ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 196, 2006

*Tomasz Jurkiewicz**, *Krzysztof Najman***

AN INFLUENCE OF CLASSIFICATION METHOD ON EFFICIENCY OF MODIFIED SYNTHETIC ESTIMATOR

Abstract. The problem of insufficient number of sample observations representing a given population domain of interest (small area) can be solved by applying such estimators, which will be able to combine sample information from the given domain with information about sample units representing other domains. One small area estimation method, called synthetic estimation technique, assumes that the distribution of the variable of interest is identical in the given domain and in the entire population. This assumption, however, is rarely met, and as a result one obtains large estimation errors.

In this paper a two-stage estimation procedure is suggested. The first stage consist in applying various classification methods 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 domains, which turned out to be similar to the domain of interest or sample units similar to units from domain of interest, are used to provide sample information with specially constructed weights.

Authors present the results of the suggested procedure in an analysis of the continuing vocational training in construction industry based on a sample survey of enterprises. A bootstrap attempt has been made to assess errors of the suggested estimation procedure.

Key words: small domain estimation, multivariate methods, neural networks.

1. INTRODUCTION

The process of economic and social development results i.a. in a growing demand for statistical information. Representative studies are one of effective ways of satisfying that demand. Because of organizational 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. Too small a number of observations coming from a particular domain may

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

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

be an obstacle in applying certain statistical conclusion generating techniques or lead to considerable errors of estimation (cf. Bracha 1996). One of possible ways of solving that problem is the construction of such estimators, which could use information about other components of a 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 to estimate parameters of a defined subpopulation.

The "small domain" (small area) is defined as a domain of studies, for which information is essential from the point of view of the data user, and it is not possible to acquire that information using the direct estimation method because the size of the sample is too small or when the information acquired with indirect methods is more credible. There is no reason for which the scope of statistics of small areas should be confined to territorial (administration) units. From a 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 aim of the paper is an attempt to evaluate an influence of classification method on efficiency of modified synthetic estimator. The parallel aim of the study is to empirically verify the modified synthetic estimator while studying a sample survey of Continuing Vocational Training (CVTS) in polish enterprises.

2. ESTIMATORS OF SMALL DOMAINS

The essence of indirect estimation consists in "borrowing the information" to strengthen the estimation in the domain being of interest to the statistician. In case of a representative study it is possible to use the following sources of additional data (Jurkiewicz 2001): other domains in the sample; information about the number of particular strata and the number of domains in the studied population; information about the values of an additional variable in a sample; information about values of an additional variable in the studied population; other available data, e.g. data from studies of other periods.

The direct estimator of an unknown parameter ΘY_d in a small domain is the Horvitz-Thompson estimator, known as the expansion estimator. It uses only 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 HT estimator is, however, unbiased, but because of the small size of the sample its variance is usually high. That estimator will have the following form for the proportion parameter:

$$_{HT}p_d = \frac{k_d}{n_d}, \tag{1}$$

where k_d and n_d symbolize the number of elements distinguished in the domain d and the size of the small domain d correspondingly.

Synthetic estimation constitutes one of the first propositions of solving the principal problem of estimation for small domains, which stems from the insufficient size of a sample. To this end an assumption is made that the structure of the studied population in the small domain and outside of it is uniform, what allows 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 proportion the estimator adopts the form of the following statistics:

$$_{syn}p_d = \frac{k}{n}, \tag{2}$$

where k and n denote the number of elements distinguished in the sample and the size of the whole sample respectively.

While applying the synthetic estimation it is very 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 has to be borne in mind at the same time, that firstly, the bias may be of considerable size, and secondly, it is in no way taken into account in formulae for mean square errors and estimators of errors.

2.1. 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, 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 proportion estimator will adopt the following form:

$$ {}_{MES}p_d = \frac{k_d + \sum_{i=1}^{n_{\sim d}} y_i w_i}{n_d + \sum_{i=1}^{n_{\sim d}} w_i}, \tag{3}$$

where:

k_d – number of elements distinguished in the sample belonging to the domain,

n_d – size of the sample in the domain,

w_i – weights for the components from outside the small domain,

y_i – values of the studied zero-one feature.

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. In the present paper the method of grouping *k*-means was used to establish which domains are similar to the studied one. As an alternative method of classification the neural network of the Self Organizing Map (SOM) type was used (Kohonen 1997), and then on the acquired neural map the grouping was carried out according to the *k*-means method.

The number of classes in the grouping process was established using as the criterion the value of the Davies-Boulding clustering evaluation index (Davis, Boulding 1979). The DB index is based on the quotient of variation within the class and the distance between classes. The establishment of the optimum number of classes consists in the calculation of the value of the index for all variants of the number of classes and selecting the variant with the minimum value of the DB index.

While establishing the weights for components from outside the small domain an assumption was made that the weight should be in direct proportion to the percentage share of units from the small domain, which were found in the given class. The weight may be written as:

$$ w_i = \frac{\frac{n_{di}}{n_d}}{\max\limits_i \left(\frac{n_{di}}{n_d}\right)}. \tag{4}$$

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

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 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 case it is also possible to calculate the estimations 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, 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. Euclidean measure of distance between components was used in this study. The weight $w_i = 1$ was assigned for (i) $2 \cdot n_d$ components with least distance to any of components from small domain; (ii) 2 nearest components to each component from small domain. All others components have had weight equal zero.

3. A RANDOM SAMPLE STUDY OF CONTINUING VOCATIONAL TRAINING

The study of the continuing vocational training was carried out in the task 1.4 of project for Ministry of Economy, Labor and Social Policy. The studied population was made of enterprises employing from 10 people registered in the REGON register in 2003. Some sectors were excluded from the population, such as public administration, health services and education. The size of the sample was calculated at the level of 15 000 enterprises. A questionnaire construed for the sake of the study included 18 wide questions provides almost 600 variables. The sample received as a result of enquiry and interviews included 15 012 components.

The building sector is one of the most essential sectors of any economy. Very often the financial results and the level of output of that sector are considered as the barometer of the economy. In publications about the economic situation changes in the level of output for the whole economy are given together with information about the level of output of the construction and building assembly industry (cf. Acs 1996).

In the studied group of enterprises 750 companies (5%) belong to the building sector (Section F, EKD code beginning from 45). This number is sufficient for a credible description of the construction sector as a whole, but is insufficient for more detailed study with the use of direct estimators. Thus the description of that sector could be based on other methods of estimation, giving more credible results. One of those possibilities is to consider that sector as a small domain and to apply the methods of estimation used for small domains.

4. EVALUATION OF PROPERTIES OF THE MES ESTIMATOR

To evaluate the MES estimator the bootstrap method was used. At the beginning 54 variables was selected for evaluating similarities. In subsequent repetitions 1000 components were drawn independently at random, considering components that were found originally in the sample as the population in question. For each simulation grouping with the use of the *k*-means method was made and grouping with the use of the SOM neural network was carried out. Also the euclidean distances between centres and between components was counted. Subsequently the values of expansion, synthetic and MES estimator ware calculated for 40 investigated variables.

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

$$BIAS_f = \frac{\sum_{i=1}^{c}(P_{f,i} - \Theta Y_d)}{s} \cdot 100 \tag{5}$$

where:

$P_{f,i}$ – value of the *f*-th estimator in the *i*-th experiment,

ΘY_d – real value of proportion of the feature *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:

$$sqr(MSE_f) = \sqrt{\frac{\sum_{i=1}^{s}(P_{f,i} - \Theta Y_d)^2}{s}} \cdot 100 \qquad (6)$$

The studied characteristics were the structural indices, that is why the bias and the mean error were expressed in percentage terms for the sake of transparency.

After the experiment the value of the third relative moment was calculated, that is the measures of the skewness of distribution of the acquired values of estimations and the Kołmogorow-Smirnow test for normality of the estimator distribution was applied.

5. RESULTS OF THE STUDY

In the domain of interest – building sector – differences between components was much higher than in other domains. Furthermore even large firms from Section F were alike small firms from other sections, not as good in vocational training. It could be likely effect of inappropriate range of variables used to investigate similarities and all those could strongly affect on results of the study.

5.1. Domain Similarities

Effective number of the small domain sample, measured as sum of weights in MSE estimator, was about 5 times higher than original. In consequence the variance of MES was much smaller than variance of expansion estimator.

When the size of neural network was enlarged, efficiency of MES estimator was increased significantly, however raising of bias could be observed too. Modified synthetic estimator was more efficient than expansion in up to 75% cases. A conclusion could be drawn that classification should be made with a large number of clusters.

Better results was obtain with the "gauss" and "cut-gauss" neighbourhood type in SOM neural network, the worse results was made using "bubble" neighbourhood. That outcome seems to be quite out of the usual run of things for applications of SOM.

Modified synthetic estimator has lower mean square error than synthetic in only 20% cases. It is a straight result of fact that the effective number of sample was 4–5 times smaller and the bias of synthetic estimator was not too large.

5.2. Components similarities

The results obtained, when the weight was assigned for two nearest components to each component from small domain were definitely better. When the weight was assigned for components with least distance to any of components from small domain, received estimations was strongly biased.

Modified synthetic estimator was more efficient than expansion estimator in up to 90% cases, however was less efficient than synthetic (only 10% cases). On the other hand synthetic estimator was higher biased than MES.

Those results could not be directly compared with results for domain similarities. It is because when the weight equal one was assigned for two nearest components the effective number of sample measured as sum of weights in MSE estimator, was only 3 times higher than original. However application of component similarities in estimation procedure appears to be more effective.

The distribution of MES estimator was as close to normal as distribution of synthetic estimator and much closer than distribution of expansion estimator. Skewness of distribution of MES estimator was higher than distribution of synthetic estimator, but there was not any influence on efficiency. Moreover when MES estimator was most efficient, skewness of distribution was higher than average.

6. CONCLUSIONS

Application of the modified synthetic estimator seems to be a good alternative to the estimation of parameters of distributions in small domains, in particular in those domains, which rather significantly differ from the population. It is characterized 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 distribution of the estimator in many cases may be considered as normal or close to normal.

The choice of the method of grouping seems to be of secondary importance, even if differences in effectiveness may be observed, the values of estimation of parameters remain, however, at a similar level. Much more important seems to be the proper choice of set of variables to similarity investigation.

An important issue is the establishment of the way of weighing additional information. In the paper weights related to the number of appearances of components from the small domain in the class and weights related to distance from components belonging to small domain were applied. 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.

REFERENCES

Acs Z. J. (ed.), (1996), *Small Firms and Economic Growth*, Vol. 1, Elgar Publishing Ltd., Brookfield.

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

Davis D. L., Boulding D. W. (1979), "A Cluster Separation Measure", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1** (2), 224–227.

Jurkiewicz T. (2001), "Efficiency of Small Domain Estimators for the Population Proportion: A Monte Carlo Analysis", *Statistics in Transition*, **5** (2).

Kohonen T. (1997), *Self-Organizing Maps*, Springer-Verlag, Berlin–New York.

Tomasz Jurkiewicz, Krzysztof Najman

WPŁYW STOSOWANEJ METODY KLASYFIKACJI NA EFEKTYWNOŚĆ MODYFIKOWANEGO ESTYMATORA SYNTETYCZNEGO

(Streszczenie)

Problem zbyt małej liczby obserwacji w próbie, reprezentującej określoną domenę populacji, może być rozwiązany m. in. poprzez estymatory wykorzystujące informacje o innych jednostkach w próbie. 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.

Problem niespełnienia założeń estymacji syntetycznej może być rozwiązany poprzez zastosowanie dwuetapowego procesu estymacji. W pierwszym etapie za pomocą metod analizy wielowymiarowej, np. za pomocą metody klasyfikacji *k*-średnich, badania odległości czy też wykorzystując sieci neuronowe typu SOM, określa się podobieństwa domen lub jednostek należących do małej domeny do jednostek z pozostałej części próby. Drugim krokiem jest wykorzystanie w estymacji, za pomocą odpowiednio skonstruowanych wag, informacji tylko o tych jednostkach lub z tych domen, które są podobne do badanej małej domeny.

W artykule autorzy przedstawiają rezultaty zastosowanej metody na przykładzie badania reprezentacyjnego kształcenia ustawicznego w branży budowlanej. Za pomocą metod bootsrtrapowych dokonano oceny wpływu stosowania różnych metod badania podobieństw między jednostkami na własności modyfikowanego estymatora syntetycznego.