# Estimation strategy for small areas, a case study

Vilma Nekrašaitė – Liegė[1]

[1]Vilniaus Gedimino Technical University,

Workshop on Survey Sampling Theory and Methodology,
Valmiera, August 24-28, 2012

## Purpose

In this research

I am interesting in an overall **strategy**

that deals with **small area** problems,

involving both planning **sample design**

and **estimation** aspects.

## Outline

- Notations;
- Sample designs: balanced sample, model-based design;
- Estimators and Models;
- Simulation;
- Conclusions.

## Notations

Finite population: $U = \{1, 2, ..., N\}$;

## Notations

Finite population: $U = \{1, 2, ..., N\}$;

Domain population: $U^{(d)} = \{1, 2, ..., N^{(d)}\}$, $d = 1, ..., D$;

## Notations

Finite population: $U = \{1, 2, ..., N\}$;

Domain population: $U^{(d)} = \{1, 2, ..., N^{(d)}\}$, $d = 1, ..., D$;

Study variable: $y(t)$;

Auxiliary variables: $\mathbf{x}(t) = \{x_1(t), x_2(t), \ldots, x_J(t)\} \in \mathbb{R}^J$;

## Notations

Finite population: $U = \{1, 2, ..., N\}$;

Domain population: $U^{(d)} = \{1, 2, ..., N^{(d)}\}$, $d = 1, ..., D$;

Study variable: $y(t)$;

Auxiliary variables: $\mathbf{x}(t) = \{x_1(t), x_2(t), \ldots, x_J(t)\} \in \mathbb{R}^J$;

**Study parameter - domain population total:**

$$TOT^{(d)}(t) = \sum_{k \in U^{(d)}} y_k(t) = \sum_{k \in U} q_k^{(d)} y_k(t), \quad d = 1, \ldots, D; \quad t = 1, 2, \ldots \quad (1)$$

## Notations

Sample: $\mathbf{S}(t) = (S_1(t), S_2(t), \ldots, S_N(t))$;

## Notations

Sample: $\mathbf{S}(t) = (S_1(t), S_2(t), \ldots, S_N(t))$;

The sample set: $s(t) = \{k : k \in U, S_k(t) \geq 1\}$;

## Notations

Sample: $\mathbf{S}(t) = (S_1(t), S_2(t), \ldots, S_N(t))$;

The sample set: $s(t) = \{k : k \in U, S_k(t) \geq 1\}$;

Inclusion probability for unit $k$: $\pi_k(t) = \mathbf{P}(k \in s(t))$.

## Balanced sample

A sample is said to be balanced if, for a vector of auxiliary variable $\mathbf{z}(t) = \{z_1(t), z_2(t), \ldots, z_L(t)\} \in \mathbb{R}^L$,

$$\sum_{k \in s(t)} \frac{\mathbf{z}_k(t)}{\pi_k(t)} = \sum_{k \in U} \mathbf{z}_k(t). \tag{2}$$

## Balanced sample

A sample is said to be balanced if, for a vector of auxiliary variable $\mathbf{z}(t) = \{z_1(t), z_2(t), \ldots, z_L(t)\} \in \mathbb{R}^L$,

$$\sum_{k \in s(t)} \frac{\mathbf{z}_k(t)}{\pi_k(t)} = \sum_{k \in U} \mathbf{z}_k(t). \tag{2}$$

## Cases of the balanced samples

1. Sampling with a fixed sample size if $\mathbf{z}(t) = \pi_k(t)$

$$\sum_{k \in s(t)} \frac{\pi_k(t)}{\pi_k(t)} = \sum_{k \in s(t)} 1 = \sum_{k \in U} \pi_k(t); \tag{3}$$

## Balanced sample

A sample is said to be balanced if, for a vector of auxiliary variable $\mathbf{z}(t) = \{z_1(t), z_2(t), \ldots, z_L(t)\} \in \mathbb{R}^L$,

$$\sum_{k \in s(t)} \frac{\mathbf{z}_k(t)}{\pi_k(t)} = \sum_{k \in U} \mathbf{z}_k(t). \tag{2}$$

## Cases of the balanced samples

1. Sampling with a fixed sample size if $\mathbf{z}(t) = \pi_k(t)$

$$\sum_{k \in s(t)} \frac{\pi_k(t)}{\pi_k(t)} = \sum_{k \in s(t)} 1 = \sum_{k \in U} \pi_k(t); \tag{3}$$

2. Stratified sample if $\mathbf{z}(t) = \delta_{kh} = \begin{cases} 1, & \text{if } k \in U_h; \\ 0, & \text{otherwise.} \end{cases}$

## Cube method (Devile and Tille 1998, 2004)

## Cube method (Devile and Tille 1998, 2004)

1. Cube method is general in the sense that the inclusion probabilities are exactly satisfied, that these probabilities may be equal or unequal and that the sample is as balanced as possible;

## Cube method (Devile and Tille 1998, 2004)

1. Cube method is general in the sense that the inclusion probabilities are exactly satisfied, that these probabilities may be equal or unequal and that the sample is as balanced as possible;
2. It is possible to get SAS/IML version of Cube method done by Chauvet and Tille (2006) and it is also available on the University of Neuchatel Web site. This software program is free, available over the Internet and is easy to use;

## Balanced sample might not help to improve estimation

1. If the sample allocation is small in some strata, balanced sampling will be only very approximate;

## Balanced sample might not help to improve estimation

1. If the sample allocation is small in some strata, balanced sampling will be only very approximate;
2. If we can reasonably assume that the balancing variables are no longer correlated to the variables of interest. This can occur when the balancing and the variables used in estimation stage are the same variables measured at different moments.

## Balanced sample might not help to improve estimation

1. If the sample allocation is small in some strata, balanced sampling will be only very approximate;
2. If we can reasonably assume that the balancing variables are no longer correlated to the variables of interest. This can occur when the balancing and the variables used in estimation stage are the same variables measured at different moments.

## Model - based design (Nekrašaitė-Liegė, Radavičius, Rudys, 2011)

## Balanced sample might not help to improve estimation

1. If the sample allocation is small in some strata, balanced sampling will be only very approximate;
2. If we can reasonably assume that the balancing variables are no longer correlated to the variables of interest. This can occur when the balancing and the variables used in estimation stage are the same variables measured at different moments.

## Model - based design (Nekrašaitė-Liegė, Radavičius, Rudys, 2011)

The suggested model-based sample design consists of three steps:

1. Model construction and estimation of it's coefficients;
2. Estimation of the variance of the prediction error;
3. Construction of the sample design $p(.)$.

## Estimators:

### Horvitz-Thompson (H-T) estimator

$$\widehat{TOT}^{(d)}_{H-T} = \sum_{k \in s(t) \cap U^{(d)}} \frac{y_k(t)}{\pi_k(t)}; \tag{4}$$

### Generalized regression (GREG) estimator

$$\widehat{TOT}^{(d)}_{GREG-\mathcal{M}_l}(t) = \sum_{k \in U^{(d)}} \hat{y}_k(t) + \sum_{k \in s(t) \cap U^{(d)}} \frac{(y_k(t) - \hat{y}_k(t))}{\pi_k(t)} \quad l = 1, \ldots 7. \tag{5}$$

## Fixed effect models

$$\mathcal{M}_1 : Y_k(W) = \beta_0(W) + \sum_{j=1}^{J} \beta_j(W) X_{j,k}(W) + \varepsilon_k(W), \quad k \in U; \tag{6}$$

$$\mathcal{M}_2 : Y_k(W) = \beta_{0,g(k)}(W) + \sum_{j=1}^{J} \beta_j(W) X_{j,k}(W) + \varepsilon_k(W), \quad k \in U;$$

$$\mathcal{M}_3 : Y_k(t) = \beta_{0,g(k)} + \sum_{j=1}^{J} \beta_{j,g(k)} X_{j,k}(t) + \varepsilon_k(t), \quad k \in U;$$

$$\mathcal{M}_4 : Y_k(t) = \beta_{0,g(k)} + a_0^{(d)} t + \mathbf{a}'^{(d)} \alpha(t) + \sum_{j=1}^{J} \beta_{j,g(k)} x_{j,k}(t) + \varepsilon_k(t), \quad k \in U.$$

## Random effect models

$$\mathscr{M}_5 : Y_k(W) = \beta_0(W) + r_{0,g(k)}(W) + \sum_{j=1}^{J} \beta_j(W) X_{j,k}(W) + \varepsilon_k(W), \quad k \in U; \tag{7}$$

$$\mathscr{M}_6 : Y_k(t) = \beta_0^{(d)} + r_{0,g(k)} + \sum_{j=1}^{J} \beta_j^{(d)} X_{j,k}(t) + \varepsilon_k(t), \quad k \in U;$$

$$\mathscr{M}_7 : Y_k(t) = \beta_0^{(d)} + r_{0,g(k)} + a_0^{(d)} t + \mathbf{a}'^{(d)} \alpha(t) + \sum_{j=1}^{J} \beta_j^{(d)} x_{j,k}(t) + \varepsilon_k(t), \quad k \in U.$$

## Simulation

| | |
|---|---|
| Population: | Lithuanian survey on short-term statistics on service ($N = 750$); |
| Time: | each quarter from 2005 till 2009; |
| Study variable $y(t)$: | income; |
| Auxiliary variables $\mathbf{x}(t)$: | number of employees ($x_1(t)$), VAT ($x_2(t)$), NACE code, region indicators ($x_{3,1}(t), x_{3,2}(t), \ldots, x_{3,11}(t)$); |

## Simulation

Parameter of interest $TOT^{(d)}(t)$: total income in domain $(d = 8 \times (10+5) = 120)$;

## Simulation

Parameter of interest $TOT^{(d)}(t)$: total income in domain ($d = 8 \times (10+5) = 120$);

Sample designs: SRS ($n = 300$, $M = 1000$);
SSRS ($n = 300$, $M = 1000$), strata - size of enterprise;
M-B ($n = 300$, $M = 1000$);

## Simulation

| | |
|---|---|
| Parameter of interest $TOT^{(d)}(t)$: | total income in domain ($d = 8 \times (10 + 5) = 120$); |
| Sample designs: | SRS ($n = 300$, $M = 1000$); |
| | SSRS ($n = 300$, $M = 1000$), strata - size of enterprise; |
| | M-B ($n = 300$, $M = 1000$); |
| Balanced variables: | $\pi_k(t), x_{1k}(t), x_{2k}(t)$; |

## Simulation

| | |
|---|---|
| Parameter of interest $TOT^{(d)}(t)$: | total income in domain ($d = 8 \times (10+5) = 120$); |
| Sample designs: | SRS ($n = 300$, $M = 1000$); |
| | SSRS ($n = 300$, $M = 1000$), strata - size of enterprise; |
| | M-B ($n = 300$, $M = 1000$); |
| Balanced variables: | $\pi_k(t), x_{1k}(t), x_{2k}(t)$; |
| Models: | $\mathcal{M}_1$ - $\mathcal{M}_7$; |
| Estimators: | H-T, GREG-$\mathcal{M}_l$ l=1, ... 7. |

## Research 1: Accuracy measures

Absolute relative bias:

$$ARB = \frac{\left| \frac{1}{M} \sum_{m=1}^{M} \widehat{TOT}^{(d)(m)}_{GREG-\mathscr{M}_l}(t) - TOT^{(d)} \right|}{TOT^{(d)}}; \quad (8)$$

Relative root means square error:

$$RRMSE = \frac{\sqrt{\frac{1}{M} \sum_{m=1}^{M} (\widehat{TOT}^{(d)(m)}_{GREG-\mathscr{M}_l}(t) - TOT^{(d)})^2}}{TOT^{(d)}}. \quad (9)$$

$\widehat{TOT}^{(d)(m)}_{GREG-\mathscr{M}_l}(t)$ - replicates of the estimates $\widehat{TOT}^{(d)}_{GREG-\mathscr{M}_l}(t)$, $m = 1, ..., M$, $l = 1, \ldots 7$.

# Simulation results1: SRS, population case

| Estimator | Sample design, balanced variables | | | | | |
|---|---|---|---|---|---|---|
| | SRS, $\pi_k(t)$ | | SRS, $\pi_k(t), x_{1k}(t)$ | | SRS, $\pi_k(t), x_{1k}(t), x_{2k}(t)$ | |
| | *MARB*,% | *MRRMSE*,% | *MARB*,% | *MRRMSE*,% | *MARB*,% | *MRRMSE*,% |
| H-T | 0.5 | 14.7 | 0.5 | 11.4 | 0.5 | 9.8 |
| GREG-$\mathcal{M}_1$ | 0.4 | 7.0 | 0.4 | 7.0 | 0.4 | 6.7 |
| GREG-$\mathcal{M}_2$ | 0.4 | 7.0 | 0.4 | 7.0 | 0.4 | 6.7 |
| GREG-$\mathcal{M}_3$ | 0.2 | 6.0 | 0.3 | 6.1 | 0.3 | 5.9 |
| GREG-$\mathcal{M}_4$ | 0.2 | 6.2 | 0.4 | 6.5 | 0.3 | 6.2 |
| GREG-$\mathcal{M}_5$ | 0.3 | 6.9 | 0.4 | 6.8 | 0.3 | 6.5 |
| GREG-$\mathcal{M}_6$ | 0.3 | 6.3 | 0.4 | 6.5 | 0.2 | 6.2 |
| GREG-$\mathcal{M}_7$ | 0.2 | 6.2 | 0.4 | 6.5 | 0.3 | 6.2 |

# Simulation results1: SRS, population case

| Estimator | Sample design, balanced variables | | | | | |
|---|---|---|---|---|---|---|
| | SRS, $\pi_k(t)$ | | SRS, $\pi_k(t), x_{1k}(t)$ | | SRS, $\pi_k(t), x_{1k}(t), x_{2k}(t)$ | |
| | *MARB*, % | *MRRMSE*, % | *MARB*, % | *MRRMSE*, % | *MARB*, % | *MRRMSE*, % |
| H-T | 0.5 | 14.7 | 0.5 | 11.4 | 0.5 | 9.8 |
| GREG-$\mathscr{M}_1$ | 0.4 | 7.0 | 0.4 | 7.0 | 0.4 | 6.7 |
| GREG-$\mathscr{M}_2$ | 0.4 | 7.0 | 0.4 | 7.0 | 0.4 | 6.7 |
| GREG-$\mathscr{M}_3$ | 0.2 | 6.0 | 0.3 | 6.1 | 0.3 | 5.9 |
| GREG-$\mathscr{M}_4$ | 0.2 | 6.2 | 0.4 | 6.5 | 0.3 | 6.2 |
| GREG-$\mathscr{M}_5$ | 0.3 | 6.9 | 0.4 | 6.8 | 0.3 | 6.5 |
| GREG-$\mathscr{M}_6$ | 0.3 | 6.3 | 0.4 | 6.5 | 0.2 | 6.2 |
| GREG-$\mathscr{M}_7$ | 0.2 | 6.2 | 0.4 | 6.5 | 0.3 | 6.2 |

## Simulation results2: GREG-$\mathscr{M}_3$ estimator

| Sample design, balanced variables | Domain sample size classes | | | | | |
|---|---|---|---|---|---|---|
| | Small 0 – 9 | | Medium 10 – 19 | | Large 20 – ... | |
| | *MARB*, % | *MRRMSE*, % | *MARB*, % | *MRRMSE*, % | *MARB*, % | *MRRMSE*, % |
| SRS, $\pi_k(t)$ | 1.7 | 43.1 | 1.2 | 14.4 | 0.6 | 10.4 |
| SRS, $\pi_k(t), x_{1k}(t)$ | 1.7 | 43.4 | 1.2 | 14.6 | 0.4 | 10.4 |
| SRS, $\pi_k(t), x_{1k}(t), x_{2k}(t)$ | 1.8 | 41.6 | 0.7 | 14.4 | 0.4 | 10.0 |
| SSRS, $\pi_k(t)$ | 1.8 | 18.1 | 0.6 | 11.6 | 0.2 | 5.5 |
| SSRS, $\pi_k(t), x_{1k}(t)$ | 1.8 | 18.3 | 0.5 | 11.2 | 0.2 | 5.5 |
| SSRS, $\pi_k(t), x_{1k}(t), x_{2k}(t)$ | 1.8 | 17.9 | 0.6 | 11.0 | 0.4 | 5.4 |
| M-B, $\pi_k(t), x_{1k}(t), x_{2k}(t)$ | 1.6 | 17.0 | 0.9 | 10.5 | 0.3 | 5.1 |

# Simulation results3: Model-based sample design

| Model for sample design and estimator | Domain sample size classes | | | | | |
|---|---|---|---|---|---|---|
| | Small 0 – 9 | | Medium 10 – 19 | | Large 20 – ... | |
| | *MARB*, % | *MRRMSE*, % | *MARB*, % | *MRRMSE*, % | *MARB*, % | *MRRMSE*, % |
| $\mathscr{M}_3$ | 1.6 | 17.0 | 0.9 | 10.5 | 0.3 | 5.1 |
| $\mathscr{M}_4$ | 1.5 | 16.7 | 0.9 | 10.4 | 0.4 | 5.1 |
| | | | | | | |
| $\mathscr{M}_6$ | 1.6 | 17.1 | 0.9 | 10.5 | 0.4 | 5.2 |
| $\mathscr{M}_7$ | 1.6 | 16.9 | 0.9 | 10.5 | 0.4 | 5.1 |

# Simulation results3: Model-based sample design

| Model for sample design and estimator | Domain sample size classes | | | | | |
|---|---|---|---|---|---|---|
| | Small 0 – 9 | | Medium 10 – 19 | | Large 20 – … | |
| | *MARB*, % | *MRRMSE*, % | *MARB*, % | *MRRMSE*, % | *MARB*, % | *MRRMSE*, % |
| $\mathscr{M}_3$ | 1.6 | 17.0 | 0.9 | 10.5 | 0.3 | 5.1 |
| $\mathscr{M}_4$ | 1.5 | 16.7 | 0.9 | 10.4 | 0.4 | 5.1 |
| $\mathscr{M}_6$ | 1.6 | 17.1 | 0.9 | 10.5 | 0.4 | 5.2 |
| $\mathscr{M}_7$ | 1.6 | 16.9 | 0.9 | 10.5 | 0.4 | 5.1 |

## In my research the results showed that:

1. the choice of model has bigger effect on estimator then the number of balanced variables;

## In my research the results showed that:

1. the choice of model has bigger effect on estimator then the number of balanced variables;
2. the sample design has bigger effect on estimator then the number of balanced variables;

## In my research the results showed that:

1. the choice of model has bigger effect on estimator then the number of balanced variables;
2. the sample design has bigger effect on estimator then the number of balanced variables;
3. the sample design has bigger effect on estimator then the choice of model.

# Literature


1. Chauvet, G., Tillé, Y. (2006). A fast algorithm of balanced sampling. *Journal of Computational Statistics* **21**, 9 - 31.
2. Deville, J.-C., Tillé, Y. (1998). Unequal probability sampling without replacement through a splitting method. *Biometrika* **85**, 89 - 101.
3. Deville, J.-C., Tillé, Y. (2004). Efficient balanced sampling: The cube method. *Biometrika* **91**, 893 - 912.
4. Ghosh, M., Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science* **9**, 55 - 93.
5. Lehtonen, R., Särndal, C.-E., Veijanen, A. (2003). The effect of model choice in estimation for domains, including small domains. *Survey Methodology* **29**, 33 - 44.
6. Nekrašaitė-Liegė, V., Radavičius, M., Rudys, T. (2011). Model-based design in small area estimation. *Lithuanian mathematical journal* **51**, 417 - 424.
7. Särndal, C., Swensson, B., Wretman, J. (1992). *Model assisted survey sampling.* Springer Verlag.
8. Tillé, Y. (2001). *Thorie des sondages : chantillonnage et estimation en populations finies.* Dunod, Paris.

Thank you