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

Title: On overall sampling plan for small area estimation

Year: 2017

Version:

Please cite the original version:

Keto, M., & Pahkinen, E. (2017). On overall sampling plan for small area estimation. Statistical Journal of the IAOS, 33(3), 727-740. https://doi.org/10.3233/SJI-170370

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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*:

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

This estimator is a combination of two estimators: the synthetic estimator $\hat{\bar{Y}}_{d(syn)} = \hat{\boldsymbol{\beta}}'\,\overline{\boldsymbol{X}}_d$, where $\hat{\boldsymbol{\beta}}$ is the estimated regression coefficient and $\overline{\boldsymbol{X}}_d$ is the area population means of auxiliary variables ***x***, and a direct estimator $\bar{y}_{dr} = \bar{y}_d + \hat{\boldsymbol{\beta}}'(\bar{\boldsymbol{x}}_d - \overline{\boldsymbol{X}}_d)$, where $\bar{y}_d$ and $\bar{\boldsymbol{x}}_d$ are the area *d* sample means of *y* and ***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

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

where $v_{d(syn)}$ is the sampling variance of the synthetic estimator $\hat{\bar{Y}}_{d(syn)}$ and $B_d = \boldsymbol{\beta}_U'\,\overline{\boldsymbol{X}}_d - \bar{Y}_d$ 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,…, *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 ***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

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

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

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

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

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

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

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

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

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 **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

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

and ARB% for area d is defined as

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

and their means over all D areas are computed as

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

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

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

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

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

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. (3nd 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*	$n_d^{\text{CAL}-g1} = \dfrac{\tilde{N}_{g1,d}\, n - (N - \tilde{N}_{g1,d}\ D - n)(1/R^2_{a,y^*} - 1)}{N + D(1/R^2_{a,y^*} - 1)}$, where $\tilde{N}_{g1,d} = SD(y^*)_d \,/\, ASD(y^*) N / D.$	Jointly area and population
MC- *q025*	$n_d^{\text{MC}} = \dfrac{n\sigma_d\ N_d^{q/2}}{\sum_{d=1}^{D}\sigma_d\ N_d^{q/2}} + \dfrac{1-\rho}{\rho}\left(\dfrac{\sigma_d\ N_d^{q/2}}{D^{-1}\sum_{d=1}^{D}\sigma_d\ N_d^{q/2}} - 1\right)$, *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”.

	Proxy data		Popu-	Model-based		Model-free			
Area (province)	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-	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	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

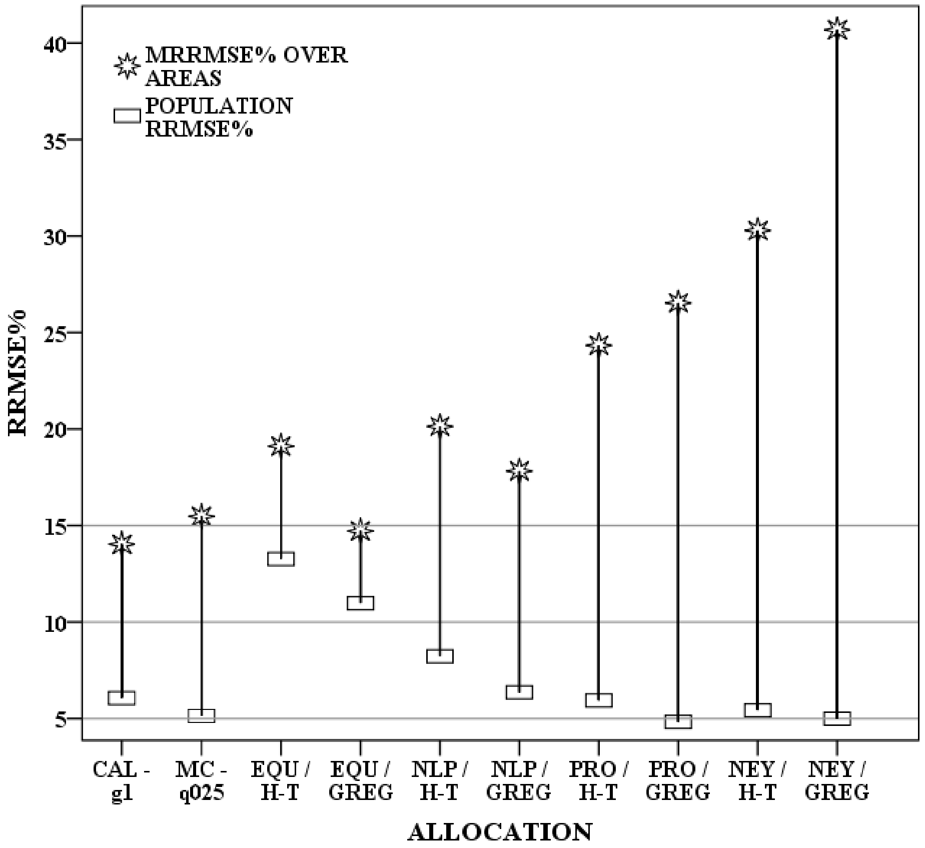

Figure 1. Means of area RRMSE$_d$ %s (MRRMSE%) and population RRMSE%s by allocation and estimation method. EBLUP estimation is applied to CAL-*g1* and MC-*q025* allocations.

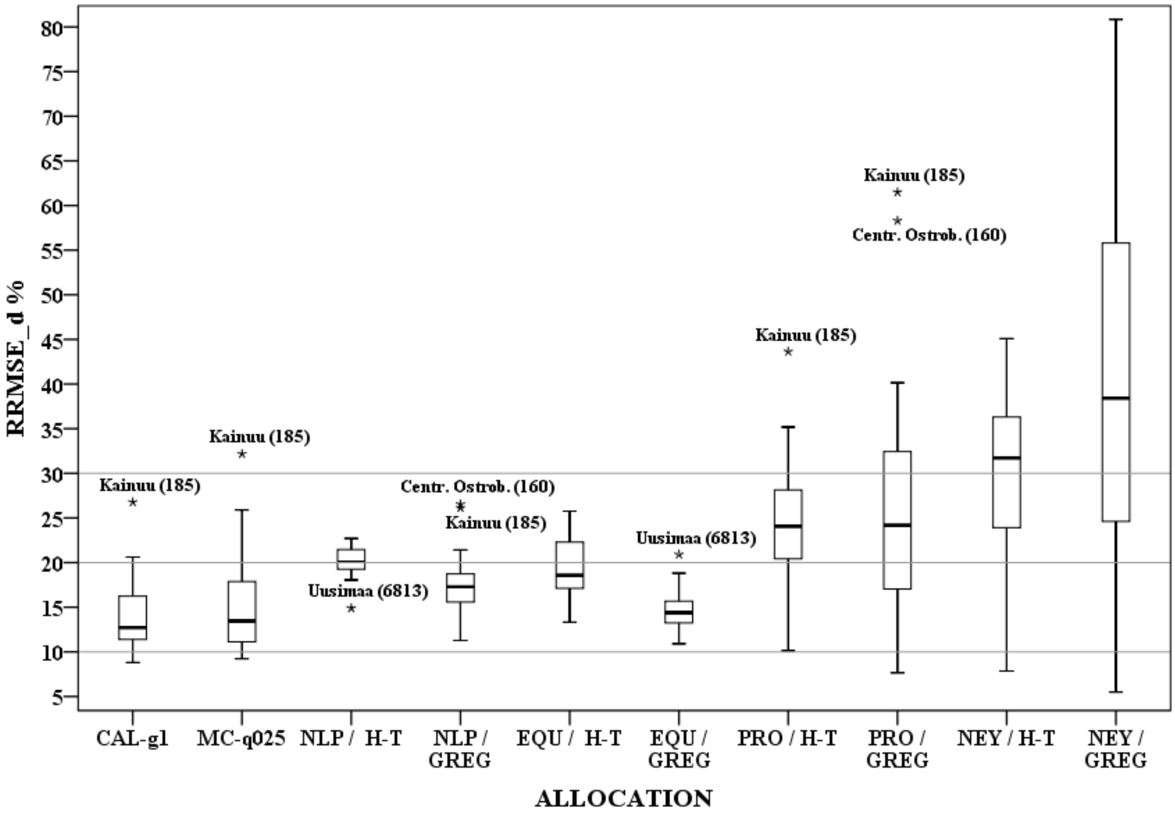

Figure 2. Distributions of area-specific RRMSE$_d$ %s by allocation and estimation method.