Application of the Ranked Set Sampling to Agricultural Data

Rasa Beržvinskaitė
Statistics Lithuania

16-20 June, 2019, Örebro

Outline

- Theoretical background
- Data and simulation study
- Results

Introduction to RSS

Ranked set sampling (RSS) is an alternative to simple random smpling (SRS) when visual perception, judgment or any auxiliary information might be used to rank the objects of interest by their size without actual measurment.

The goal of RSS is to collect observations from a population that are more likely to span a full range of values in the population than the same number of observations obtained via simple random sampling.

The method was first proposed by McIntyre (1952) as a way to estimate mean pasture and forage yields. It was later improved and modified by numerous researchers.

Selecting a sample of size k

1. Select an SRS sample of size k (called *a set*) from the population of interest.
2. Rank the sample with respect to the variable of interest X without actual measurements:
$$X_{1[1]}, X_{1[2]}, \ldots, X_{1[k]}.$$
3. Choose the element with the smallest rank $X_{1[1]}$ for the actual measurement. $X_{1[1]}$ is called a judgement order statistics.
4. Repeat steps 1-3; at the j-th iteration the element with the j-th smallest rank is chosen for actual measurement.

Selecting a sample of size k

The final sample consists of the elements at the diagonal of this matrix:

$$\begin{pmatrix} \mathbf{X_1[1]} & X_{1[2]} & X_{1[3]} & \cdots & X_{1[k]} \\ X_{2[1]} & \mathbf{X_2[2]} & X_{2[3]} & \cdots & X_{2[k]} \\ X_{3[1]} & X_{3[2]} & \mathbf{X_3[3]} & \cdots & X_{3[k]} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ X_{k[1]} & X_{k[2]} & X_{k[3]} & \cdots & \mathbf{X_k[k]} \end{pmatrix}$$

The procedure of steps 1-4 is called *a cycle*. If we want so draw a sample of size $n = mk$, we need to perform m cycles of a set size k.

RSS from finite populations

There are three different designs:

- ***level*-0.** Units in a given set are selected without replacement, but all units in the set (including the measured one) are replaced back into the population prior to selection of a next set.
- ***level*-1.** It has the same replacement policy as the *level*-0 design except that the element selected for full measurement is not returned to the population.
- ***level*-2.** None of the units in a set are replaced back into the population.

Estimators

N – population size;
k – set size;
m – number of cycles;
$X_{[i]j}$ – i-th judgement order statistics, of j-th cycle.

$$\hat{\mu}_{RSS} = \sum_{j=1}^{m} \sum_{i=1}^{k} \frac{X_{[i]j}}{km};$$

$$\hat{t}_{RSS} = N \sum_{j=1}^{m} \sum_{i=1}^{k} \frac{X_{[i]j}}{km} = N\hat{\mu}_{RSS}.$$

Survey data

Data from Statistical survey of agricultural crop area, harvest and yield of the year 2015 were used for simulation. Current sample design - stratified SRS.

There are four main variables at this survey:

- sown and planted area (D1);
- harvested area (D2);
- weight of the yield before cleaning and drying (D3);
- weight of the yield after cleaning and drying (D4).

Survey data

There are 140 different species of crops and plants. Parameters that need to be estimated are population totals of these four variables for all the species. There are many zero values in the data set.

Only most important 15 species were used for the simulation.

Data were collected from 6346 farms which were considered as the entire population for the simulation.

Ranking variable

The standard output of an agricultural product (crop), is the average monetary value of the agricultural output at farm-gate price, in euro per hectare.

The sum of all the SO per hectare of crop in a farm is a measure of its overall economic size, expressed in euro.

The SO value of each farm was used to rank the population elements with respect to their size.

Groups of correlation

Variables of different species were grouped into 6 groups with respect to their correlation to the ranking variable:

Group number	Correlation
1	Strong positive (0,7–1)
2	Moderate positive (0,5–0,7)
3	Weak positive (0,2–0,5)
4	Very weak positive (0–0,2)
5	Significantly does not differ from 0
6	Very weak negative (0– -0,2)

Simulation

There were samples of the size $n = 100$ drawn from the population using three methods: SRS, RSS *level*-1 and *level*-2. Different method parameters were used:

Number of cycles m	Set size k
1	100
2	50
4	25
5	20
10	10

25000 iterations were used with each set of parameters.

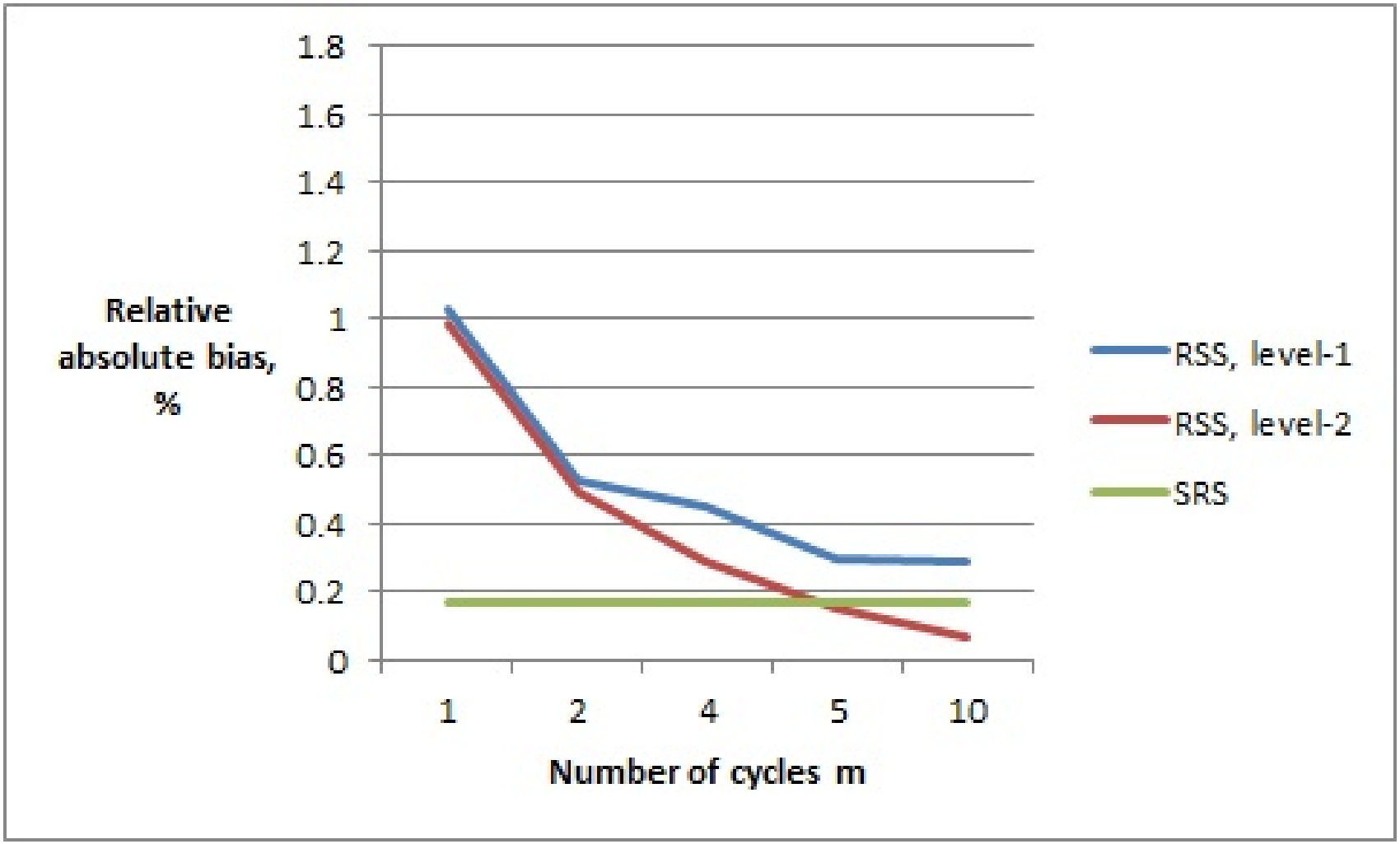
Measures of accuracy

Let's denote the true population total as t and the estimate of population total of the iteration i as $\hat{t}_i$.

- Relative absolute bias: $e_{rel} = \frac{|\frac{1}{25000}\sum_{i=1}^{25000}\hat{t}_i - t|}{t} \cdot 100\%$
- Root mean squared error: $RMSE = \sqrt{\frac{1}{25000}\sum_{i=1}^{25000}(\hat{t}_i - t)^2}$
- Relative RMSE: $RMSE_{rel} = \frac{RMSE_{RSS}}{RMSE_{SRS}} \cdot 100\%$

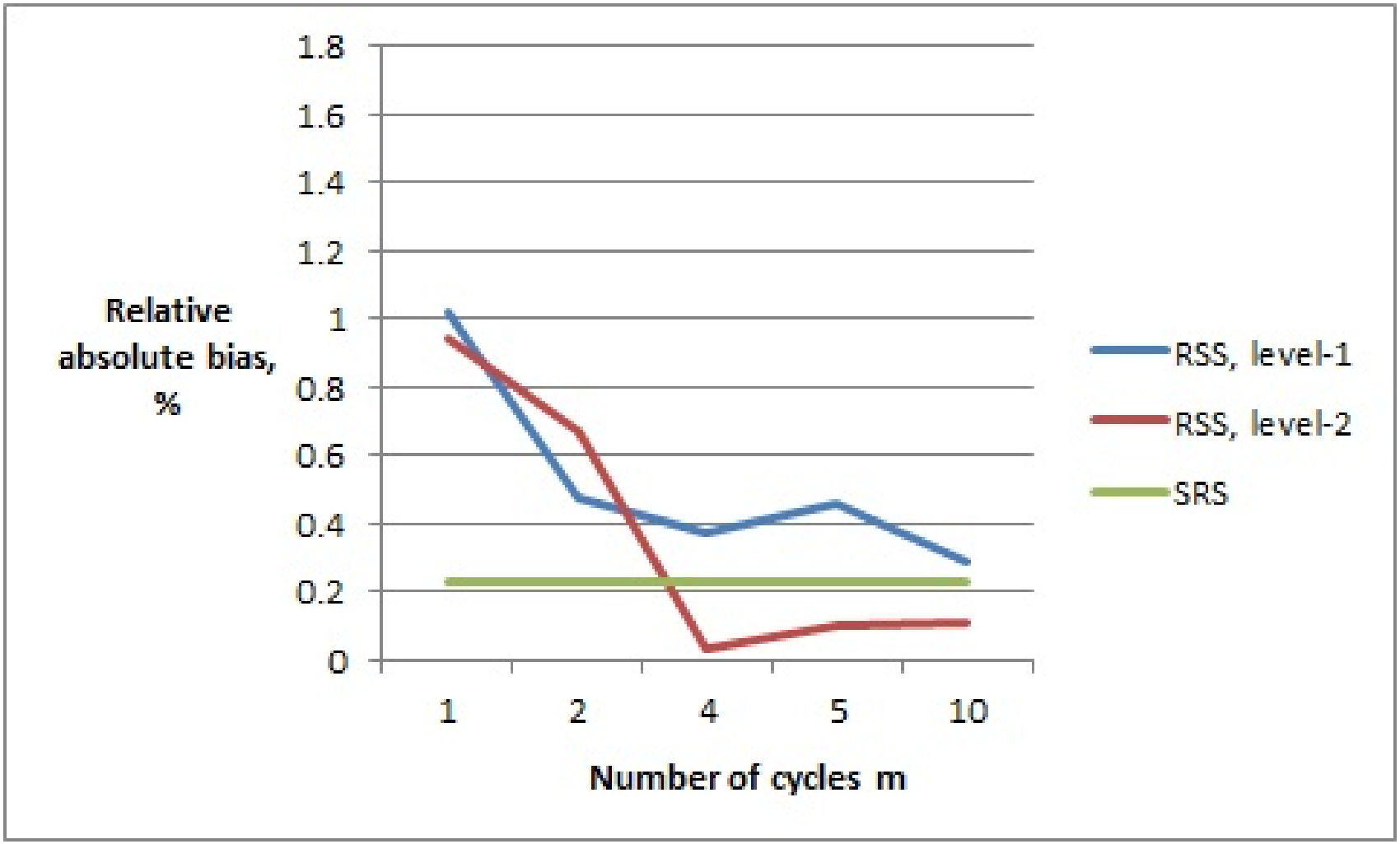
Strong positive correlation (group 1)

www.stat.gov.lt



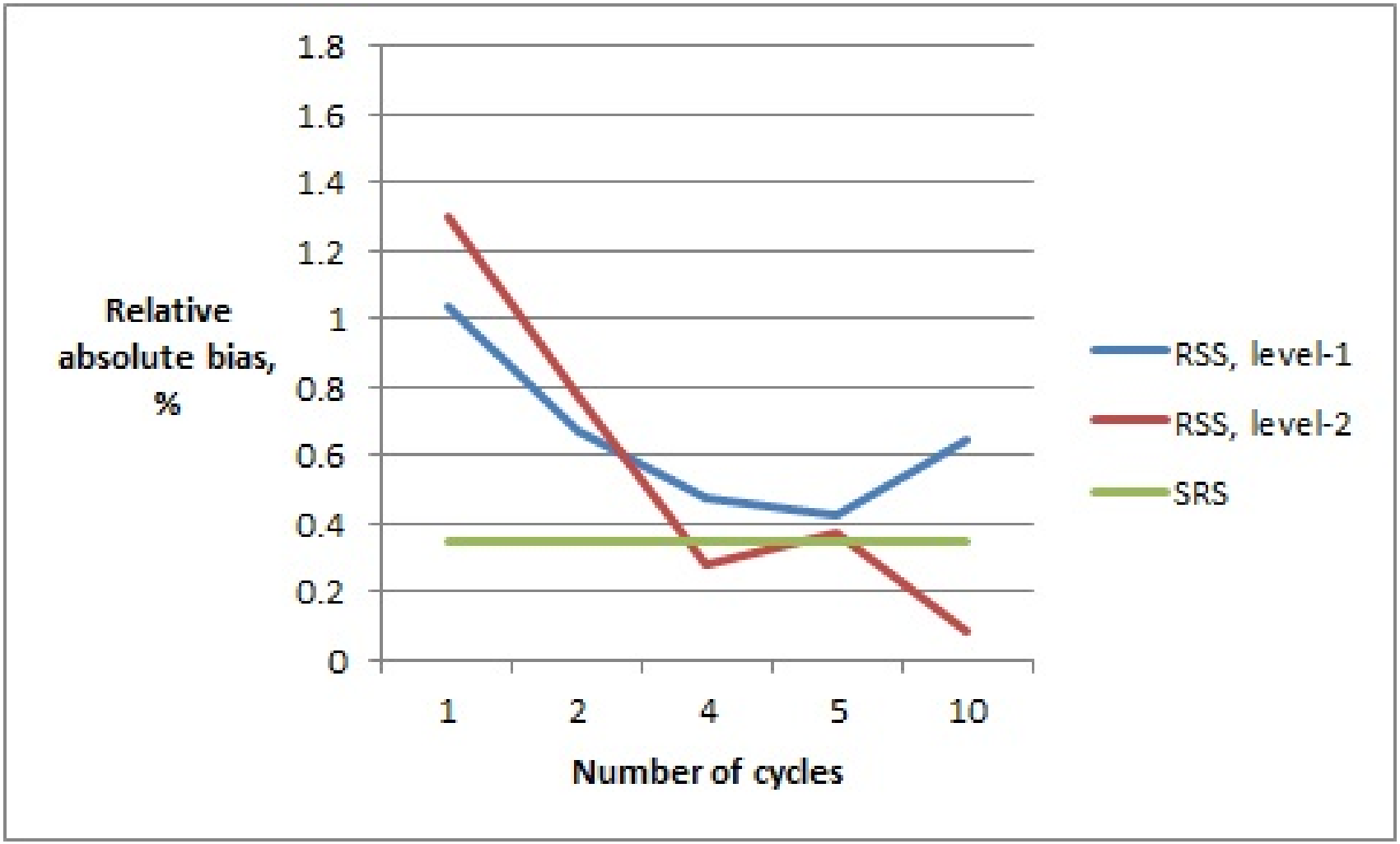
Moderate positive correlation (group 2)

www.stat.gov.lt



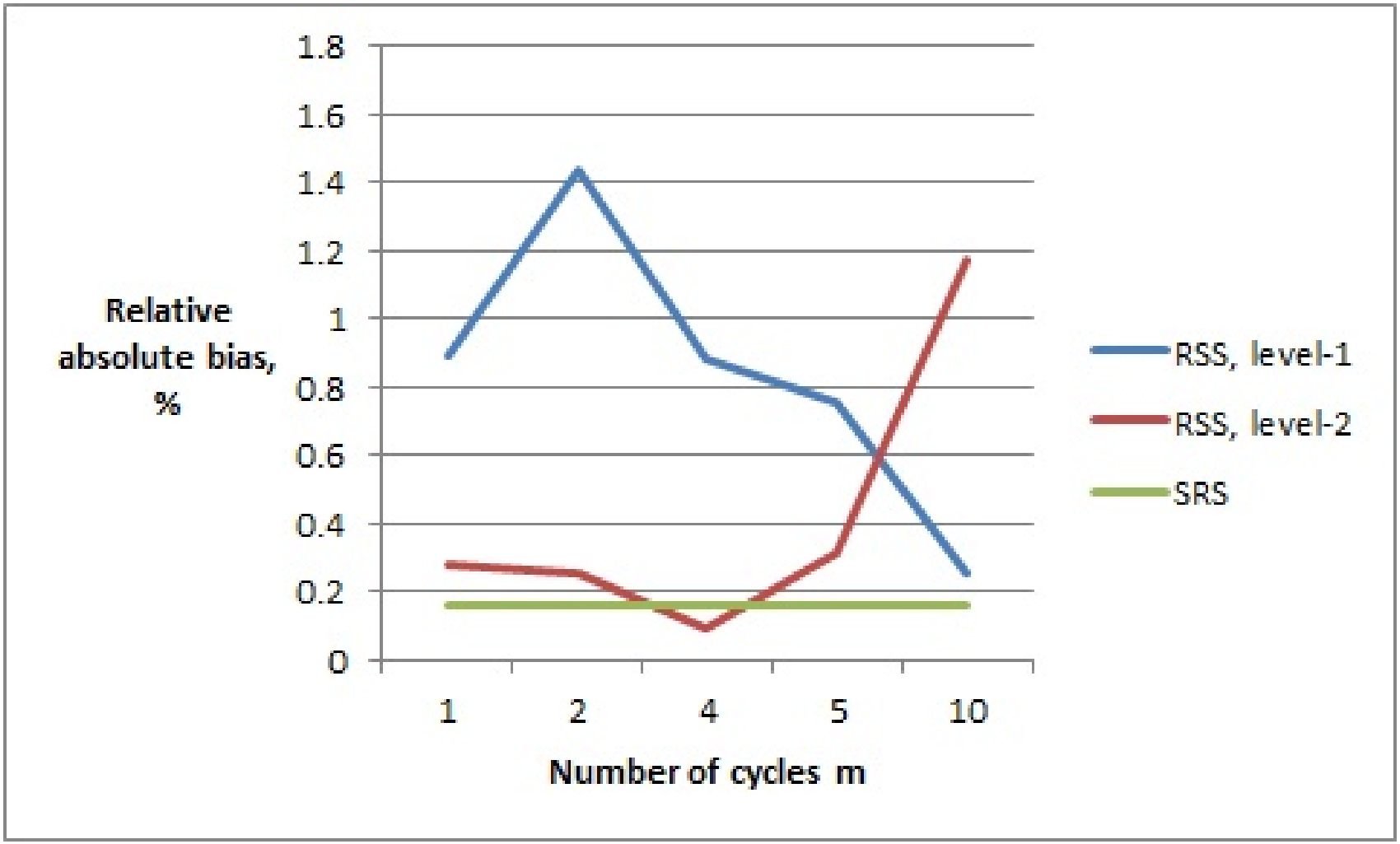
Weak positive correlation (group 3)

www.stat.gov.lt

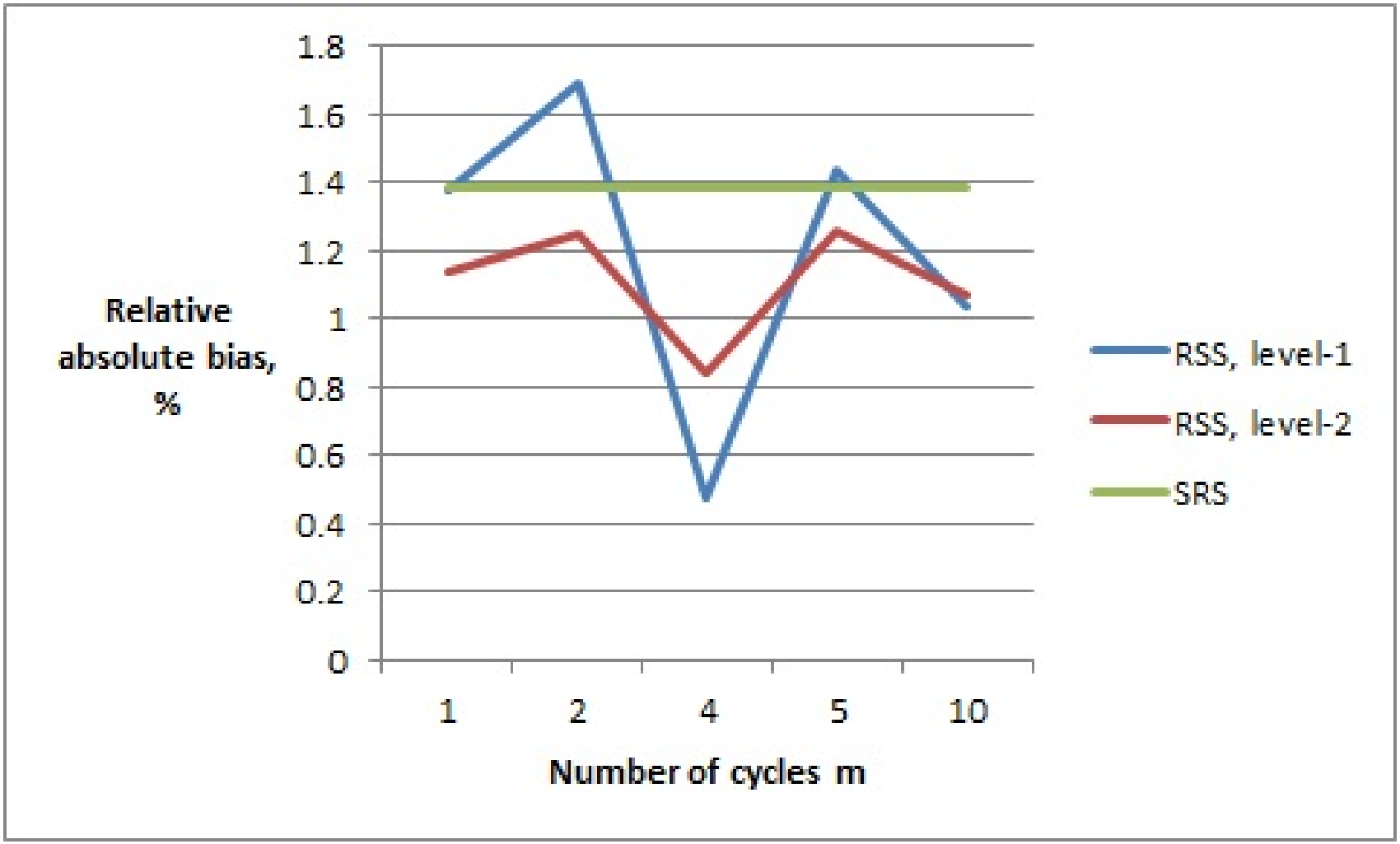


Very weak positive correlation (group 4)

www.stat.gov.lt

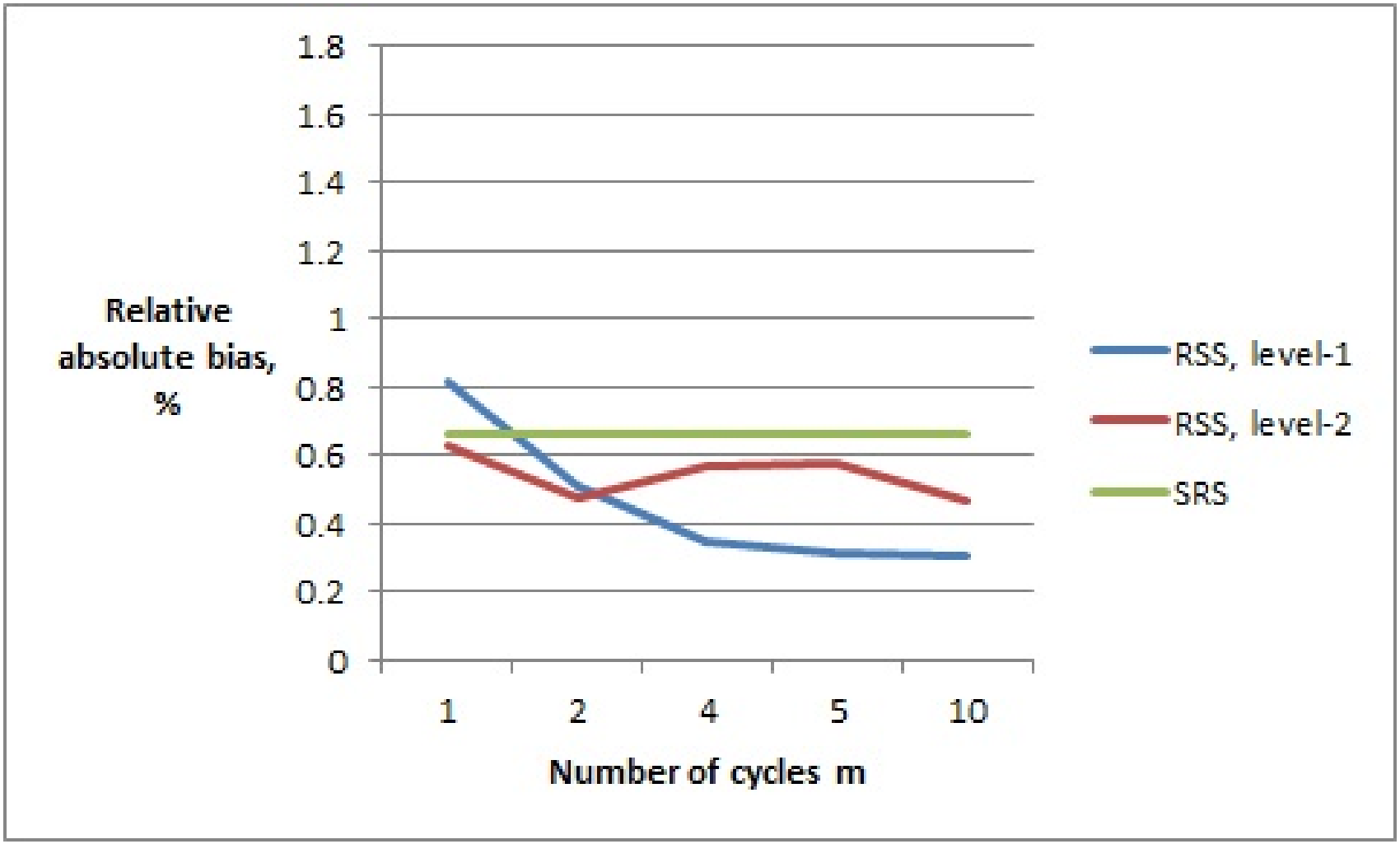


No correlation (group 5)



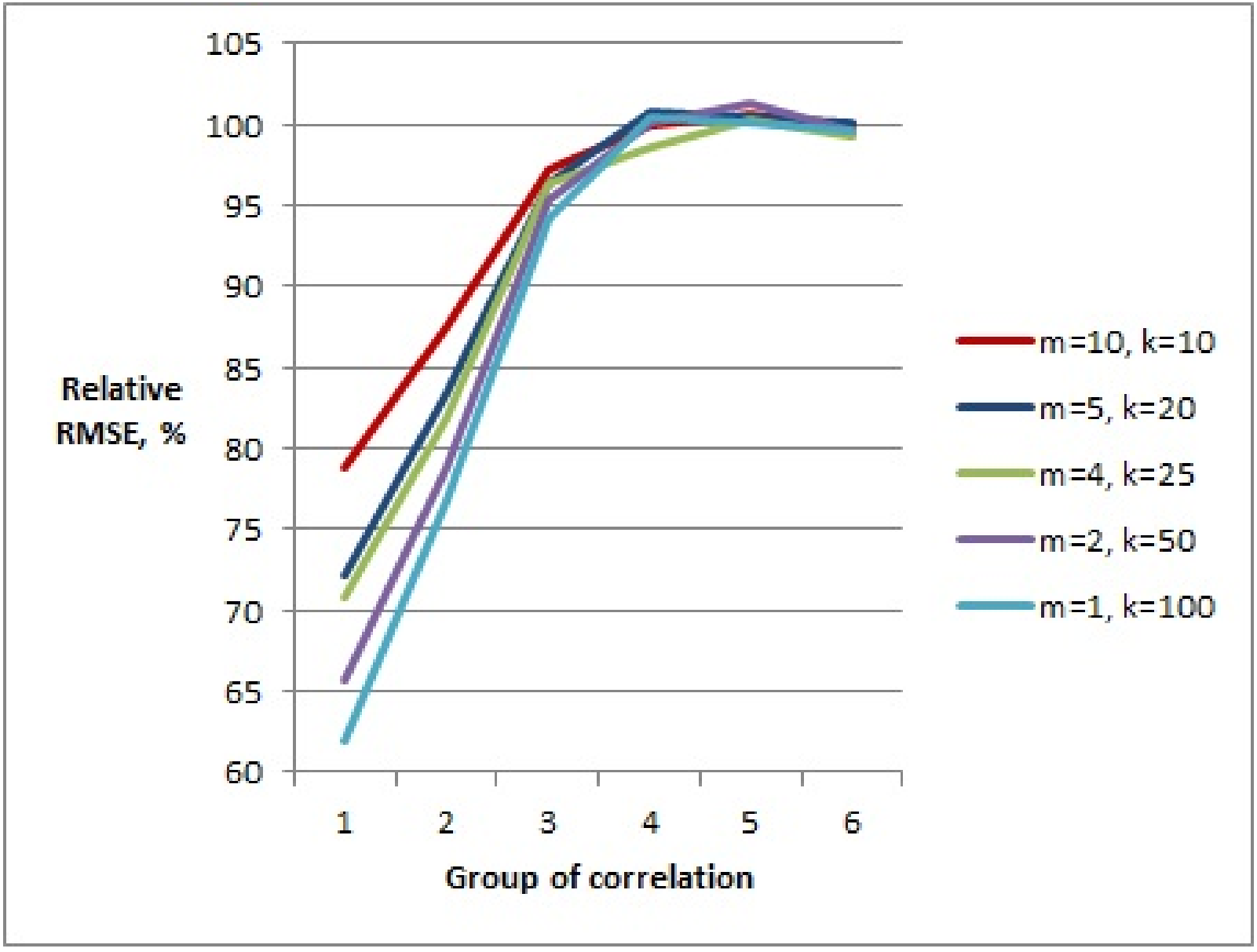
Very weak negative correlation (group 6)

www.stat.gov.lt

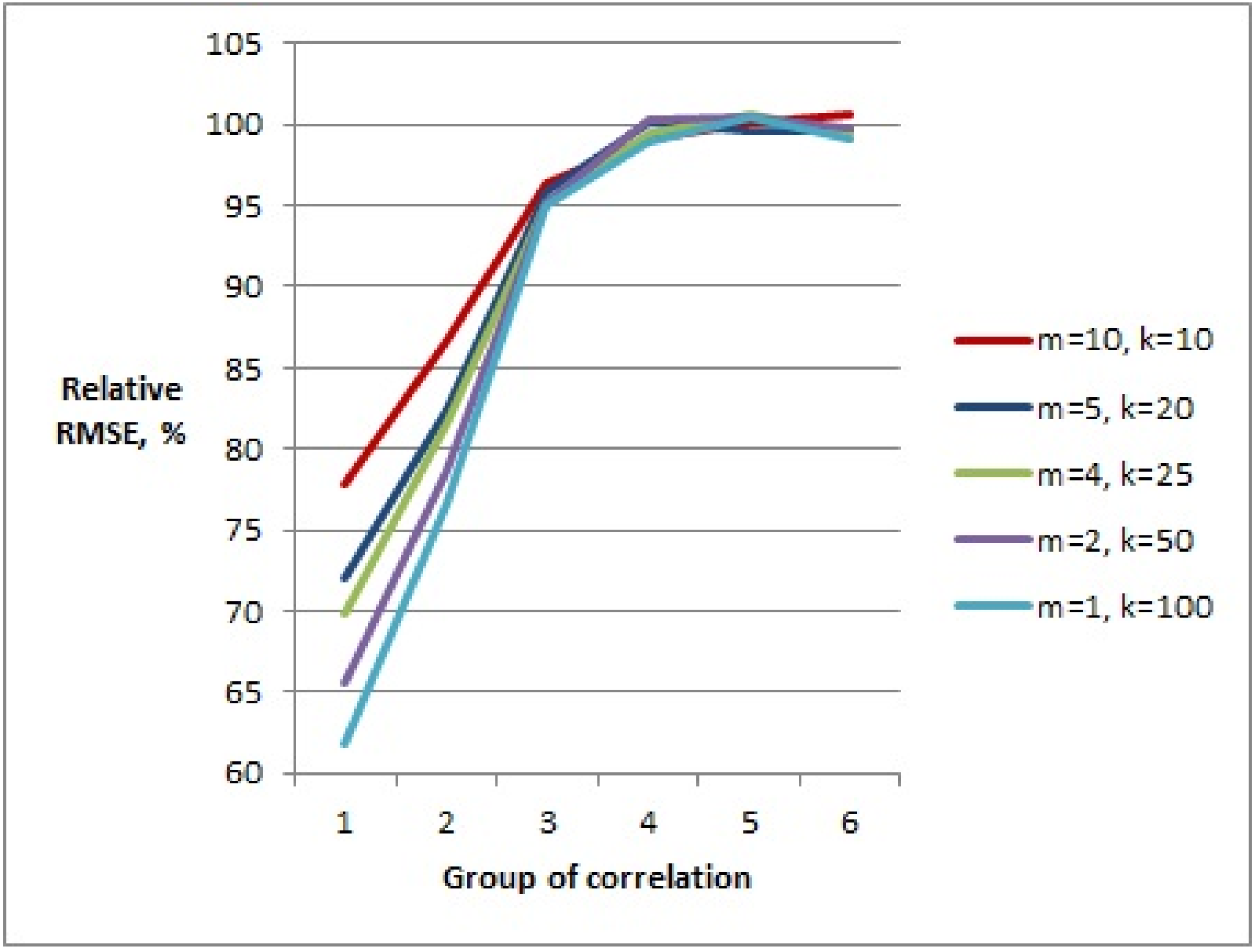


$RMSE_{rel}$, *level*-1

www.stat.gov.lt



$RMSE_{rel}$, level-2



Conclusion

- RSS estimators of both levels have significantly smaller RMSE than SRS estimator, when correlation between ranking and study variables is strong enough.
- RSS estimators with 1–2 cycles are superior to those with the greater number of cycles.
- RSS *level*-2 estimators are a little bit more accurate than *level*-1 estimators.

Thank You!