On sample allocation for effective EBLUP estimation of small area totals

Mauno Keto[1]

[1]Mikkeli University of Applied Sciences - Finland, e-mail: mauno.keto@mamk.fi

Abstract

The demand of regional or small area statistics produced from large-scale surveys has raised needs for developing the tools of optimal sample allocation on area level. The concept of optimality can of course be defined in many different ways. Most commonly used allocation methods aim at producing efficient direct areal estimates. What often happens, however, is that due to sparse sampling resources several areas receive little or none observations, and therefore indirect estimation may be necessary. These methods are often based on nested-error regression type model-based estimators. For this reason should areal sample allocation be implemented in such a way that it would lead to efficient estimation in the case of an indirect estimator. The problem has been tried to solve in this research by developing an analytical allocation method based on the main component of MSE in EBLUP estimation. The performance of this method has been tested by comparing it to various other allocations through sample simulations from real data. The effectiveness of each allocation was measured with MSE, CV and certain quality measures (ARE, ARB and RRMSE). Some results are presented here.

Key words: Planning samples sizes of small areas, indirect estimation, optimal allocation, optimization criterion, areas with none observations.

1 Introduction

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

Optimal allocation has inspired for different solutions during the last decades. Main line has prevailed to find areal allocation giving possibility to calculate direct or model-assisted direct estimators for each area. Some examples from earlier efforts are reported in Rao (2003). Recently published interesting proposition come from Longford (2006), who includes inferential priority index P_d for each area and tries then to find optimality. Another solution comes from Falorsi and Righi (2008). They assume that direct estimators should be model-assisted and their optimal allocation procedure accounts for this possibility with other prior information used in planning sample design.

The next sections describe different approaches to areal allocation, the selected model as additional information, earlier experimental studies, developing optimal sample sizes in one simple situation, the searching of an areal allocation scheme conditional to auxiliary information which includes both auxiliary variables and model for indirect estimation of fixed areal totals, and finally some results of simulations based on different allocations. Indirect or model-based estimation has been chosen because in small area calculations domains with few or none observations are general. The problem of choise of model has been profoundly investigated by Lehtonen *et al* (2003 and 2006). As a model, EBLUP has been chosen because there is a lot of evidence that this model works well in many small area estimation situations.

2 Brief overview of sample allocation methods in stratified sampling

2.1 Allocation methods developed for population and area level

Equal allocation is based only on the number of areas and doesn´t take the characteristics of areas into account at all. The sample size of each area is simply *n/D*, where *n* = overall sample size and *D* = number of areas. Especially large areas with strong variation in variables of interest suffer from the point of view of efficiency and accuracy.

Proportional allocation can be used when larger areas are expected to have higher variance compared with smaller areas. The sample size of area *d* is proportional to the number of observation units (N_d) in that area:

$$n_{d,pro} = f_d n = (N_d / N)n .$$

This allocation ensures same proportion for each area in the sample, but does not guarantee efficient estimation, especially for areas in which the response variable has high variance and significantly higher values compared with smaller areas.

Neyman´s allocation which is a special case of optimal allocation is based on sizes of areas and the variances (or standard deviations) of auxiliary variable *x* used in estimation. The *x*-value of each observation unit in the population or at least its variance in each area must be known. Sample size if area *d* is computed as follows:

$$n_{d,opt} = (N_d S_d / \sum_{d=1}^{D} N_d S_d)n ,$$

where S_d = standard deviation of auxiliary variable *x* in area *d*. This allocation favours large areas with high variance. Differences in sample sizes between areas can be significantly large compared with for example proportional allocation.

Power (Bankier) allocation is based more on internal characteristics of areas compared with previously mentioned methods. It is recommended to be used in a research where the population contains many small areas and reliable estimates must be produced for each area. Formula for ample size for area *d* is

$$n_{d,pow} = (X_d^a CV(x)_d) \sum_{d=1}^{D} X_d^a CV(x)_d)n ,$$

where X_d means the sum of values of auxiliary variable *(x)* and $CV(x)_d$ is the coefficient of variation of variable *x* in area *d*. Exponent *a* is a certain power value which can be used to regulate the significance of variable *x*. Values ½ and ⅓ are often recommended for *a*.

2.2 Approaches to optimal allocation based on certain assumptions and criteria

Some articles concerning optimal sample allocation in stratified sampling have been published during last years. Two of them are shortly referred in this paper.

Longford (2006) has introduced a method to calculate optimal sample sizes for areas when minimizing the weighted sum of sampling variances in direct estimation as a function of sample sizes n_d. Minimum can be found only under simple assumptions. The weights are called inferential priorities and their values can be changed and thus areas can be given different significance.

Falorsi and Righi (2008) have used a sampling strategy to determine sample sizes for domains that is based on balanced sampling when domains (areas) have first been divided into non-overlapping partitions in many different ways. The main goal was to produce sample sizes which guarantee the sampling errors of domain estimates to be lower than given limits. Optimization of inclusion probabilities is a crucial part of the method. Many different estimators were tested through simulations. The authors finally ended up to recommend a GREG-type model-assisted estimator.

3 Model and earlier experiments related to use of model

3.1 Model

The model used in this research is a nested-error regression, basic unit level model

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

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

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

where $\mathbf{y}$ is n×1 vector of sample observations, $\mathbf{X}$ and $\mathbf{Z}$ are known n×p and n×h matrices of full rank, and $\mathbf{v}$ and $\mathbf{e}$ are independently distributed with means $\mathbf{0}$ and covariance matrices $\mathbf{G}$ and $\mathbf{R}$ depending on some variance parameters $\boldsymbol{\delta} = (\delta_1,...,\delta_q)'$. Furthermore, Var($\mathbf{y}$) = $\mathbf{V} = \mathbf{R} + \mathbf{ZGZ}'$ is the variance-covariance matrix of $\mathbf{y}$. In model (1) y_{dk} is the kth value in area d for outcome variable (y), $\mathbf{x}_{dk}$ is the vector of auxiliary variables (x) in area d, v_d is the latent random effect of area d (d = 1,…,D) in the model and is estimated from the observations, and e_{dk} is a random error. Random effects v_d and random errors e_{dk} are assumed to be independent of each other and distributed with mean zero and variances σ_v^2 and σ_e^2 (not necessarily normally). Regression coefficients $\boldsymbol{\beta}$ are estimated from the observations.

Regression coefficients $\boldsymbol{\beta}$ and area effects $\mathbf{v}$ are estimated from the observations as follows:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \qquad \hat{\mathbf{v}} = \mathbf{GZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \tag{3}$$

EBLUP (Empirical Best Linear Unbiased Predictor) estimator for area total Y_d is the sum of sample observations and predicted values of non-sampled observations of variable y as given in Rao (2003):

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

The MSE of estimator $\hat{Y}_{d,EBLUP}$ is the sum of its variance and squared bias:

$$MSE(\hat{Y}_{d,Eblup}) = E(\hat{Y}_{d,Eblup} - Y_d)^2 = Var(\hat{Y}_{d,Eblup}) + (\hat{Y}_{d,Eblup} - Y_d)^2 \quad (5)$$

An estimator of MSE approximation in the case of finite populations is given in Rao (2003):

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

The first and most important component which is important later in this presentation is given by

$$g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = (N_d - n_d)^2(1-\gamma_d)\hat{\sigma}_v^2, \text{ where } \gamma_d = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_e^2 n_d^{-1}) = \hat{\sigma}_v^2/(n_d\hat{\sigma}_v^2 + \hat{\sigma}_e^2). \quad (7)$$

In addition, we define a specific common intrastratum or intra-area correlation

$$\hat{\rho} = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_e^2) = 1/(1 + \hat{\sigma}_e^2/\hat{\sigma}_v^2), \quad (8)$$

which measures the proportion of variation between areas and total variation (value between zero and one).

The model (1) is used as additional information when searching for optimal allocation in an analytical way which is based on the structure of model and its MSE.

3.2 Experimental allocation as the first approach under the model

The first approach to allocation problem was “Experimental allocation” which Keto and Pahkinen (2009) have introduced in a conference paper. The idea was shortly following: 1500 SRSWOR-samples were drawn from a population of 400 Finnish municipalities in 19 provinces which served also for areas and strata. Response variable (*y*) was number of unemployed people and one auxiliary variable was used. Total number of unemployed in each province (area) was estimated by using model (1) and EBLUP estimation. In the first phase MSE, CV and certain quality measures (ARE, ARB etc.) were produced for each area in every sample,and finally their means were computed in every sample. After this, samples were arranged in ascending order according to means. The following figure shows an example of MSE means. Sample sizes for the second phase were determined by using quartiles of these distributions. In the second phase the competence of this “experimental” allocation was compared with three other allocation methods.

Figure 1: Distribution of areal sample sizes in 20 “best” samples for MSE means (boxplot).

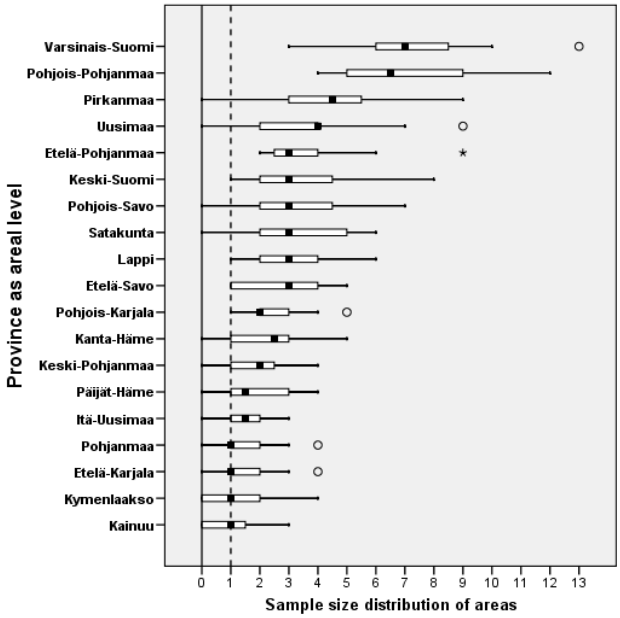


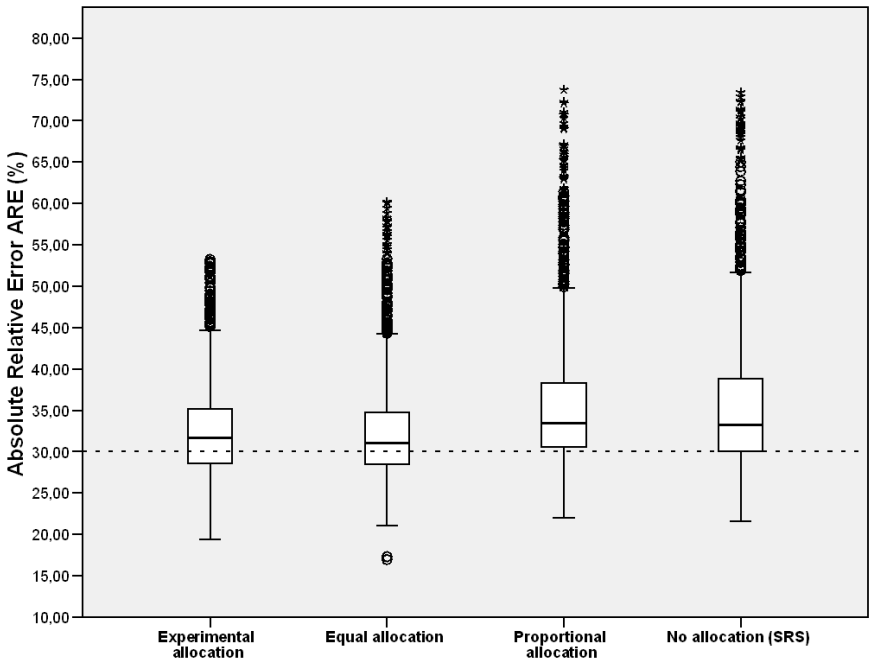
Table 1 shows final sample sizes of areas in each compared allocation. It is worth noticing that in experimental allocation as many as 7 areas have sample size zero.

Table 1: Areal sample sizes in different allocation schemes

Province	**Size of area**	**Not allocated**	**Proportional**	**Equal**	**Experimental**
Uusimaa	24		4	3	4
Varsinais-Suomi	53		7	3	7
Itä-Uusimaa	10		1	3	0
Satakunta	25	S	4	3	5
Kanta-Häme	16	R	2	3	0
Pirkanmaa	28	S	4	3	5
Päijät-Häme	12	-	2	3	0
Kymenlaakso	12	s	2	3	0
Etelä-Karjala	12	a	2	3	0
Etelä-Savo	18	m	3	3	3
Pohjois-Savo	23	p	3	3	4
Pohjois-Karjala	16	l	2	3	3
Keski-Suomi	28	e	4	3	5
Etelä-Pohjanmaa	26	s	4	3	6
Pohjanmaa	17		2	3	4
Keski-Pohjanmaa	12		2	3	0
Pohjois-Pohjanmaa	38		5	3	6
Kainuu	9		1	3	0
Lappi	21		3	3	5
TOTAL:	**400**	**57**	**57**	**57**	**57**

1 500 SRSWOR samples were simulated from population by using each of these allocations (4 times 1500 samples) and same statistics (MSE, CV) and quality measures (ARE, ARB etc.) were computed. Figure 2 shows the distributions of average absolute relative error (ARE, see appendix) in the samples.

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



Of course experimental allocation method cannot be used to prove the better performance of tested allocation, but it can show the topics to focus on in future research concerning effective sample allocation.

4 Example of finding optimal sample sizes in a simple situation

Let us assume that for each observation unit in the population there exists value x_{dk} of auxiliary variable *x*, and this value is correlated with value y_{dk} of response variable *y*, which means that following equation holds between these values: $y_{dk} = f(x_{dk}) + e_{dk}$, where e_{dk} means additive residuals. The variance of response variable can be computed with variance of explanatory variable. If variances are finite, we get equation

$$V(y_{dk}) = E(V(y_{dk}|x_{dk})) + V(E(y_{dk}|x_{dk})). \tag{9}$$

Let us consider a unit-level linear regression model

$$y_{dk} = \alpha + \beta x_{dk} + e_{dk}, e_{dk} \sim N(0,\sigma^2),$$

where α and β are least-squares regression coefficients estimated from the sample. By using rule (9) we get an estimator for variance of y_{dk} as follows: $\hat{V}(y_{dk}) = \beta^2 V(x_{dk}) + \sigma^2$. The estimator of the mean of response variable *y* in area *d* is given by expression

$$\hat{V}(\bar{y}_d) = (1 - n_d/N_d)(\beta^2 V(x_{dk}) + \sigma^2)/n_d = (1/n_d - 1/N_d)(\beta^2 V(x_{dk}) + \sigma^2)$$

that contains finite population correction. The mean of areal variances can be written as

$$(1/D)\sum_{d=1}^{D}(1/n_d - 1/N_d)(\beta^2 V(x_{dk}) + \sigma^2) = (1/D)\sum_{d=1}^{D}(1/n_d)(\beta^2 V(x_{dk}) + \sigma^2) - (1/D)\sum_{d=1}^{D}(1/N_d)(\beta^2 V(x_{dk}) + \sigma^2).$$

This expression can be minimized as a function of sample sizes $n_1,...,n_D$ by using a well-known Lagrange´s method under constraint $\sum_d n_d = n$. The result for optimal sample size is (derivation can be proved)

$$n_d = (\sqrt{1+(\beta^2/\sigma^2)V(x_{dk})} / \sum_d \sqrt{1+(\beta^2/\sigma^2)V(x_{dk})}) \times n \ .$$

If quantity $(\beta^2/\sigma^2)V(x_{dk})$ is large enough, value one is negligible, so we get an approximate value for n_d :

$$n_d \approx (\sqrt{(\beta^2/\sigma^2)V(x_{dk})} / \sum_d \sqrt{(\beta^2/\sigma^2)V(x_{dk})}) \times n = ((\beta/\sigma)S(x_{dk}) / (\beta/\sigma)\sum_d S(x_{dk})) \times n$$
$$= (S(x_{dk}) / \sum_d S(x_{dk})) \times n \ .$$

Optimal sample size is approximately proportional to areal standard deviation of covariate *x*. The same result can be obtained also without finite population correction. But the result was obtained through minimization of mean of variances. Some areal variances may remain considerably high. The purpose of this example is to show that optimality or at least approximate optimality can be reached under sufficient simple assumptions.

5 Searching for optimal allocation analytically under selected model

Let us return to model (1). It contains fixed part (regression) and a part containing random area effects. A random sample of *n* sampling units is selected from *D* areas, and sampling method is SRSWOR inside strata. Furthermore, $\sum_d n_d = n$. EBLUP estimation produces normally estimates $\hat{Y}_d$ for areal totals of response variable *(y)* according to (4) and MSE approximations (6). Also CV´s of areal estimates can be computed.

First we try to select the sample from strata so that we can minimize the arithmetic mean of areal MSE´s as a function of sample sizes $n_1, n_2,...,n_D$. But because MSE has a very complex expression, the minimization is not possible. We turn attention to the first part of MSE, g_{1d} (7). According to Nissinen (2009), this term has contributed 85 – 95 % of total MSE in many surveys, but this requires sufficient variation between areas.

Criterion for optimal allocation is now minimizing mean of areal g_{1d} values

$$1/D\sum_{d=1}^{D} g_{1d}(\hat{\sigma}_v^2, \hat{\sigma}_e^2) = 1/D\sum_{d=1}^{D}(N_d - n_d)^2(1/\hat{\sigma}_e^2 \times n_d + 1/\hat{\sigma}_v^2)^{-1} \ . \qquad (10)$$

as a function of sample sizes subject to constraint $\sum_d n_d = n$. We use method of Lagrange multipliers to solve each n_d. Derivation process is not shown here because of its length, but the result can be proved. The expression for sample size of area *d* is

$$n_{d,opt} = -\hat{\sigma}_e^2/\hat{\sigma}_v^2 + \frac{(N_d + \hat{\sigma}_e^2/\hat{\sigma}_v^2)(n + D(\hat{\sigma}_e^2/\hat{\sigma}_v^2))}{N + D(\hat{\sigma}_e^2/\hat{\sigma}_v^2)} \ . \qquad (11)$$

This expression contains ratio $\delta = \hat{\sigma}_e^2/\hat{\sigma}_v^2$ of variance components which depends on the sample. We have

earlier defined a specific intra-area correlation $\hat{\rho}$ in (8), and ratio δ is given now by expression

$$\delta = 1/\hat{\rho} - 1 . \tag{12}$$

Expression (10) can now be written in the form

$$n_d = -\delta + \frac{(N_d + \delta)(n + \delta D)}{N + \delta D} = \frac{N_d n - (N - N_d D - n)\delta}{N + D\delta} = \frac{N_d n - (N - N_d D - n)(1/\hat{\rho} - 1)}{N + D(1/\hat{\rho} - 1)}. \tag{13}$$

Some conclusions can be made immediately when examining expressions (11) or (13): 1) expression is meaningful when variance component $\hat{\sigma}_v^2 > 0$, 2) because $n + \delta D < N + \delta D$ it follows that $n_d < N_d$, 3) value of n_d depends on the ratio of variance components but not directly on the values of variance components, and 4) if all total variation consists only of variation between areas ($\delta = 0$), final result would be proportional allocation. More precise examination reveals that sample size n_d can become negative in certain situations: area *d* is small, overall sample size *n* is small and total variation is mostly within areas.

Because ratio $\delta = \hat{\sigma}_e^2 / \hat{\sigma}_v^2$ cannot be used to compute sample sizes, we have to replace it with a corresponding value that can be obtained from auxiliary variable. Because response variable *y* and auxiliary variable *x* are correlated, we assume that the variation of auxiliary variable transfers to the sample. We use so called homogeneity measure which is related to cluster sampling and which is presented for example by Särndal *et al* (1992). If the clusters have different sizes then homogeneity measure is given by expression

$$R_a^2 = 1 - R^2 = 1 - \frac{MSW}{S^2}, \tag{14}$$

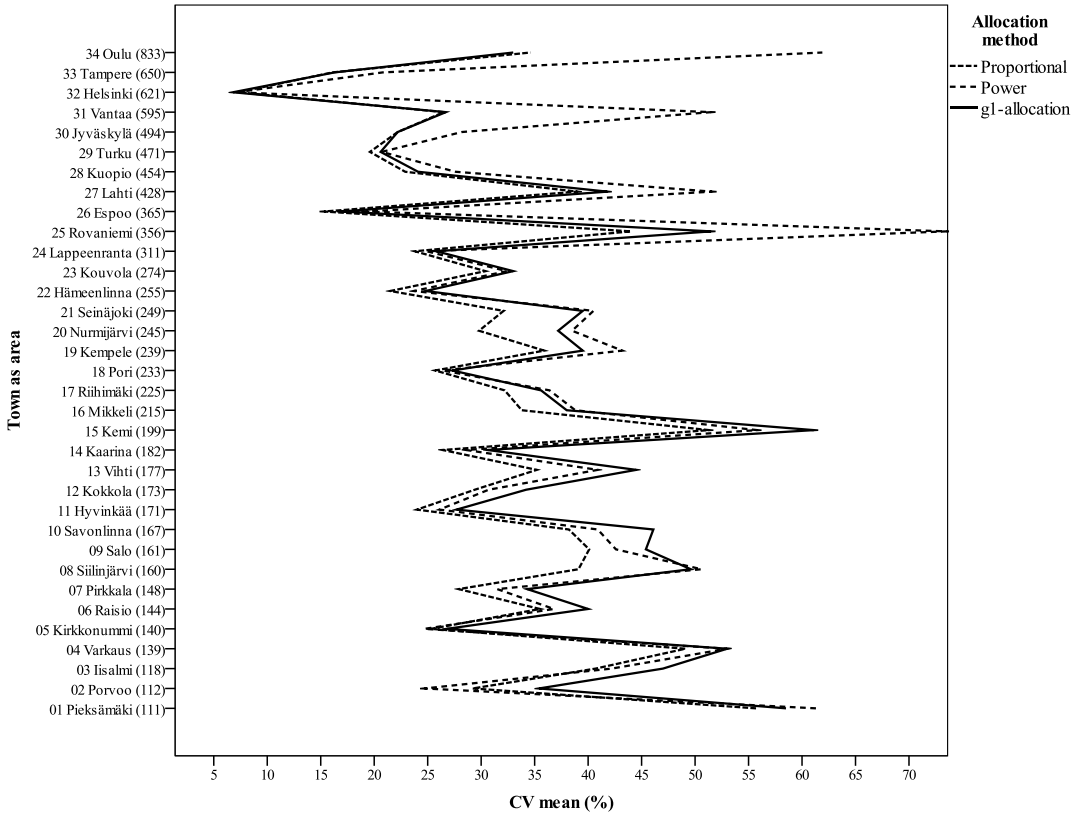
where R^2 means coefficient of determination (familiar from regression analysis), MSW is mean sum of squares within clusters (in this case strata), and S^2 is the variance of auxiliary variable. When substituting homogeneity measure (14) for intra-area correlation ($\hat{\rho}$) in (12) the sample sizes for areas can be calculated.

This method was applied to a case in which the population consists of 9 815 apartments for sale in 34 Finnish towns. The data were collected from an internet source in 2011.The size of smallest area was 111 (apartments, sampling units) and size of largest area was 833. Variable *(y)* measures the price of apartment (1 000 €) and auxiliary variable *(x)* measures size (m^2). Variables are correlated. Overall sample size *(n)* was very low 102 (3 times 34). Value of homogeneity measure for auxiliary variable *(x)* for computing sample sizes is 0.33 which means that variation between areas was high. Some compromises had to be made (for ex. negative sample sizes were turned to zero). Final areal sample sizes varied from zero (three areas) to 12.

To test the performance of developed method (we call it "g1-allocation") it was compared with five other allocations: SRSWOR, equal, proportional, optimal (Neyman) and power allocation. 1 500 samples were simulated for each method (6 times 1 500 samples). Sampling method was SRSWOR inside strata, except for first where allocation was not used. MSE and CV plus certain quality measures to discover accuracy and bias were calculated for areas in each sample, as well as the areal means of these statistics and measures.

Figure 3 presents areal CV means for calculated of 1500 samples for three different allocations (proportional, power and g1-allocation) which take areal characteristics into account. Optimal (Neyman), equal and SRSWOR alternatives are not presented. One can notice that g1-allocation had good performance in large areas, but fairly poor performance on some smaller areas. But what must be mentioned is the fact that three smallest areas were non-sampled areas in g1-allocation. Presented results bring hopefully new aspects to consider in model-based small area estimation. More investigation of the influence of areas is needed.

Figure 3: Areal CV means computed from 1 500 samples (sizes of areas inside brackets)



Appendix: formulas

Coefficient of variation (CV) for estimate of area total Y_d in EBLUP estimation:

$$CV(\hat{Y}_{d,EBLUP})\% = 100 \times (\sqrt{mse(\hat{Y}_{d,EBLUP})} / \hat{Y}_{d,EBLUP})$$

Average absolute relative error (ARE) in one sample for estimate of area total Y_d in EBLUP estimation:

$$ARE\ \% = 100 \times (1/D) \sum_{d=1}^{D} \left| \hat{Y}_{d,EBLUP} - Y_d \right| / Y_d$$

where D = number of estimated areas.

References

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

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

Khan, M.G.M., Maiti, T. and Ahsan, M.J. (2010). An Optimal Multivariate Stratified Sampling Design

Using Auxiliary Information: An Integer Solution Using Goal Programming Approach. *Journal of Official Statistics* **26,** 695-708.

Lehtonen, R., Myrskylä, M., Särndal, C.-E. and Veijanen, A. (2006). The role of models in model-assisted and model-dependent estimation for domains and small areas. *Working paper, BNU Workshop,* Ventspils, Latvia, August 2006.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2003). The Effect of Model Choice in Estimation for Domains, Including Small Domains. *Survey Methodology* **29,** 33-44.

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

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

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

Särndal, C-E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Verlag.