Mauno Keto (Lic. Phil.) and Erkki Pahkinen (Professor emeritus)
Jyväskylä University, Finland

MODEL-BASED OPTIMAL SAMPLE ALLOCATION FOR PLANNED AREAS USING EBLUP ESTIMATION

Presentation in BaNoCoSS Conference, Helsinki

August 25, 2015

1. Background and motivation

- An area (domain) estimation process begins typically in a situation where the survey data has been collected under a specific sample design, but the population is not divided into subgroups (areas later) beforehand → sample sizes of areas are determined randomly → very low or zero sample sizes are possible.
- Under previous circumstances model-assisted or model-based computing techniques are the only choice. SAE methods are most widely known and used. Latest development: see Pfeffermann (2013).
- This research uses stratified sampling where strata and areas coincide.
- Crucial question: leading principle in allocating the sample into areas and the specific allocation criterion.
- Points of view of this study: 1) Information needed about target population in a single allocation and 2) The analytical optimization criterion related to the method.

- Allocations known from literature need 1) only number-based or 2) area-level information (parameters) and can be applied regardless of estimation method.
- Only sample information of response variable (y) is available → a proper proxy variable replaces it (values from a repeating research or produced by a model).
- Model-based estimation is used very commonly, but why are the model and method NOT used as given pre-information in the allocation phase? We use two allocations that utilize the model and method, in addition to area-level information.
- Total sample size (n) is assumed to be low (budget and time restrictions).
- Main problem: How do different allocations work in model-based estimation (linear unit-level mixed model and EBLUP)?
- Comparison of performances of allocations is based on results obtained from simulation experiments which are drawn from a self-generated artificial population.
- Our earlier studies have used real research data → results can be compared.

2. Allocations based on area-level information

Number-based allocations: Equal and proportional allocation
- sample sizes are fixed and don´t depend on area characteristics

Parameter-based allocations
- use area-level information (means, totals, std. dev´s, CV´s etc.)
- Neyman and Bankier: produce area sample sizes when *n* is given
- NLP (see Choudry et al. (2012)): produce overall sample size (*n*) and area sample sizes, result of NLP optimization

Summary of number-based and parameter-based allocations

Allocation method	Computing sample size n_d for area d
Equal	$n_d^{Equ} = n / D$
Proportional	$n_d^{Pro} = W_d n = (N_d / N) n$
Neyman	$n_d^{Ney} = n(N_d S_d(y) / \sum_{d=1}^{D} N_d S_d(y))$ $S_d(y) =$ standard deviation of y (or proxy - y or x)
Bankier (power)	$n_d^{Ban} = n(X_d^a CV(y)_d / \sum_{d=1}^{D} X_d^a CV(y)_d); a = 0.5$ here. $X_d =$ total of x and $CV(y)_d$ is CV of y (or proxy - y or x) in area d.
NLP	Minimum $n = \sum_{d=1}^{D} n_d$ satisfying tolerances for y or proxy - y or x : $CV(\bar{y}_d) \le CV_{0d}; d = 1,...D$ (CV´s of stratum sample means) $CV(\bar{y}_{st}) \le CV_0$ (CV of estimated population mean)

3. Model-based allocations

3.1 Basic unit-level mixed model

Number of areas = D; size of area d (basic units) = N_d

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

v_d : random area effect, mean $= 0$, variance $= \sigma_v^2$

e_{dk} : random effect, mean $= 0$, variance $= \sigma_e^2$

$E(y_{dk}) = \boldsymbol{x}'_{dk}\boldsymbol{\beta}$ and $V(y_{dk}) = \sigma_v^2 + \sigma_e^2$ (total variance)

Ratio $\varphi = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2)$: variation between areas / total variation (intra-area correlation)

Ratio of variances: $\delta = \sigma_e^2 / \sigma_v^2 = 1/\varphi - 1.$

General theory of this model is well-known and widely used. Estimates of common regression coefficients ($\hat{\boldsymbol{\beta}}$), variance components and area effects ($\hat{\sigma}_v^2$, $\hat{\sigma}_e^2$ and $\hat{v}_d$) are obtained from the sample.

EBLUP estimate (Empirical Best Linear Unbiased Predictor) for area total Y_d of response variable y = sum of sample values and predicted sum of non-sampled values:

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

EBLUP estimator is biased → MSE is used instead of variance. Prasad-Rao approximation of MSE_d for finite populations has four components:

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

Area-specific ratio which appears in MSE and includes variance components and area sample size:

$$\gamma_d = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2 / n_d).$$

MSE is computed from sample data. Components of MSE approximation include estimates of variance components σ_v^2 and σ_e^2, estimate of ratio γ_d and asymptotic variances of estimates of σ_v^2 and σ_e^2 :

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

where estimate of ratio γ_d is $\hat{\gamma}_d = \hat{\sigma}_v^2/(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_d^*)$,

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

$$g_{3d}(\hat{\sigma}_v^2,\hat{\sigma}_e^2) = (N_d - n_d^*)^2(n_d^*)^{-2}(\hat{\sigma}_v^2 + \hat{\sigma}_e^2/n_d^*)^{-3}[\hat{\sigma}_e^4 V(\hat{\sigma}_v^2) + \hat{\sigma}_v^4 V(\hat{\sigma}_e^2)$$

$$-2\hat{\sigma}_e^2\hat{\sigma}_v^2 Cov(\hat{\sigma}_e^2,\hat{\sigma}_v^2)],$$

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

Term n_d^* is a method-specific random variable expressing sample size in area d (not fixed).

3.2 g1-allocation

This allocation uses auxiliary variable *x* (all *N* values in areas) and homogeneity coefficient known of cluster sampling. First simple ANOVA of *x* and then adjusted homogeneity measure of variation between clusters:

$$R_{ax}^2 = 1 - R^2 = 1 - MSW / S_x^2 ,$$

where R^2 is coefficient of determination (regression analysis), MSW is mean SS of clusters (strata) and S_x^2 is variance of x. In practise R_{ax}^2 measures proportion SS_B / SS_{TOT}.

Basic criterion for estimation: minimum of sum of area MSE´s subject to the constraints of fixed overall sample size n (sum of area sample sizes = n). This corresponds minimization of sum of sample variances in design-based allocation.

Because of complexity of whole MSE the optimum is impossible to reach analytically. We use only the first and most important component g_1 of MSE (section 3.1).

If variation between areas is strong enough and the model is suitable for estimation, the proportion of g_1 of whole MSE reaches 85-95 % according to many researches. We search for minimum of the sum of area g_1´s as a function of sample sizes n_d .

Expression to be mimimized with respect of n_d subject to constraint $\sum_{d=1}^{D} n_d = n$:

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

Lagrange multiplier method is used. Solution (proved in licenciate thesis of M. Keto):

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

where $\delta = \sigma_e^2 / \sigma_v^2$ and $\varphi = \sigma_v^2 / (\sigma_v^2 + \sigma_e^2)$.

Because ratio φ is unknown, it is replaced with the homogeneity measure R_{ax}^2 of known auxiliary variable *x*. Variation in x is assumed to explain variation in y.

Computational sample sizes are rounded to nearest integer. Also non-linear programming minimizing the previous sum subject to sample size constraint can be used. Area sample sizes can be forced to become non-negative integers.

Sample size increases when area size increases, but not proportionally. The homogeneity coefficient has impact on area sample sizes. If all variation is between areas, the result is proportional allocation, because ratio of variances $\delta = 0$ (and ratio $\varphi = 1$).

Computational sample sizes of 14 areas of artificial research data according to hypothetical values and true value (0.569) of homogeneity coefficient.

Area	N_d	0.1	0.2	0.3	0.4	*0.569*	0.7	**0.9**	***Prop***
Area 1	120	0	0	0	0	*1*	1	**1**	***1***
Area 2	150	0	0	0	1	*1*	1	**2**	***2***
Area 3	493	2	4	5	5	*5*	5	**6**	***6***
Area 4	500	2	4	5	5	*5*	6	**6**	***6***
Area 5	555	4	5	6	6	*6*	6	**6**	***6***
Area 6	585	4	6	6	6	*6*	7	**6**	***7***
Area 7	621	5	6	7	7	*7*	7	**7**	***7***
Area 8	735	8	8	8	8	*8*	8	**8**	***8***
Area 9	818	10	9	9	9	*9*	9	**9**	***9***
Area 10	871	11	10	10	10	*10*	10	**10**	***10***
Area 11	958	13	12	11	11	*11*	11	**11**	***11***
Area 12	1,072	15	14	13	13	*13*	12	**12**	***12***
Area 13	1,122	16	15	14	14	*13*	13	**12**	***12***
Area 14	1,400	22	19	18	17	*17*	16	**16**	***15***
Total	10,000	112	112	112	112	*112*	112	**112**	***112***

3.3 *Sim* –allocation: general principles

Step 1: Proxy-*y* variable y^* is estimated by using a model obtained from a small pre-sample → complete register of y^* and auxiliary variable *x* (*N* values) is available.

Step 2: *K* SRSWOR samples with size *n* are simulated from this population. Sample sizes of areas are determined randomly.

Step 3: EBLUP estimation is carried out for each simulated sample to estimate area totals of y^*. Variance components, areal MSE approximations etc. can be computed.

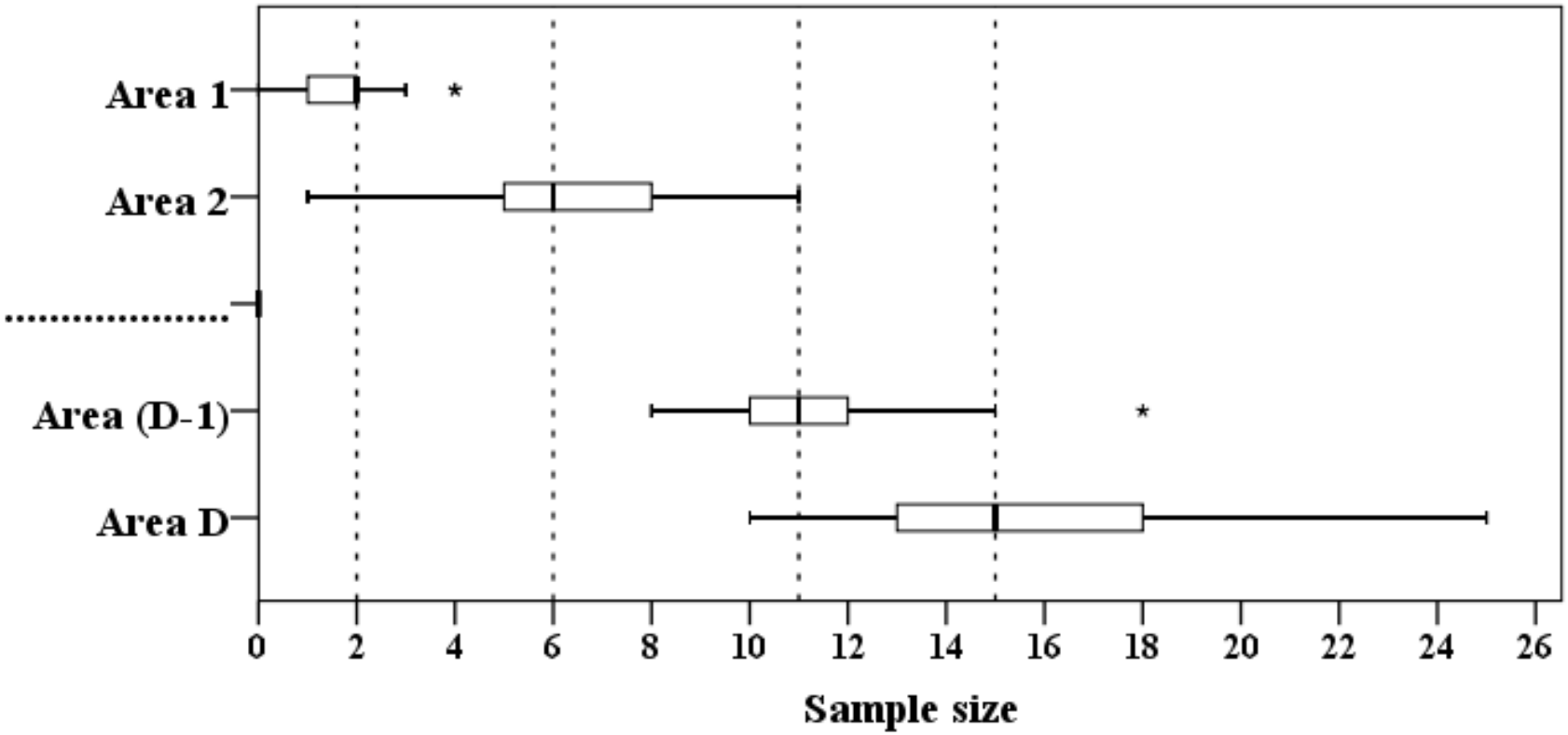
Step 4: Average of area MSE´s (= average sample-MSE) is computed for each sample. Samples are arranged in ascending order by average MSE.

Step 5: *k* samples with lowest average MSE are selected. Distributions of sample sizes n_d are created for each area.

Basic rule in selecting area sample sizes: n_d^{Sim} = cdf $^{-1}$ (0.5) (median).
Notation ”cdf”: cumulative distribution function of area sample size.

$\sum_d n_d^{Sim} < n \Rightarrow$ n_d´s of smallest areas are increased.

$\sum_d n_d^{Sim} > n \Rightarrow n_d$´s of largest areas are reduced.



4. Comparison of allocations: model-based vs. other allocations

Following allocations have been selected for comparison of performances: five previously defined and model-based allocations *g1* and *Sim*.

1,500 samples (SRSWOR inside strata = area) were simulated (SAS) for each allocation. Area estimates for totals Y_d ($d = 1,\ldots,D$) and necessary statistics were obtained by using selected model and EBLUP estimation. Quality measures were computed for each 1,500 samples´ set (SPSS). Overall sample size $n = 112$.

In each allocation: Values of quality measure $RRMSE_d\,\%$ (*Relative Root Mean Square Error*) are presented for areas, as well as *RRMSE%* for the population.

Artificial research data

Why artificial data? For evaluating allocations under different structures of areas compared with reference (real) data (sizes and prices of apartments, N = 9,815).

Population:	14 areas, 10,000 generated units (MatLab, SPSS)
Response variable (y) and auxiliary variable (x)	
Proxy-y variable:	Values estimated for *Sim* allocation
Sizes of areas:	120 – 1,400 units
Mean of X:	75.90 – 218.69 in areas, 119.34 in population
Mean of Y:	144.06 – 830.64 in areas, 459.92 in population
CV´s by area:	X: 0.107 – 0.540 and Y: 0.242 – 0.603
xy-correlation by area:	0.472 – 0.830
Homogeneity measure of x:	0.569 (in reference data 0.231; see Keto and Pahkinen (2014)).

Main steps in derivation of proxy-*y*

Two-stage cluster sampling and SRSWOR sampling are used. Cluster is one area.

1. *D* clusters (areas) are sorted in ascending order according to CV of variable *x*.
2. 3 clusters are selected randomly with SYS sampling with interval = *D*/3. Each cluster represents one CV group according to level of CV value.
3. A SRSWOR sample of 5 units is selected from each of 3 selected clusters. The result is a pre-sample of 15 units.
4. Regression model between *y* and *x* is constructed from each 5 unit set in the pre-sample. The result is 3 different models which represent CV groups.
5. Each regression model is applied to all units belonging to corresp. CV group.

$$y^{*}_{dk} = \alpha_r + \beta_r x_{dk} \; ; d = 1,2,..., D; \; k = 1,2,..., N_d; \; r = 1,2,3 \;\; (r = \text{CV - group})$$

Exception to rule: units in pre-sample. Value of proxy-*y* = real value of *y* (15 real).
After operation: *N* values (size of population) are available for proxy-*y*.

Quality measure used in evaluating the performances of allocations: Relative Root Mean Square Error (*RRMSE%*)

r = number of sample simulations in each allocation

i = number of a separate sample $(i = 1, \ldots, r)$

$RRMSE_d\,\%$: $$100 \times \sqrt{1/r \sum_{i=1}^{r} (\hat{Y}_{di,EBLUP} - Y_d)^2} \,/\, Y_d$$

Mean *RRMSE%* over areas $(MRRMSE_d\%)$: $$100 \times 1/D \sum_{d=1}^{D} \left(\sqrt{1/r \sum_{i=1}^{r} (\hat{Y}_{di,EBLUP} - Y_d)^2} \,/\, Y_d \right)$$

Estimate for population total in sample i: $$\hat{Y}_{i,EBLUP} = \sum_{d=1}^{D} \hat{Y}_{di,EBLUP}$$

RRMSE% for population total: $$100 \times \sqrt{1/r \sum_{i=1}^{r} (\hat{Y}_{di,EBLUP} - Y)^2} \,/\, Y$$

Table A. Sample sizes in simulations (g1-allocations according to homog. coeff.)

Area	N_d	Number-based		Parameter-based			Model-based			
		Equal	Prop	Ney_*x*	Ban_*xy*	Nlp_*x*	Hypoth. *g1*		True *g1*	*Sim*
							0.1	*0.2*	*0.569*	
Area 1	120	8	1	1	3	2	0	0	1	2
Area 2	150	8	2	2	4	15	0	0	1	2
Area 3	493	8	6	6	6	12	2	4	5	4
Area 4	500	8	6	9	8	8	2	4	5	6
Area 5	555	8	6	7	9	16	4	5	6	5
Area 6	585	8	7	7	7	8	4	6	6	7
Area 7	621	8	7	7	7	9	5	6	7	8
Area 8	735	8	8	3	5	2	8	8	8	9
Area 9	818	8	9	12	7	8	10	9	9	10
Area 10	871	8	10	8	6	11	11	10	10	9
Area 11	958	8	11	10	13	2	13	12	11	12
Area 12	1,072	8	12	22	21	8	15	14	13	12
Area 13	1,122	8	12	9	9	7	16	15	13	12
Area 14	1,400	8	15	9	7	4	22	19	17	14
Total	10,000	112	112	112	112	112	112	112	112	112

Table B. Area and population level *RRMSE%* values by allocation (*g1* : 3 columns)

Area	N_d	Number-based		Parameter-based			Model-based			
		Equal	Prop	Ney_*x*	Ban_*xy*	Nlp_*x*	Hypoth. *g1*		True *g1*	*Sim*
							0.1	*0.2*	*0.569*	
Area 1	120	7.23	16.80	17.80	11.18	13.53	31.37	30.14	16.65	12.99
Area 2	150	13.65	26.49	25.52	19.99	9.50	30.97	31.11	33.96	26.28
Area 3	493	14.72	19.56	19.19	20.24	10.89	38.93	26.72	22.80	24.17
Area 4	500	11.74	14.78	12.00	13.01	10.63	28.50	20.12	17.07	15.25
Area 5	555	15.43	16.76	15.45	14.12	10.59	20.19	18.09	17.23	18.83
Area 6	585	8.25	8.28	8.40	8.61	8.38	10.38	8.54	8.86	8.36
Area 7	621	11.01	13.87	13.70	14.75	9.84	19.96	17.05	14.75	12.66
Area 8	735	6.70	6.40	9.53	7.74	12.25	6.00	6.22	6.16	6.07
Area 9	818	10.89	11.95	10.08	13.73	10.29	12.78	13.41	12.68	11.10
Area 10	871	9.87	10.00	10.93	12.94	7.82	10.87	10.89	10.53	10.68
Area 11	958	8.36	7.23	7.73	6.84	18.40	6.95	7.19	7.46	7.15
Area 12	1,072	14.28	11.75	9.12	8.94	15.12	10.69	10.78	11.08	11.97
Area 13	1,122	13.63	11.09	13.51	12.85	15.40	9.44	10.12	10.98	11.57
Area 14	1,400	7.98	5.72	7.43	8.78	12.50	4.54	4.92	5.39	6.01
MRRMSE $_d$ *%*		10.98	12.91	12.89	12.41	11.80	17.25	15.38	13.97	13.08
RRMSE%		4.15	3.42	3.89	3.84	5.08	3.49	3.44	3.30	3.57

Table C. Area and population level *RRMSE%* values in reference data (n = 112)

Area		Number-based		Parameter-based			Model-based	
Label	Size N_d	Equal	Prop	Ney_x	Ban_x	Nlp_x	*g1*	*Sim*
Porvoo town	112	13.41	19.79	16.49	14.78	10.10	8.08	20.17
Pirkkala district	148	8.35	12.04	10.60	9.76	8.97	6.60	12.45
South Savo county	493	18.63	20.70	23.20	20.16	20.88	22.29	20.51
Jyväskylä town	494	13.61	14.43	20.83	18.33	21.98	15.36	14.38
Lappi county	555	19.91	21.34	25.45	23.97	22.59	21.72	19.90
South-East Finland	585	19.68	19.64	24.37	24.31	27.81	20.76	18.19
Helsinki (capital)	621	21.92	23.15	14.35	16.02	16.43	22.72	24.64
West coast district	655	20.35	19.92	21.75	20.67	18.91	21.15	19.22
Trackside district	818	12.31	11.38	13.73	12.76	13.47	11.93	11.85
Kuopio district	871	19.21	16.37	20.84	20.82	23.49	16.22	16.73
Turku district	958	20.94	17.74	21.57	22.70	26.44	17.56	17.39
Oulu district	1,072	16.96	14.34	21.22	19.00	19.81	14.39	12.90
Metropol area	1,100	12.14	9.78	10.16	10.78	11.55	9.59	10.10
Lahti-Tampere district	1,333	13.35	10.64	12.76	12.87	14.98	10.54	10.31
$MRRMSE_d$ %		16.48	16.52	18.38	17.64	18.39	15.64	16.34
RRMSE%		6.13	5.97	6.07	5.89	6.62	6.15	6.35

5. Results

Area sample sizes in allocations vary strongly. Especially in parameter-based allocations (particularly NLP) sample sizes have quite weak connection to sizes of areas.

Level of *RRMSE%* values is lower and differences between allocations are smaller than in results obtained from our reference data.

Some small, medium-sized and large areas had either a) low or b) high *RRMSE%* values, regardless of sample size. Large increment of sample size has a considerable impact only on a few areas.

All area *RRMSE%* values are below 20 % only in equal and Nlp_x allocation. Equal allocation has lowest *MRRMSE%* on area level. On the other hand, their *RRMSE%* values are highest and g1-allocation has lowest *RRMSE%* on population level.

Zero sample size causes high *RRMSE%* values (over 30 %) for two smallest areas.

6. Conclusions

Comparison of latest results with those obtained from our earlier research suggests that variation between areas and the area structure of the population have a strong impact on estimation results.

None of the allocations have good results both on area and on population level. Parameter-based allocations have better performance compared with earlier research.

g1-allocation, which uses only auxiliary variable *x*, has good results in all areas except two smallest ones. In our earlier research all areas (also small) had good results for this allocation, when homogeneity coefficient was lower. See Keto and Pahkinen (2014).

Zero sample sizes cause poor results for smallest areas in this kind of area structure.

Model-based *Sim* –allocation seems to be worth implementing and developing, as before. The contents of the pre-sample is essential. Its collection and the model used in developing proxy-*y* must be planned carefully.

7. Topics in further research

Overall sample size (n) is assumed to be low (here 112, 1.1 % of population). This must be taken into account when the model and estimation method are selected.

It is essential that the used model and EBLUP estimation are regarded as part of the pre-information which is used in determining the allocation to areas.

Because none of the allocations used here performed best both on area and population level, possibilities to develop a model-based allocation using a composite estimator must be considered. See Clark et al. (2013), Costa et al. (2004) and Longford (2006).

If auxiliary variables are utilized, one of the core questions is, how the variation of auxiliary variables inside and between areas is taken into account when explaining variation in response variable.

In addition to a unit level-model, an area model (like Fay-Herriot) must be deployed → results can be compared also between models.

It is worth studying also, in what kind of circumstances areas which receive sample size zero have good estimation results.

It is necessary to generate a new artificial research population data containing low variation (homogeneity coefficient 0.10 – 0.15) between areas for further testing of the performances of different allocations.

In the further research, also other quality measures (MSE, CV, bias, coverage of confidence intervals etc.) must be utilized when evaluating and comparing different allocations.

8. 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 effective domain estimation. *Survey Methodology* **38,** 23-29.

Clark, R. G. and Molefe, W. (2013). Model-assisted optimal allocation for planned domains using composite estimation. *Working paper 19-13, 1-27.* University of Wollongong, Australia.

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

Keto, M. (2014). *Aluekiintiöinti ositetussa otannassa*. University of Jyväskylä. Department of Mathematics and Statistics. (Licenciate thesis).

Keto, M. and Pahkinen, E. (2014). On sample allocation for efficient small-area estimation. *Small Area Estimation Conference 2014*. Poznan: Poznan University of Economics.

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

Pfeffermann, D. (2013). New Important Developments in Small Area Estimation. *Statistical Science* **28** 40–68.

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