Mauno Keto and Erkki Pahkinen
University of Jyväskylä, Finland

ONE RANDOM SAMPLE AS A TEST PLATFORM IN SEARCH FOR MORE ACCURATE ESTIMATES THROUGH SUB-SAMPLES SELECTED BY A MODEL-BASED ALLOCATION

BaNoCoss Conference, Örebro June 16-20 2019

Regurlarly repeated surveys and other types of research: continuous need for development

- Typical sample size in Finnish nationwide surveys: 1 500 – 3 000
- Improving the quality of the estimated population and area-level (domain-level) parameters (means, totals, proportions etc.)
- Reducing the implemention time and cost of the research project – without losing reliability in the results
- Factors affecting the time and cost: 1) number of contact trials before the desired sample size is reached and 2) measurement problems
- Example from opinion polls using quota sampling: 6-7 contact trials are needed to reach complete responses from one respondent
- One survey with 1 500 respondents may cost as much as 30 000 €

One option for saving time and cost: reduction of sample size by using exhaustive sample allocation

- It is easy to end up to proportional or equal allocation (or their combination Costa) – for simplicity? This may be the critical phase!
- Our experience: sample allocation is regarded often as a necessary phase carried out quickly – standard procedure which is followed
- Pre-information of variables of interest and of auxiliary variables, including past data (estimates of model parameters)
- Estimated parameters and their small area estimators with MSE
- The optimization criterion resulting into sample allocation into areas
- Importances of area and population estimates

PROBLEM: IMPROVING THE ACCURACIES OF ESTIMATES OBTAINED FROM ONE SRS SAMPLE

One SRS sample is selected from real data by proportional allocation. Area and population total estimates are obtained for the variable of interest by using three estimators:

(1) Design-based Horvitz-Thompson: $\hat{Y}_{d,HT} = \sum_{k \in s_d} y_{dk} / \pi_{dk}$

(2) Design-based model-assisted GREG: $\hat{Y}_{d,GREG} = \sum_{k \in U_d} \hat{y}_{dk} + \sum_{k \in s_d} (y_{dk} - \hat{y}_{dk}) / \pi_{dk}$

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

π_{dk} : inclusion probabilities for units $k \in s_d$

Model in (2) and (3) here: unit-level linear mixed model

$y_{dk} = \boldsymbol{x}'_{dk}\boldsymbol{\beta} + v_d + e_{dk}; k = 1,..., N_d; d = 1,..., D$ (unit-level auxiliary data available)

Real data (population) for sample simulations

National register of N = 33 429 block or row house apartments for sale, collected in April 2016 (maintained by Alma Mediapartners Ltd)

The population is divided into D = 18 provinces (areas here)

Variable of interest y	price of apartment (1 000 €)
Auxiliary variable x_1 (correl. with y)	size of apartment (m^2)
Auxiliary variable x_2 (correl. with y)	age of apartment (years)

Sizes of areas N_d : from 252 to 9 421. Large variation also in other area characteristics (totals, means, CV´s)

Estimated parameters: area and population totals of y

One random sample and simulated subsamples

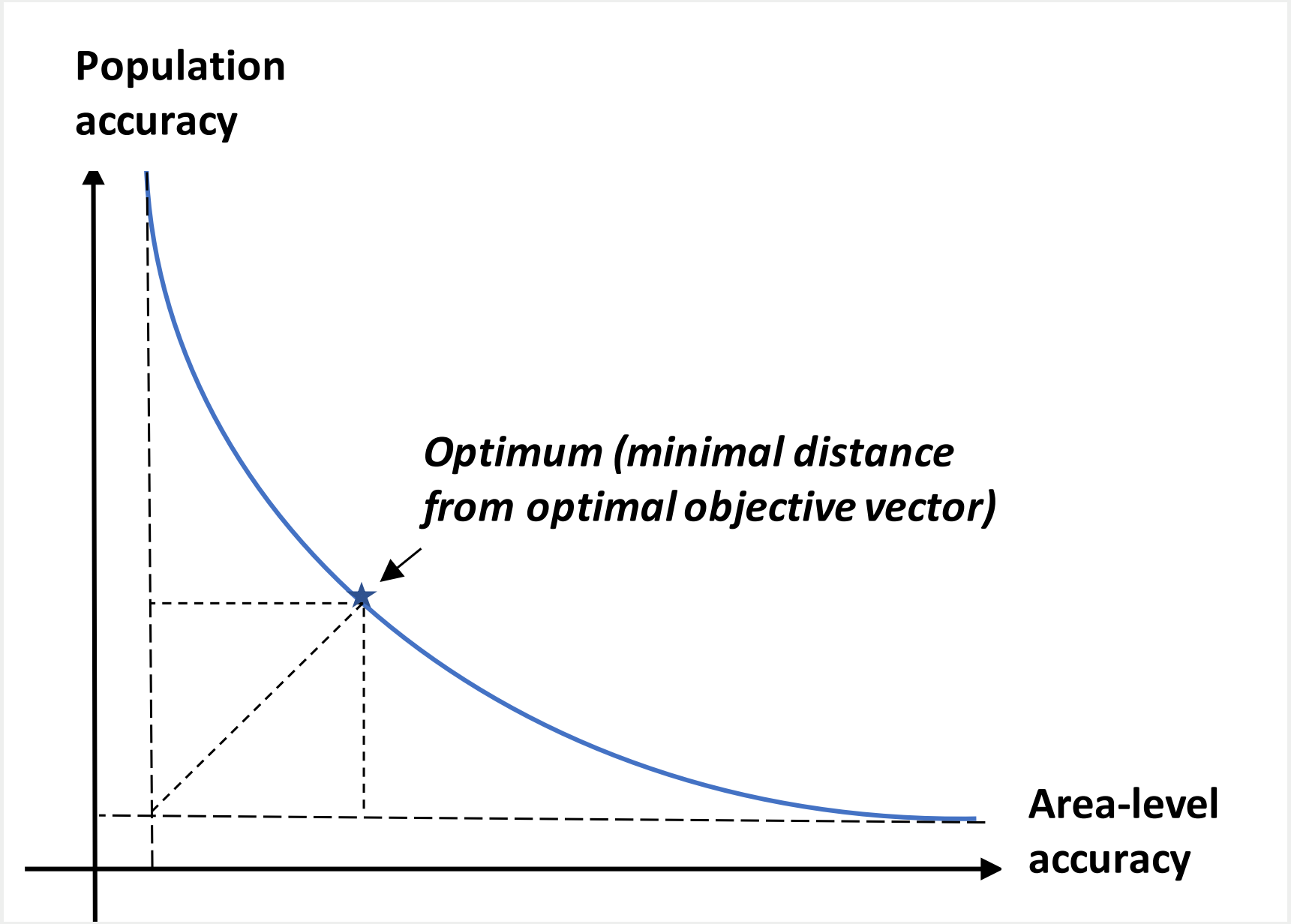
Size of original sample: n_{PRO} = 2 000 (6 % of N)
Constitutes the population for sub-samples, sizes of areas = (N_d/N) n_{PRO}

1 000 SRS subsamples are simulated from PRO-sample, n_{3TP} = 1 500
Notation ”3TP” refers to Three-term Pareto allocation for samples

3TP allocation uses three terms g_{1d}, g_{3d} and g_{4d} *of Prasad-Rao MSE estimator of estimator (3) and* multi-objective optimization (Keto-Hakanen-Pahkinen 2018). Past data is used to obtain estimates for model variances and asymptotic variances.

Area-specific sample sizes of PRO sample determine the upper limits of sample sizes for 3TP allocation

Principle of multi-objective Pareto optimization



Population characteristics of *y and allocations (PRO, 3TP)*

Area (province)	N_d	$\bar{Y}_d$	$S(y)_d$	$\mathrm{CV}(y)_d$	$n_{d,PRO}$	$n_{d,3TP}$	
Uusimaa	9 421	281.7	200.4	0.711	563	**194**	←
Pirkanmaa	3 177	161.7	91.9	0.568	190	**102**	←
North Ostrobothnia	2 456	145.7	83.0	0.570	147	147	
Varsinais-Suomi	2 405	168.2	121.2	0.720	144	**101**	←
Central Finland	1 999	138.2	73.9	0.535	120	120	
North Savo	1 831	135.5	89.5	0.661	110	110	
Satakunta	1 610	112.0	77.2	0.690	96	96	
Päijät-Häme	1 555	150.5	102.4	0.680	93	93	
Kanta-Häme	1 412	131.7	63.2	0.480	84	84	
Kymenlaakso	1 192	100.1	59.1	0.590	71	71	
South Savo	1 130	123.6	83.1	0.672	68	68	
North Karelia	1 013	142.9	78.0	0.546	61	61	
South Ostrobothnia	1 006	135.2	60.8	0.450	60	60	
Ostrobothnia	1 000	155.8	66.5	0.427	60	60	
Lapland	848	117.9	75.2	0.638	51	51	
South Karelia	819	127.0	79.7	0.628	49	49	
Kainuu	303	97.0	52.9	0.545	18	18	
Central Ostrobothnia	252	137.6	61.8	0.449	15	15	
Population	33 429	180.0	144.4	0.802	2 000	1 500	

Assessing the quality of area and population estimates: *Relative root mean square error (RRMSE) and absolute relative bias (ARB)*

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

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

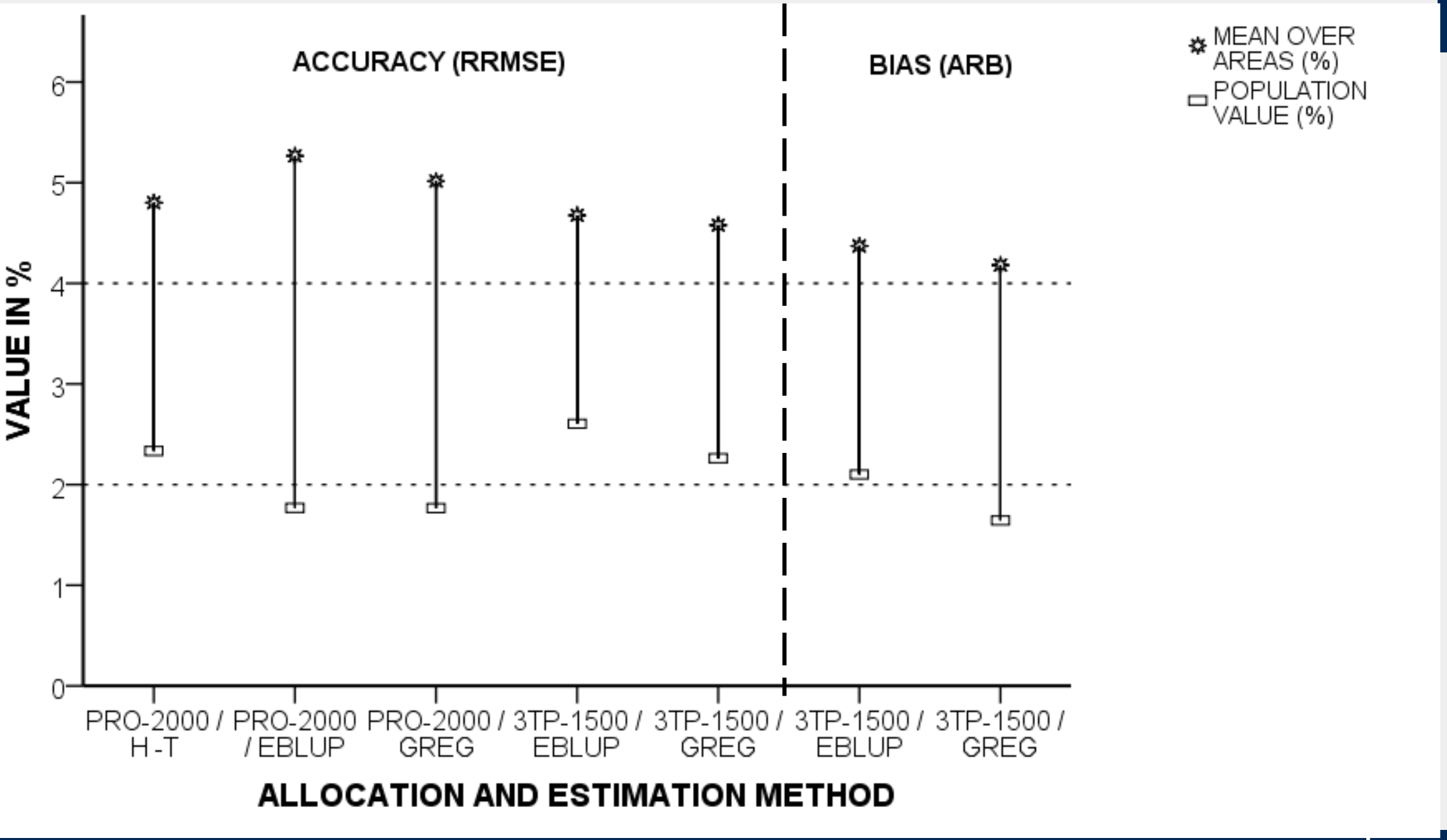
$$\mathrm{MRRMSE} = 1/D\sum_{d=1}^{D}\mathrm{RRMSE}_d \quad \text{and} \quad \mathrm{MARB} = 1/D\sum_{d=1}^{D}\mathrm{ARB}_d$$

$$\mathrm{RRMSE(pop)} = 100(1/r\sum_{i=1}^{r}(\hat{Y}_i - Y)^2)^{1/2} / Y$$

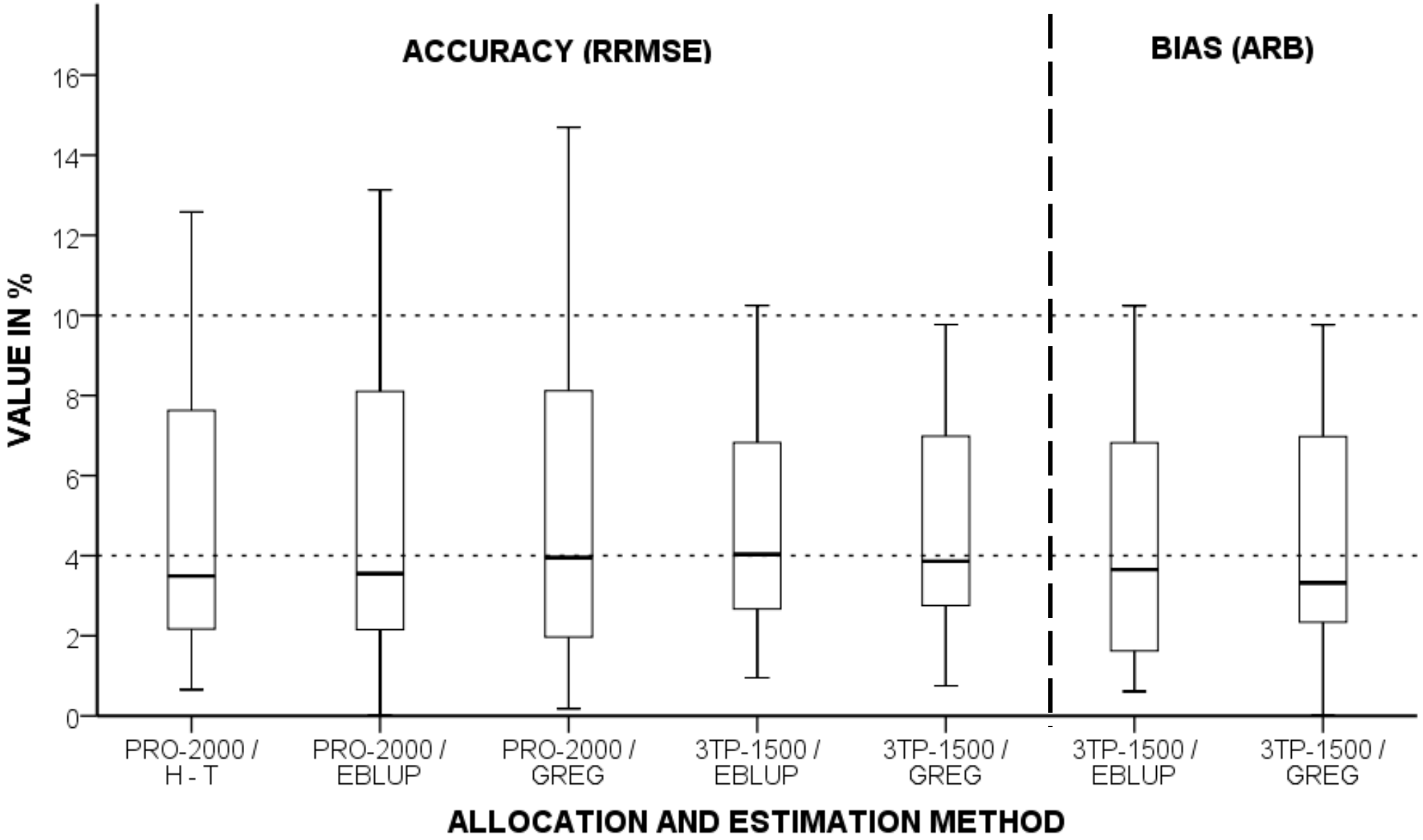
$$\mathrm{ARB(pop)} = 100\left|1/r\sum_{i=1}^{r}(\hat{Y}_i / Y - 1)\right|$$

r = number of sample simulations

Accuracy and bias (in %) for areas and for population



AREA-SPECIFIC DISTRIBUTIONS FOR ACCURACY AND BIAS



Precise values for RRMSE and for ARB and 3TP VS PRO

Area (province)		PRO-2000 sample				3TP-1500 samples			3TP vs PRO Change of RRMSE		3TP-1500 samples ARB in %	
		n_d	RRMSE in %			n_d	RRMSE in %					
Name	N_d		H-T	EBLUP	GREG		EBLUP	GREG	**EBLUP**	**GREG**	EBLUP	GREG
Uusimaa	9 421	563	2.84	3.33	2.65	194	5.05	4.16	**1.72**	**1.51**	3.75	2.34
Pirkanmaa	3 177	190	0.81	0.65	0.18	102	2.67	2.75	**2.02**	**2.57**	0.61	0.01
North Ostrobothnia	2 456	147	3.53	1.01	1.36	147	1.24	1.56	**0.23**	**0.20**	1.23	1.55
Varsinais-Suomi	2 405	144	1.70	2.15	2.75	101	3.33	3.87	**1.18**	**1.12**	1.62	2.36
Central Finland	1 999	120	2.17	3.47	4.14	120	2.83	3.27	**-0.64**	**-0.87**	2.82	3.26
North Savo	1 831	110	5.56	3.40	4.36	110	3.31	3.85	**-0.09**	**-0.51**	3.31	3.85
Satakunta	1 610	96	7.63	11.60	10.63	96	10.25	9.77	**-1.35**	**-0.86**	10.24	9.77
Päijät-Häme	1 555	93	0.66	3.46	3.77	93	2.14	2.46	**-1.32**	**-1.31**	2.11	2.44
Kanta-Häme	1 412	84	3.37	9.27	10.77	84	7.26	8.08	**-2.01**	**-2.69**	7.24	8.05
Kymenlaakso	1 192	71	1.20	8.10	5.46	71	5.90	4.23	**-2.20**	**-1.23**	5.88	4.21
South Savo	1 130	68	6.59	7.33	9.37	68	6.83	8.15	**-0.50**	**-1.22**	6.82	8.15
North Karelia	1 013	61	2.29	7.89	8.12	61	6.63	6.98	**-1.26**	**-1.14**	6.62	6.98
South Ostrobothnia	1 006	60	3.46	0.03	1.18	60	0.95	2.09	**0.92**	**0.91**	0.92	2.08
Ostrobothnia	1 000	60	8.36	3.64	2.03	60	3.57	2.86	**-0.07**	**0.83**	3.56	2.84
Lapland	848	51	9.48	0.50	1.97	51	1.48	3.41	**0.98**	**1.44**	1.41	3.38
South Karelia	819	49	4.32	4.22	6.61	49	4.49	5.97	**0.27**	**-0.64**	4.49	5.96
Kainuu	303	18	12.58	13.13	0.25	18	9.23	0.75	**-3.90**	**0.50**	9.20	0.05
Central Ostrobothnia	252	15	9.90	11.61	14.69	15	7.01	8.19	**-4.60**	**-6.50**	6.85	8.00
Population	33 429	2 000	2.33	1.77	1.77	1 500	2.60	2.26	**0.83**	**0.49**	2.10	1.64
Mean over areas			4.80	5.27	5.02		4.68	4.58	**-0.59**	**-0.44**	4.37	4.18

MAIN RESULTS

Accuracies of most EBLUP and GREG area estimates improve. EBLUP estimates of two smallest areas improve considerably. All large RRMSE values have decreased.

Area-level accuracies in general (means over areas) improve slightly and population-level accuracies decrease slightly.

Areas with diverging characteristics: RRMSE values remain quite large. These areas are also considerably biased. PRO sample does not represent the whole population.

Important: lower overall sample size leads practically to same accuracy.

CONCLUSIONS AND FURTHER RESEARCH

This experiment suggests the improvement trial and use of model-based allocation for the sub-sample.

Careful design of sample allocation may lead to substantial reduction of time and cost in the implementation of a survey. All surveys in a year carried out by one research organization → overall effect?

This problem must be studied further under different area structures.

New interest: optimal sample allocation for SAE with discrete variables
- estimated parameters: proportions for population and for areas
- model: for example logistic regression

REFERENCES

Battese, G. E., Harter, R. M., and Fuller, W. A. 1988. An error component Model for Prediction of County Crop Areas using Survey and Satellite Date. Journal of the American Statistical Association 83, 28-36.

Friedrich, U., Münnich, R., and Rupp, M. 2018. Multivariate Optimal Allocation with Box-Constraints. Austrian Journal of Statistics 47, 33–52.

Keto, M., Hakanen, J., and Pahkinen, E. 2018. Register data in sample allocations for small-area estimation. An International Journal of Mathematical Demography 25, 184-214. DOI: https://doi.org/ 10.1080/08898480.2018.1437318.

Keto, M. and Pahkinen, E. 2017. On overall sampling plan for small area estimation. Statistical Journal of the IAOS 33, 727-740. DOI: 10.3233/SJI-170370.

Lehtonen, R. and Veijanen, A. 2009. Design-Based Methods of Estimation for Domains and Small Areas. In Handbook of Statistics, Vol. 29B, 219-249. New York: Elsevier.

Miettinen, K. 1999. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, Boston.

Rao, J. N. K. and Molina, I. 2015. Small Area Estimation (2nd Edition). Hoboken, NJ: John Wiley & Sons, Inc.