HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Small area estimation by calibration methods

Risto Lehtonen, University of Helsinki
Ari Veijanen, Statistics Finland

BaNoCoSS 2015, Helsinki, August 2015

Outline

- Aims & framework
- Calibration estimators for area (domain) totals
  - Model-free calibration
  - Model calibration
  - Hybrid calibration
- Monte Carlo experiment
- Summary
- References

Aims

- General: Investigation of methods that incorporate flexible modelling into design-based estimation procedures under access to unit-level population data
- Specific: Comparison of certain more recent model-assisted calibration methods with traditional model-free calibration in the estimation of totals for population subgroups or domains (small or large)
  - Method: Empirical study with design-based simulation experiments
  - Focus: Accuracy of estimators

FRAMEWORK: Calibration methods in survey sampling

	Model-free (linear) calibration MFC	Model calibration MC	Hybrid calibration HC
Weight calibration	Calibration to reproduce known population totals of auxiliary variables	Calibration to the population total of predictions derived via specified model	Combination of MC and MFC, depending on modeling and coherence requirements
Typical study variable	Continuous	Continuous, binary, polytomous, count	
Level of auxiliary data	Aggregate level	Unit level	Unit level Aggregate level
Model specification	Linear relationships (No explicit model statement)	Many options e.g. Generalized linear (mixed) models family	
Main aims	Coherence with published statistics “Multi-purpose” weighting Accuracy improvement	Accuracy improvement Flexible modelling	Accuracy improvement Flexible modelling Coherence with published statistics
Selected literature	Deville & Särndal (1992) Estevao & Särndal (1999) Särndal (2007) Lehtonen & Veijanen (2009)	Wu & Sitter (2001) Wu (2003) Montanari & Ranalli (2005) Lehtonen & Veijanen (2012, 2015a)	Montanari & Ranalli (2009) Lehtonen & Veijanen (2014) Lehtonen & Veijanen (2015b)

Model calibration: some references

- Wu & Sitter (2001) JASA (plus corrigenda)
- Changbao Wu (2003) Biometrika
  - Optimal calibration estimators in survey sampling
- Montanari & Ranalli (2005) JASA
  - Nonparametric model calibration
  - Neural network learning and local polynomial smoothing
- Chandra & Chambers (2008)
  - CSSM Working Paper 10-08
  - Model-based framework
- Rueda, Sánchez-Borrego, Arcos and Martínez (2010)
  - distribution function using nonparametric regression
- Lehtonen & Veijanen (2012 JISAS, 2015 Wiley)
  - Model calibration in estimation of poverty indicators
  - Logistic mixed model

Estimators for domain totals

Domain totals $t_d = \sum_{k \in U_d} y_k$

Horvitz-Thompson (1952) estimator

$$\hat{t}_{dHT} = \sum_{k \in s_d} a_k y_k$$

$a_k = 1/\pi_k$ are design weights

$s_d = s \cap U_d$ sample for domain *d*

Calibration estimators

$$\hat{t}_d = \sum_{k \in s_d} w_k y_k$$

w_k are **method - specific calibration weights**

Calibration: Technical treatment

The calibration weights w_i minimize

$$\sum_{i \in s_d} \frac{\left(w_i - a_i\right)^2}{a_i} - \boldsymbol{\lambda}' \left(\sum_{i \in s_d} w_i \mathbf{z}_i - \sum_{i \in U_d} \mathbf{z}_i \right)$$

where $s_d = s \cap U_d, \quad a_i = 1/\pi_i$

$\mathbf{z}_i$ is method-specific vector of calibration variables

The calibrated weights are defined in:

$w_k = a_k(1 + \boldsymbol{\lambda}'\mathbf{z}_k)$, where $\boldsymbol{\lambda}$ is the Lagrange coefficient

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

Model-free calibration equation

$$\sum_{i \in s_d} w_i^{MFC} \mathbf{z}_i = \sum_{i \in U_d} \mathbf{z}_i = \left(N_d, \sum_{i \in U_d} x_{1i}, \ldots, \sum_{i \in U_d} x_{pi} \right)'$$

where $\mathbf{z}_i = (1, x_{1i}, \ldots, x_{pi})', \quad s_d = s \cap U_d$

NOTE: Multi-purpose weighting

No explicit model statement (linear model assumed)

Calibration of x-variable totals at the domain level

Coherence property is met

MFC estimators of domain totals are of **direct type**, when using domain-level x-totals

Model calibration equation

$$\sum_{i \in s_d} w_i^{MC} \mathbf{z}_i = \sum_{i \in U_d} \mathbf{z}_i = \left(N_d, \sum_{i \in U_d} \hat{y}_i \right)'$$

where $\mathbf{z}_i = (1, \hat{y}_i)'$, $\hat{y}_i = f(\mathbf{x}_i'(\hat{\boldsymbol{\beta}} + \hat{\mathbf{u}}_d))$ (e.g. GLMM)

NOTE: Single-purpose weighting

Separate modelling for every y-variable

Calibration of y-prediction totals at the domain level

Coherence property for x-variable totals is not met

MC estimators of domain totals are of **semi-direct** type

- modelling for the whole sample
- calibration at the domain level

Hybrid calibration equation

$$\sum_{i \in s_d} w_i^{HC} \mathbf{z}_i = \sum_{i \in U_d} \mathbf{z}_i = \left(N_d, \sum_{i \in U_d} x_{1i}, \ldots, \sum_{i \in U_d} x_{pi}, \sum_{i \in U_d} \hat{y}_i \right)'$$

where $\mathbf{z}_i = (1, x_{1i}, \ldots, x_{pi}, \hat{y}_i)'$

NOTE: Single-purpose weighting

x-variables in model part and calibration part can coincide or partially overlap or they can be separate variables

-Calibration of y-prediction totals at the domain level

-Calibration of selected x-variable totals at the domain level

-Coherence property for selected x-variable totals is met

HC estimators of domain totals are of semi-direct type

Simulation experiment

Synthetic population U of one million elements

$D = 90$ domains of interest

Domain size N_d in domain U_d determined by $exp(C)$, $C \sim Uniform(2,5)$

45 domains with linear structure: $y = 1.5 - x$

45 domains with quadratic structure: $y = 1 + 2(x - 0.5)^2$

x generated from $Beta(2,5)$

Errors $\varepsilon \sim N(0, 0.1^2)$ (added to y)

Sampling: 1,000 SRSWOR samples of size $n = 4000$ elements

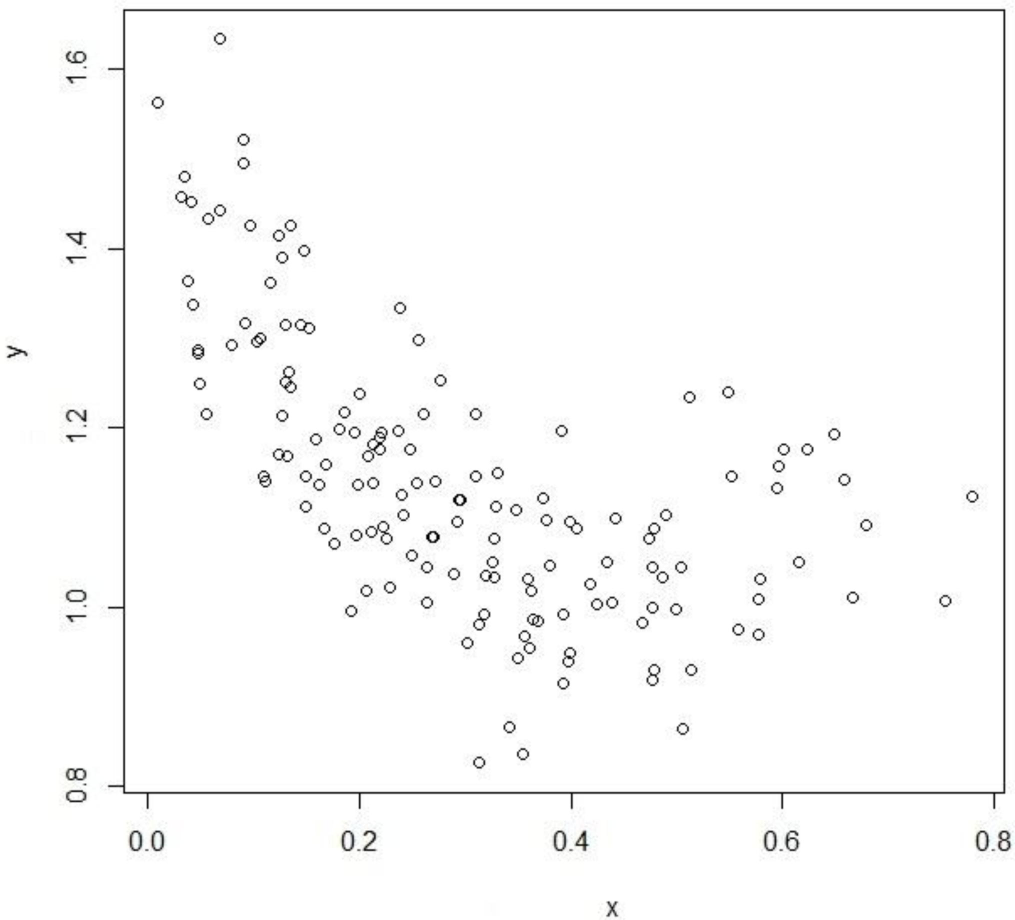
Models fitted to the sample data sets:

$$g_k = \log(y_k) = \beta_0 + \beta_1 x_k + \varepsilon_k$$

$$g_k = \log(y_k) = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + \varepsilon_k,$$

NOTE: Predictions $\hat{y}_k = \exp(\hat{g}_k)$ are needed for every $k \in U$

Population structure in domain 10



Quality measure of estimators

- **Accuracy of domain total estimator $\hat{t}_d$**
  - Relative root mean squared error RRMSE (%)
  - Medians calculated over domain sample size classes

$$RRMSE(\hat{t}_d) = \sqrt{\frac{1}{1000}\sum_{k=1}^{1000}(\hat{t}_d(s_k) - t_d)^2} \,/\, t_d$$

NOTE: All methods considered are (nearly) design unbiased

Maximum of median absolute relative bias ARB in small domains over simulations = 0.5 %

Table 1 Median relative root mean squared error (RRMSE) (%) of estimators of domain totals over domain sample size classes.
Population with **quadratic structure**

Estimator	Assisting model & domain-level calibration scheme	Expected domain sample size		
		Minor <20	Medium 20-50	Major >50
Direct estimators				
HT	None	27.2	18.1	10.4
Model-free calibration	$\log(y_k) = \beta_0 + \beta_1 x_k + \varepsilon_k$ Calibration: $\mathbf{z}_k = (1, x_k)'$	3.07	2.03	1.10
	$\log(y_k) = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + \varepsilon_k$ Calibration: $\mathbf{z}_k = (1, x_k, x_k^2)'$	4.75	1.74	0.93
Semi-direct estimators				
Model calibration	$\log(y_k) = \beta_0 + \beta_1 x_k + \varepsilon_k$ Calibration: $\mathbf{z}_k = (1, \hat{y}_k)'$	2.94	1.95	1.06
	$\log(y_k) = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + \varepsilon_k$ Calibration: $\mathbf{z}_k = (1, \hat{y}_k)'$	2.72	1.80	0.98
Hybrid calibration	$\log(y_k) = \beta_0 + \beta_1 x_k + \beta_2 x_k^2 + \varepsilon_k$ Calibration: $\mathbf{z}_k = (1, x_k, \hat{y}_k)'$	4.66	1.73	0.93

Summary

- Calibration improves accuracy substantially over the HT
- Semi-direct model calibration offers a safe choice over direct model-free calibration
  - protection against instability of model-free calibration due to small domain sample size and model misspecification
- Model calibration with more adequate model outperforms model calibration with less adequate model
- Hybrid calibration offers a realistic compromise especially under coherence requirements
  - Suffers from instability of model-free calibration if domain sample size is “too small”

References

- Deville J.-C. and Särndal C.-E. (1992) Calibration estimators in survey sampling. *JASA* 87, 376-382.
- Estevao V. M. and Särndal C.-E. (1999) The use of auxiliary information in design-based estimation for domains. *Survey Methodology* 2, 213-221.
- Chandra H. and Chambers R. (2008) Small area estimation under transformation to linearity. CSSM, University of Wollongong, Working Paper 10-08, 2008.
- Lehtonen R. and Veijanen A. (2009) Design-based methods of estimation for domains and small areas. Chapter 31 in Rao C. R. and Pfeffermann D. (Eds.) *Handbook of Statistics Vol. 29B. Sample Surveys. Inference and Analysis*. Amsterdam: 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 R. and Veijanen A. (2014) Small area estimation of poverty rate by model calibration and "hybrid" calibration. NORDSTAT Conference, Turku, June 2014
- Lehtonen R. and Veijanen A. (2015a) 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. (2015b) Small area estimation by model calibration and "hybrid" calibration. The NTTS Conference, Brussels, February 2015.
- Montanari G. E. and Ranalli M. G. (2005) Nonparametric model calibration estimation in survey sampling. *JASA* 100, 1429–1442.

References

- 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.
- Rueda M., Sánchez-Borrego I., Arcos A. and Martínez S. (2010). Model-calibration estimation of the distribution function using nonparametric regression. *Metrika,* 71, 33–44.
- 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)
- Wu C. (2003) Optimal calibration estimators in survey sampling. *Biometrika* 90, 937–951.

Thank you for your attention