p. 921). Consequently, the measurement error estimator is an improvement over the general area-level model in which erroneous covariates are ignored.

As with the small area estimators presented above, the regression coefficients β and the model variance σ_v^2 are unknown in practice and must be estimated. The model variance is estimated by a simple moment estimator, which is given by

$$\hat{\sigma}_v^2 = (D-P)^{-1} \sum_{d=1}^{D} \left(\left(\hat{\mu}_d^{Dir} - \hat{\bar{X}}_d^T \hat{\beta}_w \right)^2 - \psi_d - \hat{\beta}_w^T C_d \hat{\beta}_w \right) \tag{18}$$

where *P* is the number of used auxiliary variables. The estimation of β is also achieved by a modified least squares estimator

$$\hat{\beta}_w = \left(\sum_{d=1}^{D} w_d \left(\hat{\bar{X}}_d \hat{\bar{X}}_d^T - C_d \right) \right)^{-1} \sum_{d=1}^{D} w_d \hat{\bar{X}}_d \hat{\mu}_d^{Dir} \tag{19}$$

(Ybarra and Lohr, 2008, p. 923), provided that the inverse exists. Ybarra and Lohr (2008, p. 924) show that $\hat{\beta}_w$ and $\hat{\sigma}_v^2$ are consistent estimators for β and σ_v^2 respectively, for $D \to \infty$. Here $w_d = 1/(\sigma_v^2 + \psi_d + \beta^T C_d \beta)$ are positive finite weights. The parameters are estimated in a two-step process. First, $w_d = 1$. The β and σ_v^2 are then estimated by (18) and (19). Based on the two estimates, the weights $\hat{w}_d$ are estimated again, to finally obtain the final estimates $\hat{\beta}_w$ and $\hat{\sigma}_v^2$ (see ibid).

In the simulation study, it is assumed that the area-level auxiliary variables are estimates from another survey and that their variance-covariance-matrix C_d is known. This variance-covariance-matrix has been defined for each area separately taking into account the variables covariances across all areas and a coefficient of variation of 10%. Using the known true values of area-specific covariates $\bar{X}_d$ and the defined matrix C_d, the 'estimated' covariates are generated randomly in each iteration. The known variance-covariance-matrix C_d is then used within the estimation technique according to Ybarra and Lohr (2008).

3.4 Synthetic estimation by cluster analysis

A further possibility is to cluster municipalities, cities or other areas of interest into regions which are homogeneous with respect to selected auxiliary variables that significantly correlate with the target variable. The target variable is then estimated for each cluster separately. This can be done by means of design-based, model- assisted, or model-based estimation techniques. However, it has to be considered that the estimate is identical for the areas that belong to the same cluster and can thus be considered as some type of synthetic estimate. While dealing with this approach, it is indirectly assumed that the variable of interest is homogeneous within each cluster. If this assumption is not valid, the estimators might be biased.

As the population age and gender structures of various areas tend to be variables easily accessible and nevertheless meaningful auxiliary variables, a cluster analysis based on this criteria is investigated within the simulation. At first, the mean age and the percentage of men is calculated for each area of interest. Both variables are then standardized by centering and dividing them by two standard deviations in order to avoid that one variable significantly predominates the division into clusters. Using the k-means clustering algorithm, all areas of interest are then assigned to clusters. In the simulation study, a total of ten clusters has been proven suitable. For each cluster, a Horvitz-Thompson estimate of the cluster mean is computed using approach (3). Subsequently, the mean estimate is assigned to all areas belonging to the respective cluster.

4 Selected results of the simulation study

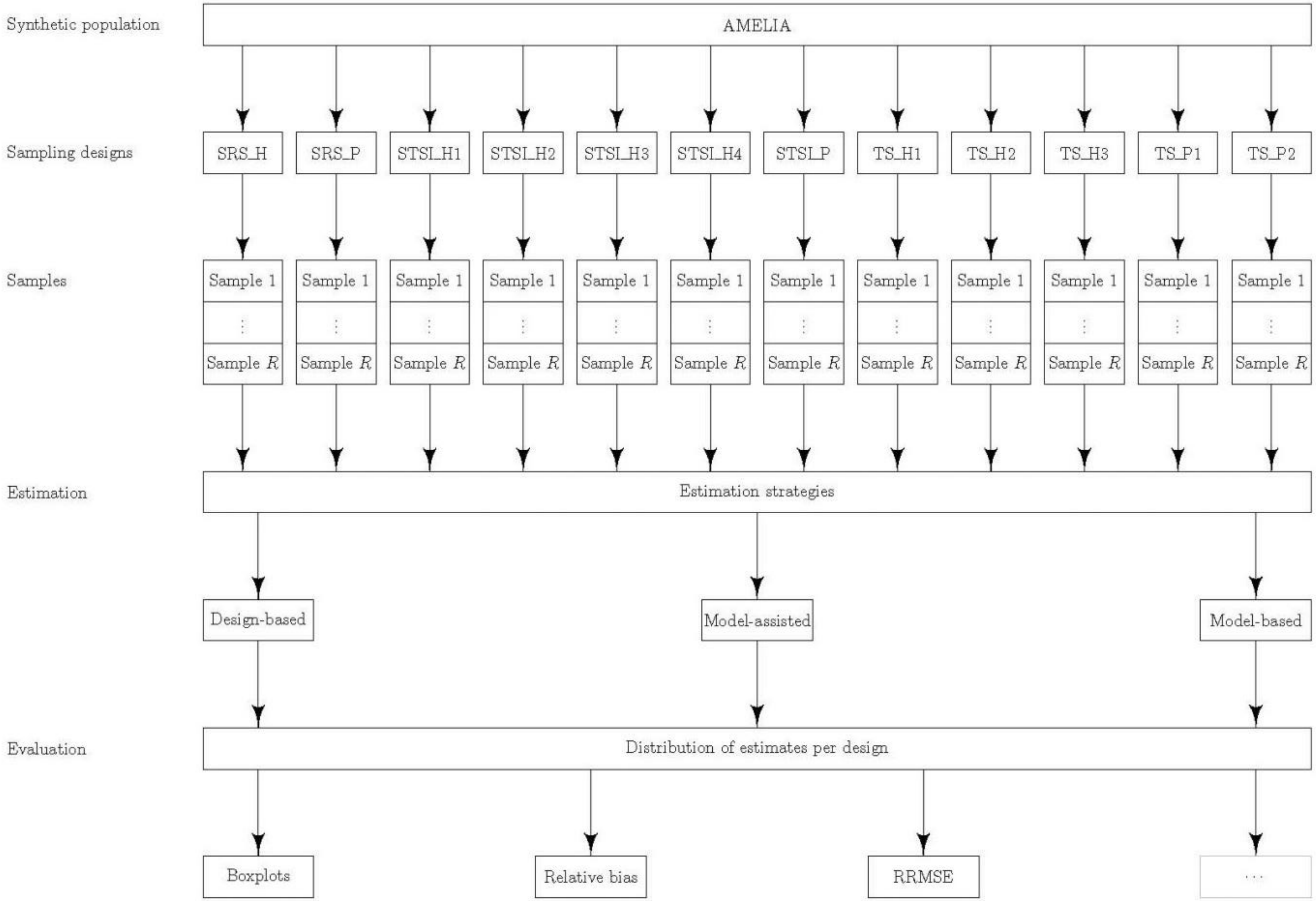
The process flow of the Monte Carlo simulation study is depicted in Figure 2. Starting with the synthetic population AMELIA (see Section 2 and Burgard et al., 2017b), the sampling designs described in Table 2 are implemented. R = 2 000 samples are drawn according to each design and the estimation strategies are applied to each sample.

The utilized auxiliary information depends on the type of the estimation approach and on the respective sampling design. For approaches using aggregated covariates, such as the Fay-Herriot estimator or the measurement error model according to Ybarra and Lohr (2008), auxiliary variables at area-level assumed to be known in practice are applied to stabilise the estimation. These include the share of persons with an ISCED-level of at least 5 (ISCED56), the unemployment rate (UER), the share of native-born persons (COB_LOC), the share of persons paying rent (RENT), the share of persons with a managerial position (SUP), the share of persons under the age of 20 (U20) as well as the AMELIA-region the respective area is belonging to (REG).

For estimation approaches at the individual data level, such as the GREG estimator or the Battese-Harter-Fuller estimator, information was utilized which seemed realistic to be available at unit-level in practice. Among others, these include data on the age of persons (AGE), the basic activity status (BAS), the country of birth (COB). In addition, again the AMELIA-region the respective area is belonging to (REG) is included as a factor variable. When persons are the final sampling units of the respective sampling designs, these unit-level information are observed at the individual level. In case households have been drawn as final sampling units, the respective unit-level information are aggregated for all members of the respective household.

There are various evaluation techniques for the large number of resulting estimates.

Figure 2: Flow chart of the Monte-Carlo simulation



In the following, the common measures of the (relative) bias

$$\mathrm{RBIAS} = \frac{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)}{\theta} \quad (20)$$

the (relative root) mean squared error

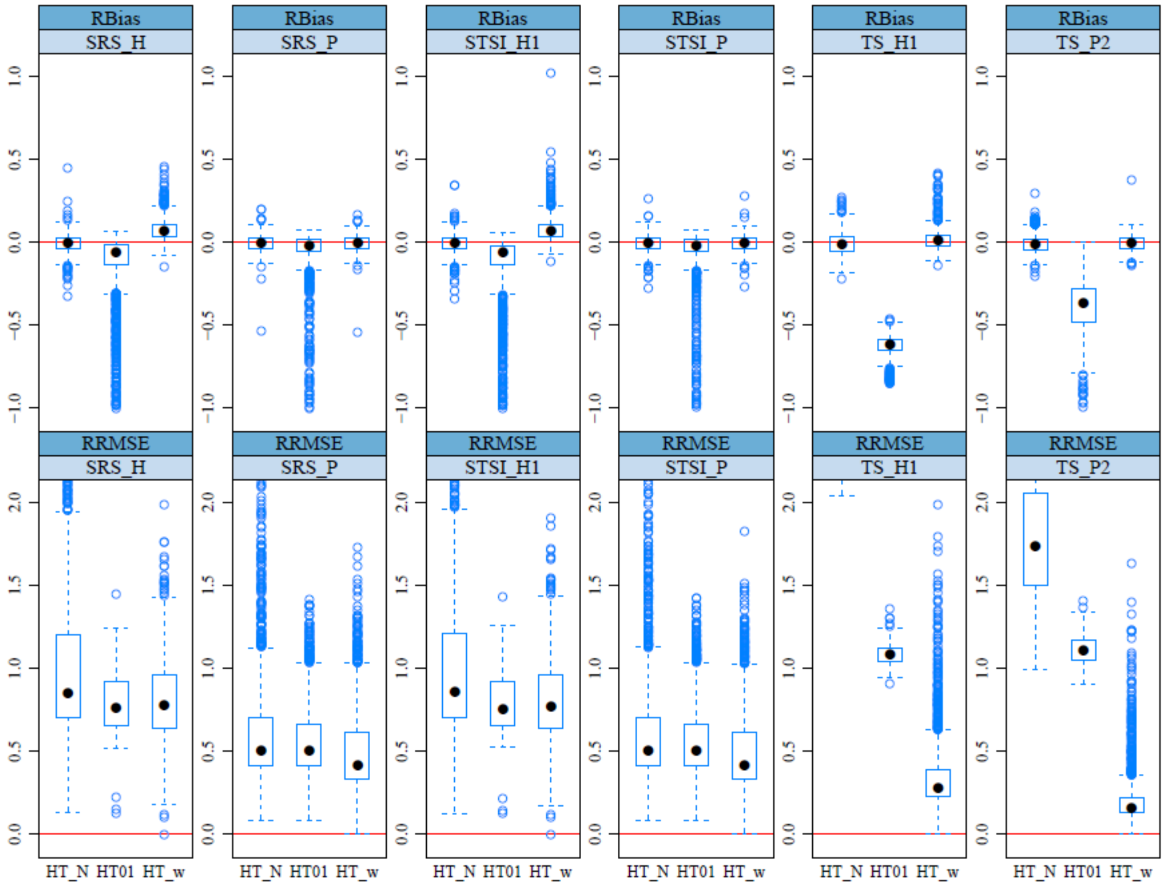
$$\mathrm{RRMSE} = \frac{\sqrt{\frac{1}{R}\sum_{r=1}^{R}(\hat{\theta}_r - \theta)^2}}{\theta} \quad (21)$$

and representations of the estimators' distributions like boxplots are used. In Equations 20 and 21, θ is the true parameter (known because we use a synthetic population and are able to compute the target parameter using *synthetic census* data) to be estimated from the samples, and $\hat{\theta}_r$ is the estimate computed using the r-th sample.

Before deriving the AROPE estimates at city-level, a reliable estimation of the target variable at municipality-level is required. Subsequently, the estimates are aggregated in order to obtain estimations for the large cities of AMELIA.

Figure 3 illustrates the results of three different versions of the Horvitz-Thompson estimator for the estimation of the AROPE rate at municipality-level, i.e. for the 1,592 municipalities corresponding to the LAU level included in AMELIA. Both the relative bias and the RRMSE are depicted depending on selected sampling designs implemented in the simulation. HT_N is the common Horvitz-Thompson mean estimator according to equation (2). It can be confirmed that the estimations according to this approach are unbiased with respect to every sampling design. Nevertheless, the estimates are subject to a remarkable RRMSE indicating that the estimations are unbiased but inefficient, especially when dealing with two-stage sampling approaches. This is due to the fact that, although the target values are proportions, the estimates can take values clearly larger than 1 while the estimates for non-sampled areas have the value 0. Two-stage sampling approaches often have municipalities or some regions consisting of several municipalities as primary sampling units. Therefore, the areas of interest are either not sampled at all or sampled at a comparatively high sampling fraction. In these cases, the estimation of the AROPE rate is certainly unbiased, but either takes the value 0 or a value far above 1. Under the approach HT01, values larger than 1 have been corrected downwards to 1. This adaptation clearly decreased the relative RRSME, even if the estimates are now no longer unbiased. An improvement of the efficiency can also achieved by the weighted sample mean HT_w (3).

Figure 3: Results: Versions of the Horvitz-Thompson estimator



Sources: see section 5 - References

This is especially apparent in the case of two-stage sampling. Here, the additional estimation of N_d causes a clear stabilisation of the estimation.

However, it has to be noted that this representation neglects the sum of non-sampled areas given the respective designs. Therefore, Table 3 lists the percentage share of non-sampled areas for the samples drawn according to the different designs in the simulation study. The large share of non-sampled areas using a two-stage design is particularly high. Hereby, purely design-based estimation strategies cannot be applied at all to the respective areas. At least in such cases, a model-assisted or model-based approach has to be utilised.

Table 3: Average share of nonsampled areas (in percent) per sampling design

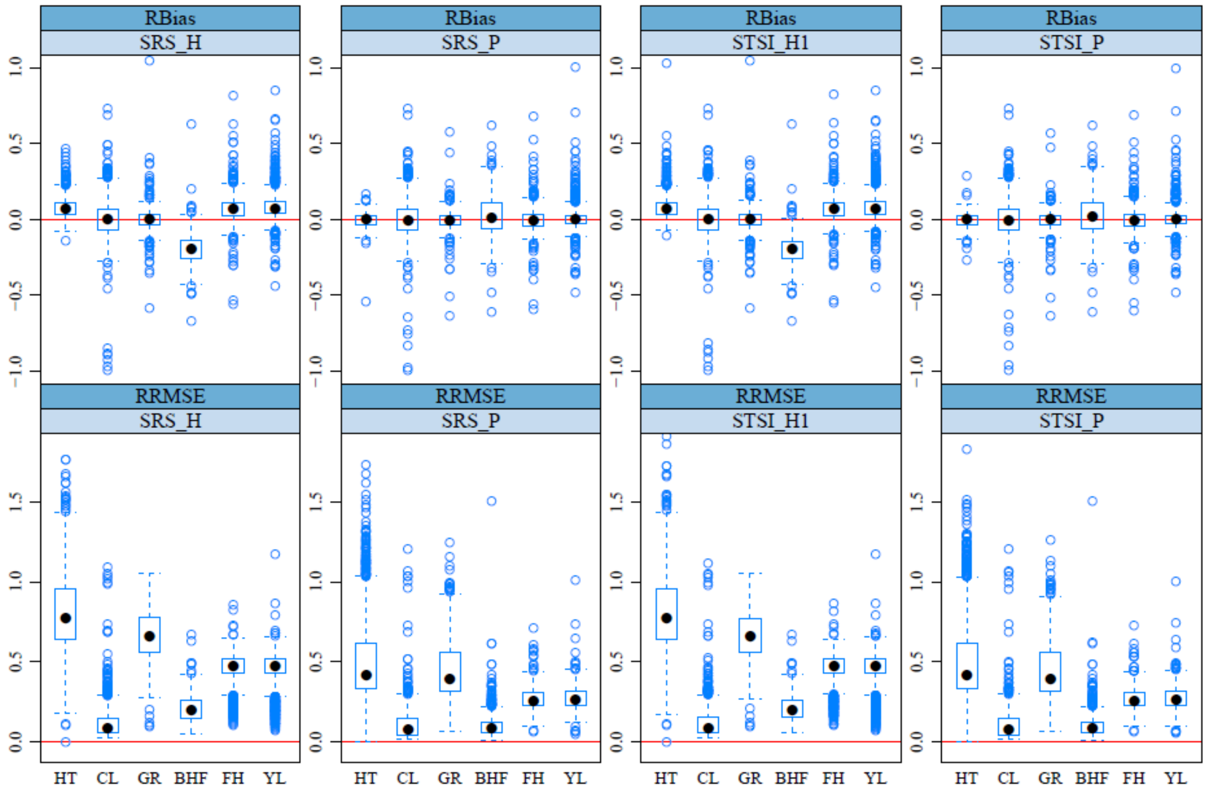
Sampling design	Municipalities	Large cities
SRS H	13.61	0.03
SRS P	5.04	0.00
STSI H1	13.59	0.03
STSI H2	13.60	0.03
STSI H3	13.58	0.03
STSI H4	13.18	0.02
STSI P	5.04	0.00
TS H1	83.98	70.24
TS H2	84.05	70.51
TS H3	84.04	83.45
TS P1	84.05	70.51
TS P2	72.64	67.94

Sources: see section 5 - References

Therefore, the results of the investigated small areas estimation approaches at municipality-level are illustrated in Figure 4. First, the results are considered given selected simple random sampling and stratified random sampling approaches. The estimation approaches comprise the weighted sample mean (HT; see equation 3), the Horvitz-Thompson estimator at cluster-level (CL), the GREG estimator (GR), the Battese-Harter-Fuller estimator (BHF), the Fay-Herriot estimator (FH) as well as the Ybarra-Lohr estimator based on the measurement error model (YL).

Focussing at the relative bias at first, slight biases can be observed occasionally, which is not surprising dealing with model-based estimation approaches. Only the Battese-Harter-Fuller estimator causes slight underestimations given a household-level sampling design. On the contrary when dealing with sampling designs at person-level, no approach stands out negatively in terms of the relative bias.

Figure 4: Results: Small area estimation at municipality-level under selected simple random sampling and stratified random sampling approaches

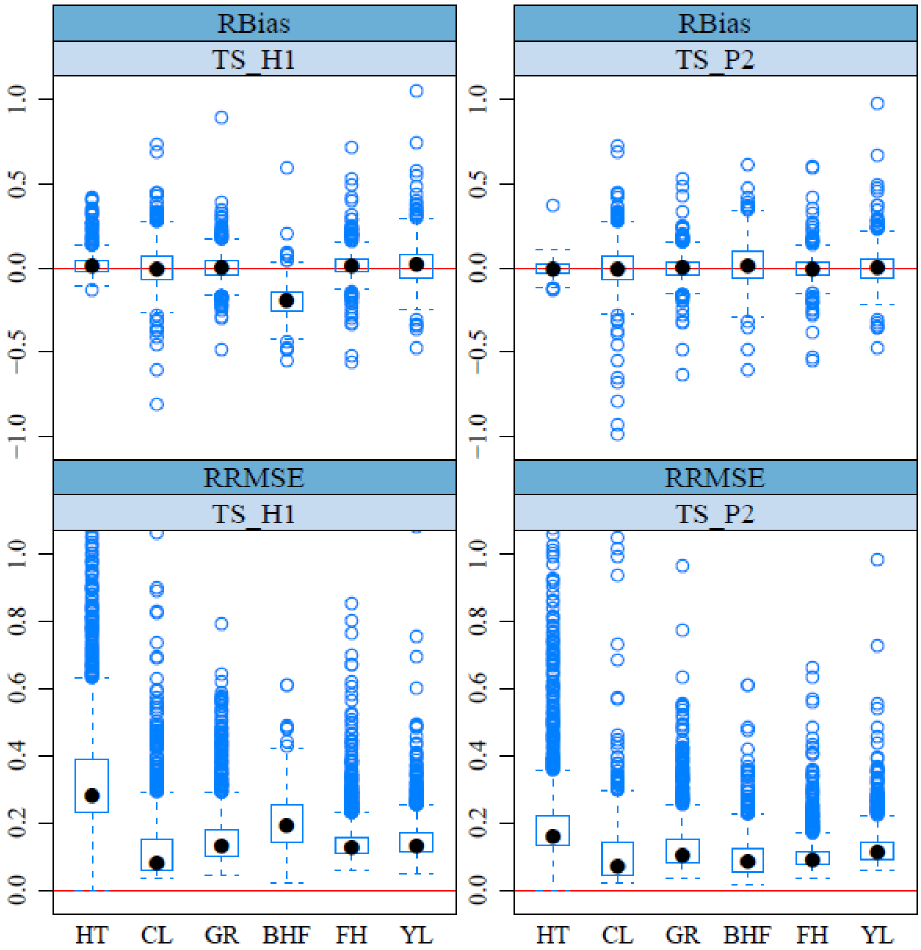


Sources: see section 5 - References

With regard to the RRMSE, all implemented small area estimation approaches induce an improvement compared to the weighted sample mean (HT). Especially the synthetic Horvitz-Thompson estimates for clustered municipalities convinces through a remarkable low RRMSE for most areas. However, these results can be explained by the synthetic nature of the population AMELIA and therefore need to be treated with caution. As AMELIA is partitioned into different regions, which have their own structures in terms of age, gender and poverty, these regions also recur in the formed clusters. In reality, the differences between clusters can be considered less hard, which also reduces the potential of the synthetic estimation approach.

The results of the Battese-Harter-Fuller estimator are likewise convincing, which emphasizes the potential contained in unit-level information. The Fay-Herriot estimator and the Ybarra-Lohr estimator are largely similar. Despite the utilization of estimated area-level auxiliary variables, the estimator based on the measurement error model (YL) therefore seems to compete with the Fay-Herriot estimator employing true covariates.

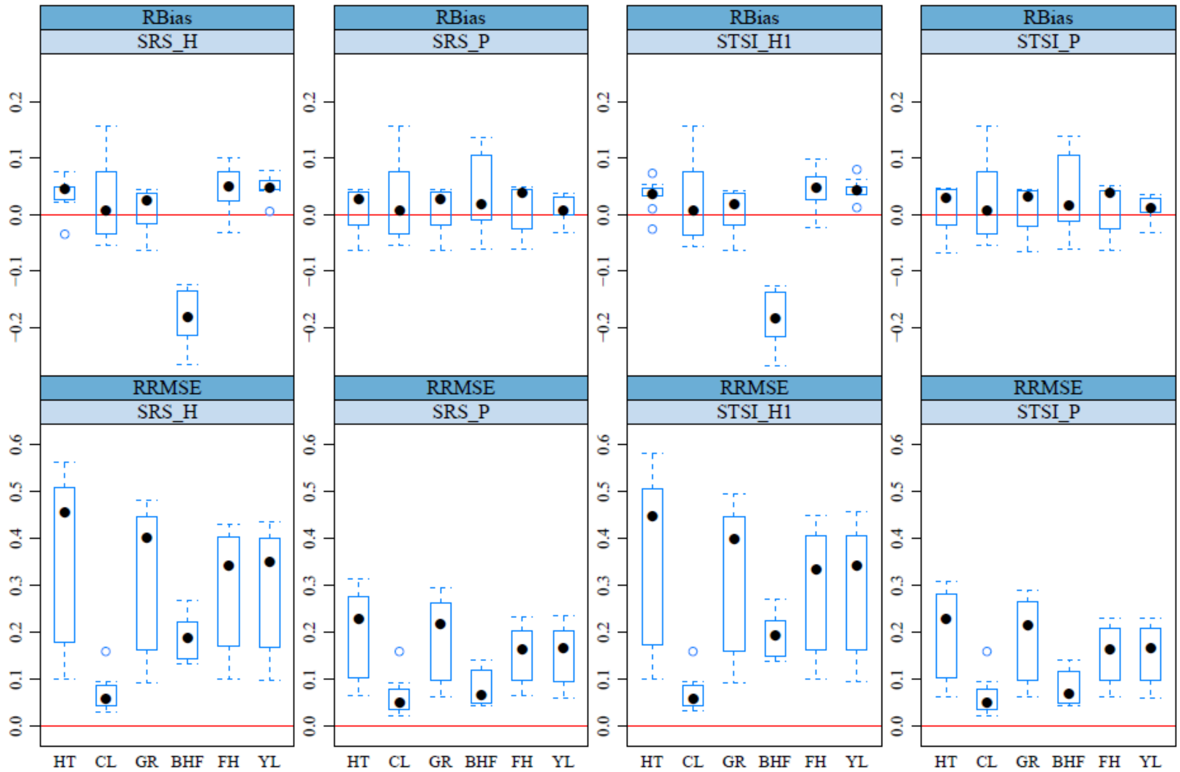
Figure 5: Results: Small area estimation at municipality-level under selected two-stage sampling approaches



Sources: see section 5 - References

In the same structure, Figure 5 gives an overview of the results of the estimation strategies at municipality-level dealing with two-stage sampling approaches. With respect to the relative bias, no noteworthy differences to the simple random sampling and stratified random sampling designs can be identified. However, the RRMSE of the estimations for the observed areas based on a two-stage sampling design has clearly decreased when utilizing the GREG estimator or an area-level model estimator, such as the Fay-Herriot or the Ybarra-Lohr estimator. This might be explained by the fact that, in two-stage designs, those areas that have been sampled as a primary sampling unit are sampled to a comparatively high extent. The relatively high sampling fraction in sampled areas might cause a more stable model estimation in case of the respective approaches.

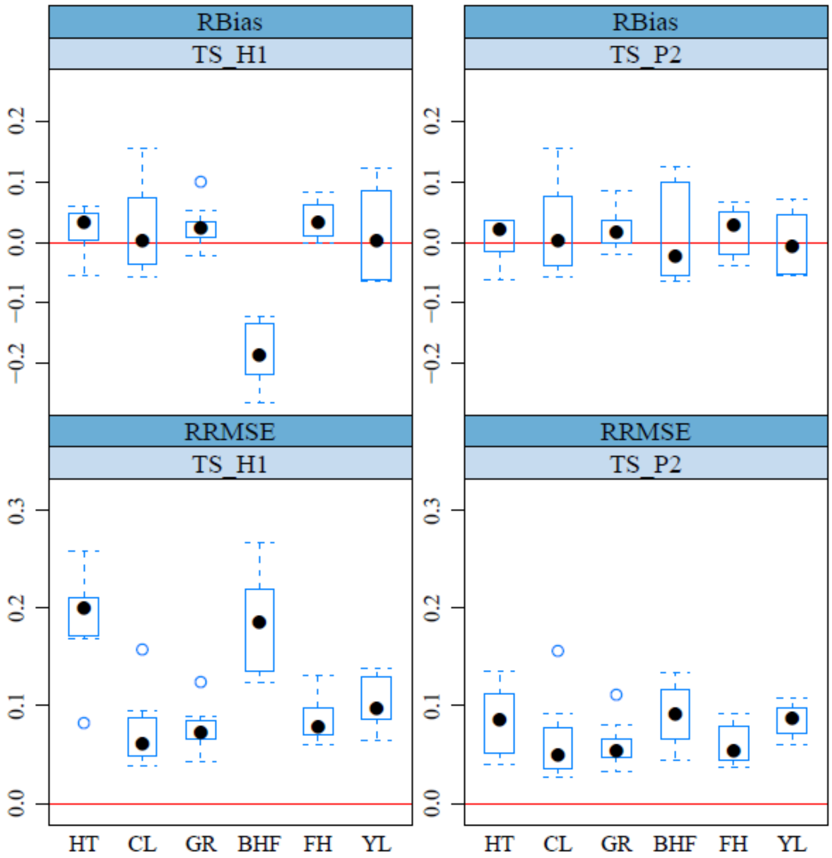
Figure 6: Results: Small area estimation at city-level under selected simple random sampling and stratified random sampling approaches



Sources: see section 5 - References

Following Figure 4, Figure 6 now outlines the results of the small area techniques for the estimation of the AROPE rate in each of the ten large cities in AMELIA. Again an underestimation of the target parameter can be observed using the Battese-Harter-Fuller estimator in combination with a sampling design at household-level. The remaining estimations are subject to a relative bias that is comparable to the estimation at municipality-level. Overall, however, it can be observed that the potential for improvement by using the implemented small area estimation approaches declines at city-level. This is due to the fact that the direct estimation using the weighted sample mean is already subject to a comparatively high quality given the increased number of area-specific sampling units. Only the clustering approach and the Battese-Harter-Fuller estimator cause a further reduction of the RRMSE.

Figure 7: Results: Small area estimation at city-level under selected two-stage sampling approaches

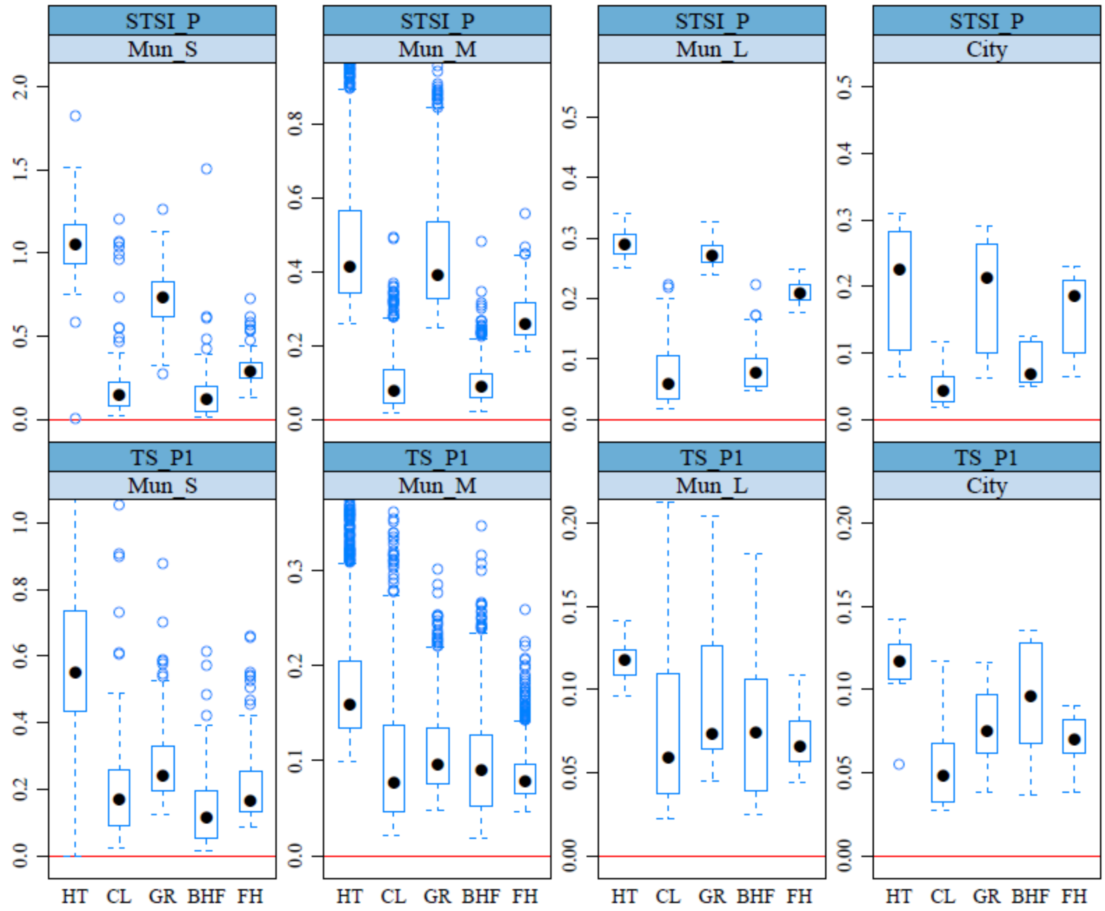


Sources: see section 5 - References

Supplementary to the previous figures, Figure 7 depicts the results of the estimation strategies at city-level dealing with two-stage sampling approaches. Here, again, the Battese-Harter-Fuller approach is not convincing in terms of both the relative bias and the RRMSE. The other estimation techniques however cause a clear improvement of the estimation efficiency under a household-level two-stage design. The decrease of the RRMSE under the two-stage sampling at person-level however is only marginal, as the estimation quality is already comparatively high in this case.

In general, the performance of different small area estimation approaches clearly depends on the respective sample size of the areas of interest. The expected sample size however also depends on the size of the area itself. To investigate the influence of different areas sizes on the estimation quality, different size categories have been classified. These comprise small AMELIA municipalities (less than 1,000 inhabitants; Mun_S), medium-sized municipalities (from 1000 to 11,000 inhabitants; Mun_M), large municipalities (more than 11,000 inhabitants; Mun_L) as well as the constructed large cities (City).

Figure 8: Results: Estimation quality in relation to the size of the target area



Sources: see section 5 – References

The RRMSE of the estimation approaches in each of the constructed size categories depending on two selected sampling designs is illustrated in Figure 8. Hereby, the different axis scalings have to be noted. In particular, it is of interest to what extent an improvement can be achieved in comparison to the weighted sample mean (HT) not including any auxiliary information. Especially when the estimation of the AROPE rate in small municipalities is of interest, the RRMSE of the weighted sample mean in most areas is unbearably high due to comparatively low expected sample sizes in these areas. By incorporating auxiliary information, the examined estimation approaches are able to clearly increase the efficiency of the estimation. The potential for improvement by using the estimation approaches slightly decreases with increasing area size. Focusing on the STSI_P sampling design, it becomes obvious that especially the improvement by using the GREG estimator or the Fay-Herriot estimator declines. The clustering estimation and the Battese-Harter-Fuller estimator however still achieve a clear improvement of the estimation quality. If however the samples have been drawn according to the two-stage design TS_P1, the potential for improvement decreases throughout all investigated estimation approaches. Especially the estimation using the weighted sample mean in large municipalities (Mun_L) is already comparatively reliable due to relatively high expected sample sizes, which enable a direct design-based estimation of sufficient precision. Thus, an estimation using the clustering approach, the GREG estimator or the Battese-Harter-Fuller estimator even causes a decline in the estimation quality for certain areas. Therefore, it has to be noted that the choice of the estimation strategy and whether to apply small area estimation approaches depend on the area-specific sample size, which tends to increase with the size of the areas given the common sampling designs of European social surveys.

5 References

Alfons, A., Filzmoser, P., Hulliger, B., Kolb, J.-P., Kraft, S., Münnich, R. and Templ, M. (2011): *Synthetic data generation of SILC data. Research Project Report WP6–D6.2*. Technical report, AMELI.

Battese, G. E., Harter, R. M. and Fuller, W. A. (1988): *An error-components model for prediction of county crop areas using survey and satellite data*. Journal of the American Statistical Association, 83 (401), pp. 28–36.

Burgard, J., Ertz, F., Merkle, H. and Münnich, R. (2017a): AMELIA - Data description v0.2.2.1. Trier University, www.amelia.uni-trier.de.
URL http://amelia.uni-trier.de/wp-content/uploads/2017/11/AMELIA_ Data_Description_v0.2.2.1.pdf

Burgard, J. P., Kolb, J.-P., Merkle, H. and Münnich, R. (2017b): *Synthetic data for open and reproducible methodological research in social sciences and official statistics*. AStA Wirtschafts-und Sozialstatistisches Archiv, 11 (3-4), pp. 233–244.

Eurostat (2010): Description of Target Variables: Cross-sectional and Longitudinal. EU-SILC 065 (2008 operation) ed.

Eurostat (2018a): *Glossary: At risk of poverty or social exclusion (AROPE)*. Last modified on 24 September 2018, at 16:13.
URL https://ec.europa.eu/eurostat/statistics-explained/index.php/Glossary:At_risk_of_poverty_or_social_exclusion_(AROPE)

Eurostat (2018b): *Glossary: At-risk-of-poverty rate*. Last modified on 24 Septem- ber 2018, at 16:13.
URL https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:At-risk-of-poverty_rate

Eurostat (2018c): *Glossary: Material deprivation*. Last modified on 24 September 2018, at 16:13.
URL https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Severe_material_deprivation_rate

Eurostat (2018d): *Glossary: Persons living in households with low work intensity*. Last modified on 24 September 2018, at 16:13.

URL https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Persons_living_in_households_with_low_work_ intensity

Fay, R. E. and Herriot, R. A. (1979): *Estimates of income for small places: an application of James-Stein procedures to census data*. Journal of the American Statistical Association, 74 (366a), pp. 269–277.

Horvitz, D. G. and Thompson, D. J. (1952): *A generalization of sampling without replacement from a finite universe*. Journal of the American Statistical Association, 47 (260), pp. 663–685.

Jiang, J. and Lahiri, P. (2006): *Mixed model prediction and small area estimation*. Test, 15 (1), pp. 1–96.

Lehtonen, R. and Veijanen, A. (2009): *Design-based methods of estimation for domains and small areas*. Handbook of statistics, 29, pp. 219–249.

Lohr, S. L. (2009): Sampling: design and analysis. Nelson Education.

Merkle, H. and Münnich, R. (2016): *The AMELIA Dataset - A Synthetic Universe for Reproducible Research*. Berger, Y. G., Burgard, J. P., 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. and Tzavidis, N. (editors) InGRID Deliverable 23.1: Case studies, WP23 – D23.1, http://inclusivegrowth.be. URL http://inclusivegrowth.be

Rao, J. N. and Molina, I. (2015): Small area estimation. John Wiley & Sons.

Särndal, C., Swensson, B. and Wretman, J. (1992): Model Assisted Survey Sampling. Springer Verlag, New York.

Ybarra, L. M. and Lohr, S. L. (2008): *Small area estimation when auxiliary information is measured with error*. Biometrika, pp. 919–931.

A Estimation results at municipality-level

Table 4: Mean relative bias of the estimation at municipality-level

	HT_N	HT01	HT_w	CL	GR	BHF	FH	YL
SS_H	-0.0054	-0.1281	0.0769	0.0049	-0.0047	-0.1956	0.0743	0.0812
SRS_P	-0.0058	-0.0483	-0.0055	0.0044	-0.0051	0.0250	-0.0015	0.0024
STSI_H1	-0.0048	-0.1275	0.0773	0.0046	-0.0042	-0.1967	0.0734	0.0803
STSI_H2	-0.0047	-0.1277	0.0769	0.0045	-0.0045	-0.1953	0.0739	0.0808
STSI_H3	-0.0061	-0.1275	0.0767	0.0044	-0.0048	-0.1963	0.0732	0.0802
STSI_H4	-0.0055	-0.1242	0.0776	0.0045	-0.0040	-0.1955	0.0731	0.0801
STSI_P	-0.0059	-0.0479	-0.0058	0.0047	-0.0049	0.0257	-0.0007	0.0029
TS_H1	-0.0059	-0.6203	0.0136	0.0060	0.0002	-0.1946	0.0157	0.0239
TS_H2	-0.0043	-0.6216	0.0157	0.0068	0.0008	-0.1951	0.0167	0.0251
TS_H3	-0.0041	-0.6239	0.0173	0.0039	0.0016	-0.1930	0.0156	0.0283
TS_P1	-0.0045	-0.6133	-0.0045	0.0055	0.0008	0.0245	0.0034	0.0112
TS_P2	-0.0061	-0.3686	-0.0047	0.0044	0.0006	0.0217	-0.0044	0.0061

Sources: see section 5 – References

Table 5: Median relative bias of the estimation at municipality-level

	HT_N	HT01	HT_w	CL	GR	BHF	FH	YL
SRS_H	-0.0027	-0.0614	0.0713	-0.0024	-0.0008	-0.1949	0.0717	0.0731
SRS_P	-0.0027	-0.0171	-0.0026	-0.0034	-0.0035	0.0139	-0.0049	-0.0026
STSI_H1	-0.0006	-0.0595	0.0700	-0.0026	-0.0007	-0.1960	0.0698	0.0720
STSI_H2	-0.0029	-0.0600	0.0729	-0.0026	-0.0004	-0.1948	0.0706	0.0726
STSI_H3	-0.0035	-0.0609	0.0690	-0.0031	-0.0006	-0.1958	0.0706	0.0732
STSI_H4	-0.0019	-0.0578	0.0714	-0.0025	-0.0001	-0.1946	0.0703	0.0716
STSI_P	-0.0029	-0.0163	-0.0025	-0.0035	-0.0015	0.0148	-0.0041	-0.0021
TS_H1	-0.0085	-0.6169	0.0123	-0.0035	0.0037	-0.1943	0.0137	0.0248
TS_H2	-0.0048	-0.6191	0.0168	-0.0020	0.0052	-0.1945	0.0137	0.0266
TS_H3	0.0010	-0.6139	0.0172	-0.0024	0.0059	-0.1922	0.0143	0.0281
TS_P1	-0.0060	-0.6125	0.0004	-0.0033	0.0045	0.0115	0.0016	0.0125
TS_P2	-0.0103	-0.3636	-0.0045	-0.0019	0.0038	0.0111	-0.0056	0.0049

Sources: see section 5 - References

Table 6: Mean RRMSE of the estimation at municipality-level

	HT_N	HT01	HT_w	CL	GR	BHF	FH	YL
SRS_H	1.1931	0.7945	0.8180	0.1163	0.6657	0.2028	0.4638	0.4606
SRS_P	0.7008	0.5726	0.5104	0.1037	0.4553	0.1007	0.2740	0.2773
STSI_H1	1.1925	0.7943	0.8177	0.1159	0.6660	0.2068	0.4636	0.4604
STSI_H2	1.1890	0.7948	0.8180	0.1159	0.6666	0.2027	0.4647	0.4613
STSI_H3	1.1777	0.7943	0.8174	0.1137	0.6654	0.2035	0.4630	0.4599
STSI_H4	1.1570	0.7896	0.8155	0.1158	0.6658	0.2017	0.4652	0.4620
STSI_P	0.7031	0.5725	0.5101	0.1037	0.4546	0.1007	0.2730	0.2763
TS_H1	2.4843	1.0853	0.3599	0.1224	0.1575	0.1974	0.1501	0.1527
TS_H2	2.4956	1.0851	0.3604	0.1223	0.1575	0.1979	0.1501	0.1530
TS_H3	2.6269	1.0814	0.3724	0.1223	0.1592	0.1957	0.1537	0.1582
TS_P1	2.3725	1.0888	0.2106	0.1096	0.1211	0.0993	0.0948	0.1154
TS_P2	1.7938	1.1134	0.2157	0.1038	0.1304	0.0980	0.1069	0.1275

Sources: see section 5 – References

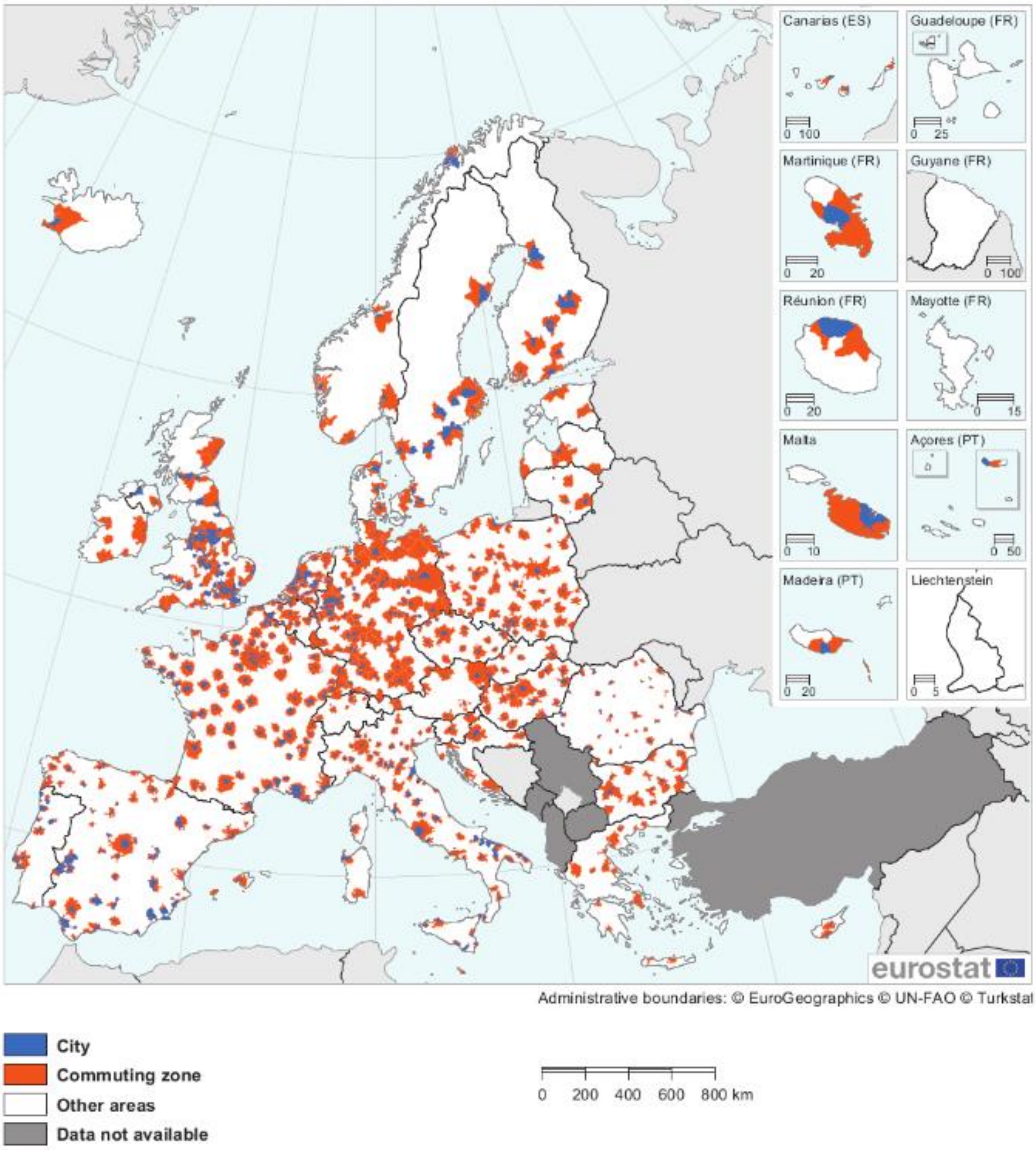
Table 7: Median RRMSE of the estimation at municipality-level

	HT_N	HT01	HT_w	CL	GR	BHF	FH	YL
SRS_H	0.8510	0.7576	0.7765	0.0826	0.6611	0.1996	0.4732	0.4747
SRS_P	0.5056	0.4992	0.4159	0.0752	0.3905	0.0894	0.2553	0.2616
STSI_H1	0.8550	0.7556	0.7703	0.0823	0.6617	0.2032	0.4743	0.4771
STSI_H2	0.8576	0.7588	0.7759	0.0826	0.6616	0.1994	0.4756	0.4771
STSI_H3	0.8520	0.7542	0.7722	0.0805	0.6614	0.2003	0.4734	0.4754
STSI_H4	0.8409	0.7504	0.7727	0.0823	0.6630	0.1985	0.4766	0.4777
STSI_P	0.5048	0.4994	0.4152	0.0750	0.3903	0.0891	0.2539	0.2606
TS_H1	2.4153	1.0811	0.2807	0.0844	0.1329	0.1954	0.1271	0.1355
TS_H2	2.4249	1.0814	0.2805	0.0846	0.1330	0.1958	0.1276	0.1356
TS_H3	2.4640	1.0785	0.2876	0.0838	0.1362	0.1936	0.1303	0.1407
TS_P1	2.3438	1.0850	0.1589	0.0771	0.0987	0.0905	0.0790	0.1016
TS_P2	1.7390	1.1077	0.1593	0.0748	0.1066	0.0886	0.0927	0.1141

Sources: see section 5 - References

B Estimation results at city-level

Figure 9: Cities and commuting zones, 2016



Note: based on population grid from 2011 to LAU 2016

Source: Eurostat, JRC and European commission, Directorate-General for Regional and Urban Policy

Table 8: Mean relative bias of the estimation at city-level

	HT_N	HT01	HT_w	CL	GR	BHF	FH	YL
SRS_H	0.0114	0.0081	0.0387	0.0220	0.0107	-0.1815	0.0477	0.0495
SRS_P	0.0106	0.0105	0.0100	0.0218	0.0106	0.0379	0.0110	0.0112
STSI_H1	0.0093	0.0062	0.0340	0.0216	0.0072	-0.1830	0.0435	0.0455
STSI_H2	0.0070	0.0038	0.0401	0.0218	0.0131	-0.1811	0.0485	0.0503
STSI_H3	0.0070	0.0035	0.0331	0.0220	0.0089	-0.1823	0.0436	0.0455
STSI_H4	0.0076	0.0046	0.0308	0.0219	0.0064	-0.1818	0.0405	0.0423
STSI_P	0.0115	0.0115	0.0112	0.0221	0.0111	0.0379	0.0118	0.0120
TS_H1	0.0157	-0.4787	0.0235	0.0202	0.0250	-0.1830	0.0372	0.0161
TS_H2	0.0047	-0.4925	0.0190	0.0208	0.0252	-0.1836	0.0376	0.0154
TS_H3	0.0013	-0.6061	0.0138	0.0215	0.0262	-0.1814	0.0372	0.0221
TS_P1	0.0044	-0.4860	0.0151	0.0201	0.0278	0.0148	0.0255	0.0051
TS_P2	0.0083	-0.2683	0.0081	0.0201	0.0209	0.0167	0.0183	0.0011

Table 9: Median relative bias of the estimation at city-level

	HT_N	HT01	HT_w	CL	GR	BHF	FH	YL
SRS_H	0.0261	0.0159	0.0443	0.0074	0.0242	-0.1821	0.0506	0.0480
SRS_P	0.0327	0.0327	0.0277	0.0072	0.0277	0.0173	0.0379	0.0074
STSI_H1	0.0232	0.0142	0.0371	0.0061	0.0192	-0.1838	0.0465	0.0430
STSI_H2	0.0159	0.0071	0.0443	0.0059	0.0250	-0.1819	0.0479	0.0521
STSI_H3	0.0226	0.0126	0.0446	0.0072	0.0279	-0.1821	0.0491	0.0451
STSI_H4	0.0237	0.0160	0.0427	0.0065	0.0237	-0.1822	0.0463	0.0446
STSI_P	0.0255	0.0255	0.0303	0.0070	0.0316	0.0156	0.0394	0.0109
TS_H1	0.0295	-0.5682	0.0347	0.0046	0.0244	-0.1847	0.0339	0.0044
TS_H2	0.0042	-0.5945	0.0403	0.0048	0.0246	-0.1853	0.0348	0.0063
TS_H3	0.0050	-0.5989	0.0367	0.0055	0.0247	-0.1830	0.0353	0.0093
TS_P1	0.0039	-0.5943	0.0321	0.0034	0.0229	-0.0261	0.0272	-0.0065
TS_P2	0.0132	-0.2906	0.0231	0.0040	0.0188	-0.0230	0.0303	-0.0044

Table 10: Mean RRMSE of the estimation at city-level

	HT_N	HT01	HT_w	CL	GR	BHF	FH	YL
SRS_H	0.4212	0.4118	0.3731	0.0698	0.3307	0.1878	0.2966	0.2974
SRS_P	0.2513	0.2513	0.1996	0.0616	0.1892	0.0817	0.1511	0.1503
STSI_H1	0.4159	0.4064	0.3715	0.0700	0.3328	0.1921	0.2952	0.2965
STSI_H2	0.4156	0.4061	0.3737	0.0705	0.3338	0.1876	0.2980	0.2990
STSI_H3	0.4159	0.4058	0.3684	0.0676	0.3333	0.1886	0.2928	0.2938
STSI_H4	0.4053	0.3963	0.3693	0.0700	0.3273	0.1872	0.2955	0.2966
STSI_P	0.2519	0.2518	0.2002	0.0619	0.1891	0.0820	0.1513	0.1505
TS_H1	2.0551	1.0621	0.1940	0.0726	0.0753	0.1842	0.0866	0.1029
TS_H2	2.0745	1.0537	0.1940	0.0733	0.0751	0.1848	0.0872	0.1032
TS_H3	2.3119	1.0804	0.1410	0.0726	0.0751	0.1826	0.0873	0.1053
TS_P1	2.0467	1.0584	0.1134	0.0657	0.0631	0.0955	0.0550	0.0858
TS_P2	1.5211	1.1071	0.0844	0.0626	0.0586	0.0901	0.0612	0.0840

Table 11: Median RRMSE of the estimation at city-level

	HT_N	HT01	HT_w	CL	GR	BHF	FH	YL
SRS_H	0.4881	0.4859	0.4547	0.0574	0.4002	0.1881	0.3412	0.3489
SRS_P	0.2719	0.2719	0.2287	0.0499	0.2175	0.0677	0.1632	0.1671
STSI_H1	0.4804	0.4784	0.4452	0.0579	0.3984	0.1923	0.3333	0.3410
STSI_H2	0.4796	0.4782	0.4524	0.0584	0.4158	0.1880	0.3378	0.3454
STSI_H3	0.4832	0.4805	0.4446	0.0555	0.4025	0.1882	0.3316	0.3387
STSI_H4	0.4626	0.4620	0.4480	0.0577	0.3961	0.1875	0.3389	0.3462
STSI_P	0.2757	0.2757	0.2270	0.0502	0.2152	0.0682	0.1623	0.1663
TS_H1	2.2162	1.0425	0.2007	0.0620	0.0737	0.1858	0.0790	0.0981
TS_H2	2.2439	1.0455	0.1958	0.0627	0.0757	0.1862	0.0775	0.1024
TS_H3	2.2544	1.0703	0.1613	0.0609	0.0742	0.1840	0.0801	0.0958
TSvP1	2.2324	1.0455	0.1167	0.0555	0.0613	0.0983	0.0470	0.0924
TS_P2	1.4733	1.0817	0.0868	0.0503	0.0545	0.0915	0.0543	0.0871

Getting in touch with the EU

In person
All over the European Union there are hundreds of Europe Direct Information Centres. You can find the address of the centre nearest you at: https://europa.eu/contact

On the phone or by e-mail
Europe Direct is a service that answers your questions about the European Union. You can contact this service
– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),
– at the following standard number: +32 22999696 or
– by electronic mail via: https://europa.eu/contact

Finding information about the EU

Online
Information about the European Union in all the official languages of the EU is available on the Europa website at: https://europa.eu

EU Publications
You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see https://europa.eu/contact)

EU law and related documents
For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: https://eur-lex.europa.eu

Open data from the EU
The EU Open Data Portal (http://data.europa.eu/euodp/en/data) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

Small area estimation for city statistics and other functional geographies

KS-TC-19-006-EN-N

The city data collection is one the regular data collections of Eurostat and the National Statistical Institutes. The demand for timely and reliable socio-economic data on cities and Functional Urban Areas has significantly increased. Since 2017, the cities and their Functional Urban Areas are legally recognised by the amended NUTS Regulation.

To produce socio-economic data coming originally from sample surveys at the level of small units such as cities, Functional Urban Areas and other functional geographies is a complex task which requires the application of small area estimation techniques since those functional geographies are usually not incorporated in the sampling design. This paper aims at investigating on the performance of different estimation strategies against the background of various sampling designs used by the National Statistical Institute. Therefore, a design-based Monte Carlo simulation study using a synthetic but close-to-reality population has been performed. The results show that all investigated estimation approaches are able to increase the efficiency and the quality of the estimates compared to the classical design-weighted estimation techniques.

For more information

https://ec.europa.eu/eurostat/

Publications Office
of the European Union

ISBN 978-92-76-11706-3