GEOCLUSTERING AS A MULTIVARIATE ANALYSIS TASK

DELIĆ Milica, (SCG)

ABSTRACT

Geoclustering is based on aggregate data, calculated for geographic units of the observed territory. By the techniques of multivariate statistical analysis, principal component analysis and cluster analysis, geographic units are grouped so that those with the similar profiles are gathered in one group, cluster. All geographic units are clustered in one of mutually exclusive clusters, which are completely covering the observed territory. Profiling of each cluster can be made by the means of exploratory data analysis.

Geoclustering offers wide possibilities of application, depending on the problem to be solved and on the data used in the analysis. In the paper, geoclustering process will be described in details, and some examples of its application will be given.

KEY WORDS

geoclustering, principal component analysis, cluster analysis, exploratory data analysis

INTRODUCTION

Contemporary fast development of all scientific areas, as well as a growth of application possibilities of the new achievements, brings changes in approach to almost every part of human living. Permanent innovations in means of collecting and processing huge amounts of data are very important too. Among all available data it is important, however, to find these relevant to solving the specific problem, and to solve the problem. Geoclustering (GC) is the methodology developed to solve some problems in marketing research, but it can be used in solving problems from the other fields where geographic units have to be clustered on the basis of similarity in their characteristics.

THE BASIC CONCEPTS OF GEOCLUSTERING

The starting point in geoclustering is the data, prepared for analysis in large matrices. The rows (observation) of matrix are geographic units. The columns are the values of observed characteristics for each unit. Every row of the matrix is the vector that represents geographic unit and the elements of the vector are the values of the variables observed in the analysis. The vector uniquely determines every observed geographic unit. Although this vector is unique, there are many similar to it, which are dispersed throughout the matrix. The target is to identify and cluster units with the similarity in data, who share the large number of characteristics. It is important to find and emphasize characteristics that are common for the members of the cluster and differ among the clusters.

The research encompasses three steps. Multivariate statistical methods used in the steps are:

1. ***Principal Component Analysis*** or ***Factor Analysis***,
2. ***Cluster Analysis*** and
3. ***Exploratory Data Analysis***.

The first task is to remove redundancy from the data and, if possible, to reduce the number of variables. The techniques, which are checking correlation between variables and removing redundancy, are factor analysis and the principal component analysis [1]. The techniques are different, but the effects of both of them are the same. The elementary data matrix is reduced so that the number of variables describing research units is reduced. New variables are the factors and each factor represents the whole group (highly correlated mutually) of variables.

Each unit then achieves its value for each of those factors. The new matrix still contains almost all unredundanced information contained in the previous matrix. The final result of this research phase represents the matrix of factor scores. The elements of this matrix are the values of the factors which were kept in the analysis, and which were computed for each unit, as the value of corresponding linear combination of original variables. The coefficients in linear combinations are correlation coefficients between original variables and factors kept in the analysis. The factor scores are at that standardised and mutually uncorrelated.

The second step is the use of matrix of factor scores for finding clusters of units. Cluster analysis is the technique used for solving problems of this kind [1]. The main target in geoclustering is to find geographic units with similar values for the factors. All units are sorted by cluster analysis in one of mutually exclusive clusters, which are completely covering the observed territory.

There exist different techniques and methods for clustering elements. They are basically divided on hierarchical and nonhierarchical methods [2]. In each step of the hierarchical procedure, the new element is joined to the “nearest” cluster. The dynamics of agglomeration can be graphically observed in dendrogram. From the same chart it can be seen how distant the elements and the clusters are. Different number of clusters can bee identified, depending on wonted hierarchical level.

The next step is the profiling of each cluster. The characteristics of the members can be reconstructed on the basis of the analysis of the values of factor scores and the characteristics used to determine the factor scores for the geographic units and many of them can be encompassed by exploratory data analysis [3].

The geoclustering does not claim that the members of the same segment are completely identical, but only that they will rather be alike than the units which were accidentally chosen from the whole territory [4].

GC can be used in the analysis with different goals and the data regarding different territories and geographic units. In this paper will be given two examples of GC research: geoclustering of Serbian market and grouping the countries of integrated Europe, based on objectives of macroeconomics.

GEOCLUSTERING OF SERBIAN MARKET

Market segmentation enables determination of the target market and the right way of satisfying its demands. Actually, market segmentation represents the process of the market division into different groups of buyers (users) who could demand special products (services) and/or marketing mixes. Company identifies different ways of market segmentation, it develops the profiles of the resulting market segments and it evaluates the attractiveness of each segment. After that, the target market is chosen (evaluation and the choice of one or more market segments) and the products are positioned in relation to the competition [5].

The GC was firstly devised as the support to the firms in direct marketing [4]. Accessibility, as one of the three rigorous principles of market segmentation, was mostly unattainable trough earlier used methods. GC enables connecting the data on demographic characteristics and other habits of consumers to the locations where they are situated. In such a way, it is possible to connect demographic and geographic market potentials.

GC, as the method of market segmentation, is often based on data collected and published by statistical offices. That will be the case with the example presented here.

Statistical institutions follow many points from various segments of life in Serbia, processes obtained data and periodically publish it, according to the given standards for collection and processing. The publication comprises data for municipalities, and that is the lowest territorial level on which the data are aggregative processed and obtainable [6]. In Serbia, without AP Kosovo and Metohija, there exist 160 municipalities. By relevant transformations of the stated demographic indicators, in order to facilitate further action, the analysis presented here was based on 31 attributes regarding demographic data (obtained in census), data on economic status of the citizens, employment structure etc. Out of this, follows, that the data matrix being used as the starting point in the analysis had 160 rows (observations – municipalities) and the 31 columns (variables – attributes) [5].

After the principal component analysis, eight factors were identified and they can be determined in a short way as follows:

FACTOR 1. – *Social-economic development;*
FACTOR 2. – *Age structure;*
FACTOR 3. – *The size of the municipality*
FACTOR 4. – *Active female population*
FACTOR 5. – *The structure of employed;*
FACTOR 6. – *Agriculture*
FACTOR 7. – *Active male population and*
FACTOR 8. – *Wealth*

The final result of this research phase represents the matrix of factor scores. The elements of this matrix are the values of the main components which were kept in the analysis, and which were computed for each municipality, as the value of corresponding linear combination of original variables. Accordingly, the factor scores matrix will have 160 rows and 8 columns, and will be the basis for the following analyses steps. At the same time the 78,74 % of the original data complexion was retained. The factors scores are standardised and mutually uncorrelated.

The Ward method of hierarchical clustering is applied on the new matrix, with the use of the squared Euclidean distance. On the basis of dendrogram analysis and corresponding attributes on similarities and differences between elements which are grouped, the dendrogram is cut in the way that the 21 clusters, i.e. market segments, were identified [5]. The segments are grouped, also on the basis of dendrogram, in six groups **GA**, **GB**, **GC**, **GD**, **GE** and **GF**, and the **GC** and **GE** are divided in subgroups. Accordingly, the final result is the hierarchical structure in which 21 segments are grouped in 9 subgroups, or 6 groups. Groups, subgroups, segments and number of municipalities in each of them are given in table 1. The common characteristics of the municipalities located in the same cluster can be identified. In this paper will be given only short description of market groups [5], obtained as the result of exploratory data analysis implementation.

To **GA** market group belong municipalities in which the agriculture is the basic activity. The group contains two segments. The market segment **S01** is composed of poorer and smaller agricultural municipalities, while the segment **S02** is composed of the larger towns in which the agriculture is also basic activity, but there are also those employed in industry. The market group **GB**, segment **S03**, is composed of “the youngest municipalities” in Serbia. The group **GC** comprise of the large number (53) of municipalities of Serbia. The average values of the most attributes for this group are near the average for the whole Republic. The group is comprised by three subgroups with altogether eight segments. The market group **GD** comprises large town centres of Serbia. The segment **S12** represents the large towns denser inhabited, with the population poorer than those living in towns from segment **S13**.

Group	Subgroup	Segment	Number of municipalities	Total
GA		S01	22	
		S02	14	36
GB		S03	5	5
GC	**GC1**	S04	2	
		S05	5	
		S06	8	
	GC2	S07	15	
		S08	4	
		S09	6	
		S10	12	
	GC3	S11	1	53
GD		S12	7	
		S13	9	16
GE	**GE1**	S14	6	
		S15	6	
		S16	8	
	GE2	S17	6	
		S18	7	
		S19	14	47
GF		S20	2	
		S21	1	3

Table 1.Clustering of municipalities

The common characteristic of the municipalities included in market group **GE** is that their population is predominantly employed in administration. The market group **GF**, composed by the Belgrade municipalities of Stari Grad, Savski Venac and Vracar, is extreme according to the most observed attributes. These municipalities represent the administrative centre of the Republic and contain the most of institutions of the national importance.

CLUSTERING OF THE EUROPEAN COUNTRIES

As the second example, here will be present some results of geoclustering of countries in Europe, based on data regarding objectives of macroeconomics. European countries included in the analysis are divided in five groups, regarding their status in integrative process in Europe. The groups and belonging countries are:

1. ***Euro area***: Austria, Belgium, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal and Spain.
2. ***Other members of EU15***: United Kingdom, Denmark and Sweden.
3. ***New 10 EU members***: Cyprus, Czech Republic, Estonia, Hungary, Latvia, Lithuania, Malta, Poland, Slovakia and Slovenia.
4. ***West European countries out of EU:*** Iceland, Norway and Switzerland.
5. ***South-east European countries out of EU (candidates):*** Albania, Bosnia and Herzegovina (B and H), Bulgaria, Croatia, Romania, Serbia and Montenegro (S and M), the former Yugoslav Republic of Macedonia (FYR M) and Turkey.

Supergroup	Group	Subgroup	Country
SGA	**GA1**	**GA11**	1 France
			1 Germany
			1 Italy
			1 Netherlands
			1 Portugal
			2 Denmark
			2 UK
			4 Norway
			4 Switzerland
		GA12	1 Austria
			1 Belgium
			1 Finland
			1 Luxembourg
			1 Spain
			2 Sweden
			3 Cyprus
			3 Hungary
			3 Malta
			3 Slovenia
			4 Iceland
	GA2		1 Ireland
			5 Albania
SGB	**GB1**	**GB11**	1 Greece
			3 Czech Republic
			3 Poland
			3 Slovakia
			5 Bulgaria
			5 Croatia
		GB12	3 Estonia
			3 Latvia
			5 Romania
		GB13	3 Lithuania
	GB2		5 B and H
			5 FYR Macedonia
SGC			5 S and M
SGD			5 Turkey

Table 2. Clusters

For these countries and the five years period, are analyzed following data [7]:

GDP -Gross domestic product *(Percentage change over the preceding year, 1999-2003);*
UIR -Standardized unemployment rates *(Per cent of civilian labour force, 1999-2003);*
CPR -Consumer prices (*Perc. ch. over the preceding year, 1999-2003);*
IOUT -Real gross industrial output *(Perc. ch. over the preceding year, 1999-2003).*

The goal of the analysis was to see how the countries will be clustered on the basses of achieved objectives of macroeconomics. The matrix used in the analysis had 36 rows (countries- research unites) and 20 columns (variables). In principal component analysis for the data, it is concluded that in the analysis can be continued with five factors. These five factors describe 88% of total original variability in data. The new matrix, which was used in farther analysis, had 36 rows, but only five columns. In the addiction, variables in this matrix are unredundanced and not correlated.

After cluster analysis, dendrogram is cut three times. The hierarchical structure of clusters (supergroups, groups, subgroups) and their members are given in table 2.

The fact that Serbia and Montenegro and Turkey are the only members of the supergroups **SGC** and **SGD,** respectively, means that macroeconomic flows in these countries differ from all other countries analyzed here. Members of subgroups are very similar in their achievements and, as the level of the clustering increases, the differences between the members of the same cluster (group or supergroup) are increasing too, but they are still more similar to the countries in their cluster then these from the others.

CONCLUSION

In geoclustering process, geographic units of the observed territory are clustered and common characteristics for the members of each cluster are identified. The methods used in GD process are multivariate statistical methods: factor analysis, cluster analysis and exploratory data analysis. Results of GC process can be used as the basis for further actions or research in different fields, depending on data and geoclustered territory.

LITERATURE

1. KOVAČIĆ, Z., *Multivarijaciona analiza*, Ekonomski fakultet, Beograd, 1998.
2. NORUSIS, M.J., *SPSS/PC+ v4.0 Advanced Statistics V. 2.0*, SPSS Inc, Chicago, 1990.
3. TUKEY, *Exploratory Data Analysis*, McGraw-Hill, New York, 1977.
4. CURRY, D., *The New Marketing Research System,* John Wiley & Sons, Inc, New York, 1993.
5. DELIC, M.- VUKMIROVIC, D.- RADOJICIC, Z.: Geoustering of Serbian Market; 22nd International Scientific Conference on Development of Organisational Sciences "*Management and Organisational Development*"; Collection of Papers, p. 499 - 506; Portoroz; Slovenia; 26-28.03.2003.
6. www.statserb.sr.gov.yu
7. http://www.unece.org/ead/survey.htm – Economics Survey of Europe 2005 No.1

CONTACT ADDRESS

Delic Milica, PhD, University of Belgrade, Faculty of Organizational Sciences, Jove Ilica 154, 11000 Belgrade, Serbia and Montenegro, mild@fon.bg.ac.yu

Recenzent: prof. Ing. Zlata Sojková, CSc.