JYVÄSKYLÄ STUDIES IN COMPUTING
279

Mauno Keto

Optimal Sample Allocation Conditioned on a Small Area Model, Estimator, and Auxiliary Data

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN COMPUTING 279

Mauno Keto

Optimal Sample Allocation Conditioned on a Small Area Model, Estimator, and Auxiliary Data

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen Gamma-salissa
toukokuun 24. päivänä 2018 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, Gamma hall, on May 24, 2018 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2018

Optimal Sample Allocation Conditioned on a Small Area Model, Estimator, and Auxiliary Data

JYVÄSKYLÄ STUDIES IN COMPUTING 279

Mauno Keto

Optimal Sample Allocation Conditioned on a Small Area Model, Estimator, and Auxiliary Data

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2018

Editors
Timo Männikkö
Faculty of Information Technology, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

Permanent link to this publication: http://urn.fi/URN:ISBN:978-951-39-7417-6

URN:ISBN:978-951-39-7417-6
ISBN 978-951-39-7417-6 (PDF)

ISBN 978-951-39-7416-9 (nid.)
ISSN 1456-5390

Jyväskylä University Printing House, Jyväskylä 2018

ABSTRACT

Keto, Mauno
Optimal sample allocation conditioned on a small area model, estimator, and auxiliary data
Jyväskylä: University of Jyväskylä, 2018, 34 p. (+ included articles)
(Jyväskylä Studies in Computing
ISSN 1456-5390; 279)
ISBN 978-951-39-7416-9 (nid.)
ISBN 978-951-39-7417-6 (PDF)
Finnish summary
Diss.

We have studied optimal sample allocation, associated with small area estimation, when the objective is to obtain as accurate estimates as possible, for the population and for the subpopulations, called as areas here. It is a question of a two-level optimization problem. The basic premise is composed of planned areas, stratified sampling, and small overall sample size predetermined by restricted time and budget resources. Low sample sizes are common in market surveys.

During this thesis, we have developed new allocation methods, based on a small area model, estimator, and auxiliary data. The final method, the three-term Pareto allocation, is based on the three terms of the mean-squared error estimator for the area total empirical best linear unbiased predictor estimator, and on the Pareto optimization technique. The performance of the final method has improved, compared with our other model-based allocations.

We compare the performances of our allocations with the reference allocations, selected from the literature, through design-based sample simulations using real data. The selection criterion is the diversity in optimality associated with the allocations. From the point of view of the performance, the most competing allocations are the nonlinear programming and the Costa allocations.

Model-based estimation produces more accurate estimates than design-based estimation under the research population structure. Our allocation leads to estimates with the best accuracies and moderately small biases.

The results support the conditioning of the sample allocation on the model and on the estimator. It is also important to consider the balance between the area level and the population level estimation, and between the accuracy and the bias of the estimates.

Keywords: small sample size, area characteristics, register data, trade-off, multi-objective optimization

Author

Mauno Keto
Faculty of Information Technology
University of Jyväskylä, Finland
mauno.j.keto@student.jyu.fi

Supervisors

Professor emeritus Erkki Pahkinen
Department of Mathematics and Statistics
University of Jyväskylä, Finland

Professor Pekka Neittaanmäki
Faculty of Information Technology
University of Jyväskylä, Finland

Doctor Jussi Hakanen
Faculty of Information Technology
University of Jyväskylä, Finland

Professor Juha Karvanen
Department of Mathematics and Statistics
University of Jyväskylä, Finland

Reviewers

Associate Professor Imbi Traat
Institute of Mathematical Statistics
University of Tartu, Estonia

Professor Mikko Myrskylä
Faculty of Social Sciences
University of Helsinki, Finland

Opponent

Professor Ralf Münnich
Economic and Social Statistics Department
University of Trier, Germany

ACKNOWLEDGEMENTS

I am deeply grateful to Professor emeritus Erkki Pahkinen, who has been my main supervisor since I started my licenciate thesis, which was the first milestone in this dissertation process. He has guided me in theoretical issues and introduced new ideas, but he has also inspired me to develop my own solutions. His wide domestic and international contacts have enriched my research work. I have had the opportunity to make the acquaintance of prominent statisticians in the field of small area estimation. During our co-operation, I have learned a lot of sampling theory and of survey methodology.

I am grateful to my second supervisor, Professor Pekka Neittaanmäki. He provided proper working conditions for the time when I completed my dissertation. He has actively followed my work and assisted in theoretical issues.

I started the co-operation with my third supervisor, Doctor Jussi Hakanen, during the preparations of the last two articles. His guidance has been significant in multi-objective optimization, but he also has assisted me in methodological and theoretical issues. He deserves my gratitude.

I thank my fourth supervisor, Professor Juha Karvanen. His proposal led to this kind of article dissertation. He has assisted me in theoretical issues.

Associate Professor Imbi Traat from University of Tartu, Estonia, and Professor Mikko Myrskylä from University of Helsinki have acted as the reviewers of my dissertation. I thank them for their constructive comments.

I thank Professor emeritus Antti Penttinen, whose support and flexibility were important during the first years of my post-graduate studies. Later, he has introduced many useful ideas and has assisted in software issues.

Professor emeritus Risto Lehtonen is a statistician whose expertise in small area estimation is appreciated among his Finnish and foreign colleagues. I am grateful to him for his constructive suggestions and comments which I received, when I prepared my licenciate thesis and this dissertation.

I thank Doctor Kari Nissinen for providing the SAS software which I used in the simulation experiments. He also has assisted in theoretical and methodological issues.

I thank Alma Mediapartners Oy Company for providing me the research data which I used in the last three articles.

I thank many of my ex-colleagues for technical support during and after my years of work in Mikkeli University of Applied Sciences.

There has been one additional feature in my research work. I live in Mikkeli, which is located over 100 km from Jyväskylä. This distance has required a lot of travelling and the organization of many practical matters at home. All this would not have been possible without the support and patience of my wife Ulla. She has encouraged me to carry on my work especially on difficult days. I also thank my daughter Iida and her family for mental support.

Jyväskylä 20.4. 2018
Mauno Keto

FIGURES

TABLES

CONTENTS

LIST OF INCLUDED ARTICLES

PI M. Keto and E. Pahkinen. On sample allocation for effective EBLUP estimation of small area totals – "Experimental Allocation", in Survey Sampling Methods in Economic and Social Research, J. Wywial and W. Gamrot (eds). Katowice: Katowice University of Economics, 27–36, 2010.

PII M. Keto and E. Pahkinen. Sample allocation for efficient model-based small area estimation. Survey Methodology, 43(1): 93–106, 2017. DOI: http://www.statcan.gc.ca/pub/12-001-x/2017001/article/14817-eng.pdf.

PIII M. Keto and E. Pahkinen. On overall sampling plan for small area estimation. Statistical Journal of the IAOS, 33: 727–740, 2017.

PIV M. Keto, J. Hakanen, and E. Pahkinen. Register data in sample allocations for small-area estimation. Mathematical Population Studies, An International Journal of Mathematical Demography, 2018 (accepted, in print).

1 INTRODUCTION

It is a common trend in sample-based surveys, that different parameters of the variables of interest are estimated both at the population and at the subpopulation (area) level. In descriptive sample surveys, an important phenomenon can be the target of estimation. For example, national investments on the research and development and on the innovations, and the industry-specific investments are estimated.

We have studied optimal sample allocation, associated with small area estimation, when the objective is to obtain as accurate estimates as possible, for the population and for the areas. This is a two-level optimization problem. The available financing and time resources are the main determinant of the overall sample size, which restricts the quality of the estimates. The basic premise in this thesis is composed of small overall sample size, stratified sampling, and planned areas, which coincide with the strata.

We estimate the area and the population totals of the variable of interest. Throughout this thesis, we use the model-based empirical best linear unbiased predictor (EBLUP) estimator, based on the unit-level linear mixed model, to obtain the estimates. We also use the design-based estimation in PIII-PIV.

During recent decades, various design- or model-based small area estimators, including also composite estimators, have been constructed (Pfeffermann 2013), but the development of sample allocations for small area estimation has not been as intensive. The contributions of Singh, Gambino, and Mantel (1994), and Marker (2001) are important in the discussion about sampling strategies. One obstacle to the development may be the complexity of the models and the estimators (Longford 2006). Examples of optimal allocation based on a composite estimator are given by Longford (2006) and Molefe and Clark (2015). Khan, Maiti and Ahsan (2010) have developed an optimal sample allocation based on multivariate ratio and regression methods of estimation. The latest developments include multi-objective Pareto optimization (Friedrich, Münnich, and Rupp 2018).

Important strategical choices are part of the sampling design. One choice concerns the balance between the population-level and the area-level estimation.

Another choice is associated with the accuracy and the bias. The mean-squared error (MSE), which measures the accuracy of an estimator or estimate, is composed of the variance and squared bias. Design-based estimators are generally design unbiased, but can have large variances, especially for areas with low sample sizes. Model-based estimators have typically lower variances, but may have a high design bias, indicating possible model misspecification. The reasonable trade-off between the variance and the bias should be considered when the model and the estimator are selected (Burgard et al. 2014).

The area-specific sample sizes for an allocation typically result from the solution of an optimization problem, which is related to measures of uncertainty like mean-squared error, variance, or coefficient of variation (CV). From this point of view, it is logical to conclude that the allocation solution must be conditioned on the underlying model and the computing technique. Sometimes a closed analytical solution exists, but in general, numerical methods must be employed, like multi-objective optimization for solving problems containing multiple conflicting objective functions.

Figure 1 illustrates the stages of a small area survey process and the key elements related to sample allocation in this thesis.

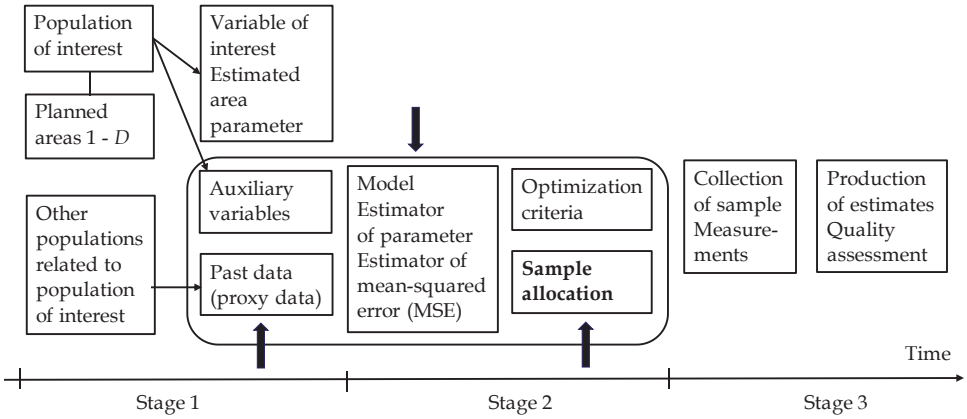

FIGURE 1 Elements of the small area survey process related to sample allocation

We have gradually developed an allocation, based on a model, estimator, auxiliary information, and multi-objective optimization. The development process ends to the allocation, where the key elements are the three terms of the MSE estimator for the area total empirical best linear unbiased predictor (EBLUP) estimator, and the Pareto optimization technique. The idea is to obtain maximal accuracy for the area and for the population estimates simultaneously.

This thesis is a collection of articles published in a conference monograph and scientific journals. In every publication, we introduce a model-based allocation and compare its performance with reference allocations, through design-based simulation experiments using real data. The selection of the reference allocations is based on the diversity in optimality. We evaluate the performanc-

es of the allocations in terms of sample-based quality measures for accuracy and bias.

Articles PI-PII use distinct real register data sets. In PIII and PIV, two periodically collected real register data sets are used, the former one for computing the sample sizes for the allocations and the latter one for simulation experiments.

In PI, we compare the performance of an experimental allocation with three reference allocations. The main objective is to find a direction for developing a model-based allocation method for stratified sampling. We examine the relationship between area-specific sample sizes and the performance of the experimental allocation, including zero sample sizes for many areas.

In PII, we derive analytically the *g1* allocation, based on the unit-level linear mixed model and the main term of the mean-squared error estimator for the area total estimator. The six reference allocations use another area model and estimator, area-specific parameter information, or only number-based information. We examine, how the quality of the area estimates is related to the characteristics of areas and the sample sizes. We show that the area estimates can be moderately accurate, despite small area sample sizes.

In PIII, we use calibrated area sizes for the *g1* allocation developed in PII, to improve the quality of the area estimates. Five reference allocations are the same as in PII, but we have changed the allocation-specific details. A new feature is the application of the design-based Horvitz-Thompson and the model-assisted regression estimation to four model-free reference allocations.

We present the model-based three-term Pareto method allocation in PIV. The assortment of five reference allocations contains a composite allocation and a modification of another allocation. To diversify the comparison of the allocations, we apply three types of estimation methods to four model-free allocations. We also consider the trade-off between the quality of the area estimates and of the population estimates, and the trade-off between accuracy and bias.

The rest of this thesis is organized as follows. In Chapter 2, we introduce the theoretical framework and the basic terminology of the small area estimators used in this thesis. Chapter 3 contains our model-based allocations, the reference allocations, and the quality measures for evaluating the allocations. In Chapter 4, we introduce our real register data sets, present the research contribution of articles PI-PIV, and compare our model-based allocations under the latest register data structure. We present the author´s contribution in Chapter 5. We finally conclude and discuss the directions for future research in Chapter 6.

2 THEORETICAL FRAMEWORK AND BASIC TERMINOLOGY OF SMALL AREA ESTIMATORS

This chapter provides the relevant theory, terminology, and notations, associated with the small area estimators used in this thesis.

2.1 Planned areas and small area

The target population in a survey or research containing measurable statistical units is partitioned into non-overlapping subpopulations, called as areas or domains. The areas are for example geographical areas, socio-demographic groups, industries of enterprises, or customer groups.

The number of statistical units in a single area is one measure of its size. From the point of view of estimation, an area is small, if the area-specific sample size is not large enough to support direct area estimates (computed of the area-specific sample data) with adequate precision (Rao and Molina 2015).

This thesis uses planned areas with known sizes and stratified simple random sampling (SRSWOR), where the strata coincide with the areas. Planned areas form a basis for the elaboration of the sampling design. Unplanned areas may occur for example in online surveys.

2.2 Variable of interest and auxiliary data

Sample-based surveys are typically carried out to provide area- and population-level estimates for different parameters of the variables of interest, like means, totals, quantiles, or proportions. This thesis focuses on estimating only one variable of interest. Lehtonen et al. (2006), Pfeffermann and Sverchkov (2007), and Burgard, Münnich, and Zimmermann (2014) are examples of the studies dealing with one variable of interest. Large-scale surveys include many variables of

interest, and different partitions of the population are possible. The study of Falorsi and Righi (2008) is an example of this case. They have developed a sampling strategy for multivariate and multidomain estimation, when the overall sample size is small.

The least amount of information for producing the estimates for the variable of interest *y* is composed of the sampled *y*-values. The only alternative is to apply direct design-based estimation, without auxiliary variables. The available past data of *y* may be applicable to the sampling design.

The design-based model-assisted or the model-based estimators can be used if unit-level or aggregate-level population data of the auxiliary variables (covariates) are available. The sampling design may benefit from the covariates if they correlate with the variable of interest. Past data may also be useful for the sampling design and for the estimation of the model parameters.

We use one auxiliary variable in PI-PII for estimation and use the auxiliary variable also for the sampling design in PII. In PIII-PIV, the two periodically collected register data sets contain the same variable of interest and the same two auxiliary variables. The former data set, called "proxy data", provides information for the sample allocation, and the variable of interest in this data set is called "proxy-*y*, denoted by y^*.

2.3 Definitions and notations

The following notations are used when defining relevant concepts and expressions associated with small area estimation.

1) The population U of N basic units is composed of D mutually exclusive and independent areas $U_1,...,U_D$, with $N_1,..,N_d$ units, and $\sum_{d=1}^{D} N_d = N$.
2) The estimated area parameters of the variable of interest y are the totals $Y_d = \sum_{k \in U_d} y_{dk}$ or means $\bar{Y}_d = Y_d / N_d$, where y_{dk} is the value for unit k in area d.
3) The estimated population parameters of y are the total $Y = \sum_{d=1}^{D} Y_d$ or the mean $\bar{Y} = Y / N$.
4) In stratified sampling, an independent subsample s_d with fixed size n_d is selected from U_d. The overall sample s is allocated to the areas so that it is the union of the subsamples s_d, and the overall sample size $n = \sum_{d=1}^{D} n_d$.
5) The area sample statistics of y are totals $\sum_{k \in s_d} y_{dk}$ or means $\bar{y}_d = \sum_{k \in s_d} y_{dk} / n_d$.

6) The population data of auxiliary variables x_i (i=1,…, p) includes unit values x_{idk}, the area totals $X_{id} = \sum_{k \in U_d} x_{idk}$, or area means $\bar{X}_{id} = X_{id} / N_d$. For the sampled units, the area totals are $\sum_{k \in s_d} x_{idk}$ and area means are $\bar{x}_{id} = \sum_{k \in s_d} x_{idk} / n_d$.

2.4 Small area estimators

The small area estimators are traditionally classified into direct and indirect estimators, or into design-based and model-based estimators.

An estimator is direct if it uses values of the variable of interest y only from the time period of interest and only from units in the domain of interest (Federal Committee on Statistical Methodology 1993). A direct estimator may also use the area-specific auxiliary information related to y (Rao and Molina 2015). An indirect estimator uses y-values from a domain other than the domain of interest, from a time period other than the period of interest, or from other domains and other time periods simultaneously (Rao and Molina 2015).

Design-based estimators use survey weights and are based on traditional sampling theory. A random sample s of size n is selected from the population U with probability $p(s)$ which is defined according to the sampling design. The unit-specific inclusion probabilities π_{dk} for units $k \in s_d$ are used when computing the estimates for the area d. The associated inferences are based on the probability distribution induced by the sampling design.

A typical design-based direct estimator is given by

$$\hat{Y}_d = \sum_{k \in s_d} w_{dk} y_{dk} , \tag{1}$$

where w_{dk} is the design weight for unit k in area d. The choice $w_{dk} = \pi_{dk}^{-1}$ leads to the well-known Horvitz-Thompson estimator

$$\hat{Y}_{d,HT} = \sum_{k \in s_d} y_{dk} / \pi_{dk} . \tag{2}$$

The estimator (1) is design unbiased, but small sample sizes may lead to inaccurate area estimates. The inclusion probability $\pi_{dk} = n_d / N_d$ in stratified SRSWOR sampling, and the estimator (2) reduces to

$$\hat{Y}_{d,HT} = N_d \bar{y}_d , \tag{3}$$

which is also a post-stratified estimator in case of unplanned areas. The design variance of estimator (3) is

$$\mathrm{V}(N_d \bar{y}_d) = N_d^2 (1 - n_d / N_d) S_{y,d}^2 / n_d = N_d^2 (1/n_d - 1/N_d) S_{y,d}^2 , \tag{4}$$

where $S^2_{y,d}$ is the population variance of y in area d. The unbiased estimator of the variance (4) can be computed with the condition that the sample size $n_d \geq 2$.

Direct model-assisted estimators are based on models fitted separately for each area. An example is a regression model between y and p auxiliary variables

$$y_{dk} = \boldsymbol{x}'_{dk}\boldsymbol{\beta}_d + e_{dk}\,, \tag{5}$$

where $\boldsymbol{x}'_{dk} = (1, x_{dk1}, ..., x_{dkp})$ is the vector of auxiliary data for unit k in area d and $\boldsymbol{\beta}_d = (\beta_{0d}, \beta_{1d}, ..., \beta_{pd})'$ is the vector of area-specific regression coefficients, and e_{dk} is the error term. A direct area estimator can incorporate auxiliary data from outside the area of interest (Lehtonen and Veijanen 2009).

Indirect estimators, based on an implicit linking model, include a synthetic (Gonzales 1973) and a composite estimator. The model in this context is more a uniting factor between the areas than an actual model. The linear combination of a direct and synthetic estimator represents a composite estimator. One of the reference allocations is based on a composite estimator

$$\hat{\bar{Y}}^{C}_{d} = (1-\phi_d)\bar{y}_{dr} + \phi_d \hat{\bar{Y}}_{d(\mathrm{syn})}\,, \tag{6}$$

where $\hat{\bar{Y}}_{d(\mathrm{syn})} = \hat{\boldsymbol{\beta}}'\bar{\boldsymbol{X}}_d$ is a synthetic estimator and $\bar{y}_{dr} = \bar{y}_d + \hat{\boldsymbol{\beta}}'(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{X}}_d)$ is a direct estimator (Molefe and Clark 2015). The vector $\hat{\boldsymbol{\beta}}$ contains the estimated regression coefficients, $\bar{\boldsymbol{X}}_d$ is the vector of the area-specific population means of auxiliary variables $\boldsymbol{x}$, and $\bar{y}_d$ and $\bar{\boldsymbol{x}}_d$ are the sample means of y and $\boldsymbol{x}$ in area d. The coefficients ϕ_d are set with the intent to minimize the design mean-squared error (MSE) of the estimator (6).

Small area models are explicit linking models containing random area-specific effects accounting for the between-area variation, instead of auxiliary variables (Rao and Molina 2015). These models are area- or unit-level models. Indirect estimators based on small area models are called model-based estimators.

In this thesis, the model-based area total estimates are produced by using the estimator based on the unit-level linear mixed model, known also as the nested error regression model (Battese, Harter, and Fuller 1988)

$$y_{dk} = \boldsymbol{x}'_{dk}\boldsymbol{\beta} + v_d + e_{dk}; k = 1, ..., N_d; d = 1, ..., D, \tag{7}$$

where $\boldsymbol{x}'_{dk} = (1, x_{dk1}, ..., x_{dkp})$ is is the vector of auxiliary data for unit k in area d, the vector $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)'$ contains the fixed regression parameters, and the area-specific effects v_d distributed as $N(0, \sigma^2_v)$ are independent of the random errors e_{dk} distributed as $N(0, \sigma^2_e)$.

The variance of y $\mathrm{V}(y_{dk}) = \sigma_v^2 + \sigma_e^2$ is decomposed into the variation between and within the areas. The common intra-area correlation (Meza and Lahiri 2005)

$$\rho = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2) \tag{8}$$

measures the relative between-area variation of y. The model parameters and area effects are estimated from the sample data. The estimate $\hat{\boldsymbol{\beta}}$ is a generalized least-squares (GLS) estimate of $\boldsymbol{\beta}$. The variance parameters are estimated by restricted maximum likelihood (REML).

The EBLUP estimator for the area total Y_d of the variable of interest y is the sum of sampled y-values and the sum of predicted y-values for non-sampled units from area d and is given by

$$\hat{Y}_{d,Eblup} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \hat{y}_{dk} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \boldsymbol{x}'_{dk} \hat{\boldsymbol{\beta}} + (N_d - n_d)\hat{v}_d \,, \tag{9}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{v}_d$ are estimates of $\boldsymbol{\beta}$ and v_d. The design mean-squared error for the estimator (9) contains the variance and squared bias and is given by

$$\mathrm{MSE}(\hat{Y}_{d,\mathrm{Eblup}}) = \mathrm{E}(\hat{Y}_{d,\mathrm{Eblup}} - Y_d)^2 = \mathrm{V}(\hat{Y}_{d,\mathrm{Eblup}}) + (\mathrm{E}(\hat{Y}_{d,\mathrm{Eblup}}) - Y_d)^2 \,. \tag{10}$$

The Prasad-Rao prediction estimator of (10) for finite populations (Rao and Molina 2015) is

$$\mathrm{mse}(\hat{Y}_{d,\mathrm{Eblup}}) = g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{4d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \,, \tag{11}$$

where the terms g_{1d}, g_{2d}, g_{3d}, and g_{4d} are functions of the variance parameter estimates $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$. The main term g_{1d} contributes generally 85–90% of the estimated MSE (Nissinen 2009).

A detailed account of the model (7) and the estimators (9) and (11) is given in PIV. The terms g_{1d}, g_{3d}, and g_{4d} are significant in this thesis. Our model-based allocations are based on either the first term or all three terms.

A design-based indirect generalized regression estimator (GREG) of the area total Y_d, assisted by a model fitted to the whole sample, is given by

$$\hat{Y}_{d,GREG} = \sum_{k \in U_d} \hat{y}_{dk} + \sum_{k \in s_d} w_{dk}(y_{dk} - \hat{y}_{dk}) \,, \tag{12}$$

where $w_{dk} = \pi_{dk}^{-1}$ and $\hat{y}_{dk}$ is the fitted value of y for unit k in area d. The unit-level linear mixed model (7) is the assisting model in this thesis. The first term of (12) is the predicted value for the area total Y_d. The second term is a bias correction term protecting against model misspecification (Lehtonen and Veijanen 2009). Design-based GREG estimators are typically design-unbiased, but their variances may be large especially for small areas.

3 SAMPLE ALLOCATIONS FOR PLANNED AREAS AND THE PERFORMANCES OF ALLOCATIONS

This chapter presents the allocations used in this thesis and the quality measures for evaluating the allocations. Most of the allocations result from the optimization of an analytical quantity, subject to pre-set constraints. Burgard, Münnich, and Zimmermann (2014) call this quantity as an explicit loss function.

3.1 Own contributions

The experimental allocation in PI is not based on an analytical solution. In the first phase, SRSWOR samples are simulated from the real data and the sample-specific area total EBLUP estimates are produced. Next, the sample-specific means of the mean-square errors and of four quality measures are computed. The sample size distributions of the samples with 20 lowest means, for each measure separately, are examined and the area-specific sample sizes are determined. The target is to find an "ideal" area sample size combination.

The model-based *g1* allocation uses only the first term g_{1d} of the MSE estimator (11). The *g1* allocation minimizes the sum of the g_{1d} terms over the areas

$$F(\boldsymbol{n}) = \sum_{d=1}^{D} g_{1d}(\sigma_v^2, \sigma_e^2) = \sum_{d=1}^{D} (N_d - n_d)(n_d / \sigma_e^2 + 1/\sigma_v^2)^{-1}, \tag{13}$$

subject to $n = \sum_{d=1}^{D} n_d$. The analytical solution yields the sample sizes

$$n_d^{g1} = \frac{N_d n - (N - N_d D - n)(1/\rho - 1)}{N + D(1/\rho - 1)}, \tag{14}$$

where ρ is the common intra-area correlation (8). It is replaced by the adjusted homogeneity coefficient obtained from y^*. A detailed account on this allocation is presented in PII. The formula (14) is an increasing function of N_d.

The *g1* allocation does not include a term for the within-area variation. In PIII, we replace the actual sizes N_d in (14) by the calibrated area sizes. The multiplication of the average area size N/D by the relative area-specific standard deviations yields the calibrated sizes.

The three-term Pareto method allocation in PIV uses three terms g_{1d}, g_{3d}, and g_{4d} of the MSE estimator (11), which reduces to the approximation

$$\text{amse}(\hat{Y}_{d,\text{Eblup}}) = g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{4d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2). \tag{15}$$

The sum of the area-specific approximations (15),

$$\text{amse}(\hat{Y}_{\text{Eblup}}) = \sum_{d=1}^{D} \text{amse}(\hat{Y}_{d,\text{Eblup}}), \tag{16}$$

is an approximate MSE estimator of the total estimator $\hat{Y}_{\text{Eblup}} = \sum_{d=1}^{D} \hat{Y}_{d,\text{Eblup}}$ for the population. The approximate design coefficients of variations for the model-based area and population total estimators are defined as

$$\begin{aligned} &\text{ACV}(\hat{Y}_{d,\text{Eblup}}) = \text{amse}(\hat{Y}_{d,\text{Eblup}})^{1/2} / Y_d, \\ &\text{ACV}(\hat{Y}_{\text{Eblup}}) = \text{amse}(\hat{Y}_{\text{Eblup}})^{1/2} / Y. \end{aligned} \tag{17}$$

The objective is to provide maximal optimal accuracy for the areas and for the population. It is not possible to increase the area-level accuracy, without decreasing the population accuracy, and conversely. The criterion to be minimized using multi-objective optimization contains the coefficients of variation (17). The variance components and the asymptotic variances in (15) are obtained through sample simulations using the proxy register data. The *y*-totals in (17) are obtained from y^*. A detailed account of the allocation is given in PIV.

3.2 Reference methods

The model-assisted allocation (Molefe and Clark 2015), is based on a composite estimator (6) for a small area mean. An assisting area model between *y* and auxiliary variables $\boldsymbol{x}$ is assumed. After simplifying assumptions, Molefe and Clark (2015) have solved the optimal weight ϕ_d in the estimator and obtained the approximate optimum anticipated mean-squared error for the estimator.

The criterion *F* is the linear combination of the anticipated MSE´s of the small area mean and of the overall mean estimators, and is given by

$$F \approx \sum_{d=1}^{D} N_d^q \sigma_d^2 \rho(1-\rho)(1+(n_d-1)\rho)^{-1} + GN_+^q \sum_{d=1}^{D} \sigma_d^2 P_d^2 n_d^{-1}(1-\rho), \tag{18}$$

where $P_d = N_d / N$, σ_d^2 is the variance of *y* in area *d*, and ρ is the common intra-class correlation of *y* between the areas.

Optimal area sample sizes are obtained by minimizing (18) subject to $n = \sum_{d=1}^{D} n_d$. The quantities N_d^q reflect the priority for area-level estimation,

with $0 \le q \le 2$ and $N_+^{(q)} = \sum_{d=1}^{D} N_d^q$. If the priority for the population-level estimation $G > 0$, the criterion F in (18) must be solved by using nonlinear optimization. If $G = 0$, the minimization of the criterion (18) has a unique solution

$$n_d^{MC} = \frac{n\sigma_d N_d^{q/2}}{\sum_{d=1}^{D} \sigma_d N_d^{q/2}} + \frac{1-\rho}{\rho}\left(\frac{\sigma_d N_d^{q/2}}{D^{-1}\sum_{d=1}^{D} \sigma_d N_d^{q/2}} - 1 \right). \tag{19}$$

The intra-class correlation ρ is replaced by the adjusted homogeneity coefficient of the proxy variable y^*, and the standard deviations σ_d are replaced by the standard deviations of y^*. If also $q = 0$, the allocation (19) depends only on ρ and on the standard deviations σ_d and is close to the calibrated *g1* allocation.

Most of the reference allocations are not related to any specific model. They use either number-based population information or the area-specific parameters of the variable of interest y, which is replaced by the proxy variable y^*.

The Neyman allocation (Tschuprow 1923) minimizes the design variance of the population total estimator, in practice the sum of the area-specific design variances (4), subject to $\sum_{d=1}^{D} n_d = n$. The sample sizes are computed as

$$n_d^{NEY} = \frac{N_d S_{d,y}}{\sum_{d=1}^{D} N_d S_{d,y}} n, \tag{20}$$

where $S_{d,y}$ is the standard deviation of y in area d. The allocation (20) performs generally well at the population level, but it may lead to inaccurate estimates for small areas. The box-constraint allocation technique (Gabler, Ganninger and Münnich 2012), used in PIV, avoids sample sizes $n_d < 2$, which prevent the unbiased variance estimation. When the lower sample size limits are increased, the population-level accuracy decreases and the area-level accuracy improves.

The nonlinear programming (NLP) allocation (Choudry, Rao, and Hidiroglou 2012) minimizes the overall sample size n, subject to fixed limits for the design coefficients of variation of the area and the population sample means $\bar{y}_d$ and $\bar{y}$, but the method works also for the total estimators. Due to the conflicting accuracies between the area and the population estimates, different combinations of the limits may lead to the same value of n, but the area-specific sample sizes are not necessarily the same each time. We apply this method conversely, by adjusting the limits until the fixed size n is reached.

Bankier (1988) introduced a "power allocation"

$$n_d^{BAN} = \frac{CV(y)_d X_d^q}{\sum_{d=1}^{D} CV(y)_d X_d^q} n, \tag{21}$$

where $\mathrm{CV}(y)_d$ is the coefficient of variation of y in area d, X_d^q is some measure of size or importance of area d, and q is an adjustable power. Allocation (21) minimizes the function $F = \sum_{d=1}^{D} (CV(\bar{y}_d) X_d^q)^2$, subject to $\sum_{d=1}^{D} n_d = n$.

Compared with one another, the equal and the proportional allocations

$$n_d^{EQU} = \frac{n}{D},$$
$$n_d^{PRO} = \frac{N_d}{N} n, \tag{22}$$

perform generally conversely. The equal allocation may lead to inaccurate population estimates, if the population includes large areas. As for the proportional allocation, especially the smallest areas may have inaccurate estimates.

Costa, Satorra, and Ventura (2004) proposed a convex combination

$$n_d^{COS} = k \frac{N_d}{N} n + (1-k) \frac{n}{D}, \tag{23}$$

where $0 \le k \le 1$. In PIV, the value of k minimizes the difference of the maximal and the minimal design coefficients of variation of the area total estimator (3), subject to $\sum_{d=1}^{D} n_d = n$. The solution requires multi-objective optimization.

3.3 Quality measures for evaluating the allocations

The performances of the allocations, combined with estimators, are evaluated in terms of relative root mean square error (RRMSE) for accuracy and absolute relative bias (ARB). They are sample-based approximations and are defined as

$$\text{RRMSE}_d = 100(1/r \textstyle\sum_{i=1}^{r} (\hat{Y}_{di} - Y_d)^2)^{1/2} / Y_d,$$
$$\text{ARB}_d = 100 \left| 1/r \textstyle\sum_{i=1}^{r} (\hat{Y}_{di} / Y_d - 1) \right|, \tag{24}$$

where $\hat{Y}_{di}$ is the design- or model-based estimate for the area total Y_d in the simulated sample i (from 1 to r). Their means over D areas are:

$$\text{MRRMSE} = 1/D \textstyle\sum_{d=1}^{D} \text{RRMSE}_d,$$
$$\text{MARB} = 1/D \textstyle\sum_{d=1}^{D} \text{ARB}_d. \tag{25}$$

The sum $\hat{Y}_i = \sum_{d=1}^{D} \hat{Y}_{di}$ is the estimate for the population total in sample i. The relative root mean square error for the population total is

$$\text{RRMSE(pop)} = 100(1/r \textstyle\sum_{i=1}^{r} (\hat{Y}_i - Y)^2)^{1/2} / Y, \tag{26}$$

where Y is the true value of the population total. The absolute relative bias is

$$\text{ARB(pop)} = 100 \left| 1/r \textstyle\sum_{i=1}^{r} (\hat{Y}_i / Y - 1) \right|. \tag{27}$$

In PI, we use two other measures (Rao 2003) for the comparisons. They are absolute relative error (ARE) and average relative efficiency (EFF). The mean absolute relative error over the areas and the overall efficiency are computed as

$$\text{MARE} = 1/D\sum_{d=1}^{D}100(1/r\sum_{i=1}^{r}\left|\hat{Y}_{di}-Y_d\right|/Y_d$$

$$\text{MEFF} = 100(\text{MMSE}(pst)/\text{MMSE}(est))^{1/2}, \tag{28}$$

where

$$\text{MMSE}(est) = 1/D\sum_{d=1}^{D}1/r\sum_{i=1}^{r}(\hat{Y}_{di,est}-Y_d)^2 \tag{29}$$

is computed of r simulated estimates $\hat{Y}_{di}$ (i = 1,…, r) for each area-specific estimator $\hat{Y}_{d,est}$. In case of post-stratified estimator (3), MMSE(*pst*) is obtained from (29) by using $\hat{Y}_{d,pst}$ instead of $\hat{Y}_{d,est}$. The quantity MEFF is the relative efficiency of an estimator, compared with the post-stratified estimator. The higher MEFF, the more efficient estimator.

4 RESEARCH CONTRIBUTION

In this chapter, we first describe our register data sets and then compare each of our model-based allocations with the reference allocations. Finally, we compare our model-based allocations under the population structure used in PIII-PIV.

4.1 Real register data sets for simulation experiments and for sample allocations

The first register data set, used in PI, is collected in 2007 and consists of 400 Finnish municipalities (units here) in 19 provinces (areas here). The area-specific number of units varies from 9 to 53. The variable of interest y is the number of unemployed people, and the auxiliary variable x is the number of private houses. The within-area variation of the variables is considerable, but they have a small between-area variation, under 10% of the total variation. The area-specific correlations between the variables are highly positive.

The second research data set, used in PII, is obtained in spring 2011 from a national Finnish register of block apartments for sale. The 9,815 units in the data set are apartments with completed construction in 14 Finnish districts, serving as areas here. The data set covers approximately 50% of all apartments in the register. The number of area-specific units varies from 112 to 1,333. The variable of interest y is the price (in 1,000 €) and the auxiliary variable x is the size (in m^2). The area-specific characteristics of the variables vary considerably. The between-area variations of the variables are moderately large. Most area-specific correlations between the variables are highly positive.

We use the same, two research data sets in PIII and in PIV. The units are block apartments, with completed construction, for sale. The apartments have been extracted in April 2015 and in October 2015 from a national register maintained by the same company as in PII. The sizes of the data sets in units are 22,230 and 21,025. Both data sets cover 18 Finnish provinces (areas here) and include the same variables. The variable of interest y is the price (in 1,000 €) and

the auxiliary variables are size in m² (x_1) and age in years (x_2). The within-area variations of y and x_2 are considerably large in both data sets. The between-area variation is quite high for y, but small for the auxiliary variables. The correlation between y and x_1 is positive and negative between y and x_2. In PIV, we justify the suitability of the former data set (proxy data) for providing data for the allocations.

4.2 Experimental allocation and three reference allocations

For the experimental allocation (Subsection 3.1), seven areas became zeros, and the sample sizes for the other areas varied from 3 to 7.

The equal, proportional, and simple random sampling (no fixed sample sizes) allocations were references. We simulated 1,500 design-based random samples for each allocation and computed the sample-specific EBLUP area total estimates. We also computed the post-stratified (PST) area estimates from the samples obtained by using the SRSWOR allocation. We used the means over the areas for absolute relative error (ARE), absolute relative bias (ARB) and absolute relative efficiency (EFF) to evaluate the allocations.

The experimental allocation performs best, despite the sample size zero for seven areas. There is a restriction, that the area-specific quality measures are not shown in PI. The re-examination of the simulation experiments reveals that the qualities of the area estimates do not generally depend on the sample sizes, but they are related to the similarity between the area-specific and population characteristics. The satisfactory estimates for some areas with zero sample sizes suggest that also the model and the estimator are important to incorporate in the allocation solution.

4.3 Analytical *g1* allocation and six reference allocations

In PII, we have developed an allocation based on the first term g_{1d} of the mean-squared error estimator (11). The adjusted homogeneity coefficient ρ, obtained of the auxiliary variable x, replaces the unknown intra-area correlation (8) in formula (14) to compute the sample sizes.

We use six reference allocations. Two of them are equal and proportional allocations. The Neyman, Bankier, and nonlinear programming (NLP) allocations use the area-specific parameter information of the auxiliary variable x. The Molefe and Clark allocation also uses parameter information and the homogeneity coefficient of x. We obtain two sample size combinations for this allocation, because of two priorities for the population-level estimation.

The overall sample size is 112 (relative size 1.1%). Two areas for the *g1* allocation have sample size zero. When the population priority is ignored for the Molefe and Clark allocation *(G*=0), the sample sizes are only slightly related to the sizes of the areas. The Neyman allocation is concordant with the sizes of the areas, unlike the Bankier and the NLP allocations.

We evaluate the performances of the allocations both at the area and at the population levels in terms of the relative root mean square error (RRMSE) for the accuracy and the absolute relative bias (ARB). The variation in the biases is larger than in the accuracies. The RRMSE means over the areas vary to some extent, but the population accuracies vary only little. Considerable differences between the bias means over the areas and between the population biases appear.

Considering the accuracies and the biases, the *g1* allocation performs the best under the structure characteristic to the register data set. The Molefe and Clark and the parameter-based allocations do not perform well. The results demonstrate that the model-based estimates can be quite accurate, despite the small or even zero sample sizes, and that the quality of the area estimates is related to the area characteristics. The fact that the *g1* allocation does not incorporate the within-area variation, may cause problems under the population structure with small between-area variation and diverging area characteristics.

4.4 Calibrated *g1* allocation and five reference allocations

In PIII, the overall sample size is 216 (1.0% of N = 21,025). A typical feature for the two research data sets is the dominating area, the province of Uusimaa, with the relative size over 32% and the highest price level among the areas.

The sample sizes of the calibrated *g1* allocation (Subsection 3.1) differ considerably from the computational sample sizes of the *g1* allocation (not compared). The five reference allocations are the equal, proportional, NLP, and Neyman allocations, and the Molefe and Clark allocation. The population-level estimation has no priority for the Molefe and Clark allocation, to avoid large sample size for the largest area. For the Neyman allocation, we raised the computational sample sizes of two areas from 1 to 2, to enable the unbiased variance estimation, and applied the formula (20) to the other areas. The NLP allocation has a loose relationship with the sizes of the areas.

To compare three types of estimators, we applied the model-based EBLUP estimation to the calibrated *g1* and to the Molefe and Clark allocations, and the design-based Horvitz-Thompson and the model-assisted regression estimation to the other allocations. The assisting model is the unit-level model (7).

The calibrated *g1* and the Molefe and Clark allocations perform best in accuracy. The model-assisted regression estimates are more accurate than the Horvitz-Thompson estimates for the equal and for the NLP allocations, contrary to the proportional and the Neyman allocations. The design-based estimates are

almost unbiased, as expected. Some considerable biases for the calibrated *g1* and for the Molefe and Clark allocations appear. Considering the integration of accuracy and bias, the equal and the NLP allocations under the model-assisted regression estimation perform best.

None of the studied allocations has the uniquely best performance. Despite the partly contradictory performances, the results support the incorporation of the model and the estimator in the allocation. The calibrated *g1* allocation takes the within-area variation into account, but for further development, more than one term of the MSE estimator (11) must be incorporated in the model-based allocation. To reach accurate area and population estimates, it is necessary to apply the multi-objective optimization technique. The new allocation must also be tested under various population structures.

4.5 Three-term Pareto allocation and five reference allocations

In PIV, we use the same register data sets for the same purposes as in PIII. We have developed the three-term Pareto method allocation, where the sample sizes result from the Pareto optimal solution (Subsection 3.1).

Two of the four model-free reference allocations are the equal and the NLP allocations. We have modified the Neyman allocation by using the box-constraint technique (Subsection 3.1), which guarantees the minimal sample size 2 for every area. The value of the constant k for the Costa allocation (23) indicates that the equal allocation has more weight than the proportional allocation.

The sample sizes of the three-term Pareto allocation are not related to the sizes of the areas. The Molefe and Clark and the NLP allocations have the same sample sizes as in PIII. The box-constraint allocation is close to the computational Neyman allocation. The Costa allocation is concordant with the sizes of the areas, but is far from the proportional allocation.

To diversify the comparison, we applied the model-based EBLUP estimation also for the model-free allocations, in addition to the design-based estimation. We evaluated 14 different allocation and estimation method combinations.

The model-based EBLUP estimates are more accurate than the design-based estimates. The three-term Pareto method performs the best in accuracy. The Molefe and Clark allocation performs almost as well. The model-assisted regression estimates are more accurate than the Horvitz-Thompson estimates only for the NLP and for the Costa allocations.

The design-based estimates are almost unbiased. Severely biased model-based estimates occur for each allocation, except for the three-term Pareto allocation with the area-specific biases under 10%. The aggregate biases obtained from the model-based estimates are the smallest for the three-term Pareto and for the Costa allocations. The integrated values of accuracy and bias indicate better performances for the Costa, the NLP, and the equal allocations under the

regression estimation than for the three-term Pareto allocation, but this is not supported by the inaccurate area estimates for the model-free allocations.

The results confirm that our model-based allocation performs better, when the mean-squared error estimator (11) is more completely incorporated in the allocation. However, it must be tested under different population structures, and the biases also need attention. Biased estimates appear also for this allocation, possibly due to model misspecification. We have tested the validity in PIV.

4.6 Comparison of the model-based allocations under the latest population structure

We compare the performances of the model-based allocations, developed in PI-PIV, under the register data sets used in PIII-PIV. We obtained the experimental allocation in the same way as in PI. Table 1 shows the area-specific sample sizes for the four allocations and for the proportional allocation, which is a reference.

TABLE 1 Area-specific sample sizes for model-based allocations and for proportional allocation

Area (province)	Size in units	Proportional	Experimental	Analytical g1	Calibrated *g1*	Three-term Pareto method
Uusimaa	6,813	69	70	90	43	36
Pirkanmaa	2,003	20	21	23	12	13
Varsinais-Suomi	1,543	16	16	17	18	11
Päijät-Häme	1,166	12	11	12	14	9
Central Finland	1,141	12	12	12	9	11
North Ostrobothnia	1,131	12	13	11	11	9
Satakunta	1,017	10	9	10	11	12
Kymenlaakso	929	10	8	9	8	14
Pohjois-Savo	923	9	10	8	12	10
Kanta-Häme	885	9	8	8	9	11
Etelä-Savo	751	8	8	6	9	10
South Karelia	553	6	4	3	10	11
North Karelia	549	6	7	3	12	6
Lapland	544	6	7	3	10	11
Ostrobothnia	421	4	6	1	8	9
South Ostrobothnia	311	3	4	0	8	9
Kainuu	185	2	2	0	5	15
Central Ostrobothnia	160	2	0	0	7	9
Total	21,025	216	216	216	216	216

The largest area (Uusimaa) has very high sample sizes for three allocations. The experimental allocation is very close to the proportional allocation. In the *g1* allocation, three smallest areas have the sample size zero. The structures of

the calibrated *g1* and of the three-term Pareto allocations are quite similar, but considerable differences between the sample sizes occur for some areas.

Figure 2 shows the allocation-specific means over the areas and the population values for relative root mean square error and for absolute relative bias.

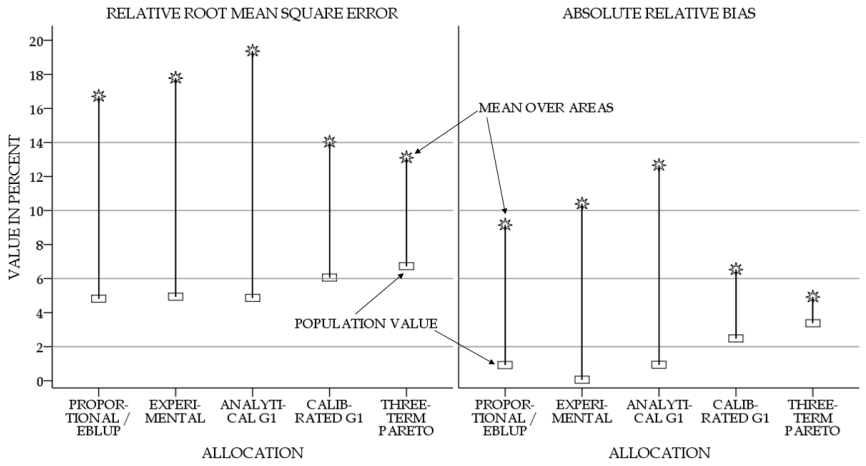

FIGURE 2 Means over the areas and population values for relative root mean square errors and for absolute relative biases, by allocation

The calibrated *g1* and the three-term Pareto allocations have little higher population values for both measures, but considerably smaller means over the areas than the three other allocations. These two allocations are also good trade-offs between the area and the population level, with respect of the quality of the estimates. Considering the accuracies and the biases simultaneously, the three-term Pareto allocation performs the best. Figure 2 also demonstrates the improvement in the model-based allocations which were developed in PI-PIV.

Figure 3 shows the area-specific distributions of the quality measures for each allocation. The randomness in the areas estimates is in the best control for the three-term Pareto allocation. This allocation has tighter distributions and has no outliers, compared with the other allocations. It is the only allocation with the area-specific relative root mean square errors under 20% and absolute relative biases less than 10%. Two small areas with sample size zero are outliers, and one of these areas is an outlier for three allocations, despite the positive sample size. The third small area with zero sample size is not an outlier. The outlier areas have diverging characteristics, compared with the population, unlike the third small area.

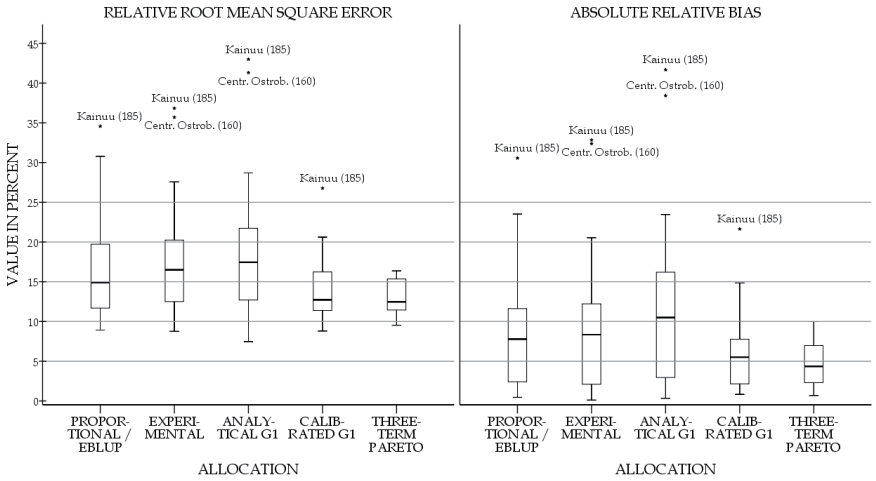

FIGURE 3 Area-specific distributions of relative root mean square errors and of absolute relative biases, by allocation. The outlier areas contain their sizes in units

The results of this subsection demonstrate that the performance of our model-based allocation has improved during the development process. The multi-objective optimization is one of the key factors in the improvement.

5 AUTHOR´S CONTRIBUTION

The author of this dissertation did the majority of work regarding each article. He acquired the real research data sets and carried out the simulation experiments and the necessary calculations. The author has analyzed the results. He has written the text of these publications.

The author presented the idea of an experimental allocation during the sixth "Survey Sampling in Economic and Social Research" Conference, arranged in 2009 by Katowice University, Poland. The presentation, article PI, was accepted and published in the printed monograph of the conference presentations in 2010. The author´s contribution to the article PII is the development of the *g1* allocation. His contribution to the article PIV is the development of the three-term Pareto allocation, the application of the Costa allocation using multi-objective optimization, and the replacement of the Neyman allocation by the box-constraint allocation.

With all articles, the first co-author assisted in the design of the articles, in methodological issues, and in reporting. His contribution to the article PI is the development of the experimental allocation, to the article PIII the development of the calibrated *g1* allocation, in co-operation with the author.

The contribution of the second co-author in PIV is the guidance in the implementation of the multi-objective optimization. He wrote the part of PIV dealing with the idea of the Pareto optimal solution.

6 DISCUSSION

A well-performing allocation is very case-specific and depends on many features of a survey. The overall sample size is fixed and small in relation to the size of the population. The problem is to resolve, what kind of area sample size combination leads to maximal quality for the area and for the population estimates, subject to boundary conditions.

We have shown that the three-term Pareto allocation, which incorporates the model and the estimator more completely, performs better than our other model-based allocations. The *g1* allocation is closely related to the between-area variation and performs generally well at the population level, but it may lead to inaccurate estimates for small areas, when the between-area variation is small. The calibrated *g1* allocation incorporates the within-area variations, but it is possible that the estimates of the areas with small variation are inaccurate.

The three-term Pareto allocation performs the best in accuracy, compared with the competing alternatives the Costa, and the nonlinear programming (NLP) allocations. The Pareto estimates are not considerably biased. The three-term Pareto and the Costa allocations are based on multi-objective optimization and use a fixed overall sample size, but the NLP allocation requires the adjusted limits for the accuracies of the area and of the population estimates, until the fixed overall sample size is reached.

The model-assisted Molefe and Clark allocation does not perform well. The fact that it is based on a composite estimator and on a different area model, may affect the performance to some extent. Furthermore, different choices of the adjustable priorities for the area- and for the population-level estimation may lead to diverging area-specific sample sizes.

The weaknesses of the equal, proportional, and the Neyman allocations have been demonstrated. They are not serious alternatives. The box-constraint allocation is an alternative to the Neyman allocation. However, the adjustment of the lower and upper sample size limits, with the intention of reaching satisfactory area and population estimates, may be laborious.

A condition for the application of the three-term Pareto allocation is the availability of past register data. In the absence of past data, another allocation

must be employed. If at least one auxiliary variable correlates highly with the variable of interest, it provides data for the allocation. The *g1* and the calibrated *g1* allocations are two alternatives in this situation.

The results support the incorporation of the model and the estimator in the allocation. Model-based estimates with satisfactory quality can be obtained, despite the small overall sample size. This is important also from the economical point of view. Tight resources are common for example in business surveys.

The multi-objective optimization is a feasible method for solving a problem which includes conflicting estimation objectives. Through this method, it is also possible to get a conception of the different weighting options, which can be set on the objectives. The three-term Pareto allocation is a trade-off between the qualities of the area- and the population-level estimates, but it is also a trade-off between the accuracy and the bias. These trade-offs have been discussed in the literature, but it seems difficult to find generally accepted balances regarding these trade-offs.

In our articles, we have not considered the possibility of nonresponse or missing survey data. If at least one of them occurs, this may cause serious problems in the estimation phase, especially when the overall sample size is small. Two common techniques for dealing with nonresponse are weighting adjustment (Särndal, Swensson, and Wretman 1992) and imputation (Rubin 1987). In this survey framework, it is also possible to produce values substituting for the missing y-values by using the model-based estimator and auxiliary data, including proxy data. The last method is a serious alternative under the assumption that the underlying model is valid.

We obtained these results under very demanding circumstances, where there is a large variety in the sizes of the areas and in the within-area characteristics. If the sizes of the areas are closer to each other, the performance of the equal allocation very likely improves. This single example demonstrates that the three-term Pareto allocation must be tested under diverse population structures, and the testing also should include the validity check of the used model. Furthermore, it would be interesting to combine a robust estimation technique and the three-term Pareto method for reducing the biases in the area estimates. This may lead to more complex multi-objective optimization.

YHTEENVETO (FINNISH SUMMARY)

Pienaluemalliin, estimaattoriin sekä apumuuttujatietoon ehdollistettu optimaalinen otoskiintiöinti

Tässä väitöskirjassa on tutkittu pienalue-estimointiin liittyvää optimaalista otoskiintiöintiä, kun tavoitteena on saada tulosmuuttujan kokonaismäärälle mahdollisimman tarkat ennusteet sekä perusjoukon että perusjoukon osajoukkojen (alueiden) tasolla. Ongelma on ratkaistava monitavoiteoptimoinnin avulla. Peruslähtökohta on ennalta suunnitellut ja toisensa poissulkevat alueet, ositettu otanta, jossa alueet ovat ositteita, sekä rajallisista aika- ja budjettiresursseista aiheutuva kiinteä ja pieni kokonaisotoskoko, joka on yleistä markkinatutkimuksissa. Jonkin perinteisen otoskiintiöintimenetelmän soveltaminen saattaa tuottaa joillekin alueille niin matalan otoskoon (jopa nolla), ettei suoraa asetelmaperusteista estimointia voida soveltaa. Tässä väitöskirjassa otoskiintiöinti perustuu valittuun malliperusteiseen estimaattoriin ja apumuuttujatietoon.

Väitöskirjassa on kehitetty uusia malliperusteisia otoskiintiöintejä, jotka perustuvat yksikkötason lineaariseen sekamalliin, malliin pohjautuvaan estimaattoriin sekä rekisteripohjaisen apumuuttujatiedon käyttöön. Viimeisin kiintiöinti perustuu tulosmuuttujan kokonaismäärän yleisesti käytetyn EBLUP-estimaattorin (empiirisesti paras lineaarinen harhaton ennustin) keskineliövirheen (MSE) estimaattorin kolmeen termiin sekä Pareto-optimointitekniikkaan.

Väitöskirjan jokaisessa artikkelissa esitellään uusi kiintiöinti, jonka suorituskykyä verrataan kirjallisuudesta poimittuihin kiintiöinteihin reaalidataa käyttävien asetelmaperusteisten otossimulointien avulla. Suorituskykyä mitataan otoksista laskettujen tarkkuutta ja harhaa ilmaisevien laatumittarien avulla. Vertailukiintiöintien valintaperusteena on niiden optimointikriteerien erilaisuus. Artikkeleissa PIII ja PIV on käytetty myös asetelmaperusteista suoraa estimointia ja malliavusteista estimointia vertailun monipuolistamiseksi.

Malliperusteinen estimointi tuottaa tarkemmat alue-ennusteet kuin asetelmaperusteinen estimointi. Malliperusteisten alue-estimaattien laatu on selkeästi yhteydessä alueiden ominaisuuksiin. Viimeksi kehitetty malliperusteinen, Pareto-optimointia käyttävä kiintiöinti, johtaa tarkimpiin alue-estimaatteihin, joiden harha on kohtalaisen vähäinen. Kilpailukykyisimmät vaihtoehdot ovat Costa-kiintiöinti sekä epälineaariseen ohjelmointiin perustuva NLP-kiintiöinti. Harkinnanvaraiset perusjoukon sekä aluetason estimoinnin painotukset rajoittavat joidenkin vertailukiintiöintien käyttökelpoisuutta.

Tulokset tukevat otoskiintiöinnin ehdollistamista käytettyyn aluemalliin sekä siihen perustuvaan estimaattoriin. Uusinta malliperusteista kiintiöintiä on kuitenkin testattava aluerakenteeltaan erilaisissa perusjoukoissa. Kiintiöinnin suunnittelun yhteydessä on tärkeää pohtia järkevää tasapainoa toisaalta aluetason ja perusjoukon tason estimoinnin välillä ja toisaalta tarkkuuden ja harhan välillä. Myös valitun mallin sopivuus tutkittavan ilmiön kuvaamiseen on tarkistettava. Olisi myös mielenkiintoista selvittää, voitaisiinko otoskiintiöinti ja robusti estimointi yhdistää tarkkuuden parantamiseksi ja harhan pienentämiseksi.

REFERENCES

Bankier, M.D. 1988. Power allocations: Determining sample sizes for subnational areas. The American Statistician 42, 174–177.

Battese, G. E., Harter, R. M., and Fuller, W. A. 1988. An error component Model for Prediction of County Crop Areas using Survey and Satellite Date. Journal of the American Statistical Association 83, 28-36.

Brackstone, G.J. 2002. Strategies and approaches for small area statistics. Survey Methodology 28, 117–123.

Burgard, J.P., Münnich, R., and Zimmermann, T. 2014. The impact of sampling designs on small area estimates for business data. Journal of Official Statistics 30 (4), 749–771.

Choudhry, G.H., Rao, J.N.K., and Hidiroglou, M.A. 2012. On sample allocation for effective domain estimation. Survey Methodology 38, 23–29. DOI: http://www.statcan.gc.ca/pub/12-001- x/2012001/article/11682-eng.pdf.

Cochran, W.G. 1977. Sampling Techniques. (3rd edition). New York: John Wiley & Sons.

Costa, A., Satorra, A., and Ventura, E. 2004. Improving both domain and total area estimation by composition. SORT 28 (1), 69–86.

Falorsi, P.D. and Righi, P. 2008. A balanced sampling approach for multi-way stratification for small area estimation. Survey Methodology 34: 223–234. DOI: http://www.statcan.gc.ca/pub/12-001-x/2008002/article/10763-eng.pdf.

Federal Committee on Statistical Methodology 1993. Indirect Estimators in Federal Programs. US Office of Management and Budget, Statistical Policy Working paper No. 21.

Friedrich, U., Münnich, R., and Rupp, M. 2018. Multivariate Optimal Allocation with Box-Constraints. Austrian Journal of Statistics 47, 33–52.

Gabler, S., Ganninger, M., and Münnich, R. 2012. Optimal allocation of the sample size to strata under box constraints. Metrika 75, 151–161.

Gonzales, M.E 1973. Use and evaluation of synthetic estimates. Proceedings of the Social Statistics Section. American Statistical Association, 33–36.

Keto, M. and Pahkinen, E. 2014. On sample allocation for efficient small area estimation. Book of Abstracts. SAE 2014, Poland: Poznan University of Economics, 50.

Khan, M.G.M., Maiti, T. and Ahsan, M.J. 2010. An Optimal Multivariate Stratified Sampling Design Using Auxiliary Information: An Integer Solution Using Goal Programming Approach. Journal of Official Statistics 26, 695–708.

Lehtonen, R., Särndal, C.E., and Veijanen, A. 2003. The effect of model choice in estimation for domains, including small domains. Survey Methodology 29, 33–44.

Lehtonen, R., Myrskylä, M., Särndal, C.-E., and Veijanen, A. 2006. The role of models in model-assisted and model-dependent estimation for domains

and small areas. Working paper, BNU Workshop, Ventspils, Latvia, August 2006.

Lehtonen, R. and Veijanen, A. 2009. Design-Based Methods of Estimation for Domains and Small Areas. In Handbook of Statistics, Vol. 29B, 219-249. New York: Elsevier.

Longford, N.T. 2006. Sample Size Calculation for Small-Area Estimation. Survey Methodology 32, 87–96. DOI: http://www.statcan.gc.ca/ pub/12-001-x/2006001/article/9259-eng.pdf.

Marker, D. A. (2001). Producing Small Area Estimates from National Surveys: Methods for Minimizing Use of Indirect Estimators. Survey Methodology 27, 183–188.

Meza, J.L. and Lahiri, P. 2005. A note on the C_p statistic under the nested error regression model. Survey Methodology 31 (1), 105-109.

Miettinen, K. 1999. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, Boston.

Molefe, W. B. and Clark, R. G. 2015. Model-assisted optimal allocation for planned domains using composite estimation. Survey Methodology 41(2), 377–387. DOI: http://www.statcan.gc. ca/pub/12-001-x/2015002/article/14230-eng.pdf.

Neyman, J. 1934. On the two different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. Journal of the Royal Statistical Society 97: 558-625. DOI: http://dx.doi.org/10.2307/2342192.

Nissinen, K. 2009. Small Area Estimation With Linear Mixed Models From Unit-Level Panel and Rotating Panel Data. Ph.D. thesis, Department of Mathematics and Statistics, University of Jyväskylä, Report 117. DOI: https://jyx.jyu.fi/dspace/handle/123456789/21312.

Pfeffermann, D. and Sverchkov, M. 2004. Prediction of Finite Population Totals Based on the Sample Distribution. Survey Methodology 30 79–92.

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

Rubin, D.B. 1987. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, Inc.

Rao, J. N. K. 2003. Small Area Estimation. Hobogen, New Jersey: Wiley.

Rao, J. N. K. and Molina, I. 2015. Small Area Estimation (2nd Edition). Hoboken, NJ: John Wiley & Sons, Inc.

Singh, M.P., Gambino, J., and Mantel, H.J. (1994). Issues and strategies for small area data. Survey Methodology 20, 3–22.

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

Tschuprow, A. A. 1923. On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. Metron 2: 461-493, 646-683.

ORIGINAL PAPERS

I

ON SAMPLE ALLOCATION FOR EFFECTIVE EBLUP ESTIMATION OF SMALL AREA TOTALS – "EXPERIMENTAL ALLOCATION"

by

Mauno Keto and Erkki Pahkinen, 2010

Published in the printed monograph of the presentations during the 6th Conference on Survey Sampling Methods in Economic and Social Research, pp. 27-36.

Published in the printed monograph prepared of the presented papers during the sixth 'Survey Sampling Methods in Economic and Social Research' Conference on 21 and 22 September 2009, in Katowice University, Poland.

Mauno Keto, Erkki Pahkinen[1]

On sample allocation for effective EBLUP estimation of small area totals – "Experimental Allocation"

ABSTRACT

The demand of regional or small area statistics produced from large-scale surveys such as Finnish unemployment survey has raised needs for developing the tools of optimal sample allocation on area level. However, most commonly used allocation methods aim at producing efficient direct areal estimates. What typically happens is that several areas receive little or none observations, and therefore one has to resort to indirect estimation methods. Best-known of these methods and perhaps most widely used are nested-error regression type model-based estimators. For this reason should areal sample allocation be implemented in such a way that it would lead to efficient estimation in the case of an indirect estimator. In this research an attempt was made to solve this problem by developing an experimental allocation method through simulations. The functioning of this new method has been tested by comparing it to equal and proportional allocation. Different allocation criteria such as values of average absolute relative bias and efficiency are examined through simulation studies with Finnish unemployment data.

Introduction

We plan sampling designs generally for efficient estimation on the population level. However, the same demand of efficiency prevails if one wants to calculate regional or small area statistics from large-scale survey data but now on the level of some subpopulation. Generally, as for basic sampling design, stratified random sampling has been chosen. Strata coincide with areas and the problem is how to allocate stratum-wise fixed sample size n.

Optimal allocation has inspired for different solutions during the last decades. Main line has prevailed to find areal allocation giving possibility to calculate direct or model-assisted direct estimators for each area. Some

[1] Mikkeli University of Applied Sciences, University of Jyväskylä, Finland.

examples from earlier efforts are reported in Rao [2003]. Recently published interesting proposition come from Longford [2006], who includes inferential priority index P_d for each area and tries then to find optimality. Another solution comes from Falorsi and Righi [2008]. They assume that direct estimators should be model-assisted and their optimal allocation procedure accounts for this possibility with other prior information used in planning sample design.

The next two sections describe simulation experiments where we have searched an areal allocation scheme conditional to auxiliary information which includes both auxiliary variables and model for indirect estimation of fixed areal totals. Indirect or model-based estimation has been chosen because in small area calculations domains with few or none observations are general. This fact has been profoundly investigated recently by Lehtonen et. al. [2003]. As a model, EBLUP has been chosen because there is a lot of evidence that this model works well in many small area estimation situations.

1. Experimental allocation

Our study data has been obtained from the Finnish unemployment survey in 2007 described in Nissinen [2009]. The population consists of all 400 (=N) local municipalities (population elements) in 19 different provinces (areas). We had originally all 20 provinces of Finland and 416 municipalities in the population, but we dropped one province out, Åland, because it is totally different from the others having own administrative autonomy. After that, the data include 19 areas and 400 municipalities.

The number of municipalities on areal level varied from 9 to 53, and average area size was 21. We used the number of unemployed people as the study or outcome variable (y) and the number of private houses in provinces as the auxiliary variable (x). The overall correlation of these variables was moderate, 0.68. The area total for variable (y) varied from 2540 to 43639 and the corresponding total for variable (x) varied from 9708 to 48720. CV value for variable (y) ranged between 1.13 and 3.80 and for auxiliary variable (x) between 0.48 and 1.18. The overall mean of the number of unemployed people (y) was 541 and for private houses (x) 1290. The differences between areas (provinces) were significant.

In our research we have used model-based or indirect estimation with one outcome variable and one auxiliary variable. Area effect is also included. The model is a nested-error regression, basic unit level model

$$y_{dk} = \mathbf{x}'_{dk}\boldsymbol{\beta} + v_d + e_{dk};\ \ k = 1,...,N_d; d = 1,...,D, \tag{1}$$

which is a special case of well-known general mixed linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{v} + \mathbf{e}, \tag{2}$$

where **y** is n×1 vector of sample observations, **X** and **Z** are known n×p and n×h matrices of full rank, and **v** and **e** are independently distributed with means **0** and covariance matrices **G** and **R** depending on some variance parameters $\boldsymbol{\delta} = (\delta_1,...,\delta_q)'$. Furthermore, Var(**y**) = **V** = **R** + **ZGZ**′ is the variance-covariance matrix of **y**.

In model (1) y_{dk} is the kth value in area d for outcome variable (y), $\mathbf{x}_{dk}$ is the vector of auxiliary variables (x) in area d, v_d is the random effect of area d (d = 1,…, D) in the model and is estimated from the observations, and e_{dk} is a random error. Random effects v_d and random errors e_{dk} are assumed to be independent of each other and (not necessarily) distributed as $N(0,\sigma_v^2)$ and $N(0,\sigma_e^2)$. Furthermore, regression coefficients $\boldsymbol{\beta}$ are estimated from the observations.

Following matrix forms are used in estimation calculations:

1. Data matrix **y** for outcome variable (y).
2. Data matrix **x** for auxiliary variable (x) (so-called random-intercept form with ones in the first column).
3. Variance-covariance matrix $\mathbf{V_{n\times n}}$ of outcome variable (y).

Variances σ_v^2 and σ_e^2 can be estimated in many ways: ML or REML method or "Fitting-of-constants" attributed to Henderson as explained in Rao [2003]. Variance σ_v^2 can be negative in which case it is set to zero.

In our research we have used the EBLUP (Empirical Best Linear Unbiased Predictor) estimator for area totals. First the BLUP (Best Linear Unbiased Predictor) estimator of area total Y_d is simply the sum of sample observations and predicted values of non-sampled observations of variable (y) as given in Rao [2003]:

$$\hat{Y}^H_{d,BLUP} = \sum_{k\in s_d} y_{dk} + \sum_{k\in \bar{s}_d} \tilde{y}_{dk} = \sum_{k\in s_d} y_{dk} + \sum_{k\in \bar{s}_d} \mathbf{x}'_{dk}\hat{\boldsymbol{\beta}} + (N_d - n_d)\tilde{v}_d\,. \tag{3}$$

In expression (3) s_d denotes the sample from area *d*, and $\bar{s}_d$ denotes the non-sampled observations from area *d*. Furthermore, $\hat{\boldsymbol{\beta}}$ is the BLUE (Best Linear Unbiased Estimator) of $\boldsymbol{\beta}$ and $\tilde{v}_d$ is the BLUP of area effect v_d. Expression (3) contains also the sum of values of auxiliary variable (x) for non-sampled observations, but individual values are not needed. Finally, when variance components σ_v^2 and σ_e^2 are substituted by their estimates, we get the EBLUP estimator of area total defined in (3).

The estimates of regression coefficients $\boldsymbol{\beta}$ and area effects **v** are obtained as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad \text{and} \quad \tilde{\mathbf{v}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \cdot \tag{4}$$

The MSE of estimator $\hat{Y}_d^H$ is the sum of its variance and squared bias:

$$MSE(\hat{Y}_d^H) = E(\hat{Y}_d^H - Y_d)^2 = Var(\hat{Y}_d^H) + (\hat{Y}_d^H - Y_d)^2 \cdot \tag{5}$$

An estimator of MSE approximation in the case of finite populations is given in Rao [2003]:

$$\begin{aligned} \mathrm{mse}(\hat{Y}_d^H) &= (N_d - n_d)^2 \left[g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \right] \\ &+ (N_d - n_d)\, \hat{\sigma}_e^2 , \end{aligned} \tag{6}$$

where the terms $g_{1d} - g_{3d}$ have quite complex expressions.

The first step in our experiments was to draw 1500 SRS samples from the population and fit them into model (1). Each sample included 57 sampling units (3×19 = 57). From these samples we computed EBLUP estimates for (y) totals, MSE´s and its components plus other necessary statistics for each area by using SAS PROC SURVEYSELECT with seed number and PROC MIXED procedures plus by SPSS. Small area estimation include special quality measures as statistics measuring accuracy and bias of samples like Absolute Relative Error (ARE), Average Squared Error (ASE) and Absolute Relative Bias (ARB) were computed with SPSS software. They are defined as follows [Rao 2003]:

$$\text{ARE in one sample} = \frac{1}{D}\sum_{d=1}^{D} \left| \hat{Y}_d - Y_d \right| / Y_d , \tag{7}$$

$$\text{ASE in one sample} = \frac{1}{D}\sum_{d=1}^{D}(\hat{Y}_d - Y_d)^2, \tag{8}$$

$$\text{ARB in one sample} = \frac{1}{D}\left|\sum_{d=1}^{D}(\hat{Y}_d - Y_d)/Y_d\right|, \tag{9}$$

$$\text{CV in one sample} = \frac{1}{D}\sum_{d=1}^{D}\sqrt{MSE(\hat{Y}_d)}/\hat{Y}_d). \tag{10}$$

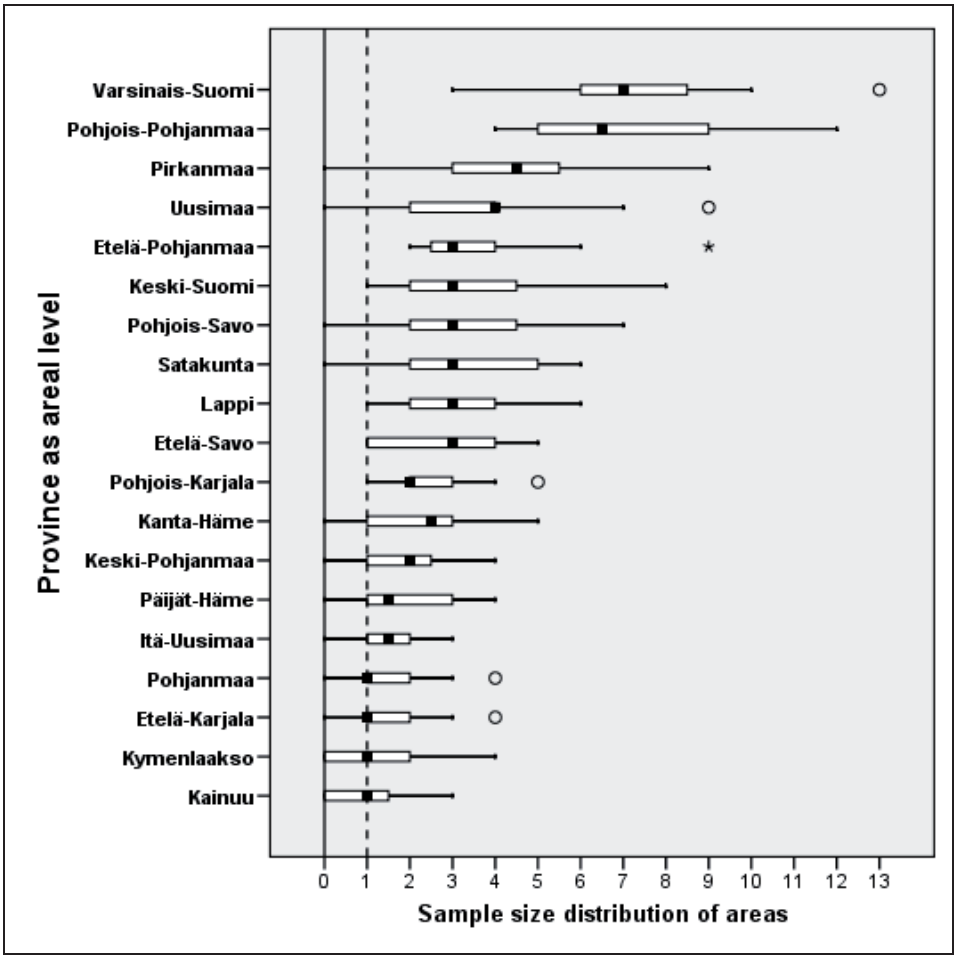

Figure 1. Distribution of areal sample sizes in 20 "best" samples for MSE means

We computed the means of MSE, CV, ARE, ASE and ARB for each sample and arranged the samples in ascending order according to each statistic mean, and so we got 5 different orders for the samples. Finally, we picked 20 "best" samples from each order (lowest means). When we examined the

distribution of areal sample sizes for MSE, CV, ARE and ASE (see Figure 1 for MSE) we could notice that the sample size distributions are quite similar with each other, which suggests that very much the same areas (7) had zero or very little observations, whereas some other areas had very many observations (as much as 10). This finding encouraged us to carry out an experimental sampling allocation in which 57 units in each sample were concentrated on only 12 areas.

In order to be able to compare the efficiency of the radical allocation scheme mentioned earlier (7 zero-areas) with other alternatives, we have used SRS and two other allocation schemes (on the basis of "gut-feeling") which are proportional stratified sampling and equal allocation. The next chapter introduces the results of our experiments.

Table 1

Areal sample sizes in different allocation schemes

Province	Size of area	Sample sizes in allocations			
		Not allocated	Proportional	Equal	Experi-mental
Uusimaa	24	.	4	3	4
Varsinais-Suomi	53	.	7	3	7
Itä-Uusimaa	10	.	1	3	0
Satakunta	25	S	4	3	5
Kanta-Häme	16	R	2	3	0
Pirkanmaa	28	S	4	3	5
Päijät-Häme	12	.	2	3	0
Kymenlaakso	12	s	2	3	0
Etelä-Karjala	12	a	2	3	0
Etelä-Savo	18	m	3	3	3
Pohjois-Savo	23	p	3	3	4
Pohjois-Karjala	16	l	2	3	3
Keski-Suomi	28	e	4	3	5
Etelä-Pohjanmaa	26	s	4	3	6
Pohjanmaa	17	.	2	3	4
Keski-Pohjanmaa	12	.	2	3	0
Pohjois-Pohjanmaa	38	.	5	3	6
Kainuu	9	.	1	3	0

Lappi	21	.	3	3	5
TOTAL:	400	57	57	57	57

The real sample size in our experiment was selected according to following principles:

$$\begin{aligned} n_d &= 0 \quad \text{if } \mathrm{cdf}^{-1}(0.25) \le 1 \\ &\approx \mathrm{cdf}^{-1}(0.75) \ \text{ if } \quad \mathrm{cdf}^{-1}(0.25) > 1. \end{aligned} \tag{11}$$

Notation "cdf" means cumulative distribution function of areal sample size. We call this kind of allocation as "experimental allocation" and its peculiar property is that a lot of areas have no observations as seen in Table 1. Some specifications were made according to other quality measures.

2. Simulation studies

Because analytical solutions for optimizations in nested-error regression model (1) are not possible in general, we have used various simulations to investigate the effect of different areal sample allocations on MSE and quality measures like ARE, ARB, ASE and EFF (Average Relative Efficiency) introduced for example by Rao [2003]. The following table shows the sample sizes of 19 areas in the four selected allocation schemes. In the simulation of samples we followed the principles used by Falorsi and Righi [2008], Longford [2007] and Nissinen [2009].

The sampling procedures (SRS and stratified sampling) as well as calculation of different EBLUP estimates and other statistics were implemented by SAS software (PROC SURVEYSELECT with seed number, PROC MIXED) and SPSS. The number of random samples was 1500 in every allocation scheme.

For each allocation scheme MSE´s, CV´s and quality measures (ARE, ASE, ARB AND EFF) were calculated and their distributions were examined. The distributions of ARE values (95% of all) for each allocation scheme are introduced in the next Figure (in boxplot form):

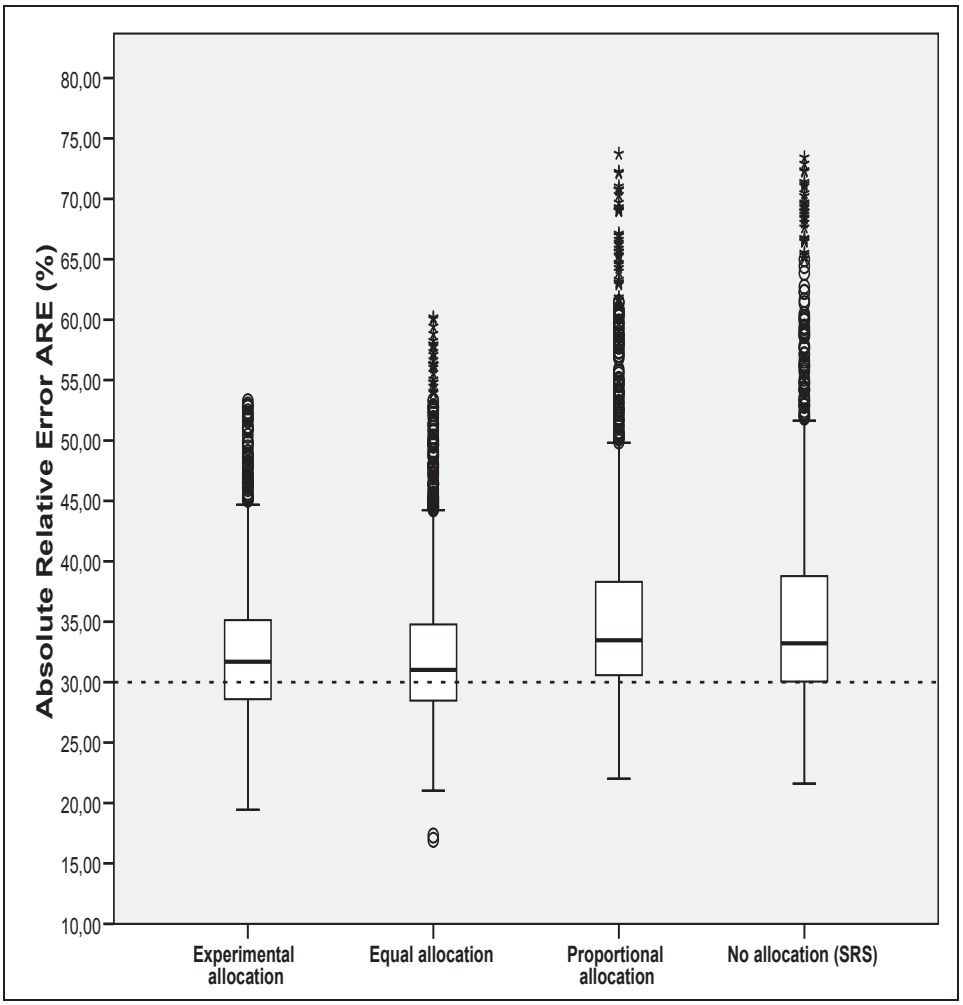

Figure 2. Distributions of 95% of ARE values of samples

As this Figure shows, ARE values in our experimental allocation scheme behave in a more controlled way (range and location) compared with the other schemes. Distributions of MSE, CV, ASE and ARB show very much similar properties for our "own" scheme.

We computed also the following quality measures for every allocation scheme (all samples in the scheme), and they can be used in assessing the accuracy and bias of an estimator of area total proposed in Rao [2003]: Overall ARE (absolute relative error), overall ARB (average absolute relative bias) and overall EFF (average relative efficiency) which compares EBLUP estimation to poststratification estimation. The formulas are:

$$\overline{ARE} = 100 \times \frac{1}{D}\sum_{d=1}^{D}\frac{1}{n}\sum_{i=1}^{n}\left|\hat{Y}_{di} - Y_d\right| / Y_d \,, \tag{12}$$

$$\overline{ARB} = 100 \times \frac{1}{D}\sum_{d=1}^{D}\left|\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_{di} - Y_d) / Y_d\right|, \tag{13}$$

$$\overline{EFF} = 100 \times \left[\overline{MSE}(PST) / \overline{MSE}(est)\right]^{1/2}. \tag{14}$$

In expression (14) $$\overline{MSE}(est) = \frac{1}{D}\sum_{d=1}^{D}\frac{1}{n}\sum_{i=1}^{n}(\hat{Y}_{di,EBLUP} - Y_d)^2 \tag{15}$$

and $\overline{MSE}(PST)$ is obtained by changing EBLUP estimator $\hat{Y}_{di,Eblup}$ to $\hat{Y}_{di,PST}$, where the latter is the value of the post-stratified estimator of the total of area d in i^{th} simulated sample. The following table shows the values for these quality measures for each allocation scheme:

Table 2

Overall ARE, EFF and ARB (%) for different allocations

Allocation scheme	ARE	EFF	ARB
Experimental	34.6%	246.7%	25.8%
Equal	35.1%	242.5%	26.3%
Proportional	39.3%	236.5%	28.6%
Not allocated (SRS)	38.7%	225.5%	27.8%
Poststratification	74.9%	100.0%	11.8%

Also these results confirm our conception of the usefulness of our allocation scheme. According to the quality measures experimental allocation seems to be the best.

Conclusions

First we have to notice that results of our analysis are based on simulations, whereupon the random-function generator used in this process may affect the results. Simulation experiments were our choice for the simple reason that our model, indirect nested-error regression model, is made up of so many variation components that it is difficult, if not impossible, to construct an analytical expression which is valid for optimization of sample sizes.

Another restriction in exploiting the results of our experiments is that they are based on computations made under indirect small-area estimation and EBLUP model.

The concept of optimization in small-area estimation includes several estimated statistics or quality measures (MSE, ARE, ARB etc.), but we treated each of them as equal. Do we have to give more weight to one quality measure and less weight to some other?

In spite of these limited possibilities it seems that experimental allocation as a whole leads to better result measured with the quality measures of small-area estimation compared with other reference allocations which were equal allocation and proportional allocation. However, this method must be tested in the environments of larger data and for example areal allocations applied in official statistics could be used as reference allocations.

Literature

Brackstone G.J. (2002): *Strategies and Approaches for Small Area Statistics.* "Survey Methodology", No. 28, pp. 117-123.

Falorsi P.D., Righi P. (2008): *A Balanced Sampling Approach for Multiway Stratification for Small Area Estimation.* "Survey Methodology", No. 34, pp. 223-234.

Lehtonen R., Särndal C.E., Veijanen A. (2003): *The Effect of Model Choice in Estimation for Domains, Including Small Domains.* "Survey Methodology", No. 29, pp. 33-44.

Longford N.T. (2006): *Sample Size Calculation for Small-Area Estimation.* "Survey Methodology", No. 32, pp. 87-96.

Nissinen K. (2009): *Small Area Estimation With Linear Mixed Models From Unit-Level Panel and Rotating Panel Data.* University of Jyväskylä Departement of Mathematics and Statistics, Report 117.

Rao J.N.K. (2003): *Small Area Estimation.* John Wiley & Sons, New York.

II

SAMPLE ALLOCATION FOR EFFICIENT MODEL-BASED SMALL AREA ESTIMATION

by

Mauno Keto and Erkki Pahkinen, 2017

Survey Methodology Journal, vol. 43(1), pp. 93-106

Survey Methodology, June 2017
Vol. 43, No. 1, pp. 93-106
Statistics Canada, Catalogue No. 12-001-X

Sample allocation for efficient model-based small area estimation

Mauno Keto and Erkki Pahkinen[1]

Abstract

We present research results on sample allocations for efficient model-based small area estimation in cases where the areas of interest coincide with the strata. Although model-assisted and model-based estimation methods are common in the production of small area statistics, utilization of the underlying model and estimation method are rarely included in the sample area allocation scheme. Therefore, we have developed a new model-based allocation named *g*1-allocation. For comparison, one recently developed model-assisted allocation is presented. These two allocations are based on an adjusted measure of homogeneity which is computed using an auxiliary variable and is an approximation of the intra-class correlation within areas. Five model-free area allocation solutions presented in the past are selected from the literature as reference allocations. Equal and proportional allocations need the number of areas and area-specific numbers of basic statistical units. The Neyman, Bankier and NLP (Non-Linear Programming) allocation need values for the study variable concerning area level parameters such as standard deviation, coefficient of variation or totals. In general, allocation methods can be classified according to the optimization criteria and use of auxiliary data. Statistical properties of the various methods are assessed through sample simulation experiments using real population register data. It can be concluded from simulation results that inclusion of the model and estimation method into the allocation method improves estimation results.

Key Words: Optimal area sample size; Criteria; Auxiliary information; Measure of homogeneity.

1 Introduction

In this paper we present a new model-based allocation method in stratified sampling where the areas of interest coincide with the strata. Our study is focused on the components of an efficient area allocation. A clear starting point for the allocation process is reached if the areas of interest are defined as early as in the design phase of the research and if it is also known how large a sample is allowed in consideration of the disposable resources (time, budget etc.). The choice of the allocation method depends on various factors such as the selected model, estimation method, available pre-information of the population and the optimization criteria set only on area or population level, or on both levels simultaneously.

We have selected six existing allocation methods and developed a new one which we call a model-based allocation. The general properties of these methods are examined in Section 2 and Section 3. Five of these allocations can be regarded as model-free. Two of them use only number-based information, such as the number of areas and the number of basic units in each area. Three other allocations need, in addition to number-based information, area level parameter information, such as area totals, standard deviation or coefficient of variation (CV). Because this information about the study variable is not available, a common solution is to replace it with a proper proxy variable. The last of the reference allocations, introduced by Molefe and Clark (MC) (2015), is a model-assisted allocation which is based on a composite estimator and a two-level model. We have named it MC-allocation.

The optimization criteria of the five model-free allocations differ from one another. Allocations based only on area-specific numbers can be computed easily, but their choice is reasonable under limited

1. Mauno Keto, University of Jyväskylä. E-mail: mauno.j.keto@student.jyu.fi; Erkki Pahkinen, Department of Mathematics and Statistics of University of Jyväskylä. E-mail: pahkinen@maths.jyu.fi.

circumstances. In each of the parameter-based allocations the optimization criterion is different. It can be set on the level of the population parameter estimate (Neyman allocation) or on area level estimates in average (Bankier allocation). The third allocation solution, which deviates from the two former ones, is the NLP allocation, in which the tolerances of estimates are set on both population and area level.

This article starts from the assumption that if model-assisted or model-based estimation is used in a survey the model and estimation method must be taken into account when the allocation of the sample into areas is designed. This was used as a starting point when the new model-based $g1-$ allocation, presented in Section 2, was derived. Also, one of the reference allocations, model-assisted allocation, is based on a given model.

The comparison of performances of different allocation methods in real situations has been implemented by using simulation experiments and is presented in Section 4. An official Finnish register of block apartments for sale serves as the population. The structure of the register is introduced in Section 4.1. An auxiliary variable has been used in place of the study variable when computing the area sample sizes for each allocation except equal and proportional allocation. The comparison demonstrates clearly that these allocations lead to different sample distributions. The same kind of variety also concerns their performances. We have applied model-based EBLUP (Empirical Best Linear Unbiased Predictor) estimation on the allocations when estimating the area totals of the study variable. For measuring and comparing the performances of allocations, a relative root mean square error RRMSE% and absolute relative bias ARB% were used.

In Section 5 empirical simulation results are discussed as concluding remarks. They support the allocation solution in which not only auxiliary information, but also the model and estimation method should be determined as early as in the design phase of a survey. A good example is the $g1-$ allocation presented in Section 2.2. The most accurate area estimates of area totals were obtained by using this method.

2 Allocations which utilize the model

2.1 Choosing the model

Pfeffermann (2013) presents a wide variety of models and methods for small area estimation. Our model is one of this assortment, a unit-level mixed model

$$y_{dk} = \mathbf{x}_{dk}'\boldsymbol{\beta} + v_d + e_{dk};\ \ k = 1,\ldots,N_d;\ \ d = 1,\ldots,D, \tag{2.1}$$

where v_d's are random area effects with mean zero and variance σ_v^2 and e_{dk}'s are random effects with mean zero and variance σ_e^2. Furthermore, $E(y_{dk}) = \mathbf{x}_{dk}'\boldsymbol{\beta}$ and $V(y_{dk}) = \sigma_v^2 + \sigma_e^2$ (total variance). Matrix $\mathbf{V}$ is the variance-covariance matrix of the study variable y. This model can be used when unit-level values are available for the auxiliary variables $\mathbf{x}$. We use one auxiliary variable in our study.

Two important measures are needed in developing one of these types of allocations. The first one is a common intra-area correlation ρ and the second one is the ratio δ between variance components. They are defined as follows:

$$\rho = \sigma_v^2 / \left(\sigma_v^2 + \sigma_e^2\right) \text{ and } \delta = \sigma_e^2 / \sigma_v^2 = 1/\rho - 1. \tag{2.2}$$

Before estimating area parameters, the variance components, regression coefficients and area effects must be estimated from the sample data. The BLUE estimator (Best Linear Unbiased Estimator) of $\boldsymbol{\beta}$, noted $\tilde{\boldsymbol{\beta}}$, is obtained according to the theory of the general linear model, and it is replaced with its EBLUP estimate $\hat{\boldsymbol{\beta}}$.

The EBLUP estimate (predicted value) for the area total Y_d of the study variable is the sum of the observed y – values and predicted y – values for units outside the sample:

$$\hat{Y}_{d,\text{Eblup}} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \hat{y}_{dk} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \mathbf{x}'_{dk} \hat{\boldsymbol{\beta}} + \left(N_d - n_d\right) \hat{v}_d. \tag{2.3}$$

We use the Prasad-Rao approximation (See Rao 2003) of MSE (Mean Squared Error) for finite populations:

$$\text{mse}\left(\hat{Y}_{d,\text{Eblup}}\right) = g_{1d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + g_{2d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + 2g_{3d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) + g_{4d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right), \tag{2.4}$$

where the four components g_{1d}, g_{2d}, g_{3d} and g_{4d} are defined as follows:

$$\begin{aligned}
g_{1d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)^2 \left(1 - \hat{\gamma}_d\right) \hat{\sigma}_v^2, \\
g_{2d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)^2 \left(\bar{\mathbf{x}}_d^* - \hat{\gamma}_d \bar{\mathbf{x}}_d\right)' \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \left(\bar{\mathbf{x}}_d^* - \hat{\gamma}_d \bar{\mathbf{x}}_d\right), \\
g_{3d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right)^2 \left(n_d^*\right)^{-2} \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 \left(n_d^*\right)^{-1}\right)^{-3} \left[\hat{\sigma}_e^4 V\left(\hat{\sigma}_v^2\right)\right. \\
&\qquad \left. + \hat{\sigma}_v^4 V\left(\hat{\sigma}_e^2\right) - 2\hat{\sigma}_e^2 \hat{\sigma}_v^2 \text{Cov}\left(\hat{\sigma}_e^2, \hat{\sigma}_v^2\right)\right], \\
g_{4d}\left(\hat{\sigma}_v^2, \hat{\sigma}_e^2\right) &= \left(N_d - n_d^*\right) \hat{\sigma}_e^2.
\end{aligned} \tag{2.5}$$

The area sample sizes n_d^* depend on the sample and are not fixed. The component g_{1d} contains the area-specific ratio $\hat{\gamma}_d = \hat{\sigma}_v^2 / \left(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 / n_d^*\right)$. According to Nissinen (2009, page 53), the g_{1d} component (later simply $g1$) contributes generally over 90% of the estimated MSE. This component represents uncertainty as regards the variation between the areas. Of course this variation must be strong enough so that such a high proportion for $g1$ exists.

Unfortunately, the idea of an analytical solution, which means minimizing the sum of MSE's over areas subject to $n = \sum_{d=1}^{D} n_d$, is difficult and laborious to accomplish because components of the MSE approximation (2.5) include sample information which is unknown, and some components contain complex matrix and variance-covariance operations. We have examined this allocation problem for the first time in an experimental study (Keto and Pahkinen 2009). Now we have developed an allocation based only on the component $g1$ and auxiliary variable x. The reasoning for this solution is that because x and y are correlated, the between-area variation in x is transferred to y.

2.2 Model-based $g1-$allocation

The $g1-$allocation utilizes the auxiliary variable x and the adjusted homogeneity coefficient (Keto and Pahkinen 2014). This coefficient is an approximation of an intra-class correlation (ICC) known of cluster sampling. We regard one area as one cluster. First, simple ANOVA between areas is carried out, and then the adjusted homogeneity measure of variation between the areas can be computed:

$$R_{ax}^2 = 1 - R^2(x) = 1 - \mathrm{MSW}/S_x^2, \tag{2.6}$$

where $R^2(x)$ is the coefficient of determination from regression analysis, MSW (Mean Square within) is the mean SS (Sum of Squares) of areas and S_x^2 is the variance of the auxiliary variable x.

Because MSE of the area total is complex, we use only the component $g1$, which appears in (2.4) and (2.5), for the reason we have given in Section 2.1. We search for the minimum for the sum of $g1$'s over areas:

$$\sum_{d=1}^{D} g_{1d}(\sigma_v^2, \sigma_e^2) = \sum_{d=1}^{D} (N_d - n_d)^2 (n_d/\sigma_e^2 + 1/\sigma_v^2)^{-1} \tag{2.7}$$

subject to $n = \sum_{d=1}^{D} n_d$.

We use Lagrange's multiplier method to find the solution. Therefore, we define the function F of sample sizes $\mathbf{n}' = (n_1, n_2, \ldots, n_D)$ and λ:

$$F(\mathbf{n}, \lambda) = \sum_{d=1}^{D} g_{1d}(\sigma_v^2, \sigma_e^2) = \sum_{d=1}^{D} (N_d - n_d)^2 (n_d/\sigma_e^2 + 1/\sigma_v^2)^{-1} + \lambda\left(\sum_{d=1}^{D} n_d - n\right). \tag{2.8}$$

We set the derivative of F with respect to the area sample size n_d to zero and solve for n_d. The expression for area sample size n_d^{g1} is as follows:

$$n_d^{g1} = \frac{(N_d + \delta)(n + \delta D)}{N + \delta D} - \delta = \frac{N_d n - (N - N_d D - n)(1/\rho - 1)}{N + D(1/\rho - 1)}, \tag{2.9}$$

where the ratio δ and the intra-area correlation ρ are defined in (2.2). The only unknown member in (2.9) is the intra-area correlation ρ. Therefore we substitute the known homogeneity measure (2.6) of the auxiliary variable x for ρ. Thus the final expression for computing area sample sizes is

$$n_d^{g1} = \frac{N_d n - (N - N_d D - n)(1/R_{ax}^2 - 1)}{N + D(1/R_{ax}^2 - 1)}. \tag{2.10}$$

It is easy to prove that $\sum_{d=1}^{D} n_d^{g1} = n$. The computed sample sizes are rounded to the nearest integer. Sometimes compromises must be made. It can be concluded by the examination of (2.10) that the sample size increases when the size of area N_d increases, but not proportionally. Under certain circumstances, such as low homogeneity coefficient, low overall sample size n or small size of area, N_d can lead to negative area sample size n_d^{g1}. In this situation the negative value is changed to zero. A special case occurs if the total variation is only between areas causing value one to the measure of homogeneity (2.6), and (2.10) is reduced to proportional allocation.

2.3 Model-assisted MC-allocation

Molefe and Clark (2015) have used the following composite estimator for estimating the mean of the study variable y for area d:

$$\tilde{y}_d^C = (1-\varphi_d)\,\overline{y}_{dr} + \varphi_d\,\hat{\boldsymbol{\beta}}'\overline{\mathbf{X}}_d. \tag{2.11}$$

This estimator is a combination of two estimators: the synthetic estimator $\hat{\overline{Y}}_{d(\text{syn})} = \hat{\boldsymbol{\beta}}'\overline{\mathbf{X}}_d$, where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient and $\overline{\mathbf{X}}_d$ is the area population means of auxiliary variables $\mathbf{x}$, and a direct estimator $\overline{y}_{dr} = \overline{y}_d + \hat{\boldsymbol{\beta}}'(\overline{\mathbf{x}}_d - \overline{\mathbf{X}}_d)$, where $\overline{y}_d$ and $\overline{\mathbf{x}}_d$ are the area d sample means of y and $\mathbf{x}$. We use one auxiliary variable in our study. The coefficients φ_d are set with the intent to minimize the MSE of the estimator (2.11). The approximated design-based MSE of the estimator under certain conditions and assumptions is given by the expression

$$\text{MSE}_p\left(\tilde{y}_d^C; \overline{Y}_d\right) \approx (1-\varphi_d)^2\, v_{d(\text{syn})} + \varphi_d^2 B_d^2, \tag{2.12}$$

where $v_{d(\text{syn})}$ is the sampling variance of the synthetic estimator $\hat{\overline{Y}}_{d(\text{syn})}$ and $B_d = \boldsymbol{\beta}_U'\,\overline{\mathbf{X}}_d - \overline{Y}_d$ is the bias when $\hat{\overline{Y}}_{d(\text{syn})}$ is used to estimate $\overline{Y}_d$, with $\boldsymbol{\beta}_U$ denoting the approximate design-based expectation of $\hat{\boldsymbol{\beta}}$.

The population contains N units and D strata defined by areas, and stratified sampling is used. A random sample SRSWOR (Simple Random Sampling without Replacement) of n_d units is selected from stratum $d\,(d = 1,\ldots,D)$ containing N_d units. The relative size of area d is $P_d = N_d / N$.

A two-level linear model ξ conditional on the values of $\mathbf{x}$ is assumed, with uncorrelated stratum random effects u_d and random effects ε_i:

$$\left.\begin{aligned} y_i &= \boldsymbol{\beta}'\mathbf{x}_i + u_d + \varepsilon_i \\ E_\xi(u_d) &= E_\xi(\varepsilon_i) = 0 \\ V_\xi(u_d) &= \sigma_{ud}^2 \\ V_\xi(\varepsilon_i) &= \sigma_{ed}^2 \end{aligned}\right\}, \tag{2.13}$$

where i refers to all units in stratum d. This model implies that $V_\xi(y_i) = \sigma_{ud}^2 + \sigma_{ed}^2$ for all population units and $\text{cov}_\xi(y_i, y_j)$ equals $\rho_d\,\sigma_d^2$ for units $i \neq j$ in the same stratum and zero for units from different strata, where $\rho_d = \sigma_{ud}^2 / (\sigma_{ud}^2 + \sigma_{ed}^2)$. A simplifying assumption that $\rho_d = \rho$ are equal for all strata is defined.

After making some other simplifying assumptions and solving the optimal weight φ_d in (2.12), the final approximate optimum anticipated MSE or approximate model assisted mean squared error is obtained of (2.12):

$$\text{AMSE}_d = E_\xi \text{MSE}_p\left(\tilde{y}_d^C\left[\varphi_{d(\text{opt})}\right]; \overline{Y}_d\right) \approx \sigma_d^2 \rho(1-\rho)\left[1 + (n_d - 1)\rho\right]^{-1}. \tag{2.14}$$

Next the criterion F using anticipated MSE's of the small area mean and overall mean estimators for model-assisted allocation is defined and developed into the final approximative form:

$$\begin{aligned} F &= \sum_{d=1}^{D} N_d^q \text{AMSE}_d + G N_+^{(q)} E_\xi\, \text{var}_p\left(\hat{\overline{Y}}_r\right) \\ &\approx \sum_{d=1}^{D} N_d^q\, \sigma_d^2\, \rho(1-\rho)\left[1 + (n_d - 1)\rho\right]^{-1} + G N_+^{(q)} \sum_{d=1}^{D} \sigma_d^2 P_d^2 n_d^{-1}(1-\rho). \end{aligned} \tag{2.15}$$

Optimal sample sizes for the areas are obtained by minimizing (2.15) subject to $\sum_d n_d = n$. Expression (2.15) follows the idea of Longford (2006). The weight N_d^q reflects the inferential priority (importance) for area d, with $0 \le q \le 2$, and $N_+^{(q)} = \sum_{d=1}^{D} N_d^q$. The quantity G is a relative priority coefficient on the population level. Ignoring the goal of estimating the population mean corresponds to $G = 0$, and the attention is then only focused on area level estimation. On the other hand, the larger the value of G, the more the second component in (2.15) dominates and the more the area level estimation is ignored.

We assume first that the population estimation has no priority $(G = 0)$ and the unit survey cost are fixed. In this case minimization of (2.15) with respect of n_d has a unique solution

$$n_{d,\text{opt}} = \frac{n\sqrt{\sigma_d^2 N_d^q}}{\sum_{d=1}^{D}\sqrt{\sigma_d^2 N_d^q}} + \frac{1-\rho}{\rho}\left(\frac{\sqrt{\sigma_d^2 N_d^q}}{D^{-1}\sum_{d=1}^{D}\sqrt{\sigma_d^2 N_d^q}} - 1\right). \tag{2.16}$$

The formula (2.16) contains two unknown parameters, the intra-class correlation ρ and the area-specific variance σ_d^2. We replace ρ with an adjusted homogeneity coefficient of the auxiliary variable x. This coefficient is an approximation of the ICC (Intra-Class Correlation) (Section 2.2). Parameter σ_d^2 is replaced with the variance of x in area d. The reason for both replacements is that y is correlated with x. If also the population estimation has a priority $(G > 0)$ then (2.16) does not apply and F must be minimized numerically by using, for example, the NLP method, as we have done (Excel Solver, NLP option).

Table 2.1
Summary of model-based and model-assisted allocations

Method	Computing sample size n_d for area d	Optimality level
Model-based $g1$	$n_d^{g1} = \dfrac{N_d n - (N - N_d D - n)(1/R_{ax}^2 - 1)}{N + D(1/R_{ax}^2 - 1)}$, where R_{ax}^2 is the adjusted homogeneity measure of auxiliary variable x.	Area
Model-assisted MCG0	$n_{d,\text{opt}} = \dfrac{n\sqrt{\sigma_d^2 N_d^q}}{\sum_{d=1}^{D}\sqrt{\sigma_d^2 N_d^q}} + \dfrac{1-\rho}{\rho}\left(\dfrac{\sqrt{\sigma_d^2 N_d^q}}{D^{-1}\sum_{d=1}^{D}\sqrt{\sigma_d^2 N_d^q}} - 1\right)$	Jointly area and population
MCG50	Minimization of $F = \sum_{d=1}^{D} N_d^q \sigma_d^2 \rho(1-\rho)\left[1+(n_d-1)\rho\right]^{-1} + GN_+^{(q)}\sum_{d=1}^{D}\sigma_d^2 P_d^2 n_d^{-1}(1-\rho)$ with respect of n_d. Parameter ρ is replaced with R_{ax}^2 and σ_d^2 with $S_d^2(x)$.	

3 Some model-free area allocations

The aim of this section is to list the five previously presented allocation methods in order to use them later as references. Depending on which kind of auxiliary information each one uses, they are divided into two groups: number-based and parameter-based allocations.

3.1 Number-based allocations

Two basic allocation solutions commonly used go under the names equal allocation and proportional allocation. Neither of these allocations contains any specific criterion on the area or population level. Their implementation requires only information on the number of strata D and the numbers of units N_d in each stratum.

In the equal area allocation the sample size n_d is simply a quotient

$$n_d^{\text{Equ}} = n/D. \tag{3.1}$$

It is recommended to choose the total sample size n so that the quotient is a whole number. This allocation method does not take differences between the areas into account in any way, which results in inaccurate area estimates. A natural lower limit of the sample size is min $n = 2D$.

Proportional allocation is a frequently used basic method. Area sample sizes are solved from

$$n_d^{\text{Pro}} = n\left(N_d / N\right). \tag{3.2}$$

If the sizes of the areas vary strongly, it can lead to situations where the allocated sample size $n_d^{\text{Pro}} < 2$ for one or more areas. This is an obstacle in calculating direct design-based estimates of standard errors. One solution is to apply the combined allocation proposed by Costa, Satorra and Ventura (2004). The idea is a weighted solution between the equal and proportional allocation depending on the situation. The combined area sample size is

$$n_d^{\text{Com}} = kn_d^{\text{Pro}} + (1-k)\,n_d^{\text{Equ}} \tag{3.3}$$

for a specified constant $k\,(0 \le k \le 1)$. A minor problem is present if for some areas $n/D > N_d$. A modified solution exists for this case.

3.2 Parameter-based allocations

These allocations use area-level information of the study variable y and in some cases of the auxiliary variable x correlated with y. The values of x are available for all population units. In practice the unknown y is replaced with a proper proxy variable y^* such as a study variable obtained from an earlier research of the same subject, or the values of y^* are generated with a suitable model developed of a small pre-sample. Also x can be substituted for y. Allocation criteria can be set on population level, only on area level or on combined population and area level.

The Neyman allocation aims at reaching an optimal accuracy concerning population parameters $\text{SD}(y)_d$ (Tschuprow 1923). The standard deviation of the study variable y or some proxy variable and the number of units in each area must be known. Allocation favors large areas with strong variation.

The Bankier or power allocation (1988) is based on a criterion set on the area level. Area CV values of y are weighted by area total transformations X_d^q which contain a tuning constant q. In practice y^* or x must be used in place of y. Allocation favors mainly large areas with high CV.

Choudhry, Rao and Hidiroglou (2012) present the NLP allocation method for direct estimation. This method uses non-linear programming to find a solution. Criteria for the allocation are defined by setting

upper limits for CV values of the study variable y in each area and in the population. In practice y^* or x replaces y. The program searches the minimum sample size $n = \sum_d n_d$ satisfying these conditions. The SAS (Statistical Analysis System) procedure NLP with Newton-Raphson option was used to find the solution. The allocation favors areas with high CV regardless of the area size N_d.

A summary of the model-free allocations and the formulas for calculating area sample sizes are presented in Table 3.1.

Table 3.1
Summary of number-based and parameter-based allocations

Allocation	Computing area sample size n_d	Optimality level
Equal	$n_d^{\text{Equ}} = n/D$	Area
Proportional	$n_d^{\text{Pro}} = n(N_d/N)$	Population
Neyman	$n_d^{\text{Ney}} = n\left(N_d S_d / \sum_{d=1}^{D} N_d S_d\right)$, where S_d is the standard deviation of y (in practise y^* or x) in area d.	Population
Bankier	$n_d^{\text{Ban}} = n\left(X_d^q \text{CV}(y)_d / \sum_{d=1}^{D} X_d^q \text{CV}_d(y)\right)$, where X_d is the area total of x, $\text{CV}_d(y) = S_d/\overline{Y}_d$ and q is a tuning constant. In practise y^* or x replace y.	Area
NLP	$n_{st}^{\text{NLP}} = \min\left(\sum_{d=1}^{D} n_d\right)$ satisfying tolerances $\text{CV}(\overline{y}_d) \le \text{CV}_{0d}$ and $\text{CV}(\overline{y}_{st}) \le \text{CV}_0$. In practise y^* or x replace y.	Jointly population and area

Some other parameter-based allocation methods are mentioned briefly. For example Longford (2006) introduced inferential priorities P_d for the strata d and G for the population and used those constraints for allocation. Another solution is presented by Falorsi and Righi (2008). This solution does not contain a direct imposition of quotas, but tries to solve the comprehensive collection of data by using a multi-stage sampling design, so that the area estimation can be implemented effectively.

4 Comparison of performances of allocations

In this section we study the performances of the allocation methods introduced in Sections 2 and 3. The estimated parameters are area and population totals of the study variable y. The overall sample size $n = 112$. Section 4.1 includes the description of the research data. Simulation experiments and comparisons of allocations are presented in Section 4.3.

4.1 Empirical data

Our research data is obtained from a national Finnish register of block apartments for sale. This register is maintained by a private company, Alma Mediapartners Ltd, whose customers are real estate agencies. They save all the necessary information of the apartments into this register as soon as they receive an assignment from the owners. The population we have used consists of 9,815 block apartments (these serve as sampling units) for sale selected from the register. They represent 14 Finnish districts, mainly towns, in spring 2011. The sizes of the smallest and largest area were 112 and 1,333, respectively. The study variable (y) measures the apartment price (1,000 €) and the auxiliary variable (x) measures the size (m^2). Area sizes (N_d), population summary statistics (totals, means, standard deviations and CVs) for y and x, as well as correlations between x and y, are given in Table 4.1. The characteristics of the areas have a wide range. The most diverging area is Helsinki.

Table 4.1
Population summary statistics

Area		Study variable y				Auxiliary variable x				Correlation
Label	N_d	Y_d	$\bar{Y}_d$	$S_d(y)$	$\mathrm{CV}_d(y)$	X_d	$\bar{X}_d$	$S_d(x)$	$\mathrm{CV}_d(x)$	r_{yx}
Porvoo town	112	25,409	226.86	207.82	0.916	8,940	79.82	50.67	0.635	0.877
Pirkkala district	148	30,323	204.88	87.82	0.429	11,149	75.33	23.78	0.316	0.823
South Savo county	493	64,863	131.57	72.90	0.554	32,644	66.22	20.25	0.306	0.437
Jyväskylä town	494	89,941	182.07	69.65	0.383	40,000	80.97	17.62	0.218	0.509
Lappi county	555	62,143	111.97	50.15	0.448	30,805	55.50	16.22	0.292	0.207
South-East Finland	585	98,504	168.38	106.78	0.634	47,750	81.62	21.68	0.266	0.601
Helsinki (capital)	621	437,902	705.16	562.38	0.798	76,931	123.88	57.98	0.468	0.753
West coast district	655	108,339	165.40	75.85	0.459	50,903	77.71	36.39	0.468	0.439
Trackside district	818	148,845	181.96	65.08	0.358	59,220	72.40	23.84	0.321	0.517
Kuopio district	871	126,867	145.66	75.79	0.520	64,103	73.60	23.27	0.324	0.580
Turku district	958	166,613	173.92	131.62	0.757	79,970	83.48	25.71	0.308	0.635
Oulu district	1,072	133,591	124.62	50.19	0.403	59,210	55.23	16.92	0.306	0.392
Metropol area	1,100	263,293	239.36	117.84	0.492	80,034	72.76	26.37	0.362	0.754
Lahti-Tampere distr.	1,333	262,400	196.85	110.76	0.563	105,804	79.37	25.54	0.322	0.602
Population	9,815	2,019,031	205.71	215.52	1.048	747,462	76.16	31.76	0.417	0.674

The adjusted measure of homogeneity of the auxiliary variable x is $R^2_{ax} = 0.231$ indicating quite strong variability between the areas.

4.2 Allocations

In general, the overall sample size depends on the available time and financial resources in the research project. This aspect has not been taken into account now, because it is a question of an experimental study.

The value of the sampling ratio was determined as $f\,(\%) = 100 \times (112/9{,}815) = 1.14\%$. Method-specific allocations were produced according to the formulas presented in Table 2.1 and Table 3.1. Some details have been taken into account. In the Bankier allocation the value of a tuning constant q is 0.5. In the NLP allocation the selected CV limits 0.1258 (12.58%) for areas and the CV limit 0.0375 (3.75%) for the population lead to the overall sample size 112. We use the Excel Solver procedure with non-linear option for solving the NLP allocation problem. We use a modified proportional allocation to obtain an area sample size which is at least two. First we allocated one unit for every area and then allocated the rest 98 units by using proportionality. We have substituted x for y in every parameter-based allocation. In the model-assisted allocations the value of q was set to 1, and the quantity G was set to zero and 50. The final sample sizes in each allocation are presented in Table 4.2. The variation of sample sizes on area level is very strong between the allocations.

Table 4.2
Area sample sizes by allocation

Area		Model-based	Composite estim. Model-assisted		Number-based allocations		Parameter-based allocations		
Label	N_d	$g1^*$	**MCG0***	**MCG50***	**EQU**	**PRO**	**Ney_X**	**Ban_X**	**NLP_X**
Porvoo town	112	0	6	3	8	2	2	6	20
Pirkkala district	148	0	2	2	8	2	2	4	6
South Savo county	493	5	4	4	8	6	4	6	6
Jyväskylä town	494	5	3	4	8	6	4	5	3
Lappi county	555	6	3	4	8	6	4	5	5
South-East Finland	585	6	6	5	8	7	6	6	4
Helsinki (capital)	621	7	21	16	8	7	16	14	14
West coast district	655	7	12	11	8	8	10	11	14
Trackside district	818	10	8	8	8	9	9	8	7
Kuopio district	871	11	8	9	8	10	9	8	6
Turku district	958	12	10	11	8	11	11	9	6
Oulu district	1,072	13	6	8	8	12	8	8	6
Metropol area	1,100	13	11	12	8	12	13	11	8
Lahti-Tampere district	1,333	17	12	15	8	14	14	11	7
Total	9,815	112	112	112	112	112	112	112	112

* based on the adjusted coefficient of homogeneity (value 0.231) computed of x.

4.3 Comparison of performances of allocations

In this section we present the results based on design-based simulation experiments. For each allocation, 1,500 independent stratified SRSWOR samples were simulated with the SAS program and necessary calculations from the simulated samples were implemented with SPSS (Statistical Package for the Social Sciences) program. We have applied model-based EBLUP estimation on the samples for each allocation. For comparison of the allocations, we have computed two quality measures: $\mathrm{RRMSE}_d\,\%$ and $\mathrm{ARB}_d\,\%$ for each allocation.

Assume that r simulated samples are drawn in each allocation, and let $\hat{Y}_{di,\mathrm{EBLUP}}$ be the EBLUP estimate of the area total Y_d in the i^{th} sample $(i = 1,\ldots,r)$. Then $\mathrm{RRMSE}_d\,\%$ and $\mathrm{ARB}_d\,\%$ are defined as

$$\text{RRMSE}_d\,\% = 100 \times \sqrt{1/r \sum_{i=1}^{r} \left(\hat{Y}_{di,\text{EBLUP}} - Y_d\right)^2} \Big/ Y_d\,,$$
$$\text{ARB}_d\,\% = 100 \times \left| 1/D \sum_{i=1}^{r} \left(\hat{Y}_{di,\text{EBLUP}} \Big/ Y_d - 1\right) \right|,$$

and their means over areas are computed as follows:

$$\text{MRRMSE}\% = 1/D \sum_{d=1}^{D} \text{RRMSE}_d\,\% \quad \text{and} \quad \text{MARB}\% = 1/D \sum_{d=1}^{D} \text{ARB}_d\,\%.$$

The estimate for the population total in the i^{th} simulated sample $(i = 1, \ldots, r)$ is the sum of the estimates of the area totals: $\hat{Y}_{i,\,\text{EBLUP}} = \sum_{d=1}^{D} \hat{Y}_{di,\,\text{EBLUP}}$. RRMSE% for the population total is computed as

$$\text{RRMSE}_{\text{pop}}\,\% = 100 \times \sqrt{1/r \sum_{i=1}^{r} \left(\hat{Y}_{i,\text{EBLUP}} - Y\right)^2} \Big/ Y\,,$$

where Y is the true value of the population total, for which ARB% is computed as

$$\text{ARB}_{\text{pop}}\,\% = 100 \times \left| 1/r \sum_{i=1}^{r} \left(\hat{Y}_{i,\text{EBLUP}} \Big/ Y - 1\right) \right|.$$

Tables 4.3 and 4.4 contain RRMSE% and ARB% values for areas, their means over areas and population RRMSE%s and ARB%s in each allocation. The evaluation of the results was based on two arguments. One was the mean value of the quality measure on the area level and the other was the value of the quality measure on the population level.

Table 4.3
Area and population RRMSE%s by allocation

Area	N_d	$g1$	MCG0	MCG50	EQU	PRO	Ney_X	Ban_X	NLP_X
Porvoo town	112	8.08	14.63	15.93	13.41	19.79	16.49	14.78	10.10
Pirkkala district	148	6.60	9.72	10.77	8.35	12.04	10.60	9.76	8.97
South Savo county	493	22.29	22.77	23.20	18.63	20.70	23.20	20.16	20.88
Jyväskylä town	494	15.36	24.55	20.70	13.61	14.43	20.83	18.33	21.98
Lappi county	555	21.72	28.19	26.19	19.91	21.34	25.45	23.97	22.59
South-East Finland	585	20.76	27.25	25.93	19.68	19.64	24.37	24.31	27.81
Helsinki (capital)	621	22.72	12.68	14.97	21.92	23.15	14.35	16.02	16.43
West coast district	655	21.15	22.43	21.57	20.35	19.92	21.75	20.67	18.91
Trackside district	818	11.93	12.86	13.63	12.31	11.38	13.73	12.76	13.47
Kuopio district	871	16.22	23.22	20.70	19.21	16.37	20.84	20.82	23.49
Turku district	958	17.56	24.75	21.66	20.94	17.74	21.57	22.70	26.44
Oulu district	1,072	14.39	25.40	21.14	16.96	14.34	21.22	19.00	19.81
Metropol area	1,100	9.59	11.31	10.86	12.14	9.78	10.16	10.78	11.55
Lahti-Tampere distr.	1,333	10.54	13.43	11.66	13.35	10.64	12.76	12.87	14.98
Mean over areas (%)		15.65	19.51	18.59	16.48	16.52	18.38	17.64	18.39
Population value (%)		6.15	6.53	5.88	6.13	5.97	6.07	5.89	6.62

The lowest RRMSE% mean over the areas (15.65%) was obtained in the $g1-$allocation developed in this study. Helsinki was an exception on area level because its RRMSE% value was clearly higher compared

with model-assisted and parameter-based allocations. Also equal and proportional allocations performed well on area level, with means 16.48% and 16.52%. The highest means were obtained in the model-assisted MC-allocations. On the population level, the lowest value for the quality measure was obtained in the model-assisted MCG50-allocation (5.88%) and the second lowest value in the Bankier allocation (5.89%), but in general, differences between the allocations on this level were small.

Table 4.4
Area and population ARB%s by allocation

Area	N_d	$g1$	MCG0	MCG50	EQU	PRO	Ney_X	Ban_X	NLP_X
Porvoo town	112	2.28	2.20	0.97	0.04	1.26	1.28	0.98	0.79
Pirkkala district	148	0.17	2.10	1.08	0.19	0.79	0.85	0.86	1.15
South Savo county	493	8.08	11.81	10.87	6.76	7.29	11.47	9.09	9.81
Jyväskylä town	494	6.09	19.78	15.36	6.10	5.82	14.33	12.16	16.31
Lappi county	555	2.08	5.27	3.14	1.45	2.70	2.44	1.22	1.44
South-East Finland	585	9.05	20.62	18.28	9.53	8.11	15.69	15.96	20.41
Helsinki (capital)	621	9.71	6.38	7.93	10.95	11.59	7.43	8.80	9.45
West coast district	655	7.83	12.34	11.60	9.07	8.16	12.69	10.52	10.87
Trackside district	818	1.21	3.11	1.78	1.76	0.96	2.61	2.10	2.94
Kuopio district	871	6.00	14.90	10.68	9.37	6.53	11.33	11.77	15.56
Turku district	958	5.26	16.46	12.59	8.48	5.78	11.54	13.27	16.91
Oulu district	1,072	0.81	10.17	6.08	1.88	1.84	6.47	4.71	4.00
Metropol area	1,100	3.06	5.84	5.11	5.29	3.37	4.39	5.12	5.76
Lahti-Tampere distr.	1,333	1.86	6.14	3.97	3.62	1.79	4.65	4.37	6.10
Mean over areas (%)		4.53	9.79	7.82	5.32	4.71	7.66	7.21	9.15
Population value (%)		0.01	3.33	2.05	0.18	0.50	2.26	1.83	3.01

The $g1-$allocation was the only allocation with absolute relative bias less than 10% on each area, and it had a practically zero bias on the population level. Also the equal and proportional allocations had low biases on both levels, but the model-assisted and parameter-based allocations had a clearly poorer performance. An interesting detail in the $g1-$allocation is that the accuracy of area estimates is fairly good and the relative bias is low also for the case of two areas with zero sample size. A common characteristic for these areas is that the means of variables y and x are close to corresponding population means. In any case, it is essential that the model-based estimation can produce reliable estimates for areas, which are not represented in the random sample.

5 Concluding remarks

This research was focused on seven different allocation solutions which were categorized into three groups according to the auxiliary data needed in their implementation. The least amount of auxiliary information is needed in equal and proportional allocation which are based on the number of areas and the number of statistical units in each area. The Neyman, Bankier and NLP allocations are based on pre-set optimization criteria, and application of these methods presumes area-specific parameter information such

as the standard deviation or CV of the study variable, and in the Bankier allocation the area totals of at least one auxiliary variable must be known. Because the study variable is unknown, it must be replaced with a suitable proxy or auxiliary variable to enable the use of these three methods. A common feature of the number-based and parameter-based allocations is that they are not based on any model, whereas the other three allocations utilize the underlying model, in addition to number-based information.

On the basis of the empirical results, the performance of the model-based $g1-$allocation can be regarded as the best compared with the other allocations tested in this research. Also equal and proportional allocations reached good results, but the model-assisted allocations and the parameter-based allocations had clearly weaker performances. The last three allocations are developed originally for direct design-based estimation, and their results can be understood from that point of view. Compared with $g1-$allocation, the MC-allocations are based on a different model and this fact seems to affect their results.

One of the characteristics of the $g1-$allocation is that when the sampling design is constructed, also the model and estimation method are fixed, meaning that they are regarded as given preliminary information. This allocation, which is based on a unit-level linear mixed model and EBLUP estimation method, needs only the homogeneity coefficient between areas which is computed by using the values of the auxiliary variable. In this respect, the $g1-$allocation differs from the other allocations used in the comparison. Also the starting point for choosing the final estimation method is different, because this allocation is focused on model-based estimation, not on direct design-based estimation using sampling weights. The choice of the model-based estimation is justified also for the reason that it is commonly used in small area estimation. On the other hand, the $g1-$allocation enables the use of small sample sizes, because information can be borrowed between areas when the model is applied. This can be significant in quick surveys or studies carried out by market research organizations, when a single measurement is expensive. However, it is important to examine the characteristics of the areas and especially the small areas, before the final sample sizes are determined.

As a recommendation, it would be justified to start a wider research to find out what advantages and disadvantages are encountered if the applicable computing technique for producing area statistics is decided as early as in the design of the research plan.

Acknowledgements

The authors thank the Editor, Associate Editor and two referees as well as Professor Risto Lehtonen for constructive comments and suggestions.

References

Bankier, M.D. (1988). Power allocations: Determining sample sizes for subnational areas. *The American Statistician,* 42, 174-177.

Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2012). On sample allocation for efficient domain estimation. *Survey Methodology,* 38, 1, 23-29. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2012001/article/11682-eng.pdf.

Costa, A., Satorra, A. and Ventura, E. (2004). Improving both domain and total area estimation by composition. *SORT,* 28(1), 69-86.

Falorsi, P.D., and Righi, P. (2008). A balanced sampling approach for multi-way stratification for small area estimation. *Survey Methodology,* 34, 2, 223-234. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2008002/article/10763-eng.pdf.

Keto, M., and Pahkinen, E. (2009). On sample allocation for effective EBLUP estimation of small area totals – "Experimental Allocation". In *Survey Sampling Methods in Economic and Social Research,* (Eds., J. Wywial and W. Gamrot), 2010. Katowice: Katowice University of Economics.

Keto, M., and Pahkinen, E. (2014). On sample allocation for efficient small area estimation. *Book of Abstracts*. SAE 2014, Poland: Poznan University of Economics, page 50.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology,* 32, 1, 87-96. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2006001/article/9259-eng.pdf.

Molefe, W.B., and Clark, R.G. (2015). Model-assisted optimal allocation for planned domains using composite estimation. *Survey Methodology,* 41, 2, 377-387. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2015002/article/14230-eng.pdf.

Nissinen, K. (2009). *Small Area Estimation with Linear Mixed Models from Unit-Level Panel and Rotating Panel Data*. Ph.D. thesis, University of Jyväskylä, Department of Mathematics and Statistics, Report 117, https://jyx.jyu.fi/dspace/handle/123456789/21312.

Pfefferman, D. (2013). New important developments in small area estimation. *Statistical Science,* 28, 40-68.

Rao, J.N.K. (2003). *Small Area Estimation*. Hobogen, New Jersey: John Wiley & Sons, Inc.

Tschuprow, A.A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron,* Vol. 2, 3, 461-493; 4, 646-683.

III

ON OVERALL SAMPLING PLAN FOR SMALL AREA ESTIMATION

by

Mauno Keto and Erkki Pahkinen, 2017

Statistical Journal of the IAOS, vol. 33, pp. 727-740

On overall sampling plan for small area estimation

Mauno Keto[a,*] and Erkki Pahkinen[b]

Abstract

The time and budget restrictions in survey sampling can impose limits on the area sample sizes. This may reduce the possibility to obtain area-specific and population parameters estimates with adequate precision. Market research companies and institutes for producing official statistics face frequently this problem. Various models and methods for small area estimation (SAE) have been developed to solve this problem. The sample allocation must support the selected model and method to ensure efficient estimation and must be implemented in the design phase of the survey. The proposed allocation is developed by incorporating auxiliary information, a model, and an estimation method. The estimated parameters are area and population totals. The performance of this allocation is assessed through design-based simulation experiments using real, regularly collected register data. Five other allocations selected from the literature serve as references. Model-based estimation is applied to two allocations and design-based Horvitz-Thompson and model-assisted GREG estimation to four model-free allocations. Four allocations are based on past register data. The allocation with uniquely best performance among all alternatives was not found, but the simulation study supports the comprehensive survey plan where the sampling design is conditioned on the available auxiliary information, selected model, and method.

Key words: Low sample size, auxiliary information, model selection, sample allocation, EBLUP estimation.

[a] Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland. E-mail: mauno.j.keto@student.jyu.fi
[b] Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland. E-mail: erkki.j.pahkinen@jyu.fi
[*] Corresponding author: Mauno Keto

1. Introduction

Many sample-based surveys in a business or an administrative environment aim at obtaining parameters estimates for the variables of interest, not only on the population level, but also on the subpopulation or area level. A fundamental survey plan contains the phases which are implemented in a specified order. The sampling design phase contains a plan for the collection of the sample data from the target population. The estimation phase uses the sample data and auxiliary information available often on unit level. The sampling design is a critical phase in the sense that one of its sub-steps, the sample allocation, may have a strong influence on the estimation results. For this reason, the sample allocation is not an independent part of the survey. It must be conditioned on the used model, estimation method and auxiliary information as well as the priorities set on the area and population level estimation. The variation of the variables of interest between and within the areas must also be considered.

The time and budget restrictions in survey sampling can impose limits on the area sample sizes. This may reduce the possibility to obtain area-specific and population parameters estimates with adequate precision. Market research companies and institutes for producing official statistics face frequently this problem. Various models and methods for small area estimation (SAE) have been developed to solve this problem. As Rao and Molina [1] present comprehensively, the assortment of different alternatives is wide. They point out the use of empirical best linear unbiased estimation methods (EBLUP). This is the main reason for applying EBLUP to the selected model. Burgard et al. [2] have studied the performances of different small area point and accuracy estimates for business data. The above sources show that the optimal solutions concerning sampling design and the choice of the model, estimator and estimation method are under intensive study.

We propose a model-based CAL-*g1* allocation for stratified sampling where the areas of interest coincide with the strata and where the overall sample size is restricted. The estimated parameters are area and population totals of the study variable y. This allocation aims at obtaining area and population estimates with sufficient accuracy. It is based on analytical optimization and the calibration of area sizes, and uses the selected model, estimation method, and the auxiliary population information, from which the variation between and within the areas can be resolved. The underlying model and the derivation of this allocation are introduced in Sections 2.1 and 2.2.

The performance of the proposed allocation method in a real situation is evaluated by using design-based simulation experiments. An official Finnish register of block apartments for sale in 18 Finnish provinces serves as the sampling population. Five other allocations selected from then literature serve as references. One of them, the MC-*q025* allocation introduced by Molefe and Clark [3], is based on a two-level area model and composite estimator, and uses the same population information as CAL-*g1* allocation. It is introduced in Section 2.3. Four other allocations are model-free and have originally been developed for design-based estimation. They are introduced in Section 3. Two of them need only number-based area information for computing the area sample sizes. The other two methods use, in addition to number-based information, area level parameter information of the study variable.

The choice of the reference allocations is based on the diversity in the optimization criteria. Among the model-free allocations, the optimality level is not defined, it is set on the area level, population level or on both levels simultaneously. The priorities for the area and population level estimation can be adjusted in MC-*q025* allocation.

Because the parameter information as well as the between-area and within-area variation concerning the study variable y are not available, it is replaced with a proxy study variable y^* obtained from the past apartment register data. Variable y^* is used when computing the area sample sizes for each allocation except for equal and proportional allocations. Section 4.1

contains the characteristics of the sampling population and the proxy population used in the allocation phase. The populations include also two auxiliary variables. The allocation-specific area sample sizes and the calculation details are presented in Section 4.2.

Different estimation methods are used for producing the estimates for the area and population totals. Model-based EBLUP estimation is applied to the simulated samples drawn according to model-based allocations. Design-based Horvitz-Thompson and model-assisted GREG estimation are applied to the samples drawn according to model-free allocations. The assisting model is the one used in EBLUP estimation. The idea in applying two methods to the same samples is to resolve how the accuracy of the estimates develops when the assisting model is included in estimation. The use of a low overall sample size (*n*) makes it easier to see how design-based and model-based estimations perform in this survey framework.

For measuring and comparing the performances of the different allocation and estimation method combinations, two quality measures are computed from the simulated samples. The relative root mean square error (RRMSE%) is a numerical approximation for the accuracy of the area-specific and population estimates, and absolute relative bias (ARB%) is a numerical approximation for the bias of the estimates. The biases of the model-based estimates can be high for some areas, indicating the model misspecification, but the design-based estimates are generally almost unbiased. The primary quality measure is RRMSE%. Section 4.3 contains the empirical simulation results. They support the strategy where the allocation is conditioned on auxiliary information, the model and estimation method, and they should be determined as early as in the design phase of a survey.

2. Allocations using the model

2.1. The model and estimation method for estimating area totals

The model for estimating the area totals of the study variable *y* is a unit-level linear mixed model, also called a nested error linear regression model

$$y_{dk} = \boldsymbol{x}'_{dk}\boldsymbol{\beta} + v_d + e_{dk};\ k = 1,\ldots,N_d;\ d = 1,\ldots,D, \tag{1}$$

where N_d is the size of area *d* and *D* is the number of the areas. The area effects v_d are assumed to be iid random variables with mean zero and variance σ_v^2, and e_{dk}´s are iid random variables with mean zero and variance σ_e^2 and they are independent of v_d´s. Furthermore, $E(y_{dk}) = \boldsymbol{x}'_{dk}\boldsymbol{\beta}$ and $V(y_{dk}) = \sigma_v^2 + \sigma_e^2$ (total variance). Matrix $\boldsymbol{V}$ is the variance-covariance matrix of the study variable *y* with a block-diagonal covariance structure. This model can be used when unit-level values are available for the auxiliary variables $\boldsymbol{x}$.

A common intra-area correlation ρ (IAC), see Meza and Lahiri [4], measures the relative variation of *y* between the areas and is computed of the variance components as

$$\rho = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2). \tag{2}$$

The variance components, regression coefficients and area effects must be estimated from the sample data before estimating the area parameters. The BLUE estimator (*Best Linear Unbiased Estimator*) of $\boldsymbol{\beta}$, noted $\tilde{\boldsymbol{\beta}}$, is obtained in accordance with the general linear model theory. It is replaced with its EBLUP (*Empirical Best Linear Unbiased Predictor*) sample estimate $\hat{\boldsymbol{\beta}}$.

The EBLUP estimate (predicted value) for the area total Y_d of the study variable is the sum of the observed *y*-values and predicted *y*-values for units outside the sample:

$$\hat{Y}_{d,EBLUP} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \hat{y}_{dk} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \boldsymbol{x}'_{dk}\hat{\boldsymbol{\beta}} + (N_d - n_d)\hat{v}_d \,. \tag{3}$$

The design MSE (mean squared error) for the estimator Eq. (3) is the sum of its variance and squared bias and is defined as

$$\text{MSE}(\hat{Y}_{d,EBLUP}) = \text{E}(\hat{Y}_{d,EBLUP} - Y_d)^2 = \text{V}(\hat{Y}_{d,EBLUP}) + (\text{E}(\hat{Y}_{d,EBLUP}) - Y_d)^2 \tag{4}$$

The second-order Prasad-Rao approximation (see Rao and Molina [1]; pp 180-181) to MSE Eq. (4) for finite populations is

$$\text{mse}\ (\hat{Y}_{d,EBLUP}) = g_{1d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) + g_{2d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) + 2g_{3d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) + g_{4d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2), \tag{5}$$

where the four terms g_{1d}, g_{2d}, g_{3d}, and g_{4d} are defined as

$$\begin{aligned}
g_{1d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) &= (N_d - n_d)^2(1-\hat{\gamma}_d)\hat{\sigma}_v^2, \\
g_{2d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) &= (N_d - n_d)^2(\bar{\boldsymbol{x}}_{\boldsymbol{d}}^{*} - \hat{\gamma}_d\bar{\boldsymbol{x}}_{\boldsymbol{d}})'(\boldsymbol{X'V^{-1}X})^{-1}(\bar{\boldsymbol{x}}_{\boldsymbol{d}}^{*} - \hat{\gamma}_d\bar{\boldsymbol{x}}_{\boldsymbol{d}}), \\
g_{3d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) &= (N_d - n_d)^2(n_d^{*})^{-2}(\hat{\sigma}_v^2 + \hat{\sigma}_e^2(n_d^{*})^{-1})^{-3}[\hat{\sigma}_e^4 V(\hat{\sigma}_v^2) \\
&\quad + \hat{\sigma}_v^4 V(\hat{\sigma}_e^2) - 2\hat{\sigma}_e^2\hat{\sigma}_v^2 Cov(\hat{\sigma}_e^2,\hat{\sigma}_v^2)], \\
g_{4d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) &= (N_d - n_d)\hat{\sigma}_e^2 \,.
\end{aligned} \tag{6}$$

The area sample sizes n_d depend on the sample and are not fixed. The main term g_{1d} contains the area-specific ratio $\hat{\gamma}_d = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_d)$. Nissinen [7, p. 53] points out that this component contributes generally over 90 % of the estimated MSE. We have reached similar proportions for g_{1d} in our simulation experiments for every allocation. The high proportion of g_{1d} suggests that the variation of the area estimates is strongly related to the variation between the areas.

2.2. Model-based calibrated CAL-*g1* allocation

One criterion for obtaining the area sample sizes in the model-based framework is to minimize the mean of MSE$_d$´s over areas subject to $n = \sum_{d=1}^{D} n_d$, but an analytical solution is difficult owing to the complexity of the MSE approximation Eq. (5). Keto and Pahkinen [8] have examined this allocation problem for the first time in an experimental study and have developed later an allocation (basic *g1* allocation) based only on the term g_{1d}. The reasoning behind this solution is the high proportion of g_{1d} in the MSE approximation. We describe first the basic *g*1 allocation and then extend it to the proposed CAL-*g1* allocation.

The basic *g1* allocation is based on the minimization of the sum of g_{1d}´s over the areas:

$$\sum_{d=1}^{D} g_{1d}(\sigma_v^2,\sigma_e^2) = \sum_{d=1}^{D}(N_d - n_d)^2(n_d/\sigma_e^2 + 1/\sigma_v^2)^{-1} \tag{7}$$

subject to $n = \sum_{d=1}^{D} n_d$.

The solution is obtained using Lagrange´s multiplier method. The function F of sample sizes $\boldsymbol{n}' = (n_1, n_2, ..., n_D)$ and λ is

$$F(\boldsymbol{n}, \lambda) = \sum_{d=1}^{D} g_{1d}(\sigma_v^2, \sigma_e^2) = \sum_{d=1}^{D} (N_d - n_d)^2 (n_d/\sigma_e^2 + 1/\sigma_v^2)^{-1} + \lambda(\sum_{d=1}^{D} n_d - n). \quad (8)$$

An analytical solution for the area sample size n_d^{g1} is

$$n_d^{g1} = \frac{N_d n - (N - N_d D - n)(1/\rho - 1)}{N + D(1/\rho - 1)}, \quad (9)$$

where the intra-area correlation ρ in Eq. (2) measuring the relative between-area variation is unknown. It is replaced with an adjusted homogeneity measure of variation, which is the approximation of an intra-class correlation (*ICC*) known of cluster sampling. One area serves as one cluster here. Because y is unknown, it is replaced with the proxy variable y^*. They are related to one another, because they measure the same numerical quantity on consecutive points of time.

The homogeneity coefficient is obtained using one-way ANOVA applied to y^* between the areas, and then the adjusted homogeneity measure between the areas is computed as

$$R_{a,y^*}^2 = 1 - \mathrm{MSW} / S_{y^*}^2, \quad (10)$$

where MSW is the mean *SS* of areas and $S_{y^*}^2$ is the variance of y^*.

Replacing ρ in Eq. (9) with the known homogeneity measure Eq. (10), the final expression for computing the area sample sizes is obtained as

$$n_d^{g1} = \frac{N_d n - (N - N_d D - n)(1/R_{a,y^*}^2 - 1)}{N + D(1/R_{a,y^*}^2 - 1)}. \quad (11)$$

The expression in Eq. (11) is an increasing function of the area size N_d. In principal, the computed sample sizes are rounded to the nearest integer. Under certain circumstances, such as low homogeneity coefficient, small overall sample size n or area size N_d, Eq. (11) may yield negative area sample sizes, which are changed to zero. An extreme case is that all variation is between the areas (ρ = 1), and Eq. (11) turns to proportional allocation. In case of equal area sizes N_d, the solution is equal allocation.

The derived *g1* allocation is efficient on the population level, but it can lead to inaccurate estimates for the areas with very small size, because they have a low sample size. This allocation does not take the within-area variation into account. This variation is included in the modified *g*1 allocation (CAL-*g1*) using calibration. The steps for the calibration are:

a) The average $ASD(y^*) = \sum_d SD(y^*)_d / D$ of the area standard deviations of y^* is computed.
b) Each true area size N_d is replaced with the constant area size $\hat{N}_d = N / D$.
c) The calibrated area sizes are computed as $\tilde{N}_{g1,d} = (SD(y^*)_d / ASD(y^*))\,\hat{N}_d$.
d) Inserting the calibrated area sizes $\tilde{N}_{g1,d}$ into Eq. (11) in place of N_d, the sample sizes for the CAL-*g1* allocation are obtained as

$$n_d^{\mathrm{CAL}-g1} = \frac{\tilde{N}_{g1,d} n - (N - \tilde{N}_{g1,d} D - n)(1/R_{a,y^*}^2 - 1)}{N + D(1/R_{a,y^*}^2 - 1)}. \quad (12)$$

This calibration ignores the true area sizes. The higher the variation in area *d*, the larger is $n_d^{\mathrm{CAL}-g1}$, and vice versa. Following the idea of Longford [12], the calibrated weight $\tilde{N}_{g1,d}$ reflects the inferential priority (importance) for area *d*.

2.3. Model-assisted MC allocation

Molefe and Clark [3] have used the following composite estimator for estimating the mean of the study variable *y* for area *d*:

$$\tilde{\bar{y}}_d^C = (1-\phi_d)\bar{y}_{dr} + \phi_d\,\hat{\boldsymbol{\beta}}'\bar{\boldsymbol{X}}_d\,. \tag{13}$$

This estimator is a combination of two estimators: the synthetic estimator $\hat{\bar{Y}}_{d(syn)} = \hat{\boldsymbol{\beta}}'\,\bar{\boldsymbol{X}}_d$, where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient and $\bar{\boldsymbol{X}}_d$ is the area population means of auxiliary variables $\boldsymbol{x}$, and a direct estimator $\bar{y}_{dr} = \bar{y}_d + \hat{\boldsymbol{\beta}}'(\bar{\boldsymbol{x}}_d - \bar{\boldsymbol{X}}_d)$, where $\bar{y}_d$ and $\bar{\boldsymbol{x}}_d$ are the area *d* sample means of *y* and $\boldsymbol{x}$. The coefficients ϕ_d are set with the intent to minimize the mean squared error (MSE) of the estimator (13). The approximated design-based MSE of the estimator under certain conditions and assumptions is given as

$$MSE_p(\tilde{\bar{y}}_d^C;\bar{Y}_d) \approx (1-\phi_d)^2 v_{d(syn)} + \phi_d^2 B_d^2\,, \tag{14}$$

where $v_{d(syn)}$ is the sampling variance of the synthetic estimator $\hat{\bar{Y}}_{d(syn)}$ and $B_d = \boldsymbol{\beta}_U'\,\bar{\boldsymbol{X}}_d - \bar{Y}_d$ is the bias when $\hat{\bar{Y}}_{d(syn)}$ is used to estimate $\bar{Y}_d$, with $\boldsymbol{\beta}_U$ denoting the approximate design-based expectation of $\hat{\boldsymbol{\beta}}$.

A random sample (SRSWOR) of n_d units is selected from stratum *d* ($d = 1,\ldots, D$) containing N_d units. The relative size of area *d* is $P_d = N_d / N$.

A two-level linear model ξ conditional on the values of $\boldsymbol{x}$ is assumed, with uncorrelated stratum random effects u_d and unit residuals ε_i:

$$\left.\begin{array}{c} y_i = \boldsymbol{\beta}'\boldsymbol{x_i} + u_d + \varepsilon_i \\ E_\xi(u_d) = E_\xi(\varepsilon_i) = 0 \\ V_\xi(u_d) = \sigma_{ud}^2 \\ V_\xi(\varepsilon_i) = \sigma_{ed}^2 \end{array}\right\}, \tag{15}$$

where *i* refers to all units in stratum *d*. This model implies that $V_\xi(y_i) = \sigma_{ud}^2 + \sigma_{ed}^2$ for all population units and $\mathrm{cov}_\xi(y_i, y_j)$ equals $\rho_d\,\sigma_d^2$ for units $i \neq j$ in the same stratum and zero for units from different strata, where $\rho_d = \sigma_{ud}^2/(\sigma_{ud}^2 + \sigma_{ed}^2)$. For simplicity, it is assumed that $\rho_d = \rho$ are equal for all strata.

After some other simplifying assumptions and solving the optimal weight ϕ_d in Eq. (14), the final approximate optimum anticipated MSE is obtained of Eq. (13) as

$$\mathrm{AMSE}_d = E_\xi \mathrm{MSE}_p(\tilde{\bar{y}}_d^C\left[\phi_{d(opt)}\right];\bar{Y}_d \approx \sigma_d^2\rho(1-\rho)\left[1+(n_d-1)\rho\right]^{-1}. \tag{16}$$

The criterion F using anticipated MSE´s of the small area mean and overall mean estimators for model-assisted allocation has the final approximative form

$$F = \sum_{d=1}^{D} N_d^q \mathrm{AMSE}_d + GN_+^{(q)} E_\xi \, \mathrm{var}_p \left(\hat{\bar{Y}}_r \right)$$

$$\approx \sum_{d=1}^{D} N_d^q \, \sigma_d^2 \, \rho(1-\rho)\left[1+(n_d-1)\rho\right]^{-1} + GN_+^{(q)} \sum_{d=1}^{D} \sigma_d^2 P_d^2 n_d^{-1}(1-\rho). \quad (17)$$

Optimal sample sizes for the areas are obtained minimizing Eq. (17) subject to $\sum_d n_d = n$, following the idea of Longford [12]. The weight N_d^q reflects the inferential priority for area *d*, with *q* as an adjustable constant ($0 \le q \le 2$), and $N_+^{(q)} = \sum_{d=1}^{D} N_d^q$. The quantity *G* is a relative priority on the population level. If *G* is set to zero, the attention is focused only on the area level estimation, and the increment in *G* diminishes the importance of area level estimation.

If also the population estimation has a priority ($G > 0$), *F* must be minimized numerically by using, for example, the NLP method. If $G = 0$ and the unit survey cost are fixed, the minimization of Eq. (17) with respect of n_d has a unique solution

$$n_d^{MC} = \frac{n\sqrt{\sigma_d^2 \, N_d^q}}{\sum_{d=1}^{D} \sqrt{\sigma_d^2 \, N_d^q}} + \frac{1-\rho}{\rho}\left(\frac{\sqrt{\sigma_d^2 \, N_d^q}}{D^{-1}\sum_{d=1}^{D} \sqrt{\sigma_d^2 \, N_d^q}} - 1 \right). \quad (18)$$

Equations (17)–(18) contain two unknown parameters, the intra-class correlation ρ and the area-specific variance σ_d^2. Parameter ρ is replaced with an adjusted homogeneity coefficient of the proxy variable y^* (Section 2.2), and σ_d^2 is replaced with the variance of y^* in area *d*. The relationship between *y* and y^* justifies both replacements.

Table 1. Summary of model-based and model-assisted allocations.

3. Model-free reference area allocations

Four allocation methods developed originally for the design-based estimation are introduced shortly in this section. They are model-free in the sense that they can be used also in other model and estimation method frameworks. Depending on which kind of auxiliary information each one uses, they are divided into two groups: number-based and parameter-based allocations.

3.1. Number-based allocations

Two basic commonly used allocations go under the names equal allocation and proportional allocation, see Cochran [5]. They don´t contain any specific criterion on the area or population level. Their implementation requires only information on the number of strata *D* and the numbers of units N_d in each stratum.

In the equal allocation (EQU), the area sample size n_d is simply

$$n_d^{EQU} = n/D. \quad (19)$$

It is recommended to choose the total sample size *n* so that the quotient is an integer. This allocation method does not take the internal characteristics of the areas into account in any way. As Choudry et al. [11] state, it can be efficient on area level, but can lead to inaccurate estimates

for very large areas, and thus for the whole population. A natural lower limit of the sample size is min $n = 2D$.

Proportional allocation (PRO) is a frequently used basic method. The area sample size n_d is proportional to the area size N_d and is computed as

$$n_d^{PRO} = (N_d / N)n . \tag{20}$$

If a stronger variation can be anticipated in large areas compared with small areas, this allocation can be a reasonable choice, but on the other hand, strong differences between the area sizes can lead to situations where $n_d^{PRO} < 2$ for the smallest areas. This is an obstacle in calculating reliable direct design-based area estimates as well as their unbiased variances. The population estimates are generally accurate, because large areas have high sample sizes, but the small area estimates are probably less accurate. Costa et al. [6] have proposed a convex combination

$$n_d^{COS} = k\, n_d^{PRO} + (1-k)\ n_d^{EQU} = k(N_d / N)n + (1-k)n / D \tag{21}$$

between equal and proportional allocation for a specified constant k $(0 \le k \le 1)$ to avoid very small sample sizes, but it can be difficult to justify the optimal value for k.

3.2. Parameter-based allocations

Parameter-based allocations use area-level information of the study variable y. In practice the unknown y is replaced with a proxy variable y^* such as a study variable measuring the same characteristics and is obtained from the past data. If the past data is not available, an auxiliary variable x correlated with y can be used as a proxy variable. The allocation criteria can be set on population level, only on area level or on combined population and area level.

The Neyman allocation (NEY) aims at reaching an optimal accuracy on the population level and uses area parameters $S(y)_d$, see Tschuprow [9] and Cochran [5]. The standard deviation of the study variable y and the number of units in each area must be known. This allocation favors large areas with strong variation and can lead to area sample sizes $n_d < 2$ preventing the unbiased estimation of the variances. An alternative to avoid this problem by using the box-constraint optimal allocation has been proposed by Gabler et al. [10].

Choudry et al. [11] present the NLP (non-linear programming) allocation for direct estimation. Criteria for the allocation are defined by setting first upper limits for CV´s of the area sample means $\bar{y}_d$ and population sample mean $\bar{y}_{st}$. The CV´s are computed as

$$\mathrm{CV}(\bar{y}_d) = \sqrt{\mathrm{V}(\bar{y}_d)} / \bar{Y}_d \ \text{ and } \ \mathrm{CV}(\bar{y}_{st}) = \sqrt{\mathrm{V}(\bar{y}_{st})} / \bar{Y} \ . \tag{22}$$

The program searches the minimum sample size $n = \sum_d n_d$ subject to pre-set tolerances for the CV´s in Eq. (22). The constraints are defined so that the function to be minimized becomes separable and convex. The SAS procedure NLP with Newton-Raphson option was used to find the solution. The allocation favors areas with high CV regardless of the area size N_d.

A summary of the model-free allocations and the formulas for calculating area sample sizes are presented in Table 2.

Table 2. Summary of number-based and parameter-based allocations.

Some other parameter-based allocation methods are mentioned briefly. Longford [12] introduces the inferential priorities P_d for the strata d and G for the population and uses those constraints for deriving sample size allocation schemes for three types of estimators. Falorsi

and Righi [13] propose an overall sampling strategy that guarantees a pre-defined precision for the domain estimators when the overall sample size is bounded. The strategy aims at controlling the area sample sizes by using a multi-stage sampling design based on a balanced sampling selection technique and a GREG-type estimation.

3.3. Estimation methods for model-free allocations

The finite population denoted $U = \{1,2,...,k,...,N\}$ is composed of D non-overlapping domains or areas $U_1,...,U_d,...,U_D$, with N_d units in each, and $\sum_d N_d = N$. A probability sample s is drawn from U, and s_d is the sample drawn from area d. The inclusion probability of unit k is denoted π_k, and the sampling weight for unit k is $w_k = 1/\pi_k$.

Two design-based estimation methods are applied to model-free allocations. The Horvitz-Thompson estimator for the area total $Y_d = \sum_{U_d} y_k$ is

$$\hat{Y}_{d,H-T} = \sum_{k \in s_d} w_k y_k = \sum_{k \in s_d} y_k / \pi_k \,. \tag{23}$$

The model-assisted GREG (Generalized Regression) estimator for area total Y_d

$$\hat{Y}_{d,GREG} = \sum_{k \in U_d} \hat{y}_k + \sum_{k \in s_d} (y_k - \hat{y}_k)/\pi_k \ (\hat{y}_k \text{ is the predicted value}), \tag{24}$$

is based on a model, and here it is the linear mixed model Eq. (1). See Lehtonen et al. [14] for details. The first part of Eq. (24) is the predicted value for Y_d when the model is applied. The predicted value for every $k \in U$ can be computed, because the unit-level values of the auxiliary variables $\boldsymbol{x}$ are known according to the model. The second term protects against model misspecification (Lehtonen et al. [14]).

4. Empirical results

This section contains the descriptions of the research data populations, the allocations and the clarifying details in computing the sample sizes, as well as the performances of the allocations based on sample simulation experiments. The estimated parameters are area and population totals of the study variable y, and the overall sample size n is fixed.

4.1. Periodically collected business register

A national Finnish register of block apartments for sale is the source of the research data. This register is maintained by a private company, Alma Mediapartners Ltd, whose customers are real estate agencies. They save all the necessary information of the apartments into this register as soon as they receive an assignment from the owners. The population for sample simulations consists of 21,025 block apartments (serve as sampling units) for sale selected from the register. They cover 18 Finnish provinces, which serve as areas, in October 2015. The smallest area contains 160 units and the largest area contains 6,813 units. The study variable *(y)* measures the apartment price (1,000 €) and the auxiliary variables (x_1 and x_2) measure the size (m^2) and age (years) of the apartments.

All the allocations except EQU and PRO allocations are based on the proxy variable y^*, which is the price variable of the proxy data register in April 2015. This register contains 22,230 apartments for sale in 18 provinces, and the variables are the same as in the sampling population. The reasoning behind the use of the proxy data for the allocations is that the

structure of this phenomenon under study has remained practically unchanged from April to October in 2015. The adjusted measure of homogeneity of the y^* is $R^2_{a,y^*} = 0.1697$ indicating a moderate variability between the areas.

Table 5 in the Appendix contains area sizes (N_d), population summary statistics (totals, means, standard deviations and CV´s) for y and the proxy variable y^*. The corresponding population statistics except totals for x-variables, as well as correlations between y- and x-variables, are given in the Appendix Table 6. The characteristics of the areas have a wide range concerning the variables price and age. There is not a very significant variation in the sizes of apartments between the areas, as can be expected. The province of Uusimaa (around capital Helsinki) is a dominating area, because its size is clearly the largest (32.4 % of the population size) and the general price level is by far the highest among the provinces. The study variable y has a strong positive correlation with x_1(size) except for one small area and a negative correlation with x_2 (age) in all areas except for the largest area (Uusimaa). The area-specific correlations between auxiliary variables are low.

4.2. Allocations

In general, the overall sample size depends on the available time and financial resources in the research project. These limitations have no significance now, because the low overall sample size (n) is an essential feature in our experimental study. The value of the sampling ratio was determined as f % = 216/21,025 = 1.03 %. Method-specific allocations are based on the formulas presented in Table 1 and Table 2.

Some details are clarified. We have substituted y^* for y in two model-free and two model-based allocations using area parameters. The Excel Solver procedure with non-linear option is used for solving the area sample sizes for NLP allocation. The selected CV limits 0.1901 (19.01 %) for areas and the CV limit 0.0800 (8.00 %) for the population lead to the overall sample size 216. Two smallest areas have a computational sample size one in NEY allocation, but they were raised to two, on the cost of Uusimaa province, to allow unbiased variance estimation. The value for the adjusted homogeneity coefficient (Section 2.2) used for CAL-*g1* and MC-*q025* allocations is 0.1697. For the MC-*q025* allocation, the value of q was set to 0.25, and the quantity G was set to zero. The reason for the choice of these values is to avoid the strong concentration of the sample on one area (Uusimaa) and a very low or zero sample size for many areas.

The allocation-specific area sample sizes, which are presented in Table 3, vary strongly between the allocations. The area sizes in the proxy population and the calibrated area sizes used for CAL-*g1* allocation are also presented. Uusimaa area dominates in three allocations, and in NEY allocation it represents almost 60 % of the overall sample. Four areas have sample size two in NEY allocation. Low area sample sizes appear also in MC-*q025* and PRO allocations.

Table 3. Area sample sizes by allocation.

4.3. Simulation experiments

The results are based on design-based simulation experiments. For each allocation, r (here r = 1,500) independent stratified SRSWOR samples were simulated using SAS program, which was used also in the computation of estimates for regression coefficients, area effects and area totals in EBLUP estimation. Other calculations from the simulated samples were implemented with SPSS program. We have applied design-based Horvitz-Thompson (H-T notation in tables and figures) and model-assisted GREG estimation to the model-free allocations and model-based EBLUP estimation to CAL-*g1* and MC-*q025* allocations.

The performances of the allocations (accuracy and bias) are evaluated in terms of two quality measures computed from the simulated samples. The relative root mean square error RRMSE% is the numerical approximation of design MSE Eq. (4) or design variance, and absolute relative bias (ARB%) is the numerical approximation of the design bias. Bias values are computed also for model-free allocations, although design-based estimators are generally design-unbiased.

The number of simulated samples is r in each allocation, and $\hat{Y}_{di}$ is a design- or model-based estimate for the area total Y_d in the i^{th} sample ($i = 1, \ldots, r$). RRMSE% for area d is defined as

$$\text{RRMSE}_d\% = 100 \times (1/r \sum_{i=1}^{r} (\hat{Y}_{di} - Y_d)^2)^{1/2} / Y_d ,$$

and ARB% for area d is defined as

$$\text{ARB}_d\% = 100 \times \left| 1/r \sum_{i=1}^{r} (\hat{Y}_{di} / Y_d - 1) \right| ,$$

and their means over all D areas are computed as

$$\text{MRRMSE\%} = 1/D \sum_{d=1}^{D} \text{RRMSE}_d\% \text{ and MARB\%} = 1/D \sum_{d=1}^{D} \text{ARB}_d\% .$$

The estimate for the population total in the i^{th} simulated sample ($i = 1, \ldots, r$) is the sum of the estimates of the area totals: $\hat{Y}_i = \sum_{d=1}^{D} \hat{Y}_{di}$. RRMSE% for the population total is computed as

$$\text{RRMSE(pop)\%} = 100 \times (1/r \sum_{i=1}^{r} (\hat{Y}_i - Y)^2)^{1/2} / Y ,$$

where Y is the true value of the population total, and the corresponding ARB% is computed as

$$\text{ARB(pop)\%} = 100 \times \left| 1/r \sum_{i=1}^{r} (\hat{Y}_{i,EBLUP} / Y - 1) \right| .$$

The evaluation of the quality measures is based on the means over the areas, the population values, and the area-specific distributions.

The RRMSE_d % means over the areas (MRRMSE%) and population RRMSE%´s are presented in Figure 1. The allocations and estimation methods are ordered so that they highlight the change in accuracy of area and population estimates when the design-based and model-assisted GREG estimation have been applied to the model-free allocations. The population level RRMSE%´s and means over the areas (MRRMSE%) have decreased clearly in EQU and NLP allocations. The corresponding changes in PRO and NEY allocations are contradictory in the sense that population RRMSE%`s have decreased slightly, but the means over the areas have increased considerably. The typical properties of the EQU, PRO and NEY allocations can be recognized from the results. The EQU allocation performs well on the area level, but poorly on the population level (H-T: 13.26 % and GREG: 10.97 %). The PRO and NEY allocations are far from good performance on the area level.

On the population level, PRO/GREG combination reaches the lowest population RRMSE% (4.82 %), but all the other allocations except EQU and NLP have almost the same accuracy. If the allocation-specific aggregate RRMSE%´s are experimentally computed as the sums of the means over the areas and population values, the allocations CAL-*g1* and MC-*q025* have the lowest sums, but their mutual differences are small.

Figure 1. Means of area RRMSE_d%s and population RRMSE%s by allocation and estimation method.

Figure 2 contains the distributions of the area-specific RRMSE_d % values for each allocation, and the precise values are presented in the Appendix Table 7. The distributions illustrate the relative variation in the area total estimates obtained from the simulated samples and express the impact of the randomness on the samples. High values and outliers exist in every distribution. The GREG estimation has different effects on the distributions of the model-free

allocations. The distributions are considerably wider in PRO and NEY allocations. The distribution level of EQU allocation falls, but on the other hand, high values (25.37 % and 20.91 %) for the largest area Uusimaa occur, regardless of the estimation method. The distribution level of NLP allocation falls also, except for two smallest areas. The model-based allocations have otherwise a tight distribution with a quite low level, but they both have one small area as an outlier case. The randomness is best controlled in the EQU/GREG combination and CAL-*g1* allocations.

Figure 2. Distributions of area-specific RRMSE$_d$ %s by allocation and estimation method.

Table 4 contains the bias (ARB%) means over areas and population ARB%´s obtained from EBLUP estimation for every allocation, together with corresponding RRMSE% values. The results concerning both quality measures in the model-based allocations are similar. CAL-*g1* allocation has lower values on the area level, and MC-*q025* performs better on the population level. As expected, the area estimates obtained for the model-free allocations are almost unbiased. The overall performances are evaluated by experimentally combining first the area and population level RRMSE% and ARB% values and then combining the two sums into overall sums. The NLP/GREG and EQU/GREG combinations have the lowest overall sums (25.59 % and 27.13 %), but CAL-*g1* and MC-*q025* allocations have only slightly higher sums.

Table 4. Means over the areas and population values for RRMSE% and ARB% by allocation. The table contains also aggregate values and overall aggregate values.

The Appendix Table 8 contains the area-specific bias (ARB%) values for each allocation and estimation method combination. As can be anticipated, the model-based allocations have considerably higher biases for most of the areas compared with the model-free allocations. The low biases occur only in the same five areas, one of which is small. Four same areas have a bias 10 % or higher, and one of them has a bias as high as over 20 %. The high area biases demonstrate that the used model is inappropriate for those areas. The CAL-*g1* allocation outperforms MC-*q025* allocation according to the area-specific biases.

NEY, PRO, and EQU allocations represent the extreme solutions in the sense that they are either very strongly or not at all related to the area sizes. These solutions lead to good estimation results only on one level. A strong connection between sample and area sizes does not occur in the rest of the allocations (CAL-*g1*, MC-*q025*, and NLP), and excluding a few exceptions, they perform moderately well on both levels. Any pre-set priorities or tolerances are not used in CAL-*g1* allocation, but NLP and MC-*q025* are based on such limitations, and it may be difficult to find proper values for them. The choice of these limitations depends on what importance is addressed to the quality of estimation on the area and population level.

Compared with Horvitz-Thompson estimation, the application of model-assisted GREG estimation improves the accuracy of estimates for EQU and NLP allocations. On the other hand, the GREG estimation leads to reduced accuracy on area level for PRO and NEY allocations, which are tightly related to the area sizes and in which one area (Uusimaa) dominates. EQU and NLP allocations do not have the same kind of dependency on the area sizes.

The two model-based allocations perform moderately well as a whole. The results for small areas indicate that model-based estimation can produce accurate estimates despite a low sample size, but sometimes a much larger sample size is necessary for reaching adequate accuracy. The available auxiliary information suggests that if the characteristics of an area deviate much from the corresponding population characteristics, it can lead to a strong underestimation or overestimation of the area totals, regardless of the area size N_d. If the area sample size n_d is

very low, the synthetic part in the estimator Eq. (3) dominates, and the area total estimate depends almost completely on the sampled units from the other areas.

5. Conclusion

The focus in this study was in resolving how area sample sizes can be controlled in stratified sampling, when the unit-level linear mixed model and EBLUP estimation is applied to the sample data and when the overall sample covers only 1 % of the population. The low overall sample size was a deliberate choice in the sense of highlighting the problems in small area estimation. The control aims at obtaining the area and population estimates with adequate accuracy and low bias. The proposed CAL-*g1* allocation method uses auxiliary information, the model, and method and is derived in the design phase of the survey.

The performance of the proposed allocation both on the area and population level was assessed through design-based sample simulations using real population data. Five allocations selected from the literature served as references. Each of them is based on a different optimization criterion and the use of auxiliary information. The MC-*q025* allocation uses another area model, whereas the other four allocations are model-free. The sample sizes except for equal and proportional allocations were calculated using the previous real register data. EBLUP estimation was applied to the samples in case of model-based allocations. The design-based Horvitz-Thompson and model-assisted GREG estimation using sampling weights were applied to the samples drawn according to model-free allocations. The results indicate that the incorporation of an assisting model does not always improve the estimation results.

The area sample sizes and estimation results have a large variability in the studied allocations. An allocation and estimation method combination with indisputably best performance does not exist among the studied alternatives, if the comparison is based on the accuracies of the area and population estimates. Every combination has high RRMSE% values, and a clear majority of the values over 20 % occur in the distributions of the design-based allocations, regardless of the estimation method.

Proportional and Neyman allocations perform well on the population level, but poorly on area level. It is also noteworthy concerning these two allocations that compared with Horvitz-Thompson estimation, the inclusion of the assisting model leads to reduced accuracies of the area estimates. It seems that under these circumstances with an uncommon area structure and the strong dependency between sample and area sizes, the model-assisted estimation can be more inefficient than Horvitz-Thompson estimation. As far as NLP and equal allocations are concerned, the application of GREG estimation improves also the accuracies of area estimates on the average, in contrast with proportional and Neyman allocations. The distribution of NLP allocation contains two smallest areas as outlier cases, and its overall performance is not the best anyway. The largest area Uusimaa is an outlier case in the distribution of equal allocation, and many other large areas have inaccurate estimates. The population level RRMSE% values which are by far the highest, demonstrate one common weakness of this allocation. As is expected, the area and population estimates are almost unbiased when the design-based estimation is applied.

Cal-*g1* and MC-*q025* allocations perform well both on the population and area level according to RRMSE% values, except for one small area as an outlier case. The population estimates are almost unbiased, but the area-specific distributions contain the same four areas with a strong bias (over 10 %). If these two allocations are evaluated in terms of area-specific bias distributions, CAL-*g1* allocation performs better compared with MC-*q025* allocation, but anyway, the same strongly biased four areas are a common problem for both allocations. This

indicates the model misspecification for these areas. The bias level of a single area remains regardless its sample size.

When analyzing the results from different standpoints, it is worth taking into consideration that they have been obtained in a quite demanding survey and area framework. Although the results are partly contradictory, they support the principle that the used model and estimation method as well as the available auxiliary information are incorporated in the sampling design implemented at the planning stage of the survey. If it is important to obtain accurate area and population estimates, the variation between and within the areas must be included in the allocation solution. Both model-based allocations satisfy these requirements, but the existence of outliers indicates deficiencies which must be corrected.

A wider conception of the performance of the proposed allocation requires, that it is tested together with the reference allocations in various other area frameworks using different study and auxiliary variables. Possible directions for further development of the proposed allocation are the use of every MSE term (not only g_{1d}) and the improvement in calibration of area sample sizes. The complexity of the MSE makes it difficult to reach an analytical solution, and for this reason, the use of software tools like nonlinear programming become necessary. It is likely that an optimization problem relating to the used model has not a closed-form solution in this situation. The question related to MC-*q025* allocation is the setting of priorities between population and area level estimation. This question arises anyway when both the area and population level parameters are estimated, regardless of the estimation method. The choice of the priorities should be a reasonable trade-off between the levels.

Acknowledgements

The authors thank Associate Editor, the referees and Professor Risto Lehtonen for constructive comments and suggestions.

References

[1] Rao JNK, Molina I. Small Area Estimation (2nd Edition) Hoboken, NJ: John Wiley & Sons, Inc.; 2015.

[2] Burgard JP, Münnich R, Zimmermann,T. The impact of sampling designs on small area estimates for Business data. Journal of Official Statistics **30**, No 4, 749–771; 2012.

[3] Molefe WB, Clark RG. Model-assisted optimal allocation for planned domain using composite estimation. Survey Methodology **41**; 2015, 377–387.

[4] Meza JL, Lahiri P. A note on the C_P statistic under the nested error regression model. Survey Methodology **31;** 2005, 105–109.

[5] Cochran, WG. (1977). Sampling Techniques. (3^{nd} edition). New York: John Wiley & Sons.

[6] Costa A, Satorra A, Ventura E. Improving both domain and total area estimation by composition. SORT **28;** 2004 (1) 69–86.

[7] Nissinen K. Small Area Estimation With Linear Mixed Models From Unit-Level Panel and Rotating Panel Data. Ph.D. thesis, University of Jyväskylä, Department of Mathematics and Statistics, Report **117**, https://jyx.jyu.fi/dspace/handle/123456789/21312 ; 2009.

[8] Keto M, Pahkinen E. On sample allocation for effective EBLUP estimation of small area totals – "Experimental Allocation". In: J. Wywial and W. Gamrot (eds.) Survey Sampling Methods in Economic and Social Research. Katowice: Katowice University of Economics; 2010.

[9] Tschuprow AA. On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. Metron **2**;1928 461-493, 646-683.

[10] Gabler S, Ganninger M, Münnich R. Optimal allocation of the sample size to strata under box constraints. Metrika **75**; 2012; 15–161.

[11] Choudhry GH, Rao JNK, Hidiroglou MA. On sample allocation for effective domain estimation. Survey Methodology **38;** 2012; 23–29.

[12] Longford NT. Sample Size Calculation for Small-Area Estimation. Survey Methodology **32**; 2006; 87–96.

[13] Falorsi PD, Righi P. A balanced sampling approach for multi-way stratification for small area estimation. Survey Methodology **34;** 2008; 223–234.

[14] Lehtonen R, Särndal CE, Veijanen A. The effect of model choice in estimation for domains, including small domains. Survey Methodology **29;** 2003; 33–44.

Tables and figures

Table 1
Summary of model-based and model-assisted allocations.

Method	Computing sample size n_d for area d	Optimality level
CAL-*g1*	$n_d^{\mathrm{CAL}-g1} = \dfrac{\widetilde{N}_{g1,d} n - (N - \widetilde{N}_{g1,d}\, D - n)(1/R^2_{a,y^*} - 1)}{N + D(1/R^2_{a,y^*} - 1)}$, where $\widetilde{N}_{g1,d} = SD(y^*)_d / ASD(y^*) N / D.$	Jointly area and population
MC-*q025*	$n_d^{\mathrm{MC}} = \dfrac{n\sigma_d\, N_d^{q/2}}{\sum_{d=1}^{D} \sigma_d\, N_d^{q/2}} + \dfrac{1-\rho}{\rho}\left(\dfrac{\sigma_d\, N_d^{q/2}}{D^{-1}\sum_{d=1}^{D} \sigma_d\, N_d^{q/2}} - 1 \right)$, G = 0 here.	Area

Table 2
Summary of number-based and parameter-based allocations.

Allocation	Computing area sample size n_d	Optimality level
Equal	$n_d^{EQU} = n / D$	Not defined
Proportional	$n_d^{PRO} = (N_d / N)n$	Population
Neyman	$n_d^{NEY} = n\,(N_d S_d / \sum_{d=1}^{D} N_d S_d)$, where S_d is the standard deviation of y (in this study y^*) in area d.	Population
NLP	$n_{st}^{NLP} = \min(\sum_{d=1}^{D} n_d)$ satisfying tolerances $CV(\bar{y}_d) \le CV_{0d}$ and $CV(\bar{y}_{st}) \le CV_0$. In this study y^* replaces y.	Jointly population and area

Table 3
Area sample sizes by allocation. The calibrated area sizes are used for calculating the sample sizes for CAL-*g1* allocation. The sampling population is denoted "Population".

Area (province)	Proxy data		Popu-	Model-based		Model-free			
	True	Calibrated	lation	CAL-	MC-	Number-based		Parameter-based	
	N_d	$\tilde{N}_d$	N_d	*g1* [1]	*q025*	EQU	PRO	NLP	NEY
Uusimaa	7,449	3,516.5	6,813	43	55	12	69	36	125
Pirkanmaa	2,121	1,256.8	2,003	12	14	12	20	11	13
Varsinais-Suomi	1,652	1,670.3	1,543	18	19	12	16	18	14
Päijät-Häme	1,103	1,368.2	1,166	14	14	12	12	13	8
Central Finland	1,219	973.8	1,141	9	8	12	12	9	6
North Ostrobothnia	1,300	1,191.4	1,131	11	11	12	12	9	7
Satakunta	962	1,189.3	1,017	11	11	12	10	15	6
Kymenlaakso	836	911.5	929	8	7	12	10	13	4
Pohjois-Savo	1,009	1,228.7	923	12	11	12	9	13	6
Kanta-Häme	755	1,021.8	885	9	9	12	9	10	5
Etelä-Savo	825	1,032.6	751	9	9	12	8	10	4
South Karelia	481	1,090.7	553	10	9	12	6	12	3
North Karelia	625	1,225.2	549	12	10	12	6	7	4
Lapland	649	1,099.2	544	10	9	12	6	12	3
Ostrobothnia	523	972.2	421	8	7	12	4	8	2
South Ostrobothnia	346	913.3	311	8	6	12	3	6	2
Kainuu	216	706.3	185	5	3	12	2	8	2
Central Ostrobothnia	159	862.3	160	7	4	12	2	6	2
Total	22,230	22,230	21,025	216	216	216	216	216	216

[1)] based on the adjusted homogeneity coefficient (value 0.1697) computed of the proxy variable y^*.

Table 4
Means over the areas and population values for RRMSE% and ARB% by allocation. The table contains also aggregate values and overall aggregate values.

Estimation method	Model-based		Design-based and model-assisted							
Allocation method	CAL-*g1*	MC-*q025*	EQU/H-T	EQU/GREG	PRO/H-T	PRO/GREG	NLP/H-T	NLP/GREG	NEY/H-T	NEY/GREG
					RRMSE%					
Mean over areas (%)	14.02	15.47	19.11	14.71	24.33	26.53	20.13	17.82	30.28	40.68
Population value (%)	6.06	5.13	13.26	10.97	5.94	4.82	8.23	6.35	5.42	4.98
Sum (%)	20.08	20.60	32.37	25.68	30.27	31.35	28.36	24.17	35.70	45.66
					ARB%					
Mean over areas (%)	6.53	7.84	0.37	0.46	0.58	0.43	0.31	0.62	0.79	1.27
Population value (%)	2.48	1.23	0.29	0.99	0.58	0.58	0.17	0.80	0.19	0.30
Sum (%)	9.01	9.07	0.66	1.45	1.16	1.01	0.48	1.42	0.98	1.57
Overall sum (%)	29.09	29.67	33.03	27.13	31.43	32.36	28.84	25.59	36.68	47.23

Appendix

Table 5
Population summary statistics of the study variable *y* obtained from the business register in October 2015 and a proxy variable y* obtained from the business register in April 2015.

Area (province)		Study variable *y* (price)					Proxy variable *y** (price)			
Name	N_d	Total	Mean	St. dev	CV	N_d	Total	Mean	St. dev	CV
Uusimaa	6,813	2,067,530	303.47	271.28	0.894	7,449	2,304,368	309.35	273.26	0.883
Pirkanmaa	2,003	311,634	155.58	106.87	0.687	2,121	332,063	156.56	97.67	0.624
Varsinais-Suomi	1,543	248,763	161.22	145.36	0.902	1,652	263,589	159.56	129.80	0.814
Päijät-Häme	1,166	174,104	149.32	107.30	0.719	1,103	170,514	154.59	106.33	0.688
Central Finland	1,141	153,693	134.70	81.07	0.602	1,219	165,102	135.44	75.67	0.559
North Ostrobothnia	1,131	180,849	159.90	98.22	0.614	1,300	215,869	166.05	92.58	0.558
Satakunta	1,017	111,409	109.55	84.94	0.775	962	118,271	122.94	92.42	0.752
Kymenlaakso	929	91,405	98.39	66.81	0.679	836	85,538	102.32	70.83	0.692
Pohjois-Savo	923	114,935	124.52	100.49	0.807	1,009	137,991	136.76	95.48	0.698
Kanta-Häme	885	106,110	119.90	73.85	0.616	755	98,418	130.36	79.40	0.609
Etelä-Savo	751	89,736	119.49	81.94	0.686	825	109,153	132.31	80.24	0.606
South Karelia	553	64,087	115.89	73.77	0.637	481	61,378	127.60	84.76	0.664
North Karelia	549	96,688	176.12	103.19	0.586	625	116,373	186.20	95.21	0.511
Lapland	544	61,867	113.73	89.11	0.784	649	83,683	128.94	85.42	0.662
Ostrobothnia	421	58,584	139.15	77.63	0.558	523	74,995	143.39	75.55	0.527
South Ostrobothnia	311	41,822	134.48	67.02	0.498	346	51,766	149.61	70.97	0.474
Kainuu	185	15,791	85.36	52.93	0.620	216	21,230	98.29	54.89	0.558
Central Ostrobothnia	160	22,403	140.02	69.53	0.497	159	23,556	148.15	67.01	0.452
Population	21,025	4,011,408	190.79	191.69	1.005	22,230	4,433,859	199.45	175.02	0.877
Mean over areas									95.97	

Table 6
Population summary statistics of the auxiliary variables and correlations between variables obtained from the business register in October 2015

Area (province)		Auxiliary variable x_1 (size)			Auxiliary variable x_2 (age)			Correlations		
Name	N_d	Mean	St. dev	CV	Mean	St. dev	CV	(y,x_1)	(y,x_2)	(x_1,x_2)
Uusimaa	6,813	70.60	28.94	0.410	33.41	30.16	0.903	0.732	0.031	-0.014
Pirkanmaa	2,003	65.02	23.75	0.365	29.63	25.04	0.845	0.649	-0.170	0.133
Varsinais-Suomi	1,543	69.26	28.10	0.406	33.83	22.22	0.657	0.573	-0.306	0.143
Päijät-Häme	1,166	66.07	23.76	0.360	30.84	22.47	0.729	0.576	-0.463	0.031
Central Finland	1,141	63.90	19.62	0.307	25.80	22.57	0.875	0.433	-0.650	0.029
North Ostrobothnia	1,131	65.41	23.11	0.353	18.17	21.90	1.205	0.625	-0.434	0.080
Satakunta	1,017	64.82	20.17	0.311	40.50	24.19	0.597	0.501	-0.163	0.059
Kymenlaakso	929	63.28	24.09	0.381	38.64	23.13	0.599	0.456	-0.508	0.165
Pohjois-Savo	923	66.07	26.19	0.396	36.90	19.28	0.523	0.535	-0.465	-0.044
Kanta-Häme	885	63.22	24.18	0.382	35.05	21.56	0.615	0.499	-0.519	-0.008
Etelä-Savo	751	62.40	20.83	0.334	34.02	20.62	0.606	0.423	-0.521	-0.009
South Karelia	553	61.91	18.08	0.292	33.83	21.31	0.630	0.458	-0.542	0.048
North Karelia	549	61.94	18.98	0.307	20.20	21.80	1.079	0.473	-0.680	0.027
Lapland	544	64.63	25.15	0.389	31.98	21.58	0.675	0.532	-0.573	0.033
Ostrobothnia	421	61.56	25.94	0.421	33.08	28.41	0.859	0.513	-0.248	0.181
South Ostrobothnia	311	64.61	24.15	0.374	25.68	22.18	0.864	0.221	-0.657	0.253
Kainuu	185	58.84	20.51	0.349	36.35	16.10	0.443	0.472	-0.590	-0.029
Central Ostrobothnia	160	75.08	40.78	0.543	40.39	26.23	0.649	0.578	-0.145	0.293
Population	21,025	66.72	25.75	0.386	32.11	25.85	0.805	0.592	-0.097	0.044

Table 7
Area and population level RRMSE%s by allocation and estimation method. The values are computed of the simulated samples drawn from the business register in October 2015.

Area (province)	N_d	Model-based		Design-based H-T and model-assisted							
		CAL-*g1*	MC-*q025*	EQU/H-T	EQU/GREG	PRO/H-T	PRO/GREG	NLP/H-T	NLP/GREG	NEY/H-T	NEY/GREG
Uusimaa	6,813	12.15	9.95	25.37	20.91	10.10	7.66	14.89	11.28	7.85	5.50
Pirkanmaa	2,003	10.14	9.72	19.56	14.66	15.01	12.08	21.21	15.57	19.20	17.86
Varsinais-Suomi	1,543	12.01	11.77	25.77	18.11	23.08	17.52	21.46	15.79	23.91	21.31
Päijät-Häme	1,166	10.14	10.38	20.42	14.02	20.92	17.03	19.68	15.33	25.26	24.62
Central Finland	1,141	11.39	12.25	17.32	11.97	16.94	16.20	20.24	16.03	23.77	29.58
North Ostrobothnia	1,131	8.80	9.22	17.97	11.51	17.35	14.55	19.86	13.73	23.25	23.15
Satakunta	1,017	16.72	17.87	22.29	18.81	24.27	24.69	19.91	18.15	31.00	35.68
Kymenlaakso	929	20.62	23.74	19.07	14.72	21.33	26.25	18.88	18.48	32.43	55.80
Pohjois-Savo	923	14.45	16.24	22.50	16.93	26.50	25.27	22.70	17.69	33.76	38.47
Kanta-Häme	885	12.90	13.76	17.17	13.25	20.42	22.61	19.14	16.90	27.32	38.37
Etelä-Savo	751	13.50	14.08	18.92	15.26	23.93	23.90	21.18	18.74	34.25	40.48
South Karelia	553	12.55	13.15	18.23	13.24	25.50	24.46	18.05	15.46	36.32	44.27
North Karelia	549	9.63	11.13	17.01	10.90	24.20	19.71	21.64	15.96	29.58	28.34
Lapland	544	16.23	19.74	22.64	15.67	30.86	32.44	22.54	18.28	45.09	55.22
Ostrobothnia	421	11.66	12.45	15.75	14.13	28.14	33.23	19.26	19.34	37.96	57.19
South Ostrobothnia	311	13.19	14.94	13.59	11.67	30.50	40.15	20.25	21.41	36.56	61.48
Kainuu	185	26.77	32.16	17.12	15.24	43.64	61.49	21.63	26.13	43.71	80.85
Central Ostrobothnia	160	19.45	25.89	13.31	13.82	35.19	58.30	19.81	26.54	33.80	72.95
Mean over areas (%)		14.02	15.47	19.11	14.71	24.33	26.53	20.13	17.82	30.28	40.68
Population value (%)		6.06	5.13	13.26	10.97	5.94	4.82	8.23	6.35	5.42	4.98

Table 8
Area and population level ARB%s by allocation and estimation method. The values are computed of the simulated samples drawn from the business register in October 2015.

Area (province)	N_d	Model-based		Design-based H-T and model-assisted							
		CAL-	MC-	EQU/	EQU/	PRO/	PRO/	NLP/	NLP/	NEY/	NEY/
		g1	*q025*	H-T	GREG	H-T	GREG	H-T	GREG	H-T	GREG
Uusimaa	6,813	7.63	5.94	0.53	1.61	0.94	1.71	0.40	1.64	0.55	0.98
Pirkanmaa	2,003	1.28	1.14	0.34	0.35	0.44	0.09	0.07	0.04	0.42	0.22
Varsinais-Suomi	1,543	0.83	0.46	0.32	1.12	0.60	0.01	0.15	0.10	0.07	0.18
Päijät-Häme	1,166	0.85	1.03	0.48	0.63	0.10	0.11	0.24	0.39	0.07	0.32
Central Finland	1,141	5.09	5.84	0.25	0.14	0.33	0.23	0.33	0.37	0.30	0.16
North Ostrobothnia	1,131	1.53	1.38	0.09	0.08	0.42	0.10	0.16	0.39	0.78	0.46
Satakunta	1,017	7.77	9.41	0.32	0.97	0.12	0.21	0.52	1.03	0.36	0.06
Kymenlaakso	929	14.84	17.66	0.60	0.75	0.06	0.49	0.40	0.06	0.68	1.37
Pohjois-Savo	923	5.40	6.54	0.39	0.59	1.68	0.02	0.45	0.73	1.04	0.52
Kanta-Häme	885	5.63	6.67	0.23	0.03	0.39	0.59	0.31	0.66	0.28	0.11
Etelä-Savo	751	5.14	5.66	0.44	0.30	0.64	1.01	0.09	0.42	0.38	3.47
South Karelia	553	5.94	6.10	0.18	0.07	1.47	0.64	0.09	0.09	1.45	1.44
North Karelia	549	4.32	6.45	0.27	0.12	0.05	0.02	0.23	0.54	0.39	0.15
Lapland	544	10.36	13.17	0.40	0.62	1.66	0.69	0.23	1.00	0.97	0.99
Ostrobothnia	421	2.15	1.68	0.12	0.24	0.17	0.00	0.69	0.01	0.82	1.41
South Ostrobothnia	311	6.58	7.84	0.39	0.43	1.21	0.35	0.01	0.59	2.53	3.49
Kainuu	185	21.64	27.14	0.76	0.18	0.21	0.55	0.92	0.22	1.48	0.90
Central Ostrobothnia	160	10.59	16.93	0.49	0.04	0.01	0.94	0.20	2.97	1.68	6.63
Mean over areas (%)		6.53	7.84	0.37	0.46	0.58	0.43	0.31	0.62	0.79	1.27
Population value (%)		2.48	1.23	0.29	0.99	0.58	0.58	0.17	0.80	0.19	0.30

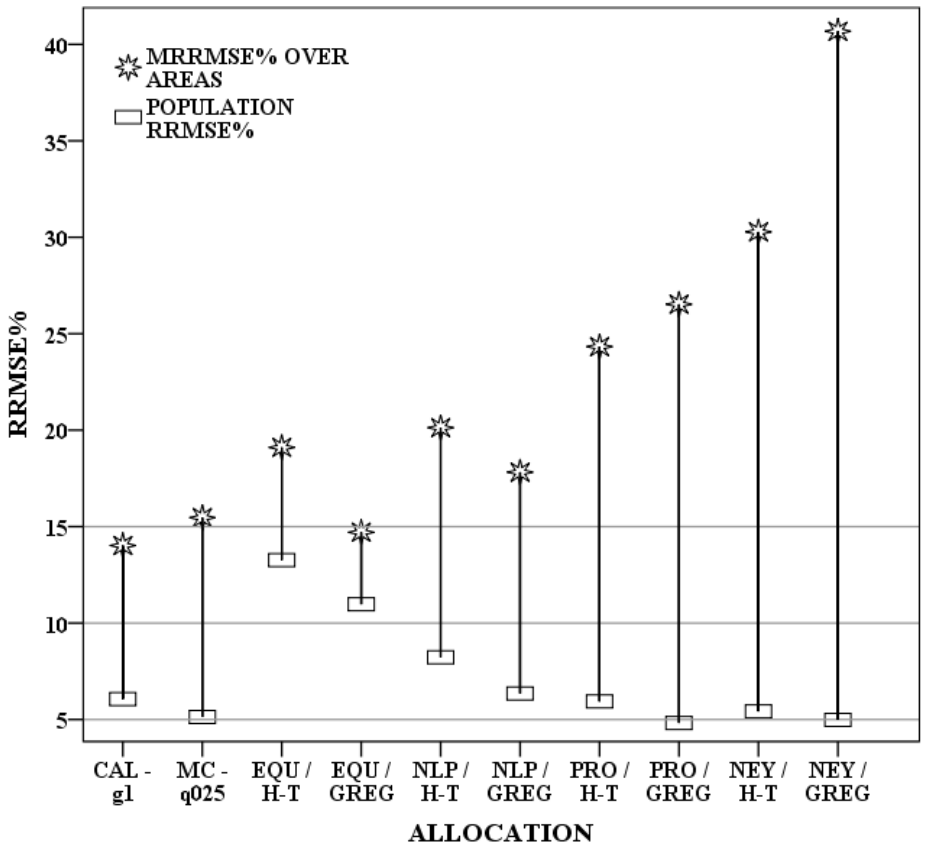

Figure 1. Means of area RRMSE$_d$%s (MRRMSE%) and population RRMSE%s by allocation and estimation method. EBLUP estimation is applied to CAL-*g1* and MC-*q025* allocations.

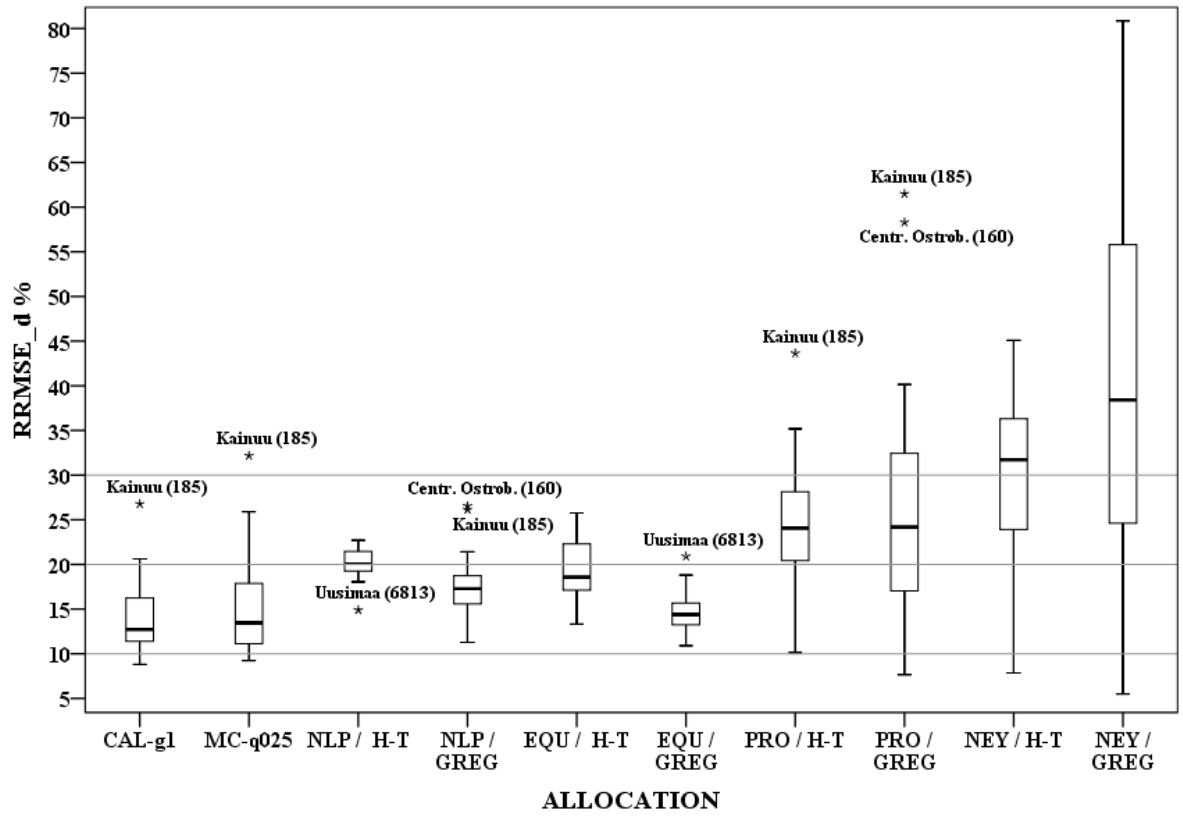

Figure 2. Distributions of area-specific RRMSE$_d$ %s by allocation and estimation method.

IV

REGISTER DATA IN SAMPLE ALLOCATIONS FOR SMALL AREA ESTIMATION

by

Mauno Keto, Jussi Hakanen, and Erkki Pahkinen, 2018

This is an accepted manuscript of an article published by Taylor & Francis in Mathematical Population Studies, An International Journal of Mathematical Demography, available online: https://doi.org/10.1080/08898480.2018.1437318

Register data in sample allocations for small-area estimation

Mauno Keto[a], Jussi Hakanen[b], and Erkki Pahkinen[c]

[a]Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland; [b]Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland; [c]Department of Mathematics and Statistics, University of Jyväskylä, Jyväskylä, Finland

Abstract

The inadequate control of sample sizes in surveys using stratified sampling and area estimation may occur when the overall sample size is small or auxiliary information is insufficiently used. Very small sample sizes are possible for some areas. The proposed allocation based on multi-objective optimization uses a small-area model and estimation method, and semi-annually collected empirical data. The assessment of its performance at the area and population levels is based on design-based sample simulations, and five previously developed allocations serve as references. The model-based estimator is more accurate than the design-based Horvitz-Thompson estimator and model-assisted regression estimator. Two trade-off issues are between accuracy and bias and between the area- and population-level qualities of estimates. If the survey uses model-based estimation, the sampling design should incorporate the underlying model and the estimation method.

Key words: Auxiliary and proxy data, model-based EBLUP, performance, multi-objective optimization, trade-off between areas and population.

CONTACT Mauno Keto. E-mail: mauno.j.keto@student.jyu.fi. Faculty of Information Technology, University of Jyväskylä, Finland. Jyväskylä, Mattilanniemi 2 – Box 35, FIN–40014, Jyväskylä, Finland.

1. Introduction

Sample surveys provide estimates of the various parameters not only for the population of interest, but also for subpopulations, referred to as "areas" here. Stratified sampling is a common design, where strata and areas coincide. How are area sample sizes controlled to provide satisfactory area and population estimates? The small overall sample size or an insufficient use of auxiliary information may lead to the fact that the areas are not defined at the planning stage of the survey. The consequence is that the area sample sizes cannot be controlled. Nonresponse as one cause of randomness is beyond the scope of the study. The lack of control can lead to small or even to null sample sizes for some areas. They are regarded as small, because the area-specific samples are small enough to hinder direct estimates of adequate precision (Rao and Molina, 2015). Various model-assisted or model-based small-area estimation techniques, which are hard to implement, have been designed to solve this problem (Pfeffermann, 2013). The World Bank uses the software *PovMap* for producing business statistics. Burgard, Münnich, and Zimmermann (2014) have used various estimators and studied the performances of small-area point and accuracy estimates under different sampling designs.

We estimate the area and population totals of the variable of interest under different sampling designs. The variable measures some quantity in business. Because the overall sample size is small and the population contains small areas, model-based estimation yields moderately accurate area estimates. The "borrowing strength" principle implies that sample information provides a higher estimation power for small areas. Two auxiliary variables correlated with the variable of interest serve as predictors. The selected model contains area-specific effects, because the variable of interest is likely to vary from one area to another. We shall compare the main estimation method, which is model-based, to the design-based Horvitz-Thompson estimator and to the model-assisted regression estimator, on the basis of model-free allocations. The model-based estimators have lower variances, but may be biased. The design-based estimators are design-unbiased, but their variances are large for small areas with small sample size. The second motivation for using these three estimators is to clarify the trade-off between accuracy and bias.

Our allocation method, called "three-term Pareto method", also uses the model and the estimation method as auxiliary information at the planning stage. It is based on multiobjective optimization, the model-based empirical best linear unbiased predictor (EBLUP) estimator for obtaining the area and population total estimates of the variable of interest, and the mean squared error estimator. We shall compare this method with five reference methods displaying various optimization criteria and using auxiliary information. The method called "Molefe and Clark", also uses an area model. We introduce model-related allocations in section 2 and four model-free allocations in Section 3: "Equal," "Costa," "nonlinear programming," and modified "box-constraint". A fixed, small overall sample size is a common restriction. We present additional numerical details related to some allocations in section 4.2.

We simulate the allocation-specific random samples from a population containing real register data, by using stratified simple random sampling without replacement. Because the variable of interest is unknown and the between-area variation of each auxiliary variable in the population is too small to support allocation, the allocation-specific sampling design, except for equal allocation, is based on previous register data, called "proxy data".

The relative root mean square error and the absolute relative bias measure the accuracy and the bias of an estimator in design-based simulations. They are sample-based approximations of the design mean squared error and of the design bias. The primary measure is the relative root mean square error, but we also compute the absolute relative bias for design-based estimates. The area-specific relative biases reflect the validity of the model in each area. There is a trade-

off between the quality of area estimates and the quality of population estimate; and a second trade-off between accuracy and bias.

The results support the sampling strategy, in which not only auxiliary information, but also the model and the estimation method should be fixed early, in the design phase of the survey. The proposed allocation uses all information available before choosing the allocation method, avoiding fixed priorities for the importance of estimation at the area and the population levels.

2. Model-related allocations

In the model-based estimation, the area parameter and often the population parameter estimates result from the statistical model and from the chosen estimator. The proposed allocation (section 2.2) is based on the model and the estimator introduced in section 2.1 and on auxiliary information. Keto and Pahkinen (2010) have used this model and this estimator to describe the relationships between area and sample sizes, estimation results, and area characteristics. One reference allocation (section 2.3) is based on a different area model and on a composite estimator, and uses auxiliary information. These two allocations are "model-related allocations". Table 1 shows the summary details of these allocations.

2.1. Model and model-based area total estimator

The area total estimator of the variable of interest is based on the linear mixed model (Battese, Harter, and Fuller, 1988):

$$y_{dk} = x'_{dk}\beta + v_d + e_{dk};\ k = 1,...,\ N_d;\ d = 1,...,\ D, \tag{1}$$

where x_{dk} is the vector of auxiliary information for unit k in area d, D is the total number of areas, N_d is the size (number of units) of area d, β is the vector of fixed regression parameters, the area-specific effects v_d are distributed as $N(0,\sigma_v^2)$, independently of the random errors e_{dk}, which are distributed as $N(0,\sigma_e^2)$. The first value of the vector x_{dk} is one, and the vector β contains the intercept term β_0. Eq. (1) is applicable when unit-level values are available for the variables x.

The expected value for the unit k in area d is $E(y_{dk}) = x'_{dk}\beta$, and the total variance

$$V(y_{dk}) = \sigma_v^2 + \sigma_e^2 \tag{2}$$

is decomposed into the variance σ_v^2 between areas and the variance σ_e^2 within areas. The common intra-area correlation (Meza and Lahiri, 2005)

$$\rho = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_e^2} \tag{3}$$

measures the relative variation of the variable of interest between the areas.

Before the area parameters, we estimate the model parameters and the area effects from the sample data. We denote $\hat{\sigma}_v^2$ and $\hat{\sigma}_e^2$ the estimated variance components, and $\hat{v}_d$ the EBLUP area effects. The estimate $\hat{\beta}$ of β is obtained using the generalized least-squares method.

The EBLUP estimator for the area total Y_d is the sum of n_d sampled y-values and the sum of predicted y-values for ($N_d - n_d$) non-sampled units:

$$\hat{Y}_{d,\text{Eblup}} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} \hat{y}_{dk} = \sum_{k \in s_d} y_{dk} + \sum_{k \in \bar{s}_d} x'_{dk}\beta + (N_d - n_d)\,\hat{v}_d \ , \tag{4}$$

where s_d and $\bar{s}_d$ denote the sampled and the non-sampled units, and the vectors x_{dk} and β are defined as in Eq. (1). The design mean squared error for the estimator in Eq. (4)

$$\text{MSE}\,(\hat{Y}_{d,\text{Eblup}}) = E(\hat{Y}_{d,\text{Eblup}} - Y_d)^2 = V(\hat{Y}_{d,\text{Eblup}}) + (E(\hat{Y}_{d,\text{Eblup}}) - Y_d)^2 \,. \tag{5}$$

is the sum of the variance and the squared bias. The Prasad-Rao prediction mean squared error estimator (Rao and Molina, 2015) for finite populations is

$$\text{mse}(\hat{Y}_{d,\text{Eblup}}) = g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{2d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + 2g_{3d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) + g_{4d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) \tag{6}$$

where the terms g_{1d}, g_{2d}, g_{3d}, and g_{4d} are functions of the variance components

$$\begin{aligned}
&g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = (N_d - n_d)^2(1 - \hat{\gamma}_d)\hat{\sigma}_v^2 \,, \\
&g_{2d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = (N_d - n_d)^2(\bar{x}_d^{\,*} - \hat{\gamma}_d \bar{x}_d)'(X'\hat{V}^{-1}X)^{-1}(\bar{x}_d^{\,*} - \hat{\gamma}_d \bar{x}_d) \,, \\
&g_{3d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = (N_d - n_d)^2 (n_d)^{-2} (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 (n_d)^{-1})^{-3} (\hat{\sigma}_e^4 \text{V}(\hat{\sigma}_v^2) + \hat{\sigma}_v^4 \text{V}(\hat{\sigma}_e^2) \\
&\quad -2\hat{\sigma}_e^2 \hat{\sigma}_v^2 \text{Cov}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)) \,, \\
&g_{4d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = (N_d - n_d)\hat{\sigma}_e^2 \,.
\end{aligned} \tag{7}$$

The terms g_{1d} and g_{2d} include the shrinkage factor

$$\hat{\gamma}_d = \hat{\sigma}_v^2 (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 n_d^{-1})^{-1} \,. \tag{8}$$

The matrix X contains the sampled values of the auxiliary variables, and the vectors $\bar{x}_d$ and $\bar{x}_d^*$ contain the area-specific means for the sampled and the non-sampled x-values. The variance-covariance matrix $V = V(y)$ has a block diagonal form, with the blocks V_d defined as (Meza and Lahiri, 2005):

$$V_d = (1 - \rho) I_{n_d} + \rho J_{n_d} \,, \tag{9}$$

where ρ is defined in Eq. (3), I_{n_d} is the $n_d \times n_d$ identity matrix, and J_{n_d} is the $n_d \times n_d$ matrix, whose all entries are equal to 1. The term g_{3d} contains the asymptotic variances $\text{V}(\hat{\sigma}_v^2)$ and $\text{V}(\hat{\sigma}_e^2)$, and the asymptotic covariance $\text{Cov}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)$. If these parameters are estimated by restricted maximum likelihood, the estimator in Eq. (6) is approximately unbiased (Nissinen, 2009). The area-specific mean squared error estimates are obtained when the variance parameter estimates are inserted into Eq. (7).

Nissinen (2009) states that the term g_{1d} contributes for 85–90% of the estimated mean squared error, that the proportion of g_{4d} is seldom over 1%, that the proportion of g_{2d} is between 4 and 6%, and that the proportion of g_{3d} is between 6 and 10%. We obtained similar percentages in our simulations. The high proportion of g_{1d} indicates that the variation in the area estimates is mostly related to the uncertainty about the area effects (Nissinen, 2009).

The proposed allocation in Eq. (17) uses three terms of the mean squared error estimator in Eq. (6). The term g_{2d} is excluded because of its small proportion of the estimated mean squared

error and because it involves complex matrix operations and auxiliary variables, whose values depend on the sample.

2.2. Model-based three-term Pareto method allocation using multiobjective optimization

A sample allocation is often based on the solution of an optimization problem subject to given restrictions. It is related to the sample design and to the variance, mean squared error, and the coefficient of variation of the estimator.

Our allocation uses the approximation of the mean squared error (amse) in Eq. (6):

$$\begin{aligned}&\operatorname{amse}(\hat{Y}_{d,\mathrm{Eblup}}) = g_{1d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) + 2g_{3d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) + g_{4d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2)\\&= (N_d - n_d)^2(1-\hat{\gamma}_d)\hat{\sigma}_v^2\\&+2(N_d - n_d)^2(n_d)^{-2}(\hat{\sigma}_v^2 + \hat{\sigma}_e^2(n_d)^{-1})^{-3}(\hat{\sigma}_e^4\mathrm{V}(\hat{\sigma}_v^2) + \hat{\sigma}_v^4\mathrm{V}(\hat{\sigma}_e^2) - 2\hat{\sigma}_e^2\hat{\sigma}_v^2\mathrm{Cov}(\hat{\sigma}_e^2,\hat{\sigma}_v^2))\\&+(N_d - n_d)\hat{\sigma}_e^2. \qquad (10)\end{aligned}$$

Eq. (10) contains the fixed area sizes N_d, the area sample sizes n_d to be found by optimization, and the unknown variance and covariance parameters. Their values are estimated through sample simulations drawn from the register of proxy data (section 1), together with auxiliary variables. The estimates of the variance and covariance parameters depend on the sample. The means of their sample estimates are inserted into Eq. (10). The sum of the area-specific approximations in Eq. (10)

$$\operatorname{amse}(\hat{Y}_{\mathrm{Eblup}}) = \sum_{d=1}^{D} \operatorname{amse}(\hat{Y}_{d,\mathrm{Eblup}}) \qquad (11)$$

is an approximation for the mean squared error estimator of the population total estimator $\hat{Y}_{\mathrm{Eblup}} = \sum_{d=1}^{D} \hat{Y}_{d,\mathrm{Eblup}}$.

The design-based direct estimator $\hat{Y}_d = N_d\bar{y}_d$ ($\bar{y}_d$ is the sample mean) is the estimator for the area total Y_d and $\hat{Y} = \sum_d N_d\bar{y}_d$ is the estimator for the population total *Y*. The design coefficients of variation (CV) of these estimators are

$$\begin{aligned}&\mathrm{CV}(\hat{Y}_d) = \frac{\mathrm{V}(N_d\bar{y}_d)^{\frac{1}{2}}}{Y_d},\\&\mathrm{CV}(\hat{Y}) = \frac{(\sum_d \mathrm{V}(N_d\bar{y}_d))^{\frac{1}{2}}}{Y}. \qquad (12)\end{aligned}$$

In the model-based estimation, the mean squared error replaces the variance, and in accordance with the design-based estimation, the approximate coefficient of variation (ACV) for the area total estimates $\hat{Y}_{d,\mathrm{Eblup}}$ and the population total estimate $\hat{Y}_{\mathrm{Eblup}}$ are:

$$\begin{aligned}&\mathrm{ACV}(\hat{Y}_{d,\,\mathrm{Eblup}}) = \frac{\operatorname{amse}(\hat{Y}_{d,\,\mathrm{Eblup}})^{\frac{1}{2}}}{Y_d},\\&\mathrm{ACV}(\hat{Y}_{\mathrm{Eblup}}) = \frac{\operatorname{amse}(\hat{Y}_{\mathrm{Eblup}})^{\frac{1}{2}}}{Y}, \qquad (13)\end{aligned}$$

where Y_d and Y are obtained from the variable of interest in the proxy data. We denote this variable "y^*".

This allocation should provide the optimal accuracy both on area and population levels. This is the reason why the optimal area sample sizes result from a multi-objective optimization, yielding the minimal approximate population coefficient of variation and the minimal mean of approximate coefficients of variation over areas. For multi-objective optimization, there may exist several solutions, so-called Pareto optimal solutions, where none of the objectives can be improved without impairing another one (Miettinen, 1999). In this case, the Pareto optimal solutions are such that smaller values for the approximate population coefficient of variation cannot be obtained without letting the mean of the approximate coefficient of variation over areas increase, and conversely. For two objectives, the Pareto front consisting of all optimal solutions is a curve in the two-dimensional objective space. Then all solutions on the Pareto front are candidates for the final solution, in the absence of information on preference. A multi-objective optimization problem is solved either by approximating the whole Pareto front or by identifying a preferred solution from the Pareto front. In the first alternative, a set of Pareto optimal solutions is generated through optimization. It approximates the whole set, which can be infinite, of Pareto optimal solutions. In the second alternative, we take account of information on preference in the optimization and identify a Pareto-optimal solution as close as possible to this information. We develop both alternatives. The functions to be optimized are too complicated to yield closed-form solutions, so that nonlinear numerical optimization method is mandatory. The area sample sizes are the variables in the multi-objective optimization subject to the constraints

$$\begin{aligned} &\textstyle\sum_{d=1}^{D} n_d = n, \\ &n_d \geq 1 \text{ and } n_d \ (d = 1,\ldots, D) \text{ are integers} \\ &n_d \leq N_d \ (N_d \leq n \text{ is possible for the smallest areas}). \end{aligned} \tag{14}$$

To approximate the Pareto front, we use the ε-constraint method (Miettinen, 1999), where one objective is minimized while the other one is converted into a constraint with a fixed upper bound ε. The solutions on the Pareto front are then obtained by solving the resulting single objective optimization problems where we use different values for the upper bound ε. If the resulting single objective problems are not convex, then the globally optimal solutions may be intractable and we resort to an appropriate single objective optimization method. If the solutions are only locally Pareto optimal, they are Pareto optimal in some neighborhood of the solution. We use the ε-constraint method also in the nonlinear programming allocation (section 3.3), because it corresponds to a multi-objective minimization of the overall sample size n, of the coefficient of variation for each area, and of the coefficient of variation for the whole population. This problem includes D+2 objective functions.

Figure 1 shows an example of the approximated Pareto front, where the approximate population coefficient of variation is minimal under the constraints of 48 upper bounds for the approximate mean coefficient of variation over areas, corresponding to 48 Pareto optimal solutions (denoted by the star symbols). Each solution represents an allocation with corresponding area sample sizes. The Pareto front allows the selection of the allocation. It shows the trade-offs between the two objectives.

The second alternative is to use preference information for identifying the preferred trade-off, without computing all Pareto optimal solutions. We have used the method of global criterion (Miettinen, 1999). The principle is to minimize the distance to the vector whose components are the optimal values for each objective. First we compute the minimum of the approximate population coefficient of variation in Eq. (13), subject to the constraints of Eq.

(14). The mean approximate coefficient of variation over the areas is ignored in this optimization. Second, we compute the minimal mean over the areas:

$$\text{MACV} = \frac{\sum_{d=1}^{D} \text{ACV}(\hat{Y}_{d,\,\text{Eblup}})}{D} \quad , \tag{15}$$

subject to the constraints of Eq. (14), while ignoring the approximate population coefficient of variation. The resulting area sample sizes in these two optimizations are only by-products. These two minima form the ideal objective vector and are denoted

$$\text{Min}_{\,\text{pop}} = \min(\text{ACV}(\hat{Y}_{\text{Eblup}})),$$

$$\text{Min}_{\,\text{are}} = \min(\text{MACV}) \quad , \tag{16}$$

subject to constraints of Eq. (14).

We set the initial values on the area sample sizes n_d, and minimize the sum of squares

$$S = (\text{ACV}(\hat{Y}_{\text{Eblup}}) - \text{Min}_{\,\text{pop}})^2 + (\text{MACV} - \text{Min}_{\,\text{are}})^2, \tag{17}$$

subject to the constraints of Eq. (14). We obtain the preferred area sample sizes. The solution of Eq. (17) is a trade-off between the estimation efficiencies at the area and at the population levels. Figure 1 shows the solution obtained by using this allocation, which belongs to the Pareto front and is the closest to the objective vector. The dotted lines indicate the values of the vector constituting the objective.

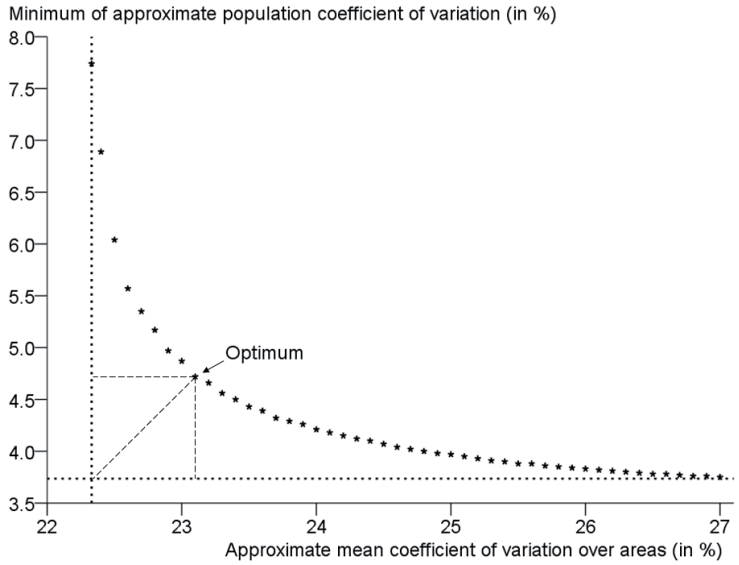

Figure 1: The approximated Pareto front minimizing the mean of approximate coefficients of variation over areas and of the approximate population coefficient of variation. The label “Optimum” denotes the Pareto optimal solution.

The accuracy at the population level improves to the detriment of accuracy at the area level. The optimal allocation corresponding to the three-term Pareto method allocation has a minimal distance to the objective vector. We use the Excel Solver with the option “generalized reduced

gradient nonlinear" to provide the full Pareto optimal solutions to the single objective optimization problems.

2.3. Model-assisted Molefe and Clark´s allocation

Molefe and Clark (2015) have developed an allocation based on a composite estimator for estimating the area-specific means of the variable of interest. A simple random sample of n_d units is selected from each stratum $d = 1, \ldots, D$, defined by small areas and containing N_d units. The relative size of the area d is $P_d = N_d / N$.

The estimator

$$\hat{\bar{y}}_{C} = (1-\phi_d)\bar{y}_{dr} + \phi_d \hat{\beta}' \bar{X}_d \tag{18}$$

combines a synthetic estimator $\hat{\bar{Y}}_{d(\text{syn})} = \hat{\beta}' \bar{X}_d$, where $\hat{\beta}$ is the coefficient in the regression Eq. (18) and $\bar{X}_d$ the vector of area-specific means of auxiliary variables, and a direct estimator $\bar{y}_{dr} = \bar{y}_d + \hat{\beta}'(\bar{x}_d - \bar{X}_d)$, where $\hat{\beta}$ and $\bar{X}_d$ are the same as in the estimator in Eq. (18), and $\bar{y}_d$ and $\bar{x}_d$ are the sample means of the variable of interest and of auxiliary variables in the area d. The coefficients ϕ_d minimize the design mean squared error of the estimator in Eq. (18). Under the conditions given by Molefe and Clark (2015), the approximate design-based mean squared error estimator of Eq. (18) is

$$\text{MSE}_p(\tilde{\bar{y}}_d^C; \bar{Y}_d) \approx (1-\phi_d)^2 v_{d(\text{syn})} + \phi_d^2 B_d^2, \tag{19}$$

where $v_{d(\text{syn})}$ is the sampling variance of the synthetic estimator $\hat{\bar{Y}}_{d(\text{syn})}$. The bias is $B_d = \hat{\beta}_U' \bar{X}_d - Y_d$, where $\hat{\bar{Y}}_{d(\text{syn})}$ is used to estimate $\bar{Y}_d$, with β_U denoting the approximate design-based expectation of $\hat{\beta}$.

Molefe and Clark (2015) assume a two-level linear model ξ, conditional on the values of the auxiliary variables x, with uncorrelated stratum random effects u_d and unit residuals ε_i:

$$\begin{cases} y_i = \beta' x_i + u_d + \varepsilon_i \\ E_\xi(u_d) = E_\xi(\varepsilon_i) = 0 \\ V_\xi(u_d) = \sigma_{ud}^2 \\ V_\xi(\varepsilon_i) = \sigma_{ed}^2 \end{cases}, \tag{20}$$

where i refers to the unit i in the stratum d. This model implies that the area-specific variance of the variable of interest according to Eq. (20) is $V_\xi(y_i) = \sigma_{ud}^2 + \sigma_{ed}^2 = \sigma_d^2$ and holds for all population units. The covariance of y-values between two units i and $j \neq i$ is $\text{cov}_\xi(y_i, y_j) = \rho_d \sigma_d^2$ for units in the same stratum and zero otherwise, where

$$\sigma_d = \frac{\sigma_{ud}^2}{\sigma_{ud}^2 + \sigma_{ed}^2} \tag{21}$$

is the intra-class correlation in the area *d*. Molefe and Clark (2015) assume that the areas have a common intra-class correlation $\rho_d = \rho$ for all *d*. The ratio of between-area variation to the total variation of *y* is constant.

After computing the optimal weight ϕ_d in Eq. (19), we obtain the approximate optimal anticipated mean squared error:

$$\mathrm{AMSE}_d = E_\xi \mathrm{MSE}_p(\tilde{y}_d^C(\phi_{d(opt)}); \bar{Y}_d) \approx \sigma_d^2 \rho(1-\rho)(1+(n_d-1)\rho)^{-1}. \tag{22}$$

The criterion *F* using anticipated mean squared errors of the small-area mean and the overall mean estimators for the model-assisted allocation has the approximative form:

$$\begin{aligned} F &= \sum_{d=1}^D N_d^q \mathrm{AMSE}_d + G N_+^{(q)} E_\xi \,\mathrm{var}_p\left(\hat{\bar{Y}}_r\right) \\ &\approx \sum_{d=1}^D N_d^q \sigma_d^2 \rho(1-\rho)(1+(n_d-1)\rho)^{-1} + G N_+^{(q)} \sum_{d=1}^D \sigma_d^2 P_d^2 n_d^{-1}(1-\rho). \end{aligned} \tag{23}$$

The optimal area sample sizes minimize Eq. (23) subject to $\sum_{d=1}^D n_d = n$, and the solution is consistent with Longford (2006). The weight N_d^q reflects the inferential priority for area *d*, with $0 \le q \le 2$ and $N_+^{(q)} = \sum_d N_d^q$. The quantity *G* is a relative priority coefficient at the population level. When *G* is null, we focus on area-level estimation. The larger *G*, the less important the area-level estimation. The values of *q* and *G* depend on these priorities.

When the population estimation has no priority (*G* = 0) and the cost of the survey are fixed, the minimization of Eq. (23) with respect of n_d has the unique solution

$$n_d^{\mathrm{MC}} = \frac{n\sigma_d N_d^{\frac{q}{2}}}{\sum_{d=1}^D \sigma_d N_d^{\frac{q}{2}}} + \frac{1-\rho}{\rho}\left(\frac{\sigma_d N_d^{\frac{q}{2}}}{D^{-1}\sum_{d=1}^D \sigma_d N_d^{\frac{q}{2}}} - 1\right). \tag{24}$$

In Eq. (23) and (24), both the intra-class correlation ρ and the area-specific standard deviation σ_d of the variable of interest *y* are unknown. We replace the intra-class correlation ρ by the adjusted homogeneity coefficient obtained from the proxy variable of interest y^*:

$$R_{a,y^*}^2 = 1 - \frac{\mathrm{MSW}}{S_{y^*}^2}, \tag{25}$$

where MSW is the mean sum of squares of areas, provided by a one-way analysis of variance between the areas in the proxy population, and $S_{y^*}^2$ is the variance of y^*. We replace the parameter σ_d by the standard deviation of the proxy variable y^* in the area *d*.

The reason for both replacements is the link between *y* and y^*. The allocation favors large areas with large variances of y^*: the higher the value of the constant *q*, the more likely the occurrence of negative sample sizes for small areas with small variances. Also, if the population estimate has a strictly positive priority *G*, then *F* in Eq. (23) must be minimized numerically; theoretical values of *q* and *G* are out of reach.

Table 1: Summary of model-based and model-assisted allocations.

Allocation	Computing sample size for area $d = 1,\ldots, D$	Optimality level
Three-term Pareto method	n_d^{Pareto} : sample sizes minimize the sum of squares $S = (\text{ACV}(\hat{Y}_{\text{Eblup}}) - \text{Min}_{\text{pop}})^2 + (\text{MACV-Min}_{\text{are}})^2$, based on the approximate coefficients of variation according to Eq. (13), at the area and population level. The register of proxy data is used.	Jointly area and population
Molefe and Clark	$n_d^{\text{MC}} = \frac{n\sigma_d N_d^{\frac{q}{2}}}{\sum_{d=1}^{D}\sigma_d N_d^{\frac{q}{2}}} + \frac{1-\rho}{\rho}\left(\frac{\sigma_d N_d^{\frac{q}{2}}}{D^{-1}\sum_{d=1}^{D}\sigma_d N_d^{\frac{q}{2}}} - 1\right)$, where q is an adjustable constant ($0 \le q \le 2$), ρ is the common intra-area correlation, and σ_d is the area-specific standard deviation obtained from the proxy variable y^*.	Area

3. Model-free reference allocations

One of the model-free reference allocations, equal allocation, uses only number-based information. Others use both number-based and parameter-based information on the variable of interest, which is unknown and is replaced by a proxy variable y^*. It can be the same variable obtained from an earlier research of the same subject. An auxiliary variable correlated with the variable of interest also can serve as a proxy variable if its area characteristics are available. Table 2 shows the summary details of these allocations introduced in sections 3.1–3.4.

3.1. Equal allocation

In equal allocation, the sample size is

$$n_d^{\text{EQU}} = \frac{n}{D} \ . \tag{26}$$

The expression of this allocation in Eq. (26) includes neither the area-specific characteristics nor the between-area variation. It may perform well at the area level, but may lead to poor estimates for very large areas and for the population size. The total sample size n should be an integer multiple of the total number of areas D. The minimal overall sample size $n = 2D$ allows the unbiased estimation of area-specific sampling variances.

3.2. The Costa allocation

Costa, Satorra, and Ventura (2004) introduce a convex combination

$$n_d^{\text{COS}} = k\frac{N_d}{N}n + (1-k)\frac{n}{D} \tag{27}$$

of proportional and equal allocations, where $0 \le k \le 1$. Value 0 for k yields equal allocation and value 1 yields proportional allocation. The equal allocation at the area level and the proportional allocation at the population level perform satisfactorily. The choice of k depends on the wished qualities of estimates at each level. The design coefficient of variation for the estimator $\hat{Y}_d = N_d \bar{y}_d$ of the area total Y_d according to Eq. (12) is

$$C_d = \mathrm{CV}(\hat{Y}_d) = \frac{1}{Y_d}\left(N_d^2\left(\frac{1}{n_d} - \frac{1}{N_d}\right)S_{y,d}^2\right)^{\frac{1}{2}}, \tag{28}$$

where N_d is the size of the area d counted in statistical units, $S_{y,d}^2$ is the variance of y and Y_d the total of y on the area d, and the sample size n_d is defined according to Eq. (27). The area-specific coefficients of variation C_d depend on the value of k, because the area-specific totals and variances, and the area sizes are fixed.

The optimal value for k minimizes the difference

$$\max(C_d) - \min(C_d);\ d = 1,\ldots, D, \tag{29}$$

subject to the constraints

$$0 \le k \le 1,$$
$$n_d \ge 2, n = \sum_{d=1}^{D} n_d\ . \tag{30}$$

The idea of this solution is to obtain at least moderately accurate area estimates for the areas and for the population size.

We use the area statistics of the proxy variable y^* instead of the unknown variable of interest and Excel Solver with the option "generalized reduced gradient nonlinear". We insert the optimal value of k from Eq. (29) into Eq. (27) to compute the area-specific sample sizes, rounded to the closest integer.

3.3. Allocation using nonlinear programming

The allocation for the design-based direct estimation of area-specific and population means (Choudhry, Rao, and Hidiroglou, 2012) uses nonlinear programming and the area-specific and population coefficients of variation for the variable of interest:

$$\begin{aligned} \mathrm{CV}(\bar{y}_d) &= \frac{1}{\bar{Y}_d}\mathrm{V}(\bar{y}_d)^{\frac{1}{2}}, \\ \mathrm{CV}(\bar{y}) &= \frac{1}{\bar{Y}}\mathrm{V}(\bar{y})^{\frac{1}{2}} \end{aligned}\ . \tag{31}$$

The criterion is the minimization of the overall sample size $n = \sum_{d=1}^{D} n_d$, subject to the fixed upper limits for the coefficients of variation in Eq. (31) and $n_d \ge 2$. This allocation favors areas with a high coefficient of variation, regardless of the area size N_d. Many combinations of upper limits may lead to the same minimum overall sample size. This allocation is also applicable for the total estimators $\hat{Y}_d = N_d \bar{y}_d$ and $\hat{Y} = \sum_{d=1}^{D} N_d \bar{y}_d$, because $\mathrm{CV}(\hat{Y}_d) = \mathrm{CV}(\bar{y}_d)$ and $\mathrm{CV}(\hat{Y}) = \mathrm{CV}(\bar{y})$ under stratified simple random sampling.

Our allocation by nonlinear programming is based on finding the upper limits, which lead to the fixed overall sample size *n*. We use the area and population statistics of the proxy variable y^*, and Excel Solver with the option "generalized reduced gradient nonlinear".

3.4. Allocation using box constraints

Tschuprow (1923) and Neyman (1934) introduced the allocation for minimizing the variance

$$V(\hat{Y}) = \sum_{d=1}^{D} N_d^2 (\frac{1}{n_d} - \frac{1}{N_d}) S_{y,d}^2 \tag{32}$$

for the population total estimator $\hat{Y} = \sum_{d=1}^{D} N_d \bar{y}_d$ under stratified simple random sampling. The minimization of Eq. (32) subject to $n = \sum_{d=1}^{D} n_d$ leads to the Neyman allocation

$$n_d = \frac{N_d S_{y,d}}{\sum_{d=1}^{D} N_d S_{y,d}} n, \tag{33}$$

where the area-specific standard deviations $S_{y,d}$ of the variable of interest or in its absence, of a proxy variable, and the number of units must be available. This allocation favors large areas with high variation and can lead to area sample sizes $n_d < 2$ or even to over-allocation $n_d > N_d$. When $n_d < 2$, the unbiased estimation of the sample variance is impossible. The box-constraint optimal allocation avoids these difficulties, by allowing the control of the sample sizes or of the sampling fractions and the design weights. The allocation minimizes Eq. (32) subject to constraints

$$L_d \le n_d \le U_d, d = 1,\ldots, D$$

$$\sum_{d=1}^{D} n_d \le n, \tag{34}$$

where L_d is the lower limit and U_d is the upper limit for the sample size of domain *d*. The limits are adjusted according to the desired accuracies for the area total estimates, but the choices affect the precision of the population total estimate. The lower limit is $L_d = 2$ and the upper limit $U_d = N_d$. We call this allocation "box-constraint" (BCO). We use an R program (Gabler, Ganninger, and Münnich, 2012) and the R software (http://www.R-project.org) to compute the sample sizes.

Longford (2006) introduces inferential priorities for the areas and for the population. He uses those constraints for deriving sample size allocation schemes for direct, composite, and empirical Bayes estimators. Molefe and Clark´s (2015) reference allocation uses the allocation idea of Longford for a composite estimator, but Longford´s other solutions are not applicable here. Falorsi and Righi (2008) propose a sampling strategy for multi-variate and multi-domain estimation guaranteeing a pre-defined precision for the domain estimators when the overall sample size is small. The point is to collect the sample data by using a multi-stage sampling design based on a balanced sampling technique and on generalized regression. This solution can be extended with indirect small-area estimators, but we cannot apply it because variables of interest are too many.

Table 2: Summary of number-based and parameter-based allocations.

Allocation	Computing sample size for area $d = 1,\ldots, D$	Optimality level
Equal	$n_d^{\mathrm{EQU}} = \frac{n}{D}$	not defined
Costa	$n_d^{\mathrm{COS}} = k\frac{N_d}{N}n + (1-k)\frac{n}{D}$. The constant k is the solution of the minimization problem $\max(C_d) - \min(C_d)$; $d = 1,\ldots, D$, where the coefficient of variation $C_d = \mathrm{CV}(\hat{Y}_d)$ is defined in Eq. (28).	jointly population and area
Nonlinear programming	n_d^{NLP}: minimize $n = \sum_{d=1}^{D} n_d$ subject to limits for coefficients of variation in Eq. (31) $\mathrm{CV}(\bar{y}_d) \le \mathrm{CV}_{0d}$ and $\mathrm{CV}(\bar{y}) \le \mathrm{CV}_0$.	jointly population and area
Box-constraint	n_d^{BCO}: minimize the variance of the population total estimator $V(\hat{Y}) = \sum_{d=1}^{D} N_d^2(\frac{1}{n_d} - \frac{1}{N_d})S_{y,d}^2$ subject to constraints $L_d \le n_d \le U_d$ and $\sum_{d=1}^{D} n_d \le n$. $L_d = 2$ and $U_d = N_d$ here.	population

3.5. Design-based estimation methods for model-free allocations

We apply the three estimation methods to model-free allocations. The design-based Horvitz-Thompson method and the model-assisted generalized regression method use survey weights, which are the inverses of the inclusion probabilities.

The finite population U is composed of D non-overlapping domains or areas, with N_d units in each, and $\sum_{d=1}^{D} N_d = N$. A probability sample is drawn from U for estimating the area totals $Y_d = \sum_{k=1}^{N_d} y_{dk}$, where y_{dk} is the variable of interest for unit k in area d.

The Horvitz-Thompson estimator for the area total Y_d is

$$\hat{Y}_{d,\mathrm{HT}} = \sum_{k=1}^{n_d} w_{dk} y_{dk} = \sum_{k=1}^{n_d} \frac{y_{dk}}{\pi_{dk}}, \tag{35}$$

where n_d is the sample size for area d, π_{dk} is the inclusion probability of unit k in area d, and $w_{dk} = \pi_{dk}^{-1}$ is the sampling weight for the same unit.

The model-assisted generalized regression estimator for the area total Y_d is

$$\hat{Y}_{d,\mathrm{GREG}} = \sum_{k=1}^{N_d} \hat{y}_{dk} + \sum_{k=1}^{n_d} \frac{y_{dk} - \hat{y}_{dk}}{\pi_{dk}}, \tag{36}$$

where the predicted value $\hat{y}_{dk} = x_{dk}'\hat{\beta} + \hat{v}_d$ is based on Eq. (1), and π_{dk} is the inclusion probability (Lehtonen, Särndal, and Veijanen, 2003). The first part of Eq. (36) is the predicted value for Y_d when the assisting model is applied. The predicted values $\hat{y}_{dk}$ can be computed, because the unit-level values of the auxiliary variables x are available. The second term protects against model mis-specification (Lehtonen, Särndal, and Veijanen, 2003).

4. Application: Finnish business register

The estimated parameters are area and population totals of the variable of interest, and the overall sample size n is fixed at 216 individuals.

4.1. Business registers for sampling and allocations

A national Finnish register of block apartments for sale constitutes the data set. This register is maintained by the private company Alma Mediapartners Ltd. Its customers are real estate agencies. They deposit all the appropriate information about the apartments in this register as soon as they receive an assignment from the owners. The population for sample simulations consists of 21,025 sampling units, which are block apartments for sale, selected from the register. In October 2015, they cover 18 Finnish provinces, which are treated as areas. The smallest area contains 160 units and the largest one contains 6,813 units. The variable of interest y measures the price (1,000 €) of the apartment and two auxiliary variables measure the size (in m^2) and age (in years) of the apartment.

All allocations except equal allocation are based on the proxy variable y^*, which is the price of apartment in the register of April 2015. This proxy register contains 22,230 apartments for sale in 18 provinces, and the variables are the same as in the sample population. Table 5 in the Appendix contains the sizes N_d of the areas, population summary statistics for the variable of interest y, and statistics on the differences between y and y^*. The area characteristics of these variables have a wide range. The differences between area sizes, area totals, and area means are mostly negative, in contrast to the differences in area standard deviations and coefficients of variation. This indicates a slight increase in the variation of the prices from April to October 2015.

Table 6 in the Appendix shows the population statistics for the auxiliary variables and correlations between the variables. The between-area variations of the auxiliary variables are very small (1.7% for size and 3.9% for age of total variation, according to a one-way analysis of variance), which means that the allocations cannot be based on the present auxiliary variables. The province of Uusimaa (near capital Helsinki) is a dominating area, because it contains the largest number of apartments (32.4% of the population) and the price level there is the highest among all provinces. The variable of interest has a strong positive correlation with the size of apartment except for one small area, and a negative correlation with the age of apartment except for the largest area. The auxiliary variables are not correlated to one another. The area-specific changes between the correlations (Table 7 in Appendix) are small, except between auxiliary variables for some areas.

Considering the reported changes in the variables between the business registers in April and October 2015, we consider the structures of these registers to be sufficiently similar. This justifies our using the register of April 2015 as the population, which provides the data for computing the allocation-specific sample sizes.

4.2. Allocations

The small overall sample size (n = 216, sampling ratio 1.0%) is a key feature in our procedure. The proxy variable y^* replaces the variable of interest in the model-free allocations using area parameters. The implementation of the Excel Solver with the option “nonlinear generalized reduced gradient” yielded a weight of 0.3528 for k used in the Costa allocation. We use the same Excel option for solving the area sample sizes in the nonlinear programming allocation. The selected limit of 19.01% for the coefficient of variation for areas and the 8.00% limit for

the population size lead to the overall sample size 216. The adjusted homogeneity coefficient of 0.1697 computed with the proxy variable y^* replaces the unknown intra-class correlation in the Molefe and Clark allocation. The low value 0.25 for the constant q and zero for the quantity G in this allocation avoid the concentration of sampling units in a single area (here the province of Uusimaa). The three-term Pareto method allocation is based on simulations and multi-objective optimization. We estimated the unknown variance and covariance parameters in Eq. (7) using the 1,500 simulated simple random samples drawn from the proxy data register, before running the actual allocation-specific simulations. The minimum value of 3.74% for the approximate population coefficient of variation and the minimum value of 22.33% for the mean approximate coefficient of variation over the areas result from the first optimization in Eq. (16). The solution of the optimization criteria in Eq. (17) yields the area sample sizes.

The area sample sizes (Table 3) vary much between the allocations. The largest area, the province of Uusimaa, dominates in two allocations. For the box-constraint allocation, this area contributes for almost 60% of the overall sample size. Four smallest areas have sample size 2, which allow the computation of standard errors for the area total estimates in design-based estimation. The other allocations contain no very small area-specific sample sizes. The structures of the four other allocations have common features. The three-term Pareto method allocation favors the smallest areas and one larger area (the province of Kymenlaakso). It favors less one area (the province of North Karelia). The sample sizes for the Costa allocation are concordant with the area sizes. The nonlinear programming allocation favors areas with a high coefficient of variation, which is characteristic of this allocation.

Table 3: Area sample sizes by allocation.

Area (province)	Size in units	Model-related		Model-free			
		Three-term Pareto method	Molefe and Clark	Equal	Costa	Nonlinear programming	Box-constraint
Uusimaa	6,813	36	55	12	33	36	125
Pirkanmaa	2,003	13	14	12	15	11	13
Varsinais-Suomi	1,543	11	19	12	13	18	14
Päijät-Häme	1,166	9	14	12	12	13	8
Central Finland	1,141	11	8	12	12	9	6
North Ostrobothnia	1,131	9	11	12	12	9	7
Satakunta	1,017	12	11	12	11	15	6
Kymenlaakso	929	14	7	12	11	13	4
Pohjois-Savo	923	10	11	12	11	13	6
Kanta-Häme	885	11	9	12	11	10	5
Etelä-Savo	751	10	9	12	11	10	4
South Karelia	553	11	9	12	10	12	3
North Karelia	549	6	10	12	10	7	4
Lapland	544	11	9	12	10	12	3
Ostrobothnia	421	9	7	12	9	8	2
South Ostrobothnia	311	9	6	12	9	6	2
Kainuu	185	15	3	12	8	8	2
Central Ostrobothnia	160	9	4	12	8	6	2
Total	21,025	216	216	216	216	216	216

4.3. Comparison of the allocations

The results are based on design-based simulation experiments. For each allocation, we simulated $r = 1{,}500$ independent stratified simple random samples and estimated the area totals, variance parameters, mean-squared error approximations, and the allocation-specific quality measures (relative root mean square error and absolute relative bias), using the SAS software (www.sas.com/en_us/home.html) or the IBM SPSS software (www.ibm.com/analytics/data-science/predictive-analysis/spss-statistical-software). We computed design-based Horvitz-Thompson and model-assisted regression estimates for the model-free allocations and model-based EBLUP estimates for every allocation. We compare the allocations, combined with estimators, on the basis of the accuracy and bias, which we measure with the relative root mean square error and absolute relative bias. We compute these quantities, in percent, as sample-based approximations of the expressions in Eq. (5).

The area-specific relative root mean square error and the absolute relative bias in percent are

$$\mathrm{RRMSE}_d = 100 \frac{(\frac{1}{r}\sum_{i=1}^{r}(\hat{Y}_{di} - Y_d)^2)^{\frac{1}{2}}}{Y_d},$$

$$\mathrm{ARB}_d = 100 \left| \frac{1}{r}\sum_{i=1}^{r} \frac{\hat{Y}_{di} - Y_d}{Y_d} \right|, \tag{37}$$

where $\hat{Y}_{di}$ is the design- or the model-based estimate of the area total Y_d for the simulated sample $i = 1,\ldots, r$. Their means over D areas, in percent, are:

$$\mathrm{MRRMSE} = \frac{1}{D}\sum_{d=1}^{D}\mathrm{RRMSE}_d ,$$

$$\mathrm{MARB} = \frac{1}{D}\sum_{d=1}^{D}\mathrm{ARB}_d . \tag{38}$$

The sum $\hat{Y}_i = \sum_{d=1}^{D}\hat{Y}_{di}$ is the estimate for the population total in sample $i = 1,\ldots, r$. The relative root mean square error for the population total, in percent, is

$$\mathrm{RRMSE(pop)} = 100\,\frac{1}{Y}(\frac{1}{r}\sum_{i=1}^{r}(\hat{Y}_i - Y)^2)^{\frac{1}{2}}, \tag{39}$$

where Y is the true value of the population total, and the corresponding absolute relative bias, in percent, is

$$\mathrm{ARB(pop)} = 100 \left| \frac{1}{r}\sum_{i=1}^{r} \frac{\hat{Y}_i - Y}{Y} \right|. \tag{40}$$

We evaluate two measures of quality: the mean over the areas and the mean over the population level. Tables 8 and 9 in the Appendix show the values for these measures at the area and at the population levels.

Figure 2 shows the means of area-specific relative root mean square errors and population relative root mean square errors for each combination of allocation and estimation method. The model-based estimation by EBLUP leads to more accurate area estimates than those obtained from the design-based estimation (Horvitz-Thompson and generalized regression), whatever the three estimation methods applied to whatever of the four model-free allocations. The population values among these allocations are the lowest for the model-assisted regression

estimate. The relative root mean square errors are in stark contrast between the equal and the box-constraint allocations. The equal allocation has the lowest mean over areas (12.3%) and the highest population value (12.2%) for the estimation by EBLUP. The box-constraint allocation performs satisfactorily at the population level, as expected (between 5.0 and 5.6%, depending on the estimation method), but poorly at the area level (mean between 22.3% and 40.6%). The highest mean is obtained for the model-assisted regression estimation, in contrast with other model-free allocations. At the population level, the smallest value is for the Molefe and Clark allocation (5.1%). The allocations provided either by the three-term Pareto method, the Costa method, or by nonlinear programming are good trade-offs, provided the criterion is accurate enough at both the area and at the population levels. No allocation has an optimal accuracy at both levels at the same time. Figure 1 shows the trade-offs for the area and population levels, in the shape of the approximated Pareto front of the bi-objective optimization.

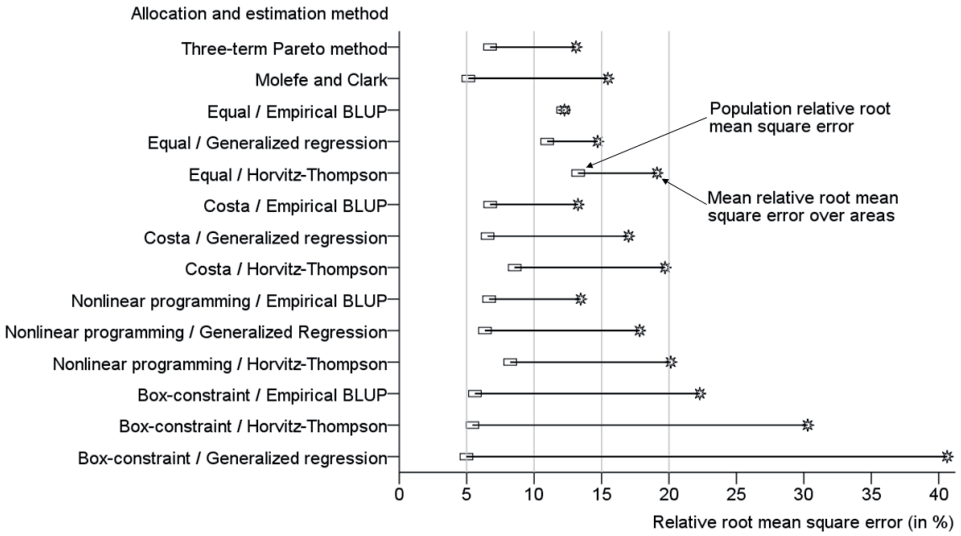

Figure 2: Means of the area-specific relative root mean square errors and of the population relative root mean square errors (in percent) for design- and model-based estimates, by allocation.

On Figure 3, the distributions of the area-specific relative root mean square errors for each allocation show the relative variation of the area total estimates and the presence of randomness in the simulated samples. The model-free allocations are more accurate with model-based estimation. Randomness is the smallest in the three-term Pareto method allocation (lowest median and range of values without outliers). The nonlinear programming allocation has the smallest area as an outlier. The means over the areas of these three allocations are close to each other (Figure 2), although they come from different area-specific distributions. The equal allocation has the lowest median, although a narrow range of variation, and a single outlier (23.4%) for the largest area, the province of Uusimaa. This is a difficulty inherent in this allocation. The area estimates in the box-constraint allocation are the least accurate, regardless of the estimation method. The model-assisted regression estimation is the least accurate.

The EBLUP estimates of the four areas, where the sample size is 2 in the box-constraint allocation, have high relative root mean square errors, excluding the province of Ostrobothnia (14.4%, close to the median). The model-based estimation then can produce at least moderately accurate estimates for a single area, in spite of a small sample size.

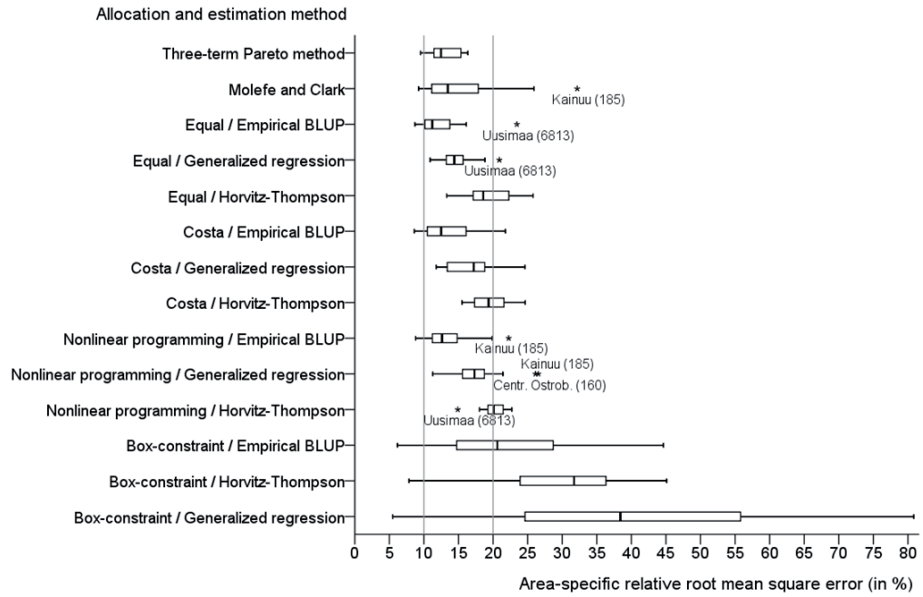

Figure 3: Allocation-specific distributions of area-specific relative root mean square errors (in percent) for design- and model-based estimates.

Table 9 in the Appendix shows the simulation biases for the design-based estimates. As expected, these estimates are almost unbiased. The area-specific biases of the Horvitz-Thompson and of the regression estimates are under 2%, except for three areas in the box-constraint allocation. The area-specific bias distributions for each allocation (Figure 4) demonstrate the similarity between accuracy and bias in the case of the estimation by EBLUP. As for the distributions of the relative root mean square errors, the model-based three-term Pareto method allocation has the narrowest range and is the only allocation with biases under 10%. In the distribution of the equal allocation, the upper quartile is under 4%, but four outliers appear, including the largest area (almost 15%). The distributions of the Costa and of the nonlinear programming allocations are similar, ranging to over 15%. Molefe and Clark´s and the box-constraint allocations are the most dispersed. The contrast between the equal and the box-constraint allocations is similar for the biases and for the relative root mean square errors. The three-term Pareto method, the Costa, and the nonlinear programming allocations with moderately low biases on both levels are satisfactory trade-offs. The population estimate is almost unbiased for Molefe and Clark´s allocation (1.2%), but most of the area estimates are seriously biased, regardless of the sample size. Five areas have important biases for most of the allocations, which indicates that the model is not up the task.

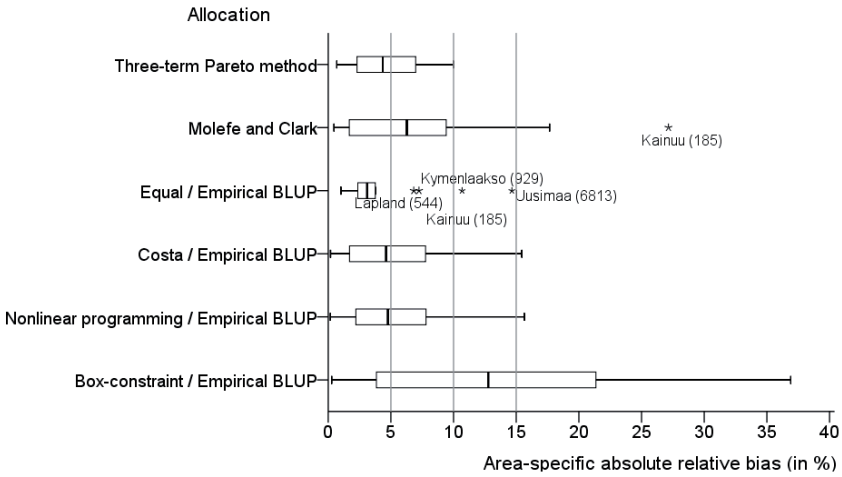

Figure 4: Area-specific absolute relative bias distributions (in %) for model-based empirical best linear unbiased predictor (EBLUP) estimates, by allocation.

Table 4 presents the allocation-specific means over the areas, the population values, and their aggregate values (sums), for the relative root mean square errors and the relative biases. The aggregate relative root mean square errors are the lowest for the EBLUP estimates, except for the equal and the box-constraints allocations. The Horvitz-Thompson estimates are less accurate. The model-assisted regression estimates are more accurate than the Horvitz-Thompson ones, except for the box-constraint allocation, which is high (45.6%). The Horvitz-Thompson and the regression estimates are almost unbiased for the model-free allocations, but the box-constraint allocation is an exception. For the EBLUP estimates, the three-term Pareto method, Molefe and Clark´s, Costa´s, and the nonlinear programming allocations have the smallest aggregate biases, which are close to each other; the box-constraint allocation has the largest aggregate bias.

Table 4: Means over areas, population values, and aggregate values for quality measures (in percent), by allocation. Estimation methods for model-free allocations: 1=Horvitz-Thompson, 2=regression estimation, and 3=empirical best linear unbiased predictor.

	Model-related		Model-free											
	Three-term Pareto	Molefe and Clark	Equal			Costa			Nonlinear programming			Box-constraint		
			1	2	3	1	2	3	1	2	3	2	1	3
	Relative root mean square error													
Mean over areas	13.1	15.5	19.1	14.7	12.3	19.7	17.0	13.3	20.1	17.8	13.5	40.6	30.3	22.3
Population value	6.7	5.1	13.3	11.0	12.2	8.6	6.6	6.8	8.2	6.4	6.7	5.0	5.4	5.6
Sum	19.8	20.6	32.4	25.7	24.4	28.3	23.6	20.0	28.4	24.2	20.1	45.6	35.7	27.9
	Absolute relative bias													
Mean over areas	4.9	7.8	0.4	0.5	4.2	0.5	0.5	5.3	0.3	0.6	5.5	1.3	0.8	13.8
Population value	3.4	1.2	0.3	1.0	7.3	0.4	0.4	3.0	0.2	0.8	3.2	0.3	0.2	2.2
Sum	8.3	9.1	0.7	1.5	11.5	0.8	0.9	8.3	0.5	1.4	8.7	1.6	1.0	16.0
	Integrated accuracy and bias													
Overall sum	28.1	29.7	33.0	27.1	35.9	29.1	24.5	28.3	28.8	25.6	28.8	47.2	36.7	43.9

We evaluate the allocations by integrating the aggregate values for the relative root mean square error and the absolute relative bias. The model-assisted regression estimates of Costa´s, of the nonlinear programming, and of the equal allocations have the smallest values (24.5%, 25.6%, and 27.1%). The three-term Pareto method allocation has the second smallest value (28.1%), which includes a high aggregate bias. The aggregate values indicate that the model-assisted regression estimation performs the best for the three model-free allocations, although not supported by the area-specific relative root mean square errors (Table 8 in Appendix).

The box-constraint and the equal allocations are extreme, in the sense that they are strongly or not at all associated with the area sizes. These solutions lead to satisfactory estimates only at one level, either population or area. The three-term Pareto method, Costa´s, and the nonlinear programming allocations take both the between-area and the within-area variations into account. They perform well at both levels, when the model is included. The three-term Pareto method and Costa´s allocations do not use fixed priorities or limits for the area-level and the population-level estimation, unlike the nonlinear programming and Molefe and Clark´s allocations.

For small areas, the model-based estimation produces area estimates of moderate accuracy, despite a small sample size (provinces of North Karelia and Ostrobothnia). Large sample sizes, however, do not guarantee high accuracy (provinces of Satakunta, Kymenlaakso, and Kainuu). The accuracy of the area estimates seems to be related to the area-specific means and to the coefficients of variation of the variables. Large deviations from the corresponding population statistics may bias the estimation of the area totals. The skewness of the variable of interest usually confuses the EBLUP estimation, as the important biases for some areas indicate.

We examined the validity of the unit-level linear mixed model in Eq. (1) by testing the null hypothesis that the error terms v_d and e_{dk} are normally distributed. We computed the transformed residuals $(y_{dk} - \hat{\tau}_d \bar{y}_d) - (x_{dk} - \hat{\tau}_d \bar{x}_d)'\beta$, where $\hat{\tau}_d = 1 - (1 - \hat{\gamma}_d)^{\frac{1}{2}}$ and the factor $\hat{\gamma}_d$ is defined in Eq. (8) (Rao and Molina, 2015). Under the null hypothesis, the residuals are approximately identically and independently distributed as $N(0, \sigma_e^2)$. We applied the test to a simple random sample, of n = 5,000 individuals, selected from the population. We took σ_v^2 = 1,570 and σ_e^2 = 17,550. The Shapiro-Wilks test yielded a p-value of 0.00, leading us to reject the null hypothesis. We also computed the allocation-specific means for the variance parameters, and the regression coefficients and the area effects of the area total estimator in Eq. (4), for the simulated samples. The means for Molefe and Clark´s and the box-constraint allocations differ from those for the other allocations. Our model has deficiencies when its parameters are estimated by generalized least-squares or by restricted maximum likelihood. It is possible, before the estimation phase, to make the distribution of the variable of interest more symmetric by an algebraic transformation such as the lognormal method, but we have not done that.

5. Conclusion

We compared six allocation methods in stratified sampling, when applying model-based estimation and design-based estimation for obtaining area and population estimates. The fixed and small total sample size is a common restriction. Our three-term Pareto method allocation uses auxiliary information, the model, and an estimation method. Accuracy at both the area and at the population levels are optimized, which requires multi-objective optimization techniques. We chose the reference allocations on the basis of the variety of information, which the allocations use: model and estimator, optimization criteria, fixed limits or priorities, and

auxiliary information. The allocation-specific area sample sizes are various. The sample is concentrated on the largest area for two allocations, a situation which may lead to inaccurate and biased estimates for small areas.

We computed the area sample sizes for five allocations using the previous register data, because the auxiliary variables are insufficient to support the allocations. The distance between apartment and the town center has a predictive power, but it is not available.

We applied design- and model-based estimations and evaluated the allocations in terms of accuracy and bias obtained from design-based sample simulations. We confirm that, in this survey framework, the model-based estimates are more accurate than the design-based estimates. The "borrowing strength" principle may be significant in surveys where some areas have too small sample sizes to allow direct estimates of satisfactory quality. The model-free allocations have similar performance structures at different levels, regardless of the estimation method.

The studied allocations have all pros and cons, depending on the estimation level (area and population). Considering the aggregate values, the EBLUP estimates for the three-term Pareto method, the Costa, and the nonlinear programming allocations are most accurate. The randomness associated with the area estimates is best controlled in the three-term Pareto method allocation, from the viewpoint of the area-specific distributions of relative root mean square errors.

The bias results for the EBLUP estimates demonstrate that the allocations have very different performances. The three-term Pareto method and the Costa allocations perform better, with respect to aggregate values and area-specific distributions.

By considering accuracy and bias, we showed that the Costa, the nonlinear programming, and the equal allocations under model-assisted regression estimation perform the best, and that the three-term Pareto method allocation performs very close. This comes from the fact that the design-based estimates are almost unbiased, but that many of these estimates are inaccurate. The model-based estimation suffers from an important bias, leading to try methods likely to improve accuracy and reduce bias. The applicable software is also necessary.

Getting a well-performing allocation is not an easy task; it is very case-specific and depends on the objectives of a survey and on the availability of auxiliary information. Accurate estimates, both at the area and at the population levels, are made obtainable by multi-objective optimization. The model and the estimation method have become part of the sampling design.

The first trade-off is between the quality of the area estimates and the quality of population estimates. We showed the impossibility of obtaining maximum quality at both levels simultaneously. The fixed priorities or limits at the area and at the population levels, which some allocations use, do not guarantee the maximum quality.

The second trade-off is between accuracy and bias of the estimates. Model-based estimators are usually more accurate than design-based estimators when the sample size is small, but model-based estimators may be importantly biased. The sample allocation affects accuracy and bias, but the increment of the area sample size does not correct the bias entirely. This trade-off appears commonly in the literature, but the discussion has seldom concerned the priorities of these measures.

Acknowledgements

The authors thank the two referees as well as Professor Risto Lehtonen for constructive comments and suggestions.

References

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988). An error component Model for Prediction of County Crop Areas using Survey and Satellite Date. *Journal of the American Statistical Association 83*: 28-36.

Burgard, J.P., Münnich, R., and Zimmermann, T. (2014). The impact of sampling designs on small area estimates for business data. *Journal of Official Statistics 30*(4): 749–771.

Choudhry, G.H., Rao, J.N.K., and Hidiroglou, M.A. (2012). On sample allocation for effective domain estimation. *Survey Methodology 38*: 23–29.

Costa, A., Satorra, A., and Ventura, E. (2004). Improving both domain and total area estimation by composition. *SORT 28*(1): 69–86.

Falorsi, P.D. and Righi, P. (2008). A balanced sampling approach for multi-way stratification for small area estimation. *Survey Methodology 34*: 223–234.

Gabler, S., Ganninger, M., and Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika 75*: 15–161.

Keto, M. and Pahkinen, E. (2010). On sample allocation for effective EBLUP estimation of small area totals – "Experimental Allocation", in *Survey Sampling Methods in Economic and Social Research*, J. Wywial and W. Gamrot (eds). Katowice: Katowice University of Economics, 27–36.

Lehtonen R., Särndal C.E., and Veijanen A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology 29*: 33–44.

Longford, N.T. (2006). Sample Size Calculation for Small-Area Estimation. *Survey Methodology 32*: 87–96.

Meza, J.L. and Lahiri, P. (2005). A note on the C_p statistic under the nested error regression model. *Survey Methodology 31*(1): 105-109.

Miettinen, K. (1999). *Nonlinear Multiobjective Optimization*. Kluwer Academic Publishers, Boston.

Molefe, W. B. and Clark, R. G. (2015). Model-assisted optimal allocation for planned domain using composite estimation. *Survey Methodology 41*(2): 377–387.

Neyman, J. (1934). On the two different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society 97*: 558-625. DOI: http://dx.doi.org/10.2307/2342192.

Nissinen, K. (2009). Small Area Estimation With Linear Mixed Models From Unit-Level Panel and Rotating Panel Data. Ph.D. thesis, Department of Mathematics and Statistics, University of Jyväskylä, Report 117. DOI: https://jyx.jyu.fi/dspace/handle/123456789/21312.

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

Rao, J. N. K. and Molina, I. (2015). *Small Area Estimation* (2nd Edition). Hoboken, NJ: John Wiley & Sons, Inc.

Tschuprow, A. A. (1923). On the mathematical expectation of the moments of frequency distributions in the case of correlated observations. *Metron 2*: *461-493, 646-683.*

Appendix

Table 5: Population statistics of the variable of interest y (price) in October 2015 business register and the changes between y and proxy variable y^* (price in April 2015 business register).

	Variable of interest y (price)				Difference y - y^*			
Area (province)	Size in units	Total	Mean	Coefficient of variation	Size in units	Total	Mean	Coefficient of variation
Uusimaa	6,813	2,067,530	303.5	0.89	-636	-236,839	-5.88	0.01
Pirkanmaa	2,003	311,634	155.6	0.69	-118	-20,429	-0.98	0.06
Varsinais-Suomi	1,543	248,763	161.2	0.90	-109	-14,826	1.66	0.09
Päijät-Häme	1,166	174,104	149.3	0.72	63	3,589	-5.27	0.03
Central Finland	1,141	153,693	134.7	0.60	-78	-11,410	-0.74	0.04
North Ostrobothnia	1,131	180,849	159.9	0.61	-169	-35,020	-6.15	0.06
Satakunta	1,017	111,409	109.5	0.78	55	-6,862	-13.40	0.02
Kymenlaakso	929	91,405	98.4	0.68	93	5,866	-3.93	-0.01
Pohjois-Savo	923	114,935	124.5	0.81	-86	-23,056	-12.24	0.11
Kanta-Häme	885	106,110	119.9	0.62	130	7,692	-10.46	0.01
Etelä-Savo	751	89,736	119.5	0.69	-74	-19,417	-12.82	0.08
South Karelia	553	64,087	115.9	0.64	72	2,709	-11.71	-0.03
North Karelia	549	96,688	176.1	0.59	-76	-19,685	-10.08	0.07
Lapland	544	61,867	113.7	0.78	-105	-21,816	-15.22	0.12
Ostrobothnia	421	58,584	139.2	0.56	-102	-16,411	-4.24	0.03
South Ostrobothnia	311	41,822	134.5	0.50	-35	-9,944	-15.14	0.02
Kainuu	185	15,791	85.4	0.62	-31	-5,439	-12.93	0.06
Central Ostrobothnia	160	22,403	140.0	0.50	1	-1,153	-8.13	0.04
Population	21,025	4,011,408	190.8	1.00	-1,205	-422,451	-8.66	0.13

Table 6: Population summary statistics of the auxiliary variables "size" (m^2) and "age" (years) and correlations between variables in the business register in October 2015.

	Auxiliary variable x_1 (size)			Auxiliary variable x_2 (age)			Correlations		
Area (province) and size in units	Total	Mean	Coefficient of variation	Total	Mean	Coefficient of variation	Price, size	Price, age	Size, age
Uusimaa (6,813)	481,026	70.6	0.41	227,623	33.4	0.90	0.73	0.03	-0.01
Pirkanmaa (2,003)	130,232	65.0	0.37	59,354	29.6	0.85	0.65	-0.17	0.13
Varsinais-Suomi (1,543)	106,871	69.3	0.41	52,196	33.8	0.66	0.57	-0.31	0.14
Päijät-Häme (1,166)	77,040	66.1	0.36	35,962	30.8	0.73	0.58	-0.46	0.03
Central Finland (1,141)	72,908	63.9	0.31	29,438	25.8	0.87	0.43	-0.65	0.03
North Ostrobothnia (1,131)	73,978	65.4	0.35	20,549	18.2	1.21	0.63	-0.43	0.08
Satakunta (1,017)	65,924	64.8	0.31	41,189	40.5	0.60	0.50	-0.16	0.06
Kymenlaakso (929)	58,788	63.3	0.38	35,892	38.6	0.60	0.46	-0.51	0.17
Pohjois-Savo (923)	60,985	66.1	0.40	34,057	36.9	0.52	0.54	-0.47	-0.04
Kanta-Häme (885)	55,949	63.2	0.38	31,023	35.1	0.62	0.50	-0.52	-0.01
Etelä-Savo (751)	46,865	62.4	0.33	25,547	34.0	0.61	0.42	-0.52	-0.01
South Karelia (553)	34,235	61.9	0.29	18,709	33.8	0.63	0.46	-0.54	0.05
North Karelia (549)	34,005	61.9	0.31	11,090	20.2	1.08	0.47	-0.68	0.03
Lapland (544)	35,156	64.6	0.39	17,396	32.0	0.67	0.53	-0.57	0.03
Ostrobothnia (421)	25,915	61.6	0.42	13,925	33.1	0.86	0.51	-0.25	0.18
South Ostrob. (311)	20,093	64.6	0.37	7,986	25.7	0.86	0.22	-0.66	0.25
Kainuu (185)	10,886	58.8	0.35	6,724	36.3	0.44	0.47	-0.59	-0.03
Central Ostrob. (160)	12,013	75.1	0.54	6,463	40.4	0.65	0.58	-0.15	0.29
Population (21,025)	1,402,870	66.7	0.39	675,123	32.1	0.81	0.59	-0.10	0.04

Table 7: Changes in the auxiliary variables and in correlations between October 2015 and April 2015 (´*´denotes auxiliary variables in the proxy register April 2015).

	Changes x_1-x_1^* in size			Changes x_2-x_2^* in age			Correlation changes		
Area (province) and size in units	Total	Mean	Coefficient of variation	Total	Mean	Coefficient of variation	Price, size	Price, age	Size, age
Uusimaa (6,813)	-46,084	-0.16	0.01	1,726	3.08	-0.10	0.00	-0.03	-0.07
Pirkanmaa (2,003)	-6,154	0.72	-0.00	1,916	2.55	-0.08	0.04	0.07	-0.01
Varsinais-Suomi (1,543)	-4,632	1.76	0.04	-412	1.98	-0.07	-0.01	0.08	0.07
Päijät-Häme (1,166)	2,567	-1.45	0.01	2,158	0.19	-0.02	0.02	0.07	0.04
Central Finland (1,141)	-2,566	1.98	0.03	233	1.84	-0.05	0.00	0.03	0.04
North Ostrob. (1,131)	-7,082	3.06	-0.02	2,365	4.18	-0.26	0.02	-0.03	-0.02
Satakunta (1,017)	2,752	-0.85	-0.03	5,391	3.29	-0.11	0.04	0.11	-0.00
Kymenlaakso (929)	6,606	0.86	-0.00	3,538	-0.06	-0.03	0.01	0.04	0.04
Pohjois-Savo (923)	-5,640	0.04	0.05	2,452	5.58	-0.20	-0.01	0.09	-0.01
Kanta-Häme (885)	6,754	-1.94	0.02	6,091	2.03	-0.05	-0.03	0.04	0.02
Etelä-Savo (751)	-3,232	1.67	0.04	1,638	5.04	-0.18	0.05	0.01	-0.08
South Karelia (553)	3,453	-2.09	-0.01	3,398	2.00	-0.04	-0.06	0.14	0.17
North Karelia (549)	-4,025	1.09	-0.00	888	3.88	-0.24	0.02	-0.05	-0.05
Lapland (544)	-6,000	1.21	0.04	2,294	8.71	-0.29	0.05	0.07	-0.09
Ostrobothnia (421)	-5,547	1.40	-0.00	904	8.18	-0.22	-0.04	-0.02	-0.11
South Ostrob. (311)	-1,555	2.04	0.00	1,347	6.49	-0.29	-0.04	-0.02	-0.02
Kainuu (185)	-2,189	-1.69	0.01	-252	4.05	-0.15	0.09	0.10	-0.11
Central Ostrob. (160)	415	2.13	-0.02	902	5.41	-0.13	0.07	0.17	0.07
Population (21,025)	-72,160	0.37	0.01	36,577	3.39	-0.12	-0.00	-0.01	-0.04

Table 8: Relative root mean square errors (in percent) for areas and population, by allocation. Estimation methods for model-free allocations: 1=Horvitz-Thompson, 2=regression estimation, 3=empirical best linear unbiased predictor (EBLUP).

Area (province) and size in units	Model-related		Model-free											
	Three-term Pareto	Molefe and Clark	Equal			Costa			Nonlinear programming			Box-constraint		
			1	2	3	1	2	3	1	2	3	1	2	3
Uusimaa (6,813)	12.6	10.0	25.4	20.9	23.4	15.5	11.8	12.9	14.9	11.3	12.9	7.9	5.5	6.2
Pirkanmaa (2,003)	10.4	9.7	19.6	14.7	11.0	17.6	13.3	9.9	21.2	15.6	10.5	19.2	17.9	11.9
Varsinais-Suomi (1,543)	14.2	11.8	25.8	18.1	13.8	24.7	18.0	13.2	21.5	15.8	12.4	23.9	21.3	15.6
Päijät-Häme (1,166)	11.1	10.4	20.4	14.0	10.4	20.2	14.9	10.5	19.7	15.3	11.0	25.3	24.6	14.7
Central Finland (1,141)	10.2	12.3	17.3	12.0	9.5	17.3	13.4	10.0	20.2	16.0	11.1	23.8	29.6	16.8
North Ostrob. (1,131)	9.5	9.2	18.0	11.5	8.7	17.3	12.0	8.6	19.9	13.7	8.8	23.3	23.2	12.5
Satakunta (1,017)	16.4	17.9	22.3	18.8	14.7	22.9	20.9	16.1	19.9	18.2	14.8	31.0	35.7	28.7
Kymenlaakso (929)	15.9	23.7	19.1	14.7	13.5	20.7	18.8	17.1	18.9	18.5	16.7	32.4	55.8	38.2
Pohjois-Savo (923)	14.5	16.2	22.5	16.9	12.9	23.8	19.3	14.0	22.7	17.7	13.9	33.8	38.5	25.4
Kanta-Häme (885)	12.2	13.8	17.2	13.3	10.1	18.8	16.2	12.0	19.1	16.9	12.1	27.3	38.4	21.5
Etelä-Savo (751)	12.9	14.1	18.9	15.3	11.7	20.9	18.1	13.4	21.2	18.7	13.0	34.3	40.5	20.8
South Karelia (553)	11.5	13.2	18.2	13.2	10.5	19.6	15.9	11.8	18.1	15.5	11.6	36.3	44.3	20.5
North Karelia (549)	11.7	11.1	17.0	10.9	9.1	18.1	13.3	10.1	21.6	16.0	11.2	29.6	28.3	16.5
Lapland (544)	15.4	19.7	22.6	15.7	13.8	24.0	18.7	16.1	22.5	18.3	15.0	45.1	55.2	32.0
Ostrobothnia (421)	12.4	12.5	15.8	14.1	10.6	19.1	18.1	11.6	19.3	19.3	11.8	38.0	57.2	14.4
South Ostrob. (311)	12.3	14.9	13.6	11.7	9.4	15.8	16.4	12.0	20.3	21.4	13.4	36.6	61.5	21.5
Kainuu (185)	16.3	32.2	17.1	15.2	16.1	21.6	24.6	21.8	21.6	26.1	22.3	43.7	80.9	39.4
Central Ostrob. (160)	16.4	25.9	13.3	13.8	11.4	16.9	22.3	17.3	19.8	26.5	19.9	33.8	73.0	44.7
Mean over areas	13.1	15.5	19.1	14.7	12.3	19.7	17.0	13.3	20.1	17.8	13.5	30.3	40.6	22.3
Population value	6.7	5.1	13.3	11.0	12.2	8.6	6.6	6.8	8.2	6.4	6.7	5.4	5.0	5.6

Table 9: Absolute relative biases (in percent) for areas and population, by allocation. Estimation methods for model-free allocations: 1=Horvitz-Thompson, 2=regression estimation, 3= empirical best linear unbiased predictor (EBLUP).

Area (province) and size in units	Model-related		Model-free											
	Three-term	Molefe and	Equal			Costa			Nonlinear programming			Box-constraint		
	Pareto	Clark	1	2	3	1	2	3	1	2	3	2	1	3
Uusimaa (6,813)	7.9	5.9	0.5	1.6	14.7	0.9	0.7	7.8	0.4	1.6	8.2	1.0	0.6	3.4
Pirkanmaa (2,003)	1.9	1.1	0.3	0.4	2.4	0.5	0.2	1.4	0.1	0.0	1.7	0.2	0.4	0.3
Varsinais-Suomi (1,543)	2.2	0.5	0.3	1.1	3.5	0.1	0.9	1.7	0.2	0.1	0.9	0.2	0.1	3.8
Päijät-Häme (1,166)	0.7	1.0	0.5	0.6	1.3	0.1	0.1	0.2	0.2	0.4	0.2	0.3	0.1	4.3
Central Finland (1,141)	3.7	5.8	0.3	0.1	2.9	0.7	0.4	3.4	0.3	0.4	4.6	0.2	0.3	7.5
North Ostrob. (1,131)	0.7	1.4	0.1	0.1	1.0	0.0	0.3	1.0	0.2	0.4	1.5	0.5	0.8	1.6
Satakunta (1,017)	5.6	9.4	0.3	1.0	3.3	0.4	0.6	6.4	0.5	1.0	5.1	0.1	0.4	21.4
Kymenlaakso (929)	9.6	17.7	0.6	0.8	7.2	0.3	0.4	11.1	0.4	0.1	9.8	1.4	0.7	30.8
Pohjois-Savo (923)	3.9	6.5	0.4	0.6	2.7	0.4	0.4	4.2	0.5	0.7	4.7	0.5	1.0	15.8
Kanta-Häme (885)	4.3	6.7	0.2	0.0	2.6	0.1	0.4	4.5	0.3	0.7	4.8	0.1	0.3	12.6
Etelä-Savo (751)	4.4	5.7	0.4	0.3	2.5	0.6	1.0	4.6	0.1	0.4	4.6	3.5	0.4	12.9
South Karelia (553)	4.3	6.1	0.2	0.1	3.4	0.2	0.2	5.1	0.1	0.1	4.4	1.4	1.5	13.5
North Karelia (549)	6.8	6.5	0.3	0.1	3.6	0.3	0.3	4.7	0.2	0.5	6.1	0.2	0.4	11.8
Lapland (544)	8.5	13.2	0.4	0.6	6.8	0.7	0.5	9.7	0.2	1.0	7.8	1.0	1.0	25.3
Ostrobothnia (421)	2.3	1.7	0.1	0.2	1.9	1.3	0.6	1.5	0.7	0.0	2.2	1.4	0.8	1.7
South Ostrob. (311)	4.9	7.8	0.4	0.4	3.8	0.7	0.7	5.5	0.0	0.6	6.5	3.5	2.5	13.1
Kainuu (185)	10.0	27.1	0.8	0.2	10.7	0.6	0.0	15.5	0.9	0.2	15.6	0.9	1.5	32.5
Central Ostrob. (160)	7.0	16.9	0.5	0.0	2.0	0.3	1.3	7.8	0.2	3.0	10.2	6.6	1.7	36.9
Mean over areas	4.9	7.8	0.4	0.5	4.2	0.5	0.5	5.3	0.3	0.6	5.5	1.3	0.8	13.8
Population value	3.4	1.2	0.3	1.0	7.3	0.4	0.4	3.0	0.2	0.8	3.2	0.3	0.2	2.2