Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Sample allocation for efficient model-based small area estimation

SURVEY
METHODOLOGY

Catalogue No. 12-001-XPB

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

Statistics Canada Statistique Canada

Canada

by Mauno Keto and Erkki Pahkinen

Release date: June 22, 2017

Statistics Canada Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

email at STATCAN.infostats-infostats.STATCAN@canada.ca

telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following toll-free numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-877-287-4369

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Standard table symbols

The following symbols are used in Statistics Canada publications:

.	not available for any reference period
..	not available for a specific reference period
...	not applicable
0	true zero or a value rounded to zero
0^{s}	value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p}	preliminary
^{r}	revised
x	suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E}	use with caution
F	too unreliable to be published
*	significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

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

Sample allocation for efficient model-based small area estimation

Mauno Keto and Erkki Pahkinen[1]

Abstract

We present research results on sample allocations for efficient model-based small area estimation in cases where the areas of interest coincide with the strata. Although model-assisted and model-based estimation methods are common in the production of small area statistics, utilization of the underlying model and estimation method are rarely included in the sample area allocation scheme. Therefore, we have developed a new model-based allocation named $g1$-allocation. For comparison, one recently developed model-assisted allocation is presented. These two allocations are based on an adjusted measure of homogeneity which is computed using an auxiliary variable and is an approximation of the intra-class correlation within areas. Five model-free area allocation solutions presented in the past are selected from the literature as reference allocations. Equal and proportional allocations need the number of areas and area-specific numbers of basic statistical units. The Neyman, Bankier and NLP (Non-Linear Programming) allocation need values for the study variable concerning area level parameters such as standard deviation, coefficient of variation or totals. In general, allocation methods can be classified according to the optimization criteria and use of auxiliary data. Statistical properties of the various methods are assessed through sample simulation experiments using real population register data. It can be concluded from simulation results that inclusion of the model and estimation method into the allocation method improves estimation results.

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

1 Introduction

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

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

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

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

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

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

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

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

2 Allocations which utilize the model

2.1 Choosing the model

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2.2 Model-based $g1-$allocation

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

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

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

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

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

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

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

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

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

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

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

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

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

2.3 Model-assisted MC-allocation

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

3 Some model-free area allocations

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

3.1 Number-based allocations

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

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

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

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

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

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

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

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

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

3.2 Parameter-based allocations

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

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

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

Choudhry, Rao and Hidiroglou (2012) present the NLP allocation method for direct estimation. This method uses non-linear programming to find a solution. Criteria for the allocation are defined by setting upper limits for CV values of the study variable y in each area and in the population. In practice y^* or x replaces y. The program searches the minimum sample size $n = \sum_d n_d$ satisfying these conditions. The SAS (Statistical Analysis System) procedure NLP with Newton-Raphson option was used to find the solution. The allocation favors areas with high CV regardless of the area size N_d.

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

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

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

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

4 Comparison of performances of allocations

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

4.1 Empirical data

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

Table 4.1
Population summary statistics

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

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

4.2 Allocations

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

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

Table 4.2
Area sample sizes by allocation

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

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

4.3 Comparison of performances of allocations

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

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

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

and their means over areas are computed as follows:

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

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

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

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

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

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

Table 4.3
Area and population RRMSE%s by allocation

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

The lowest RRMSE% mean over the areas (15.65%) was obtained in the $g1-$allocation developed in this study. Helsinki was an exception on area level because its RRMSE% value was clearly higher compared with model-assisted and parameter-based allocations. Also equal and proportional allocations performed well on area level, with means 16.48% and 16.52%. The highest means were obtained in the model-assisted MC-allocations. On the population level, the lowest value for the quality measure was obtained in the model-assisted MCG50-allocation (5.88%) and the second lowest value in the Bankier allocation (5.89%), but in general, differences between the allocations on this level were small.

Table 4.4
Area and population ARB%s by allocation

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

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

5 Concluding remarks

This research was focused on seven different allocation solutions which were categorized into three groups according to the auxiliary data needed in their implementation. The least amount of auxiliary information is needed in equal and proportional allocation which are based on the number of areas and the number of statistical units in each area. The Neyman, Bankier and NLP allocations are based on pre-set optimization criteria, and application of these methods presumes area-specific parameter information such as the standard deviation or CV of the study variable, and in the Bankier allocation the area totals of at least one auxiliary variable must be known. Because the study variable is unknown, it must be replaced with a suitable proxy or auxiliary variable to enable the use of these three methods. A common feature of the number-based and parameter-based allocations is that they are not based on any model, whereas the other three allocations utilize the underlying model, in addition to number-based information.

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

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

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

Acknowledgements

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

References

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

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

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

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

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

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

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

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

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

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

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

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