ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 228, 2009

*Tomasz Jurkiewicz**

EFFICIENCY OF THE MODIFIED SYNTHETIC ESTIMATOR – MONTE CARLO ANALYSIS

Abstract. The problem of insufficient number of sample observations representing a population domain of interest (small area) can be solved by applying estimators which will be able to combine sample information from the given domain with information about sample units representing other domains. Modified Synthetic Estimator (MES) can be regarded as one of the proposals in this field.

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. 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.

Author presents and discusses some results of Monte Carlo analysis aimed at comparing efficiency between the MES and other estimators.

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

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 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

* Ph.D., Department of Statistics, University of Gdańsk, jurkiewicz@wzr.pl.

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.

Small area statistics provides estimation methods which tend to be complex and sometimes difficult for applications, e.g. in business research. Therefore, there is a need for developing techniques which will be efficient in one hand, and reasonably simple in applications on the other hand.

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.

The essence of indirect estimation consists in "borrowing information" from other domains or other sources in order to improve efficiency of estimation in the domain of interest. In case of random sample surveys it is often possible to use various sources of additional information or statistical data (see: Domański, Pruska (2001), Jurkiewicz (2001), Kordos (1999)).

The main purpose of this paper is to compare efficiency of the modified synthetic estimator with some other estimators recommended as reasonably simple ones in applications.

II. ESTIMATORS OF SMALL DOMAINS

The direct estimator (1) of an unknown mean $\overline{Y}_d$ in a small domain is the simple domain estimator:

$$_{DIR}\overline{y}_d = \frac{\sum_{i=1}^{n_d} y_i}{n_d} \tag{1}$$

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.

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.

In synthetic estimation it is assumed that the structures of the studied population in the small domain and outside the domain are uniform. However, in the case of inference about the arithmetic mean, it is sufficient to require the similarity of the means in the population and in the investigated domain. The mean estimator has the following form:

$$_{SYN}\overline{y}_d = \frac{\sum_{i=1}^{n} y_i}{n} \tag{2}$$

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 incompatibility between the assumptions on which the estimation technique is based 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.

Both the ratio estimator (3) and the synthetic ratio estimator (4) incorporate information about an auxiliary population variable which is strongly correlated with the investigated variable, and whose actual value in the domain is known. Moreover, for the synthetic ratio estimator, by analogy to the synthetic estimator, an assumption is made that there is constant ratio of the investigated and auxiliary variables in the population and in the domain of interest:

$$_{RAT}\overline{y}_d = \frac{_{DIR}\overline{y}_d}{_{DIR}\overline{x}_d}\overline{X}_d \tag{3}$$

$$_{SYN\_RAT}\overline{y}_d = \frac{\overline{y}}{\overline{x}}\overline{X}_d \tag{4}$$

The composite estimator (5) and the composite ratio estimator (6) are linear combinations of the direct or the ratio estimator and an appropriate synthetic estimator:

$$_{COMP}\overline{y}_d = \alpha \cdot {}_{DIR}\overline{y}_d + (1-\alpha)\,{}_{SYN}\overline{y}_d \tag{5}$$

$$_{COMP(RAT)}\overline{y}_d = \alpha \cdot {}_{RAT}\overline{y}_d + (1-\alpha)\,{}_{SYN\_RAT}\overline{y}_d \tag{6}$$

Coefficient α, considered as weight in the above combination, is given the value which minimizes the mean square error of the estimator. The optimal weights, taking into account some simplifications, are given by the following formulae:

$$\alpha = \frac{MSE(_{SYN}\bar{y}_d)}{MSE(_{SYN}\bar{y}_d) + MSE(_{DIR}\bar{y}_d)}$$

or

$$\alpha = \frac{MSE(_{SYN\_RAT}\bar{y}_d)}{MSE(_{SYN\_RAT}\bar{y}_d) + MSE(_{RAT}\bar{y}_d)} \tag{7}$$

However, the practical problem which occurs in applications relates to the establishment of actual value of MSE of synthetic estimators, which would also cover the bias that arises as a result of inadequate or imprecise assumptions. As a consequence, the following variants are considered in the paper:

1) for composite estimator the weight $\alpha = 0{,}5$ (denoted by COMP(0,5))[1] is arbitrary set up;

2) on the basis of all performed (COMP(opt)) experiments the optimal weight was obtained, which can be considered as a benchmark, although very difficult to be established in practice;

3) the optimal weight was obtained for the composite ratio estimator (COMP(RAT_opt)) on the basis of all conducted experiments;

4) the coefficient α for composite estimator was established on the basis of sample results (which means that the assumption of synthetic estimation is met, and as a consequence the bias does not exist).

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 peculiar domains, which 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 (see: Jurkiewicz (2008)). The proposed procedure of estimation is carried out in two stages. The first step consists in establishing 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 rela-

[1] This way of determining the weight is suggested by Bracha (2003).

tively lower weight or will not be taken into account at all. The mean estimator will adopt the following form:

$$_{MES}\overline{y}_d = \frac{\sum_{i=1}^{n_d} y_i + \sum_{i=1}^{n-n_d} y_i w_i}{n_d + \sum_{i=1}^{n-n_d} w_i} \tag{8}$$

where w_i stand for weights for the components from outside the small domain d.

Establishing the degree of similarity between the studied domain and the other domains in the population may be carried out i.a. using the method of multidimensional analysis (see: Jurkiewicz, Najman (2004, 2006)). The 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 (MES(1)) or to the centre of small domain (MES(2)).

The weights for the components were calculated using the following formula:

$$w_i = \begin{cases} 1 - \dfrac{d_i - \min(d_i)}{Q_{s,200}(d_i) - \min(d_i)} & d_i < Q_{s,200} \\ 0 & d_i \geq Q_{s,200} \end{cases} \tag{9}$$

where: d_i – Euclidean distance; $Q_{s,200}$ – s-th 200-quantile of distances; $s = 1, 2, ..., 200$.

Parameter s influences the scope of information „borrowed" from outside the small domain.

III. PROCEDURE OF A MONTE CARLO ANALYSIS

For the sequence of nine covariance matrices with the average values[2] of correlation coefficient $r_{ij} = 0.1, 0.2, ..., 0.9$ in subsequent 10.000 repetitions, in each repetition 1000 units ($n = 1000$) were generated from an 11-dimension

[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.

multivariate normal distribution[3] with a given covariance matrix[4]. For each of ten domains ($n_{d(i)}$ = 100) the distribution was shifted in such a way that the mean values in small domains differ one from another by the quantities ranged from -20% to 20% of the standard deviation (e.g. in the first domain "D1" the mean value was –0,2019, in the second "D2" –0,1566, ..., in the last one "D10" +0,2019). This means that the assumptions of the synthetic estimation were not satisfied. The first variable out of 11 was the variable of interest, and the second one was the auxiliary variable used by ratio estimators. The remaining variables have been transformed[5] into variables in ordinal scale with 4 variants (variables from 3 to 5) and into dummy variables (variables from 6 to 11). In each repetition values for all estimators were calculated (including all variants of MES(1) and MES(2) estimators for parameter s ranging from 1 to 200) and after completing the repetitions the bias and the mean square error were calculated for all domains[6].

IV. RESULTS OF THE STUDY

On the basis of the experiments carried out, one can say that synthetic estimators applied for the domains which differ from the whole population by more than 10% of the standard deviation are completely inefficient (e.g. for D1, D2, D3, see Figure 1)[7]. Composite estimators exhibit higher efficiency, however their major weakness in practical situations lies in difficulties with establishing the weight. Estimators COPM(0.5) and COMP(var) with practically established weights tend to be less efficient. MES estimators (especially MES2) with optimal scope of information „borrowed" from outside the small domain, are most efficient in nearly all domains.

[3] All variables had the standard normal distribution.

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

[5] The purpose of this transformation is to make the simulation experiments more adequate to practical business problems in which qualitative variables are most frequently considered. Correlations among variables in the sample was slightly different than reflected in covariance matrix, which was the result of the transformation mentioned above.

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

[7] Due to space limitations of article, only some major parts of the results are presented in this paper. The entire set of results obtained can be found on: http://www.jurkiewicz.pl/pliki/wyniki_MSA07.xls

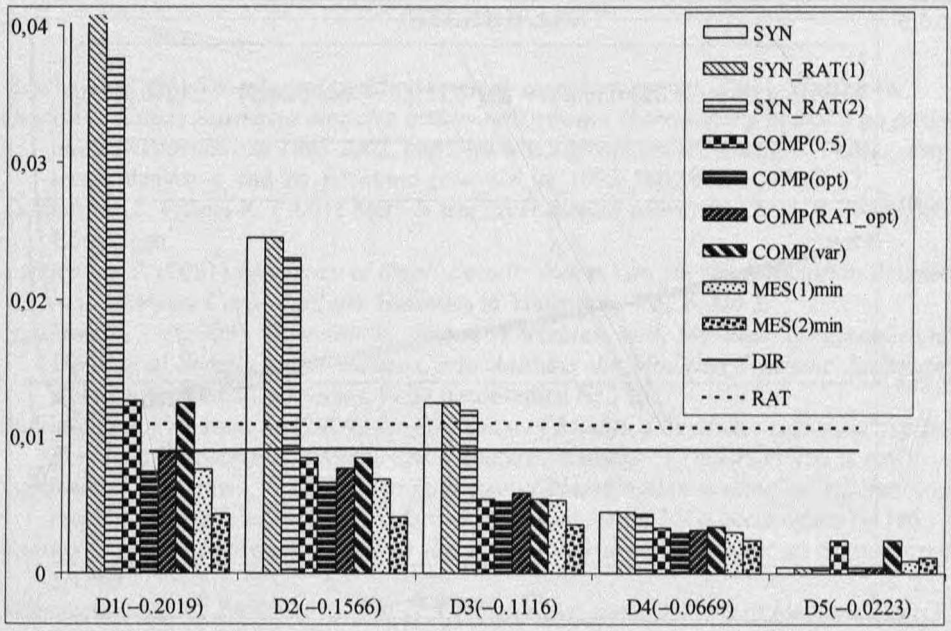

Figure 1. The absolute mean square error of small domain estimators for the covariance matrix ($r_{ij} = 0.4$)*

* In order to be legible, the graph present results obtained for only five small domains. In brackets, the number of the domain is accompanied by the size of the difference between the mean values in the domain and the whole population. Results for the remaining five domains were similar.

Source: own study

It can be observed that if the amount of "borrowed" information increases, the efficiency of MES(2) increases up usually to a certain limit (see Figure 2). The choice of the appropriate quantile ($Q_{s,200}$ parameter) for the borrowed information is determined by the degree of association among the variables and the size of difference between the actual mean value in the analyzed domain and the whole population. This is the problem, which remains to be solved in further studies.

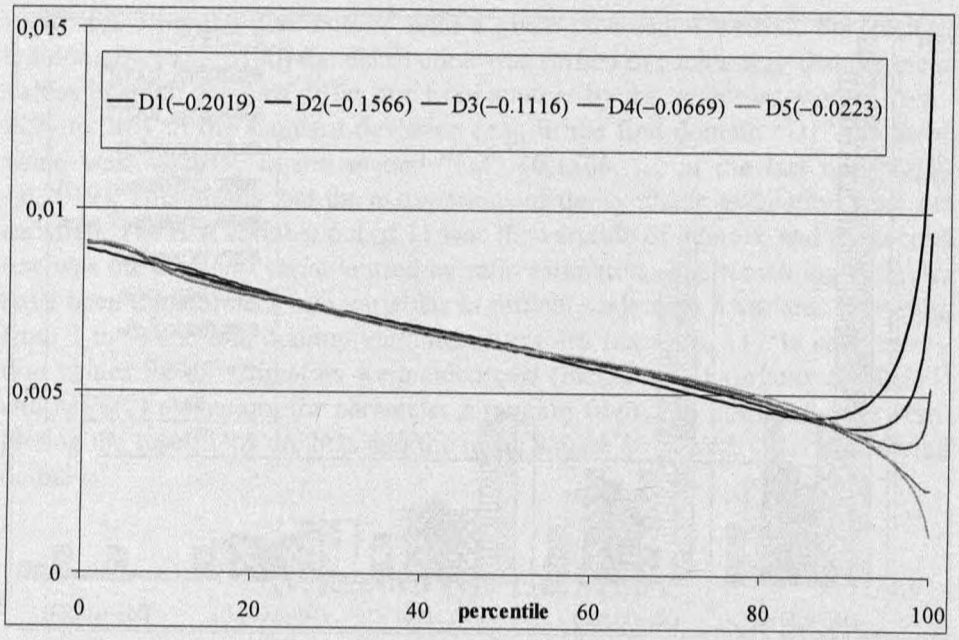

Figure 2. The mean square error of MES(2) estimator for $r_{ij} = 0.4$

Source: own study.

V. 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 and in such circumstances when relatively weak correlation is observed among auxiliary variables. An important issue is an establishment of the way of weighing additional information. It seems that a possibly good solution for establishing the weight is to determine it on the basis of distances between centers of particular domains, except for those domains which lie very close to the centre of the analyzed population. Moreover, the degree of correlation among variables influences significantly both the size and the structure of the mean square error of the estimator. One of the problems, which remains to be tackled is a way of determining an optimal amount of information, which in practice means an unknown number of units "borrowed' from outside the considered domain.

BIBLIOGRAPHY

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

Bracha C. (2003) *Estymacja danych z badania aktywności ekonomicznej ludności na poziomie powiatów dla lat 1995-2002*, http://www.stat.gov.pl/cps/rde/xbcr/gus/ PUBL_ estymacja_danych_z_bad_na_poziomie_pow_dla_lat_1995_2002.pdf; 2007-12-27.

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.

Jurkiewicz T. (2008) *Correlation Among Variables and Methods of Establishing Weights of Sample Units. Monte Carlo Analysis the Modified Synthetic Estimator*, Acta Universitatis Lodziensis, Folia oeconomica Nr 216.

Jurkiewicz T., Najman K. (2004) *An Efficiency of Modified Synthetic Estimator for the Population Proportion: A Monte Carlo Analysis*, Statistics in Transition Vol. 6, No 5.

Jurkiewicz T., Najman K. (2006) *An influence of classification method on efficiency of modified synthetic estimator*, Acta Universitatis Lodziensis, Folia oeconomica Nr 196.

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

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.

Rozwijane przez statystykę małych obszarów metody estymacji są często skomplikowane i trudne do praktycznego zastosowania np. w badaniach biznesowych. Stąd też istnieje po-trzeba rozwijania także metod, które będą łatwe w aplikacji i wystarczające efektywne. Jedną z takich propozycji może być zmodyfikowany estymator syntetyczny (MES).

Zastosowanie estymatora 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.

Autor przedstawia wyniki porównania efektywności estymatora MES z innymi estyma-torami na bazie eksperymentów symulacyjnych.