Twostep cluster analysis: Segmentation of largest companies in Macedonia

Marija Trpkova[1] - Dragan Tevdovski[2]

One of the important procedures for segmentation and classification of the largest Macedonian companies is twostep cluster analysis. This clustering method is very efficient in classification of large data sets, has the ability to create groups using categorical and continuous variables and it is provided with automatic selection of number of clusters. These are all advantages of twostep analysis compared to the traditional clustering methods.

The goal of this paper is to present valuable application of the twostep cluster analysis in segmentation of the Macedonian companies. Every year, the Central Register of Republic of Macedonia and Euro Business Centre - Macedonia present a publication that reveals the 200 largest and most successful companies in Macedonia. In order to reveal the structure of the Macedonian companies, twostep cluster analysis is performed using the following continuous variables: total revenue in 2007, total revenue in 2006, earnings before taxes in 2007, revenue growth rate 2007/2006 and number of employees. Also, one categorical variable is included, type of industry.

The analysis successfully manages to create solution of four clusters or four different types of companies on the Macedonian market. The first type represents the most successful companies with significantly high revenues, earnings and stabile growth. These companies come from industries such as communications, electricity and manufacturing, and provide significant employment of the work force. The second type represents companies with relatively smaller revenues and earnings compared to the first type, but yet higher than the country's average. These are all manufacturing companies with steady growth. The third group has slightly smaller revenues and earnings than the second group, but the difference is that this group represents companies with high revenue growth rate, representing developing companies with significant potential. These are companies that mostly provide services, companies that provide telecommunication and transport, and also few companies from other industries. The last group represents the smallest companies from the analyzed 200 largest companies, having the smallest revenues, earnings and number of employees. These companies will develop further, but with much smaller rate than the companies in the third group. These are all companies that deal with retail and wholesale trade.

These findings are useful because mainly they provide the general structure of the largest Macedonian companies. For the potential foreign investors this analysis is an insight

[1] Marija Trpkova, Teaching and research assistant, Faculty of Economics, University "Ss. Cyril and Methodius" (Skopje)

[2] Dragan Tevdovski, MSc, Teaching and research assistant, Faculty of Economics, University "Ss. Cyril and Methodius" (Skopje)

to the most lucrative industries in the country. For the government, the presented results give information about which industries dominate in the most successful companies, in order to invest in their development through infrastructure, university education, tax relief and deduction of other expenditures.

Keywords: Twostep cluster analysis, log-likelihood distance measure, Schwarz Bayesian information criterion, segmentation

1. Introduction

In order to achieve the desired economic growth, Macedonian economy needs to attract foreign investments. The inflow of the foreign capital means growth of the gross domestic product, increase of the employment rate, improvement of the overall standard of living of Macedonian population and faster integration in the European Union. This paper represents an effort to create the general structure of the most successful Macedonian companies. For the potential investors it will act as indicator to the most lucrative industries in Macedonia and stimulus for increased inflow of foreign investments.

The twostep cluster analysis proves to be important procedure in clustering of the large data base consisted of 200 companies – observations with five continuous and one categorical variable.

Structure of the paper is following: the first part of the paper briefly explains the twostep cluster algorithm, determining the number of clusters and assignment of the observations in the most appropriate newly created clusters. The second part of the paper shows the application of the procedure in clustering of the 200 most successful Macedonian companies. Used data set and the empirical results are elaborated. Final conclusion about the findings of the analysis is given.

2. Twostep cluster analysis algorithm

Twostep cluster analysis is method of the statistical software package SPSS used for large data bases, since hierarchical and k-means clustering do not scale efficiently when n is very large (Garson 2009). This analysis can be used both for categorical and continuous variables, and has its application when there are categorical variables with three or more categories.

Twostep cluster analysis represents method that requires only one pass throughout the data. The process is consisted of two major steps: first step, where initial clustering of observations into small subclusters is performed and further on these subclusters are treated as separate observations. The decision whether the observation is joined in already formed cluster or a new cluster shall be formed is made on the bases on the distance criteria. The grouping of these new observations

is done by hierarchical cluster method. It is possible for the algorithm of the twostep cluster analysis to determine the number of clusters, or the number of clusters can be assigned previously. The second step is groping, where the subclusters are bases for the analysis, and they are grouped into the required number of clusters. Since the number of subclusters is significantly smaller than the number of observations, the traditional grouping methods are easy to be used. The method is more precise if there are more subclusters (Zhang et al. 1996).

In this analysis, if one or more variables are categorical, the log-likelihood distance measure is used, in such manner that the observations are grouped in the cluster which has the highest values of this measure, using a method developed by Meila and Heckerman (1998). If all of the variables are continuous, the Euclidean distance is used, so that the observations are grouped in the cluster that has the smallest Euclidean distance. SPSS algorithm uses a decrease in the log-likelihood distance measure for combining clusters as the distance measure because the log-likelihood method is compatible with categorical and continuous variables.

The procedure of twostep cluster analysis that uses log-likelihood distance measure assumes normal distribution for continuous variables and multinomial distribution for categorical variables. The twostep cluster analysis gives good results even if the normality assumption is not met. Another assumption of this analysis is that the sample is large (> 200).

The distance measure is needed in both steps, or in the step of the initial clustering and in the clustering step. There are two distance measures available, the first one is log-likelihood distance measure which represents the distance based on probability. The distance between two clusters is in a relation with the decrease of the value of the log-likelihood distance measure, when two clusters are joined in one (Banfield-Raftery 1993). While calculating the log-likelihood distance measure, normal distribution for continuous variables and multinomial for categorical variables is assumed. Also, the independence of the variables and independence of observations is also assumed. The distance between clusters R and S is defined as

$$d_{(R)(S)} = \xi_R + \xi_S - \xi_{(R,S)}$$

where

$$\xi_v = -N_v \cdot \left(\left(\sum_{k=1}^{K^A} \frac{1}{2} \cdot \log(\hat{\sigma}_k^2 + \hat{\sigma}_{v \cdot k}^2) \right) + \left(\sum_{k=1}^{K^B} \hat{E}_{v \cdot k} \right) \right)$$

and where

$$\hat{E}_{v \cdot k} = -\sum_{l=1}^{L_k} \left(\frac{N_{v \cdot k \cdot 1}}{N_v} \cdot \log\left(\frac{N_{v \cdot k \cdot 1}}{N_v} \right) \right)$$

where

K^A is the total number of the continuous variables in the analysis; K^B is the total number of the categorical variables in the analysis ; R_k is the interval or range of the

k continuous variable; N is the number of observations in the data base; N_k is the number of objects in k cluster; $\hat{\sigma}_k^2$ is the estimated variance of the k continuous variable for all data; $\hat{\sigma}_{Rk}^2$ is the estimated variance of the k continuous variable in the R cluster; N_{Rkl} is the number of objects in the R cluster, where k categorical variable takes the l category; $d_{(R)(S)}$ is the distance between the R and the S clusters; (R,S) is the index that represents cluster which is formed by joining of the clusters R and S (Chiu et al. 2001).

If the $\hat{\sigma}_k^2$ is ignored in the equation, the distance between the clusters R and S will be equal to the decreased value of the log-likelihood distance measure when two clusters are joined. The expression $\hat{\sigma}_k^2$ is given as a solution of the rising problem, if $\hat{\sigma}_{v\cdot k}^2 = 0$, by which undefined values for natural logarithm are reached. This problem occurs if the clusters have only one observation.

The other distance measure, the Euclidean distance, can be used only in a situation when all of the variables are continuous. The Euclidean distance between two points is clearly defined. The distance between two clusters is defined by the Euclidean distance between their centroids. The centroid of the clusters is defined as vector consisted of the means of all variables for a given cluster.

The procedure of the twostep cluster analysis begins with the first step, which is creation of initial cluster. This step uses method of sequential clustering. It analyzes the observations and decides if the given observation will join in one of the already formed cluster, or whether it will form a new cluster. This decision is based on the distance criteria.

2.1. Determining the number of clusters

When the process of clustering is started, the question is how many clusters should be formed. The answer depends on the data base. Characteristic of the hierarchical cluster analysis is to form a set of possible solutions from one pass throughout the date, with one, two, three or more cluster. K-means cluster algorithm has to be performed several times (each time for different number of clusters) so that a set of solutions is generated.

For automatic determination of number of clusters SPSS has developed the twostep procedure which is compatible with hierarchical cluster analysis. In the first step, *BIC* or Bayes information criterion or *AIC* or Akaike information criterion statistics is calculated for each different cluster solution with different number of clusters. In the second step, the initial estimate is improved by finding the highest distance increase between the two closest clusters during each stage in the hierarchical clustering.

The statistics BIC and AIC for R clusters is defined as

$$BIC_R = -2 \cdot \sum_{i=1}^{R} \xi_R + m_R \cdot \log(N)$$

$$AIC_R = -2 \cdot \sum_{i=1}^{R} \xi_R + 2 \cdot m_R$$

where

$$m_R = R \cdot \left\{ 2 \cdot K^A + \sum_{k=1}^{K} (L_k - 1) \right\}$$

where L_k is the number of groups in k categorical variable.

2.2. Assignment of the observations

The assignment of the observations into cluster is done by assigning the observations to the nearest cluster, if there is no transformation of the outliers. If there is an outlier transformation, then log-likelihood distance is used.

Let us assume that the extreme observations follow a normal distribution. Two separate likelihood functions are calculated, one when the observation is assigned to unstandardized cluster, and the other one to the nearest cluster of the unstandardized cluster. The observation then is assigned to the cluster that has the highest value of the log-likelihood function. The procedure is equal to assignment of the observation to the nearest cluster (which is not unstandardized) if the distance from the cluster is smaller than the critical value $C = \log(V)$ where $V = \left(\prod_k R_k \right) \cdot \left(\prod_m L_m \right)$. In rest of the cases, the observation is classified as outlier.

The object is assigned in the nearest cluster (which is not unstandardized) if the Euclidean distance is smaller than the critical value $C = 2 \cdot \sqrt{\frac{\sum_{l=1}^{K^A} \hat{\sigma}_{k \cdot l}^2}{K^A}}$, otherwise is classified as outlier.

Missing values are not allowed. The observations with missing values are excluded from the analysis.

3. Application of the twostep cluster analysis in clustering of 200 Macedonian companies

Twostep cluster analysis creates subclusters using hierarchical methods. If the analysis includes large data base, twostep cluster analysis is recommended, as well in situation when the categorical variables are included. The analysis performed with statistical software SPSS gives significant output, as well as variable importance charts.

3.1. Data set

The data set is consisted of 200 Macedonian companies. The data are provided from the Central registry of Macedonia and they refer to the most successful 200 companies. The main purpose of this analysis is to define clusters of companies on the base on the following continuous variables: total income in 2007, total income in 2006, income growth rate 2007/2006, earning before taxes in 2007 and number of employees. Also, there is one categorical variable, type of industry. The final clusters and their profiles will provide with the structure of the main groups of Macedonian companies. Analysis is performed with the statistical software SPSS 15.

3.2. Empirical results

The first step in the analysis is to examine the data and make the necessary transformations. The starting point is standardization of the variables. The twostep analysis performs standardization of the continuous variables, yet the final results are given in original values of the variables. The question about the outliers is solved by the Mahalanobis D^2 distance measure (Hair et al. 2005). This measure shows that less than ten companies may be potential outliers. The decision is to keep these companies in the analysis.

The correlation matrix calculates the Pearson coefficient of correlation between the variables and confirms that the correlation exists. This is expected, especially between the variables income and earnings. This analysis does not have a situation where several variables represent a factor, where if multicollinearity exists, it is necessary to reduce the variables for each factor or set of correlated variables. Because of this, the analysis can be continued, without fulfilment of this assumption.

The analysis includes five continuous variables and one categorical variable. If one or more variables are categorical, then a log-likelihood distance measure is used. If all variables are continuous, then the Euclidean distance is used. For determination of number of cluster Bayes or Akaike information criteria is used. The number of clusters also can be automatically assigned by the researcher. The analysis gives the following results.

Table 1 is automatic clustering table.

The presented statistics are values of the Bayesian criterion and value of BIC change for solutions with different number of clusters. When performing automatic clustering, different criteria are given, so that the best cluster solution is the one that has the lowest value of the Schwarz's Bayesian Criterion, or the lowest vale of Akaike information criterion. The statistical software SPSS chooses the solution that has significantly high value of BIC change and high value of ratio of distance measures. The performed simulation studies confirm that the ratio of BIC changes, which is combined criterion, givers better results than the individual values of Bayesian and Akaike information criterion.

In the presented example, the best solution is the one with 4 clusters, because this solution gives highest value for the ratio of distance measures and the lowest value of the Schwarz's Bayesian Criterion. The SPSS algorithm need not agree with the BIC criterion used alone, though it does in this example. When it differs, in essence the SPSS algorithm judges that the gain in information from having more than the number of clusters specified by BIC alone is not worth the increased complexity (diminution of parsimony) of the model. The researcher has the option to override this default and specify 6 or some other number of clusters (Garson 2009).

Table 1. Automatic clustering

Number of clusters	Schwarz's Bayesian Criterion (BIC)	BIC Change (a)	Ratio of BIC Changes (b)	Ratio of Distance Measures (c)
1	1272,747			
2	990,722	-282,026	1,000	1,839
3	884,535	-106,186	0,377	1,152
4	806,071	-78,464	0,278	2,168
5	825,571	19,500	-0,069	1,906
6	884,966	59,395	-0,211	1,212
7	952,054	67,088	-0,238	1,394
8	1029,408	77,354	-0,274	1,034
9	1107,607	78,199	-0,277	1,146
10	1189,023	81,416	-0,289	1,209
11	1274,238	85,216	-0,302	1,015
12	1359,726	85,488	-0,303	1,318
13	1449,533	89,808	-0,318	1,373
14	1543,036	93,503	-0,332	1,189
15	1638,112	95,076	-0,337	1,126

(a) The changes are from the previous number of clusters in the table.
(b) The ratios of changes are relative to the change for the two cluster solution.
(c) The ratios of distance measures are based on the current number of clusters against the previous number of clusters.
Source: Own creation

The Table 2 presents the distribution of the observations in the cluster, or the number of observations in each cluster. This is the first indicator of the size of clusters. Also, the number of the excluded observations is also given, or in this example 24 observations are excluded because they to not have sufficient data for the chosen variables so they can be grouped in of the clusters.

Table 2. Cluster distribution

Cluster	Number of observations	% of combined cluster	% of total
1	7	4,0%	3,5%
2	64	36,4%	32,0%
3	43	24,4%	21,5%
4	62	35,2%	31,0%
Combined clusters	176	100,0%	88,0%
Excluded observations	24		12,0%
Total	200		100,0%

Source: Own creation

The centroids Table 3 represent descriptive statistics for the continuous variables. The mean values for all continuous variables for each cluster are presented.

Table 3. Cluster centroids

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Total income in EUR (2007)	301.044.003	25.714.168	21.771.629	16.452.301
Total income in EUR (2006)	248.912.033	21.215.261	16.912.291	13.322.724
Income growth rate 2007/2006	27	33	6.119	129
Earnings before tax EUR (2007)	48.756.644	2.103.284	1.714.597	845.832
Number of employees (2007)	1.872	298	371	88

Source: Own creation

The interpretation of the clusters profiles shows that the first cluster has 7 companies. According to the cluster centroids, these are the largest and the most profitable companies, with the highest number of employees and steady growth rate. The main industries of these companies are manufacturing, electricity and

telecommunication. These are the largest and most successful companies in Macedonia.

The second cluster is the biggest and has 64 companies that have significantly lower income and earnings from the companies in the first cluster. These companies have steady growth and the number of employees is also lower than the number of employees in the first cluster. These are middle size companies with normal growth.

Table 4. Frequencies

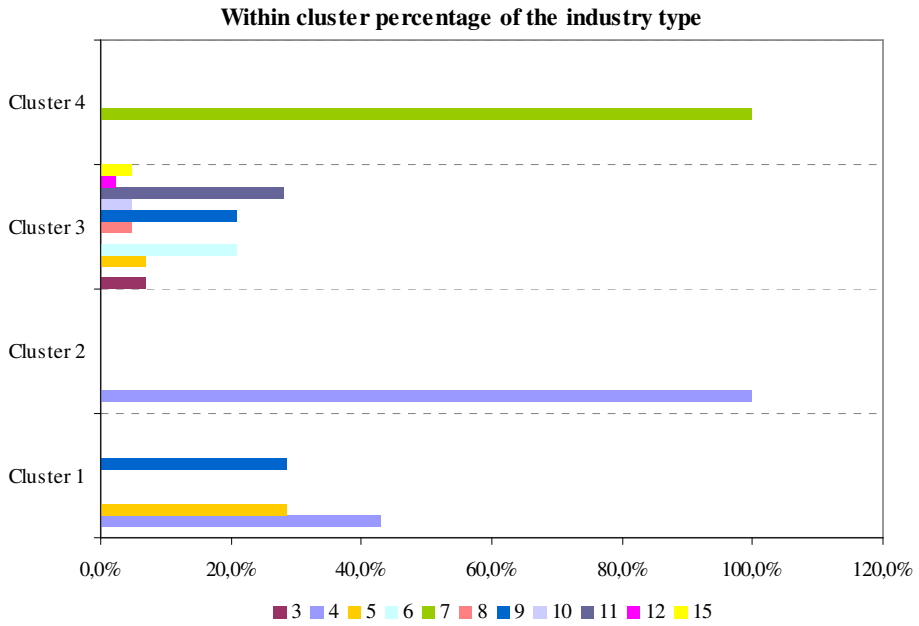
		Industry code										
		3	4	5	6	7	8	9	10	11	12	15
Frequency	Cluster 1	0	3	2	0	0	0	2	0	0	0	0
	Cluster 2	0	64	0	0	0	0	0	0	0	0	0
	Cluster 3	3	0	3	9	0	2	9	2	12	1	2
	Cluster 4	0	0	0	0	62	0	0	0	0	0	0
	Combined	3	67	5	9	62	2	11	2	12	1	2
Percent	Cluster 1	0	4,5	40	0	0	0	18,2	0	0	0	0
	Cluster 2	0	95,5	0	0	0	0	0	0	0	0	0
	Cluster 3	100	0	60	100	0	100	81,8	100	100	100	100
	Cluster 4	0	0	0	0	100	0	0	0	0	0	0
	Combined	100	100	100	100	100	100	100	100	100	100	100

Source: Own creation

The third cluster has 43 companies and has slightly lower income and earnings from the companies in the second cluster. The number of employees is greater than the number of employees in the second cluster. The main difference here is in the income growth rate. The companies in this cluster have very high growth rate, indicating that these are developing companies, with potential for further development in even more successful companies.

The fourth cluster has 62 companies, with lowest income, earnings and number of employees. It is important to say that these companies also have high income growth rate. This is a cluster with companies significantly smaller then the companies in the other clusters, yet since the growth rate is high, further growth can be also expected here.

Figure 1. Graphical presentation of the frequencies (in percentage)



Source: Own creation

The frequency Table 4 uses the descriptive statistics for categorical variables. For each variable separate table is created. In this table is clearly shown that in the first cluster the companies that have industry code 4 – manufacturing, industry code 5 – electricity and industry code 9 – telecommunications dominate. In the second cluster all companies have industry code 4 – manufacturing. In the third cluster there are companies from different industries, practically all industries are present here. In the fourth cluster all companies have industry code 7 – trade.

The industry codes are: 1 – agriculture, forestry and hunting, 2 – fishing, 3 – mining and quarrying, 4 – manufacturing, 5 – electricity, gas, water supply, 6 – construction, 7 – wholesale and retail trade, 8 – hotels and restaurants, 9 – transport, storage and communications, 10 – financial intermediation, 11 – real estate, renting and business activities, 12 – public administration and defence, compulsory social security, 13 – education, 14 – health and social work, 15 – other community, social and personal service activities.

Frequency table does not have data about the industries with codes 1, 2, 13 and 14. The reason for this is because these types of industries are not present in the companies from the observed data base, or they are excluded because of the missing data.

The frequency distribution in the cluster is also shown on a chart on Figure 1, so that the group structure of the categorical variable in each cluster is clearer. The distribution is shown in percentage.

Figure 2 shows the within cluster variation. There is one error bar chart for each categorical variable which shows the arithmetic mean for all clusters for that particular variable. Since a sample is used, there is 95% confidence interval. The first chart represents the mean for the variable total income in 2007. It is clearly shown that the first cluster has significantly higher value from the other clusters, which have approximately same value for this variable. The same situation is for the variable total income in 2006, earnings before tax and number of employees. Only the variable growth rate has significantly higher confidence interval in the third cluster than the other cluster, which has moderate growth.

SPSS offers another group of charts as an outcome in the twostep cluster analysis – charts that show the significance of the variables (Figure 3). For the continuous and categorical variables, on special charts, on the X axis the χ^2 value is given, and on the Y axis the particular variable. Bar lines that are longer than the critical value show that the variable is statistically significant in differentiation of the clusters. The charts are shown in Figure 2.

The first variable is the categorical variable Industry. The value of χ^2 shows that this variable is statistically significant for differentiation of the clusters, especially for clusters 3,4 and 2, while less important for cluster 1. The second variable, total income in 2007 and the third variable, total income in 2006 are variables that significantly contribute in differentiation of the clusters 4, 3 and 1, while they do not contribute in differentiation of the cluster 2.

The fourth variable, growth rate of the total income, has particularly significant influence in differentiation of clusters 2 and 1, small influence in differentiation of cluster 4, and no influence in differentiation of cluster 3. The variable earnings before taxes in 2007 significantly differentiates cluster 4 and 3, while it is insignificant for clusters 2 and 1. The last variable, number of employees in 2007 is significant only for differentiation of cluster 4, while there is no influence on clusters 3, 2 and 1.

Table 5 represents all 200 companies and their cluster membership.

4. Conclusion

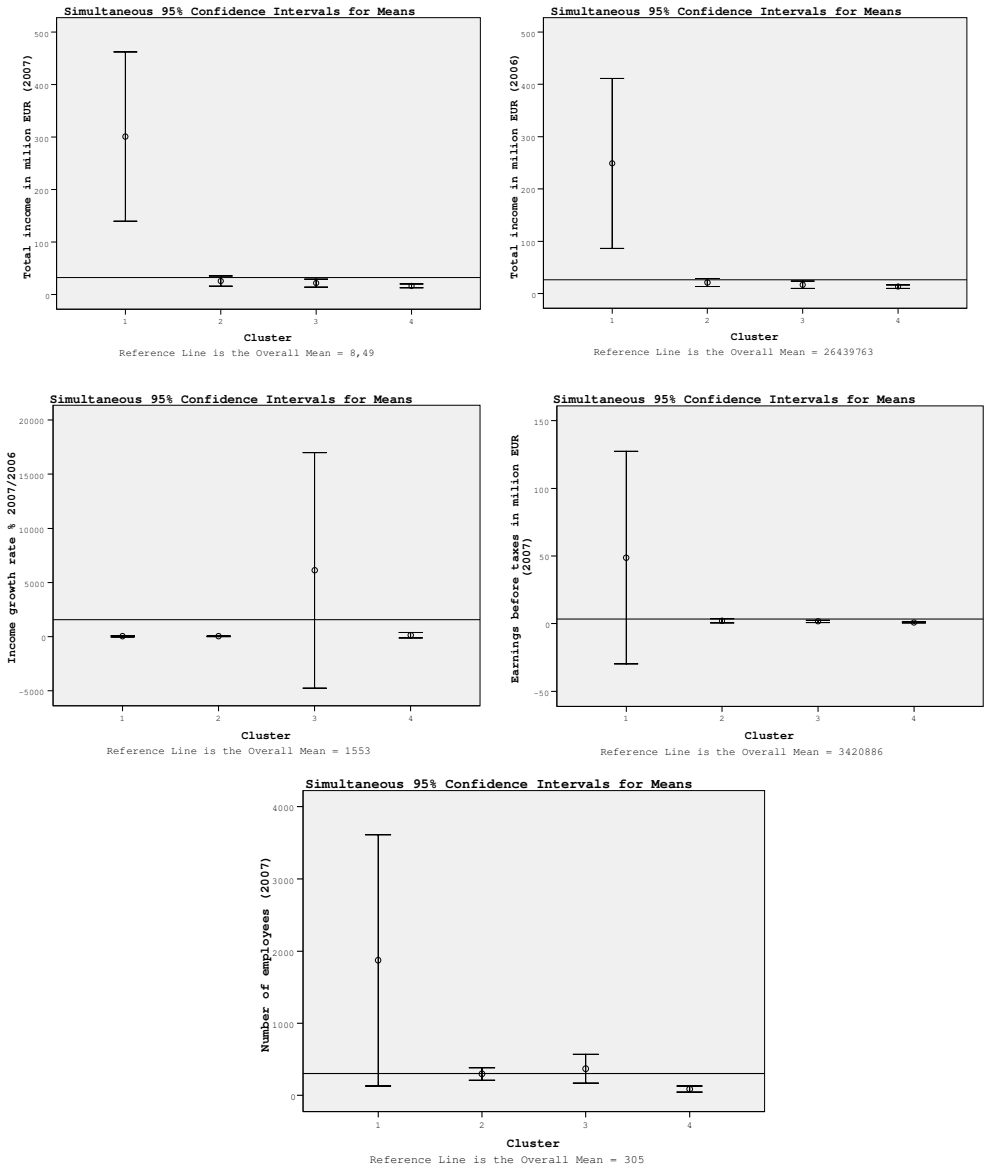
The main idea of this paper is to examine and to prove that the twostep cluster analysis is appropriate for segmentation of the Macedonian companies, so that the formed clusters are represents of the country's main company structure.

As presented in the empirical results, the four cluster solution seems to be a good represent of the main structure of the Macedonian companies. Distribution of

the observations in the clusters can be evaluated as satisfactory, since the number of companies in the clusters is relatively equal, except for the first cluster. The first cluster has small number of companies, yet these are the companies that are the base of the Macedonian company because they create a significant amount of revenues and also they provide the country with energy resources and telecommunication. Most of them were privatized from the state ownership and today they are joint stock companies partly owned by foreign investors. The second cluster is the largest cluster that includes some large companies and mostly middle sized companies. The industry of these companies is manufacturing. These are companies with steady growth, they have been present on the Macedonian market for number of years and can be qualified as firm companies. Except for the few large companies, most of the middle sized companies are owned by the domestic owners. The third cluster can be named as "the growing companies cluster" since here the income growth rate is significantly higher than the other clusters. Most of these companies are successfully companies that are more likely to grow in the years to come. Also, these companies employ significant part of the countries work force. The cluster is consisted of 43 companies. They come from different industries, yet mostly are from construction, transport, storage, telecommunications and other business activities. Investing in some of these companies may seem like a good decision, since most of these companies are likely to expand their business. The last cluster is consisted of the smallest, yet successful companies in wholesale and retail trade. Since the industry is trade, the earnings before tax are smaller than in clusters with companies from industries like manufacturing or services. Similar to the second cluster, majority of the companies are in domestic ownership, and can be classified as small and medium enterprises.

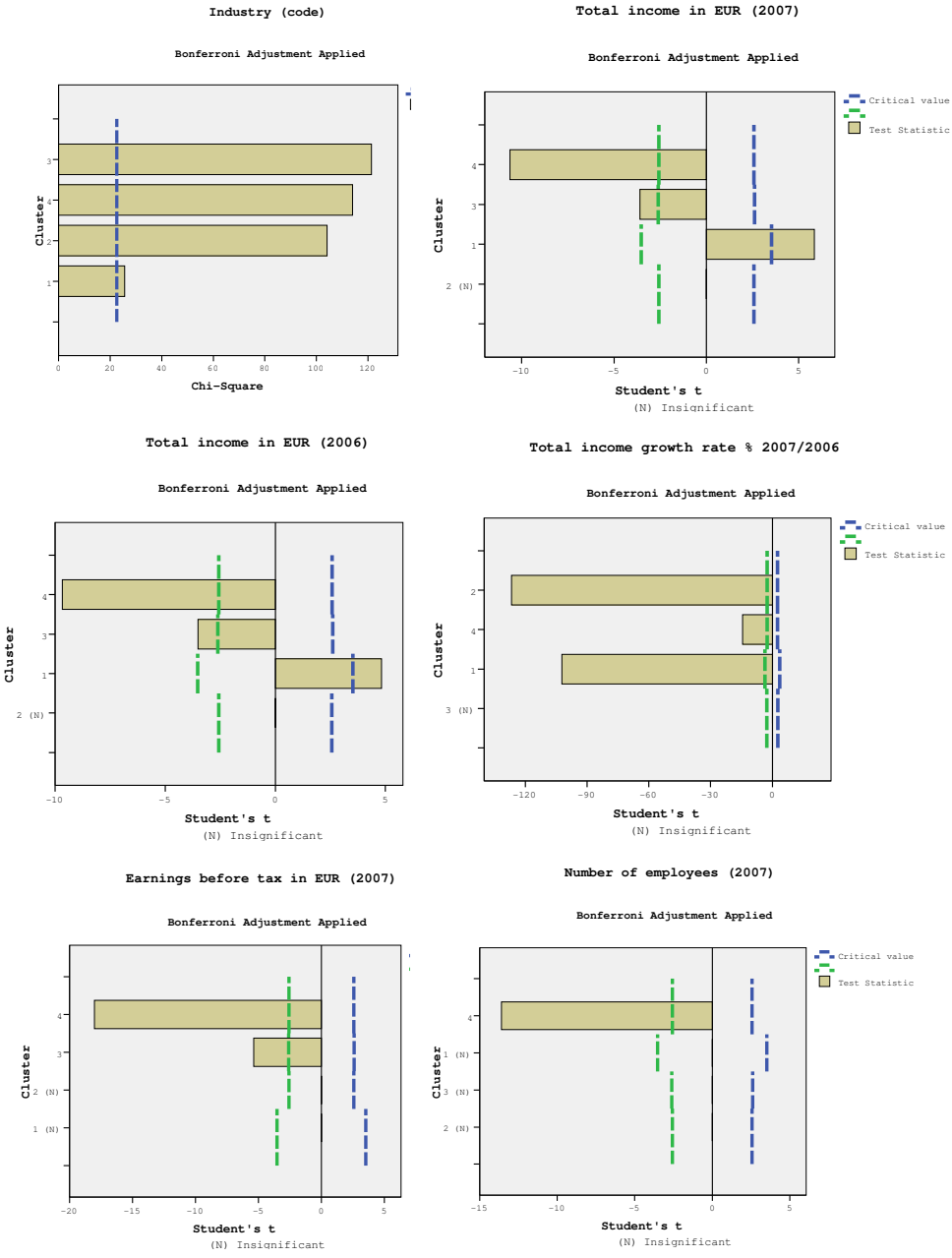
The given explanation of the empirical results proves that the twostep analysis is practical tool for segmentation of the large data base of 200 Macedonian companies. Findings of the analysis are useful because mainly they provide the general structure of the largest Macedonian companies. For the potential foreign investors this analysis is an insight to the most lucrative industries in the country. For the government, the presented results give information about which industries dominate in the most successful companies, in order to invest in their development through infrastructure, university education, tax relief and deduction of other expenditures. For further research, the cluster membership of the twostep cluster analysis can be used as dependent categorical variable in discriminant analysis.

Figure 2. Within cluster variation for all continuous variables



Source: Own creation

Figure 3. Variablewise importance



Source: Own creation

Table 5. Cluster membership of the companies

Company / Cluster		Company / Cluster	
Okta AD Skopje	1	Adrijus DOOEL Skopje	2
Feni industries AD Kavadarci	1	MZT Hepos AD Skopje	2
Makpetrol AD Skopje	1	ADE Skopsko pivo Tetovo	2
Makedonski telekom AD Skopje	1	Swisslion-Agroplod AD Resen	2
EVN Macedonia AD Skopje	1	Ekstra-Skopsko Kosel DOOEL Ohrid	2
T-Mobile Macedonia AD Skopje	1	Metro AD Skopje	2
Macedonian power plants JSC Skopje	1	Bas tuti friuti DOOEL Skopje	2
Arcelormittal Skopje (CRM) AD Skopje	2	Brako DOO Veles	2
Arcelormittal Skopje (HRM) AD Skopje	2	Fabika Karpos AD Skopje	2
Makstil AD Skopje	2	Mlekara Zdravje Radovo, Strumica	2
Usje AD Skopje	2	Metalopromet DOOEL Strumica	2
Pivara Skopje AD Skopje	2	Prototip DOOEL Skopje	2
Alkaloid AD Skopje	2	MZT Learnica AD Skopje	2
Igm-trade DOO Kavadarcki	2	Helmateks AD Strumica	2
Dojran stil DOO Dojran	2	Tondach-Makedonija AD Vinica	2
Skopski leguri DOOEL Skopje	2	Kiro Kucuk AD Veles	2
11 Oktomvri AD Kumanovo	2	Zdenka DOOEL Negotino	2
Tutunski kombinat AD Skopje	2	Fersped AD Skopje	3
Dil Petrol DOOEL Stip	2	Euro tabak DOO Skopje	3
Toplifikacija AD Skopje	2	Cosmofon AD Skopje	3
Brilijant DOOEL Stip	2	Granit AD Skopje	3
Silmak DOOEL Tetovo	2	Makedonski aviotransport AD Skopje	3
Swisslion DOO Skopje	2	Tec Negotino AD Negotino	3
Mlekara AD Bitola	2	NLB Lizing DOOEL Skopje	3
Zito luks AD Skopje	2	Knauf-radika AD Debar	3
Vinarska vizba Tikves AD Skopje	2	JP za stop. so stanben i deloven prostor Skopje	3
Bomex DOO Skopje	2	Terna AD Skopje	3
Pekabesko AD Skopje	2	Makosped AD Skopje	3
ZK Pelagonija AD Bitola	2	Makoten DOOEL Gevgelija	3
F.I. Vitaminka AD Prilep	2	EFT Makedonija DOOEL Skopje	3
EMO AD Ohrid	2	Sasa DOOEL Makedonska kamenica	3
Teteks AD Tetovo	2	Alma-m DOO Skopje	3
Mega DOOEL Skopje	2	Makedonska posta AD Skopje	3
EMO DOOLE Ohrid	2	JP Vodovod i kanalizacija Skopje	3
Prilepska pivarnica AD Prilep	2	Bucim DOOEL Radovis	3
Droga kolinska DOOEL Skopje	2	Beton AD Skopje	3
Vest DOOEl Bitola	2	Media print Makedonija DOO Skopje	3
Strumica tabak AD Strumica	2	Pexim DOOEL Skopje	3
Promes DOO Skopje	2	Mavrovoinzenerig DOOEL Skopje	3
Alayans uan Makedonija AD Kavadarci	2	Peas Macedonia Skopje	3
FI Blagoj Gorev JSC Veles	2	JP Komunalna higiena Skopje	3

MIK Sveti Nikole DOO Sveti Nikole	2	DS Iskra steel construction DOO Kumanovo	3
Imperijal-tabako AD Valandovo	2	Indo minerals&metals DOOEL Skopje	3
Kontihidroplast DOOEL Gevgelija	2	Neocom AD Skopje	3
Ideal sipka DOO Bitola	2	Makinvest DOO Skopje	3
Pucko petrol DOO Makedonski brod	2	Dzesas insaat tidzared i sanaji AS Podruznica SK	3
Komuna AD Skopje	2	Alfeks inzenering DOO Skopje	3
Riomk bomeks - refraktori AD Pehcevo	2	DGU Pelister Bitola DOO Bitola	3
TGS Tehnicki gasovi AD Skopje	2	On.net DOO Skopje	3
Evropa AD Skopje	2	Pakom kompani DOOEl Skopje	3
Leov kompani DOOEL Veles	2	Dauti komerc AD Skopje	3
Makprogres DOO Vinica	2	Mlaz AD Bogdanci	3
FHL Mermeren kombinat AD Prilep	2	Makedonija Turist AD Skopje	3
Pavor DOOEL Veles	2	Kiro D. Dandaro AD Bitola	3
Bunar petrol DOO Gostivar	3	Centro union DOO Skopje	4
Kvalitet-prom DOOEl Kumanovo	3	Gorenje DOOEL Skopje	4
Senker DOOEL Skopje	3	Podravka DOOEL Skopje	4
Pelagonija Inzinering DOOEL Skopje	3	Energomarket DOO Skopje	4
Publicis DOO Skopje	3	Tabako-promet BM DOOEL Valandovo	4
Int trejd DOOEL Kocani	3	Eurotrejd DOO Skopje	4
Lukoil Macedonia LTD Skopje	4	Montenegro DOO Gostivar	4
Tinex-mt DOOEL Skopje	4	Ka-dis DOO Skopje	4
Veroupulos DOOEL Skopje	4	Kolid kompani AS DOO, s. Kolesino Novo Selo	4
Porshe Makedonija DOOEL Skopje	4	German PX DOO Skopje	4
Gemak-trade DOOEL Skopje	4	Gamatroniks DOOEL Skopje	4
ZEGIN DOO Skopje	4	Automakedonija AD Skopje	4
Skopski Pazar AD Skopje	4	Marija treid DOO Veles	4
Euro aktiva DOO Skopje	4	Gross prom DOO Skopje	4
KAM DOOEL Skopje	4	Grosist DOOEL Bitola	4
Makautostar DOOEL Skopje	4	Agrokumanovo AD Kumanovo	4
Nelt st DOOEL Skopje	4	Swisslion Mak DOO Skopje	4
AD D-r Panovski Skopje	4	Agroefodia DOOEL Strumica	4
Pharmacy Zegin farm Skoopje	4	Kemo-farm DOOEL Skopje	4
Promedika DOO Skopje	4	Avto kuka DOO Skopje	4
Euro media DOO Skopje	4	Krka-farma DOOEL Skopje	4
Euroimpex DOO Skopje	4	Zito DOOEL Veles	4
KIK DOO Skopje	4	Mepso AD Skopje	*
Ekspanda DOOEL Skopje	4	Kameni most komunikacii AD Skopje	*
Makoil DOOEL Skopje	4	ADG Mavrovo Skopje	*
Automobile sk DOOEL Skopje	4	Zito vardar AD Veles	*
Elektroelement DOO Skopje	4	JP Makedonski sumi Skopje	*
Jaka 80 AD Radovis	4	Jaka tabak AD Radovis	*
Replek AD Skopje	4	Hypo-alpe-adria-lizing DOOEL Skopje	*
Ramstore Macedonia DOO Skopje	4	4 Noemvri AD Bitola	*

Filip Moris Skopje DOOEL Skopje	4	Tutunski kombinat - cigari DOOEL Prilep	*
Toyota avto centar DOOEL Skopje	4	Public transportation enterprise Skopje	*
Mako-market DOO Skopje	4	Haier Makedonija trejd DOOEL Skopje	*
Merkur Makedonija DOO Skopje	4	JP Makedonijapat Skopje	*
Makedonija Lek DOO Skopje	4	Germanos Telekom AD Skopje	*
Mi-da motors DOO Skopje	4	Lek DOOEL Skopje	*
Libra 1 AG Skopje	4	Lotarija na Makedonija AD Skopje	*
Euromilk DOO Skopje	4	JP Makedonska radio televizija Skopje	*
Tediko super DOOEL Skopje	4	TCG learnica DOOEL Ohrid	*
Internacional food bazar DOO Skopje	4	AD Negotino Negotino	*
Avtonova DOO Skopje	4	Prima.mk DOO Skopje	*
Rudine-mm DOO Skopje	4	Fruktal mak AD Skopje	*
JUS MB DOO Skopje	4	Fabrika za kvasec i alkohol AD Bitola	*
Makkar DOO Skopje	4	SAF Energiferzorgungslezngen GmbH podr. SK	*
Inter tobako DOOEL Skopje	4	Alfa kopi DOOEL Skopje	*
Kola DOOEL Skopje	4	MIA Beverages DOO Skopje	*

** Excluded observations due to the missing data*
Source: Own creation

References

Banfield, J. D. - Raftery, A. E. 1998: Model based Gaussian and non Gaussian clustering. *Biometrics,* 49, 803-821. p.

Chiu, T. - Fang, D.- Chen, J. - Wang, Y. - Jeris, C. 2001: A Robust and Scalable Clustering Algorithm for Mixed Type Attributes in Large Database Environment. *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining,* SanFrancisco, CA: ACM, 263–268. p.

Euro Business Centre, Central Registry of Republic of Macedonia, 2008: *200 Largest*. Euro Business Centre.

Hair, J. F. - Black, B. - Babin, B. - Anderson, R. E. 2005: *Multivariate data analysis*. Upper Saddle River, NJ , Prentice Hall.

Garson, D. 2009: *Cluster analysis from Statnotes: Topics in Multivariate analysis* retrieved from http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm.

Meila, M. - Heckerman, D. 1998: *An experimental comparison of several clustering and initialization methods.* Microsoft Research Technical Report MSR-TR-98-06.

SPSS 2007: *SPSS 16.0 Command syntax reference*. SPSS Inc., Chicago.

Zhang, T. - Ramakrishnon, R. - Livny, M. 1996: BIRCH: Method for very large databases. *Proceedings of the ACM. Management of Data,* 103–114. p. Montreal, Canada.

Different Clustering Techniques – Means for Improved Knowledge Discovery

Olivera Grljević[1] - Saša Bošnjak[2] - Zita Bošnjak[3]

Application of different clustering techniques can result in different basic data set partitions emphasizing diversified aspects of resulting clusters. Since analysts have a great responsibility for the successful interpretation of the results obtained through some of the available tools, and for giving meaning to what forms a qualitative set of clusters, additional information attained from different tools is of a great use to them.

In this article we presented the clustering results of small and medium sized enterprises' (SMEs) data, obtained in DataEngine, iData Analyzer and Weka tools for intelligent analysis.

Keywords: Data mining, clustering, DataEngine, iData Analyzer, Weka

1. Introduction

The idea of *Knowledge Discovery in Databases* (KDD) is to search for relations and global schemes that exist in large databases and are hidden in the vast amount of data. *Data mining*, as the part of KDD, is the process of using one or more computational techniques in automated search for hidden information and relationships among data. As such, it represents indivisible part of qualitative research. Knowledge discovered through different data mining methods and techniques reveal behavioral patterns, profiles of entities, and similar regularities in data. Using solely statistical methods, qualitative data model can not be built.

[1] Olivera Grljević, MSc, University of Novi Sad, Faculty of Economics Subotica, Serbia
[2] Saša Bošnjak, PhD, associate professor, University of Novi Sad, Faculty of Economics Subotica, Serbia
[3] Zita Bošnjak, PhD, full professor, University of Novi Sad, Faculty of Economics Subotica, Serbia

Acknowledgement: This paper is the result of a research on the project titled: "Comparative Advantages of Intelligent Data Analysis Methods in Strengthening the Sector of Small and Medium Sized Enterprises", No. 114-451-01092/2008-01, funded by the Ministry of Sciences and Technology Development of Province of Vojvodina, Republic of Serbia.

Besides large databases, sophisticated algorithms are needed, which are subject of knowledge discovery in databases.

As proven by now (Liao- Triantaphyllou 2008), (Harrison-Llado 2000) each clustering algorithm, sometimes even the same algorithm applied several times on the initial dataset, can result in different basic dataset partitions, putting an accent on a specific aspect of the resulting clusters. Apart from diverse outputs, clustering algorithms use different visualization techniques to represent the derived clusters, which enable better insight into their structure and grouping relationships of similar entities. Furthermore, they can denote cluster centers, typical and least typical representatives of clusters, etc.

The selection of a subset of attributes of the database for clustering data, as well as the determination of the most adequate number of clusters, is under subjective appraisal of analysts. Furthermore, they have a great responsibility to carry out the interpretation of the results gained through some of the available tools successfully, and to give meaning to what forms a qualitative set of clusters. Consequently, additional information attained from different tools that support clustering techniques is of great use in clusters shaping.

Collecting and compounding various information about defined clusters, contributes to qualitative decision making on optimal cluster number and elements that constitute them. Consequently, to obtain as qualitative results as possible, and to facilitate cluster interpretation, analysts should combine different tools in the process of data clustering.

In our paper, we described a *composite approach* that implies diversity of tools and obtained results that significantly simplify the work of analysts in knowledge discovery, helps the interpretation of results, and facilitates the derivation of detail and clear conclusions. We present the results of clustering small and medium sized enterprises' (SMEs) data in Vojvodina province using *DataEngine, iData Analyzer* and *Weka tools* for intelligent analysis. Each tool supports a different clustering algorithm.

2. Techniques used in the empirical examination

The broader goal of our research was to determine discriminators between successful and less successful enterprises, and to distinguish the profile of businesses that will succeed in their goals from those that are likely to fail. These tasks are classification and clustering tasks, respectively. (Bošnjak et al. 2009) provides more detailed presentation of these problems. In this article we presented only the results of clustering techniques utilization, since they are common to all three tools we used, and are in compliance with the goals of our research.

In Komem- Schneider (2005), Bratko et al. (1998), Jiawei-Kamber (2001) it is defined that clustering is a process of grouping feature space vectors into classes in