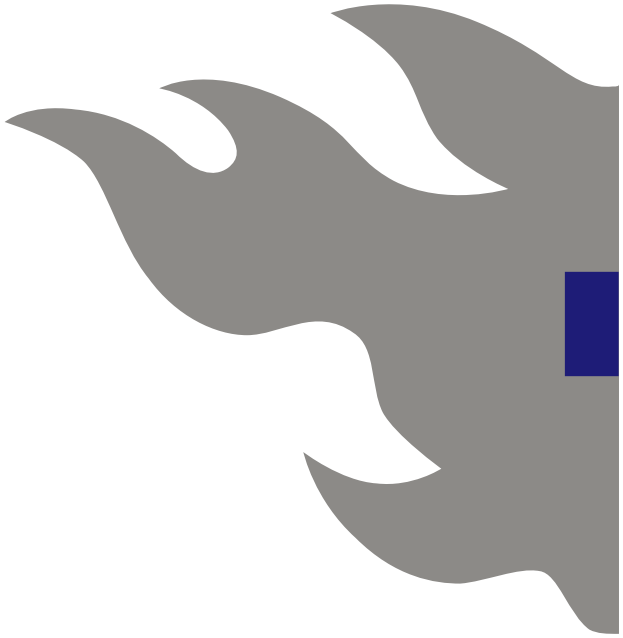
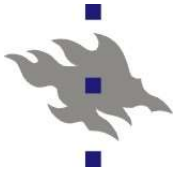
HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Model-assisted calibration and MRP methods for small area estimation: an empirical comparison

Risto Lehtonen (University of Helsinki)
Ari Veijanen (University of Helsinki)



Workshop on Survey Statistics
Tartu, August 2022



Outline

Introduction

MRP: Multilevel regression and poststratification

Gelman and Little (1997), Gelman and Hill (2006)

SAE with MRP under informative sampling

Poisson probability proportional to size sampling (PPS)

Ignoring informativeness can cause bias in estimation and invalid inferences

Si (2022): MRP design consistent if the poststratification cell structure fully accounts for design information

Our study

Behavior of MRP when ignoring sampling informativeness

Does inclusion of sampling information in poststratification help?

MC: Model calibration

Design consistent reference method

Wu and Sitter (2001), Lehtonen and Veijanen (2019)

Parameters of interest

Domain proportions

$$p_d = \frac{\sum_{k \in U_d} Y_k}{N_d}, \quad d = 1, \ldots, D,$$

where

Y_k population values of binary Y, $k \in U$

U_d domains of interest, $U_d \subset U$

N_d size of domain, $\sum_{d=1}^{D} N_d = N$

MRP estimators

$$\hat{p}_d^{MRP} = \sum_{c=1}^{C_d} \frac{N_{dc}}{N_d} \overline{p}_{dc} = \sum_{c=1}^{C_d} w_{dc}^{MRP} \overline{p}_{dc}, \; d = 1,...,D$$

MRP estimator in domain d, where

$\overline{p}_{dc} = \sum_{k \in U_{dc}} \hat{y}_k \; / \; N_{dc}, \; c = 1,...,C_d$ poststratum cell means

$w_{dc}^{MRP} = N_{dc} \; / \; N_d$ MRP weights

$\hat{y}_k$ predictions from logistic model, $k \in U$

NOTE: In usual (design-based) poststratification

$\hat{p}_d^{POST} = \sum_{c=1}^{C_d} \frac{N_{dc}}{N_d} \overline{p}_{dc}^{POST}$ poststratified estimator in domain d

$\overline{p}_{dc}^{POST} = \sum_{k \in s_{dc}} y_k \; / \; n_{dc}$ poststratum sample cell means

y_k sample observations, $k \in s$

MC estimators (Hájek type)

$$\hat{p}_d^{MC} = \frac{\sum_{k \in s_d} w_{dk}^{MC} y_k}{\sum_{k \in s_d} w_{dk}^{MC}}, \; d = 1, \ldots, D$$

Weights w_{dk}^{MC} are obtained by minimizing a chi-square distance function subject to calibration equations

$\sum_{k \in s_d} w_{dk}^{MC} \hat{y}_k = \sum_{k \in U_d} \hat{y}_k$, where

w_{dk}^{MC} model calibration weights

$\hat{y}_k$ predictions from logistic model, $k \in U$

NOTE: MC uses $\hat{y}_k$ in calibration equations instead of the auxiliary variable values $\mathbf{x}_k = (x_{1k}, \ldots, x_{pk})'$ as in the traditional calibration of Deville and Särndal (1992), see Wu and Sitter (2001)

Monte Carlo design

Hierarchical population U, size N= 10000 elements

5 regions U_r, 15 subregions U_t

Binary Y_k, mean in population 0.0988

Population generated with logistic mixed model

$$\eta_k = \text{logit}(E(Y_k)) = \beta_0 + \sum_{i=1}^{2} \beta_{1i} x_{1ik} + \sum_{i=1}^{2} \beta_{2i} x_{2ik} + \beta_3 x_{3k} + \beta_4 x_{4k} + \gamma z_k + u_{r[k]} + v_{t[k]},\ u_r \sim N(0, \sigma_u^2),\ v_t \sim N(0, \sigma_v^2)$$

x_1 and x_2 unit-level three-category variables

x_3 continuous unit-level variable, x_4 continuous subregion-level variable

z size variable in Poisson sampling, $cor(\eta, z) = 0.8$

$D = 45$ domains U_d created by cross-classifying subregions U_t ($t = 1,...,15$) with three-category x_1

Independent Poisson samples s_j ($j = 1,...,1000$) of size $n = 1000$ drawn from the population in the simulation experiments

Logistic models

Predicted values $\hat{y}_k = \exp(\hat{\eta}_k) \,/\, (1 + \exp(\hat{\eta}_k)),\ k \in U$

MRP:

Model 1 $\quad \hat{\eta}_k = \hat{\beta}_0 + \hat{\beta}_4 x_{4k} + \hat{\alpha}^{x_1}_{i[k]} + \hat{\alpha}^{x_2}_{j[k]} + \hat{u}_{d[k]}$

Model 2 $\quad \hat{\eta}_k = \hat{\beta}_0 + \hat{\beta}_4 x_{4k} + \hat{\alpha}^{x_1}_{i[k]} + \hat{\alpha}^{x_2}_{j[k]} + \hat{\alpha}^{z_q}_{l[k]} + \hat{u}_{d[k]}$

where $\quad \alpha^{x_1} \sim N(0, \sigma^2_{x_1}),\ \alpha^{x_2} \sim N(0, \sigma^2_{x_2}),\ \alpha^{z_q} \sim N(0, \sigma^2_{z_q})$

$u_d \sim N(0, \sigma^2_u), \quad z_q$ quartiles of Poisson size variable z

R-package autoMrP (Broniecki, Leemann and Wuest 2022)

MC:

Model 3 $\quad \hat{\eta}_k = \hat{\beta}_0 + \sum_{i=1}^{2} \hat{\beta}_{1i} x_{1ik} + \sum_{i=1}^{2} \hat{\beta}_{2i} x_{2ik} + \hat{\beta}_4 x_{4k} + \hat{u}_{d[k]}$

Model 4 $\quad \hat{\eta}_k = \hat{\beta}_0 + \sum_{i=1}^{2} \hat{\beta}_{1i} x_{1ik} + \sum_{i=1}^{2} \hat{\beta}_{2i} x_{2ik} + \hat{\beta}_4 x_{4k} + \hat{\gamma} z_k + \hat{u}_{d[k]}$

where $u_d \sim N(0, \sigma^2_u),\ z_k$ values of continuous z

ARB and RRMSE

Absolute Relative Bias

$$ARB(\hat{p}_d) = \frac{1}{1000}\sum_{j=1}^{1000} \mid \hat{p}_d(s_j) - p_d \mid / p_d, \;\; d = 1,\ldots,45$$

Relative Root Mean Squared Error

$$RRMSE(\hat{p}_d) = \sqrt{(1 \,/\, 1000)\sum_{j=1}^{1000} (\hat{p}_d(s_j) - p_d)^2} \;\; / \;\; p_d$$

where

$\hat{p}_d(s_j)$ estimated domain proportion for sample s_j

p_d known population proportion in the domain

Table 1 Scenarios implemented in Monte Carlo experiments

Strategies for accounting for Poisson sampling in inference	
Multilevel regression and poststratification MRP	
Scenario 1	Predictions $\hat{y}_k$: Design information not included Weights w_{dc}^{MRP} : Design information not included 135 poststrata, 3 in each domain
Scenario 2	Predictions $\hat{y}_k$: Size variable in model Weights w_{dc}^{MRP} : Size variable in poststratification 540 poststrata, 12 in each domain
Model calibration MC	
Scenario 3	Predictions $\hat{y}_k$: Design information not included Weights w_{dk}^{MC} : Size variable contributes via design weight $a_k = 1 / \pi_k$
Scenario 4	Predictions $\hat{y}_k$: Size variable in model Weights w_{dk}^{MC} : Size variable contributes via design weight and model

Table 2 Median ARB (%) and median RRMSE (%) of domain proportion estimators of binary Y for 45 domains under unequal probability sampling (Generated population, 1,000 Poisson PPS samples of 1,000 elements

	Design bias Median ARB (%)		Design accuracy Median RRMSE (%)	
	Expected domain sample size		Expected domain sample size	
	Minor < 20 (30 domains)	Major ≥ 20 (15 domains)	Minor < 20 (30 domains)	Major ≥ 20 (15 domains)
Multilevel calibration and poststratification MRP				
Scenario 1	29.4	31.5	65.9	54.9
Scenario 2	15.6	8.2	47.0	37.7
Model calibration MC				
Scenario 3	2.6	1.2	115.4	45.7
Scenario 4	7.8	1.2	109.5	41.5

Discussion

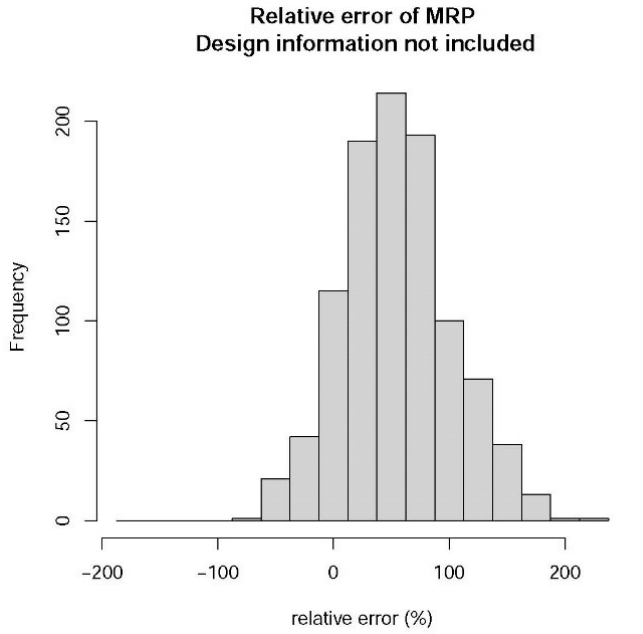
MRP provides an interesting model-based approach for small area estimation

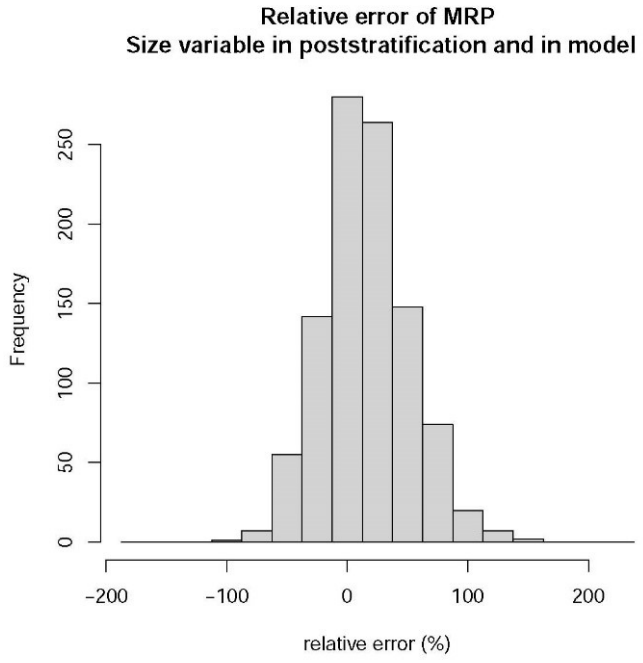
Sampling informativeness should be accounted for in estimation

The success of MRP prominently depends on the availability of auxiliary information strongly related to the outcome and the inclusion mechanism

Our limited empirical exercise suggests further studies on the limitations and potentials of MRP in relation to other SAE methods

Relative error of MRP in Domain 1 (minor domain)





References

Broniecki P., Leemann L. and Wuest R. (2022) Improved multilevel regression with post-stratification through Machine Learning (autoMrP). The Journal of Politics, 84, 1. https://www.duo.uio.no/bitstream/handle/10852/85772/broniecki-etal-jop-2020.pdf?sequence=2&isAllowed=y

Deville J.-C. and Särndal C.-E. (1992) Calibration estimators in survey sampling. JASA 87, 376–382.

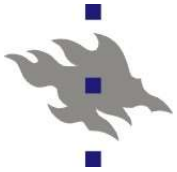
Gelman, A. and Hill J. (2006) Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge university press.

Gelman A. and Little T.C. (1997) Poststratification into many categories using hierarchical logistic regression. Survey Methodology, 23, 127–35.

Lehtonen R. and Veijanen A. (2019) Hybrid calibration methods for small domain estimation. Statistica & Applicazioni, XVII, 2, 201–235. https://www.researchgate.net/publication/346657782_Hybrid_calibration_methods_for_small_domain_estimation

Si Y. (2022) On the use of auxiliary variables in multilevel regression and poststratication. https://arxiv.org/abs/2011.00360

Wu C. and Sitter R.R. (2001) A model-calibration approach to using complete auxiliary information from survey data. JASA 96, 185–193.



Thank you!