Statistics Finland

Effect of register errors on quality of survey estimates

Ari Veijanen August 2015
Helsinki

Outline

- Study variable from sample survey
  - combined with other data from registers (personal ID)
  - are registers reliable?
- Regional means
  - errors in auxiliary register variables
  - sensitivity of estimators
- Class means of the survey variable
  - classes defined in the register
  - misclassification causes bias

What is a register?

- Administrative register
  - unit-level data
  - maintained by state administration, for example
  - updated whenever a change occurs in population
    - e.g. information from a person
- Statistical register
  - constructed from administrative registers
    - screening and editing
  - applied in this study

Errors in register

- Registers are usually perceived as **reliable**
- **But** error rates of 10 % have been observed
- Literature: Wallgren & Wallgren, Zhang, Zhang & Fosen
- Typical errors in register
  - coding errors
  - reporting errors by a person
  - delayed update
    - may cause misclassification

Statistics Finland

Point of view

- The effect of register errors on estimators
- Compare two sets of estimates
  - results with errors in register
  - results with error-free register (inaccessible)

Regional means involving auxiliary variables

- Estimators incorporating auxiliary variables
  - ordinary calibration without a model
  - model-assisted and model-dependent methods
- Problem: auxiliary register variable X contains errors
- “Contamination" produces outliers
  - observed $X^* = X + e$ e independent of Y
- Regression model:

$$\hat{\beta}_{X^*} \to 0, \text{ as } \mathrm{Var}(e) \to \infty$$

- Effects on calibration, GREG, EBLUP?

Impact of misclassification on class means

- Means of response Y over classes of a register variable
  - example: small area estimation with errors in area codes
- Misclassification:
  - class label C* not always same as true class C
- Estimator of class mean is biased
- How large can this bias be?

Domain estimators of population means

- (1) No auxiliary data, domain sizes known
  - HT, Hajek
- (2) Auxiliary data from registers
  - (a) no explicit model in model-free "ordinary" calibration
  - (b) model fitted to whole sample
    - GREG
    - EBLUP
    - model calibration

Notation

- Domain in sample s_d
- Domain in population U_d
- Domain size in population N_d
- Design weights a_k

Definitions of domain estimators of means

- HT $$\hat{\hat{Y}}_{d;HT} = \frac{1}{N_d}\sum_{k\in s_d} a_k y_k$$
- Hajek $$\hat{\hat{Y}}_{d;Hajek} = \frac{\sum_{k\in s_d} a_k y_k}{\sum_{k\in s_d} a_k}$$
- GREG $$\hat{\hat{Y}}_{d;GREG} = \frac{1}{N_d}\left(\sum_{k\in U_d} \hat{y}_k + \sum_{k\in s_d} a_k\left(y_k - \hat{y}_k\right)\right)$$
- EBLUP $$\hat{\hat{Y}}_{d;EBLUP} = \frac{1}{N_d}\left(\sum_{k\in U_d - s_d} \hat{y}_k + \sum_{k\in s_d} y_k\right).$$

Model-free domain level calibration

– Estimator $\hat{Y}_{d;CAL} = \frac{1}{N_d} \sum_{k \in s_d} w_{dk} y_k .$

– Conditions on weights

  – calibration equation

$$\sum_{k \in s_d} w_{dk} \begin{pmatrix} 1 \\ x_{1k} \\ \vdots \\ x_{pk} \end{pmatrix} = \sum_{k \in U_d} \begin{pmatrix} 1 \\ x_{1k} \\ \vdots \\ x_{pk} \end{pmatrix}$$

  – minimize distance to design weights $\sum_{k \in s_d} \frac{(w_{dk} - a_k)^2}{a_k}$

Model-free calibration weights

$$w_{dk} = a_k \left(1 + \left(\sum_{i \in U_d} \mathbf{x}_i - \sum_{i \in s_d} a_i \mathbf{x}_i \right)' \left(\sum_{i \in s_d} a_i \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \mathbf{x}_k \right)$$

Model calibration

- Model predictions instead of auxiliary variables
- Calibration equations

$$\sum_{k \in s_d} w_{dk} \begin{pmatrix} 1 \\ \hat{y}_k \end{pmatrix} = \sum_{k \in U_d} \begin{pmatrix} 1 \\ \hat{y}_k \end{pmatrix}.$$

Simulation experiments

- Auxiliary data in a synthetic register
  - continuous X,Z
  - categorical C
  - 40 regions D
- Response Y depends on the values of X, Z and C
  - mixed model: regional random intercepts, random slopes
- Errors in X and C are generated after creating Y
- Design-based simulation: 1000 SRSWOR samples
  - model fitted: mixed model, regional random effects

Experiment 1. Effects of contamination

- Contaminate 1% randomly chosen units in the population

  $$X_k^* = X_k + M$$

  - M=20 or M=500 (note: X* ranges from -10 to 26)
- Estimation uses X*,Z and C
- Sample size n=4000

Statistics Finland

Absolute relative bias (ARB)

- Domain estimates from 1000 simulated samples
  - estimated bias

$$bias = (mean\ of\ estimates) - (true\ value)$$

  - absolute relative bias $ARB = \left| \frac{bias}{true\ value} \right|$

Statistics Finland

Mean squared error (MSE)

$$\mathrm{MSE}\left(\hat{\theta}\right) = \frac{1}{1000}\sum_{k=1}^{1000}\left(\hat{\theta}\left(s_k\right) - \theta\right)^2$$

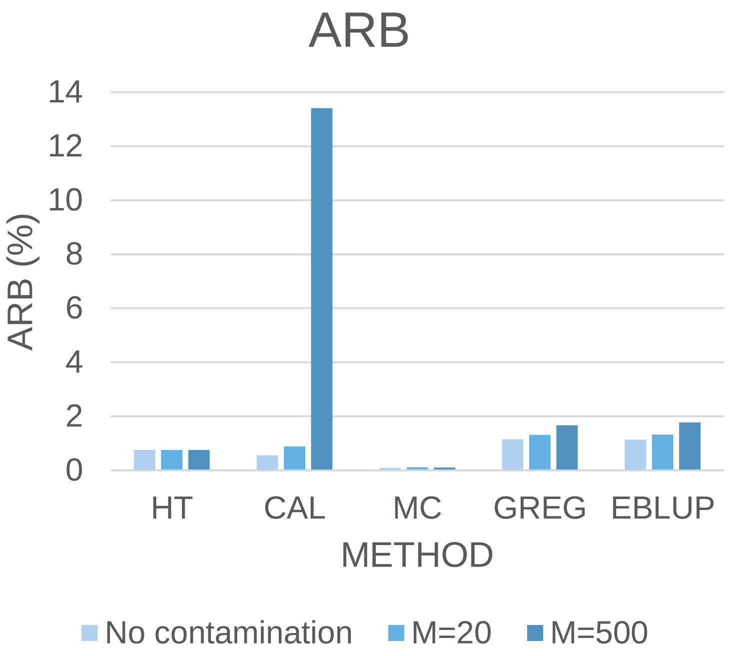
- Relative root mean squared error (RRMSE)

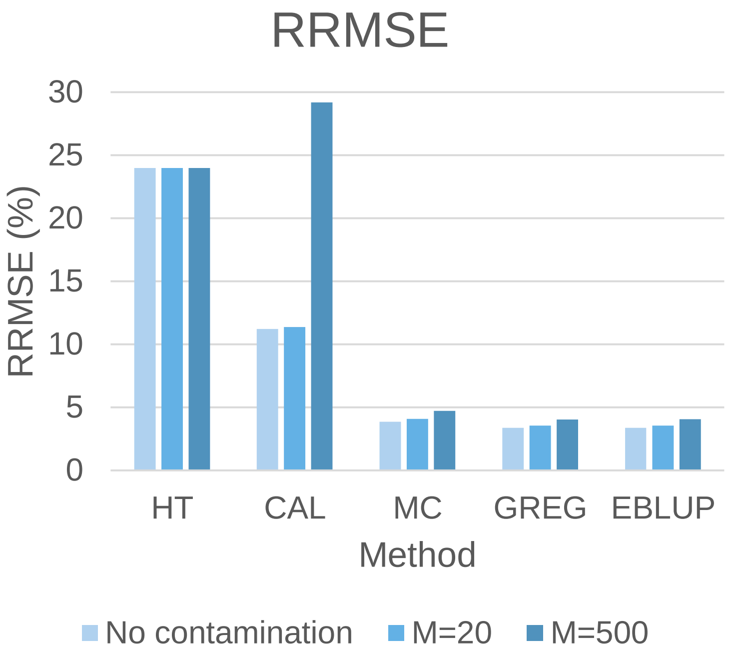
$$\mathrm{RRMSE}\left(\hat{\theta}\right) = \frac{\sqrt{\mathrm{MSE}\left(\hat{\theta}\right)}}{\theta}$$

Averages over a domain size class

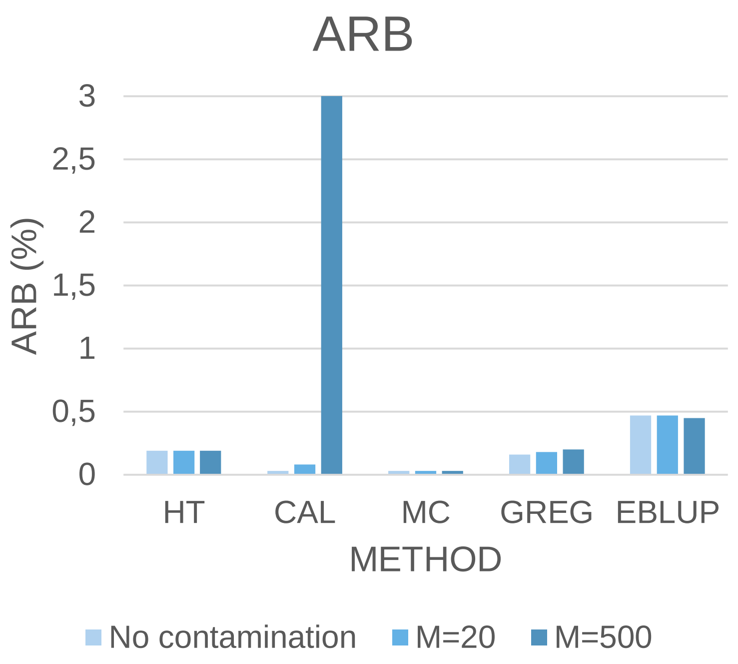
- Averages of ARB and RRMSE calculated over
  - small domains (expected sample size smaller than 30)
  - large domains (larger than 100)

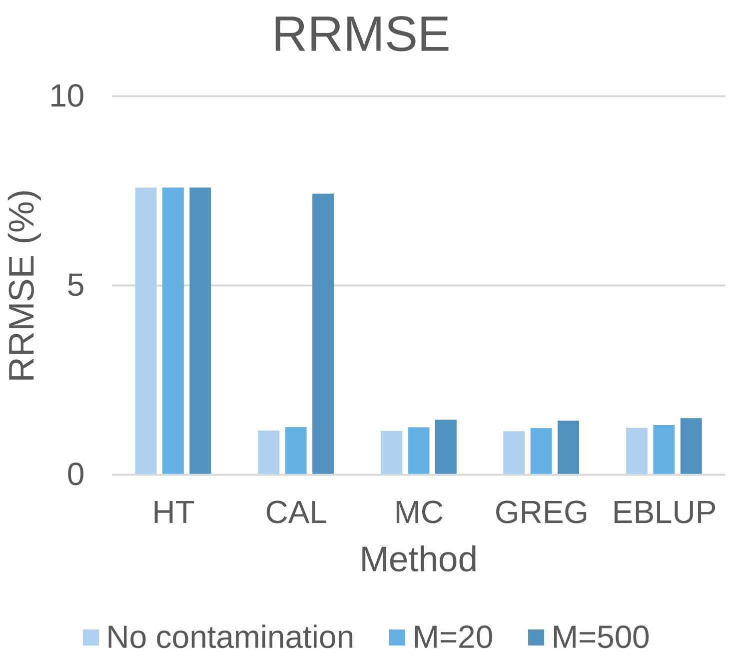
Effects of contamination in small domains





Effects of contamination in large domains





Comparison of methods

- HT is not affected (no auxiliary data)
- Model-free calibration is sensitive (direct estimator)
- Model calibration less sensitive (indirect estimator)
- GREG remains design unbiased
- MSE of GREG and EBLUP increases slightly

Effect of contamination on GREG and EBLUP

- Extreme contamination with M=500
- Estimated slope for X* close to zero
- Most predictions almost as in a model that excludes X*

$$\text{GREG}(X^*,Z,C) \approx \text{GREG}(Z,C)$$

$$\text{EBLUP}(X^*,Z,C) \approx \text{EBLUP}(Z,C)$$

Robust methods could be used

- Effect of outliers is reduced
  - robust EBLUP estimator (Sinha and Rao, 2009)
  - robust GREG (Lee and Patak, 1998)
- These handle outliers in both Y and in X
- Not commonly found in statistical packages

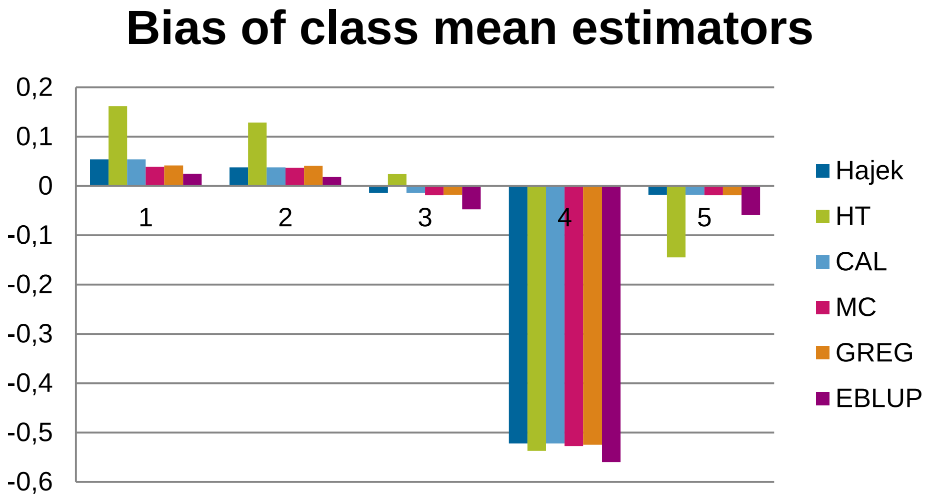
Experiment 2. Effects of misclassification on class means

– Means of Y in classes of C

Class	1	2	3	4	5
Share (%)	6.7	13.3	20.0	26.7	33.3
Mean of Y	17.3	23.2	28.8	34.8	40.5

Biased class mean estimators

- 10% of units in class 2 classified to class 4
- Observed mean in class 4 decreases, relative bias -1.5%



Upper bound for bias

- Impose superpopulation model
  - random variables: response Y, true class C, observed class C*
  - assume that classification does not depend on Y given C
- Estimator for class mean converges to E(Y|C*), not E(Y|C)
- Asymptotic bias E(Y|C*)-E(Y|C)

Difference between expectations

$$\left|E\left(Y \mid C^* = b\right) - E\left(Y \mid C = b\right)\right|$$

$$\leq \left(1 - P\left\{C = b \mid C^* = b\right\}\right)\max_i \left|\mu_i - \mu_b\right| \qquad \left(\mu_i = E\left(Y \mid C = i\right)\right)$$

$$\left(1 - P\left\{C = b \mid C^* = b\right\}\right) \leq 1 - \frac{1}{Q \max_t \dfrac{\pi_t p_{tb}}{\pi_b p_{bb}}}$$

- Q classes
- Class probabilities (proportions) $\pi_t = P\left\{C = t\right\}$
- Classification probabilities $p_{tb} = P\left\{C^* = b \mid C = t\right\}$

Statistics Finland

Applying the equation in experiment

- Plug in true values of probabilities and $\mu_2 - \mu_4$
  - upper bound 0.549
  - observed absolute bias 0.537

Approximations in practice

- Preliminary approximations
  - Subjective estimate of misclassification probability (like 0.01)
  - Plug in class proportions
  - Plug in maximum difference of class means of Y
- Later
  - upper bound for $\max_i |\mu_i - \mu_b|$ that holds with probability 0.99

Statistics Finland

References

- **Fuller, W. A. (1987).** Measurement error models. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons. New York.
- **Lee, H. and Patak, Z. (1998).** Outlier robust generalized regression estimator. Proceedings of the Survey Methods Section. SSC Annual Meeting, June 1998.
- **Lehtonen, R. & Veijanen, A. (2012)**. Small area poverty estimation by model calibration. Journal of the Indian Society of Agricultural Statistics, 66, 125-133.

References 2

- **Sinha, S. K. and Rao, J. N. K. (2009).** Robust small area estimation. Can. J. Stat. 37, 381-399.
- **Wallgren, A. & Wallgren, B. (2014)**. Register-based statistics: statistical methods for administrative data. 2nd edition.
- **Zhang, L.-C. (2011).** A unit-error theory for register-based household statistics. Journal of Official Statistics, 27, 415-432.
- **Zhang, L.-C. & Fosen, J. (2012).** A modeling approach for uncertainty assessment of register-based small area statistics. Journal of the Indian Society of Agricultural Statistics, 66, 91-104.

Thank You!

Ari Veijanen ari.veijanen@stat.fi

BaNoCoSS 2015