Author(s): Keto, Mauno; Hakanen, Jussi; Pahkinen, Erkki

Title: Register data in sample allocations for small-area estimation

Year: 2018

Version: Accepted version (Final draft)

Please cite the original version:

Keto, M., Hakanen, J., & Pahkinen, E. (2018). Register data in sample allocations for small-area estimation. Mathematical Population Studies, 25(4), 184-214. 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):

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

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

$$g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = (N_d - n_d)^2 (1 - \hat{\gamma}_d)\hat{\sigma}_v^2 ,$$

$$g_{2d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = (N_d - n_d)^2 (\bar{x}_d^{\ *} - \hat{\gamma}_d \bar{x}_d)'(X'\hat{V}^{-1}X)^{-1}(\bar{x}_d^{\ *} - \hat{\gamma}_d \bar{x}_d) ,$$

$$\begin{aligned} g_{3d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = {} & (N_d - n_d)^2 (n_d)^{-2} (\hat{\sigma}_v^2 + \hat{\sigma}_e^2 (n_d)^{-1})^{-3} (\hat{\sigma}_e^4 \text{V}(\hat{\sigma}_v^2) + \hat{\sigma}_v^4 \text{V}(\hat{\sigma}_e^2) \\ & -2\hat{\sigma}_e^2 \hat{\sigma}_v^2 \text{Cov}(\hat{\sigma}_e^2, \hat{\sigma}_v^2)) \end{aligned} ,$$

$$g_{4d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = (N_d - n_d)\hat{\sigma}_e^2 \,. \tag{7}$$

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):

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

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)

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

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

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

$$\text{CV}(\hat{Y}_d) = \frac{\text{V}(N_d \bar{y}_d)^{\frac{1}{2}}}{Y_d} ,$$

$$\text{CV}(\hat{Y}) = \frac{(\sum_d \text{V}(N_d \bar{y}_d))^{\frac{1}{2}}}{Y} . \qquad (12)$$

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 $\hat{Y}_{d,\text{Eblup}}$ and the population total estimate $\hat{Y}_{\text{Eblup}}$ are:

$$\text{ACV}(\hat{Y}_{d,\text{ Eblup}}) = \frac{\text{amse}(\hat{Y}_{d,\text{ Eblup}})^{\frac{1}{2}}}{Y_d} ,$$

$$\text{ACV}(\hat{Y}_{\text{Eblup}}) = \frac{\text{amse}(\hat{Y}_{\text{Eblup}})^{\frac{1}{2}}}{Y} , \qquad (13)$$

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:

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

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

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

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

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.

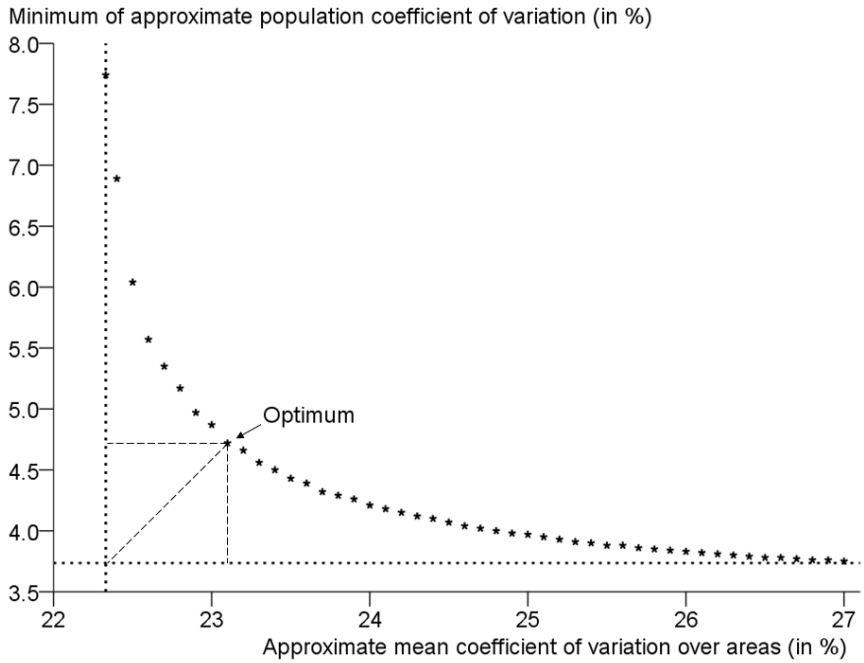

Figure 1: The approximated Pareto front minimizing the mean of approximate coefficients of variation over areas and of the approximate population coefficient of variation. The label “Optimum” denotes the Pareto optimal solution.

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, …, 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

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

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

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

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:

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

$$\approx \sum_{d=1}^{D} N_d^q \sigma_d^2 \rho(1-\rho)(1+(n_d-1)\rho)^{-1} + G N_+^{(q)} \sum_{d=1}^{D} \sigma_d^2 P_d^2 n_d^{-1}(1-\rho). \tag{23}$$

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

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

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,…, 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

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

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

where $\hat{Y}_{di}$ is the design- or the model-based estimate of the area total Y_d for the simulated sample i = 1,…, 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,…, r. The relative root mean square error for the population total, in percent, is

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

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.

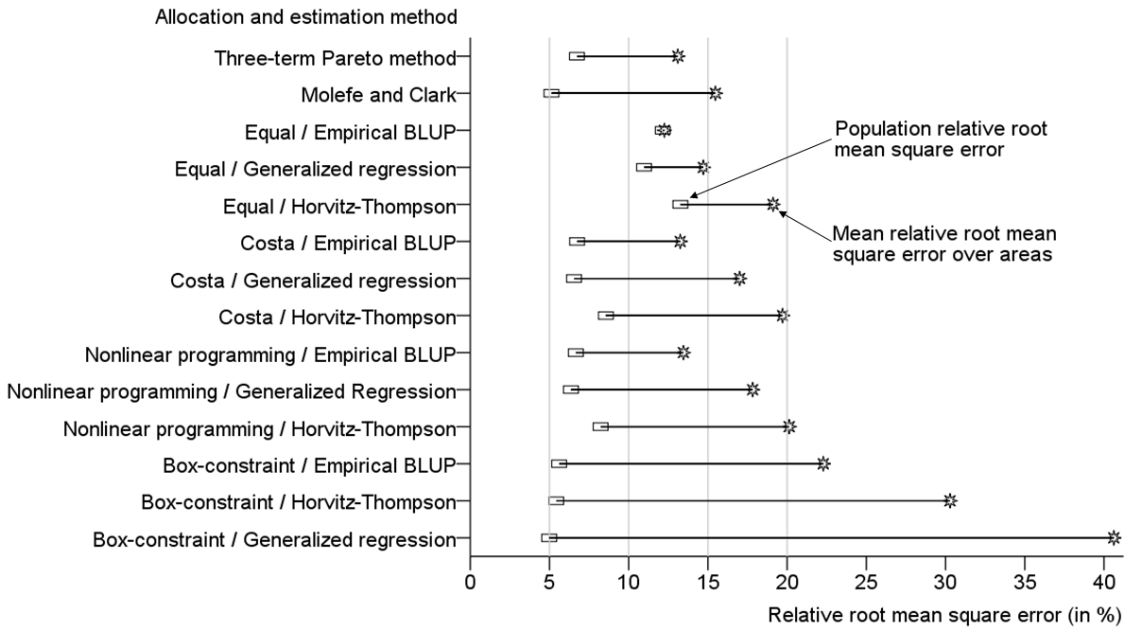

Figure 2: Means of the area-specific relative root mean square errors and of the population relative root mean square errors (in percent) for design- and model-based estimates, by allocation.

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.

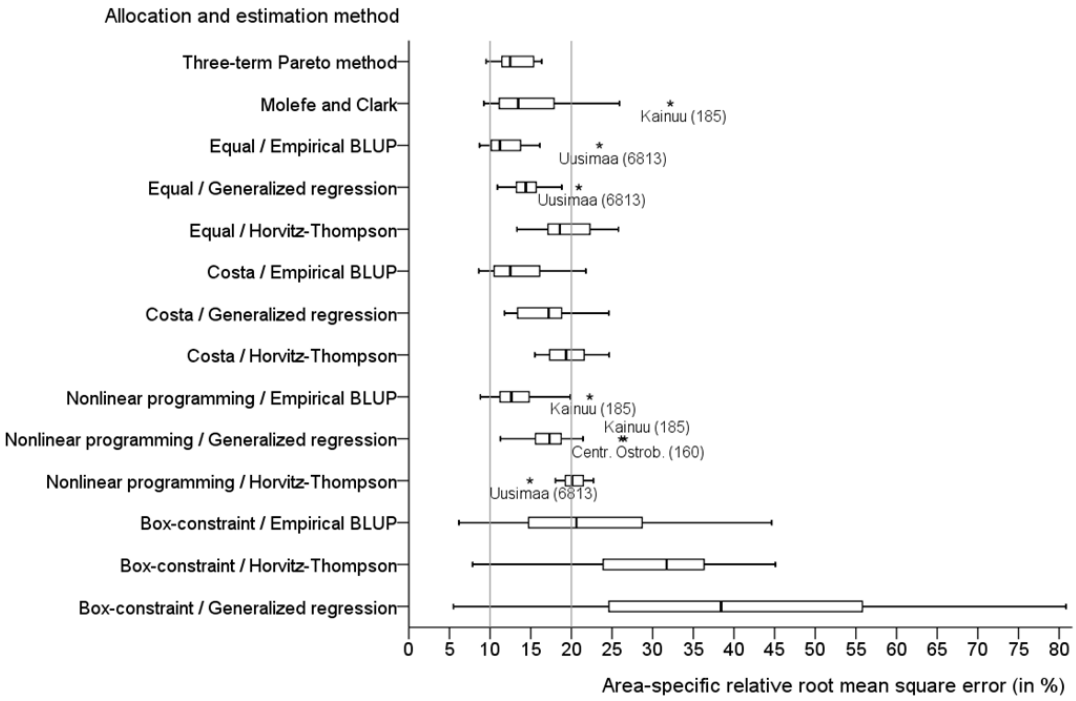

Figure 3: Allocation-specific distributions of area-specific relative root mean square errors (in percent) for design- and model-based estimates.

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.

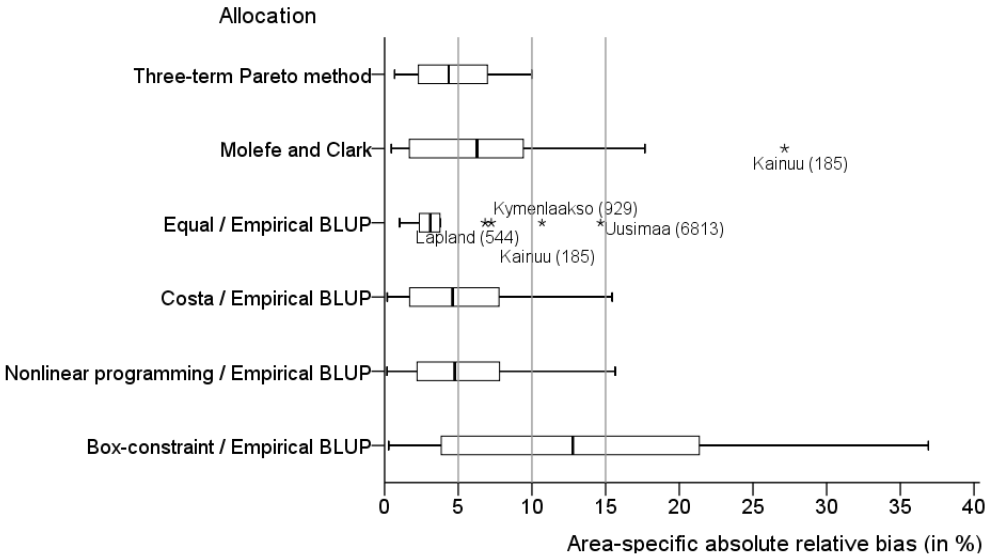

Figure 4: Area-specific absolute relative bias distributions (in %) for model-based empirical best linear unbiased predictor (EBLUP) estimates, by allocation.

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 Pareto	Molefe and Clark	Equal			Costa			Nonlinear programming			Box-constraint		
			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