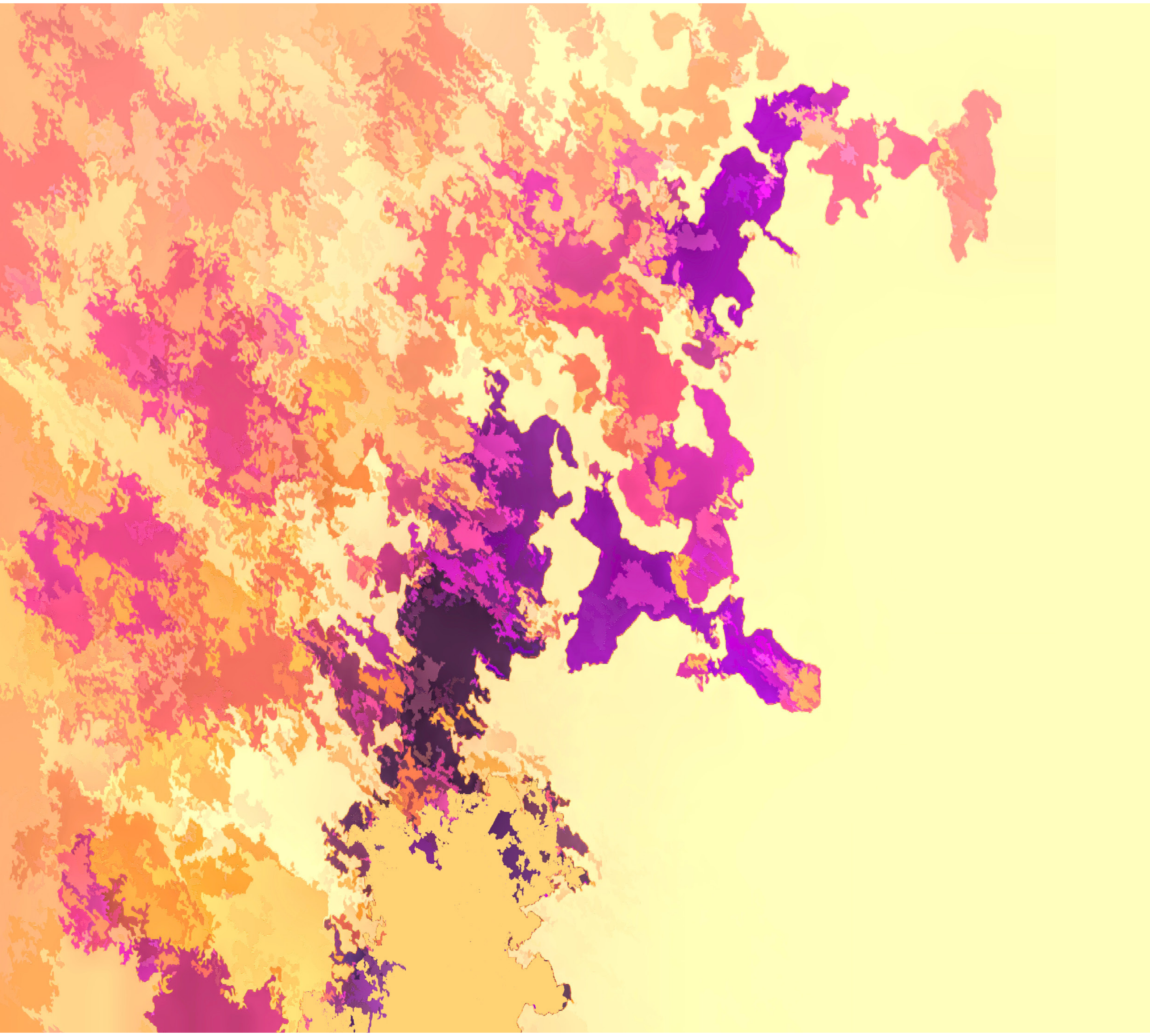
Guidelines on small area estimation for city statistics and other functional geographies

2019 edition



MANUALS
AND GUIDELINES

eurostat

Guidelines on small area estimation for city statistics and other functional geographics

2019 edition

Manuscript completed in October 2019.

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of the following information.

Luxembourg: Publications Office of the European Union, 2019
The European Commission is not liable for any consequence stemming from the reuse of this publication.

Copyright for the cover photo: © Shutterstock/Ken Schulze

The information and views set out in this publication are those of the author(s) and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

Collection: Manuals and guidelines
Theme: General and regional statistics

Print:	ISBN 978-92-76-11708-7	ISSN 2363-197X	doi:10.2785/822325	KS-GQ-19-011-EN-C
PDF:	ISBN 978-92-76-11709-4	ISSN 2315-0815	doi: 10.2785/763467	KS-GQ-19-011-EN-N

Acknowledgements

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.

Authors:
Economic and Social Statistics Department Trier University, Faculty IV, Economics
Ralf Münnich, Jan Pablo Burgard, Florian Ertz, Simon Lenau, Julia Manecke, Hariolf Merkle.

Contents

Figures

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 Introduction

1.1 Motivation

The demand for reliable information on the level of cities and functional urban areas has increased significantly. However, the increasing need for detailed information in politics and economics is largely offset by unchanged or even lower budgets for data collection. Often, the sampling designs of social surveys are only designed for a reliable design-based estimation at the state or federal level due to maximum permissible sample sizes. By contrast, cities and functional urban areas are usually not incorporated in the sampling design and are, therefore, referred to as unplanned areas. As the relevant areas might have unplanned and, hence, random as well as small sample sizes, the estimation of the respective parameters of interest might be challenging. A direct design-weighted estimation method such as the well-known estimator by Horvitz and Thompson (1952), which includes only sampled units from each area of interest itself, might lead to unbiased estimates. However, unplanned and small sample sizes due to the disregard of cities and functional urban areas at the stage of designing the sample might result in imprecise estimates with large standard errors. It might even be the case that some areas of interest may not be sampled at all. Thus, the user needs often exceed the limits of traditional estimation methods.

Small area estimation methods may be used to improve the quality of estimates for the respective areas of interest. 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 and even the estimation for areas which have not been sampled at all.

The purpose of the guidelines at hand is to help national statistical institutes to produce small area estimates. In addition to the reliability and precision of the estimates, coherence and comparability across all member states is of particular importance. This, however, might be a challenge as the sampling designs of suit- able surveys as well as the availability of auxiliary variables differs largely, which underlines the need for harmonization. Given each respective survey and data situation, the guidelines are supposed to provide support to identify the current best practice to implement small area estimation. The common step-by-step procedure of specifying, implementing and evaluating various small area estimation techniques up to the final selection of an optimal approach, therefore, is an essential step towards a comparability of the estimation process and the respective results.

1.2 Additional information

The field of small area estimation has grown fast in the past years. The guidelines cannot provide a full overview on all the methods that are available. For ensuring a certain degree of autonomy of this document, some basic small area estimation methods are described in the Appendix A. Further, an exemplary application that follows these guidelines is performed in Section 6 to enhance the applicability of the guidelines.

For the interested reader, wanting to deepen the knowledge in the field of small area estimation beyond the scope of these guidelines, the following references are a good starting point. The most prominent book on small area estimation is the one by Rao and Molina (2015). This book gives a concise overview of up-to-date small area methods from a mathematical statistical point of view. It also includes some examples on the use of the different estimators. Pfeffermann (2013) provides a review article on small area estimation methods, giving a very compact insight on the broad range of possible directions of the methodological advancements. In the book edited by Pratesi (2016), a large collection of different small area estimation techniques applied to poverty measurement is offered.

In cases where investigations and comparisons are required for specific situations to be encountered in official statistics, reports generated by or in cooperation with statistical offices might provide reasonable starting points. Examples are the work by Szymkowiak et al. (2017), Sõstra and Aru (2013), Strzalkowska and Molina (2018) as well as the EURAREA project (EURAREA Consortium, 2015). Nevertheless, it has to be emphasised that project reports designed for specific issues are not generalisable to each situation, in which small area estimates are desired.

1.3 Scope

The guidelines aim at providing a consistent framework for implementing small area estimation in the context of city statistics. In the course of this, they cover relevant issues associated with the specification, implementation and evaluation of methods as well as with the choice of a final optimal approach. Consequently, issues related to the different stages of the production process are covered and different options at each step are described. In contrast, the methods themselves are not discussed in detail. The interested reader will find further information in the appendix.

The guidelines in Sections 3 through 5 follow a standardized structure. They start with a description presenting each respective problem in the small area estimation process. This problem description is followed by a list of options, which present possibilities to deal with the specific problem at hand. On this basis, each guideline closes with a list of three ranked alternatives. Alternative (A) is considered to be the best choice that should be targeted in each case. Alternative (B) is an acceptable option to be chosen if certain constraints prevent the implementation of alternative (A). Alternative (C), however, is an alternative that is not recommended for further implementation and therefore should be avoided by the producers.

2 Terms and basic definitions

Small area estimation problems deal with the simultaneous estimation of several parameter values, which are partitioned according to regions or content.

Areas denote subgroups defined by regions.

Examples: federal states, sampling points, districts or municipalities, NUTS or LAU regions

Small areas/domains denote areas/domains in which the sample size is not large enough for a direct design-based estimation of sufficient precision. Further information from outside the area/domain are required for a reliable estimation. The term is independent of the actual size of the area/domain (e.g. the area of interest might be a rather large city, but the sample size within this city is too small for a precise direct estimation).

Direct estimators only use data from the area of interest itself.

Examples: mean estimators, separate regression estimators

Indirect estimators denote all estimators which also use information from outside the area of interest. In this context, one strives to borrow strength from additional information outside the area of interest. Thereby, similarities and differences between different areas or subpopulations are modelled using implicit or explicit models. These models are then used for the prediction of population characteristics.

Examples: national sample mean, model-based estimation approaches

Design-based estimators are approaches whose inference is based on the probability distribution generated by the underlying sampling design. The estimators refer to a fixed and finite population. Design-based methods are (approximately) design-unbiased. However, they might be subject to a large estimation variance if the (area/domain-specific) sample size is small. Note that design-based estimators may use models to improve the accuracy of the estimator. They are then also called model-assisted estimators (cf. Appendix A.2).

Examples: Horvitz-Thompson estimator, GREG estimator, calibration estimator

Model-based estimators include additional auxiliary information. Through the statistical modelling of a relation across all areas, it may be possible to achieve a stabilisation of the estimation. The objective is to predict unobserved values, which are then used to estimate the parameters. The finite population thereby is considered to be a random realisation of a super population. Model-based methods may be subject to a bias. Their estimation variance however tends to be small in comparison to design-based approaches. The (area/domain-specific) sample size only has a minor influence on the estimation variance.

Examples: synthetic approach, EBLUP (empirical best linear unbiased predictor) methods

Unit-level models can be used if information about the variable of interest, the auxiliary variables and the area membership is available for all units of the population. In the case of a linear model, the knowledge of the variable of interest and the auxiliary variables for the sample elements only is sufficient. Furthermore, the area-specific aggregated values of the auxiliary variables must be known for the population.

Area-level models can be used if no access is possible to unit-level information or suitable auxiliary information is not available at unit level. It requires direct estimators for the variables of interest in the areas and the aggregated values of the auxiliary variables in the areas. Area-level models have advantages in terms of computing time that can be used to estimate more complex models.

Planned areas exist if the underlying sampling design is a stratified random sample, whereby the areas of interest constitute the strata. The requirement is that the sampling design is based on the area membership of the population units. Areas can be considered to be separate subgroups, within which traditional estimation approaches can be applied. The area-specific sample size is often fixed and known.

Unplanned areas exist if the area membership is not taken into account within the sampling design. Thereby, the area-specific sample sizes are random.

Non-sampled areas are areas from which no sample units have been drawn. This is the case when the areas of interest are unplanned a priori, which results in a random sample size in single (or all) areas. Therefore, in extreme cases, certain areas are non-sampled.

3 Specification and evaluation of needs and resources

3.1 Definition of target areas

Description

The definition of the target areas significantly impacts all following steps connected to the production of small area estimates. It may seem desirable to estimate at a level which is as small as possible. However, decreasing the size of the target areas increases the extent and precision of the auxiliary information required for the estimation. Therefore, the aim should be to find a target level with a granularity that both meets the users' requirements and is supported by the data availability (Tzavidis et al., 2018). In addition, the question of whether the areas of interested are planned or unplanned areas is of vital importance, as it affects the required methodology.

Moreover, it should be decided whether estimates for superordinate geographical or content-related areas are of interest in addition to estimates for the target geography. With regard to the coherence of estimates, it may be desirable that an aggregation of subgroup-specific estimates is equivalent to the estimate for the superordinate area. This would require so-called benchmarking approaches described in Section 4.5.

Options

- Define planned areas only, which correspond to the strata of a sample survey suitable for the estimation.
- Define different target areas and compare the feasibility of estimation at each level.parameters: related to the distribution of target variables in the population,
- Define areas in accordance with the respective user needs.

Alternatives

(A) At first, define a target area at a sufficiently high level of aggregation and continue with further disaggregated levels after the estimation at the respective level has proven to be feasible.

(B) Directly target a level that appears to be sufficiently included in the design of the sample survey and offers enough auxiliary information to be used for the estimation.

(C) Define the target geography straightforward considering user needs only and regardless of the data availability and granularity.

3.2 Definition of target parameter

Description

In addition to the target areas, the definition of the target parameter to be estimated is also of vital importance. The target parameter needs to be well-supported by the available data. An increasing complexity of the indicators of interest simultaneously increases the granularity of the data that are needed for the estimation (Tzavidis et al., 2018).

In this context, one has to differentiate between linear and non-linear target parameters. Linear parameters comprise totals, means and proportions. The estimation of these indicators as well as the estimation of their variances is relatively uncomplicated. In contrast, indicators such as the at-risk-of-poverty rate, the at-risk-of-poverty threshold, the Gini coefficient or percentiles of income distributions are non-linear functions of survey variables. Often, their estimation is not possible using the common small areas estimation approaches without further adaptation (Münnich et al., 2013). Moreover, access to microdata in the form of a census or register might be necessary for a successful estimation. In addition, approximations are needed in order to estimate the variance of these indicators. Therefore, these parameters should only be aimed at if they are well-supported by the data availability.

Options

- Define the target parameter in consideration of the data situation.
- Define the target parameter in consideration of user needs.

Alternatives

(A) Define the target parameter in consideration of the current data situation and granularity.

(B) Define the target parameter in consideration of certain user needs which are compatible with the data situation.

(C) Define the target parameter straightforward considering user needs only and regardless of the data availability and granularity.

3.3 Distinction of sampling designs

Description

The estimation of the parameter of interest for small areas is usually based on the use of survey data. The sampling designs employed by the individual national statistical institutes in different social surveys are very diverse, which implies considerably different conditions for the application of small area estimation techniques and the choice of the estimation approach across member states.

Especially when using design-based estimation approaches, the sampling design has a considerable impact on the accuracy of the resulting estimators. For a detailed overview on the influence of sampling designs on small area estimation, the reader is referred to Burgard et al. (2016).

It is important to examine to what extent the small areas of interest are covered by the sample to be used and whether observations are available for every target area. In the case of two-stage sampling designs, it might often be the case that the areas of interest have been used as primary sampling units. Hereby, the sampling fraction of secondary sampling units within sampled primary units, i.e. the areas of interest in this case, tends to be notably higher than the sampling fraction of primary sampling units in simple random sampling or stratified random sampling approaches. This would result in a large number of non-sampled areas of interest, i.e. the target areas themselves are either sampled to a comparatively large extent or not sampled at all. In this case, the small area estimation will rely heavily on the underlying model assumptions (Tzavidis et al., 2018), which has to be considered at the stage of the implementation.

Options

- Identify the sampling design used.
- Identify the level of primary and secondary sampling units (if any).
- Identify the target areas in relation to the stages of the sampling design.
- Examine the distribution of the sample size across areas.

Alternatives

(A) Implement all options mentioned above and identify the relation between the target areas and the primary and secondary sampling units of the underlying sampling design.

(B) Implement certain options only.

(C) Disregard the underlying sampling design.

3.4 Definition of data availability

Description

Small area estimation approaches heavily rely on the availability of auxiliary information in the form of further survey data, administrative registers or census data. These sources contain variables that might be correlated with the target variable and thus might be highly suitable to be used for the stabilisation of the estimation. Therefore, a differentiation between area-level and unit-level auxiliary data has to be made. Access to unit-level auxiliary data might be challenging but necessary for non-linear target parameters.

Options

- Compile unit-level and/or area-level auxiliary information.
- Examine the quality of the data.
- Examine the explanatory power of the auxiliary variables, e.g. using model selection criteria such as the Akaike Information Criterion (AIC) and the conditional AIC criterion (see for example Saefken et al., 2014).
- If estimates are used as area-level auxiliary information, estimate their variance.

Alternatives

(A) Compile auxiliary data that strongly supports the target parameter. Examine the quality and explanatory power of the auxiliary variables.

(B) Compile auxiliary data considering the target parameter but without examining their quality and explanatory power.

(C) Compile auxiliary data without considering the target areas and the target parameter.

4 Implementation of small area estimation

4.1 Choice of small area estimation approach

Description

The traditional way of obtaining population estimates used within national statistical institutes is to use design-based methods. That is, the sample is taken to be a random subset of a finite and fixed population. The observed values themselves are, therefore, not considered to be random variables. Only the inclusion of a unit of the population into the sample is random. Based on this randomization, which is fully defined by the survey design, properties such as design-unbiasedness and design-consistency can be postulated.

A design-unbiased estimator is the Horvitz-Thompson estimator proposed by Horvitz and Thompson (1952). Unfortunately, the variance of this estimator tends to be quite large, especially in the case of small sample sizes within an area or domain. This problem is aggravated if sample sizes are not fixed but a random number themselves. In this case, the Hájek estimator proposed by Hájek (1971) is preferable. See also p. 182 of Särndal et al. (1992) and Dorfman and Valliant (1997) for a deeper discussion of the advantages of the Hájek estimator.

If some additional information on the population is available at population level, that is, not only for the sample but also for the rest of the population, model-assisted estimation methods can be applied, for example the GREG estimator proposed by Cassel et al. (1976). Here, a model is used to reduce the unexplained variability of the variable of interest. The resulting predictor typically has a variance that is lower or equal to the one of the Horvitz-Thompson estimator. Still, for applying this approach, the sample size needs to be large enough in the domains and areas of interest. The necessary sample size is determined by the level of precision required for the produced population estimates. If the variance estimates for the GREG estimates indicate too high a variability, the design-based and model-assisted approaches are not suited for the production of these estimates. In this case model-based small area estimation methods should be applied.

Oftentimes, the areas and domains of interest are not planned in the design phase of a survey. Therefore, it is common to observe that some areas have no or only a negligible number of sampled units. This leads to the necessity to produce estimates for these areas or domains without having any area/domain-specific information on the variable of interest. In this case, synthetic estimation procedures can be applied. The synthetic estimation methods produce their estimates solely from a model prediction point of view. That is, a model is fitted to the data and a prediction that is unconditional on the variable of interest is used. This typically leads to very low variability in the predictions. On the other hand, if the model does not perfectly describe the variable of interest, which should generally be assumed to be the case, the predictions can lead to large biases.

The empirical best predictors try to use *the best of both worlds* by conditioning the prediction on the variable of interest, and thus using a convex combination of a direct and a synthetic estimation procedure for the population values of interest. Among many, there are the two standard estimators, the Fay-Herriot estimator (Fay and Herriot, 1979) as an area-level estimator and the Battese-Harter-Fuller estimator (Battese et al., 1988) as a until-level estimator. These empirical best predictors have shown to have a much lower variability than model-assisted estimation procedures at the cost of accepting some possible bias under the design randomization. However, if the model is adequate, the bias tend to remain moderate. In contrast, if the model does not have a large explanatory power, bias tends to be dominant in the mean squared error (MSE) of the empirical best predictor.

Note that depending on the distribution of the dependent variable, the (assisting) model has to be chosen accordingly. For continuous data typically linear models with Gaussian errors are used. For binary dependent variables often times logit models are preferred, see González-Manteiga et al. (2007) and Lehtonen and Veijanen (1998), and for multinomial data multinomial logit models like in López-Vizcaíno et al. (2015).

Options

- Use design-based estimation methods.
- Use model-assisted estimation methods.
- Use synthetic estimation methods.
- Use empirical best prediction methods

Alternatives

(A) If sample sizes are large enough and the necessary auxiliary information is given for the population, the use of a model-assisted estimator is a good approach.

(B) If the precision in (A) is not good enough, an empirical best prediction approach should be applied.

(C) If no sampled units are available in some areas of interest, the only remedy left is to use a synthetic estimation procedure.

4.2 Unit- versus area-level empirical best prediction

Description

When estimating using empirical best predictors, two major approaches exist. First, one could use a unit-level estimator, such as the BHF estimator (Battese et al., 1988). This estimator takes as input values the unit-level characteristics of the sampled units. Then, a statistical model, in this case a linear mixed model, is fitted on the sampled data and extrapolated using the known population values for the auxiliary variables in the model. This, however, requires the knowledge of additional auxiliary variables in the population. These variables are available from registers in some countries. In the case of using a linear (mixed) model, aggregated totals of the auxiliary variables on area or domain level are sufficient. In the case of non-linear models, e.g. a logistic (mixed) model, the auxiliary information has to be available for each single unit in the population separately. This imposes a very high degree of dependence on available information for the researcher that applies small area estimation methods on unit level.

Alternatively, area-level models such as the FH estimator (Fay and Herriot, 1979) can be applied. These area-level models need both the variable of interest and the auxiliary variables to be aggregated totals or means on area level only. This information is typically much easier to gain. Even if there are registers for the population, for confidentiality reasons, these may not in fact be usable in many cases.

In general, if the unit-level information is available, one can suspect to obtain more precise predictions with unit-level estimators than with area-level estimators. Vogt (2008) describes a situation where the opposite is the case. Therefore, the choice between area- and unit-level estimators depends mostly on the availability of data.

Options

- Use unit-level empirical best predictors
- Use area-level empirical best predictors

Alternatives

(A) If unit-level data is available for the auxiliary variables for all units in the population unit-level estimators should be used.

(B) If the auxiliary variables are only known at aggregated level for the areas or domains, but not on unit-level for the out-of-sample units, unit-level estimators may still be applied based on linear mixed models.

(C) If data is only available as aggregates on area-level, then area-level empirical best predictors should be used.

4.3 Auxiliary variables measured with (out) errors

Description

One of the core assumptions of the estimation methods described above is that the auxiliary variables are measured without errors. If the auxiliary variables are deduced from registers, the measurement error is typically taken to be negligible. However, registers are also prone to errors, and these could lead to reduced precision of the estimators.

In the case where measurement errors are observed, this should be accounted for in the estimator. For unit-level models with measurement errors in the auxiliary variables, there has not been much research done yet to our knowledge. For the case of area-level models Ybarra and Lohr (2008) proposed an extension to the FH model allowing for measurement errors in the auxiliary variables. Burgard et al. (2019) propose an analytical MSE estimator for the case where the measurement errors are normally distributed, which is a plausible assumption due to the central limit theorem.

Options

- Use estimators not accounting for measurement errors.
- Use estimators accounting for measurement errors.

Alternatives

(A) If there are measurement errors in the auxiliary variables, then the estimator should be chosen such that they can be accounted for.

(B) If there is no appropriate estimator accounting for measurement errors, e.g. unit-level empirical best predictor, then the estimates have to be taken with caution.

(C) The auxiliary data contain errors, but these are ignored without further mentioning.

4.4 Testing for area-specific random effects

Description

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 stabilization of the estimation. Sometimes the synthetic part of the empirical best predictors have very high explanatory power. In other words, in a linear mixed model the area-specific random effects variance tends against zero. If it is zero, a simple linear regression model would suffice for the area predictions. Therefore, it is of interest to test for the need of the more complex linear mixed model.

Datta et al. (2011) and Molina et al. (2015) propose a method for testing for the need of the random effect. Molina et al. (2015) provide a new MSE estimator specifically suitable for the case of a small number of areas. An exemplary test that follows the recommendations of this article is performed in Section 6.

Options

- Use the empirical best predictor without accounting for possibly neglectable random effects variance.
- Test for the need for a random effect for the empirical best predictor.

Alternatives

(A) Test for the need for a random effect
  - If the random effect is significant, then use EBPs.
  - If the random effect is not significant, then either use EBPs or synthetic estimators.

(B) Simply apply the empirical best predictor based on the linear mixed model.

4.5 Assurance of coherence

Description

An important aspect in official statistics also contained in the *European Statistics Code of Practice* **(EUROPEAN STATISTICAL SYSTEM COMMITTEE, 2017)** is the assurance of internal coherence and consistency. In the context of small area estimation, it is of interest that the small area estimates satisfy the benchmarking property, i.e. they are vertically coherent with estimates at larger scale. Vertical coherence describes the aggregation of estimated values at small scale to superordinate estimated values in terms of hierarchical levels.

The ratio benchmarking approach described by Rao and Molina (2015, p. 159 f.) is a simple method to ensure that small-scale estimates $\hat{\mu}_d$ add up to a reliable superordinate direct estimate $\hat{\mu}_+^{Dir} = \sum_{l=1}^{m} W_l \hat{\mu}_l^{Dir}$ with $W_l = N_l/N$ as known proportion of units in area l and provided that all areas are sampled $(\sum_{l=1}^{m} W_l = 1)$.

Thereby, each area-specific mean estimate $\hat{\mu}_d$ is multiplied by a common adjustment factor $\hat{\mu}_+^{Dir}/\sum_{l=1}^{m} W_l \hat{\mu}_l$. Thus, the ratio benchmark estimator is given by

$$\hat{\mu}_d^{RB} = \hat{\mu}_d(\sum_{l=1}^{m} W_l \hat{\mu}_l^{Dir} \Big/ \sum_{l=1}^{m} W_l \hat{\mu}_l)$$

and $\sum_{d=1}^{m} W_d \hat{\mu}_d^{RB} = \sum_{l=1}^{m} W_l \hat{\mu}_l^{Dir} = \hat{\mu}_+^{Dir}$ (Rao and Molina, 2015, p. 160) One drawback of this approach is that all area-specific estimates $\hat{\mu}_d$ are multiplied by one common adjustment factor regardless of the area-specific size and the precision of each estimate. In addition, the estimation is not design-consistent and the second-order unbiased MSE estimates are not available, as the MSE estimation is non-trivial.

An alternative method that avoids the mentioned shortcomings is the pseudo-EBLUP according to You and Rao (2002), also referred to as You-Rao estimator, which automatically satisfies the benchmarking property by including survey weights. The approach is described in Section A.3.3 in the appendix.

In addition to the pseudo-EBLUP, You and Rao (2003) derived a pseudo hierarchical Bayes estimator, which also utilises the design weights and automatically satisfies the benchmarking property. For detailed information on this approach, it is referred to (Rao and Molina, 2015) as well as to the underlying paper.

Options

- Utilisation of the ratio benchmarking approach.
- Utilisation of the You-Rao estimator (i.e. the pseudo-EBLUP).
- Utilisation of the pseudo hierarchical Bayes estimator.
- Disregard vertical coherence and do not use any approach.

Alternatives

(A) Utilisation of the You-Rao estimator (i.e. the pseudo-EBLUP).

(B) Utilisation of the ratio benchmarking approach.

(C) Disregard vertical coherence and do not use any approach.

5 Model diagnostics and evaluation

5.1 Model diagnostics

Description

Small area estimation approaches largely depend on the quality of the utilized models. Incorrect specifications may lead to a strong bias of the estimators and thus to a misleading basis of information in applications. Therefore, after implementing the small area estimation, the utilised model needs to be cautiously examined and checked for a violation of the underlying assumptions and a potential bias. For a comprehensive overview of model diagnostic techniques the reader is referred to Pfeffermann (2013).

When implementing model-based small area estimation techniques, a bias is accepted for the sake of a decreased variance of the estimator. However, it is preferable that this bias resulting from the model-based prediction is as small as possible. Brown et al. (2001) suggest a graphical diagnostic for a visual examination of this potential bias. It is assumed that if the model-based estimators were unbiased estimates of the true parameter of interest, the equally unbiased direct estimates would fluctuate randomly around the values of the corresponding model-based estimates. Hence, a scatter plot, in which the relation between direct and model-based estimates is illustrated, would be evenly distributed around the bisector of the plot so that a linear regression would yield the parameter $\beta_0 = 0$ for the intercept and $\beta_1 = 1$ for the slope.

In addition, the analysis of the residuals is a further important diagnostic tool. It has to be examined whether the assumption of normal distribution of the sampling errors and the random effects applies. A normal quantile-quantile plot (Q-Q plot), which illustrates the relation between the sample quantiles of the standardised residuals and the theoretical quantiles of a normal distribution, is a graphical diagnostic tool that tests this assumption. If the standardised residuals are normally distributed, they will lie on a straight line. If the normal Q-Q plot for the standardised residuals of the model-based estimation indicates a slight deviation from the normal distribution, an additional Shapiro-Wilk test for normality is a further tool to check whether the null hypothesis of normality can be rejected. In this context, both unit-level and area-level residuals should be taken into account.

Furthermore, it is assumed that the sampling errors have a constant variance. A further diagnostic graphical tool to check this requirement is to plot the residuals against the model-based estimates, i.e. the fitted values, in order to detect systematic patterns.

Options

- Examine whether a potential bias exists, e.g. by outlining the relation between the direct and the model-based estimates.
- Examine whether the assumption of normal distribution of the sampling errors and random effects applies, e.g. by means of a Q-Q plot.
- Examine whether the assumption of a constant variance of the sampling errors applies, e.g. by outlining the relation between the residuals and the model-based estimates.

Alternatives

(A) Implement all listed model diagnostic tools in order to ensure that the model estimation is not harmed by any violation of the model assumptions.

(B) Implement those diagnostic tools that cover model assumptions which are likely to be violated given the present model.

(C) Disregard the listed model diagnostic tools altogether.

5.2 Evaluation of approaches

Description

The stage of model evaluation serves to find out whether the set of estimates can be considered to be of acceptable precision.

The mean squared error (MSE) is the most common measure to assess the uncertainty associated with the area-specific prediction under the model that has been assumed (Tzavidis et al., 2018). However, the calculation of an expression for the MSE is significantly more difficult than the derivation of the predictors themselves. Although it is not possible to specify a closed form solution for the MSE of the EBLUP, approximations may be used, which decisively depend on the normality assumption of the random effects and the sampling errors. A common analytic MSE estimator developed by Prasad and Rao (1990) consists of three components. These reflect the uncertainty resulting from the prediction of the random effects, the variability resulting from the fixed effects term and the uncertainty due to the estimation of the variance components. A further approach for the estimation of the MSE is the usage of a resampling method, such as the parametric bootstrap (Efron, 1982) or a jackknife approach.

The efficiency gain which can be achieved by the application of small area estimation methods may be analysed by calculating the ratio between the estimated MSE of the small area estimate and the variance of the direct estimate. A further insight may be gained by illustrating this ratio in relation to the area-specific sample size. In doing so, it may be detected whether an increasing number of observations per area results in a decreased improvement by using the small area estimation approach. This may be expected, as an increasing area-specific sample size tends to result in direct estimates with a comparatively low variance.

Furthermore, the methodology used for the estimation and its behaviour in different situations may be evaluated using a simulation study. Here, one must not confuse design-based and model-based simulation studies. In a design-based simulation study, samples are repeatedly drawn from a finite population according to one or several sampling designs. The parameters of interest and the auxiliary data are treated as fixed and the randomness results from the inclusion of the elements into the sample. This approach serves to evaluate the influence of the sampling designs and to validate the performance of the estimators under realistic conditions. In contrast, model-based simulation studies are based on a so-called data generating process. Each sample is a realisation of this process, where the randomness results from the underlying structure of the error term resulting from the process. This approach enables a control over the model characteristics and an assignment of the results to specific causes.

Options

- Estimate the MSE of the estimates using an analytical approach.
- Estimate the MSE of the estimates using a resampling approach.
- Evaluate the uncertainty of the estimation using a simulation study.

Alternatives

(A) Perform an uncertainty assessment using the presented procedures and decide whether the estimates are satisfactory. If not, repeat the listed small area estimation processes under alternative conditions, such as altered target areas, target indicators, data sources or small area estimation approaches, until a final set of estimates of sufficient precision is found.

(B) Perform an uncertainty assessment using the presented procedures and decide whether the estimates are satisfactory. If not and if it is not possible to repeat the estimation process under alternative conditions, keep the set of estimates but clearly communicate to the data users that there is considerable uncertainty associated with the estimation in the case at hand.

(C) Disregard the process of uncertainty assessment.

6 Exemplary application

Section 6 is an exemplary application, giving guidance for the implementation of each single step in practice. This example is based on the synthetic population data set AMELIA, see Merkle and Münnich (2016) and Burgard et al. (2017).

Specification and evaluation of needs and resources

The target areas in this example are the 1,592 municipalities of the synthetic population AMELIA. The granularity of the data at this level supports an estimation of sufficient precision. In addition, it is assumed that estimates at a higher aggregation level have proven to be feasible.

The target parameter is the share of persons at risk of poverty or social exclusion (AROPE) (cf. Eurostat, 2018). As the sub-indicators of the AROPE indicator are additionally included in the dataset, it can be concluded that both the data situation and granularity support the estimation of the target parameter.

As sampling design, a stratified random sampling of persons, which are the primary sampling units, is chosen in this case, using ten age classes as stratification criteria. No secondary sampling units are included in this design. As the target areas (municipalities) are not related to the stages of the sampling design, they are unplanned areas, making it likely to have substantial amounts of non-sampled areas. And indeed, with a sampling fraction of 0.16%, the sample size per area ranges from 0 to 418 units with 84 of the 1,592 municipalities (5.27%) not being sampled at all.

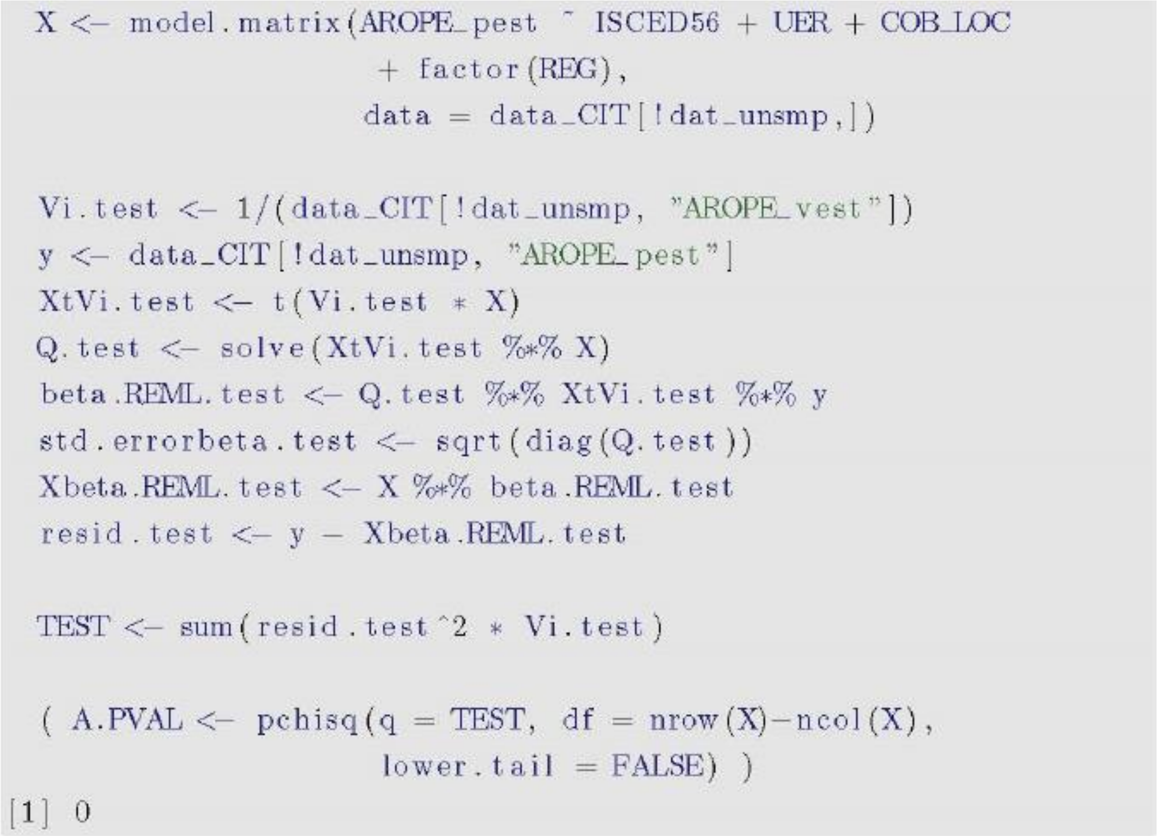
In the current application, it is assumed that no unit-level information strongly supporting the target parameter AROPE are available. As auxiliary area-level information for the estimation, the share of native-born persons (**COB_LOC**), the share of persons with an ISCED-level of at least 5 (**ISCED56**), the share of unemployed persons (**UER**) and the region within AMELIA (**REG**) is used. Their explanatory power appears to be sufficient, according to simple correlation tests on the target parameter.

Implementation of small area estimation

The sampling fraction of 0.16% and the resulting low sample sizes of maximum 84 units is insufficient for either design- or model-assisted estimation procedures of sufficient precision. In addition, the rather high fraction of 5.27% non-sampled areas makes a design-based estimation unfeasible in these areas.

In this case, an empirical best prediction approach should be chosen instead, where estimators for sampled areas are derived by combining direct and synthetic estimation procedures. In non-sampled areas, however, the prediction will be made using synthetic estimation approaches. Due to the extensive availability of auxiliary information on the area-level, the Fay-Herriot estimator seems appropriate. Thereby, it is not assumed that the auxiliary variables are measured with error.

The significance of a Random effect can be measured according to Molina and Marhuenda (2015), Datta et al. (2011):



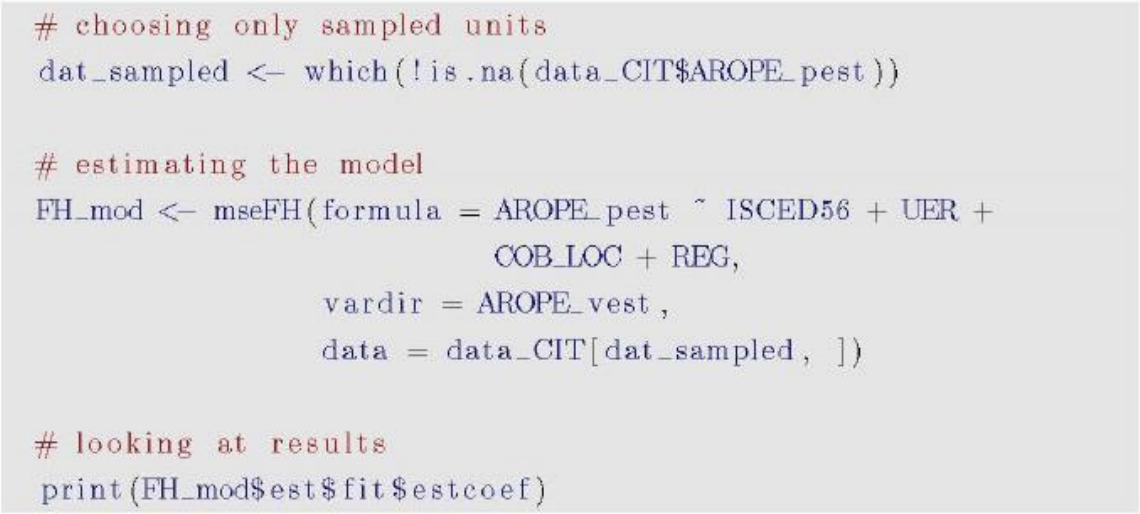
```
 X <- model.matrix(AROPE_pest ~ ISCED56 + UER + COB_LOC
                        + factor(REG),
                      data = data_CIT[!dat_unsmp,])

 Vi.test <- 1/(data_CIT[!dat_unsmp, "AROPE_vest"])
 y <- data_CIT[!dat_unsmp, "AROPE_pest"]
 XtVi.test <- t(Vi.test * X)
 Q.test <- solve(XtVi.test %*% X)
 beta.REML.test <- Q.test %*% XtVi.test %*% y
 std.errorbeta.test <- sqrt(diag(Q.test))
 Xbeta.REML.test <- X %*% beta.REML.test
 resid.test <- y - Xbeta.REML.test

 TEST <- sum(resid.test^2 * Vi.test)

 ( A.PVAL <- pchisq(q = TEST, df = nrow(X)-ncol(X),
                        lower.tail = FALSE) )
[1] 0
```

The resulting p-value is numerically 0. Therefore, the null hypothesis that the random effects variance is equal to zero can be rejected at the 5% significance level. We should use an empirical best predictor and not a synthetic estimator.

Model diagnostics and evaluation

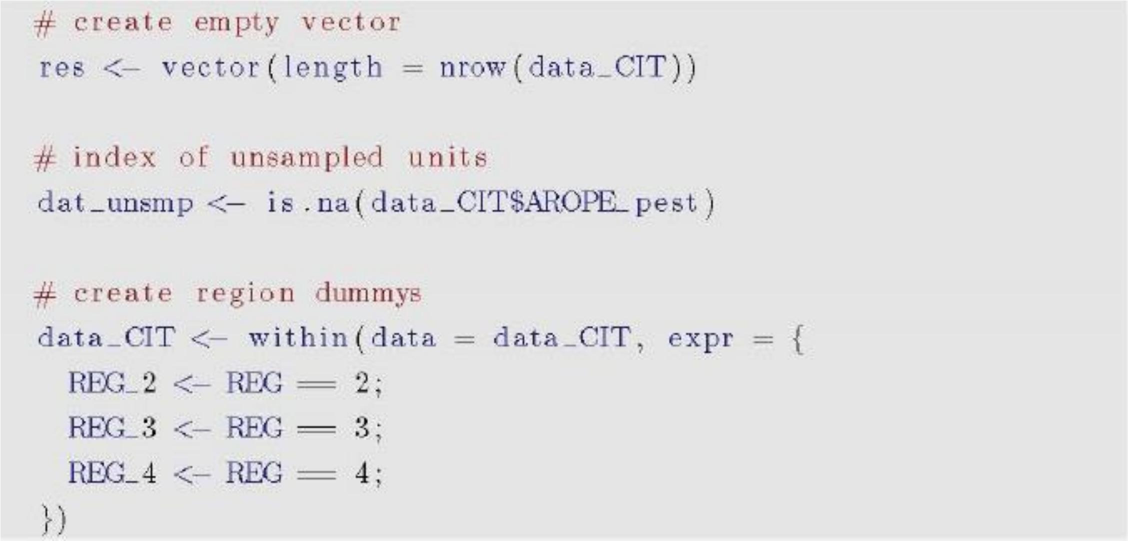
The direct Horvitz-Thompson estimator of the AROPE rate is given by the variable **AROPE_pest**. Of course, this estimator can only be derived for sampled areas. Its variance is designated by **AROPE_vest**. As already mentioned, the auxiliary information includes the share of native-born persons (**COB_LOC**), the share of persons with an ISCED-level of 5 or higher (**ISCED56**), the share of unemployed persons (**UER**) and the region within AMELIA (**REG**).

The model according to Fay and Herriot (1979) is estimated using only sampled areas:

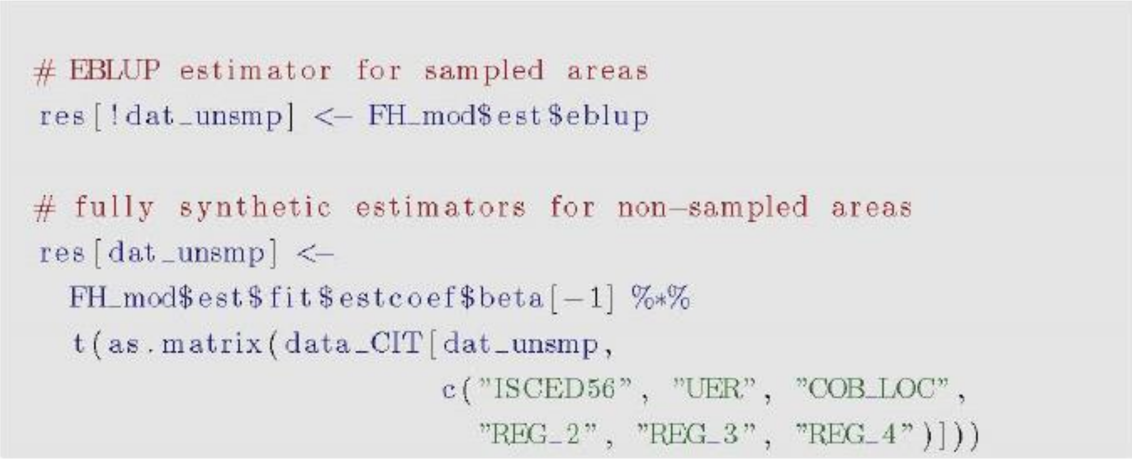


```
# choosing only sampled units
dat_sampled <- which(!is.na(data_CIT$AROPE_pest))

# estimating the model
FH_mod <- mseFH(formula = AROPE_pest ~ ISCED56 + UER +
                          COB_LOC + REG,
                vardir = AROPE_vest,
                data = data_CIT[dat_sampled, ])

# looking at results
print(FH_mod$est$fit$estcoef)
```

Subsequently, the EBLUP estimators are derived. These are composite EBLUP estimates for the sampled areas, whereas fully synthetic estimators are assigned to the non-sampled areas:



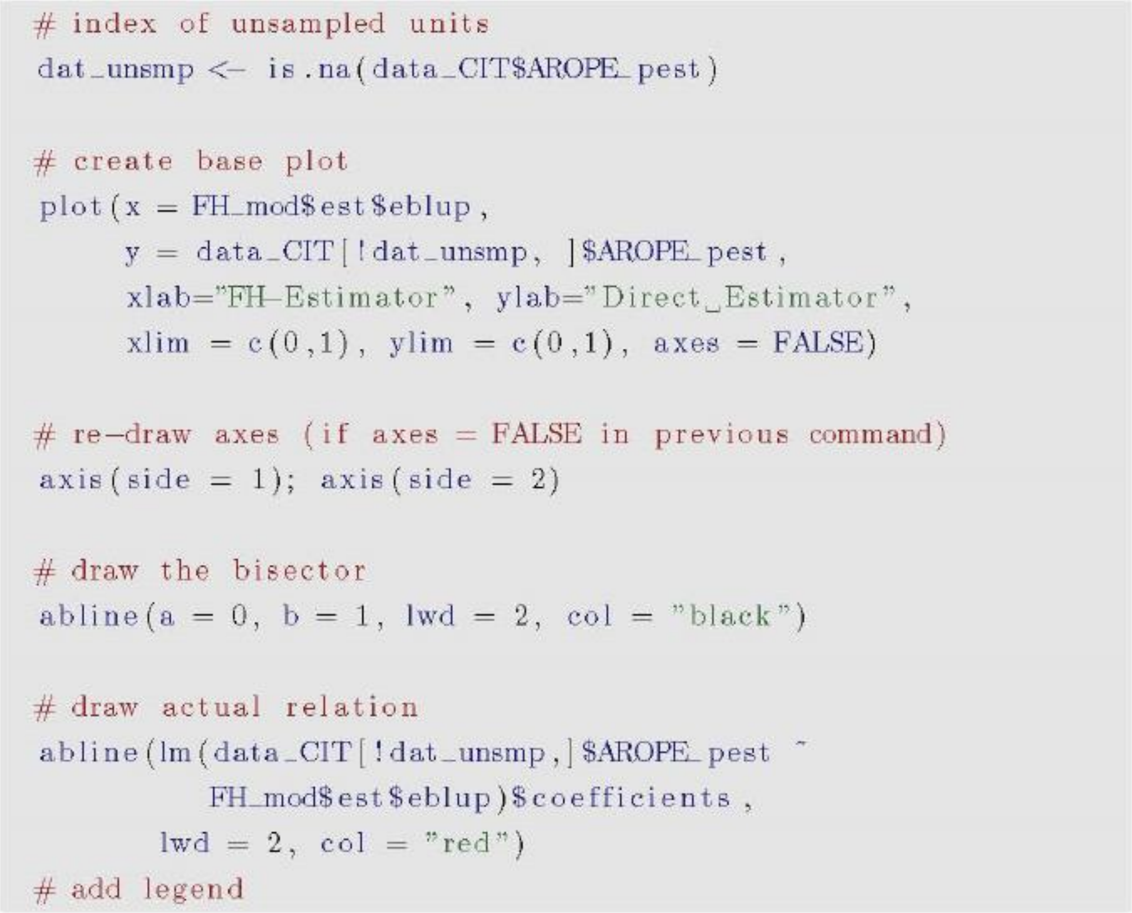
```
# create empty vector
res <- vector(length = nrow(data_CIT))

# index of unsampled units
dat_unsmp <- is.na(data_CIT$AROPE_pest)

# create region dummys
data_CIT <- within(data = data_CIT, expr = {
  REG_2 <- REG == 2;
  REG_3 <- REG == 3;
  REG_4 <- REG == 4;
})
```



```
# EBLUP estimator for sampled areas
res[!dat_unsmp] <- FH_mod$est$eblup

# fully synthetic estimators for non-sampled areas
res[dat_unsmp] <-
  FH_mod$est$fit$estcoef$beta[-1] %*%
  t(as.matrix(data_CIT[dat_unsmp,
                        c(”ISCED56”, ”UER”, ”COB_LOC”,
                          ”REG_2”, ”REG_3”, ”REG_4”)]))
```

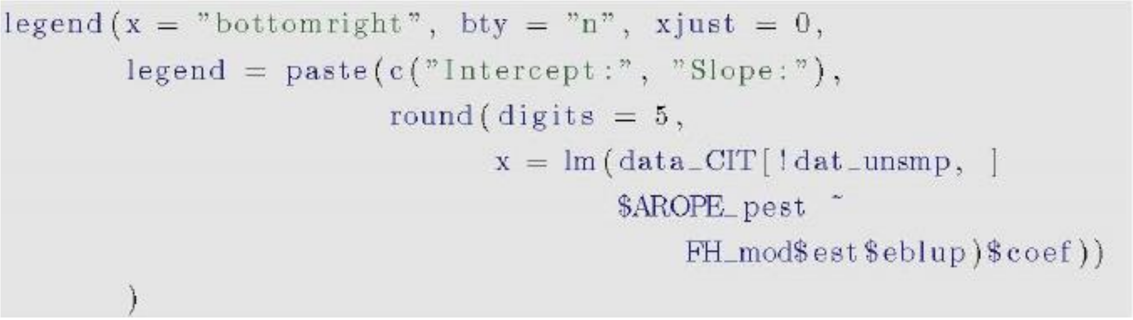
Model diagnostics

It is important to check the model for a violation of the model assumptions and a potential bias. In order to examine whether a substantial bias exists, the graphical diagnostic suggested by Brown et al. (2001) is produced and illustrated in Figure 1: Relation between direct estimator and Fay-Herriot



```
# index of unsampled units
dat_unsmp <- is.na(data_CIT$AROPE_pest)

# create base plot
plot(x = FH_mod$est$eblup,
     y = data_CIT[!dat_unsmp, ]$AROPE_pest,
     xlab=”FH-Estimator”, ylab=”Direct␣Estimator”,
     xlim = c(0,1), ylim = c(0,1), axes = FALSE)

# re-draw axes (if axes = FALSE in previous command)
axis(side = 1); axis(side = 2)

# draw the bisector
abline(a = 0, b = 1, lwd = 2, col = ”black”)

# draw actual relation
abline(lm(data_CIT[!dat_unsmp,]$AROPE_pest ~
          FH_mod$est$eblup)$coefficients,
       lwd = 2, col = ”red”)
# add legend
```

estimator (below).



```
legend(x = ”bottomright”, bty = ”n”, xjust = 0,
       legend = paste(c(”Intercept:”, ”Slope:”),
                      round(digits = 5,
                            x = lm(data_CIT[!dat_unsmp, ]
                                   $AROPE_pest ~
                                      FH_mod$est$eblup)$coef))
       )
```

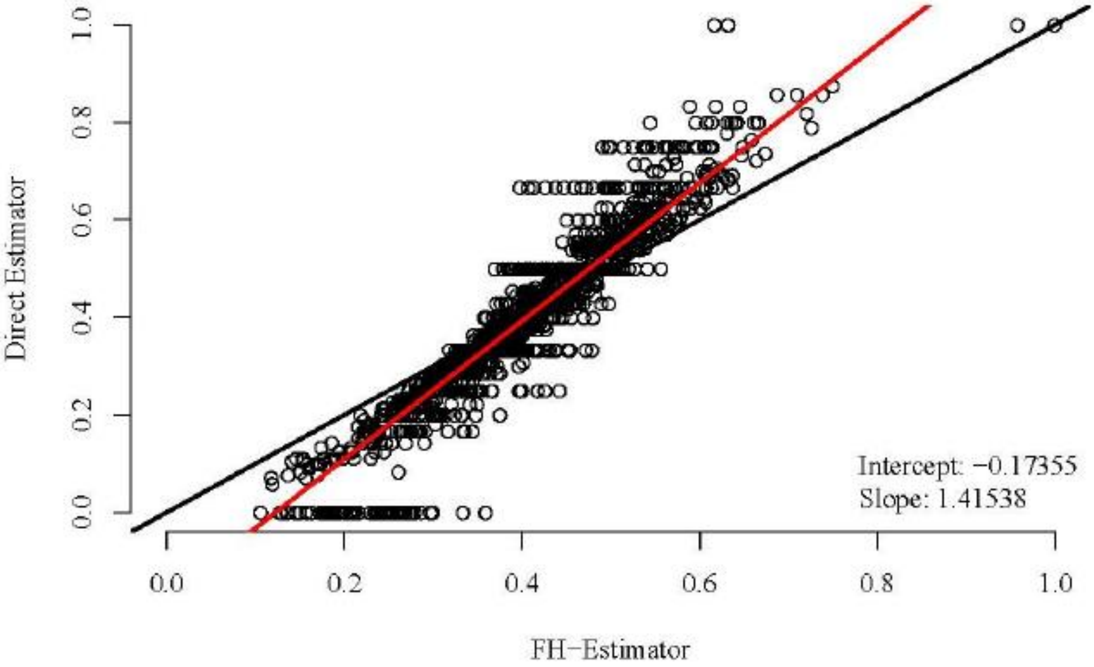


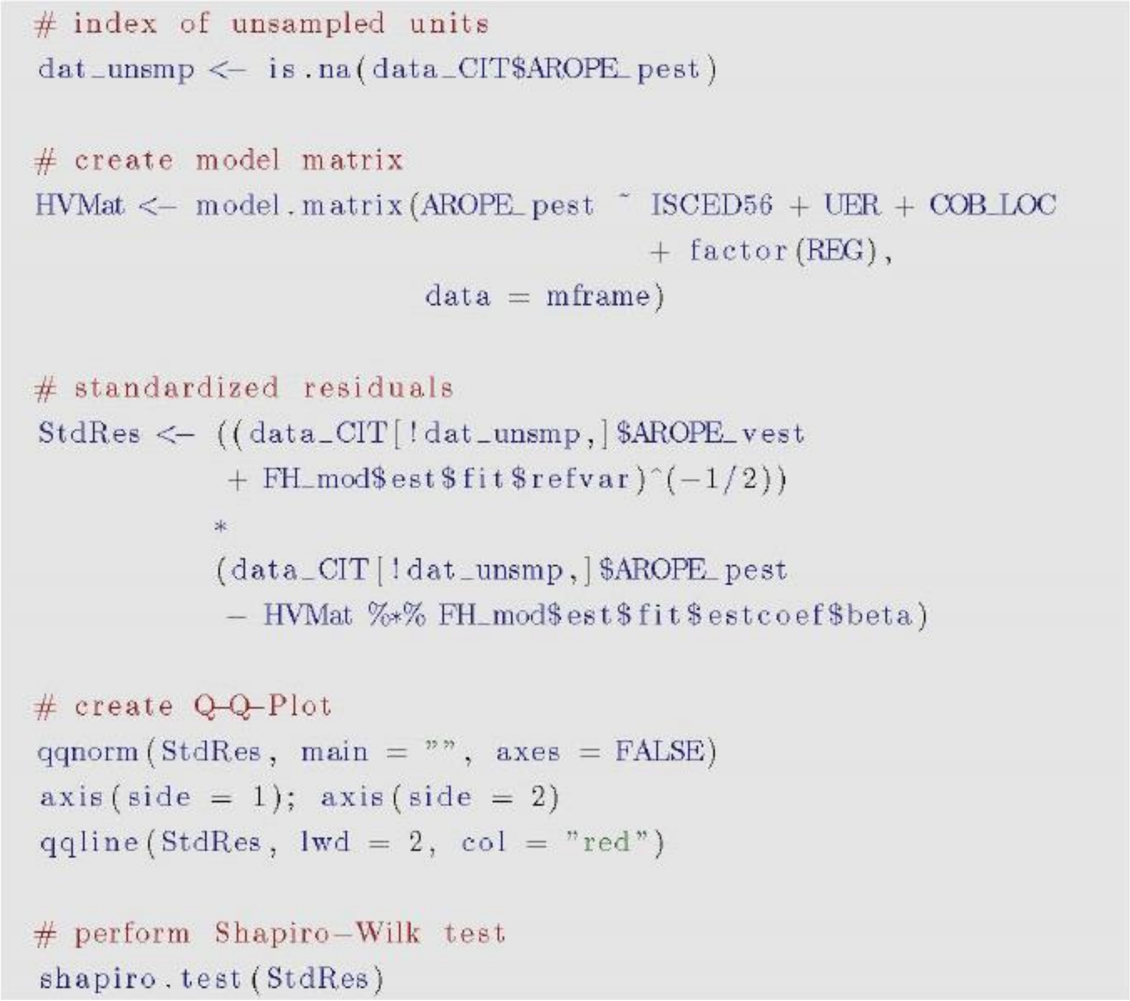
Figure 1: Relation between direct estimator and Fay-Herriot estimator

Source: see section 7 - References

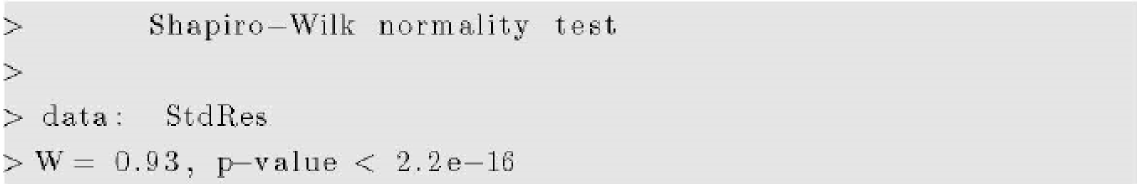
It is assumed that the equally unbiased direct estimates would fluctuate randomly around the value of the model-based estimate, if the latter were unbiased estimates of the true value. The produced scatter plot would be evenly distributed around the biscetor of the plot. Thereby, a linear regression would yield the parameter $\beta_0 = 0$ for the intercept and $\beta_1 = 1$ for the slope. In Figure 1: Relation between direct estimator and Fay-Herriot estimator the bisector is delineated in black, whereas the regression line is red. The intercept of the linear regression is $\beta_0 = 0.174$ and the parameter for the slope is $\beta_1 = 1.415$.

Therefore, model-based estimators of areas with a high AROPE rate tend to be smaller than the direct estimators of these regions. In areas with a low AROPE rate, however, the model predictions are higher than the direct estimators. It seems as if these models lead to biased estimates and in this case it would be recommended to re-specify the respective model.

In addition, it is recommended to check whether the assumption of normal distribution of the sampling errors applies. A graphic diagnostic tool that tests this assumption is a normal Quantile-Quantile-Plot (Q-Q-Plot), which illustrates the relation between the sample quantiles of the



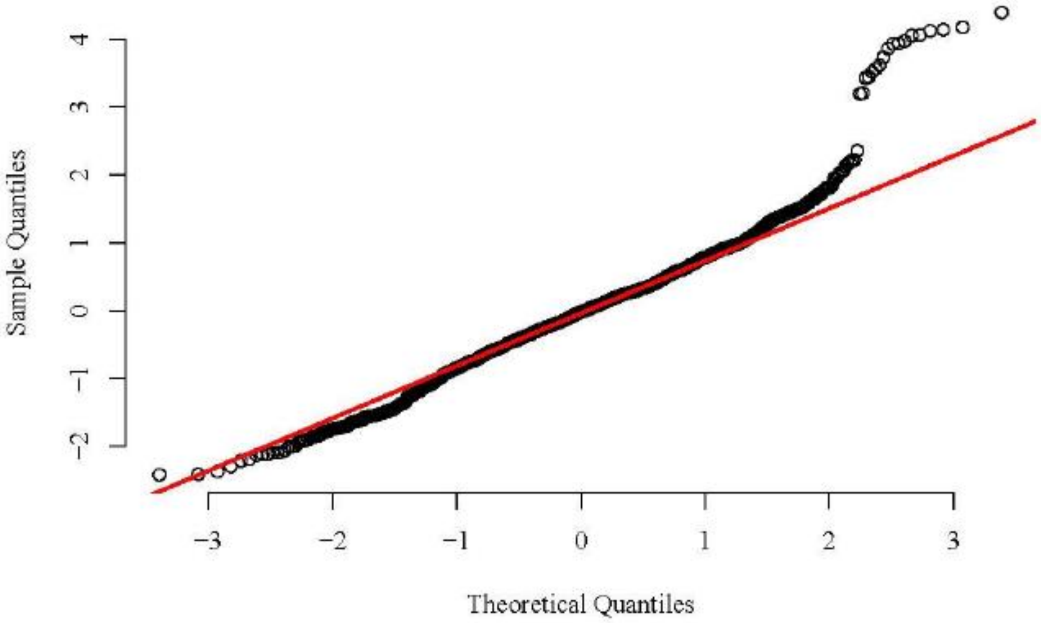
```
# index of unsampled units
dat_unsmp <- is.na(data_CIT$AROPE_pest)

# create model matrix
HVMat <- model.matrix(AROPE_pest ~ ISCED56 + UER + COB_LOC
                                   + factor(REG),
                      data = mframe)

# standardized residuals
StdRes <- ((data_CIT[!dat_unsmp,]$AROPE_vest
            + FH_mod$est$fit$refvar)^(-1/2))
           *
           (data_CIT[!dat_unsmp,]$AROPE_pest
            - HVMat %*% FH_mod$est$fit$estcoef$beta)

# create Q-Q-Plot
qqnorm(StdRes, main = "", axes = FALSE)
axis(side = 1); axis(side = 2)
qqline(StdRes, lwd = 2, col = "red")

# perform Shapiro-Wilk test
shapiro.test(StdRes)
```



```
>        Shapiro-Wilk normality test
>
> data:   StdRes
> W = 0.93, p-value < 2.2e-16
```

standardised residuals $r_d = (\psi_d + \sigma_v^2)^{-\frac{1}{2}}(\hat{\mu}_d^{Dir} - \overline{X}_d^T\beta)$ and the theoretical quantiles of a normal distribution. This plot has been produced and is shown in Figure 2. Furthermore, the Shapiro-Wilk test has been conducted as a further useful metric to test for normality.

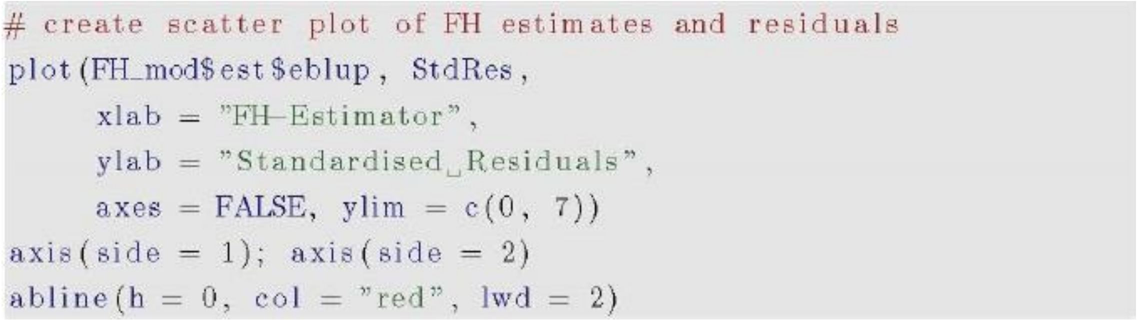
Figure 2: Normal Q-Q-Plot of standardised residuals



Source: see section 7 - References

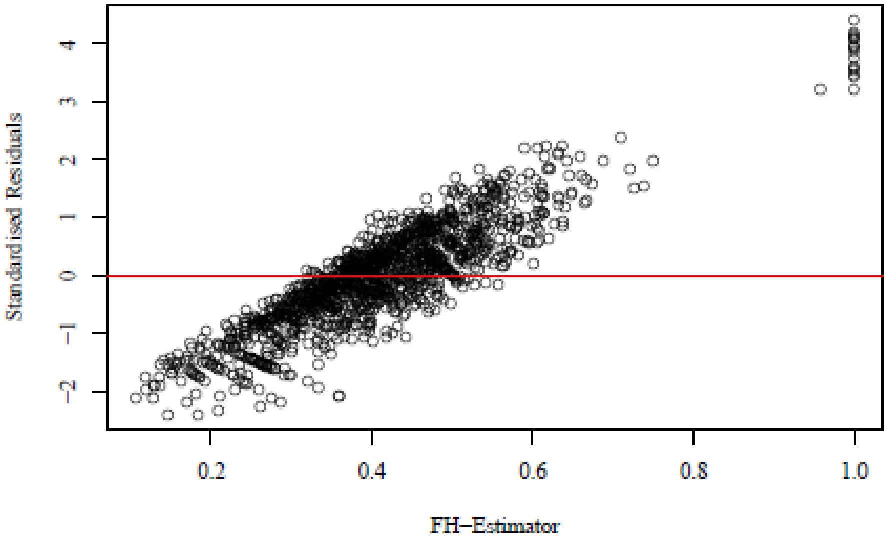
The Q-Q-Plot in Figure 2: Normal Q-Q-Plot of standardised residuals indicates a notable deviation from the normal distribution. Moreover, the additional Shapiro-Wilk test for normality suggests that the distribution of standardised residuals significantly differs from a normal distribution.

A further model-assumption is that the sampling errors have a constant variance. This can be checked by plotting the residuals against the model-based estimates, which has been done and is shown in Figure 3: Fay-Herriot estimator in relation to standardised residuals.



```
# create scatter plot of FH estimates and residuals
plot(FH_mod$est$eblup, StdRes,
     xlab = ”FH–Estimator”,
     ylab = ”Standardised␣Residuals”,
     axes = FALSE, ylim = c(0, 7))
axis(side = 1); axis(side = 2)
abline(h = 0, col = ”red”, lwd = 2)
```

Figure 3: Fay-Herriot estimator in relation to standardised residuals

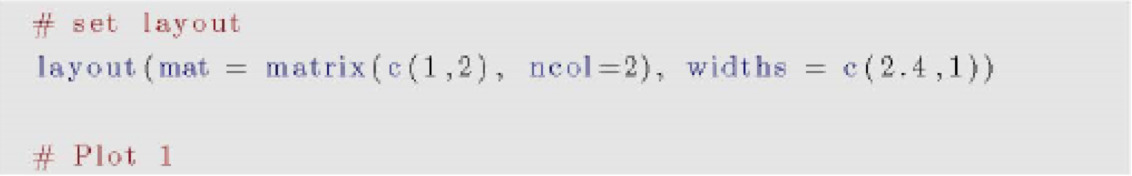


Source: see section 7 - References

Here however, a connection can be observed between the residuals and the Fay-Herriot estimates. This again suggests to re-specify the respective model due to the observed violations of the model assumptions.

Evaluation of approaches

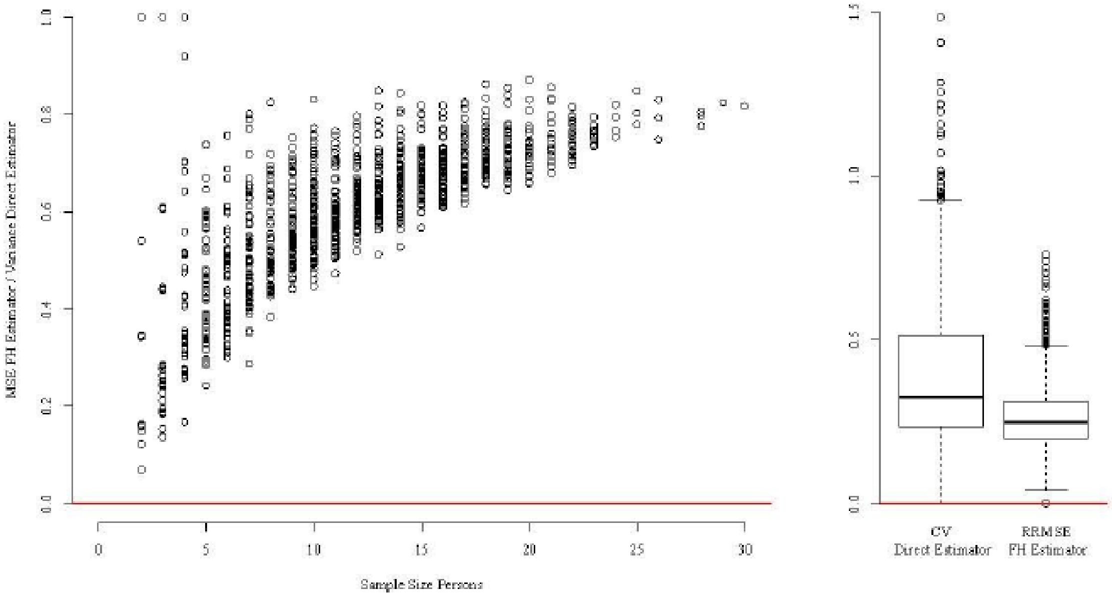
The efficiency gain, which can be achieved by the application of the Fay-Herriot estimator, is illustrated in Figure 4: Efficiency gain through the Fay-Herriot estimator. Thereby, the area-specific sample size is plotted against the ratio between the mean square error (MSE) of the Fay-Herriot estimate and the variance of the direct AROPE estimate. In addition, two box plots visualize the relation between the coefficient of variation of the direct estimator and the relative root mean square error of the Fay-Herriot estimator over all areas observed.



```
# set layout
layout(mat = matrix(c(1,2), ncol=2), widths = c(2.4,1))

# Plot 1
```

```
plot(x = cbind(data_CIT[!dat_unsmp,]$nh),
     y = FH_mod$mse/data_CIT[!dat_unsmp,]$AROPE_vest,
     xlab = "Sample␣Size␣Persons",
     ylab = "MSE␣FH␣Estimator␣/␣Variance␣Direct␣Estimator",
     ylim = c(0, 1),
     xlim = c(0, 30),
     axes = FALSE)
axis(side = 1); axis(side = 2) # re-draw axes
abline(h = 0, col = "red", lwd = 2) # draw bottom-line

# Plot 2
boxplot(x = list(x = (sqrt(data_CIT[!dat_unsmp,]$AROPE_vest) /
                      data_CIT[!dat_unsmp,]$AROPE_pest),
                 y = (sqrt(FH_mod$mse) / FH_mod$est$eblup)),
        xaxt = "n", axes = FALSE)
# re-draw axes
axis(side = 2)
axis(side = 1, at = 1:2, tick = FALSE,
     labels = c("CV\n␣Direct␣Estimator",
                "RRMSE\n␣FH␣Estimator"))
abline(h = 0, col = "red", lwd = 2) # draw bottom-line
```

Figure 4: Efficiency gain through the Fay-Herriot estimator



Source: see section 7 - References

It can be observed that, even though the model-assumptions are not fulfilled, the model-based Fay-Herriot approach is able to achieve a gain in precision with respect to the **estimated MSE of** the area-specific estimates. This improvement tends to increase with decreasing area-specific sample size. This is evident as the direct estimator is likely to be more volatile in weakly surveyed regions. Here, the application of small area estimation approaches may accomplish a significant efficiency gain.

7 References

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.

Brown, G., Chambers, R., Heady, P. and Heasman, D. (2001): *Evaluation of small area estimation methods-An application to unemployment estimates from the UK LFS.* Proceedings of Statistics Canada Symposium.

Burgard, J., Ertz, F., Merkle, H. and Münnich, R. (2017): 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., Esteban, M. D., Morales, D. and Pérez, A. (2019): *A Fay– Herriot model when auxiliary variables are measured with error.* TEST, ISSN 1863-8260, doi:10.1007/s11749-019-00649-3.
URL https://doi.org/10.1007/s11749-019-00649-3

Burgard, J. P., Münnich, R. and Zimmermann, T. (2016): *Impact of sampling designs in small area estimation with applications to poverty measurement.* Analysis of Poverty Data by Small Area Estimation, pp. 83–108.

Cassel, C. M., Särndal, C. E. and Wretman, J. H. (1976): *Some Results on Generalized Difference Estimation and Generalized Regression Estimation for Finite Populations.* Biometrika, 63 (3), pp. 615–620, ISSN 00063444.
URL http://www.jstor.org/stable/2335742

Datta, G. S., Hall, P. and Mandal, A. (2011): *Model Selection by Testing for the Presence of Small-Area Effects, and Application to Area-Level Data.* Journal of the American Statistical Association, 106 (493), pp. 362–374, doi: 10.1198/jasa.2011.tm10036.
URL https://doi.org/10.1198/jasa.2011.tm10036

Dorfman, A. H. and Valliant, R. (1997): *The Hájek Estimator Revisited.* ASA Joint Statistical Meetings: Section on Survey Research Methods.

Efron, B. (1982): The jackknife, the bootstrap, and other resampling plans, vol. 38. Siam.

EURAREA Consortium (2015): *Enhancing small area estimation techniques to meet European needs.* Accessed: 2019-08-05.
URL https://www.ine.es/en/docutrab/eurarea/eurarea_05_en.pdf

European Statistical System Committee (2017): *European Statistics Code of Practice.*
URL https://ec.europa.eu/eurostat/documents/4031688/8971242/ KS-02-18-142-EN-N.pdf/e7f85f07-91db-4312-8118-f729c75878c7

Eurostat (2018): *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)

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.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2007): *Estimation of the mean squared error of predictors of small area linear parameters under a logistic mixed model.* Computational Statistics and Data Analysis, 51, pp. 2720–2733.

González-Manteiga, W., Lombardía, M. J., Molina, I., Morales, D. and Santamaría, L. (2008): *Bootstrap mean squared error of a small-area EBLUP.* Journal of Statistical Computation and Simulation, 78 (5), pp. 443–462.

Hájek, J. (1971): *Comment on 'An essay on the logical foundations of survey sampling, part one'.* The foundations of survey sampling, 236.

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. (1998): *Logistic generalized regression estimators.* Survey Methodology, 24, pp. 51–56.

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

López-Vizcaíno, E., Lombardía, M. J. and Morales, D. (2015): *Small area estimation of labour force indicators under a multinomial model with correlated time and area effects.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 178 (3), pp. 535–565, doi:10.1111/rssa.12085. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssa.12085

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

Molina, I. and Marhuenda, Y. (2015): *sae: An R package for small area estimation.* R Journal, in print.

Molina, I., Rao, J. N. K. and Datta, G. S. (2015): *Small area estimation under a Fay–Herriot model with preliminary testing for the presence of random area effects.* Surv Methodol, 41 (1), pp. 1–19.

Münnich, R., Burgard, J. P. and Vogt, M. (2013): S*mall Area-Statistik: Methoden und Anwendungen.* AStA Wirtschafts-und Sozialstatistisches Archiv, 6 (3-4), pp. 149–191.

Pfeffermann, D. (2013): *New important developments in small area estimation.* Statistical Science, 28 (1), pp. 40–68.

Prasad, N. N. and Rao, J. N. K. (1990): *The estimation of the mean squared error of small-area estimators.* Journal of the American statistical association, 85 (409), pp. 163–171.

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

R Core Team (2017): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org

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

Sõstra, K. and Aru, J. (2013): *Regional poverty mapping.* Narusk, E. and Loode, H. (editors) Regional Development in Estonia, pp. 111–115, Statistics Estonia.

Saefken, B., Kneib, T., van Waveren, C.-S. and Greven, S. (2014): *A unifying approach to the estimation of the conditional Akaike information in generalized linear mixed models.* Electron. J. Statist., 8 (1), pp. 201–225, doi: 10.1214/14-EJS881.
URL https://doi.org/10.1214/14-EJS881

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

Strzalkowska, E. and Molina, I. (2018): S*mall area estimation (communes) of economic activity rate in the structural survey.* Accessed: 2019-08-05.
URL https://www.experimental.bfs.admin.ch/en/sae.html

Szymkowiak, M., Mlodak, A. and Wawrowski, L . (2017): *Mapping Poverty at the Level of Subregions in Poland Using Indirect Estimation.* Statistics in Transition, 18 (4), pp. 609–635.

Tzavidis, N., Zhang, L.-C., Luna, A., Schmid, T. and Rojas-Perilla, N. (2018): *From start to finish: a framework for the production of small area official statistics.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 181 (4), pp. 927–979.

Vogt, M. (2008): Die Schätzer von Fay-Herriot und Battese-Harter-Fuller. Diploma thesis, Trier University.

Ybarra, L. M. and Lohr, S. L. (2008): *Small area estimation when auxiliary information is measured with error.* Biometrika, 95 (4), pp. 919–931.

You, Y. and Rao, J. N. K. (2002): *A pseudo-empirical best linear unbiased prediction approach to small area estimation using survey weights*. Canadian Journal of Statistics, 30 (3), pp. 431–439.

You, Y. and Rao, J. N. K. (2003): *Pseudo Hierarchical Bayes Small Area Esti- mation Combining Unit Level Models and Survey Weights.* Journal of Statistical Planning and Inference, 111, pp. 197–208.

8 Appendix

A.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 units in area d. For each area with the running index $d = 1, \ldots, D$ 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

$$\hat{\tau}_d^{HT} = \sum_{k \in S_d} \frac{y_k}{\pi_k} = \sum_{k \in S_d} w_k \, y_k \qquad (1)$$

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.

A.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). Typically, the assisting model is a linear regression. In general, a wide number of different assisting model can be used, such as logistic or multinomial models and regularized regressions.

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

$$\hat{\tau}_d^{GREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in S_d} w_k(y_k - \hat{y}_k) \qquad (2)$$

(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 (2) 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.3 Model-based estimation

A.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 $d = 1, \dots, D$, is given by

$$\hat{\mu}_d^{Dir} = \mu_d + e_d \qquad (3)$$

with a direct estimator $\hat{\mu}_d^{Dir}$. In the present case $\hat{\mu}_d^{Dir}$ is the direct estimator for the respective area of interest d. An estimator is designated to be direct if only data from the respective area of interest have been used for the estimation. In addition, μ_d is the true but usually unknown parameter of interest in region d. It is assumed that the sampling errors e_d, are independent and $e_d \sim N(0, \psi_d)$. Therefore, it is supposed that $\hat{\mu}_d^{Dir}$ is a design-unbiased estimator for μ_d, and that ψ_d sampling variance of the estimator, is known.

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,

$$\mu_d = \overline{\mathrm{X}}_d^T \beta + v_d \tag{4}$$

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

$$\hat{\mu}_d^{Dir} = \overline{\mathrm{X}}_d^T \beta + v_d + e_d \tag{5}$$

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

$$\hat{\mu}_d^{FH} = \overline{\mathrm{X}}_d^T \hat{\beta} + \hat{v}_d \tag{6}$$

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

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 $\psi_d + \sigma_v^2$ 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

$$\hat{\beta} = (\sum_{d=1}^{D} \frac{\overline{\mathrm{X}}_d \overline{\mathrm{X}}_d^T}{(\psi_d + \sigma_v^2)})^{-1} (\sum_{d=1}^{D} \frac{\overline{\mathrm{X}}_d \hat{\mu}_d^{Dir}}{(\psi_d + \sigma_v^2)}) \tag{7}$$

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

$$\hat{\mu}_d^{FH} = \gamma_d \hat{\mu}_d^{Dir} + (1 - \gamma_d) \overline{\mathrm{X}}_d^T \hat{\beta} \tag{8}$$

As a result of the transformation, it is visible that 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 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 practice however σ_v^2, is unknown and has to be estimated as well For this purpose a number of fitting methods exist. By replacing the model variance σ_v^2 by the estimated variance of the random effects $\hat{\sigma}_v^2$ in (6) and (7), the empirical best linear unbiased predictor (EBLUP) is obtained.

Software

In the statistical software R (R Core Team, 2017), the package sae (Molina and Marhuenda, 2015) provides routines for the Fay-Herriot estimator. The functions **eblupFH()** and **mseFH()** calculate the Fay-Herriot point estimates and the point estimates in addition to analytical MSE estimates respectively. The function calls are given by

```
eblupFH(formula, vardir, method = "REML", MAXITER = 100, PRECISION = 0.0001, data)
```

and

```
mseFH(formula, vardir, method = "REML", MAXITER = 100, PRECISION = 0.0001, data).
```

The fixed part of the model is specified in the formula object. The variance ψ_d of the direct estimator $\hat{\mu}_d^{Dir}$ is specified in the vardir part. method sets the default fitting method used to estimate σ_v^2, where the default is the restricted maximum likelihood (**REML**) approach. **MAXITER** and **PRECISION** are optional arguments and specify the default maximum number of iterations and the convergence tolerance criteria of the Fisher-scoring algorithm respectively. By means of the data part, the object containing the respective data, i.e. direct estimator and its variance, can be specified.

The functional output of both function calls comprises a list with the EBLUPs for the defined areas (**eblup**) and the output from the fitting process (**fit**). If the function **mseFH()** has been chosen, the output additionally contains the analytical estimates for the MSE of the EBLUPs (**mse**).

A.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

$$y_{dk} = x_{dk}^T\beta + v_d + e_{dk}, d = 1, \dots, D, k = 1, \dots, N_d \qquad (9)$$

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 (5), 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 (9) 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):

$$\hat{\mu}_d^{BHF} = \overline{\mathrm{X}}_d^T\hat{\beta} + \hat{v}_d \qquad (10)$$

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

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

(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 $\overline{\mathrm{X}}_d$, 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:

$$\hat{\mu}_d^{BHF} = \gamma_d\left(\bar{y}_d + (\overline{\mathrm{X}}_d - \bar{x}_d)^T\hat{\beta}\right) + (1-\gamma_d)\overline{\mathrm{X}}_d^T\hat{\beta} \tag{11}$$

Here, it has to be recognised that the Battese-Harter-Fuller estimator is a weighted average of the direct sample regression estimator $\bar{y}_d + (\overline{\mathrm{X}}_d - \bar{x}_d)^T\hat{\beta}$ and the regression synthetic component $\overline{\mathrm{X}}_d^T\hat{\beta}$. 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 $(n_d = 0, \text{so } \gamma_d = 0)$ the BLUP consists entirely of the synthetic estimator. However, this assumes that the auxiliary characteristics of the units of this area are known, so that the area-specific average value $\overline{\mathrm{X}}_d$ 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).

Software

The package sae (Molina and Marhuenda, 2015) also provides routines for the Battese-Harter-Fuller estimator. The functions **eblupBHF()** and **pbmseBHF()**

provide the respective point estimates for the area mean values as well as point esti- mates in addition to parametric bootstrap MSE estimates according to Gonza´lez- Manteiga et al. (2008), respectively. The function calls are given by

```
eblupBHF(formula, dom, selectdom, meanxpop, popnsize,method = "REML", data)
```

and

```
pbmseBHF(formula, dom, selectdom, meanxpop, popnsize, B = 200, method = "REML", data).
```

As with the Fay-Herriot estimator in the sae package, the fixed part of the model again is specified in the formula object. The area codes for the sample elements are specified in the dom object. selectdom specifies a selected subset of areas, for which the mean values should be estimated. In addition, **meanxpop** contains the population means of the auxiliary variables in a named data frame. The population sizes of the areas, N_d, are listed as a data frame in popnsize. The object method again sets the fitting method, where the default is the restricted maximum likelihood (**REML**) approach. The object containing the respective data, i.e. direct estimator and its variance, can by specified by means of the data part. The number of parametric bootstrap replicates for the MSE estimation can be set by the parameter B in the **mseFH()** function.

Again, the functional output of the function calls consists of a list with the EBLUP values for the specified areas (**eblup**) and the results from the fitting process (**fit**). If the function **pbmseBHF()** has been chosen, the output additionally returns the estimated MSE values for the specified areas (**mse**).

A.3.3 You-Rao estimator

The Battese-Harter-Fuller estimator introduced in the previous section is fully model-based and therefore does not use the design weights resulting from the sampling process. A potential alternative is the pseudo-EBLUP developed by You and Rao (2002). The estimator is based on the aggregated design-weighted area-level model

$$\bar{y}_{d.w} = \bar{x}_{d.w}^T\beta + v_d + \bar{e}_{d.w} \tag{12}$$

using the standardised weights $\tilde{w}_{dk} = w_{dk}/\sum_{k=1}^{n_d} w_{dk}$ (You and Rao, 2002, p. 433). The weighted averages are calculated using unit-level information. Thereby, it is $\bar{y}_{d.w} = \sum_{k=1}^{n_d} \tilde{w}_{dk} y_{dk}$ and $\bar{x}_{d.w} = \sum_{k=1}^{n_d} \tilde{w}_{dk} x_{dk}$. Furthermore, $\bar{e}_{d.w} = \sum_{k=1}^{n_d} \tilde{w}_{dk} e_{dk}$ applies with $E(\bar{e}_{d.w}) = 0$ and $Var(\bar{e}_{d.w}) = \sigma_e^2 \sum_{k=1}^{n_d} \tilde{w}_{dk}^2$ (cf. ibid.).

For known parameters β, σ_v^2 and σ_e^2 the BLUP of the domain-specific mean μ_d is given by

$$\hat{\mu}_d^{YR} = \gamma_{d.w}(\bar{y}_{d.w} + (\bar{X}_d - \bar{x}_{d.w})^T\beta) + (1 - \gamma_{d.w})\bar{X}_d^T\beta \tag{13}$$

$$\text{with } \gamma_{d.w} = \sigma_v^2/(\sigma_v^2 + \sigma_e^2 \textstyle\sum_{k=1}^{n_d} \tilde{w}_{dk}^2)$$

(You and Rao, 2002, p. 435). The regression parameter β unknown in practical applications, can be estimated by

$$\hat{\beta}_w(\sigma_v^2, \sigma_e^2) = \left[\sum_{d=1}^{D}\sum_{k=1}^{n_d} w_{dk}x_{dk}(x_{dk} - \gamma_{d.w}\bar{x}_{dw})^T\right]^{-1}\left[\sum_{d=1}^{D}\sum_{k=1}^{n_d} w_{dk}(x_{dk} - \gamma_{d.w}\bar{x}_{dw})y_{dk}\right] \quad (14)$$

using the design weights. The variance components σ_v^2 and σ_e^2 contained in $\hat{\mu}_d^{YR}$ are also estimated in practice. For this respect, You and Rao (2002, p. 433) apply the method of moments, so that the parameters are estimated by

$$\hat{\sigma}_e^2 = (n - D - P + 1)^{-1} \sum_{d=1}^{D}\sum_{k=1}^{n_d} \hat{\epsilon}_{dk}^2 \quad (15)$$

and $\hat{\sigma}_v^2 = \max(\tilde{\sigma}_v^2, 0)$ with

$$\tilde{\sigma}_v^2 = \left[n - tr\left((\mathrm{X}^T\mathrm{X})^{-1}\sum_{d=1}^{D} n_d^2 \bar{\mathrm{x}}_d\bar{\mathrm{x}}_d^T\right)\right]^{-1}\left[\sum_{d=1}^{D}\sum_{k=1}^{n_d} \hat{u}_{dk}^2 - (n - P)\hat{\sigma}_e^2\right] \quad (16)$$

Thereby, $P$is the number of auxiliary variables including the intercept. $\hat{\epsilon}_{dk}^2$ are the residuals of the least-squares regression of $y_{dk} - \bar{y}_d$ on $x_{dk1} - \bar{x}_{d.1}, \ldots, x_{dkP} - \bar{x}_{d.P}$. Furthermore, $\hat{u}_{dk}$ are the residuals of the least squares regression of y_{dk} on $x_{dk1}, \ldots, x_{dkP}$ (ibid).

The so-called pseudo-EBLUP according to You and Rao (2002) is finally created by replacing β, σ_v^2 and σ_e^2 in (13) by $\hat{\beta}_w(\hat{\sigma}_v^2, \hat{\sigma}_e^2), \hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$. In In contrast to the simple unit-level EBLUP according to Battese et al. (1988), it is design-consistent with increasing n_d. In addition, the pseudo-EBLUP automatically fulfils the benchmarking.

property when aggregating domain-specific total value estimates across all domains This implies that the sum of these estimates $\sum_{d=1}^{D} N_d\hat{\mu}_d^{YR}$ corresponds to a GREG estimate for the total population. The prerequisite for this is that the design weights have been calibrated according to the known domain size in the population, so that $N_d = \sum_{k=1}^{n_d} w_{dk}$ applies (see Rao and Molina, 2015, p. 208). This so-called vertical coherence is a property which is of great importance especially in official statistics. For the proof of the benchmarking property as well as the approximation of the MSE of the pseudo-EBLUP, it is referred to You and Rao (2002, p. 436 f.) and Rao and Molina (2015, p. 208 f.).

A.3.4 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. It is often 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 $\widehat{\overline{X}}_d$.

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 $\widehat{\overline{X}}_d \overset{ind}{\sim} N(\overline{X}_d, C_d)$, where C_d is the known variance-covariance matrix of the estimated mean values of the register variables. Furthermore, $\widehat{\overline{X}}_d$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:

$$\hat{\mu}_d^{ME} = \gamma_d \hat{\mu}_d^{Dir} + (1-\gamma_d)\widehat{\overline{X}}_d^T \beta \tag{17}$$

The weighting factor γ_d depend not 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.

$$\gamma_d = \frac{\sigma_v^2 + \beta^T C_d \beta}{\sigma_v^2 + \beta^T C_d \beta + \psi_d} \tag{18}$$

The more inexactly $\widehat{\overline{X}}_d$ 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 $\widehat{\overline{X}}_d$ is made without error $(C_d = 0)$, $\hat{\mu}_d^{ME}$ is reduced to the Fay-Herriot 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 (17) is

$$MSE(\hat{\mu}_d^{ME}) = \gamma_d \psi_d \tag{19}$$

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

$$\hat{\sigma}_v^2 = (D-P)^{-1} \sum_{d=1}^{D} \left(\left(\hat{\mu}_d^{Dir} - \hat{\bar{\mathrm{X}}}_d^T \hat{\beta}_w \right)^2 - \psi_d - \hat{\beta}_w^T C_d \hat{\beta}_w \right) \qquad (20)$$

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

$$\hat{\beta}_w = \left(\sum_{d=1}^{D} w_d \left(\hat{\bar{X}}_d \hat{\bar{X}}_d^T - C_d \right) \right)^{-1} \sum_{d=1}^{D} w_d \, \hat{\bar{\mathrm{X}}}_d \hat{\mu}_d^{Dir} \qquad (21)$$

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 $D \rightarrow \infty$. Here $w_d = 1/(\sigma_v^2 + \psi_d + \beta^T C_d \beta)$ are positive finite weight. The parameters are estimated in a two-step process. First $w_d = 1$. The β and σ_v^2 are then estimated by (20) and (21). Based on the two estimates, the weights $\widehat{w}_d$ are estimated again, to finally obtain the final estimates $\hat{\beta}_w$ and $\hat{\sigma}_v^2$ (see ibid.).

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: https://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 (https://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.

Guidelines on small area estimation for city statistics and other functional geographies

KS-GQ-19-011-EN-C

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. The aim of these guidelines is to provide the National Statistical Institutes with a common framework for production of small area estimates. In addition to the timeliness and the quality of the estimates, to assure comparability across Europe and coherence with the estimates at larger territorial scale would be essential. Therefore, the proposed step-by-step procedure of specifying, implementing and evaluating various small area estimation techniques up to the final selection of an optimal approach has a great potential for improvement of quality in official statistics.

For more information
https://ec.europa.eu/eurostat/

Publications Office
of the European Union

ISBN 978-92-76-11708-7