ACTA UNIVERSITATIS LODZIENSIS
FOLIA OECONOMICA 194, 2005

*Tomasz Jurkiewicz**, *Krzysztof Najman**

PROPOSITION OF APPLYING *K*-MEANS CLASSIFICATION METHOD AND THE SOM TYPE NEURAL NETWORK TO IMPROVE THE EFFICIENCY OF SMALL DOMAINS ESTIMATION IN A REPRESENTATIVE STUDY OF SMALL AND MEDIUM-SIZED ENTERPRISES

Abstract

The problem of a too small number of observations of a sample, representing a defined domain of a population may be solved *inter alia* thanks to the application of estimators which would use information about other components of the sample (derived from outside the defined part of the population) to estimate parameters in a given subpopulation (small area, domain). One of estimation methods for small domains – the synthetic estimation – assumes, that the distribution of the studied small domain is identical with the distribution of the whole population. This assumption remains usually unfulfilled, in particular in case of specific domains, what results in large estimation errors.

The authors present a proposition of two-stage estimation process. In the first stage, using the SOM-type neural networks and using the *k*-means classification method the similarity of components belonging to the small domain with the components belonging to the remaining part of the sample is determined. The second step consists in using the information only from those domains, which are similar to the studied small domain with the help of appropriately construed weights. Authors present the results of the above procedure in the analysis of the building industry on the basis of a representative study of small and medium-sized enterprises. They have also undertaken an attempt to estimate the errors of the synthetic estimation method modified in such a way.

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

I. INTRODUCTION

The process of economic and social development results i.a. in a growing demand for statistical information. One of effective ways of satisfying that demand are representative studies. Because of organisational and financial

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

constraints those studies, however, are not able to supply credible data for a more detailed division of the population into subpopulations (domains of studies). Too small a number of observations coming from a particular domain may be an obstacle in applying certain statistical conclusion generating techniques or lead to considerable errors of estimation (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 or additional information from outside of the sample to estimate parameters of a defined subpopulation (small area, domain).

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 principal aim of the paper is an attempt to determine the qualities of a synthetic estimator after a modification consisting in using only the information about components similar to the ones found in the small domain in the estimation process. The parallel aim of the study is to empirically verify the modified synthetic estimator while studying a concrete sample.

II. 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, strongly related to the studied variable and at least as credible as the variable in question;

– 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. 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 symbolise 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 correspondingly.

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.

III. 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:

$$MSE\,p_d = \frac{k_d + \sum_{i=1}^{n-d} y_i w_i}{n_d + \sum_{i=1}^{n-d} w_i} \tag{3}$$

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. 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 in the form (Davis, Boulding, 1979):

$$DB = \frac{\sum_{k=1}^{c} \max_{i=1}^{k} \left(\max_{j \neq i} \left(\frac{S_i + S_j}{M_{ij}} \right) \right)}{c} \tag{4}$$

where:

S_i – standard deviation in the i-th class,

M_{ij} – distance between classes,

c – number of classes.

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:

$$w_i = \frac{\frac{n_{di}}{n_d}}{\max_i \left(\frac{n_{di}}{n_d}\right)} \gamma \tag{5}$$

where:

n_{di} – number of components belonging to the domain d which were found in the class i,

γ – standardising coefficient from the range (0, 1) defining the maximum value of the weight.

For instance, if in the i-th class twice as many components 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.

IV. EVALUATION OF PROPETIES OF THE *MES* ESTIMATOR

To evaluate the MES estimator the bootstrap method was used. In subsequent repetitions 224 components were drawn independently at random, considering components, that were found originally in the sample as the

population in question. 1000 simulations were made. For each simulation grouping with the use of the k-means method for 5 classes and 20 iterations was made and grouping with the use of the SOM neural network was carried out, assuming the 12×12 neurones with the "bubble" neighbourhood function and the number of clusters established from the (2, 9) range on the basis of the DB index. The above assumptions were optimal for the data from the original sample. Searching of optimum parameters of grouping for each bootstrap sample might improve final results of estimation, but because of a long period of each simulation it was decided to retain uniform parameters in all simulations.

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:

$$BIAS_f = \frac{\sum_{i=1}^{s} (P_{f,i} - \Theta Y_d)}{s} \cdot 100 \tag{6}$$

where:

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

ΘY_d – the 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:

$$sqr(MSE_f) = \sqrt{\frac{\sum_{i=1}^{s} (P_{f,i} - \Theta Y_d)^2}{s}} \cdot 100. \tag{7}$$

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 fourth relative moment, being the measure of flatness of the distribution.

V. A REPRESENTATIVE STUDY OF SMALL ENTERPRISES IN THE POMERANIAN VOIVODSHIP

The study of the small business sector in the Pomeranian and the Lublin voivodships was carried out by an international team of scientists[1]. The studied population was made of small enterprises in the Pomeranian voivodship employing between 10 and 49 people registered in the REGON register on 30th June 1999. Some sectors were excluded from the population, such as the E sector – power generation industry as well as public administration, health services and education.

The size of the sample for the Pomeranian voivodship was calculated at the level of 237 enterprises, i.e. about 5% of the studied population. A questionnaire construed for the sake of the study included 58 questions and was divided into six sections. The sample received as a result of interviews included 239 components. For the sake of the present paper we excluded 15 components, which did not meet certain criteria set at the moment of designing the study project. These were firms, which – according to their REGON number carried activities in other areas than selected for the project and firms, which failed to give answers to many questions included in the questionnaire.

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 (Acs, 1996).

In the studied group of 224 enterprises in the Pomeranian voivodship 19 companies (8.5%) belong to the building sector (EKD code beginning from 45). This number is far insufficient for a credible description of the construction sector with the use of direct estimators. It results from the potentially very high value of the average error of estimation which may even reach the level of 11.5%. Thus the description of that sector should 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.

VI. RESULTS OF THE STUDY

In the Table 2 values of the *MSE* root for estimations of exemplary six variables are given:

– percentage of firms which have been established since 1994,

[1] Phare ACE Programme 1997 contract no. P97–8123–R.

– firms perceiving the advantage over their competitors in attractiveness of their products,

– firms perceiving the advantage over their competitors in high quality of their output,

– firms, that incurred capital investment outlays in 1999,

– firms perceiving their chance in high skills of their employees,

– firms perceiving their chance in good knowledge of the market.

The two last variables were characterized with relatively close levels in the population and in the domain. The first four variables were characterized with quite high a difference between the value in the domain and in the population reaching in the case of the second feature the level of over 20 percentage points.

It may be perceived that even if the variance of the *MES* estimator (Table 1) is much lower in relation to the *HT* estimator, yet because of the bias the mean square error is usually larger. Only in the case of estimations of the last two last variables the *MES* estimator appears to be more effective, but only in the case of parameter γ smaller than 0.5.

Table 1. Variance of estimators (root) using the neural networks of the SOM type depending on the maximum weight

γ	$MESp_1$	$MESp_2$	$MESp_3$	$MESp_4$	$MESp_5$	$MESp_6$
1	3.89%	5.06%	5.18%	5.98%	5.24%	4.88%
0.5	3.99%	4.90%	5.28%	5.78%	5.19%	4.86%
0.3	4.28%	5.07%	5.65%	5.86%	5.40%	5.06%
0.2	4.68%	5.49%	6.18%	6.17%	5.79%	5.41%
0.1	5.60%	6.66%	7.48%	7.22%	6.89%	6.33%
HT	8.59%	10.64%	11.72%	11.50%	10.83%	9.73%

Table 2. Mean square error (root) of estimators using the SOM type neural networks depending on the maximum weight

γ	$MESp_1$	$MESp_2$	$MESp_3$	$MESp_4$	$MESp_5$	$MESp_6$
1	17.73%	15.81%	21.29%	13.30%	11.05%	11.40%
0.5	15.74%	13.95%	18.94%	11.86%	9.84%	10.19%
0.3	13.74%	12.10%	16.56%	10.40%	8.62%	8.96%
0.2	11.89%	10.42%	14.35%	9.04%	7.49%	7.80%
0.1	8.55%	7.41%	10.32%	6.53%	5.42%	5.68%
*HT**	7.5%	9.0%	8.2%	9.5%	9.4%	9.8%
*syn**	12.4%	12.2%	20.6%	17.0%	3.3%	3.8%

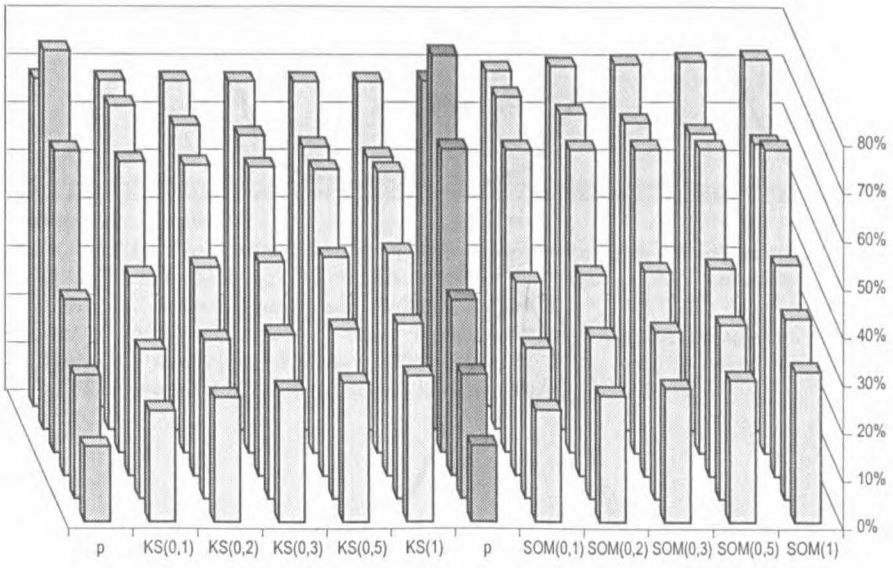
* Approximate values.

Table 3. Mean square error (root) of estimators using the grouping method of k-means depending on the maximum weight

γ	$MESp_1$	$MESp_2$	$MESp_3$	$MESp_4$	$MESp_5$	$MESp_6$
1	17.12%	14.11%	24.22%	14.12%	10.83%	11.28%
0.5	15.47%	12.76%	21.87%	12.78%	9.82%	10.21%
0.3	13.74%	11.34%	19.40%	11.35%	8.76%	9.08%
0.2	12.07%	9.98%	17.03%	9.98%	7.73%	7.99%
0.1	8.90%	7.37%	12.53%	7.35%	5.74%	5.91%

Table 4. Difference in errors of estimators calculated according to the SOM method and estimators calculated with the use of the k-means method

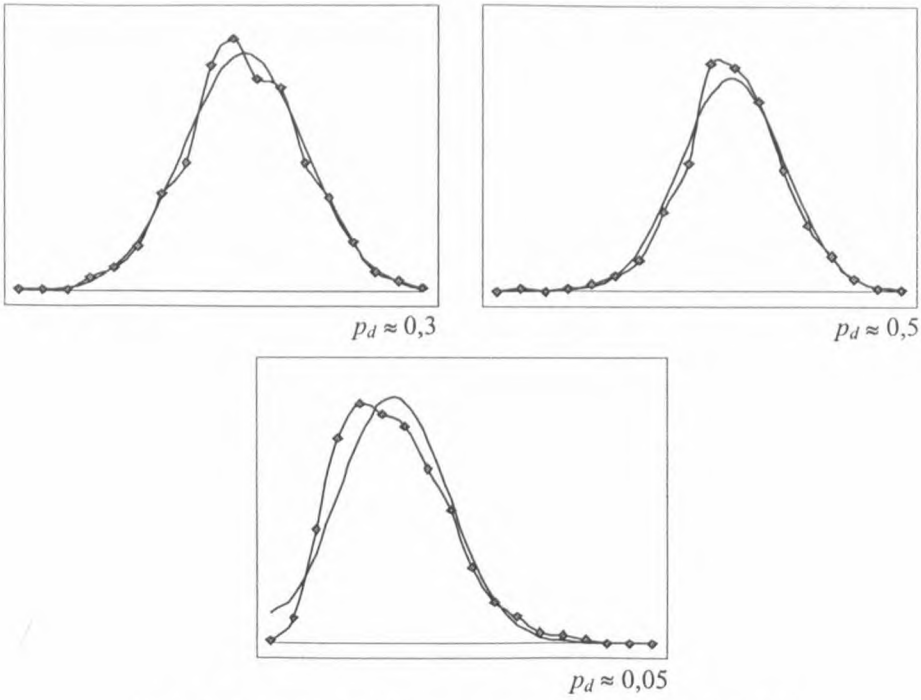
γ	$MESp_1$	$MESp_2$	$MESp_3$	$MESp_4$	$MESp_5$	$MESp_6$
1	0.61%	1.70%	–2.93%	–0.82%	0.22%	0.12%
0.5	0.27%	1.18%	–2.93%	–0.91%	0.02%	–0.02%
0.3	0.00%	0.76%	–2.84%	–0.95%	–0.13%	–0.13%
0.2	–0.18%	0.44%	–2.68%	–0.94%	–0.23%	–0.19%
0.1	–0.35%	0.04%	–2.21%	–0.82%	–0.32%	–0.23%



Graph 1. Values of estimators using the neural networks of the SOM type and the method of grouping of k-means depending on the maximum weight

Comparing results acquired while using two so different methods of grouping it may be said that the calculated values of estimated parameters do not differ too much (cf. Graph 1). The highest observed value amounts to 4.8 percentage points for the fifth variable at the parameter $\gamma = 1$. The effectiveness of estimators remains as well at a similar level, although in the case of the third variable the method of grouping k-means appeared to be definitely less effective.

Examples of received distributions of *MES* estimators for various values of the estimated parameter p_d. were presented in Graph 2. The distributions of estimators are characterised with relatively normal flatness, a_4 in most cases was close to zero and in a great majority of cases had a positive value, which means that the distributions of estimators are slimmer than the normal distribution. For the flattest distribution the value of a_4 was equal to about –0.2. The acquired distributions were also approximately symmetrical, while the value of the asymmetry increased in line with the decrease of the parameter γ. Besides that, at relatively high values of the parameter γ the distributions could be considered as normal (χ^2 test at the division into 18 classes). Certain distortions visible in the graph result from a small number of repetitions of the simulation.



Graph 2. Exemplary distributions of modified synthetic estimators and approximation of the normal distribution

VII. CONCLUSIONS

Application of the *MES* 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 characterised with a relatively low variation, even if its bias may be quite considerable, in a vast majority of cases it is 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.

An important issue is the establishment of the way of weighing additional information. The change in parameter γ, defining the maximum value of the weight resulted in quite meaningful changes both in the estimation of parameters and the effectiveness of estimators. In the paper weights related to the number of appearances of components from the small domain in the class were applied. It seems that a better solution would be 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. (red.) (1996), *Small Firms and Economic Growth*, vol. 1, Elgar Publishing Ltd, Cheltenham, England.

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

Davis D.L., Boulding D.W. (1979), A cluster separation measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-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, New York.

Tomasz Jurkiewicz, Krzysztof Najman

PROPOZYCJA ZASTOSOWANIA METODY KLASYFIKACJI *K*-ŚREDNICH ORAZ SIECI NEURONOWEJ TYPU SOM DO POPRAWY EFEKTYWNOŚCI ESTYMACJI DLA MAŁYCH DOMEN W REPREZENTACYJNYM BADANIU MAŁYCH I ŚREDNICH PRZEDSIĘBIORSTW

Streszczenie

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 supopulacji (małym obszarze, domenie) mogłyby wykorzystać informacje o innych jednostkach w próbie, które pochodzą spoza określonej części populacji. 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.

Autorzy przedstawiają propozycję dwuetapowego procesu estymacji. W pierwszym etapie za pomocą sieci neuronowych typu SOM oraz za pomocą metody klasyfikacji *k*-średnich 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 z tych domen, które są podobne do badanej małej domeny. Autorzy przedstawiają rezultaty zastosowania podanej procedury w analizie branży budowlanej na podstawie wyników reprezentacyjnego badania małych i średnich przedsiębiorstw. Podjęli także próbę oszacowania błędów tak zmodyfikowanej metody estymacji syntetycznej.