HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Small domain estimation with calibration methods

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

ITACOSM 2019 – Florence, Italy, 5-7 June 2019

Outline

Preliminaries

Notation and methods

Monte Carlo experiments

Discussion

Literature

This is joint work with my colleague Dr Ari Veijanen of Statistics Finland

Calibration methods to be discussed

Traditional *model-free* calibration (MFC)

- Deville J.-C. & Särndal C.-E. (1992)
- Särndal C.-E. (2007)
- Lehtonen & Veijanen (2009) (domain estimation)

Model-assisted calibration (MC)

- Wu & Sitter (2001) (Model calibration)
- Montanari & Ranalli (2005)
- Lehtonen & Veijanen (2012, 2016) (domain estimation)

Hybrid calibration HC)

- Montanari & Ranalli (2009) (Multiple model calibration)
- Lehtonen & Veijanen (2015) (domain estimation)

Two-level hybrid calibration (HC2)

- Lehtonen & Veijanen (2017)

Questions of interest

Relative design-based properties of MFC, MC, HC and HC2
- Horvitz-Thompson (HT) vs. Hájek (HA) type calibration estimators of totals and proportions for domains
  - Accuracy properties
  - Distributional properties of calibrated weights

Main interest: What happens in small domains (with small domain sample size)?

Empirical framework
- Design-based simulation experiments
- Generated & real populations
  - GLMMs

Target parameters

Domain totals for continuous target variable *y*

$$t_d = \sum_{k \in U_d} y_k, \quad d = 1,...,D$$

Domain proportions

$$p_d = \frac{t_d}{N_d} = \frac{\sum_{k \in U_d} y_k}{N_d}, \quad d = 1,...,D$$

where the target variable y is binary (1: in poverty, 0: otherwise)

$U = \{1,...,k,...,N\}$ unit-level population

$U_d \subset U,\ d = 1,...,D$ domains of interest

$R_d \supset U_d,\ d = 1,...,D$ supersets, higher-level regions (for HC2)

We assume access to auxiliary data vectors

$\mathbf{x}_k = (x_{1k},...,x_{jk},...,x_{Jk})'$ for every $k \in U$

Sample data & estimator types

$s \subset U$ sample from U with sampling design $p(s)$

$r_d = s \cap R_d$ part of sample s falling in higher-level area R_d

$s_d = s \cap U_d$, $r_d \supset s_d$, part of sample s falling in domain U_d

$a_k = 1/\pi_k$ design weight, π_k inclusion probability

HT and HA type calibration estimators for domain totals of continuous y:

$$\hat{t}_{dHT} = \sum_{k \in s_d} w_{dk} y_k \text{ and } \hat{t}_{dHA} = \frac{N_d}{\hat{N}_d} \hat{t}_{dHT},\ \hat{N}_d = \sum_{k \in s_d} w_{dk}$$

Proportions for the binary y:

$$\hat{p}_{dHT} = \frac{\hat{t}_{dHT}}{N_d} \text{ and } \hat{p}_{dHA} = \frac{\hat{t}_{dHT}}{\hat{N}_d}$$

where w_{dk} are method-specific calibration weights

Calibration at domain level U_d

Calibration equations for MFC, MC and HC

$$\sum_{i \in s_d} w_{di} \mathbf{z}_i = \sum_{i \in U_d} \mathbf{z}_i, \quad d = 1,...,D \qquad (1)$$

w_{di} calibration weight for element i in domain d

$\mathbf{z}_i$ generic calibration vector for unit i

We minimize $$\sum_{k \in s_d} \frac{(w_{dk} - a_k)^2}{a_k} - \boldsymbol{\lambda}'_d \left(\sum_{i \in s_d} w_{di} \mathbf{z}_i - \sum_{i \in U_d} \mathbf{z}_i \right) \qquad (2)$$

subject to calibration equations (1)

Equation (2) is minimized by weights $w_{dk} = a_k \left(1 + \boldsymbol{\lambda}'_d \mathbf{z}_k\right)$

$$\boldsymbol{\lambda}'_d = \left(\sum_{i \in U_d} \mathbf{z}_i - \sum_{i \in s_d} a_i \mathbf{z}_i \right)' \left(\sum_{i \in s_d} a_i \mathbf{z}_i \mathbf{z}'_i \right)^{-1}, \quad d = 1,...,D$$

Calibration vectors for Horvitz-Thompson (HT) and Hájek (HA) type MFC, MC and HC estimators

MFC	HT type: Calibration vector $\mathbf{z}_i = (1, \mathbf{x}'_{Ci})'$, $i \in U_d$ HA type: Calibration vector $\mathbf{z}_i = \mathbf{x}_{Ci}$, $i \in U_d$ $\mathbf{x}_{Ci} = (x_{1i}, ..., x_{Ji})'$ calibration x-vector
MC	HT type: Calibration vector $\mathbf{z}_i = (1, \hat{y}_i)'$, $i \in U_d$ HA type: Calibration vector $\mathbf{z}_i = \hat{y}_i$, $i \in U_d$ $\hat{y}_i = f(\mathbf{x}'_{Mi}(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d))$, $i \in U_d$, predictions from the mixed model $\mathbf{x}_{Mi} = (1, x_{1i}, ..., x_{Ji})'$ model x-vector
HC	HT type: Calibration vector $\mathbf{z}_i = (1, \hat{y}_i, \mathbf{x}'_{Ci})'$, $i \in U_d$ HA type: Calibration vector $\mathbf{z}_i = (\hat{y}_i, \mathbf{x}'_{Ci})'$, $i \in U_d$ $\hat{y}_i = f(\mathbf{x}'_{Mi}(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d))$, $i \in U_d$, $\mathbf{x}_{Ci}$ calibration x-vector, $\mathbf{x}_{Mi}$ model x-vector $\mathbf{x}_{Ci}$ and $\mathbf{x}_{Mi}$ are separate or overlapping sub-vectors of $\mathbf{x}_i$

Calibration at domain U_d & region R_d levels

Calibration equations for HC2

$$\sum_{i \in r_d} w_{ri} \mathbf{z}_i^{(1)} = \sum_{i \in U_d} \mathbf{z}_i^{(1)} \quad \text{MC part (lower level, domains)} \qquad (3)$$

$$\sum_{i \in r_d} w_{ri} \mathbf{z}_i^{(2)} = \sum_{i \in R_d} \mathbf{z}_i^{(2)} \quad \text{MFC part (higher level, regions)} \qquad (4)$$

where $r_d = s \cap R_d$, supersets (regions) $R_d \supset U_d$ (domains)

With suitably defined $\mathbf{z}_i^{(1)}$ and $\mathbf{z}_i^{(2)}$, we minimize

$$\sum_{k \in r_d} \frac{\left(w_{rk} - a_k\right)^2}{a_k} - \boldsymbol{\lambda}_r' \left(\sum_{i \in r_d} w_{ri} \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix} - \begin{pmatrix} \sum_{i \in R_d} \mathbf{z}_i^{(1)} \\ \sum_{i \in R_d} \mathbf{z}_i^{(2)} \end{pmatrix} \right) \qquad (5)$$

subject to calibration equations (3) and (4)

Equation (5) is minimized by weights $w_{rk} = a_k \left(1 + \boldsymbol{\lambda}_r' \mathbf{z}_k\right)$

$$\boldsymbol{\lambda}_r' = \left(\sum_{i \in R_d} \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix} - \sum_{i \in r_d} a_i \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix} \right)' \left(\sum_{i \in r_d} a_i \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix} \begin{pmatrix} \mathbf{z}_i^{(1)} \\ \mathbf{z}_i^{(2)} \end{pmatrix}' \right)^{-1}$$

NOTE: Weights for $k \in r_d$ outside s_d tend to be small in absolute value

Calibration vectors for HT and Hájek type two-level hybrid (HC2) estimators

Level 1 (domains) calibration vectors for MC

HT type: $\mathbf{z}_i^{(1)} = (x_{0i}^{(1)}, \hat{y}_i^{(1)})'$, $i \in R_d$

HA type: $\mathbf{z}_i^{(1)} = \hat{y}_i^{(1)}$, $i \in R_d$

$x_{0i}^{(1)} = 1$, $\hat{y}_i^{(1)} = \hat{y}_i$, $i \in U_d$

$x_{0i}^{(1)} = 0$, $\hat{y}_i^{(1)} = 0$, $i \in R_d \setminus U_d$

$\hat{y}_i = f(\mathbf{x}'_{Mi}(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d))$, $i \in U_d$, $\mathbf{x}_{Mi}$ model x-vector

Level 2 (regions) calibration vectors for MFC

$\mathbf{z}_i^{(2)} = \mathbf{x}_{Ci}$, $i \in R_d$, $\mathbf{x}_{Ci}$ calibration x-vector

$\mathbf{x}_{Ci}$ and $\mathbf{x}_{Mi}$ are separate or overlapping sub-vectors of $\mathbf{x}_i$

NOTE: $\hat{t}_{dHT-HC2} = \sum_{k \in r_d} w_{rk} y_k$ (similar for other HC2 estimators)

EXAMPLE 1: Domain totals for simulated population

Design-based simulation experiment with generated data

$$y_k = (5 + \alpha_{0r}) + u_{0d} + (3 + u_{1d})x_{1k} + (1 + u_{2d})x_{2k} + \beta_{3d}x_{3k} + \varepsilon_k$$

Fixed finite population of 1,000,000 units

40 domains of interest: Minor (15 domains)
Medium-sized (15 domains)
Major (10 domains)

Higher level regions: 4 regions, 10 domains per region

Monte Carlo experiments

K = 10,000 SRSWOR samples of n = 2000 units

Fitted models: Linear mixed models

$$y_k = \beta_0 + u_{0d} + \beta_1 x_{1k} + \beta_2 x_{2k} + \beta_3 x_{3k} + \varepsilon_k$$

Calibration vectors for HT & HA type estimators

MFC	HT calibration vector $\mathbf{z}_i = (1, x_{1i}, x_{2i}, x_{3i})'$ HA calibration vector $\mathbf{z}_i = (x_{1i}, x_{2i}, x_{3i})'$, $i \in U_d$
Predictions for HT and HA type MC, HC, HC2 $\hat{y}_i = \hat{\beta}_0 + \hat{u}_{0d} + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \hat{\beta}_3 x_{3i}$, $i \in U_d$	
MC	HT calibration vector $\mathbf{z}_i = (1, \hat{y}_i)'$ HA calibration vector $\mathbf{z}_i = \hat{y}_i$
HC	HT calibration vector $\mathbf{z}_i = (1, \hat{y}_i, x_{3i})'$, $i \in U_d$ HA calibration vector $\mathbf{z}_i = (\hat{y}_i, x_{3i})'$, $i \in U_d$
HC2	Calibration vectors: HT Level 1: $\mathbf{z}_i^{(1)} = (x_{0i}^{(1)}, \hat{y}_i^{(1)})'$, $i \in R_d$ HA Level 1: $\mathbf{z}_i^{(1)} = \hat{y}_i^{(1)}$, $i \in R_d$ $x_{0i}^{(1)} = 1$, $\hat{y}_i^{(1)} = \hat{y}_i$, $i \in U_d$ $x_{0i}^{(1)} = 0$, $\hat{y}_i^{(1)} = 0$, $i \in R_d \setminus U_d$ HT and HA Level 2: $\mathbf{z}_i^{(2)} = x_{3i}$, $i \in R_d$

Accuracy of estimators

Relative root mean squared error (RRMSE)

$$\mathrm{RRMSE}(\hat{t}_d) = \sqrt{\frac{1}{K}\sum_{i=1}^{K}(\hat{t}_d(s_i) - t_d)^2} \,/\, t_d, \quad d = 1,\ldots,D$$

where

$\hat{t}_d(s_i)$ estimate from sample s_i for domain d

t_d known parameter value in domain d

K number of simulated samples

NOTE: MFC, MC, HC and HC2: Nearly design unbiased

Largest $ARB(\hat{t}_d) < 0.2\%$

Table 1 Median RRMSE (%) of design-based HT and Hájek type calibration estimators for totals for 40 domains in three domain sample size classes (Generated population)

	Expected domain sample size			All	Expected domain sample size			All
	Minor 12	Medium 40	Major 122		Minor 12	Medium 40	Major 122	
	RRMSE (%) HT type estimators				RRMSE (%) Hájek type estimators			
Model-free (direct) method								
MFC	8.82	1.62	0.78	1.72	6.39	1.89	0.91	1.98
Model-assisted (indirect) methods								
MC	4.29	1.58	0.78	1.67	4.53	1.85	0.91	1.96
HC	5.49	1.60	0.78	1.69	4.90	1.88	0.92	1.99
HC2	4.25	1.58	0.78	1.67	4.55	1.86	0.91	1.96

Distributional properties of calibrated weights

Problems of practical concern in *model-free* calibration:

- Possible large variation of weights
- Weights smaller than one
- Positive but extremely small weights
- Negative weights

To what extent can *model-assisted* calibration methods help?

Any differences between HT type vs. Hájek type methods?

Small simulation experiment:

- 100 SRSWOR samples of size 2,000 elements from *U*

Results:

- Median of interquartile range (IQR) of weights
- Distribution of weights by domain size

Table 2 Median interquartile range (IQR) of calibrated weights w_{dk} for HT and w_{HAdk} for Hájek relative (%) to the IQR of HT-type MFC in minor domains

Method	Expected domain sample size					
	Minor	Medium	Major	Minor	Medium	Major
	HT type w_{dk}			Hájek type w_{HAdk}		
Model-free (direct) method						
MFC	100	51	30	79	43	26
Model-assisted (indirect) methods						
MC	61	35	21	43	28	14
HC	78	43	25	60	36	21
HC2	61	36	21	47	28	16

$$w_{HAdk} = N_d \times \frac{w_{dk}}{\sum_{k \in s_d} w_{dk}}$$

Fig. 1 Distribution of weights by domain size class
HT type calibration estimators

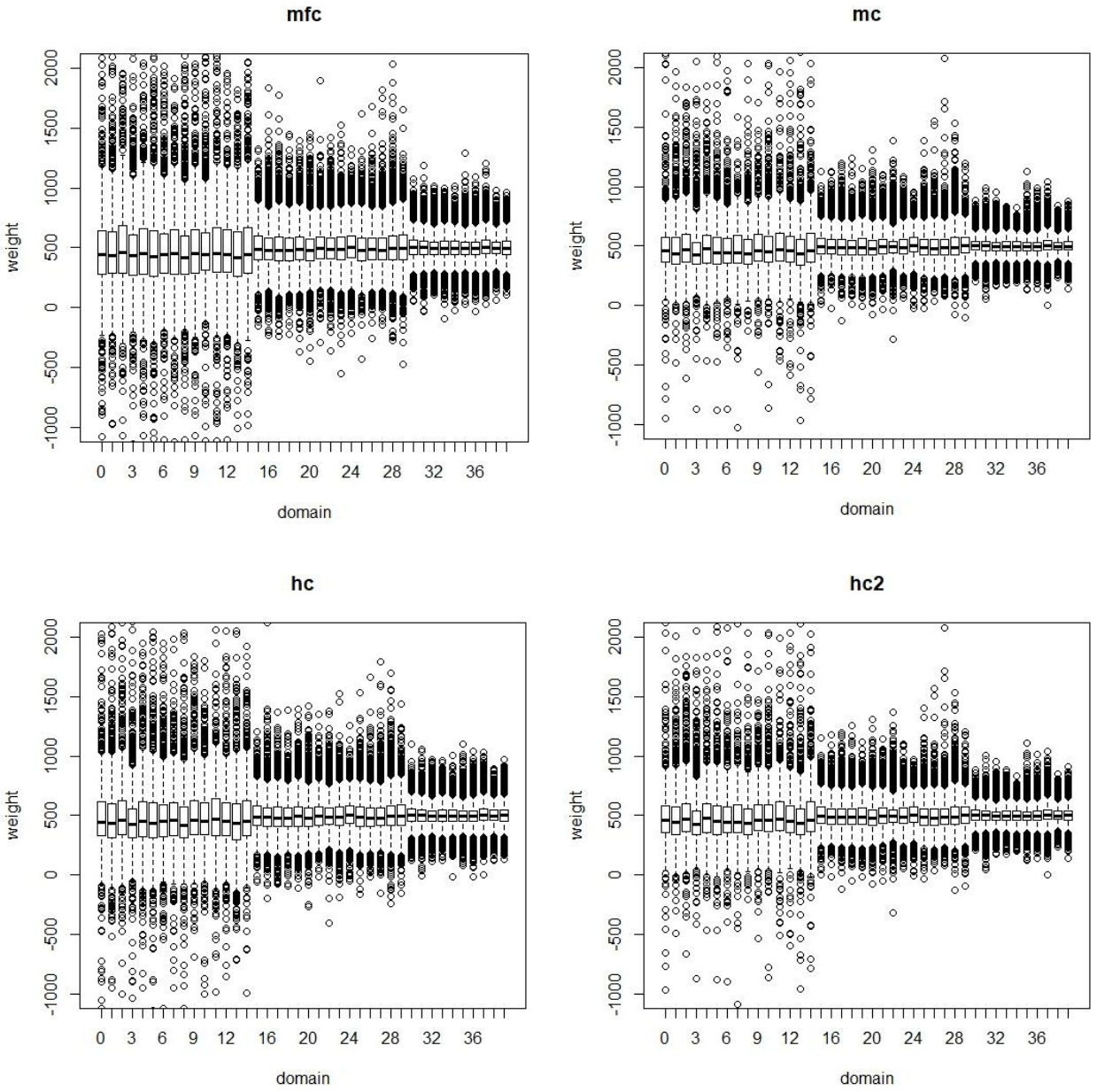
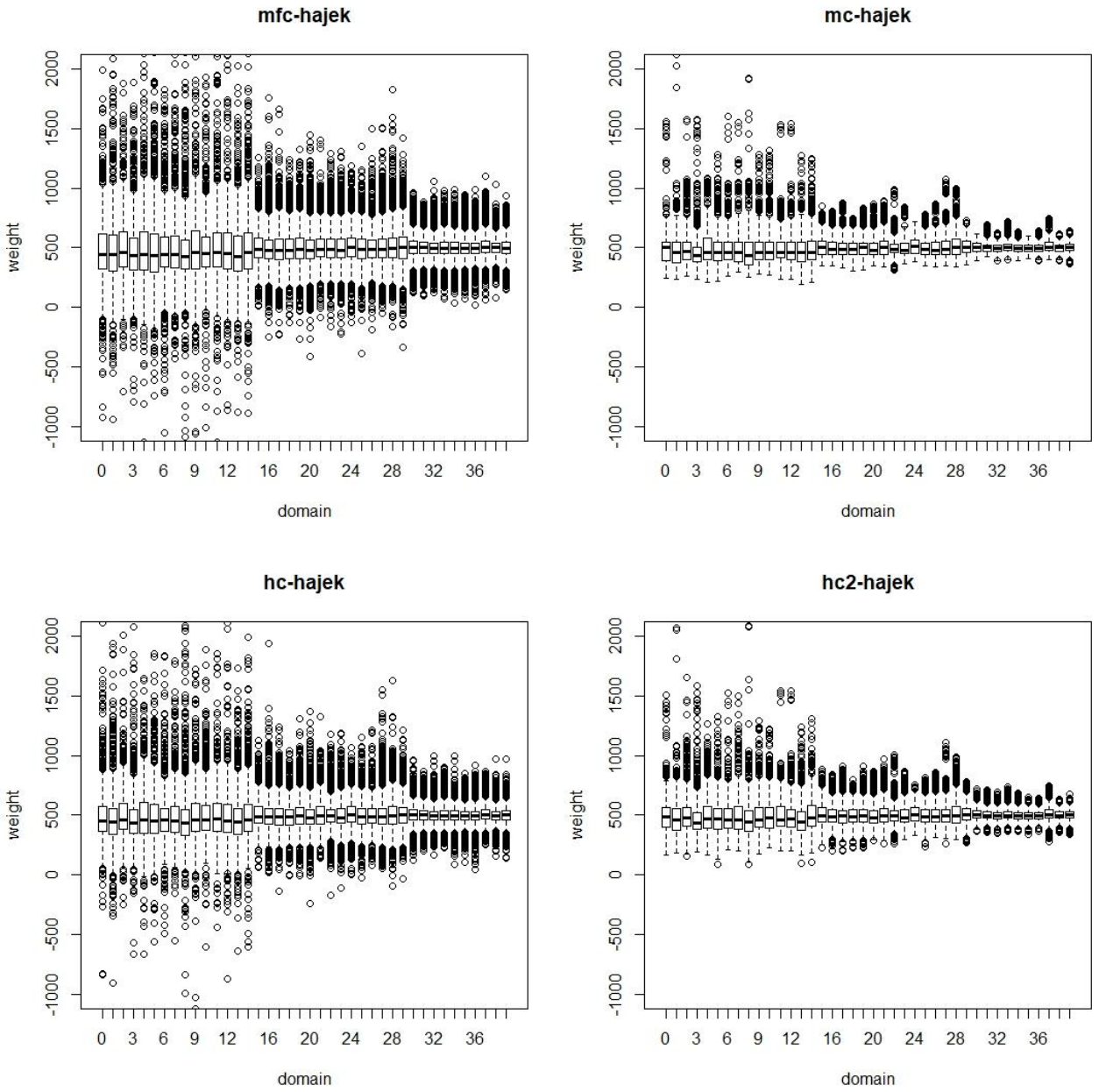


Fig. 2 Distribution of weights by domain size class
Hájek type calibration estimators



EXAMPLE 2: Domain poverty rates for real population

Design-based simulation experiment with real data of 795,000 adults

Regional hierarchy: LAU1 regions within NUTS3 regions
Domains of interest: 36 LAU1 regions
Higher level regions: 7 NUTS3 regions

Estimation of poverty rate for the D = 36 domains
Overall poverty rate in population: 14.3% (9.9 - 22.4%)

Monte Carlo experiments: K = 1000 simulated samples
SRSWOR sampling of n = 2000 persons

Auxiliary information

$\mathbf{x}_k = (x_{1k}, x_{2k}, x_{3k}, x_{4k}, x_{5k}, x_{6k})'$ auxiliary data vector

x_1 and x_2 Indicators for three-category labor force status

x_3 Indicator for sex class

x_4, x_5 and x_6 Indicators for four-category age

Calibration vectors for HT and HA type methods

MFC	HT calibration vector $\mathbf{z}_i = (1, \mathbf{x}'_{Ci})'$; HA calibration vector $\mathbf{z}_i = \mathbf{x}_{Ci}$, $i \in U_d$ $\mathbf{x}_{Ci} = (x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i})'$
MC, HC, HC2: $\hat{y}_i = \exp(\mathbf{x}'_{Mi}\hat{\boldsymbol{\beta}} + \hat{u}_d) / (1 + \exp(\mathbf{x}'_{Mi}\hat{\boldsymbol{\beta}} + \hat{u}_d))$, $i \in U_d$	
MC	Model x-vector $\mathbf{x}_{Mi} = (1, x_{1i}, x_{2i}, x_{3i}, x_{4i}, x_{5i}, x_{6i})'$ HT calibration vector $\mathbf{z}_i = (1, \hat{y}_i)'$; HA calibration vector $\mathbf{z}_i = \hat{y}_i$, $i \in U_d$
HC	Model x-vector $\mathbf{x}_{Mi} = (1, x_{1i}, x_{2i})'$ for LF status Calibration x-vector $\mathbf{x}_{Ci} = (x_{3i}, x_{4i}, x_{5i}, x_{6i})'$ for gender and age group HT calibration vector $\mathbf{z}_i = (1, \hat{y}_i, \mathbf{x}'_{Ci})'$; HA calibration vector: $\mathbf{z}_i = (\hat{y}_i, \mathbf{x}'_{Ci})'$, $i \in U_d$
HC2	Model x-vector $\mathbf{x}_{Mi}$ and calibration x-vector $\mathbf{x}_{Ci}$ as in HC Calibration vectors HT Level 1: $\mathbf{z}_i^{(1)} = (x_{0i}^{(1)}, \hat{y}_i^{(1)})'$, $i \in R_d$ HA Level 1: $\mathbf{z}_i^{(1)} = \hat{y}_i^{(1)}$, $i \in R_d$ $x_{0i}^{(1)} = 1$, $\hat{y}_i^{(1)} = \hat{y}_i$, $i \in U_d$; $x_{0i}^{(1)} = 0$, $\hat{y}_i^{(1)} = 0$, $i \in R_d \setminus U_d$ HT and HA Level 2: $\mathbf{z}_i^{(2)} = \mathbf{x}_{Ci}$, $i \in R_d$

Table 3 Median RRMSE (%) of design-based HT and Hájek type calibration estimators of poverty rates for 36 domains in three domain sample size classes (Real population)

	Expected domain sample size			All	Expected domain sample size			All
	Minor <25	Medium 25-50	Major >50		Minor <25	Medium 25-50	Major >50	
	RRMSE (%) HT type estimators				RRMSE (%) Hájek type estimators			
Model-free (direct) method								
MFC	70.8	48.7	30.9	48.7	64.6	47.7	30.6	47.7
Model-assisted (indirect) methods								
MC	54.6	44.0	29.9	44.0	53.9	43.6	30.2	43.6
HC	69.5	47.1	30.6	47.1	64.1	47.5	30.9	47.5
HC2	55.2	44.2	30.0	44.2	54.2	44.1	30.4	44.1

Summary of results

All estimators considered appeared nearly design unbiased

Model-assisted calibration estimators outperform direct model-free calibration in accuracy, in small domains in particular

Model-assisted calibration shows best overall accuracy

Hybrid calibration offers coherence property for selected x-variables but can suffer from instability in small areas

Two-level hybrid calibration decreases instability and can provide a good compromise

Model-assisted calibration and two-level hybrid calibration indicate best weight performance, for Hájek type calibration in particular

Hájek type model-assisted calibration MC and two-level HC may offer safe choices when negative weights are not allowed

References

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

Lehtonen R. and Veijanen A. (2009) Design-based methods of estimation for domains and small areas. In Rao C.R. and Pfeffermann D. (Eds.) Handbook of Statistics 29B. Elsevier, 219-249.

Lehtonen R. and Veijanen A. (2012) Small area poverty estimation by model calibration. Journal of the Indian Society of Agricultural Statistics 66, 125-133.

Lehtonen and Veijanen (2015) Small area estimation by calibration methods. WSC 2015 of the ISI, Rio de Janeiro, August 2015.

Lehtonen R. and Veijanen A. (2016) Design-based methods to small area estimation and calibration approach. In: Pratesi M. (Ed.) Analysis of Poverty Data by Small Area Estimation. Chichester: Wiley.

Lehtonen R. and Veijanen A. (2017) A two-level hybrid calibration technique for small area estimation. SAE2017 Conference, Paris, June 2017.

Montanari G.E. and Ranalli M. G. (2005) Nonparametric model calibration estimation in survey sampling. JASA 100, 1429–1442.

Montanari G.E. and Ranalli M.G. (2009) Multiple and ridge model calibration. Proceedings of Workshop on Calibration and Estimation in Surveys 2009. Statistics Canada.

Särndal C.-E. (2007) The calibration approach in survey theory and practice. SMJ 33, 99–119.

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

Thank you!