AN OPTIMAL ALLOCATION FOR MULTIVARIATE SURVEYS WHEN AN AUXILIARY INFORMATION IS USED

V. Nekrašaitė-Liegė[1]

[1]Vilnius Gediminas Technical University, Lithuania

16-20 June, 2019.
The 5th Baltic-Nordic Conference on Survey Statistics
(BaNoCoSS-2019)

Notations

$\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ – a finite population of N units;

$y^{(1)}, y^{(2)}, \ldots, y^{(m)}$ – study variables;

$y^{(j)} : y_1^{(j)}, y_2^{(j)}, \ldots, y_N^{(j)}, j = 1, 2, \ldots, m$ – values of study variables;

$x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ – auxiliary variables;

$x^{(j)} : x_1^{(j)}, x_2^{(j)}, \ldots, x_N^{(j)}, j = 1, 2, \ldots, m$ – values of auxiliary variables.

Notations

$\mathcal{U} = \{u_1, u_2, \ldots, u_N\}$ – a finite population of N units;

$y^{(1)}, y^{(2)}, \ldots, y^{(m)}$ – study variables;

$y^{(j)} : y_1^{(j)}, y_2^{(j)}, \ldots, y_N^{(j)}, j = 1, 2, \ldots, m$ – values of study variables;

$x^{(1)}, x^{(2)}, \ldots, x^{(m)}$ – auxiliary variables;

$x^{(j)} : x_1^{(j)}, x_2^{(j)}, \ldots, x_N^{(j)}, j = 1, 2, \ldots, m$ – values of auxiliary variables.

Suppose the population consists of H non-overlapping strata:

$$\mathcal{U}_1, \mathcal{U}_2, \ldots, \mathcal{U}_H, \quad \mathcal{U} = \bigcup_{h=1}^{H} \mathcal{U}_h, \quad N_h, h = 1, 2, \ldots, H.$$

A simple random sample $\mathbf{s}_h \subset \mathcal{U}_h$ of size n_h is drawn without replacement in each strata. Thus:

$$\mathbf{s} = \bigcup_{h=1}^{H} \mathbf{s}_h, \quad N = \sum_{h=1}^{H} N_h, \quad n = \sum_{h=1}^{H} n_h.$$

Notations

The parameters of interest:

$$t_{y^{(j)}} = \sum_{i=1}^{N} y_i^{(j)}, \quad j = 1, 2, \ldots, m.$$

Horvitz-Thompson estimators:

$$\hat{t}_{y^{(j)}} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}^{(j)}, \quad j = 1, 2, \ldots, m,$$

The variance of $\hat{t}_{y^{(j)}}, j = 1, 2, \ldots, m,$:

$$V(\hat{t}_{y^{(j)}}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{hy^{(j)}}^2}{n_h} = \sum_{h=1}^{H} \left(\frac{N_h^2 s_{hy^{(j)}}^2}{n_h} - N_h s_{hy^{(j)}}^2 \right)$$

here $s_{hy^{(j)}}^2$ denotes the variance of $y^{(j)}$ in the stratum $\mathcal{U}_h$.

Notations

The parameters of interest:

$$t_{y^{(j)}} = \sum_{i=1}^{N} y_i^{(j)}, \quad j = 1, 2, \ldots, m.$$

Horvitz-Thompson estimators:

$$\hat{t}_{y^{(j)}} = \sum_{h=1}^{H} \frac{N_h}{n_h} \sum_{i=1}^{n_h} y_{hi}^{(j)}, \quad j = 1, 2, \ldots, m,$$

The variance of $\hat{t}_{y^{(j)}}, j = 1, 2, \ldots, m,$:

$$V(\hat{t}_{y^{(j)}}) = \sum_{h=1}^{H} N_h^2 \left(1 - \frac{n_h}{N_h}\right) \frac{s_{hy^{(j)}}^2}{n_h} = \sum_{h=1}^{H} \left(\frac{N_h^2 s_{hy^{(j)}}^2}{n_h} - N_h s_{hy^{(j)}}^2 \right)$$

here $s_{hy^{(j)}}^2$ denotes the variance of $y^{(j)}$ in the stratum $\mathcal{U}_h$.

Tasks for achieving better accuracy.

- A total cost C of the survey is minimized, in order to have coefficients of variation, for the estimates of totals of the m survey variables, below specified thresholds;

Tasks for achieving better accuracy.

- A total cost C of the survey is minimized, in order to have coefficients of variation, for the estimates of totals of the m survey variables, below specified thresholds;
- A weighted sum of variances of the estimates of totals for the m survey variables is minimized, while the total cost C of the survey is below or equal to the specified amount.

Minimization of a total cost. (De Moura Brito et al, 2015).

To find such sample size in each strata $n_1, n_2, \ldots, n_H$*, that total cost* C *of the survey*

$$\sum_{h=1}^{H} c_h n_h$$

is minimum

Minimization of a total cost. (De Moura Brito et al, 2015).

To find such sample size in each strata $n_1, n_2, \ldots, n_H$, that total cost C of the survey

$$\sum_{h=1}^{H} c_h n_h$$

is minimum and satisfies such requirements:

$$n_{min} \leq n_h \leq N_h, \quad h = 1, \ldots, H,$$

$$\frac{\sqrt{V(\hat{t}_{y^{(j)}})}}{t_{y^{(j)}}} \leq cv_{y^{(j)}}, \quad j = 1, \ldots, m,$$

$$n_h \in Z_+, \quad h = 1, \ldots, H,$$

here n_{min} – minimum sample size in strata ($n_{min} = 2$) and $cv_{y^{(j)}}, j = 1, \ldots, m$, predefined coefficients of variation.

Minimization of a total cost. (De Moura Brito et al, 2015).

To find such sample size in each strata $n_1, n_2, \dots, n_H$, that total cost C of the survey

$$\sum_{h=1}^{H} c_h n_h$$

is minimum and satisfies such requirements:

$$n_{min} \leq n_h \leq N_h, \quad h = 1, \dots, H,$$

$$\frac{\sqrt{V(\hat{t}_{y^{(j)}})}}{t_{y^{(j)}}} \leq cv_{y^{(j)}}, \quad j = 1, \dots, m,$$

$$n_h \in Z_+, \quad h = 1, \dots, H,$$

here n_{min} – minimum sample size in strata ($n_{min} = 2$) and $cv_{y^{(j)}}, j = 1, \dots, m$, predefined coefficients of variation.

Minimization of a total cost. (De Moura Brito et al, 2015).

To find such sample size in each strata $n_1, n_2, \ldots, n_H$, that total cost C of the survey

$$\sum_{h=1}^{H} c_h n_h$$

is minimum and satisfies such requirements:

$$n_{min} \leq n_h \leq N_h, \quad h = 1, \ldots, H,$$

$$\frac{\sqrt{V(\hat{t}_{y^{(j)}})}}{t_{y^{(j)}}} \leq cv_{y^{(j)}}, \quad j = 1, \ldots, m,$$

$$n_h \in Z_+, \quad h = 1, \ldots, H,$$

here n_{min} – minimum sample size in strata ($n_{min} = 2$) and $cv_{y^{(j)}}, j = 1, \ldots, m$, predefined coefficients of variation.

Minimization of a total cost. (De Moura Brito et al, 2015).

To find such sample size in each strata $n_1, n_2, \ldots, n_H$, *that total cost* C *of the survey*

$$\sum_{h=1}^{H} c_h n_h$$

is minimum and satisfies such requirements:

$$n_{min} \leq n_h \leq N_h, \quad h = 1, \ldots, H,$$

$$\frac{\sqrt{V(\hat{t}_{y^{(j)}})}}{t_{y^{(j)}}} \leq cv_{y^{(j)}}, \quad j = 1, \ldots, m,$$

$$n_h \in Z_+, \quad h = 1, \ldots, H,$$

here n_{min} – *minimum sample size in strata* ($n_{min} = 2$) *and* $cv_{y^{(j)}}, j = 1, \ldots, m$, *predefined coefficients of variation.*

New notations

Let us define a new binary variable with the values

$$z_{hk} = \begin{cases} 1, & \text{if the sample size } k \in \{n_{min}, \dots, N_h\}, \\ & \quad h = 1, \dots, H, \text{ is allocated to stratum } U_h; \\ 0, & \text{otherwise.} \end{cases}$$

Note that

$$n_h = \sum_{k=n_{min}}^{N_h} k z_{hk}, \quad h = 1, 2, \dots, H.$$

New notations

Let us define a new binary variable with the values

$$z_{hk} = \begin{cases} 1, & \text{if the sample size } k \in \{n_{min}, \ldots, N_h\}, \\ & \quad h = 1, \ldots, H, \text{ is allocated to stratum } U_h; \\ 0, & \text{otherwise.} \end{cases}$$

Note that

$$n_h = \sum_{k=n_{min}}^{N_h} k z_{hk}, \quad h = 1, 2, \ldots, H.$$

New formulation. (De Moura Brito et al, 2015)

Find z-variable, which minimize total survey cost function

$$C(z) = \sum_{h=1}^{H} c_h \sum_{k=n_{min}}^{N_h} k z_{hk}$$

Note. This algorithm is actualized in MultAlloc package in R.

New formulation. (De Moura Brito et al, 2015)

Find z-variable, which minimize total survey cost function

$$C(z) = \sum_{h=1}^{H} c_h \sum_{k=n_{min}}^{N_h} k z_{hk}$$

Note. This algorithm is actualized in MultAlloc package in R.

New formulation. (De Moura Brito et al, 2015)

Find z-variable, which minimize total survey cost function

$$C(z) = \sum_{h=1}^{H} c_h \sum_{k=n_{min}}^{N_h} k z_{hk}$$

and satisfy the following constraints:

$$1) \quad \sum_{k=n_{min}}^{N_h} z_{hk} = 1, \quad h = 1, 2, \ldots, H;$$

$$2) \quad \sum_{h=1}^{H} \frac{N_h^2 s_{hy^{(j)}}^2}{t_{y^{(j)}}^2 cv_{y^{(j)}}^2} \sum_{k=n_{min}}^{N_h} \frac{z_{hk}}{k} - \sum_{h=1}^{H} \frac{N_h s_{hy^{(j)}}^2}{t_{y^{(j)}}^2 cv_{y^{(j)}}^2} \leq 1, \quad j = 1, 2, \ldots, m;$$

$$3) \quad z_{hk} \in \{0, 1\}, \quad k = n_{min}, \ldots, N_h, \quad h = 1, 2, \ldots H.$$

Note. This algorithm is actualized in MultAlloc package in R.

New formulation. (De Moura Brito et al, 2015)

Find z-variable, which minimize total survey cost function

$$C(z) = \sum_{h=1}^{H} c_h \sum_{k=n_{min}}^{N_h} k z_{hk}$$

and satisfy the following constraints:

$$1) \quad \sum_{k=n_{min}}^{N_h} z_{hk} = 1, \quad h = 1, 2, \ldots, H;$$

$$2) \quad \sum_{h=1}^{H} \frac{N_h^2 s_{hy^{(j)}}^2}{t_{y^{(j)}}^2 cv_{y^{(j)}}^2} \sum_{k=n_{min}}^{N_h} \frac{z_{hk}}{k} - \sum_{h=1}^{H} \frac{N_h s_{hy^{(j)}}^2}{t_{y^{(j)}}^2 cv_{y^{(j)}}^2} \leq 1, \quad j = 1, 2, \ldots, m;$$

$$3) \quad z_{hk} \in \{0, 1\}, \quad k = n_{min}, \ldots, N_h, \quad h = 1, 2, \ldots H.$$

Note. This algorithm is actualized in MultAlloc package in R.

New formulation. (De Moura Brito et al, 2015)

Find z-variable, which minimize total survey cost function

$$C(z) = \sum_{h=1}^{H} c_h \sum_{k=n_{min}}^{N_h} k z_{hk}$$

and satisfy the following constraints:

$$1) \quad \sum_{k=n_{min}}^{N_h} z_{hk} = 1, \quad h = 1, 2, \ldots, H;$$

$$2) \quad \sum_{h=1}^{H} \frac{N_h^2 s_{hy^{(j)}}^2}{t_{y^{(j)}}^2 cv_{y^{(j)}}^2} \sum_{k=n_{min}}^{N_h} \frac{z_{hk}}{k} - \sum_{h=1}^{H} \frac{N_h s_{hy^{(j)}}^2}{t_{y^{(j)}}^2 cv_{y^{(j)}}^2} \leq 1, \quad j = 1, 2, \ldots, m;$$

$$3) \quad z_{hk} \in \{0, 1\}, \quad k = n_{min}, \ldots, N_h, \quad h = 1, 2, \ldots H.$$

Note. This algorithm is actualized in MultAlloc package in R.

Bethel (1985, 1989) algorithm

To find such sample size in each strata $n_1, n_2, \ldots, n_H$, that total cost C of the survey

$$\sum_{h=1}^{H} c_h n_h$$

is minimum and satisfies such requirements:

$$\frac{\sqrt{V(\hat{t}_{y(j)})}}{t_{y(j)}} \leq cv_{y(j)}, \quad j = 1, \ldots, m,$$

Note. This algorithm is actualized in bethel package in R.

Simulated populations. Options.

- Population size $N = 1000$, number of strata $H = 4$.
- Auxiliary variables: $x^{(1)} \sim \mathcal{E}(0.01)$, $x^{(2)} \sim \mathcal{N}(50, 15)$, $x^{(3)} \sim \mathcal{N}(75, 15)$, $x^{(4)} \sim \mathcal{N}(100, 20)$.
- Study variables where generated using such formula:

$$y^{(j)} = 50 + x^{(j)} + \alpha\sqrt{x^{(j)}}z, \quad j = 1, 2, 3, 4,$$

  where $z \sim \mathcal{N}(0, 1)$, and α is chosen so, that correlation between study and auxiliary variables are in a given interval.
- $c_1 = c_2 = c_3 = c_4$.

Simulated populations. Options.

- Population size $N = 1000$, number of strata $H = 4$.
- Auxiliary variables: $x^{(1)} \sim \mathcal{E}(0.01)$, $x^{(2)} \sim \mathcal{N}(50, 15)$, $x^{(3)} \sim \mathcal{N}(75, 15)$, $x^{(4)} \sim \mathcal{N}(100, 20)$.
- Study variables where generated using such formula:

$$y^{(j)} = 50 + x^{(j)} + \alpha\sqrt{x^{(j)}}z, \quad j = 1, 2, 3, 4,$$

  where $z \sim \mathcal{N}(0, 1)$, and α is chosen so, that correlation between study and auxiliary variables are in a given interval.
- $c_1 = c_2 = c_3 = c_4$.

Simulated populations. Strategies.

4 different strategies are chosen:

Strategy	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.1	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.2	$s_{hx^{(j)}}$	$t_{x^{(j)}}$
St.3	$s_{h\hat{y}^{(j)}}$	$t_{\hat{y}^{(j)}}$
St.4	$s_{h\hat{\varepsilon}^{(j)}}$	$t_{\hat{y}^{(j)}}$

Here $\hat{y}_k^{(j)} = a + bx_k^{(j)}$
and $\hat{\varepsilon}_k^{(j)} = \hat{y}_k^{(j)} - y_k^{(j)}$, $j = 1, ..., 4,\ k = 1, ..., N.$

Simulated populations. Strategies.

4 different strategies are chosen:

Strategy	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.1	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.2	$s_{hx^{(j)}}$	$t_{x^{(j)}}$
St.3	$s_{h\hat{y}^{(j)}}$	$t_{\hat{y}^{(j)}}$
St.4	$s_{h\hat{\varepsilon}^{(j)}}$	$t_{\hat{y}^{(j)}}$

Here $\hat{y}_k^{(j)} = a + bx_k^{(j)}$
and $\hat{\varepsilon}_k^{(j)} = \hat{y}_k^{(j)} - y_k^{(j)}$, $j = 1, ..., 4,\ k = 1, ..., N$.

Simulated populations. Strategies.

4 different strategies are chosen:

Strategy	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.1	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.2	$s_{hx^{(j)}}$	$t_{x^{(j)}}$
St.3	$s_{h\hat{y}^{(j)}}$	$t_{\hat{y}^{(j)}}$
St.4	$s_{h\hat{\varepsilon}^{(j)}}$	$t_{\hat{y}^{(j)}}$

Here $\hat{y}_k^{(j)} = a + bx_k^{(j)}$
and $\hat{\varepsilon}_k^{(j)} = \hat{y}_k^{(j)} - y_k^{(j)}$, $j = 1, ..., 4,\ k = 1, ..., N.$

Simulated populations. Strategies.

4 different strategies are chosen:

Strategy	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.1	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.2	$s_{hx^{(j)}}$	$t_{x^{(j)}}$
St.3	$s_{h\hat{y}^{(j)}}$	$t_{\hat{y}^{(j)}}$
St.4	$s_{h\hat{\varepsilon}^{(j)}}$	$t_{\hat{y}^{(j)}}$

Here $\hat{y}_k^{(j)} = a + bx_k^{(j)}$
and $\hat{\varepsilon}_k^{(j)} = \hat{y}_k^{(j)} - y_k^{(j)}$, $j = 1, ..., 4,\ k = 1, ..., N$.

4 different strategies are chosen:

Strategy	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.1	$s_{hy^{(j)}}$	$t_{y^{(j)}}$
St.2	$s_{hx^{(j)}}$	$t_{x^{(j)}}$
St.3	$s_{h\hat{y}^{(j)}}$	$t_{\hat{y}^{(j)}}$
St.4	$s_{h\hat{\varepsilon}^{(j)}}$	$t_{\hat{y}^{(j)}}$

Here $\hat{y}_k^{(j)} = a + bx_k^{(j)}$
and $\hat{\varepsilon}_k^{(j)} = \hat{y}_k^{(j)} - y_k^{(j)}$, $j = 1, ..., 4,\ k = 1, ..., N$.

Simulation results.

Minimizing sample size, when $cv_{y^{(j)}} \leq 0.05$.

$0.9 \leq cor(x^{(j)}, y^{(j)}) \leq 0.98$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	3	10	2	2
n_2	10	20	7	4
n_3	33	52	26	8
n_4	18	30	16	3
Total	**64**	**112**	**51**	**17**

$0.65 \leq cor(x^{(j)}, y^{(j)}) \leq 0.75$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	16	10	2	13
n_2	64	20	7	63
n_3	155	52	21	150
n_4	60	30	14	55
Total	**295**	**112**	**44**	**281**

$0.3 \leq cor(x^{(j)}, y^{(j)}) \leq 0.5$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	24	10	4	21
n_2	100	20	6	100
n_3	237	52	19	237
n_4	89	30	12	88
Total	**450**	**112**	**41**	**446**

Here $j = 1, ..., 4$.

Simulation results.

Minimizing sample size, when $cv_{y^{(j)}} \leq 0.05$.

$0.9 \leq cor(x^{(j)}, y^{(j)}) \leq 0.98$	Strategy			
	St.1	St.2	St.3	St.4
n_1	3	10	2	2
n_2	10	20	7	4
n_3	33	52	26	8
n_4	18	30	16	3
Total	**64**	**112**	**51**	**17**

$0.65 \leq cor(x^{(j)}, y^{(j)}) \leq 0.75$	Strategy			
	St.1	St.2	St.3	St.4
n_1	16	10	2	13
n_2	64	20	7	63
n_3	155	52	21	150
n_4	60	30	14	55
Total	**295**	**112**	**44**	**281**

$0.3 \leq cor(x^{(j)}, y^{(j)}) \leq 0.5$	Strategy			
	St.1	St.2	St.3	St.4
n_1	24	10	4	21
n_2	100	20	6	100
n_3	237	52	19	237
n_4	89	30	12	88
Total	**450**	**112**	**41**	**446**

Here $j = 1, ..., 4$.

Simulation results.

Minimizing sample size, when $cv_{y^{(j)}} \leq 0.05$.

$0.9 \leq cor(x^{(j)}, y^{(j)}) \leq 0.98$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	3	10	2	2
n_2	10	20	7	4
n_3	33	52	26	8
n_4	18	30	16	3
Total	**64**	**112**	**51**	**17**

$0.65 \leq cor(x^{(j)}, y^{(j)}) \leq 0.75$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	16	10	2	13
n_2	64	20	7	63
n_3	155	52	21	150
n_4	60	30	14	55
Total	**295**	**112**	**44**	**281**

$0.3 \leq cor(x^{(j)}, y^{(j)}) \leq 0.5$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	24	10	4	21
n_2	100	20	6	100
n_3	237	52	19	237
n_4	89	30	12	88
Total	**450**	**112**	**41**	**446**

Here $j = 1, \ldots, 4$.

Simulation results.

Minimizing sample size, when $cv_{y^{(j)}} \leq 0.05$.

$0.9 \leq cor(x^{(j)}, y^{(j)}) \leq 0.98$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	3	10	2	2
n_2	10	20	7	4
n_3	33	52	26	8
n_4	18	30	16	3
Total	**64**	**112**	**51**	**17**

$0.65 \leq cor(x^{(j)}, y^{(j)}) \leq 0.75$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	16	10	2	13
n_2	64	20	7	63
n_3	155	52	21	150
n_4	60	30	14	55
Total	**295**	**112**	**44**	**281**

$0.3 \leq cor(x^{(j)}, y^{(j)}) \leq 0.5$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	24	10	4	21
n_2	100	20	6	100
n_3	237	52	19	237
n_4	89	30	12	88
Total	**450**	**112**	**41**	**446**

Here $j = 1, ..., 4$.

Simulation results.

Minimizing sample size, when $cv_{y^{(j)}} \leq 0.05$.

$0.9 \leq cor(x^{(j)}, y^{(j)}) \leq 0.98$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	3	10	2	2
n_2	10	20	7	4
n_3	33	52	26	8
n_4	18	30	16	3
Total	**64**	**112**	**51**	**17**

$0.65 \leq cor(x^{(j)}, y^{(j)}) \leq 0.75$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	16	10	2	13
n_2	64	20	7	63
n_3	155	52	21	150
n_4	60	30	14	55
Total	**295**	**112**	**44**	**281**

$0.3 \leq cor(x^{(j)}, y^{(j)}) \leq 0.5$				
	Strategy			
	St.1	St.2	St.3	St.4
n_1	24	10	4	21
n_2	100	20	6	100
n_3	237	52	19	237
n_4	89	30	12	88
Total	**450**	**112**	**41**	**446**

Here $j = 1, \ldots, 4$.

Population: Household Budget Survey

- 2016 y. data with $N = 3443$ elements.
- Study parameters:
  - $y^{(1)}$ – total consumption expenditure (monthly);
  - $y^{(2)}$ – food and non-alcoholic beverages;
  - $y^{(3)}$ – alcoholic beverages, tobacco and narcotics;
  - $y^{(4)}$ – clothing and footwear;
  - $y^{(5)}$ – housing, water, electricity, gas and other fuels;
  - $y^{(6)}$ – furnishings, household equipment and routine maintenance of the house;
  - $y^{(7)}$ – health;
  - $y^{(8)}$ – transport;
  - $y^{(9)}$ – communication;
  - $y^{(10)}$ – recreation and culture;
  - $y^{(11)}$ – education;
  - $y^{(12)}$ – restaurants and hotels;
  - $y^{(13)}$ – miscellaneous goods and services.
- Strata: Type of dwelling (3) * Income quintile groups (5) * Place of residence (3) $= 45$.

Population: Household Budget Survey

- 2016 y. data with $N = 3443$ elements.
- Study parameters:
  - $y^{(1)}$ – total consumption expenditure (monthly);
  - $y^{(2)}$ – food and non-alcoholic beverages;
  - $y^{(3)}$ – alcoholic beverages, tobacco and narcotics;
  - $y^{(4)}$ – clothing and footwear;
  - $y^{(5)}$ – housing, water, electricity, gas and other fuels;
  - $y^{(6)}$ – furnishings, household equipment and routine maintenance of the house;
  - $y^{(7)}$ – health;
  - $y^{(8)}$ – transport;
  - $y^{(9)}$ – communication;
  - $y^{(10)}$ – recreation and culture;
  - $y^{(11)}$ – education;
  - $y^{(12)}$ – restaurants and hotels;
  - $y^{(13)}$ – miscellaneous goods and services.
- Strata: Type of dwelling (3) * Income quintile groups (5) * Place of residence (3) $= 45$.

Comparison of Brito and Bethel algorithms, when auxiliary variables are used.

Minimizing sample size, when $cv_{y(j)}$ can't exceed predefined threshold.

	Brito algorithm			Bethel algorithm		
$cv_{y(j)}$	0.03	0.05	0.10	0.03	0.05	0.10
n	**2656**	**2158**	**1217**	**2464**	**2150**	**1236**
$cv(\hat{t}_{y(1)})$	0.01	0.01	0.02	0.01	0.01	0.02
$cv(\hat{t}_{y(2)})$	0.01	0.01	0.02	0.01	0.01	0.02
$cv(\hat{t}_{y(3)})$	0.02	0.03	0.04	0.02	0.03	0.04
$cv(\hat{t}_{y(4)})$	0.03	0.04	0.06	0.03	0.04	0.05
$cv(\hat{t}_{y(5)})$	0.01	0.01	0.02	0.01	0.01	0.02
$cv(\hat{t}_{y(6)})$	0.02	0.03	0.04	0.02	0.03	0.04
$cv(\hat{t}_{y(7)})$	0.02	0.03	0.04	0.02	0.03	0.04
$cv(\hat{t}_{y(8)})$	0.02	0.03	0.05	0.02	0.03	0.05
$cv(\hat{t}_{y(9)})$	0.01	0.02	0.02	0.01	0.02	0.02
$cv(\hat{t}_{y(10)})$	0.02	0.02	0.04	0.02	0.02	0.03
$cv(\hat{t}_{y(11)})$	0.03	0.05	0.10	0.04	0.05	0.10
$cv(\hat{t}_{y(12)})$	0.03	0.04	0.07	0.03	0.04	0.06
$cv(\hat{t}_{y(13)})$	0.02	0.02	0.04	0.02	0.02	0.04

Comparison of Brito and Bethel algorithms, when auxiliary variables are used.

Minimizing sample size, when $cv_{y(j)}$ can't exceed predefined threshold.

	Brito algorithm			Bethel algorithm		
$cv_{y(j)}$	0.03	0.05	0.10	0.03	0.05	0.10
n	**2656**	**2158**	**1217**	**2464**	**2150**	**1236**
$cv(\hat{t}_{y(1)})$	0.01	0.01	0.02	0.01	0.01	0.02
$cv(\hat{t}_{y(2)})$	0.01	0.01	0.02	0.01	0.01	0.02
$cv(\hat{t}_{y(3)})$	0.02	0.03	0.04	0.02	0.03	0.04
$cv(\hat{t}_{y(4)})$	0.03	0.04	0.06	0.03	0.04	0.05
$cv(\hat{t}_{y(5)})$	0.01	0.01	0.02	0.01	0.01	0.02
$cv(\hat{t}_{y(6)})$	0.02	0.03	0.04	0.02	0.03	0.04
$cv(\hat{t}_{y(7)})$	0.02	0.03	0.04	0.02	0.03	0.04
$cv(\hat{t}_{y(8)})$	0.02	0.03	0.05	0.02	0.03	0.05
$cv(\hat{t}_{y(9)})$	0.01	0.02	0.02	0.01	0.02	0.02
$cv(\hat{t}_{y(10)})$	0.02	0.02	0.04	0.02	0.02	0.03
$cv(\hat{t}_{y(11)})$	0.03	0.05	0.10	0.04	0.05	0.10
$cv(\hat{t}_{y(12)})$	0.03	0.04	0.07	0.03	0.04	0.06
$cv(\hat{t}_{y(13)})$	0.02	0.02	0.04	0.02	0.02	0.04

Comparison of Brito and Bethel algorithms, when auxiliary variables are used.

Minimizing sample size, when $cv_{y(j)} \leq 0.03$.

Strategy	Brito algorithm St.1	Brito algorithm St.2	Bethel algorithm St.1	Bethel algorithm St.2
n	**2656**	**2739**	**2464**	**2564**
$cv(\hat{t}_{y(1)})$	0.008	0.005	0.008	0.007
$cv(\hat{t}_{y(1a)})$	0.053	0.030	0.048	0.034
$cv(\hat{t}_{y(1b)})$	0.033	0.030	0.034	0.030
$cv(\hat{t}_{y(1c)})$	0.021	0.022	0.024	0.025
$cv(\hat{t}_{y(1d)})$	0.082	0.030	0.074	0.032
$cv(\hat{t}_{y(1e)})$	0.130	0.030	0.129	0.032
$cv(\hat{t}_{y(1f)})$	0.095	0.030	0.097	0.034
$cv(\hat{t}_{y(1g)})$	0.005	0.006	0.007	0.008
$cv(\hat{t}_{y(1h)})$	0.010	0.010	0.013	0.013
$cv(\hat{t}_{y(1i)})$	0.068	0.030	0.063	0.030

Comparison of Brito and Bethel algorithms, when auxiliary variables are used.

Minimizing sample size, when $cv_{y(j)} \leq 0.03$.

Strategy	Brito algorithm St.1	Brito algorithm St.2	Bethel algorithm St.1	Bethel algorithm St.2
n	**2656**	**2739**	**2464**	**2564**
$cv(\hat{t}_{y(1)})$	0.008	0.005	0.008	0.007
$cv(\hat{t}_{y(1a)})$	0.053	0.030	0.048	0.034
$cv(\hat{t}_{y(1b)})$	0.033	0.030	0.034	0.030
$cv(\hat{t}_{y(1c)})$	0.021	0.022	0.024	0.025
$cv(\hat{t}_{y(1d)})$	0.082	0.030	0.074	0.032
$cv(\hat{t}_{y(1e)})$	0.130	0.030	0.129	0.032
$cv(\hat{t}_{y(1f)})$	0.095	0.030	0.097	0.034
$cv(\hat{t}_{y(1g)})$	0.005	0.006	0.007	0.008
$cv(\hat{t}_{y(1h)})$	0.010	0.010	0.013	0.013
$cv(\hat{t}_{y(1i)})$	0.068	0.030	0.063	0.030

Comparison of Brito and Bethel algorithms, when auxiliary variables are used.

Minimizing sample size, when $cv_{y(j)} \leq 0.03$.

Strategy	Brito algorithm St.1	Brito algorithm St.2	Bethel algorithm St.1	Bethel algorithm St.2
n	**2656**	**2739**	**2464**	**2564**
$cv(\hat{t}_{y(1)})$	0.008	0.005	0.008	0.007
$cv(\hat{t}_{y(1a)})$	0.053	0.030	0.048	0.034
$cv(\hat{t}_{y(1b)})$	0.033	0.030	0.034	0.030
$cv(\hat{t}_{y(1c)})$	0.021	0.022	0.024	0.025
$cv(\hat{t}_{y(1d)})$	0.082	0.030	0.074	0.032
$cv(\hat{t}_{y(1e)})$	0.130	0.030	0.129	0.032
$cv(\hat{t}_{y(1f)})$	0.095	0.030	0.097	0.034
$cv(\hat{t}_{y(1g)})$	0.005	0.006	0.007	0.008
$cv(\hat{t}_{y(1h)})$	0.010	0.010	0.013	0.013
$cv(\hat{t}_{y(1i)})$	0.068	0.030	0.063	0.030

Main literature

J. Bethel. An optimum allocation algorithm for multivariate surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1985, 209-212.

Bethel, J. (1989). Sample allocation in multivariate surveys. *Survey Methodology*, Vol. 15, No. 1, 47-57.

J. A. de Moura Brito, P. L. do Nascimento Silva, G. S. Semaan, N. Maculan. Integer programming formulations applied to optimal allocation in stratified sampling. *Survey Methodology*, 2015, Vol. 41, No. 2, 427-442.

Cochran, W. G. (1977) *Sampling Techniques*. New York: John Wiley & Sons.

D. Krapavickaitė, A. Plikusas. *Imčių teorijos pagrindai*. Vilnius: Technika, 2005.

C. E. Särndal, B. Swensson J. Wretman. *Model Assisted Survey Sampling*, Springer-Verlag, New York, 1992.