Uniwersytet
Ekonomiczny
w Katowicach

blisko

międzynarodowo

przez całe życie

Properties of selected strategies for estimating the population total using sequential fixed-cost sampling schemes

Krzysztof Szymoniak-Książek
College of Management
Department of Statistics, Econometrics, and Mathematics
szymoniak-ksiazek@ue.katowice.pl

Outline

- Purpose of the presentation
- Known fixed-cost sampling schemes
- Proposed modifications to the schemes
- Comparison of properties
- Summary

Purpose of the presentation

Comparison of properties of sampling schemes incorporating data acquisition cost.

Sampling schemes:

1. Pathak scheme
2. Greedy scheme
3. Modified greedy scheme

Notation

- Finite population:

$$U = \{1, \dots, N\}$$

- Study variable:

$$\boldsymbol{y} = [y_1, \dots, y_N]$$

- Population total:

$$\tilde{y} = \sum_{i=1}^{N} y_i$$

Pathak scheme (S1)

- Cost vector:

$$\boldsymbol{c} = [c_1, \dots, c_N]$$

- Research budget B:

$$B > \max_{i \neq j \in U} \{c_i + c_j\}$$

Sampling until the sum of costs for the selected elements exceeds or reaches the research budget. The element for which this occurs is not included in the sample.

[Pathak, 1976]

Pathak scheme (S1)

Consider sampling scheme for sample $s = \{s_1, \dots, s_n\}$. Let

$$S_k = \sum_{i=1}^{k} c_{s_i}$$

First step ($k = 1$):

Randomly select (SRS) an element s_1 with cost c_{s_1} from set $U_1 = U$.

Next steps ($k = 2, 3, \dots$):

Randomly select (SRS) element s_k with cost c_{s_k} from set $U_k = U_{k-1} \setminus \{s_{k-1}\}$. If $S_k = S_{k-1} + c_{s_k} \geq B$ then the sample is a set $s = \{s_1, \dots, s_{k-1}\}$. Otherwise, we go to step $\mathrm{k} + 1$.

S1– inclusion probabilities

$\Omega = \{\omega_1, \omega_2, \ldots, \omega_{N!}\}$ - permutations of U

Pathak sampling as a two-step process:

1. Randomly select (SRS) permutation ω from Ω.
2. Select the longest subsequence of initial elements from ω, such that the cumulative cost is less than the research budget B.

Let $\kappa_P : \Omega \to \Omega_P$ be a function such that $\kappa_P(\omega) = \omega^*$, where ω^* is the subsequence selected according to the second point.

First-order inclusion probabilities:

$$\pi_i = \frac{\#\{\omega \in \Omega : i \in \kappa_P(\omega)\}}{N!}$$

First-order inclusion probabilities for S1 - an example.

$U = \{1,2,3,4,5,6\}$

$\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5, y_6]$

$\boldsymbol{c} = [1,1,1,2,2,3]$

$B = 6$

$\#\Omega = 720$

Example values of κ_P:

$\kappa_P((1,2,3,4,5,6)) = (1,2,3,4)$

$\kappa_P((6,5,4,3,2,1)) = (6,5)$

$\kappa_P((4,1,2,6,5,3)) = (4,1,2)$

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3, \pi_4, \pi_5, \pi_6) = \left(\frac{360}{720}, \frac{360}{720}, \frac{360}{720}, \frac{324}{720}, \frac{324}{720}, \frac{276}{720}\right)$$

Greedy scheme (S2)

Cost vector: $\boldsymbol{c} = [c_1, \dots, c_N]$

Research budget: $B > \max_{i \neq j \in U} \{c_i + c_j\}$

Sampling until the total cost for selected elements exceeds the budget. The element for which this occurs is not included in the sample. Among the remaining elements, we choose those whose selection will not exceed the budget.

We repeat the procedure until no such element remains.

Greedy scheme (S2)

$$T \subset U$$

$$c(T) = \sum_{i \in T} c_i$$

An algorithm implementing a greedy scheme [Gamrot, 2014]:

1. Randomly order all elements of the population U into a $N-$element sequence $(A_1, \dots, A_N)$.
2. Assume that $s_0 = \emptyset$.
3. For $i = 1, \dots, N$, perform the following N steps sequentially:
   - if $c_{A_i} \leq B - c(s_{i-1})$, then $s_i = s_{i-1} \cup \{\mathrm{A_i}\}$
   - otherwise $s_i = s_{i-1}$
4. The set s_N is a sample.

S2– inclusion probabilities

$\Omega = \{\omega_1, \omega_2, \dots, \omega_{N!}\}$ - permutations of U

Let $\kappa_Z\colon \Omega \to \Omega_Z$ be a function such that $\kappa_P(\omega) = \omega^*$, where ω^* is a subsequence selected according to steps 2 - 4 of the algorithm.

Greedy sampling as a two-step process:

1. Randomly select (SRS) permutation ω from Ω.
2. Calculate $\kappa_Z(\omega)$.

First-order inclusion probabilities:

$$\pi_i = \frac{\#\{\omega \in \Omega\colon i \in \kappa_Z(\omega)\}}{N!}$$

First-order inclusion probabilities for S2 - an example

$U = \{1,2,3,4,5,6\}$

$\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5, y_6]$

$\boldsymbol{c} = [1,1,1,2,2,3]$

$B = 6$

$\#\Omega = 720$

Example values κ_Z:

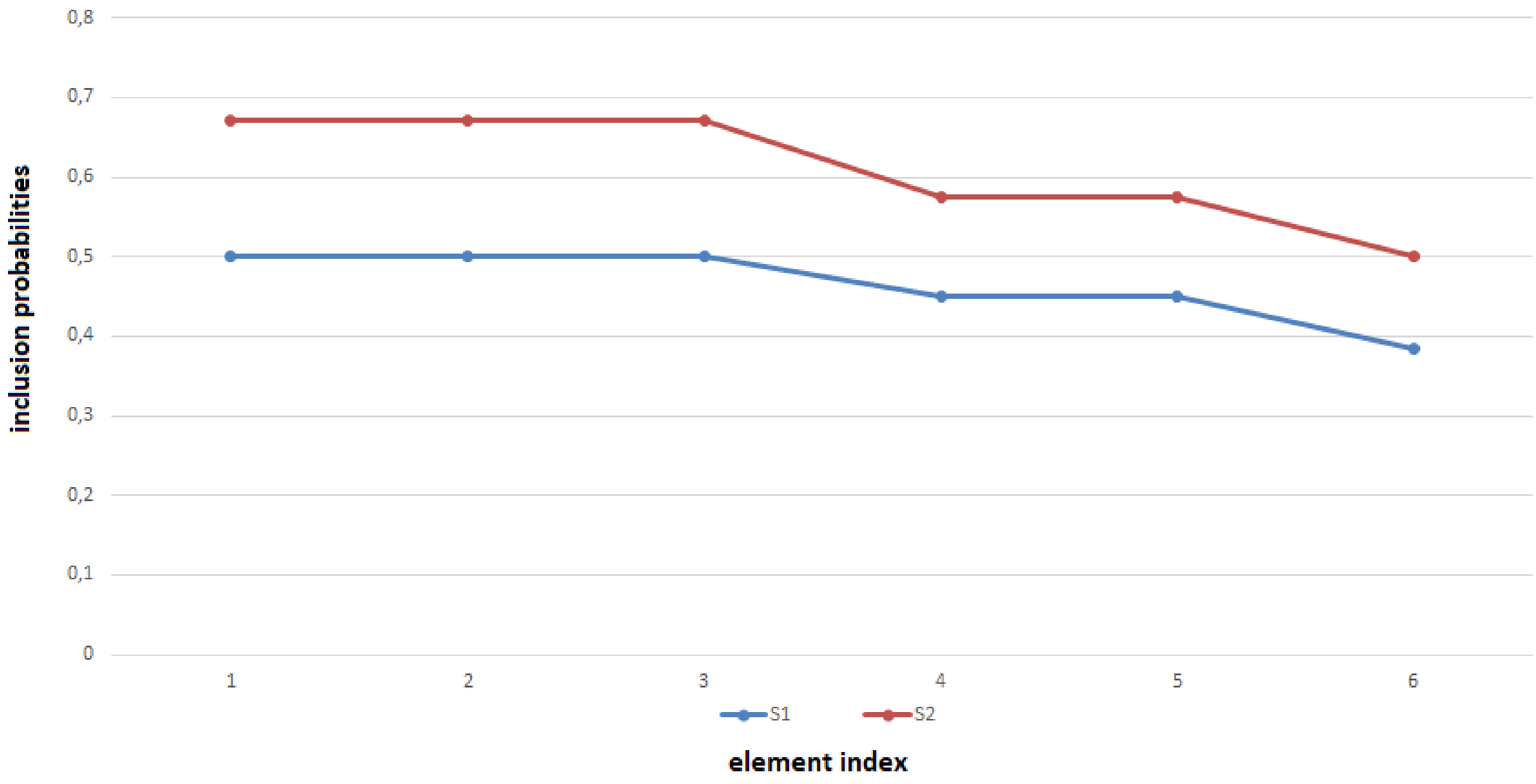
$\kappa_Z\big((1,2,3,4,5,6)\big) = (1,2,3,4)$

$\kappa_Z\big((6,5,4,3,2,1)\big) = (6,5,3)$

$\kappa_Z((4,1,2,6,5,3)) = (4,1,2,5)$

$$\boldsymbol{\pi} = \left(\frac{484}{720}, \frac{484}{720}, \frac{484}{720}, \frac{414}{720}, \frac{414}{720}, \frac{360}{720}\right) = \left(\frac{121}{180}, \frac{121}{180}, \frac{121}{180}, \frac{23}{40}, \frac{23}{40}, \frac{1}{2}\right)$$

Comparison



0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
inclusion probabilities
1
2
3
4
5
6
S1
S2
element index

Horvitz-Thompson (HT) population total estimator

$\pi_1, \ldots, \pi_N$ – first-order inclusion probabilities

$d_i = \frac{1}{\pi_i}$ – weights

$s = \{s_1, \ldots, s_k\}$ – sample

HT estimator for known π_i's:

$$\tilde{y}_{HT} = \sum_{i \in s} d_i y_i$$

[Narain, 1951], [Horvitz, Thompson, 1952]

Empirical inclusion probabilities

When calculating π_i's is impossible, replace them with estimates.

$s_1^{(1)}, \ldots, s_1^{(R)}$ - sample replications

$f_1, \ldots, f_N$ - the number of occurrences of each item in R sample replications

$\hat{\pi}_i = \frac{f_i}{R}$ - empirical inclusion probabilities

$\hat{d}_i = \frac{1}{\hat{\pi}_i}$ - empirical weights

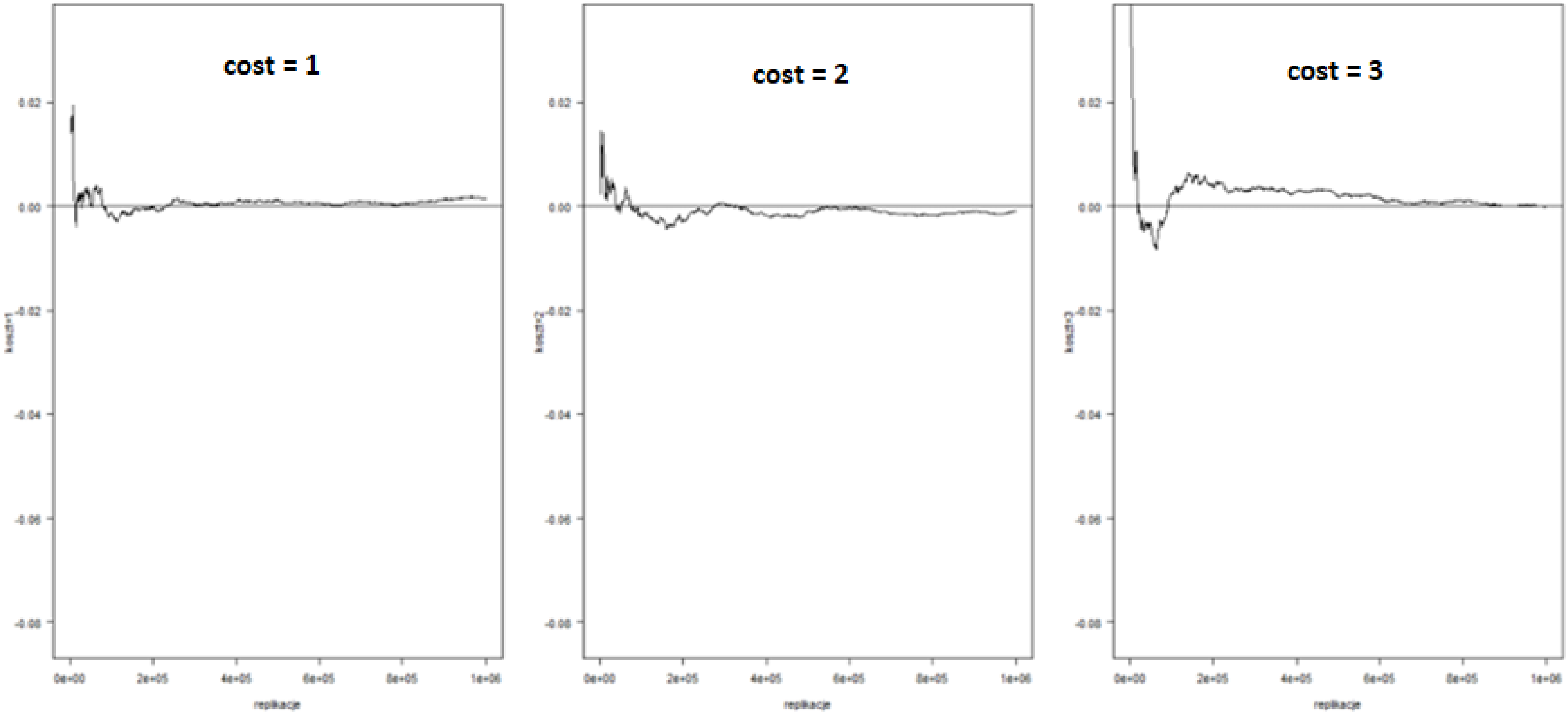
The empirical HT population total estimator

$s = \{s_1, \dots, s_k\} -$ sample

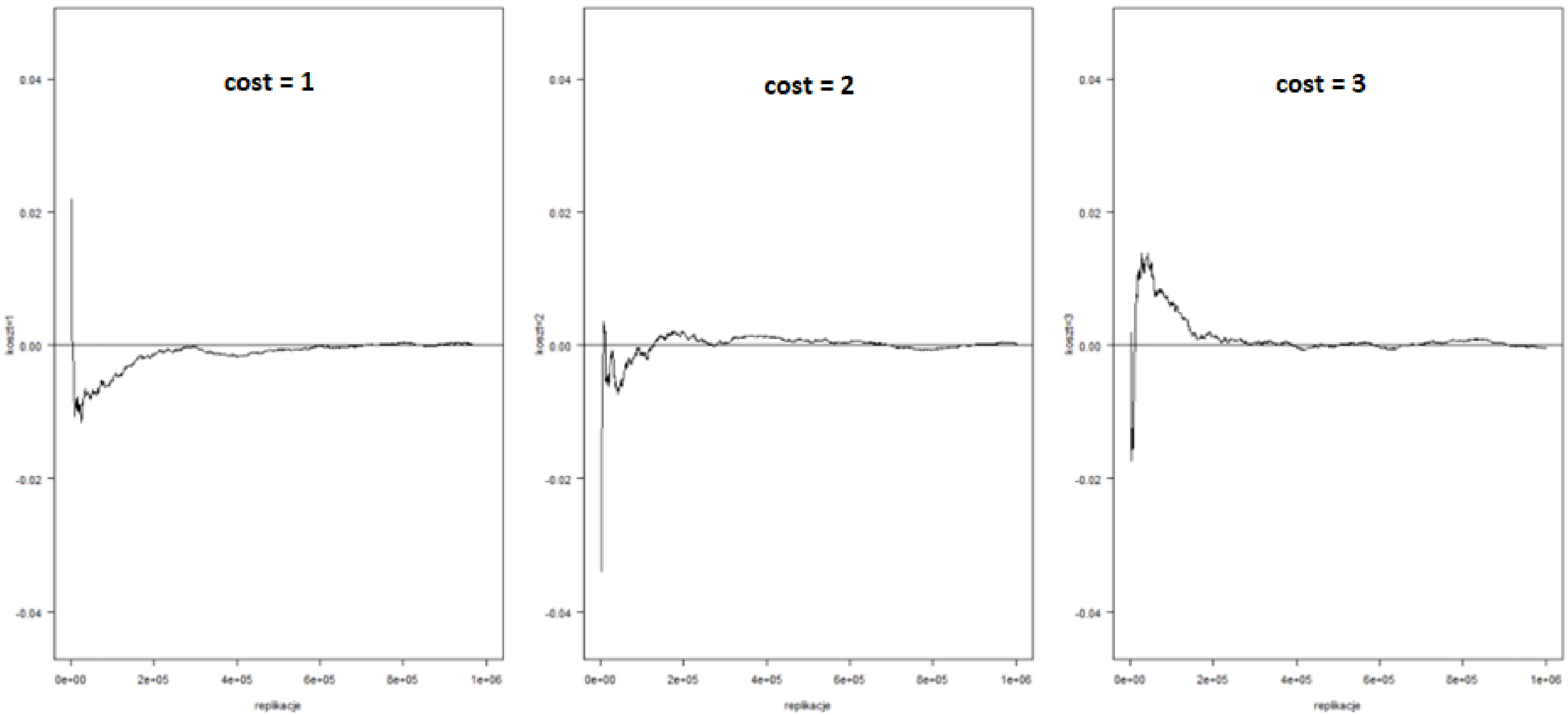
$\hat{d}_i -$ empirical weights

$$\tilde{y}_{EHT} = \sum_{i \in s} \hat{d}_i y_i$$

Empirical inclusion probabilities for S1 - relative biases



Empirical inclusion probabilities for S2 - relative biases



Modified Pathak scheme (S3)

$\boldsymbol{x} = [x_1, \dots, x_N]$ – auxiliary variable values

$\boldsymbol{c} = [c_1, \dots, c_N]$ – costs

$B > \max_{i \neq j \in U}\{c_i + c_j\}$ – research budget

Sampling (with probabilities proportional to $\boldsymbol{x}$) until the total cost for selected elements exceeds the budget. The element for which this occurs is not included in the sample.

Modified Pathak scheme (S3)

Consider sampling scheme for sample $s = \{s_1, \dots, s_n\}$. Let

$$S_k = \sum_{i=1}^{k} c_{s_i}$$

First step ($k = 1$):

Randomly select s_i from set $U_1 = U$ with probabilities

$$p_i^{(1)} = \frac{x_i}{\sum_{i \in U_1} x_i}$$

Next steps ($k = 2, 3, \dots$):

Randomly select s_k from set $U_k = U_{k-1} \setminus \{s_{k-1}\}$ with probabilities $p_i^{(k)} = \frac{x_i}{\sum_{i \in U_k} x_i}$.

If $S_k = S_{k-1} + c_{s_k} > B$, then the sample is a set $s = \{s_1, \dots, s_{k-1}\}$. Otherwise, we go to step $k + 1$.

Modified greedy scheme (S4)

$\boldsymbol{x} = [x_1, \dots, x_N]$ – auxiliary variable values

$\boldsymbol{c} = [c_1, \dots, c_N]$ – costs

$B > \max_{i \neq j \in U} \{c_i + c_j\}$ – research budget

Sampling (with probabilities proportional to $\boldsymbol{x}$) until the total cost for selected elements exceeds the budget. The element for which this occurs is not included in the sample. Among the remaining elements, we choose those whose selection will not breach the budget.

Repeat the procedure until no such element remains.

Modified greedy scheme (S4)

$$T \subset U$$

$$c(T) = \sum_{i \in T} c_i$$

$$U(T) = \{i \in U - T : c_i \leq B - c(T)\}$$

1. Randomly select s_1 from population U with the modified Pathak scheme with a budget constraint B.
2. For $l = 1,2,3,\ldots$ create a set s_{l+1} by adding through random selection s_l with scheme S3, elements from the set $U(s_l)$ with the constraint of $B_{l+1} = B - c(s_l)$, until some $l = L$ for which set $U(s_L)$ is empty. Here, the assumption that the budget is greater than the sum of the two largest costs is disregarded.
3. The set s_L is a sample.

First-order inclusion probabilities for S4 - an example

$U = \{1,2,3,4,5,6\}$

$\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5, y_6]$

$\boldsymbol{x} = [1,1,1,2,2,3]$

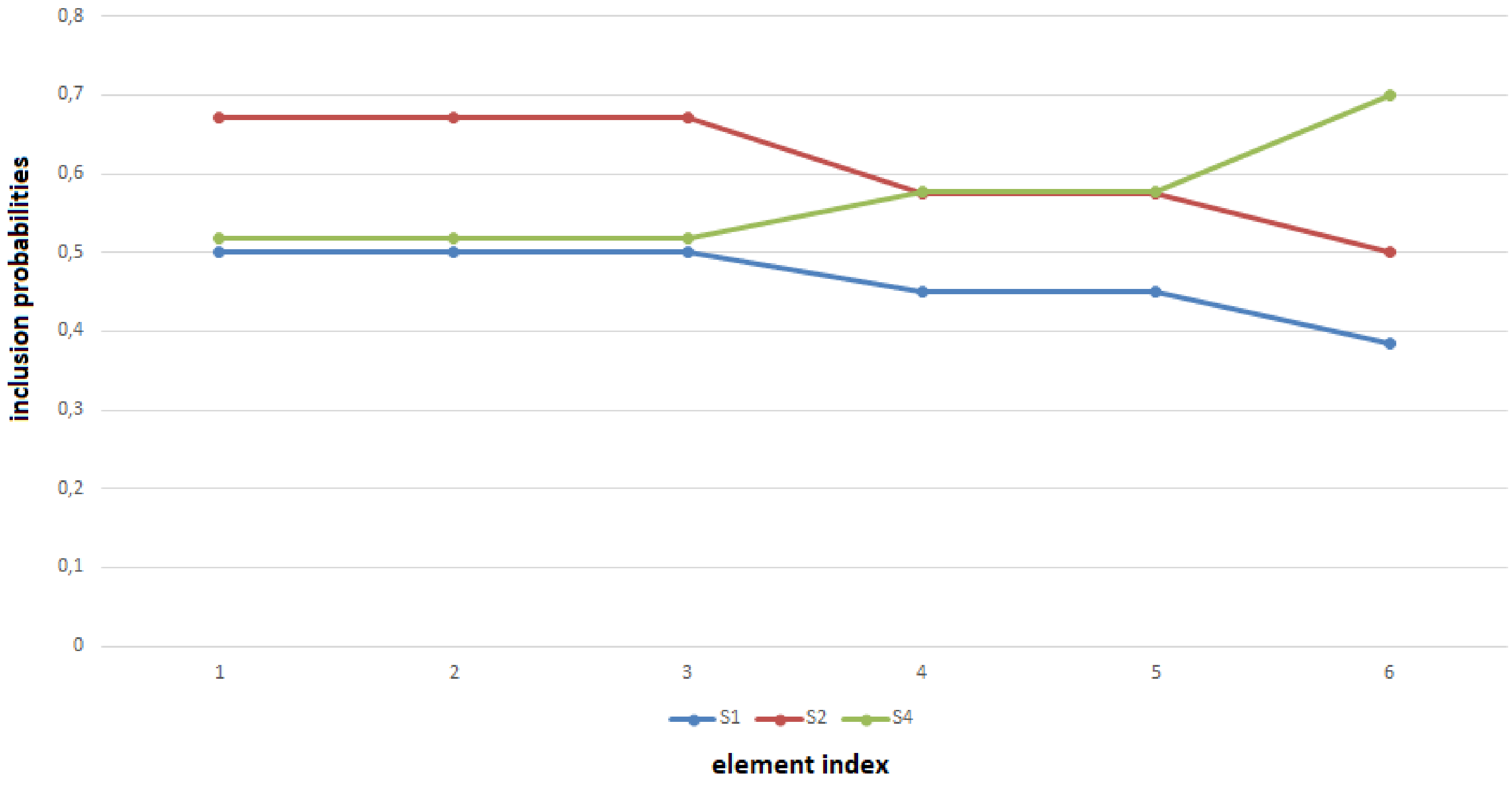
$\boldsymbol{c} = [1,1,1,2,2,3]$

$B = 6$

The inclusion probabilities calculated by analyzing all possible outcomes of the random selection process:

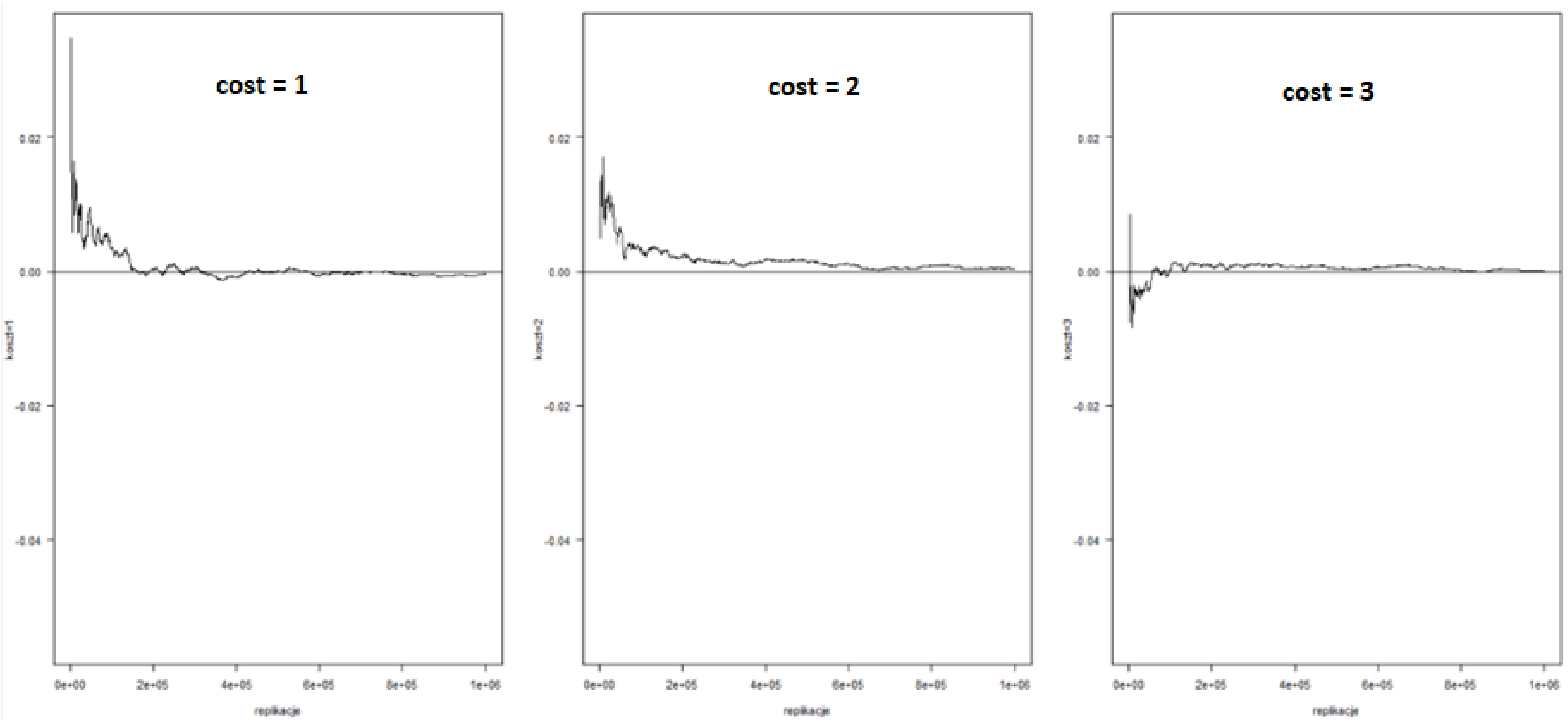
$$\boldsymbol{\pi} = \left(\frac{163}{315}, \frac{163}{315}, \frac{163}{315}, \frac{121}{210}, \frac{121}{210}, \frac{7}{10}\right)$$

Comparison



0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
inclusion probabilities
1
2
3
4
5
6
S1
S2
S4
element index

Empirical inclusion probabilities for S4 - relative biases



The average amount of unused funds

In the following part, the average amount of unused budget funds during the sampling using schemes S1, S2, and S4 is considered.

Data: agricultural census from 1996 - Dąbrowa Tarnowska county

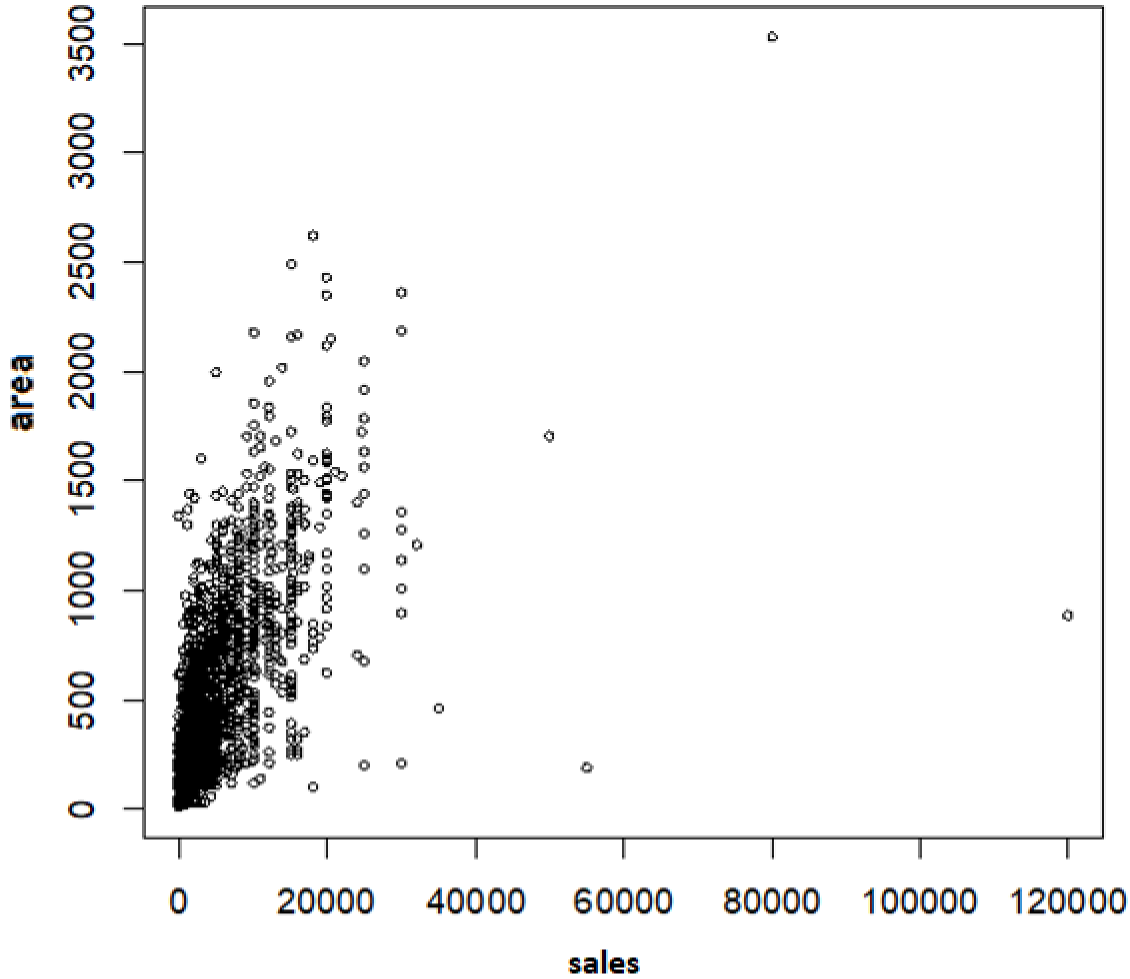
$\boldsymbol{y}$ – farm sales

$\boldsymbol{x}$ – farm area

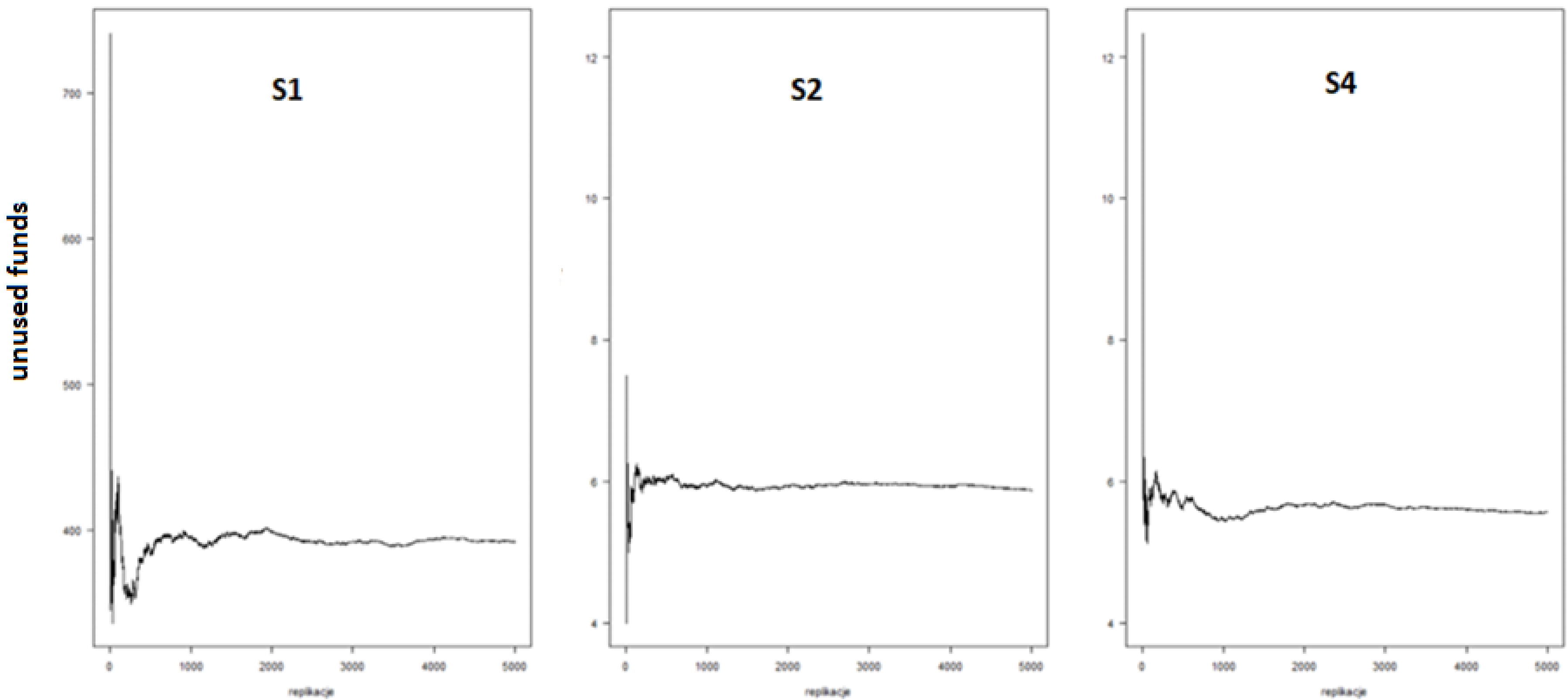
$\boldsymbol{c} \sim \boldsymbol{x}$

$B = 100\,000\ (8{,}62\%\ \tilde{x});\ 200\,000\ (17{,}24\%\ \tilde{x})$

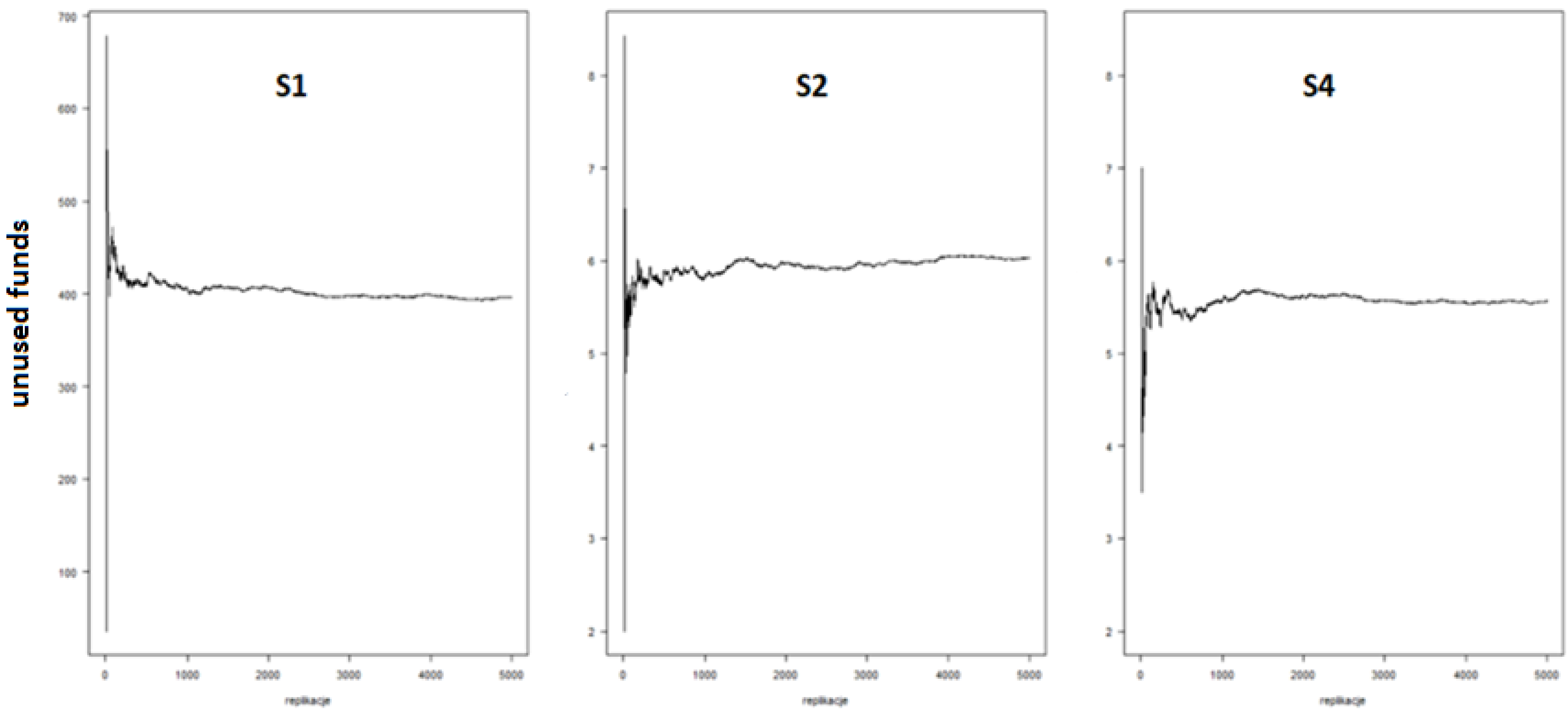
Area and sales



The average amount of unused funds (B=100 000)



The average amount of unused funds (B=200 000)



Introducing true auxiliary variable

$U = \{1, 2, 3, 4, 5, 6\}$
$\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5, y_6]$
$\boldsymbol{x} = [1, 1, 1, 1.492047, 1.492047, 2]$
$\boldsymbol{c} = [1, 1, 1, 2, 2, 3]$
$B = 6$

$$\boldsymbol{\pi} = (0.572, 0.572, 0.572, 0.572, 0.572, 0.638)$$

Introducing true auxiliary variable

$$U = \{1, 2, 3, 4, 5, 6\}$$
$$\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5, y_6]$$
$$\boldsymbol{x} = [1, 1, 1, 1.492047, 1.492047, 1.8]$$
$$\boldsymbol{c} = [1, 1, 1, 2, 2, 3]$$
$$B = 6$$

$$\boldsymbol{\pi} = (0.581, 0.581, 0.581, 0.587, 0.587, 0.606)$$

Example

$U = \{1,2,3,4,5,6\}$

$\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5, y_6]$

$\boldsymbol{x}(A, B) = [1,1,1, A, A, B]$

$\boldsymbol{c} = [1,1,1,2,2,3]$

$B = 6$

$$Q(A, B) = \sum_{i=1}^{6} |\pi_i(A, B) - \bar{\pi}(A, B)|$$

$$M = \min_{\substack{A=0.1,0.2,\ldots,10 \\ B=0.1,0.2,\ldots,10}} Q(A, B)$$

Example

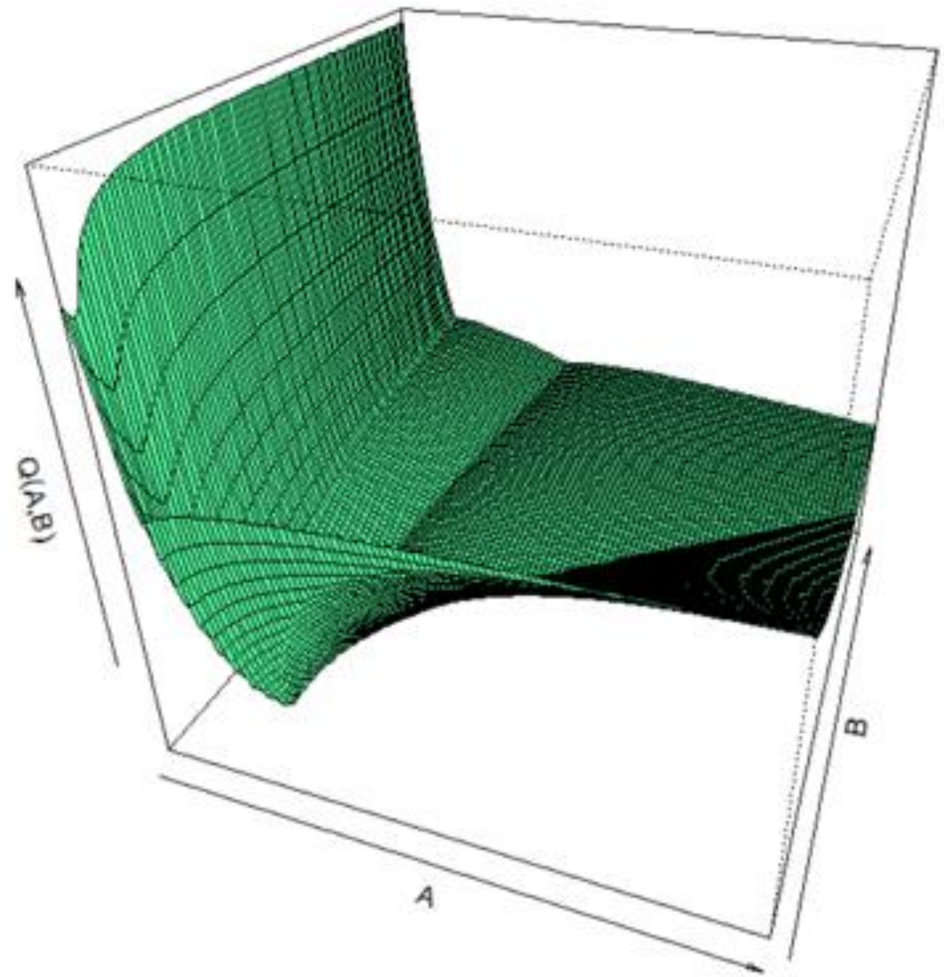
$U = \{1,2,3,4,5,6\}$
$\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5, y_6]$
$\boldsymbol{x}(A, B) = [1, 1, 1, A, A, B]$
$\boldsymbol{c} = [1,1,1,2,2,3]$
$B = 6$

$$M = \min_{\substack{A=0.1,0.2,\dots,10 \\ B=0.1,0.2,\dots,10}} \sum_{i=1}^{6} Q(A, B)$$

$$A = 1.5$$
$$B = 1.7$$
$$Q(1.5, 1.7) = M = 0.028$$

$$\boldsymbol{\pi}(A, B) = (0.586, 0.586, 0.586, 0.597, 0.597, 0.587)$$

The objective function



Empirical inclusion probabilities

$U = \{1,2,3,4,5,6\}$

$\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5, y_6]$

$\boldsymbol{x}(A, B) = [1,1,1, A, A, B]$

$\boldsymbol{c} = [1,1,1,2,2,3]$

$B = 6$

Number of replications $R = 500$

$$\hat{Q}(A, B) = \sum_{i=1}^{6} |\hat{\pi}_i(A, B) - \hat{\bar{\pi}}(A, B)|$$

$$EM = \min_{\substack{A=0.1,0.2,\dots,10 \\ B=0.1,0.2,\dots,10}} \hat{Q}(A, B)$$

Empirical inclusion probabilities

$U = \{1,2,3,4,5,6\}$
$\boldsymbol{y} = [y_1, y_2, y_3, y_4, y_5, y_6]$
$\boldsymbol{x}(A, B) = [1,1,1, A, A, B]$
$\boldsymbol{c} = [1,1,1,2,2,3]$
$B = 6$

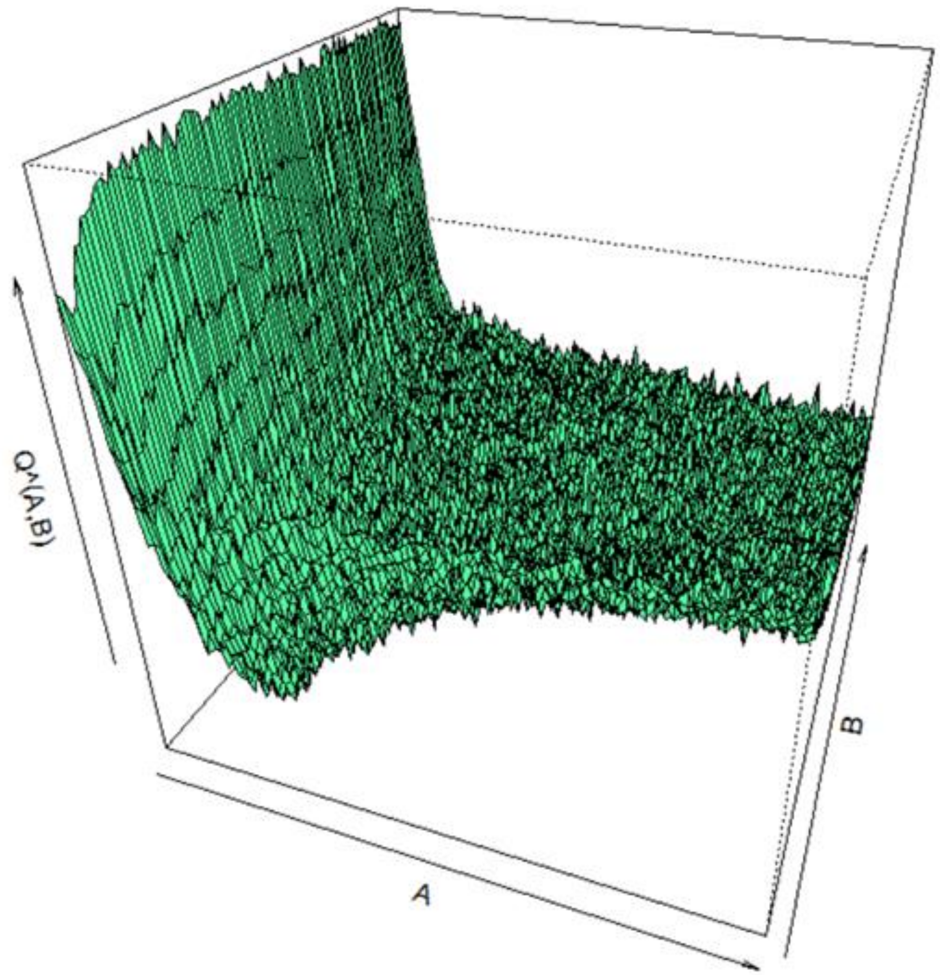
$$EM = \min_{\substack{A=0.1,0.2,\ldots,10 \\ B=0.1,0.2,\ldots,10}} \hat{Q}(A, B)$$

$A = 1.5$
$B = 1.7$
$Q(1.5, 1.7) = M = 0.028$
$Q(1.6, 1.7) = 0.085$

$A = 1.6$
$B = 1.7$
$\hat{Q}(1.6, 1.7) = EM = 0.046$
$\hat{Q}(1.5, 1.7) = 0.126$

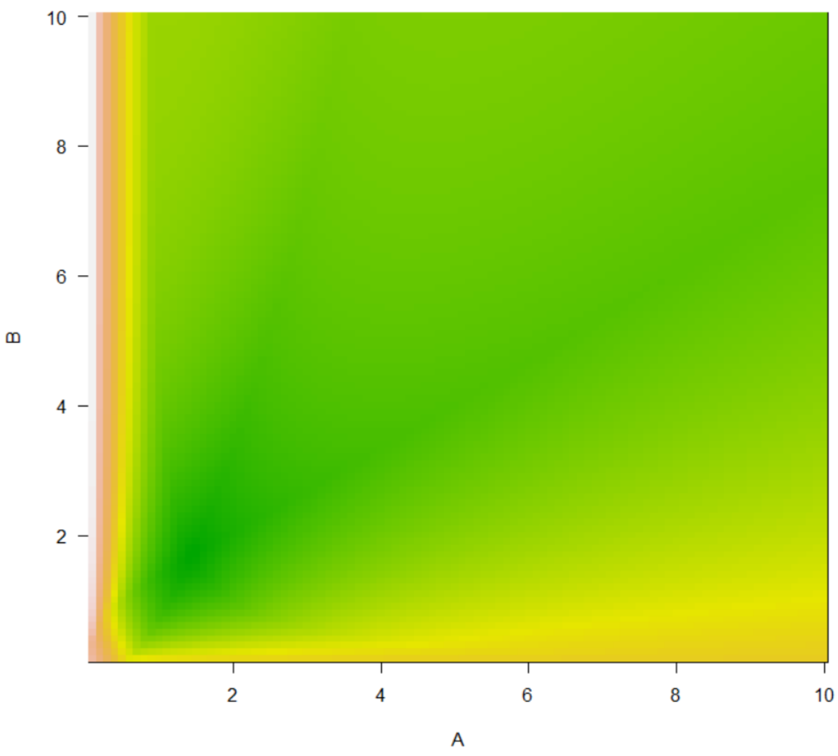
(Overall, 5,000,000 random selections)

The estimated objective function

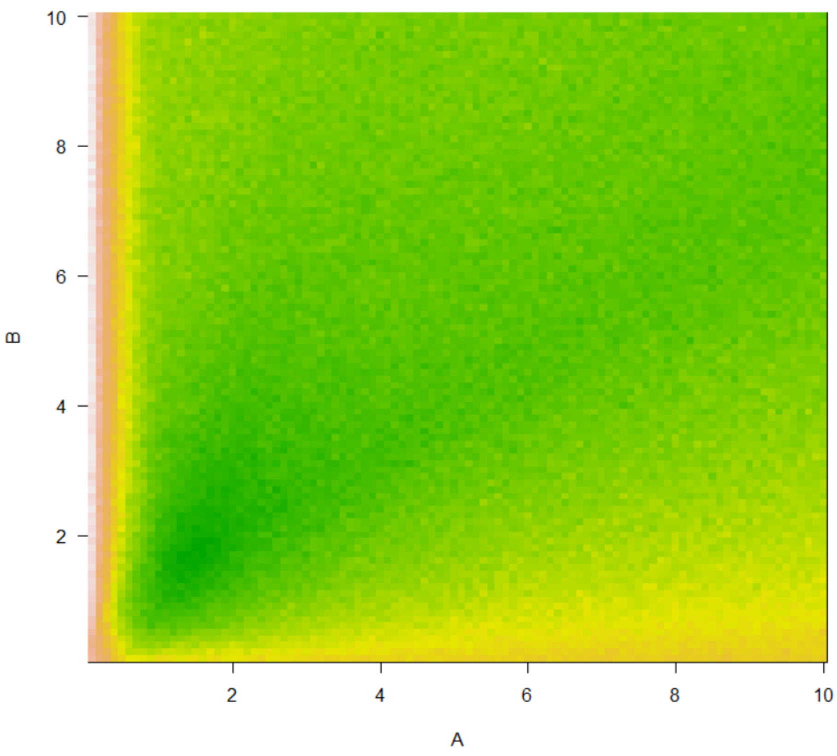


Comparison

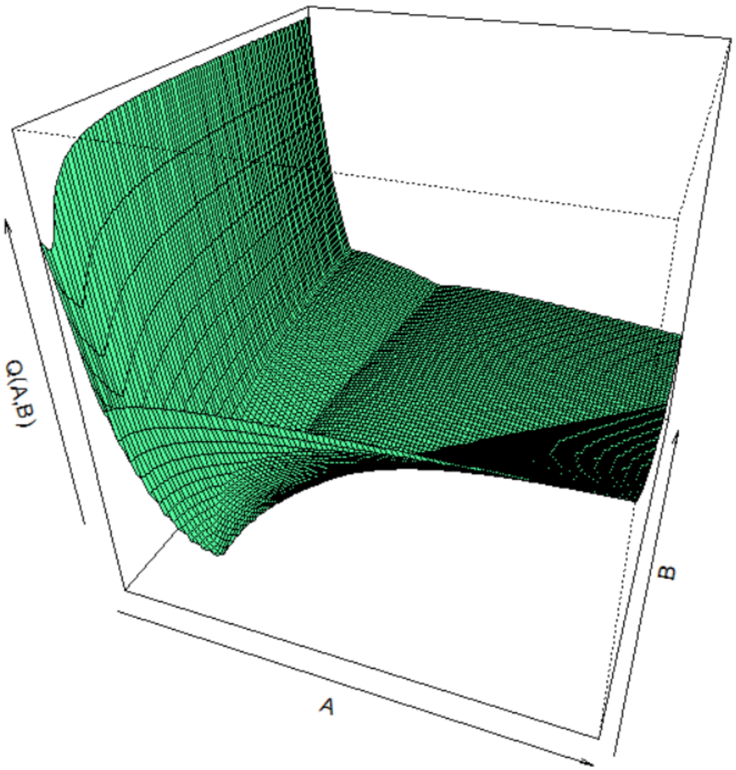
$Q(A,B)$

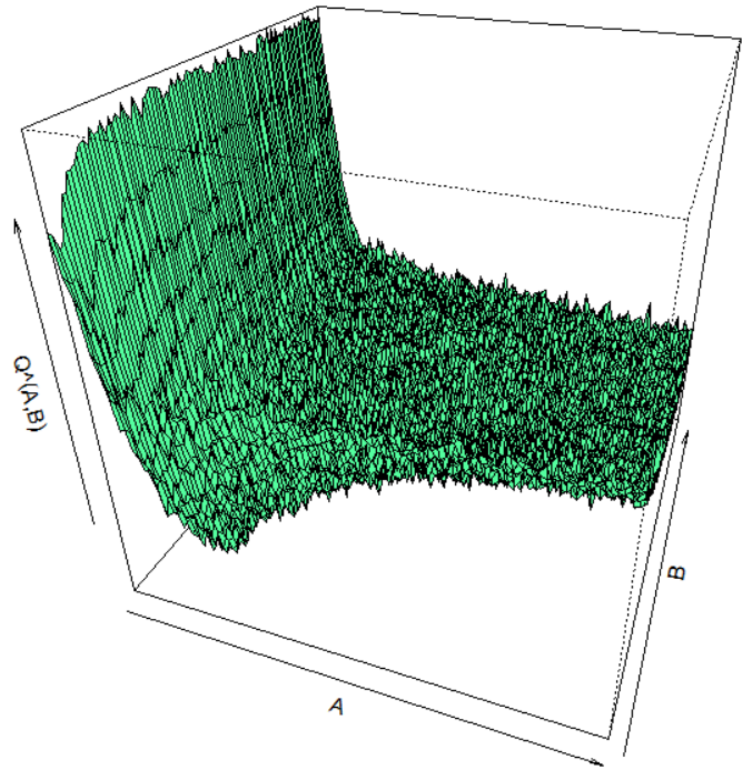


$\hat{Q}\ (A,B)$



Comparison





Summary

Challenges:

- How many sample replications to generate for calculating empirical inclusion probabilities?
- Is it possible to indicate better criteria than those proposed?
- How to find a solution in larger populations?

References

- Fattorini L. (2006): Applying the Horvitz-Thompson Criterion in Complex Design: A Computer-Intensive Perspective for Estimating Inclusion Probabilities, „Biometrica”, 93(2), 269-278
- Fattorini L. (2009): An Adaptive Algorithm for Estimating Inclusion Probabilities and Performing the Horvitz-Thompson Criterion in Complex Designs, „Computational Statistics”, 24, 623-639
- Gamrot W. (2014): Estymacja wartości przeciętnej uwzględniająca koszt pozyskania danych, Wydawnictwo UE, Katowice
- Horvitz D.G., Thompson D.J. (1952): A Generalization of Sampling Without Replacement from a Finite Universe, „Journal of the American Statistical Association”, 47, 663-685
- Narain R.D. (1951): On Sampling Without Replacement with Varying Probabilities, „Journal of the Indian Society of Agricultural Statistics”, 4, 169-175
- Pathak K. (1976): Unbiased Estimation in Fixed Cost Sequential Sampling Scheme, „Annals of Statistics”, 4(5), 1012-1017

Uniwersytet
Ekonomiczny
w Katowicach

www.ue.katowice.pl