Journal of Statistical Theory and Practice (2021) 15:61
https://doi.org/10.1007/s42519-021-00196-x

ORIGINAL ARTICLE

The Influence of Clustering Population on Estimation Accuracy of Population Totals Vector

Janusz L. Wywiał[1] · Grzegorz Sitek[1]

Accepted: 16 April 2021 / Published online: 21 May 2021

Abstract
The estimation accuracy of a population totals vector based on a simple cluster sample is considered. The variance-covariance matrix of estimators depends on the intra-cluster spread of variables under study. The spread depends on the partition of the population into clusters. The variance-covariance matrix is evaluated under several variants of clustering algorithm. This lets us find the clustering algorithm providing the most accurate estimation of population vector totals.

Keywords Vector estimation · Cluster sample · Clustering algorithms · Homogeneity coefficient · Generalized variance · Spectral radius

1 Introduction

Research into practical survey sampling is usually based on vector parameters. The purpose of this paper is to simultaneously estimate the population totals of at least two variables. The well-known vector estimator from a simple cluster sample drawn without replacement is considered. Its accuracy is compared with the ordinary vector estimator from a simple random sample drawn without replacement using the variance-covariance matrix. We analyse accuracy of vector estimators using the methodology proposed by Borovkov [2], Jensen [4], Rao [7] and the generalized relative efficiency coefficient proposed by Rao [8]. This coefficient is defined as the maximal eigenvalue of the product of two matrices. In our case, one of the matrices is the variance-covariance matrix of the vector estimator from the cluster sample and the other is the inverse of the variance-covariance matrix of the vector estimator from the simple random sample. Let us add that Wywial [14] analysed the

✉ Janusz L. Wywiał
janusz.wywial@ue.katowice.pl

Grzegorz Sitek
grzegorz.sitek@ue.katowice.pl

[1] Department of Statistics, Econometrics and Mathematics, University of Economics in Katowice, Katowice, Poland

accuracies of estimation procedures of the total based on samples selected according to several sampling designs from a population clustered using several algorithms. This paper is in some sense a generalization of their considerations into simultaneous estimation of totals of at least two variables under study.

The accuracy of the vector estimation of totals based on the cluster sample depends on partitioning a population into clusters. The influence of the partitioning a population on accuracy of the estimation is considered. This problem was analysed by means of the several methods of measuring vector estimation.

The obtained results should be useful in survey sampling conducted, e.g. by statistical offices. In this case, census data are usually available. Variables under study (observed during a census) can be used as auxiliary variables in a survey sampling on a subsequent occasion. In this situation, appropriate clustering of the population considered in this paper should contribute to improving future sampling strategies.

One of the aspects of big-data analysis is the problem of reduction in the data (observations of variables) number. The results of this paper partially contribute to solve this problem because considered methods provide partitions of a population into such clusters that each of them is similar to the population as close as it is possible. More precisely, the spreads of data observations in the clusters are not less then the spread of the data in the population.

The main results of this paper are as follows:

- the variance-covariance matrix of the vector estimation of totals from the simple cluster sample is shown as a function of the matrix of homogeneity coefficients, Sect. 2.2 and “Appendix”,
- properties of the homogeneity matrix let us show when the vector estimator from the cluster sample is more accurate than the vector estimator from the simple sample , Sect. 2.3 and “Appendix”,
- several algorithms for partitioning a population into mutually disjoint and non-empty clusters are proposed, Sect. 2.4,
- using these algorithms to partition of the population of Swedish municipalities into clusters, Sect. 3,
- for theses partitions values of the generalized coefficient of the relative efficiency of the vector estimator from the cluster sample are evaluated what let us analyse influence the partition of the population on the accuracy estimation, Sect. 3.

2 Estimation Based on Cluster Sample

2.1 Basic Notations

Let U be a population of size N. The number of variables observed in U is denoted by m. Observations of a vector variable will be denoted by $\boldsymbol{y}_k = [y_{k,1} \dots y_{k,m}]$ where $k \in U$. Let us assume that the population is partitioned into disjoint subpopulation U_h of sizes N_h, $h = 1, \dots, G$, called clusters. Hence, $N = \sum_{h=1}^{G} N_h$ and $\bar{N} = G^{-1} \sum_{h=1}^{G} N_h$. Let $\bar{N} = M$, if all the clusters are of the same size. Let

$\mathcal{U} = \{U_1, ..., U_h, ..., U_G\}$ be a partition of the population elements into clusters. Hence, $\mathcal{U}$ is the set of G-mutually disjoint and non-empty clusters. Let

$$\bar{\boldsymbol{y}} = [\bar{y}_1 ... \bar{y}_m] = \sum_{k \in U} \boldsymbol{y}_k / N, \quad \boldsymbol{y}_U = N\bar{\boldsymbol{y}} = \sum_{k \in U} \boldsymbol{y}_k = [y_{U,1} ... y_{U,m}],$$

$$y_{U,i} = \sum_{k \in U} y_{k,i}, \quad \boldsymbol{C} = [c_{i,j}], \quad c_{i,j} = \sum_{k \in U} (y_{k,i} - \bar{y}_i)(y_{k,j} - \bar{y}_j)/(N-1),$$

$$\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{C}\boldsymbol{D}^{-1/2} = [r_{i,j}], \quad \boldsymbol{D} = [v_i], \quad r_{i,j} = \frac{c_{i,j}}{\sqrt{v_i v_j}}, \quad v_i = c_{i,i}$$

where $\bar{\boldsymbol{y}}$ is the vector of population means, $\boldsymbol{y}_U$ is the vector of totals, $\boldsymbol{C}$ is the matrix of variances and covariances, $\boldsymbol{R}$ is the matrix of correlation coefficients and $\boldsymbol{D}$ is the diagonal matrix of variances. Moreover, let

$$\bar{\boldsymbol{y}}_{U_h} = \sum_{k \in U_h} \boldsymbol{y}_k / N_h, \quad \bar{\boldsymbol{y}}_{U_h} = [\bar{y}_{U_h,1} ... \bar{y}_{U_h,m}], \quad \bar{y}_{U_h,i} = \sum_{k \in U_h} y_{k,i}/N_h,$$

$$\boldsymbol{y}_{U_h} = N_h \bar{\boldsymbol{y}}_{U_h} = \sum_{k \in U_h} \boldsymbol{y}_k = [y_{U_h,1} ... y_{U_h,m}], \quad y_{U_h,i} = \sum_{k \in U_h} y_{k,i},$$

$$\boldsymbol{C}_{U_h} = [c_{U_h,i,j}], \quad c_{U_h,i,j} = \sum_{k \in U_h} (y_{k,i} - \bar{y}_{U_h,i})(y_{k,j} - \bar{y}_{U_h,j})/(N_h - 1),$$

$$\bar{\boldsymbol{y}}_{\mathcal{U}} = \sum_{h=1}^{G} \boldsymbol{y}_{U_h}/G = \boldsymbol{y}_U/G = [\bar{y}_{\mathcal{U},1} ... \bar{y}_{\mathcal{U},m}], \;\; \bar{y}_{\mathcal{U},i} = \sum_{h=1}^{G} y_{U_h,i}/G = y_{U,i}/G,$$

$$\boldsymbol{y}_{\mathcal{U}} = G\bar{\boldsymbol{y}}_{\mathcal{U}} = \sum_{h=1}^{G} \boldsymbol{y}_{U_h} = \boldsymbol{y}_U, \quad \boldsymbol{C}_{\mathcal{U}} = [c_{\mathcal{U},i,j}],$$

$$c_{\mathcal{U},i,j} = \sum_{h=1}^{G} (y_{U_h,i} - \bar{y}_{\mathcal{U},i})(y_{U_h,j} - \bar{y}_{\mathcal{U}}, j)/(G-1)$$

where $k \in U$, $i = 1, ..., m$, $j = 1, ..., m$, $h = 1, ..., G$. $\bar{y}_{U_h,i}$ is the mean of the i-th variable in the h-cluster, $y_{U_h,i}$ is the total of the i-th variable in the h-th cluster, $\boldsymbol{C}_h$ is the variance-covariance matrix in the h-th cluster, and $\bar{\boldsymbol{y}}_{\mathcal{U}}$ is the vector of the means of the cluster totals, $\boldsymbol{C}_{\mathcal{U}}$ is the variance-covariance matrix of the cluster totals.

2.2 Simple Cluster Sampling

Let sample s be drawn from population U or partition $\mathcal{U}$. The random sample will be denoted by the capital letter S while its observation by s.

Cluster sample s is defined as a g-element set of clusters U_h drawn from partition $\mathcal{U}$. The well-known simple cluster sampling design is defined as $P_1(s) = \binom{G}{g}^{-1}$ where $s \in \boldsymbol{S}_{\mathcal{U}}$ and $\boldsymbol{S}_{\mathcal{U}}$ is the sampling space generated for set $\mathcal{U}$. The vector of the unbiased estimator of the population total $\boldsymbol{y}_U$ is as follows:

$$\tilde{\boldsymbol{y}}_S = \frac{G}{g}\sum_{h\in S}\sum_{k\in U_h}\boldsymbol{y}_k = \frac{G}{g}\sum_{h\in S}\boldsymbol{y}_{U_h}, \tag{1}$$

Its variance-covariance matrix is:

$$\boldsymbol{V}(\tilde{\boldsymbol{y}}_S) = \frac{G(G-g)}{g}\boldsymbol{C}_{\mathcal{U}} \tag{2}$$

where $\tilde{\boldsymbol{y}}_S$ is evaluated on the basis of the simple cluster sample drawn without replacement. Generalizing the results of [9, pp. 129–133] into multidimensional case, we derived in "Appendix" the following expression (see also [11, 12]):

$$\boldsymbol{V}(\tilde{\boldsymbol{y}}_S) = \frac{G(G-g)}{g}\bar{N}\boldsymbol{C}\left(\boldsymbol{I}_m + \frac{N-G}{G-1}\boldsymbol{\Delta}\right) + \frac{G(G-g)}{g}\boldsymbol{A} \tag{3}$$

where:

$$\begin{aligned} &\boldsymbol{\Delta} = \boldsymbol{I}_m - \boldsymbol{C}^{-1}\boldsymbol{C}_*,\\ &\boldsymbol{A} = [a_{i,j}]; \quad a_{i,j} = \frac{1}{G-1}\sum_{h=1}^{G}(N_h - \bar{N})N_h\bar{y}_{U_h,i}\bar{y}_{U_h,j}, \quad i \neq j = 1,\dots,m \end{aligned} \tag{4}$$

or $\boldsymbol{A} = \boldsymbol{A}_1 + \boldsymbol{A}_2 + \boldsymbol{A}_3$,

$$\begin{aligned} &\boldsymbol{A}_1 = [a_{i,j}(111)], \quad \boldsymbol{A}_2 = [\bar{y}_{\mathcal{U},i}a_{i,j}(101)], \quad \boldsymbol{A}_3 = [\bar{y}_j a_{i,j}(110)],\\ &a_{i,j}(bed) = \frac{1}{G-1}\sum_{h=1}^{G}(N_h - \bar{N})^b(y_{U_h,i} - \bar{y}_{\mathcal{U},i})^e(\bar{y}_{U_h,j} - \bar{y}_j)^d,\\ &\boldsymbol{C}_* = [c_{*i,j}], \quad c_{*i,j} = \frac{1}{N-G}\sum_{h=1}^{G}\sum_{k\in U_h}(y_{k,i} - \bar{y}_{U_h,i})(y_{k,j} - \bar{y}_{U_h,j}), \end{aligned} \tag{5}$$

or

$$c_{*i,j} = \sum_{h=1}^{G} w_h c_{*,U_h,i,j}, \quad c_{*,U_h,i,j} = \frac{1}{N_h - 1}\sum_{k\in U_h}(y_{k,i} - \bar{y}_{U_h,i})(y_{k,j} - \bar{y}_{U_h,j})$$

and $w_h = \frac{N_h-1}{N-G}$. Parameter $\boldsymbol{\Delta}$ is the matrix of the coefficients of intra-cluster data spread homogeneity or simply the homogeneity matrix. The intra-cluster variance-covariance matrix is denoted by $\boldsymbol{C}_*$. Let us underline that when $N_h = M$ for all $h = 1,\dots,H$, then $\boldsymbol{A} = \boldsymbol{O}$. Sarndal et al. [9] proved that all diagonal elements of $\boldsymbol{\Delta}$ take values from $\left[-\frac{G-1}{N-G};1\right]$. Let δ be an eigenvalue of $\boldsymbol{\Delta}$. In the last part of the "Appendix" is proved the following inequality:

$$-\frac{G-1}{N-G} \leq \delta \leq 1. \tag{6}$$

Kish [5] provided sound advices on grouping problems that might be encountered in practical surveys.

2.3 Relative Efficiency

Let $\boldsymbol{t}_{1s}$ and $\boldsymbol{t}_{2s}$ be the unbiased estimators of vector parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Borovkov [2] proposed comparing the accuracy of vector estimators using the following definition (see also [7] or [12], pp. 28–29):

Definition 1 *Estimator* $\boldsymbol{t}_{1s}$ *is not worse than* $\mathbf{t}_{2s}$*if and only if:*

$$\forall_{\boldsymbol{\alpha}\neq\mathbf{0}}\forall_{\theta\in\Theta} \;\; v(\boldsymbol{t}_{1s}\boldsymbol{\alpha}^T) \leq v(\boldsymbol{t}_{2s}\boldsymbol{\alpha}^T)$$

where $\boldsymbol{\alpha} = [\alpha_1 ... \alpha_m]$,

$$v(\boldsymbol{t}_{2s}\boldsymbol{\alpha}^T) = \boldsymbol{\alpha}\boldsymbol{V}(\boldsymbol{t}_{is})\boldsymbol{\alpha}^T, \quad i = 1, 2.$$

Estimator $\mathbf{t}_{1s}$ *is better than* $\mathbf{t}_{2s}$ *if and only if* $\boldsymbol{t}_{1s}$ *is not worse than* $\mathbf{t}_{2s}$*and the above inequality becomes sharp for at least one fixed parameter* $\boldsymbol{\theta}$.

This definition directly leads to the following, see [7] and Borovkov [2]:

Theorem 1 *Let the variance-covariance matrices* $\boldsymbol{V}(\boldsymbol{t}_{is})$, $i = 1, 2$ *be positive definite. If estimator* $\mathbf{t}_{1s}$ *is not worst than* $\mathbf{t}_{2s}$ *, then* $\boldsymbol{V}(\boldsymbol{t}_{2s}) - \boldsymbol{V}(\boldsymbol{t}_{1s})$ *is non-negative definite and:*

$$\begin{aligned}
&tr\big(\boldsymbol{V}(\boldsymbol{t}_{1s})\big) \leq tr\big(\boldsymbol{V}(\boldsymbol{t}_{2s})\big),\\
&det\big(\boldsymbol{V}(\boldsymbol{t}_{1s})\big) \leq det\big(\boldsymbol{V}(\boldsymbol{t}_{2s})\big),\\
&\lambda\big(\boldsymbol{V}(\boldsymbol{t}_{1s})\big) \leq \lambda\big(\boldsymbol{V}(\boldsymbol{t}_{2s})\big),\\
&\forall_{j=1,\ldots,m} \;\; v(t_{1,is})) \leq v(t_{2,is})
\end{aligned}$$

where $tr\big(\boldsymbol{V}(\boldsymbol{t}_{is})\big)$, $det\big(\boldsymbol{V}(\boldsymbol{t}_{is})\big)$ and $\lambda\big(\boldsymbol{V}(\boldsymbol{t}_{is})\big)$ *are called the mean square radius, the generalized variance and the the spectral radius (maximal eigenvalue of* $\boldsymbol{V}(\boldsymbol{t}_{is})$*) of the vector estimator* $\mathbf{t}_{is}$*, while* $v(t_{i,js})$ *is variance of* j-th *component of* $\mathbf{t}_{is}$*. The above inequalities become sharp, when* $\boldsymbol{V}(\boldsymbol{t}_{1s}) - \boldsymbol{V}(\boldsymbol{t}_{2s})$ *is positive definite.*

The accuracy of estimator $\tilde{\boldsymbol{y}}_S$ is compared with the accuracy of the following well-known estimator of the vector of totals from an ordinary simple random sample drawn without replacement from a whole population:

$$\boldsymbol{y}_S = \frac{N}{n}\sum_{k\in S}\boldsymbol{y}_k, \qquad \boldsymbol{V}(\boldsymbol{y}_S) = \frac{N(N-n)}{n}\boldsymbol{C} \tag{7}$$

where S is drawn without replacement according to sampling design: $P_0(s) = \binom{N}{n}^{-1}$, $s \in \boldsymbol{S}$ and $\boldsymbol{S}$ is the sampling space generated for U. Under the assumption that $n = g\bar{N}$, we have:

$$\boldsymbol{V}(\tilde{\boldsymbol{y}}_s) - \boldsymbol{V}(\boldsymbol{y}_S) = \Big(\boldsymbol{I}_m + \frac{N-G}{G-1}\boldsymbol{\Delta}\Big)\big(\boldsymbol{C} + \bar{N}\boldsymbol{C}\big) = \frac{G(G-g)}{g}\big(\boldsymbol{C}_{\mathcal{U}} - \bar{N}\boldsymbol{C}\big). \tag{8}$$

According to Theorem 1, the estimator $\tilde{\boldsymbol{y}}_s$ is not worse than $\boldsymbol{y}_s$, when $\boldsymbol{C}_{\mathcal{U}} - \bar{N}\boldsymbol{C}$ is non-positive definite.

Particularly, when $N_h = M$ for all $h = 1, ..., G$, expressions (4), (7) and (8) let us write:

$$\begin{aligned} \boldsymbol{V}(\tilde{\boldsymbol{y}}_s) - \boldsymbol{V}(\boldsymbol{y}_S) = \frac{N(N-n)}{N}\frac{N-G}{G-1}\boldsymbol{C}\boldsymbol{\Delta} = \\ = \frac{N(N-n)}{N}\frac{N-G}{G-1}(\boldsymbol{C} - \boldsymbol{C}_*). \end{aligned} \tag{9}$$

If $N_h = M$ for all $h = 1, ..., G$, the estimator $\tilde{\boldsymbol{y}}_s$ is not worse than $\boldsymbol{y}_s$, when $\boldsymbol{C}\boldsymbol{\Delta}$ is non-positive definite.

The following theorem is proved in the "Appendix"

Theorem 2 *Let the variance-covariance matrices* $\boldsymbol{V}(\boldsymbol{t}_{is})$, $i = 1, 2$ *be positive definite. If estimator* $\boldsymbol{t}_{1s}$ *is not worse than* $\boldsymbol{t}_{2s}$, *then* $\boldsymbol{V}(\boldsymbol{t}_{2s}) - \boldsymbol{V}(\boldsymbol{t}_{1s})$ *is non-negative definite and:*

$$\lambda\left(\boldsymbol{V}(\boldsymbol{t}_{2s})\boldsymbol{V}^{-1}(\boldsymbol{t}_{1s})\right) = \lambda\left(\boldsymbol{V}^{-1}(\boldsymbol{t}_{1s})\boldsymbol{V}(\boldsymbol{t}_{2s})\right) \geq 1,$$
$$\lambda\left(\boldsymbol{V}^{-1}(\boldsymbol{t}_{2s})\boldsymbol{V}(\boldsymbol{t}_{1s})\right) = \lambda\left(\boldsymbol{V}(\boldsymbol{t}_{1s})\boldsymbol{V}^{-1}(\boldsymbol{t}_{2s})\right) \leq 1$$
$$\lambda_1\left(\boldsymbol{V}(\boldsymbol{t}_{2s})\boldsymbol{V}^{-1}(\boldsymbol{t}_{1s})\right) \leq \frac{\boldsymbol{\alpha}\boldsymbol{V}(\boldsymbol{t}_{2s})\boldsymbol{\alpha}^T}{\boldsymbol{\alpha}\boldsymbol{V}(\boldsymbol{t}_{1s})\boldsymbol{\alpha}^T} = \frac{\boldsymbol{V}(\boldsymbol{t}_{2s}\boldsymbol{\alpha}^T)}{\boldsymbol{V}(\boldsymbol{t}_{1s}\boldsymbol{\alpha}^T)} \leq \lambda\left(\boldsymbol{V}^{-1}(\boldsymbol{t}_{1s})\boldsymbol{V}(\boldsymbol{t}_{2s})\right).$$

for all $\boldsymbol{\alpha} \neq \boldsymbol{0}$ *where* $\lambda_1(...)$ *is the minimal eigenvalue of a matrix. The above inequalities become sharp, when* $\boldsymbol{V}(\boldsymbol{t}_{1s}) - \boldsymbol{V}(\boldsymbol{t}_{2s})$ *is positive definite.*

When the clusters are of the same size, Theorem 1 and expressions (9) let us conclude that when matrix $\boldsymbol{\Delta}$ is non-positive (non-negative) definite, then estimator $\tilde{\boldsymbol{y}}_s$ is not worse (not better) than $\boldsymbol{y}_s$.

Rao and Scott [8, pp. 223], define the generalized relative efficiency coefficient as follows:

$$deff(\boldsymbol{t}_S) = \lambda\left(\boldsymbol{V}(\boldsymbol{y}_S)^{-1}\boldsymbol{V}(\boldsymbol{t}_S)\right). \tag{10}$$

where $\boldsymbol{V}(\boldsymbol{y}_S)$ is non-singular. When $n = g\bar{N}$, expressions (2), (3) and (10) lead to the following:

$$deff(\tilde{\boldsymbol{y}}_S) = \frac{G(G-g)n\bar{N}}{N(N-n)g}\lambda\left(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}}\right) = 1 + \lambda\left(\frac{N-G}{G-1}\boldsymbol{\Delta} + \frac{1}{\bar{N}}\boldsymbol{C}^{-1}\boldsymbol{A}\right). \tag{11}$$

Hence, $deff(\tilde{\boldsymbol{y}}_S) = minim$ when the population is partitioned into set $\mathcal{U}$ of clusters in such a way that $\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}}) = minim$.

In particular, expressions (3) and (4) show that when $N_h = M$ for all $h = 1, ..., H$, then $\boldsymbol{A} = \boldsymbol{0}$. Inequality $-\frac{G-1}{N-G} \leq \lambda(\boldsymbol{\Delta}) \leq 1$ leads to the following:

$$0 \leq deff(\tilde{\boldsymbol{y}}_S) = 1 + \frac{N-G}{G-1}\lambda(\boldsymbol{\Delta}) \leq \frac{N-1}{G-1}. \tag{12}$$

When $\lambda(\boldsymbol{\Delta}) \leq 0$, then $\tilde{\boldsymbol{y}}_S$ is more efficient than $\boldsymbol{y}_S$. Hence, we should partition the population into clusters of the same size in such a way that coefficient $\lambda(\boldsymbol{\Delta})$ takes the minimal negative value.

2.4 Clustering Algorithms

We can expect that variables observed in a finite and fixed population in a past occasion are highly correlated with the appropriate variables observed in a current occasion or in future occasions. Therefore, census data could be used to construct reasonable sampling design for future occasion.

The above considerations lead to the conclusion that the population has to be clustered in such a way that the maximal eigenvalue of $\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}}$ takes the minimal value. Additionally, when we assume that the population has to be partitioned into clusters of the same size, then minimization of $\lambda(\boldsymbol{\Delta})$ is the criterion for population clustering. The following clustering algorithms will be considered:

Systematic algorithm 1:

Let us assume that $\boldsymbol{y}_k > \boldsymbol{0}$ for all $k = 1, ..., N$. Next, we evaluate squared distances $d_k = \boldsymbol{y}_k \boldsymbol{y}_k^T$ of $\boldsymbol{y}_k$ from the zero vector $\boldsymbol{0}$ for all $k \in U$. Let us assume that $d_k \leq d_{k+1}$ for $k = 1, ..., N-1$. The h-th cluster is identified by the unit labels $k \in U_h$ that $k = (i-1)G + h$, for $i = 1, ..., M$ and $h = 1, ..., G$. This leads to inequalities: $d_{U_h} \leq d_{U_{h+1}}$ for $h = 1, ..., G-1$ where $d_{U_h} = \sum_{k \in U_h} d_k$. The result of this clustering algorithm will be denoted by $\mathcal{U}_1$. In some sense, this result is the well-known systematic simple sample space.

Systematic algorithm 2: Let $d_k = (\boldsymbol{y}_k - \bar{\boldsymbol{y}})(\boldsymbol{y}_k - \bar{\boldsymbol{y}})^T$ be the squared distance of $\boldsymbol{y}_k$ from vector $\bar{\boldsymbol{y}}$ for all $k \in U$. Let us assume that $d_k \leq d_{k+1}$ for $k = 1, ..., N-1$. Let $M = 2$ and $N = MG$. In this case, $U_h = \{h; N - h + 1\}$ for $h = 1, ..., G$. In general, when M is even and $N = MG$, then $U_h = \{(h-1)\frac{M}{2} + i; N - (h-1)\frac{M}{2} - i + 1\}$ for $h = 1, ..., G$ and $i = 1, ...M/2$. The result of this clustering algorithm will be denoted by $\mathcal{U}_2$.

Permutation algorithm 3: Let $\mathcal{U}^{(0)} = \{U_1^{(0)}, ..., U_G^{(0)}\}$ be any start partition of a population partitioned into clusters of the same sizes, $M \geq m$. In the t-th (t=0,1,...) iteration partition $\mathcal{U}^{(t)} = \{U_1^{(t)}, ..., U_G^{(t)}\}$ is generated through permutating population elements. For an assumed $t = T$, $\mathcal{U}^{(T)}$ is treated as optimal when

$$\lambda_*(\mathcal{U}^{(T)}) = min_{\{t=1,..,T\}}(\lambda(\Delta(\mathcal{U}^{(t)}))). \tag{13}$$

Iteration algorithm 4: Let $\mathcal{U}^{(0)} = \{U_1^{(0)}, ..., U_G^{(0)}\}$ be any start partition of the population partitioned into clusters which are not necessary of the same size. Let $\mathcal{U}^{(t)} = \{U_1^{(t)}, ..., U_G^{(t)}\}$ be the partition of the population obtained as result of the t-th iteration and let $\lambda_t = \lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}})$ be the maximal eigenvalue of the variance-covariance matrix of cluster totals. Moreover, let $f : U \to \mathcal{U}^{(t)}$, $f_t(k) = h$, if and only if $k \in U_h^{(t)}$.

In iteration $t + 1$, we randomly choose number k_* of data observation from the sequences 1, ..., N. Next, element k_* is moved from the cluster $h_\# = f_t(k_*)$ to cluster h_* where h_* is randomly drawn from set $\{h : h = 1, ..., G; h \neq h_\#\}$. This leads to the new partition $\mathcal{U}^{(t+1)}$. Finally, we count $\lambda_{t+1} = \lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t+1)}})$. If $\lambda_{t+1} < \lambda_t$, then $\mathcal{U}^{(t+1)}$ is the current partition and we start iteration $t+2$ of the algorithm. If $\lambda_{t+1} \geq \lambda_t$, then we start stage $t+2$ of the algorithm from partition $\mathcal{U}^{(t)}$. The algorithm of the partition is stopped when the number of iterations reaches the assumed level T. This algorithm leads to the minimization of $deff(\tilde{\boldsymbol{y}}_S)$. The population clustered according to this algorithm will be denoted by $\mathcal{U}_4$.

Iteration algorithm 5: The clustering procedure described below is similar to the above one and also leads to minimization of $\boldsymbol{V}(\tilde{\boldsymbol{y}}_S)$.

Let $\mathcal{U}^{(t)} = \{U_1^{(t)}, ..., U_G^{(t)}\}$ be the partition of the population obtained as result of the t-th iteration where $t = (l-1)N + k$, $k = 1, ..., N$, $l = 1, 2, ...$ and let $\lambda_t = \lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}})$ be the maximal eigenvalue evaluated on the basis of $\mathcal{U}^{(t)} = \{U_1^{(t)}, ..., U_G^{(t)}\}$. Let $f : U \to \mathcal{U}^{(t)}$, $f_t(l) = h$, if and only if $l \in U_h^{(k,t)}$.

In stage $t+1$, the population element $k \in U_h^{(t)}$, where $h = f_t(k)$, is moved to clusters $U_z^{(t)}$, $z \neq h$, $z = 1, ..., G$ and calculated using the following

$$(k, \underline{z}) = arg\big(min_{\{z=1,...,G,z\neq f_t(k)\}}\big(\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}}(k,z))\big)\big) \tag{14}$$

where $\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}}(k,z))$ is evaluated for partition $\mathcal{U}^{(t)}$ in which clusters $U_z^{(t)}$, $U_h^{(t)}$ are replaced by $\{U_z^{(t)} \cup \{k\}\}$ and $\{U_h^{(t)} - \{k\}\}$, respectively, and $h = f_t(k)$. If $\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}}(\underline{z})) < \lambda_t$, then $\lambda_{t+1} = \lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t+1)}})$ and $\mathcal{U}^{(t+1)}$ is equal to $\mathcal{U}^{(t)}$ where clusters $U_{\underline{z}}^{(t)}$ and $U_{\underline{h}}^{(\overline{t})}$ are replaced by $U_{\underline{z}}^{(t+1)} = \{U_{\underline{z}}^{(t)} \cup \{k\}\}$ and $U_h^{(t+1)} = \{U_h^{(t)} - \{k\}\}$, respectively. If $\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}}(\underline{z})) \geq \lambda_t$, then $\mathcal{U}^{(t+1)} = \mathcal{U}^{(t)}$ and $\lambda_{t+1} = \lambda_t$.

The iteration clustering process is stopped when $\lambda_{t+N} = \lambda_t$ or the number of iterations reaches the assumed level T. This algorithm will be denoted by $\mathcal{U}_5$.

Iteration algorithm 6: We keep the notation introduced earlier. In iteration $t+1$, the population element $k \in U_h^{(t)}$, where $h = f_t(k)$, is moved to clusters $U_z^{(t)}$, $z \neq h$, $z = 1, ..., G$. Next, we calculate the following

$$(\underline{k}, \underline{z}) = arg\big(min_{\{k\in U\}}min_{\{z\neq f_t(k),z=1,...,G\}}\big(\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}}(k,z))\big)\big) \tag{15}$$

where $\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}}(k,z))$ is evaluated for partition $\mathcal{U}^{(t)}$ in which clusters $U_z^{(t)}$ and $U_h^{(t)}$ are replaced by $\{U_z^{(t)} \cup \{k\}\}$ and $\{U_h^{(t)} - \{k\}\}$, respectively, and $h = f_t(k)$. If $\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}}(\underline{k}, \underline{z})) < \lambda_t$, then $\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t+1)}}) = \lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}}(\underline{k}, \underline{z}))$ and $\mathcal{U}^{(t+1)}$ is equal to $\mathcal{U}^{(t)}$ where clusters $U_z^{(t)}$ and $U_h^{(t)}$ are replaced by $U_{\underline{z}}^{(t+1)} = \{U_{\underline{z}}^{(t)} \cup \{k\}\}$ and $U_h^{(t+1)} = \{U_h^{(t)} - \{k\}\}$, respectively. The iteration clustering process is stopped when $\lambda(\boldsymbol{C}^{-1}\boldsymbol{C}_{\mathcal{U}^{(t)}}(\underline{k}, \underline{z})) \geq \lambda_t$. The population clustered according to this algorithm will be denoted by $\mathcal{U}_6$.

3 Accuracy Analysis

Data about Swedish municipalities published in the monograph by [9] will be considered. Variables y_1 and y_2 are the real estate values (according to the 1984 assessment, in millions of kronor) and number of the municipal employees (in millions of kronor), respectively. Their population correlation coefficient is $\rho_{y_1,y_2} = 0.9924$. The population size (without outliers) is $N = 280$. Moreover, $\bar{y}_{1,U} = 51945.99$, $\bar{y}_{2,U} = 378859$, $v_{y_1} = 35954.39$, $v_{y_2} = 2008981$. The partitions obtained as results of the above clustering algorithms will be denoted by $\mathcal{U}_j$, $j = 1, ..., 6$. We will consider the sample sizes $g = 2, 4, 8, 12, 14, 24$ and cluster sizes $M = 2, 4, 8, 14$. The relative efficiency coefficient is evaluated according to expression (10) for estimation strategy $(\tilde{\boldsymbol{y}}_S)$.

Analysis of Table 1 leads to the following conclusions. Only under clustering algorithms $\mathcal{U}_1$ and $\mathcal{U}_2$, the accuracy of estimator $\boldsymbol{y}_S$ is approximately not less than the accuracy of estimator $\tilde{\boldsymbol{y}}_S$ for all considered combinations (M, g).

Partition $\mathcal{U}_4$ leads to the most efficient estimation based on $\tilde{\boldsymbol{y}}_S$. When we also assume that the population is split into sub-populations of the same sizes, estimator $\tilde{\boldsymbol{y}}_S$ based on the sample drawn from the population clustered according to algorithm $\mathcal{U}_3$ is the most efficient.

For algorithms $\mathcal{U}_1$ and $\mathcal{U}_2$, the estimation efficiency based on $\tilde{\boldsymbol{y}}_S$ decreases (or equivalently $deff(\tilde{\boldsymbol{y}}_S)$ increases) when the number of clusters g decreases under the fixed sample size n. For algorithms $\mathcal{U}_3$ - $\mathcal{U}_6$, the situation is reversed. Under the fixed sample size n, the estimation efficiency based on $\tilde{\boldsymbol{y}}_S$ increases when number of clusters g decreases. For instance, under partition $\mathcal{U}_4$ when $(M, g) = (2, 14)$ and $(M, g) = (14, 2)$, the accuracy of $\tilde{\boldsymbol{y}}_S$ is almost two times and fifty times better than the accuracy of $\boldsymbol{y}_S$, respectively.

4 Conclusions

In this paper, we have shown that it is possible to significantly increase the accuracy of estimating population totals using vector estimator from a simple cluster sample drawn without replacement by considering specific partition of a population into clusters. In the analysed empirical example, algorithm 5 and 6 lead to the optimal partition of the population. These algorithms should work quickly when a population size is large. The results could be useful for panel or census survey sampling repeated on more than one occasion. The results of paper could be applied to partitioning a population into clusters based on census data. This

Table 1 Relative efficiency for the population partitioned into clusters. *Source*: Own calculations.

n	(M,g)	$\mathcal{U}_1$	$\mathcal{U}_2$	$\mathcal{U}_3$	$\mathcal{U}_4$	$\mathcal{U}_5$	$\mathcal{U}_6$
1	2	3	5	6	7	8	9
16	(2,8)	0.992	1.108	0.824	0.561	0.639	0.771
16	(4,4)	1.095	2.147	0.749	0.182	0.444	0.578
16	(8,2)	1.171	4.245	0.621	0.049	0.182	0.442
28	(2,14)	0.992	1.108	0.824	0.561	0.639	0.771
28	(4,7)	1.095	2.147	0.749	0.182	0.444	0.578
28	(14,2)	1.309	7.353	0.504	0.019	0.041	0.196
48	(2,24)	0.992	1.108	0.824	0.561	0.639	0.771
48	(4,12)	1.095	2.147	0.749	0.182	0.444	0.578
48	(8,6)	1.171	4.245	0.621	0.049	0.182	0.442

could improve accuracy of estimation of population total vector. Moreover, the results could be useful in some aspects of big-data analysis.

This paper could be treated as a contribution to comparison of vector estimators. Several properties of the generalized relative efficiency coefficient are considered in Theorem 2. The generalized coefficient of intra-cluster data spread homogeneity was defined, its properties were considered and its values were interpreted. The generalized *deff* coefficient was also written as the function of matrix of coefficients of intra-cluster homogeneity. The proposed procedures could be developed in several ways. Other clustering algorithms could be considered. In particular, the clustering procedures based on multivariate variables that are proposed in this paper could be reduced to one-dimensional cases. For instance, these variables could be replaced with their principal component. In this case, the several clustering procedures based on one-dimensional variables that have been proposed by [14] could be adopted in our considerations.

In addition, many of the clustering algorithms available in the statistical literature (see, e.g. [1, 6]) divide the population into homogeneous clusters. Typically, these procedures can be modified to algorithms that ensure the maximum spread of multivariate observations within the cluster. This seems to the well-known nearest (farthest) neighbour criteria. Properties of some sampling designs used in spatial statistics could inspirate for the construction of clustering algorithms. For example, the criteria considered by Thompson and Seber [10] or [13] can be adapted to divide the spatial population into clusters composed of non-neighbours.

Appendix

Decomposition of Matrix $\boldsymbol{C}_{\mathcal{U}}$

[9] decomposed the diagonal element of $\boldsymbol{V}(\tilde{\boldsymbol{y}}_S)$ defined by expression (2) as the function of matrix $\boldsymbol{C}_{\mathcal{U}}$. Their result can be generalized as follows. Using their result, we transform elements of $\boldsymbol{C}_{\mathcal{U}} = [c_{\mathcal{U},i,j}]$ in the following way (see [12], pp. 139–150):

$$
\begin{aligned}
(G-1)c_{\mathcal{U},i,j} &= \sum_{h=1}^{G}(y_{U_h,i}-\bar{y}_{\mathcal{U},i})(y_{U_h,j}-\bar{y}_{\mathcal{U},j}) \\
&= \sum_{h=1}^{G}(N_h\bar{y}_{U_h,i}-\bar{N}\bar{y}_i)(N_h\bar{y}_{U_h,j}-\bar{N}\bar{y}_j) \\
&= \sum_{h=1}^{G}((N_h-\bar{N})\bar{y}_{U_h,i}+\bar{N}(\bar{y}_{U_h,i}-\bar{y}_i))((N_h-\bar{N})\bar{y}_{U_h,j}+\bar{N}(\bar{y}_{U_h,j}-\bar{y}_j)) \\
&= \sum_{h=1}^{G}(N_h-\bar{N})^2\bar{y}_{U_h,i}\bar{y}_{U_h,j}+\bar{N}^2\sum_{h=1}^{G}(\bar{y}_{U_h,i}-\bar{y}_i)(\bar{y}_{U_h,j}-\bar{y}_j) \\
&+ \bar{N}\sum_{h=1}^{G}(N_h-\bar{N})(\bar{y}_{U_h,j}-\bar{y}_j)\bar{y}_{U_h,i}+\bar{N}\sum_{h=1}^{G}(N_h-\bar{N})(\bar{y}_{U_h,i}-\bar{y}_i)\bar{y}_{U_h,j} \\
&= \sum_{h=1}^{G}(N_h-\bar{N})^2\bar{y}_{U_h,i}\bar{y}_{U_h,j}+\bar{N}^2\sum_{h=1}^{G}(\bar{y}_{U_h,i}-\bar{y}_i)(\bar{y}_{U_h,j}-\bar{y}_j)+ \\
&+ 2\bar{N}\sum_{h=1}^{G}(N_h-\bar{N})(\bar{y}_{U_h,i}-\bar{y}_i)(\bar{y}_{U_h,j}-\bar{y}_j)+\bar{N}\bar{y}_i\sum_{h=1}^{G}(N_h-\bar{N})(\bar{y}_{U_h,j}-\bar{y}_j)+ \\
&+ \bar{N}\bar{y}_j\sum_{h=1}^{G}(N_h-\bar{N})(\bar{y}_{U_h,i}-\bar{y}_i) = \sum_{h=1}^{G}(N_h-\bar{N})^2\bar{y}_{U_h,i}\bar{y}_{U_h,j}+ \\
&- \bar{N}^2\sum_{h=1}^{G}(\bar{y}_{U_h,i}-\bar{y}_i)(\bar{y}_{U_h,j}-\bar{y}_j)+2\bar{N}\sum_{h=1}^{G}(\bar{y}_{U_h,i}-\bar{y}_i)(\bar{y}_{U_h,j}-\bar{y}_j)N_h \\
&+ \bar{N}\bar{y}_i\sum_{h=1}^{G}(N_h-\bar{N})(\bar{y}_{U_h,j}-\bar{y}_j)+\bar{N}\bar{y}_j\sum_{h=1}^{G}(N_h-\bar{N})(\bar{y}_{U_h,i}-\bar{y}_i) =
\end{aligned}
$$

$$= \sum_{h=1}^{G}(N_h - \bar{N})^2 \bar{y}_{U_h,i}\bar{y}_{U_h,j} + \bar{N}\sum_{h=1}^{G}(N_h - \bar{N})(\bar{y}_{U_h,i} - \bar{y}_i)(\bar{y}_{U_h,j} - \bar{y}_j)$$

$$+ \bar{N}\sum_{h=1}^{G}(\bar{y}_{U_h,i} - \bar{y}_i)(\bar{y}_{U_h,j} - \bar{y}_j)N_h + \bar{N}\bar{y}_i\sum_{h=1}^{G}(N_h - \bar{N})(\bar{y}_{U_h,j} - \bar{y}_j)$$

$$+ \bar{N}\bar{y}_j\sum_{h=1}^{G}(N_h - \bar{N})(\bar{y}_{U_h,i} - \bar{y}_i) = \sum_{h=1}^{G}(N_h - \bar{N})^2 \bar{y}_{U_h,i}\bar{y}_{U_h,j}$$

$$+ \bar{N}\sum_{h=1}^{G}(N_h - \bar{N})(\bar{y}_{U_h,i} - \bar{y}_i)(\bar{y}_{U_h,j} - \bar{y}_j + \bar{y}_j)$$

$$+ \bar{N}\sum_{h=1}^{G}(\bar{y}_{U_h,i} - \bar{y}_i)(\bar{y}_{U_h,j} - \bar{y}_j)N_h + \bar{N}\bar{y}_i\sum_{h=1}^{G}(N_h - \bar{N})(\bar{y}_{U_h,j} - \bar{y}_j)$$

$$= \sum_{h=1}^{G}(N_h - \bar{N})^2 \bar{y}_{U_h,i}\bar{y}_{U_h,j} + \bar{N}\sum_{h=1}^{G}(\bar{y}_{U_h,i} - \bar{y}_i)(\bar{y}_{U_h,j} - \bar{y}_j)N_h +$$

$$+ \bar{N}\sum_{h=1}^{G}(N_h - \bar{N})\bar{y}_{U_h,i}\bar{y}_{U_h,j}.$$

Finally, we have:

$$(G-1)c_{\mathcal{U},i,j} = \sum_{h=1}^{G}(N_h - \bar{N})N_h\bar{y}_{U_h,i}\bar{y}_{U_h,j} + \bar{N}\sum_{h=1}^{G}(\bar{y}_{U_h,i} - \bar{y}_i)(\bar{y}_{U_h,j} - \bar{y}_j)N_h. \quad (16)$$

The decomposition of the ordinary covariance is:

$$(N-1)c_{i,j} = \sum_{h=1}^{G}\sum_{k\in U_h}(y_{k,i} - \bar{y}_i)(y_{k,j} - \bar{y}_j) =$$

$$= \sum_{h=1}^{G}\sum_{k\in U_h}(y_{k,i} - \bar{y}_{U_h,i})(y_{k,j} - \bar{y}_{U_h,j}) + \sum_{h=1}^{G}(\bar{y}_{U_h,i} - \bar{y}_i)(\bar{y}_{U_h,j} - \bar{y}_j)N_h =$$

$$= (N-G)c_{*i,j} + \sum_{h=1}^{G}(\bar{y}_{U_h,i} - \bar{y}_i)(\bar{y}_{U_h,j} - \bar{y}_j)N_h.$$

This and expression (16) lead to the following:

$$(G-1)c_{i,j,\mathcal{U}} = \sum_{h=1}^{G}(N_h - \bar{N})N_h\bar{y}_{U_h,i}\bar{y}_{U_h,j} + \bar{N}(N-1)c_{i,j} - \bar{N}(N-G)c_{*i,j}. \quad (17)$$

Moreover:

$$\begin{aligned}
&\frac{1}{G-1}\sum_{h=1}^{G}(N_h-\bar{N})N_h\bar{y}_{U_h,i}\bar{y}_{U_h,j} = \frac{1}{G-1}\sum_{h=1}^{G}(N_h-\bar{N})y_{U_h,i}\bar{y}_{U_h,j}\\
&= \frac{1}{G-1}\sum_{h=1}^{G}(N_h-\bar{N})((y_{U_h,i}-\bar{y}_{\mathcal{U},i})+\bar{y}_{\mathcal{U},i})((\bar{y}_{U_h,j}-\bar{y}_j)+\bar{y}_j) =\\
&= \frac{1}{G-1}\sum_{h=1}^{G}(N_h-\bar{N})(y_{U_h,i}-\bar{y}_{\mathcal{U},i})(\bar{y}_{U_h,j}-\bar{y}_j)\\
&+ \frac{\bar{y}_j}{G-1}\sum_{h=1}^{G}(N_h-\bar{N})(y_{U_h,i}-\bar{y}_{\mathcal{U},i}) + \frac{\bar{y}_{\mathcal{U},i}}{G-1}\sum_{h=1}^{G}(N_h-\bar{N})(\bar{y}_{U_h,j}-\bar{y}_j).
\end{aligned} \tag{18}$$

This leads to the decomposition of matrix $\boldsymbol{A}$ shown in Sect. 2.2.

Derivation of Expression (3)

Expression (17) leads to the following:

$$\begin{aligned}
\boldsymbol{C}_{\mathcal{U}} &= \boldsymbol{A} + \frac{\bar{N}(N-1)}{G-1}\boldsymbol{C} - \frac{\bar{N}(N-G)}{G-1}\boldsymbol{C}_*\\
&= \boldsymbol{A} + \bar{N}\boldsymbol{C}\left(\boldsymbol{I}_m + \boldsymbol{I}_m\frac{N-G}{G-1} - \frac{N-G}{G-1}\boldsymbol{C}^{-1}\boldsymbol{C}_*\right)\\
&= \boldsymbol{A} + \bar{N}\boldsymbol{C}\left(\boldsymbol{I}_m + \frac{N-G}{G-1}(\boldsymbol{I}_m - \boldsymbol{C}^{-1}\boldsymbol{C}_*)\right).
\end{aligned} \tag{19}$$

This result, expressions (2) and (4) let us evaluate expression (3).

Proof of Theorem 2

Let $\boldsymbol{F}$ and $\boldsymbol{L}$ be $m \times m$ symmetric matrices. $\boldsymbol{F}$ is non-negative definite, while $\boldsymbol{L}$ is positive definite. There is a non-singular matrix $\boldsymbol{G}$ such that $\boldsymbol{L} = \boldsymbol{G}^T\boldsymbol{G}$ (see, e.g. [3], p. 218-219. Moreover, [3], p. 563) shows that:

$$|\boldsymbol{F} - \lambda\boldsymbol{L}| = |\boldsymbol{G}^2||\boldsymbol{M} - \lambda\boldsymbol{I}| = |\boldsymbol{L}||\boldsymbol{F}\boldsymbol{L}^{-1} - \lambda\boldsymbol{I}| = |\boldsymbol{L}||\boldsymbol{L}^{-1}\boldsymbol{F}\boldsymbol{L}^{-1} - \lambda\boldsymbol{I}| \tag{20}$$

where

$$\boldsymbol{M} = (\boldsymbol{G}^T)^{-1}\boldsymbol{F}\boldsymbol{G}^{-1}. \tag{21}$$

Therefore, matrices $\boldsymbol{M}$, $\boldsymbol{F}\boldsymbol{L}^{-1}$ and $\boldsymbol{L}^{-1}\boldsymbol{F}$ have the same eigenvalues as roots of equation $|\boldsymbol{F} - \lambda\boldsymbol{L}| = 0$. If $\boldsymbol{F}$ is non-negative (positive) definite, then $\boldsymbol{M}$ is non-negative (positive) definite.

Let $\boldsymbol{V}(\boldsymbol{t}_{1s}) = L$ and $\boldsymbol{V}(\boldsymbol{t}_{2s}) = F$. This and the above properties let us immediately prove the first two equalities of Theorem 2.

Expression (20) let us rewrite $|\boldsymbol{F}\boldsymbol{L}^{-1} - \lambda\boldsymbol{I}| = 0$ as follows:

$$|\boldsymbol{F} - \lambda\boldsymbol{L}| = 0, \qquad \boldsymbol{F} - \boldsymbol{L} - (\lambda - 1)\boldsymbol{L}| = 0, \qquad |(\boldsymbol{G}^T)^{-1}(\boldsymbol{F} - \boldsymbol{L})\boldsymbol{G}^{-1} - \kappa\boldsymbol{I}| = 0$$

where $\kappa = \lambda - 1$. If $\boldsymbol{F} - \boldsymbol{L}$ is non-negative definite, then matrix $(\boldsymbol{G}^T)^{-1}(\boldsymbol{F} - \boldsymbol{L})\boldsymbol{G}^{-1}$ is non-negative definite (see, e.g. [3], pp. 213) and $\kappa = \lambda - 1 \geq 0$. This leads to $\lambda \geq 1$. Hence, the first inequality of Theorem 2 is proved.

If matrix $\boldsymbol{F} - \boldsymbol{L}$ is non-positive definite, then $\boldsymbol{L} - \boldsymbol{F}$ is non-negative definite and $|\boldsymbol{F} - \lambda\boldsymbol{L}| = 0$ is equivalent to $|\boldsymbol{F} - \boldsymbol{L} - (\lambda - 1)\boldsymbol{L}| = 0, |\boldsymbol{L} - \boldsymbol{F} - (1 - \lambda)\boldsymbol{L}| = 0$ and

$$|(\boldsymbol{G}^T)^{-1}(\boldsymbol{L} - \boldsymbol{F})\boldsymbol{G}^{-1} - \kappa\boldsymbol{I}| = 0, \qquad \kappa = 1 - \lambda.$$

Therefore, matrices $\boldsymbol{L} - \boldsymbol{F}$ and $(\boldsymbol{G}^T)^{-1}(\boldsymbol{L} - \boldsymbol{F})\boldsymbol{G}^{-1}$ are non-negative definite and $\kappa = 1 - \lambda \geq 0, \lambda \leq 1$. Hence, the second inequality of Theorem 2 is proved.

Let us consider the following ratio of the quadratic forms:

$$\frac{\boldsymbol{\alpha F \alpha}^T}{\boldsymbol{\alpha L \alpha}^T} = \frac{\boldsymbol{\beta M \beta}^T}{\boldsymbol{\beta \beta}^T}$$

where $\boldsymbol{\alpha} = \boldsymbol{\beta}(\boldsymbol{G}^T)^{-1}$ and $\boldsymbol{L} = \boldsymbol{G}^T\boldsymbol{G}$. This and (16) and (17) let us write the following:

$$\lambda_1 \leq \frac{\boldsymbol{\beta M \beta}^T}{\boldsymbol{\beta \beta}^T} = \frac{\boldsymbol{\beta F L}^{-1}\boldsymbol{\beta}^T}{\boldsymbol{\beta \beta}^T} \leq \lambda_m,$$

where λ_1 and λ_m are minimal and maximal eigenvalue of matrix $\boldsymbol{FL}^{-1} = \boldsymbol{V}(\boldsymbol{t}_{2s})\boldsymbol{V}^{-1}(\boldsymbol{t}_{1s})$, respectively. This directly leads to last expression of Theorem 2.

Proof of Expression (6) About Eigenvalues of the Homogeneity Matrix

The eigenvalues of the homogeneity matrix, given by (4), are roots of the following equation:

$$|\boldsymbol{\Delta} - \delta\boldsymbol{I}| = 0, \qquad |\boldsymbol{C} - \boldsymbol{C}_* - \delta\boldsymbol{C}| = 0, \qquad |\boldsymbol{C}_* - (1 - \delta)\boldsymbol{C}| = 0. \tag{22}$$

According to the properties of matrix (see, e.g. [3], pp. 219) $\boldsymbol{C} = \boldsymbol{G}^T\boldsymbol{G}$, where $\boldsymbol{G}$ is symmetric because matrix $\boldsymbol{C}$ is symmetric and positive definite. Therefore:

$$|\boldsymbol{M} - (1 - \delta)\boldsymbol{I}| = 0 \tag{23}$$

where $\boldsymbol{M} = (\boldsymbol{G}^{-1})^T\boldsymbol{C}_*\boldsymbol{G}^{-1}$. Matrix $\boldsymbol{M}$ is non-negative definite (see, e.g. [3], pp. 213). This let us write $1 - \delta \geq 1$ and $\delta \leq 1$.

The well-known decomposition of $\boldsymbol{C}$ is:

$$(N - 1)\boldsymbol{C} = G\boldsymbol{C}_\# + (N - G)\boldsymbol{C}_*$$

Matrix $(N - 1)\boldsymbol{C} - (N - G)\boldsymbol{C}_*$ is non-negative definite because $\boldsymbol{C}$ is positive definite and $\boldsymbol{C}_*$, $\boldsymbol{C}_\#$ are non-negative definite. Moreover, $\frac{N-1}{N-G}\boldsymbol{C} - \boldsymbol{C}_* = \boldsymbol{C} + \frac{G-1}{N-G}\boldsymbol{C} - \boldsymbol{C}_*$ is non-negative definite. This and expression (22) let us write that equation $|\boldsymbol{\Delta} - \delta\boldsymbol{I}| = 0$ is equivalent to the following

$$\left|\boldsymbol{C} + \frac{G-1}{N-G}\boldsymbol{C} - \boldsymbol{C}_* - \left(\delta + \frac{G-1}{N-G}\right)\boldsymbol{C}\right| = 0$$

Similarly to expression (23), there is symmetric matrix $\boldsymbol{K}$ making the above equation equivalent to the following

$$\left|\boldsymbol{E} - \left(\delta + \frac{G-1}{N-G}\right)\boldsymbol{I}\right| = 0$$

where

$$\boldsymbol{E} = (\boldsymbol{K}^{-1})^T(\boldsymbol{C} + \frac{G-1}{N-G}\boldsymbol{C} - \boldsymbol{C}_*)\boldsymbol{K}^{-1}$$

is non-negative definite. This leads to inequality $\delta \geq -\frac{G-1}{N-G}$. Hence, inequality $-\frac{G-1}{N-G} \leq \delta \leq 1$ where δ is the eigenvalue of $\boldsymbol{\Delta}$ is proved.

Acknowledgements This paper is a result of a grant supported by the *National Science Centre, Poland*, No. 2016/21/B/HS4/00666

References

1. Bock HH (2002) Clustering methods: from classical models to new approaches. Stat Trans 5:725–728
2. Borovkov AA (1984) Mathematical statistics. Estimation. Testing hypotheses. Nauka Moskva (in Russian)
3. Hardville DA (1997) Matrix algebra from a statistician's perspective. Springer, New York
4. Jensen DR (1991) Vector efficiency in multiparameter estimation. Linear Algebra Appl 151:143–155
5. Kish L (1995) Survey sampling. Wiley, New York
6. Mosler K (2012) Multivariate dispersion, central regions, and depth. Springer, New York
7. Rao CR (1973) Linear statistical inference and its applications. Wiley, New York
8. Rao JNK, Scott AJ (1981) The analysis of categorical data from complex sample surveys: chi-squared tests of goodness of fit and independence in two-way tables. J Am Stat Assoc 76(374):221–230
9. Särndal CE, Swensson B, Wretman J (1992) Model assisted survey sampling. Springer, Berlin
10. Thompson SK, Seber GAF (1996) Adaptive sampling. Wiley, New York
11. Wywiał JL (2002) On estimation of population average on the basis of cluster sample. In: Jajuga K, Sokołowski A, Bock HH (eds) Classification, clustering, and data analysis. Springer, Berlin, pp 271–277
12. Wywiał J (2003) Some contributions to multivariate methods in survey sampling. University of Economics in Katowice, Katowice. https://www.ue.katowice.pl/fileadmin/user_upload/wydaw

nictwo/Darmowe_E-Booki/Wywial_Some_Contributions_To_Multivariate_Methods_In_Survey_Sampling.pdf

13. Wywiał JL (2013) On space sampling. Acta Universitatis Lodziensis Folia Oeconomica 292:21–35
14. Wywiał JL, Sitek G (2020) On influence of clustering population on accuracy of population total estimation. J Stat Comput Simul 90(2):234–251. https://doi.org/10.1080/00949655.2019.1676426