COMPARING COMBINATIONS OF ESTIMATOR AND SAMPLE ALLOCATION WHEN ESTIMATING POPULATION AND DOMAIN-SPECIFIC PROPORTIONS FOR A BINARY VARIABLE

Mauno Keto[1] and Risto Lehtonen[2]

[1] University of Jyväskylä, Finland
e-mail: maujohketo@gmail.com

[2] University of Helsinki, Finland
e-mail: risto.lehtonen@helsinki.fi

Abstract

In many surveys, both population-level and domain-level statistics are estimated for the target variables which can be continuous or discrete, and the objective is to obtain reliable estimates for each level. The domains may have very diverse sizes and other relevant characteristics. For this reason, it is important to plan stratified sampling and domain estimation carefully, so that the objective is possible to reach. Very small sample sizes are possible for some domains. In this situation, small area estimation, although its basic idea is to utilize information from other domains, does not necessarily produce high-quality estimates for every domain. Sometimes it may be reasonable to set limits to domain-specific sample sizes in order to obtain even moderate estimates for the domains. The concept of optimal allocation for domains depends on the situation. It is a solution of an optimization problem, but all relevant objectives can scarcely be reached. The selected estimator of the target variable may have a strong impact on estimation results, and the combination of sample allocation and estimator is worth studying also. It is possible to develop a model- and estimator-based allocation.

We use planned domains in our study. Our main interest is focused on estimating population and domain-specific proportions for a binary variable by using three model-assisted estimators based on a logistic regression model which use auxiliary data. We compute the domain-specific sample sizes according to six different allocation principles. Five of these allocations are developed by utilizing earlier collected proxy data and the logistic model. But we also test the performances of three other estimators: a direct Horvitz-Thompson estimator, a model-based EBLUP and a model-assisted GREG estimator. The last two estimators are based on the same regression model with random domain-specific effects. We use two allocations with these three estimators. The assessment of the performances of the estimators and allocations at the domain and population levels are based on design-based sample simulations. We measure the performances of the allocations and estimators with quality indicators. We introduce four different R-square measures to assess the suitability of the logistic models in estimation.

The estimators based on the logistic models outperform the design-based and model-related regression estimators, but the performances of different logistic models are close to each other. One allocation can be regarded as slightly more effective than the others. The predictive power of the logistic models can be regarded as moderate.

Keywords: Auxiliary and proxy data, model-assisted logistic regression, direct estimator, model-based and model-assisted regression estimator, performance, optimization, limitation of sample size, trade-off between domains and population, predictive power.

References

Allison, P. (2013). What's the Best R-Squared for Logistic Regression? *Statistical Horizons* (*https://statisticalhorizons.com/r2logistic*).

Cox, D.R. and E.J. Snell (1989) *Analysis of Binary Data*. Second Edition. Chapman & Hall.

Demidenko, E. (2008). Sample size and optimal design for logistic regression with binary interaction. *Statistics in Medicine* 27: 36–46.

Duchesne, P. (2003). Estimation of a Proportion with Survey Data. *Journal of Statistics Education* 11(3).

Gabler, S., Ganninger, M., and Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika 75*: 15–161.

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.

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.

Lehtonen, R., Särndal, C.-E. and Veijanen, A. (2005). Does the model matter? Comparing model-assisted and model-dependent estimators of class frequencies for domains. *Statistics in Transition* 7: 649–673.

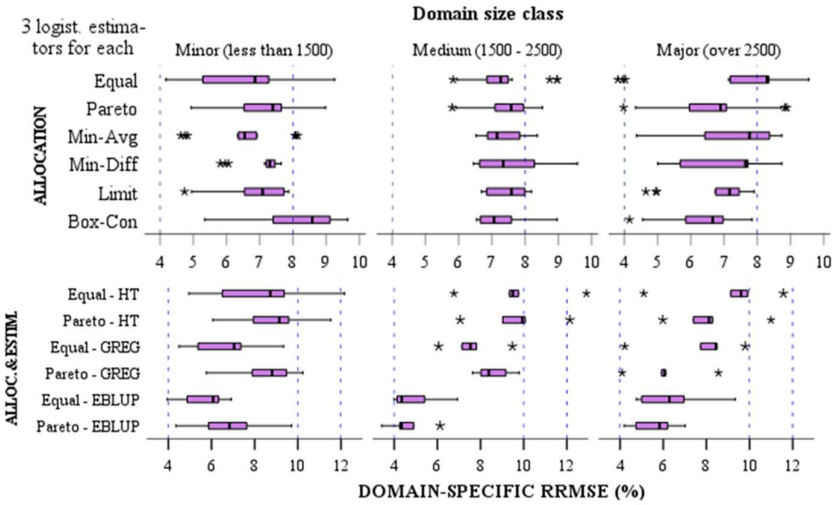
Lehtonen R. and Veijanen A. (2016). Model-assisted methods for Small Area Estimators of Poverty Indicators. In *Analysis of Poverty Data by Small Area Estimation, Pratesi M. (ed.). Wiley and Sons*: 109–127.

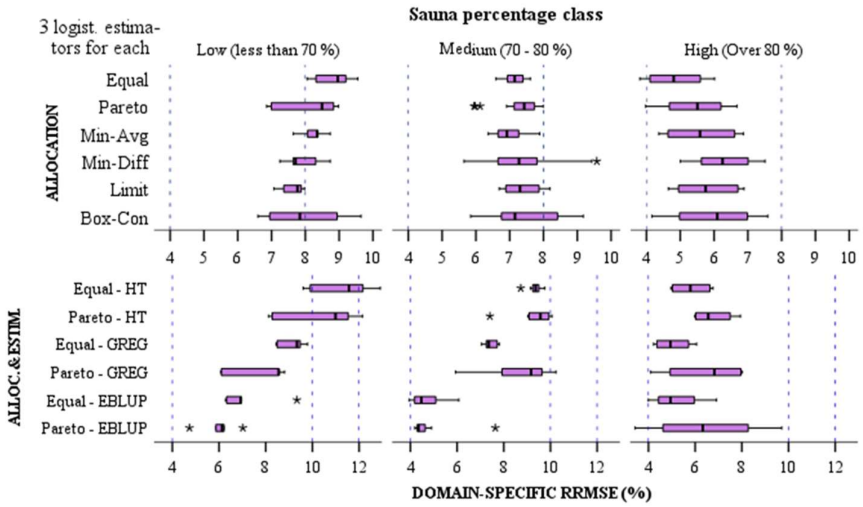
Miettinen, K. 1999. Nonlinear Multiobjective Optimization. Kluwer Academic Publishers, Boston.

Nagelkerke, N.J.D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* 78: 691–692.

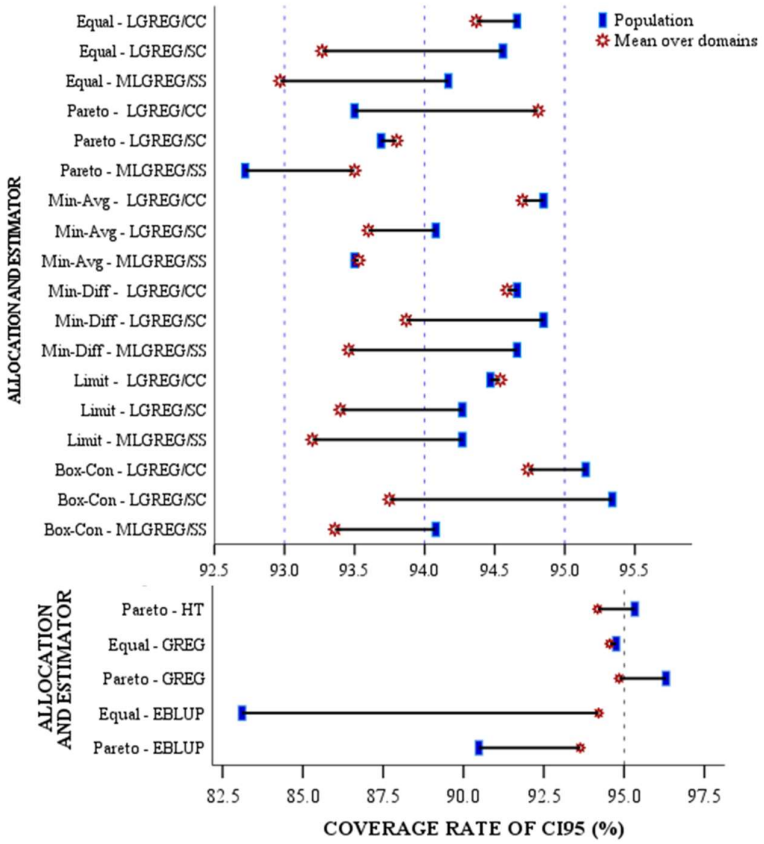
Rao, J. N. K. and Molina, I. 2015. Small Area Estimation (2nd Edition). Hoboken, NJ: John Wiley & Sons, Inc.

Figures describing results (accuracy RRMSE, CI95 coverage, and R-square measures)

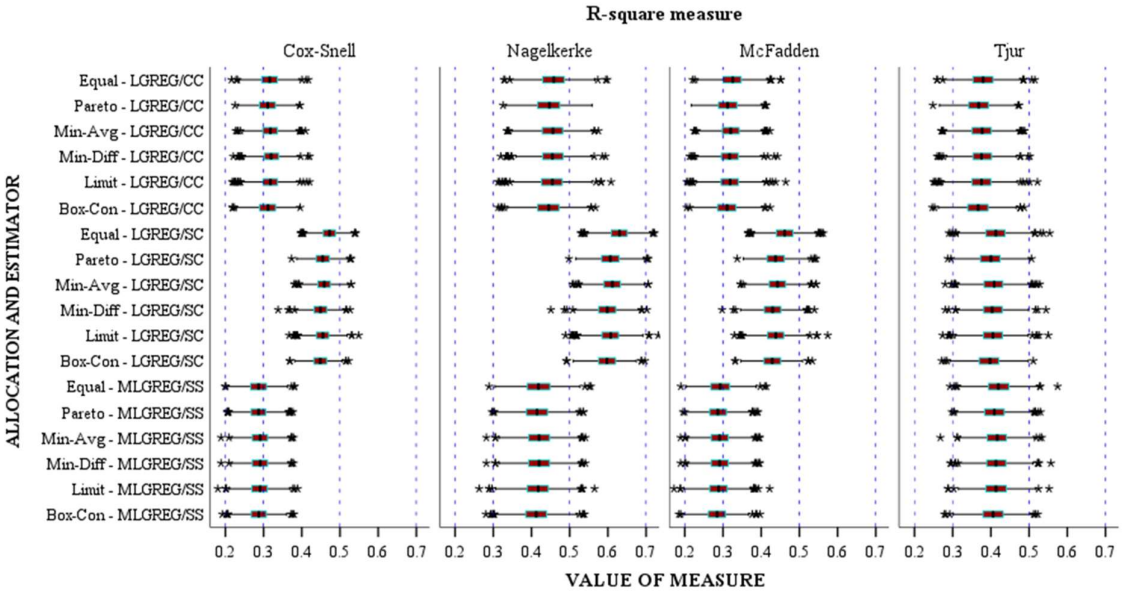




CI95 coverage rates by allocation and estimator



Distributions of different R-square measures in samples by allocation and logistic estimator


R-square measure
Cox-Snell
Nagelkerke
McFadden
Tjur
ALLOCATION AND ESTIMATOR
Equal - LGREG/CC
Pareto - LGREG/CC
Min-Avg - LGREG/CC
Min-Diff - LGREG/CC
Limit - LGREG/CC
Box-Con - LGREG/CC
Equal - LGREG/SC
Pareto - LGREG/SC
Min-Avg - LGREG/SC
Min-Diff - LGREG/SC
Limit - LGREG/SC
Box-Con - LGREG/SC
Equal - MLGREG/SS
Pareto - MLGREG/SS
Min-Avg - MLGREG/SS
Min-Diff - MLGREG/SS
Limit - MLGREG/SS
Box-Con - MLGREG/SS
0.2 0.3 0.4 0.5 0.6 0.7
VALUE OF MEASURE