ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 225, 2009

*Tomasz Jurkiewicz**

AN INFLUENCE OF DISTANCE MEASURE AMONG SAMPLE UNITS ON EFFICIENCY OF THE MODIFIED SYNTHETIC ESTIMATOR: MONTE CARLO ANALYSIS

Abstract

The problem of insufficient number of sample observations representing a given 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. Synthetic estimation technique assumes that the distribution of the variable of interest is the same in the given domain and in the entire population. This assumption, however, is rarely met, and as a result, one can obtain large estimation errors.

Use of 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.

Chosen distance measure is one of the crucial factors in using *MES* estimator. Author presents Monte Carlo analysis of the efficiency of *MES* estimator using different distance measures between sample units.

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

1. Introduction

Probability sample surveys seem to be an efficient way of satisfying a growing demand for statistical information covering economic and social developments. Because of organisational and financial constraints those surveys, however, are not able to supply reliable data for a more detailed division of the

* Ph.D., Department of Statistics, University of Gdańsk.

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 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 "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. There is no reason for which the scope of statistics of small areas should be confined to territorial (administration) units. 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 main purpose of the paper is an attempt to evaluate an influence of suggested distance measure on efficiency of the modified synthetic estimator.

2. Estimators of small domains

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

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. That estimator will have the following form for the mean parameter:

$$_{SD}\overline{x}_d = \frac{\sum_{i=1}^{n_d} x_i}{n_d} \tag{1}$$

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.

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. To this end an assumption is made that the structure of the studied population in the small domain and outside of it 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 feature in the population and in the domain do not essentially differ. For the mean value of the estimator one can adopt the following statistics:

$$_{syn}\overline{x}_d = \frac{\sum_{i=1}^{n} x_i}{n} \tag{2}$$

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 further the assumptions laying at the base of the estimation are from 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 unfulfilled, in particular in case of specific domains, what results in large estimation errors. The solution to the 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 establishing what 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 proposition estimator will adopt the following form:

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

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.

The establishment of the similarity of the studied feature to other features in the population may be carried out i.a. using the method of multidimensional analysis, like a k-means grouping method. In this paper, to establish which units are similar to the units representing the studied domain, different distance measures are used.

While establishing the weights for components from outside the small domain, when k-means method is used, an assumption could be made that the weights should be in direct proportion to the percentage share of units from the small domain, which were found in the given class. The weights may be written as:

$$w_j = \frac{\frac{n_{dj}}{n_d}}{\max\limits_j \left(\frac{n_{dj}}{n_d} \right)} \qquad (4)$$

where n_{dj} – number of units belonging to the domain d which were found in the class j.

For instance, if in the i-th class twice as many components from small domain were found than in the j-th class, then all components from outside the small domain in the i-th class will have the same weight and it will be a weight twice as high as the one used for components from the j-th class.

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 it is possible to use data from completely different, e.g. earlier studies or the available information about the population. In such a case, it is possible to calculate estimators of parameters for a domain, which is not represented in the sample.

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. Weights for components from outside the small domain should be calculated proportionally to the distance from component to the nearest grouping centre.

The second proposal, which was applied in this paper, is based on individual distances between all units in the sample. The presumption was undertaken that the weight of component from outside domain of interest should be run on the distance to the nearest component from small domain. The weight $w_i = 1$ was assigned for 2 nearest components to each component from small domain. All others components had weights equal to zero.

In the study the following distance measures are compared:

"euclidean"	Euclidean distance
"seuclidean"	Standardized Euclidean distance, each coordinate in the sum of squares is inverse weighted by the sample variance of that coordinate
"cityblock"	City Block distance
"mahalanobis"	Mahalanobis distance
"minkowski"	Minkowski distance with different exponents
"cosine"	One minus the cosine of the included angle between observations (treated as vectors)
"correlation"	One minus the sample linear correlation between observations (treated as sequences of values).
"spearman"	One minus the sample Spearman's rank correlation between observations (treated as sequences of values).
"hamming"	Hamming distance, percentage of coordinates that differ
"jaccard"	One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ
"chebychev"	Chebychev distance (maximum coordinate difference)

3. 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 estimator in all experiments was used, calculated according to the following formula:

$$BIAS_f = \frac{\sum_{i=1}^{s}(T_{f,i} - \Theta Y_d)}{s} \qquad (5)$$

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.

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

$$sqr(MSE_f) = \sqrt{\frac{\sum_{i=1}^{s}(T_{f,i} - \Theta Y_d)^2}{s}} \qquad (6)$$

4. Procedure of A Monte Carlo analysis

To evaluate an influence of distance measure on efficiency of the *MES* estimator three simulation experiments[1] were carried out.

Simulation 1. For the sequence of six covariance matrix with mean[2] value of correlation coefficient r_{ij} = 0, 0.1, 0.2, 0.3, 0.4, 0.5 in subsequent 1000 repetitions, in each repetition 1000 units (n = 1000) were generated from 11-dimension multivariate normal distribution[3] with a given covariance matrix[4]. Units with first variable values from 0.385 to 0.842 were assumed as a small domain units. The small domain number covered about 150, 15% of the population. Subsequently the values of expansion, synthetic and *MES* estimator were calculated for variables from 2 to 11. After all repetitions bias and mean square errors were calculated for all variables. For the obtained results average values were counted.

The correlation between variables, reflected in covariance matrix, was strongly combined with specificity of small domain. When the level of

[1] All simulations 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).

correlation was higher, units from the small domain distinguish stronger from other units. Because of that limitation, the way of defining small domain units was changed in the second simulation.

Simulation 2. For the sequence of four covariance matrix with mean value of correlation coefficient r_{ij} = 0.1, 0.2, 0.3, 0.4 and for the sequence of three values of difference parameter (diff = 0.2, 0.5, 0.8) in subsequent 1000 repetitions 1000 units were generated from 7-dimension multivariate normal distribution with established covariance matrix. First 100 units were assigned to small domain. For those units to all variables "diff" parameter given above was added.

Subsequently the values of expansion, synthetic and *MES* estimator were calculated for variables from 1 to 7. After all repetitions bias and mean square errors were calculated for all variables and similarly to first simulation from those results average values were counted.

In respect of some results third simulation was made. The simulation was similar to the second one, except that the diff parameter was now equal to 0.2, 0.4, 0.6. The second difference in this study was that only Minkowski distance was used with exponents 1.75, 2.5, 3.25, 4.0, 4.75.

In the second and third simulation specificity of small domain and level of correlation between variables were independent, but the number of units similar to the units from small domain was inversely proportional to the diff parameter.

5. Results of the study

Table 1

Root of mean square errors of small domain estimators for different covariance matrixes

Estimator (distance for *MES*)	Level of correlation					
	0.0	0.1	0.2	0.3	0.4	0.5
chebychev	0.11279	0.11182	0.11745	0.12304	0.13144	0.14491
cityblock	0.05935	0.05981	0.06024	0.05962	0.05771	0.05608
correlation	0.27326	0.49928	0.69388	0.86759	1.02724	1.17724
cosine	0.06992	0.07264	0.08443	0.10567	0.14087	0.19625
hamming	0.46896	0.47422	0.48020	0.49857	0.51346	0.54286
jaccard	0.46896	0.47422	0.48020	0.49857	0.51346	0.54286
mahalanobis	0.02830	0.06109	0.11014	0.16143	0.21536	0.26637
minkowski	0.05991	0.06014	0.06044	0.05997	0.05778	0.05607
seuclidean	0.06026	0.06056	0.06058	0.06009	0.05803	0.05619
spearman	0.23654	0.41845	0.57175	0.70432	0.81914	0.91705
SD	0.07585	0.07531	0.07460	0.07258	0.06916	0.06599
SYN	0.00226	0.06141	0.12075	0.18021	0.24204	0.29990

Source: own study.

The proper choice of a distance measure seems to be a crucial factor of efficiency of the modified synthetic estimator. Application of some distances (e.g. correlation, jaccard, hamming) results in high error levels. Very good results were obtained with all Minkowski distances, *MES* in all cases in the first simulation was more efficient than the expansion estimator. When the level of correlation was increased the efficiency of *MES* was increased too. Values of mean square error for all estimators are presented in Table 1.

In the second simulation the results turned out to be similar. Minkowski's measures were usually better than others. Modified synthetic estimator was less efficient than expansion when small domain was specific and the correlation level between variables was not too high. The ranking of estimators are presented in Table 2.

Table 2

Rank of root of mean square error of small domain estimators for various diff parameters

Estimator	diff = 0.2				diff = 0.5				diff = 0.8			
	level of correlation											
	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
chebychev	9	9	9	9	7	7	7	7	7	7	7	7
cityblock	1	1	1	1	6	4	3	1	6	6	5	5
correlation	13	13	13	13	13	13	13	13	13	13	13	13
cosine	6	6	6	6	1	6	6	6	2	2	6	6
euclidean	2	3	2	2	3	1	1	2	3	3	2	2
hamming	10	10	10	10	10	10	10	10	10	9	10	9
jaccard	10	10	10	10	10	10	10	10	10	9	10	9
mahalanobis	7	7	7	7	8	8	8	8	8	8	8	8
minkowski	4	2	4	4	5	3	4	4	5	5	4	4
seuclidean	2	3	2	2	3	1	1	2	3	3	2	2
spearman	12	12	12	12	12	12	12	12	12	12	12	12
SD	5	5	5	5	2	5	5	5	1	1	1	1
SYN	8	8	8	8	9	9	9	9	9	11	9	11

Source: own study.

There are no significant differences between efficiency of MES estimators using various exponent in Minkowski measure. As Figure 1 shows, for all values of exponent efficiency of MES estimators it stays on the same level.

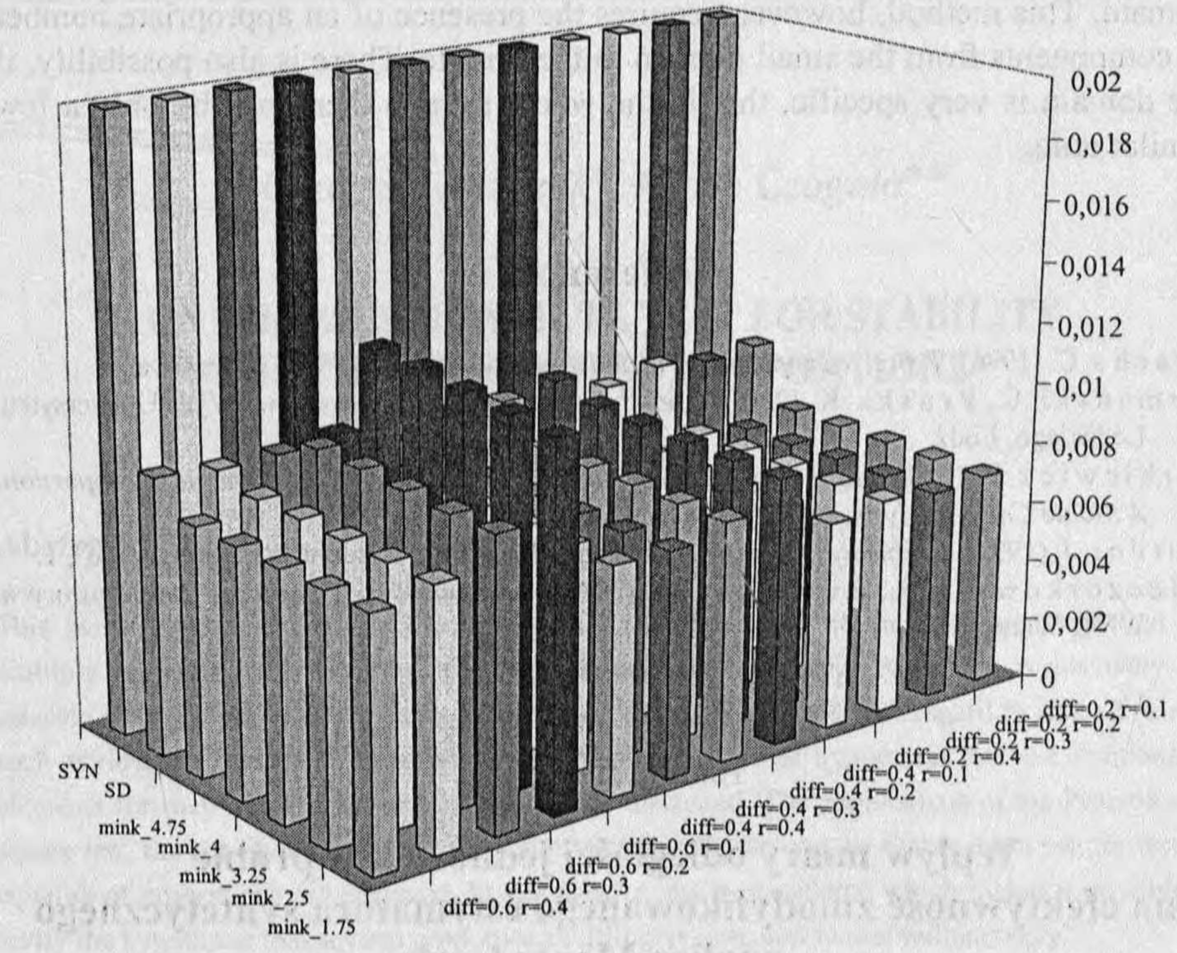

Fig. 1. Root of mean square error of small domain estimators for various covariance matrixes and diff parameters

Source: own study.

6. Conclusions

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 a vast majority of cases it is usually smaller than the bias of the synthetic estimator.

The choice of the distance measure seems to be of primary importance. Some of distance measures give really poor results.

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. 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, Łódź.

Jurkiewicz T. (2001), *Efficiency of small domain estimators for the population proportion: A Monte Carlo analysis*, "Statistics in Transition", **5**, 2.

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

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

Tomasz Jurkiewicz

Wpływ miary odległości jednostek w próbie na efektywność zmodyfikowanego estymatora syntetycznego – analiza Monte Carlo

Problem zbyt małej liczby obserwacji w próbie, reprezentującej określoną domenę populacji, może być rozwiązany m. in. 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, zakłada, że rozkład 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 zastosowana miara odległości. Autor przedstawia wyniki symulacyjnego badania efektywności estymatora *MES* przy zastosowanych różnych miarach odległości do badania podobieństwa jednostek.