ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 216, 2008

*Tomasz Jurkiewicz**

CORRELATION AMONG VARIABLES AND METHODS OF ESTABLISHING WEIGHTS OF SAMPLE UNITS. MONTE CARLO ANALYSIS OF THE MODIFIED SYNTHETIC ESTIMATOR

ABSTRACT. In many statistical surveys one faces the problem of insufficient number of sample observations to make reliable inference about a given population domain of interest (small area). One possible solution, which has been discussed in statistical publications consists in applying estimators, which will be able to combine sample information from the given domain with information about sample units representing other domains. Synthetic estimation technique is particularly efficient, if the distribution of the variable of interest is the same in the given domain and in the entire population. When this assumption is far from being met, one can obtain, as a consequence, large estimation errors.

Using modified synthetic estimator requires an application of a two-stage estimation procedure. The first stage consists in applying some distance measures in order to identify the degree of similarity between the sample units from the investigated domain and sample units representing other domains. In the second stage, those units, which turned out to be similar to units from the domain of interest, are used to provide sample information with specially constructed weights.

A method of establishing weights is one of the crucial factors in using MES estimator. Author presents results of Monte Carlo analysis of the efficiency of MES estimator using different weights.

Key words: small domain estimation, multivariate methods, distance measures.

I. INTRODUCTION

It is widely observed that the processes of economic and social developments result in an increasing demand for statistical information. Statistical surveys, and representative surveys in particular, have recently become one of the

* Ph.D., Department of Statistics, University of Gdańsk, t.jurkiewicz@zr.univ.gda.pl.

most popular ways of collecting data and information needed to make decisions in various areas of human activity. Because of organisational and financial constraints those studies, however, are not able to supply credible data for a more detailed division of the population into smaller domains of studies. An insufficient number of observations representing a particular domain may be an obstacle in applying certain statistical techniques and tools, or may lead to considerable errors of estimation (cf. Bracha (1996)). One possible way of solving this problem is an attempt to construct estimators, which could use some additional information i.e. information about other components of the sample, namely those coming from outside a particular part of the population. The other possibility is to use additional information from outside of the sample (prior information) to estimate parameters of a defined subpopulation.

The notion of "small domain" (small area) is defined as a domain of studies, for which information is essential for the data user, and cannot be obtained by using a direct estimation method because of insufficient sample size. Also, a small domain could be understood as a domain of studies, for which the information acquired with indirect methods is more reliable. From the methodological point of view it does not make any difference whether we consider a subpopulation of one territory or a subpopulation isolated according to any other method.

The essence of indirect estimation consists in "borrowing information" from other domains or other sources in order to improve the estimation efficiency in the domain of interest. In case of a representative study it is possible to use the following sources of additional data (see: Domański, Pruska (2001), Jurkiewicz (2001), Kordos (1999)): some other domains in the sample; information about the number of particular strata, and the number of domains in the studied population; information about additional variables in the sample; information about an additional variable in the studied population; other available prior data, e.g. data from studies carried out in other periods.

The main purpose of this paper is to evaluate an influence of methods of establishing weights of sampling units on efficiency of the modified synthetic estimator.

II. ESTIMATORS OF SMALL DOMAINS

The direct estimator of an unknown parameter ΘY_d in a small domain is the simple domain (SD) estimator, known as the expansion estimator. It uses entirely the data about randomly drawn components of a sample belonging to the small domain, that way is not a truly small domain estimator, but it is a datum for other estimators. The SD estimator is unbiased, but because of the small size

of the sample its variance is usually high. This estimator will have the following form for the mean value:

$$_{SD}\overline{x}_d = \frac{\sum_{i=1}^{n_d} x_i}{n_d}. \quad (1)$$

where: x_i stands for the variable values of units in the domain d and n_d is the size of the small domain d.

Synthetic estimation constitutes one of the first propositions of solving the principal problem of estimation for small domains, which stems from an insufficient sample size. In order to do this an assumption is made that the structure of the studied population in the small domain and outside the domain is uniform, which enables us to use the information from the whole sample to estimate the value for the domain. This assumption may be limited in some cases to the similarity of only certain parameters in the population and in the domain. For instance, the basis for construction of the common synthetic estimator is the assumption that the means of the studied characteristic in the population and in the domain do not essentially differ from each other. For the mean value of the estimator one can adopt the following statistics:

$$_{syn}\overline{x}_d = \frac{\sum_{i=1}^{n} x_i}{n}. \quad (2)$$

where n is the sample size.

While applying the synthetic estimation, it is important to pay careful attention to the problem of efficiency of the adopted model. The larger discrepancies between the assumptions laying at the base of this estimation technique and the reality, the more biased will be the estimators. It must be borne in mind that firstly, the bias may be of considerable size, and secondly, in no way it is taken into account in formulae for the mean square error and estimators of errors.

Modified Synthetic Estimator (MES)

The assumption about the compatibility of structures of the population and the domain remains usually unsatisfied, in particular in case of specific domains, what results in large estimation errors. A possible solution of this problem may be to strengthen the estimation process by modifying the estimator with information from components or domains similar to the studied one. The proposed procedure of estimation is carried out in two stages. The first step consists in estab-

lishing which components or domains are similar to the studied one. Weights for additional information are calculated in relation to the degree of similarity. Thus, data from similar components will imply a relatively high value of the weight, while data from distant components will have a relatively lower weight or will not be taken into account at all. The mean estimator will adopt the following form:

$$_{MES}\overline{x}_d = \frac{\sum_{i=1}^{n_d} x_i + \sum_{i=1}^{n_{\sim d}} x_i w_i}{n_d + \sum_{i=1}^{n_{\sim d}} w_i}. \tag{3}$$

where w_i stand for weights for the components from outside the small domain d, $n_{\sim d}$ is the size of all domains except for domain d.

It is worth to pay attention to one of the advantages of the MES estimator, which consists in the possibility of using prior information derived from outside the study. Namely, while establishing the similarity between domains one can use data from completely different, e.g. earlier studies or other available information about the population. In such a case, it is possible to derive estimators of parameters for the given domain, which is not represented in the sample.

The establishment of the similarity of the studied domain to other domains in the population may be carried out i.a. using the method of multidimensional analysis, like a k-means grouping method. A different possibility to use additional information about units from outside the small domain gives as evaluation of similarities between units. The first proposal based on a k-means grouping method. Components belonging to the domain of study have to be classifying into k centres. The weights for components from outside the small domain should be calculated proportionally to the distance from the component to the nearest grouping centre.

The second proposal, which was applied in this paper, is based on individual distances among all units in the sample. In this study the Euclidean distance measure is used. The presumption was undertaken that the weight of component from outside the domain of interest should be run on the distance to the nearest component from small domain.

There were used four different methods of establishing weights:

1. The weight $w_i = 1$ was assigned for n nearest components form outside small domain to each individually component from small domain ($n = 1, 2, 3, 4$). All others components have weights equal to zero. These variants of weights establishing were labelled as n1, n2, n3, n4.

2. The weight $w_i = 1$ was assigned for p (p = 5%, 10%, ..., 30%) components form outside small domain with smallest distances to any component from small domain. All others components have weights equal to zero. These variants of weights establishing were labelled as pn5, pn10, ..., pn30.

3. This variant was similar to previous, but the weight was proportional to the smallest distance to any component from small domain, $w_i = 1$ for the nearest component, $w_i = 0$ for all components with distance measure higher than k-th percentile (k = 5, 10, ..., 50) of minimal distances. These variants of weights establishing were label as pnw5, pnw10, ..., pnw50.

4. This variant was similar to previous, but the weight was proportional to the square of smallest distance to any component from small domain, $w_i = 1$ for the nearest component, $w_i = 0$ for all components with distance measure higher than k-th decile (k = 1, 2, ..., 8) of minimal distances. These variants of weights establishing were label as pnwk1, pnwk2, ..., pnwk8.

III. EVALUATION OF PROPERTIES OF THE MES ESTIMATOR

To evaluate the properties of estimators of the ΘY_d parameter in this study the mean bias of the estimator in all experiments was used, calculated according to the following formula:

$$BIAS_f = \frac{\sum_{i=1}^{s}(T_{f,i} - \Theta Y_d)}{s}. \tag{4}$$

where: $T_{f,i}$ is the value of the f-th estimator in the i-th experiment; ΘY_d is the real value of mean of the variable Y in domain d; s is the number of simulations.

The second element of the evaluation was the mean square error, calculated according to the following formula:

$$MSE_f = \frac{\sum_{i=1}^{s}(T_{f,i} - \Theta Y_d)^2}{s}. \tag{5}$$

IV. PROCEDURE OF A MONTE CARLO ANALYSIS

To evaluate an influence of selecting particular weights on efficiency of the MES estimator, some simulation experiments[1] were carried out.

For the sequence of three covariance matrix with mean[2] value of correlation coefficient r_{ij} = 0.2, 0.3, 0.4 in subsequent 1000 repetitions, in each repetition 1000 units (n = 1000) were generated from 7-dimension multivariate normal distribution[3] with a given covariance matrix[4]. One variable is considered as the variable of interest (i.e. the inference relates to this particular variable), and the other six variables are regarded auxiliary ones. First 100 units were assigned to small domain and were generated from 7-dimension multivariate normal distribution with marginal distribution $N(0.1, 0.8)$. Others 900 units assigned to other domains were generated from 7-dimension multivariate normal distribution with marginal distribution $N(0.0, 1.0)$. Subsequently the values of expansion, synthetic and MES estimator were calculated for variables from 1 to 7.

After all repetitions the bias and the mean square error were calculated for all variables. For the obtained results average values were computed.

In the simulation specificity of small domain and level of correlation between variables were mutually independent. The correlation between variables in the sample was about 0,5%–1% higher than reflected in covariance matrix. It resulted from combining of units from small and other domains. The small domain was differ from others with mean level of variable and was more homogeneous.

V. RESULTS OF THE STUDY

The proper choice of the method of establishing weights seems to be a crucial factor of efficiency of the modified synthetic estimator. Very good results were obtained with all the methods, but the last two of them were most effective. MES estimator in all cases was more efficient than the expansion estimator and synthetic estimator. Values of the mean square error for all estimators are presented on figure 1. The comparison of methods of establishing weights is presented on figure 2.

---

[1] All simulations quoted in this paper were carried out using Mathlab 7.1

[2] All correlation coefficients were established at the same value, but because of appearing correlations between randomly generated variables, the final covariance matrix could be slightly different than the established one.

[3] All variables had the standard normal distribution.

[4] Algorithm from Wieczorkowski, Zieliński (1997).

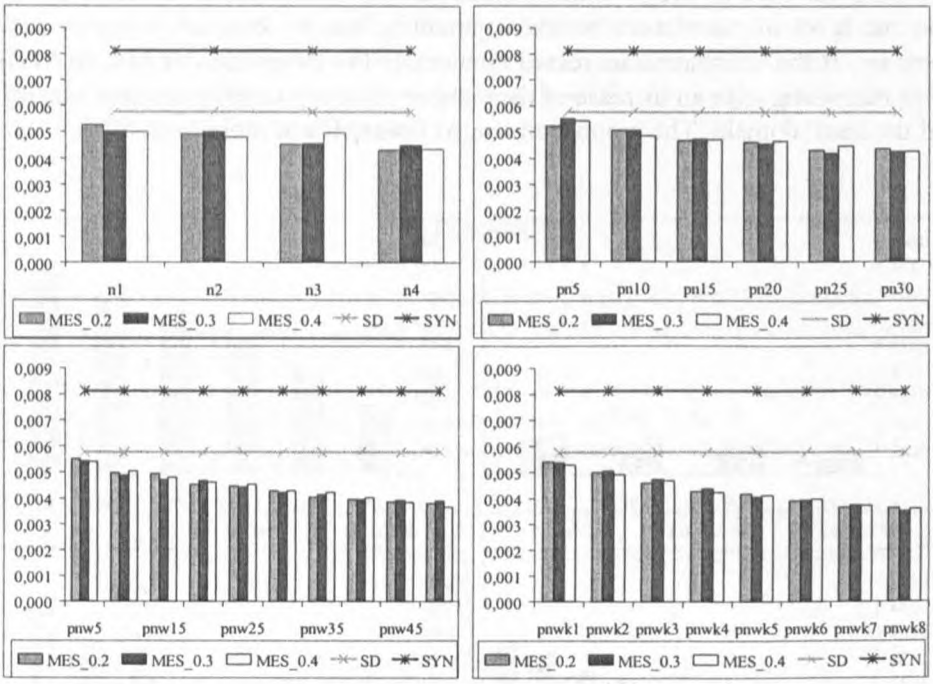

Figure 1. The mean square error of small domain estimators for various covariance matrixes
Source: own study.

It can be observed that with the number of units increasing, the efficiency of MES increases too, but the increase is smaller with each following unit (fig. 2).

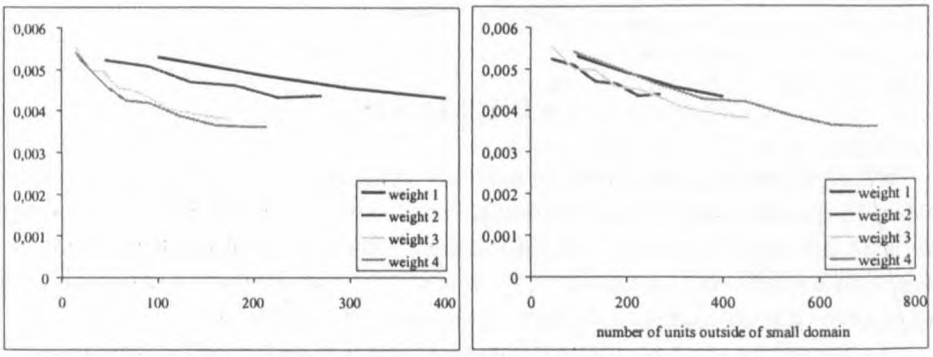

Figure 2. The mean square error of MES estimator for $r_{ij} = 0.2$
Source: own study.

The variance of MES estimator in simulation turned out to be independent on the level of correlation between variables, but the bias of estimator was smaller, if the correlation increased (figure 3). The proportion of bias in MSE was increasing with an increase of the number of incorporated units from outside of the small domain. The proportion ranged from 0.1% to more than 40%.

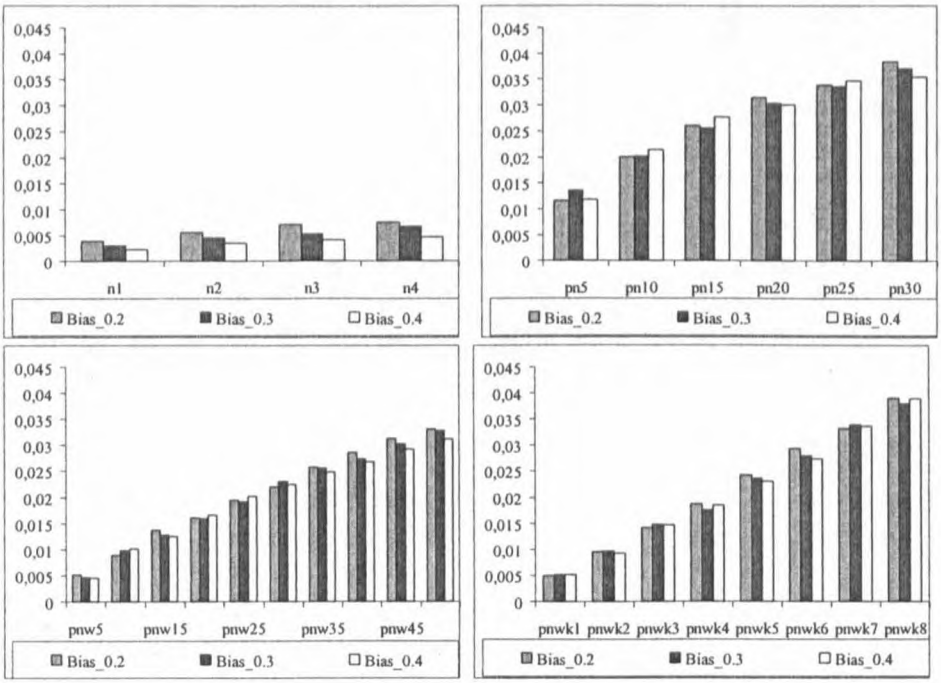

Figure 3. Bias of MES estimator for various covariance matrices
Source: own study.

CONCLUSIONS

An application of the modified synthetic estimator seems to be a good alternative to the estimation of distribution parameters in small domains, in particular in those domains, which differ significantly from the population. It is characterised with a relatively low variation, even if its bias may be quite considerable, in all of cases it is smaller than the bias of the synthetic estimator.

An important issue is an establishment of the way of weighing additional information. It seems that a better solution is to establish the weight for each observation derived from outside of the small domain individually, on the basis of

the distance of each component from components belonging to the small domain (3-rd and 4-th presented method). This method, however, requires the presence of an appropriate number of components from the small domain in the sample. There is also possibility, if the domain is very specific, that in the whole sample there will be only a few similar units.

REFERENCES

Bracha C. (1996) *Teoretyczne podstawy metody reprezentacyjnej*, PWN, Warszawa

Domański C., Pruska K. (2001) *Metody statystyki małych obszarów*, Wyd. Uniwersytetu Łódzkiego

Jurkiewicz T. (2001) *Efficiency of Small Domain Estimators for the Population Proportion: A Monte Carlo Analysis*, Statistics in Transition, Vol. 5, No 2

Kordos J. (1999) *Problemy estymacji dla małych obszarów*, Wiadomości Statystyczne 1/1999

Wieczorkowski R., Zieliński R. (1997) *Komputerowe generatory liczb losowych*, WNT, Warszawa

Tomasz Jurkiewicz

WPŁYW POZIOMU ZALEŻNOŚCI MIĘDZY ZMIENNYMI I SYSTEMU USTALANIA WAG NA EFEKTYWNOŚĆ ZMODYFIKOWANEGO ESTYMATORA SYNTETYCZNEGO – ANALIZA MONTE CARLO

Problem zbyt małej liczby obserwacji w próbie, reprezentującej określoną domenę populacji, może być rozwiązany między innymi poprzez zastosowanie takich estymatorów, które do szacowania parametrów w określonej subpopulacji (małym obszarze, domenie) wykorzystują dodatkowe informacje z pozostałej części próby. Jedna z metod estymacji dla małych domen zwana estymacją syntetyczną sprawdza się przy założeniu, że rozkład (albo któryś z parametrów rozkładu) w badanej małej domenie jest identyczny z rozkładem całej populacji. Założenie to pozostaje zazwyczaj niespełnione, zwłaszcza w przypadku specyficznych domen, co skutkuje dużymi błędami estymacji.

Zastosowanie zmodyfikowanego estymatora syntetycznego (MES) zakłada dwuetapowy proces estymacji. W pierwszym etapie za pomocą metod klasyfikacji lub badania podobieństw określa się podobieństwa jednostek należących do małej domeny do jednostek z pozostałej części próby. Drugim krokiem jest wykorzystanie w estymacji, za pomocą odpowiednio skonstruowanych wag, informacji tylko od tych jednostek, które są podobne do jednostek z małej domeny.

Ważnym czynnikiem wpływającym na efektywność zmodyfikowanego estymatora syntetycznego jest dobór metod ustalania wag dla poszczególnych jednostek badanej zbiorowości. Au-tor przedstawia wyniki symulacyjnego badania efektywności estymatora MES przy zastosowanych różnych sposobach ustalania wag.