Balkans JETSS (2019) 1: 88-96
DOI: https://doi.org/10.31410/Balkans.JETSS.2019.2.1.88-96

STATISTICAL ANALYSIS OF THE EUROPEAN UNION COUNTRIES ON THE BASIS OF SELECTED SOCIO-ECONOMIC AND DEMOGRAPHIC INDICATORS

Ľubica Hurbánková[1]
Dominika Krasňanská[2]

Received: November 7, 2018 / Revised: January 10, 2019/ Accepted: March 14, 2019

Abstract: *The aim of the paper is to compare the European Union countries on the basis of selected socio-economic and demographic indicators for the year 2016. The following indicators are selected for analysis: gross domestic product per capita, government gross debt as a percentage of gross domestic product, inflation rate, unemployment rate, total fertility rate, infant mortality rate and crude divorce rate.*

The contribution of the paper is a division of the countries of the European Union into several groups using cluster analysis so that the countries belonging to the same cluster are as similar as possible and the countries belonging to different clusters are the least similar, or rather the most different. The cluster analysis consists of several steps: a selection of the type of clustering process (hierarchical and non-hierarchical, the hierarchical can be agglomerated or divisive), a selection of the aggregation method (the nearest neighbour method, the furthest neighbour method, the average distance method, the centroid method, the median method, the Ward method, the typical points method, the k-means method, a method of optimum centers or medoids and fuzzy clustering, all of which can be used as the aggregation method), a selection of similarity rate (such as the Euclidean distance, the Hamming distance, the Minkow distance, the Mahalabonis distance), a specification of the number of significant clusters (based on the standard deviation of variables creating one cluster, the determination coefficient, the semi partial coefficient of determination, the distances of clusters, the cubic clustering criterion), a cluster interpretation (the description of each cluster based on the observed characteristics).

The application of individual statistical methods is implemented through the statistical programme SAS Enterprise.

Keywords: *Cluster analyses, European Union Countries, Method*

JEL Classification C40 · E24 · J01 · J13

This paper was presented at the Second International Scientific Conference on IT, Tourism, Economics, Management and Agriculture – ITEMA 2018 - November 8, 2018, Graz, Austria, www.itema-conference.com

✉ Ľubica Hurbánková
lubica.hurbankova@gmail.com

[1] University of Economics in Bratislava, Dolnozemská cesta 1, 852 35 Bratislava, Slovak Republic
[2] University of Economics in Bratislava, Dolnozemská cesta 1, 852 35 Bratislava, Slovak Republic

1. INTRODUCTION

In most cases, the statistical research focuses on the analysis of only one observed statistical character and its only characteristics in the analyzed file. In many cases, however, it is necessary to examine a statistical file from a lot of aspects and to take into consideration its multiple characteristics displayed by multiple statistical characters. In this analysis, it is necessary to use multidimensional statistical methods, including cluster analysis.

The above-mentioned cluster analysis is used in the paper to compare the countries of the European Union based on selected socio-economic and demographic indicators (gross domestic product - GDP - per capita, government gross debt as a percentage of GDP, inflation rate, unemployment rate, total fertility rate, infant mortality rate and crude divorce rate). The goal of this method is to divide the set of objects into several relatively homogeneous clusters so that the objects, in our case the EU countries belonging to different clusters, are the least similar and objects belonging to the same cluster are as similar as possible.

2. CLUSTER ANALYSIS

The cluster analysis is a basic research tool that sorts data vectors into similar groups (Wilks, 2011). It includes a wide range of procedures and methods used to solve object typology problems and their classification. The goal of cluster analysis is to divide a set of objects into several relatively homogeneous clusters so that objects belonging to different clusters are the least similar and objects belonging to the same cluster are as homogeneous as possible. We get a few clusters (relatively homogeneous subsets) from the object file. Its application does not permit to us to determine in advance which object will be in which cluster or the total number of clusters (Kubanová, 2003).

There are several names for this kind of methods in Slovakia, such as composite analysis, aggregate analysis, trice analysis, or analysis of nests, but the name cluster analysis provides the closest reflection of the ultimate goal of the method and corresponds to the English concept Cluster Analysis, which was used for the first time in 1939 by R. C. Tryon to describe a method of dividing a set of objects into several mutually exclusive subsets. The cluster analysis was developed independently of statistics in such sectors as education, biology and psychology. Due to the lack of exchange of information between science departments, the same methods have often been discovered several times. The same techniques, discovered as duplicates, have different names. Statisticians started to be involved in cluster analysis only around 40 years ago, which resulted in the development of a cluster analysis as a non-theoretical branch using ad hoc methods for a long time (Stankovičová, Vojtková, 2007).

The cluster analysis consists of several steps that need to be followed:

- Selection of the type of clustering process - we recognize hierarchical and non-hierarchical clustering processes. Hierarchical can be depicted easily using a hierarchical tree - dendrogram, which shows the exact sequence of decomposition at the individual clustering levels. Hierarchical procedures may be agglomerate or divisive (for details see in (Kubanová, 2003).
- Selection of the aggregation method - the nearest neighbour method, the furthest neighbour method, the average distance method, the centroid method, the median method, the Ward method, the typical points method, the k-means method, a method of optimum centers or medoids and fuzzy clustering (a description of these methods is given in (Kubanová, 2003), (Stankovičová, Vojtková, 2007), Chajdiak, Komorník, Komorníková, 1999), (Meloun, Militký, 2004) can all be used as the aggregation method.

- Selection of similarity (or non-similarity) rate – the rates to which the similarity is found out are divided into four groups - distance measures (such as the Euclidean distance, the Hamming distance, the Minkow distance, the Mahalabonis distance), association coefficient, correlation coefficient, probability similarity rates (for details see in (Stankovičová, Vojtková, 2007)).
- Specification of the number of significant clusters - there are two basic approaches to the specification of the number of clusters - heuristic procedures and formal tests. The principle of a heuristic approach is to determine the number of clusters based on the subjective opinion of the investigator. In the second approach we use formal tests, or more precisely the quality indicators of clustering (the standard deviation of variables creating one cluster, the determination coefficient, the semi partial coefficient of determination, the distance of clusters, the graph of the number of clusters and cubic clustering criterion - CCC) (see (Stankovičová, Vojtková, 2007) for a description of the cluster quality indicators).
- Cluster interpretation - when formulating conclusions about the results and the quality of clustering, it is always necessary to take into consideration the factual aspect of the problem and to thoroughly assess whether the results of cluster analysis have a practical meaning, and whether they are interpretable in and acceptable for practice (Stankovičová, Vojtková, 2007). Interpretation of clusters means making a description of each cluster based on the observed characteristics (Řezánková, Húsek, Snášel, 2009).

3. INPUT DATA

We have selected 28 member countries of the European Union for the analysis. We will make a comparison of the selected countries with the use of 7 socio-economic and demographic indicators for the year 2016. The Eurostat website will serve as a source of the data. We will also briefly define the selected indicators:

Gross domestic product per capita – the ratio of gross domestic product and average population in the year. Gross domestic product is an indicator for a nation´s economic situation. It reflects the total value of all goods and services produced less the value of goods and services used for intermediate consumption in their production. Calculations on a per head basis allows for the comparison of economies significantly different in absolute size.[3]

General gross debt as a percentage of gross domestic product – represents the total general debt as a share of GDP in percentage. It is made up of government commitments and is generated by a deficit financing of the state budget (Gola, 2009).

Inflation rate is defined as the devaluation of the monetary unit, which is manifested by the persistent growth in the price level of products and services in the economy (Šenkýřová, 2010).

Unemployment rate **represents unemployed persons as a percentage of the labour force. The labour force is the total number of people employed and unemployed. The indicator is based on the EU Labour Force Survey.**[4]

Total fertility rate – the mean number of children that would be born alive to a woman during her lifetime if she were to survive and pass through her childbearing years conforming to the fertility rates by age of a given year (Jurčová, 2002).

[3] https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tec00001&plugin=1

[4] https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&plugin=1&pcode=tepsr_wc170&language=en

Infant mortality rate – the ratio of the number of deaths of children under one year of age during the year to the number of live births in that year. The value is expressed per 1 000 live births.[5]

Crude divorce rate **is the ratio of the number of divorces during the year to the average population in that year. The value is expressed per 1 000 persons.**[6]

4. APPLICATION OF CLUSTER ANALYSIS

Since the analyzed indicators are expressed in different units, we need to transform them by standardization. In this cluster analysis we used the Ward's method, which was supposed to help us create stable clusters of approximately the same size.

The number of significant clusters was determined based on the semi partial coefficient of determination by which we tried to achieve a minimum value. The decrease of this characteristic occurs already in the 3rd stage of clustering, but its value is not sufficiently low. In the 4th stage of clustering, the value of the semi-partial coefficient of determination is 0.0722, which is considered sufficiently low, since there is only a minimal decrease in this characteristic (figure 1) on other levels. Based on the results of the Ward cluster method, we divided the selected countries into four clusters.

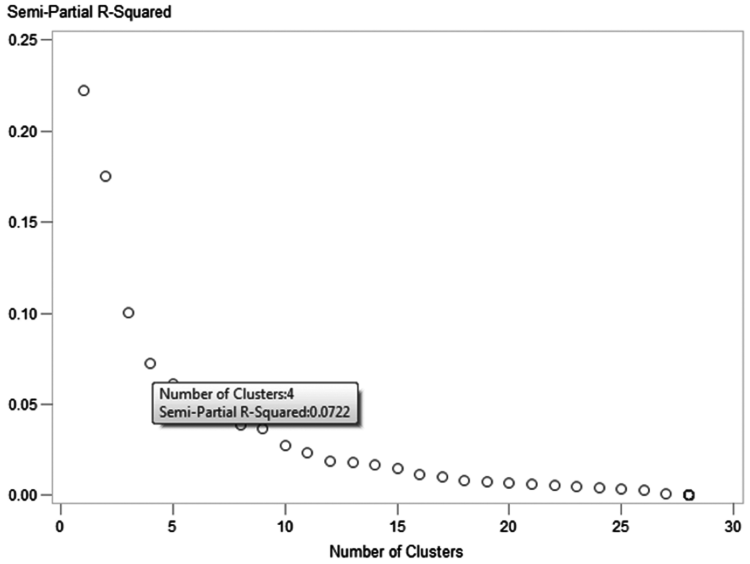

Figure 1: Trend line of the semi partial coefficient of determination

A graphical representation of the clustering of selected countries on individual levels is depicted using the hierarchical tree - dendrogram in figure 2. The *y* axis shows selected countries.

[5] http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_minfind&lang=en

[6] https://ec.europa.eu/eurostat/tgm/refreshTableAction.do?tab=table&plugin=1&pcode=tps00206&language=en

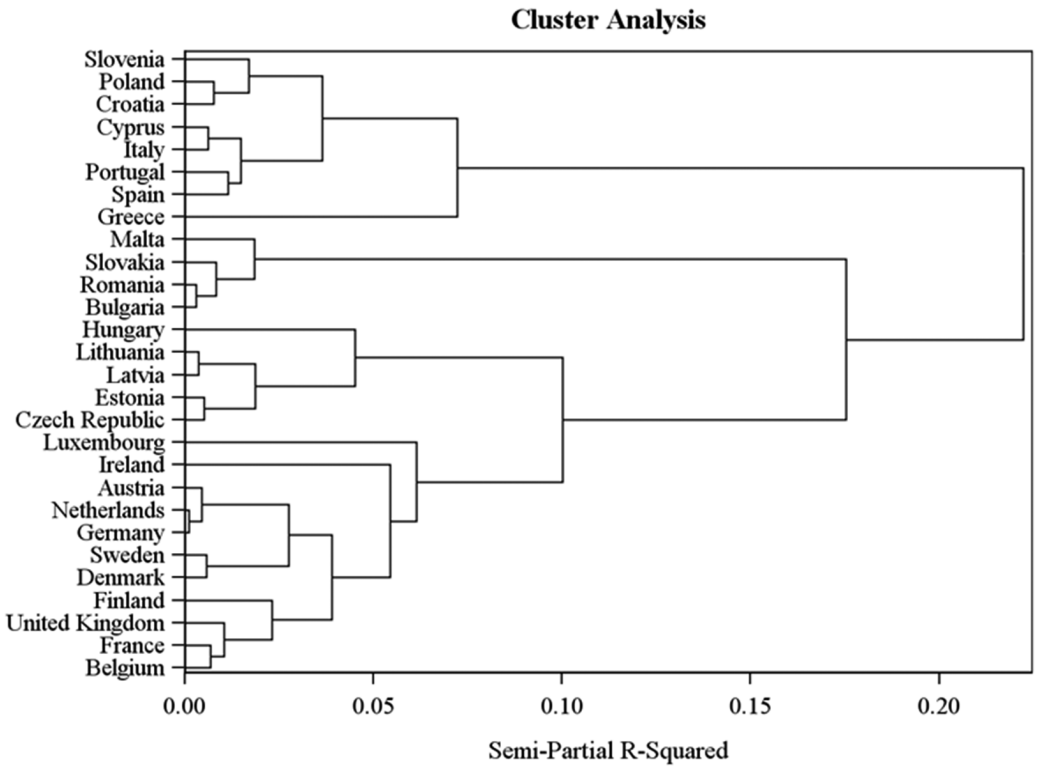

Figure 2: Dendrogram of clustering of selected countries

In the following tables (Table 1 - Table 4), we will integrate individual countries into particular clusters that we have obtained from the results of the Ward method. Each cluster contains a list of countries that belong to it together with the indicators which served for the comparison of selected European Union countries. For easier interpretation, we have replaced the standardized data with the original data. However, we have worked with standardized data at the time of implementation of the procedure.

The last step of the cluster analysis is cluster interpretation. For better interpretation of the obtained results, we used cluster centroids which represent the average level of selected indicators in the given cluster (table 5).

The use of cluster analysis, which served for the division of a set of objects into several relatively homogeneous clusters, allowed us to obtain four clusters.

The first cluster includes 11 countries: Belgium, Denmark, Germany, France, Luxembourg, Netherlands, Austria, Finland, Sweden, Ireland and the United Kingdom. This cluster is characterized by the highest average GDP, the second highest average government gross debt and the highest average fertility rate.

The second cluster consists of the following countries: Bulgaria, Malta, Romania and Slovakia. These countries reach the highest average infant mortality rate and the lowest average crude divorce rate.

In **the third cluster**, there are the Czech Republic, Estonia, Latvia, Lithuania and Hungary. These countries have the lowest average government gross debt and the lowest average unemployment rate. When comparing the average GDP, these countries are in the third place. This cluster includes countries that reach the highest values of inflation rate and crude divorce rate.

The fourth cluster is made up of Greece, Spain, Croatia, Italy, Cyprus, Poland, Portugal and Slovenia. This cluster is characterized by the highest average government gross debt and the highest average unemployment rate. However, the average inflation rate is the lowest among all clusters.

Table 1: Division of the EU countries to the first cluster

CLUSTER=1							
COUNTRY	**GDP per capita (EUR)**	**General gross debt (% of GDP)**	**Inflation rate (%)**	**Unemployment rate (%)**	**Total fertility rate (‰)**	**Infant mortality rate (‰)**	**Crude divorce rate (‰)**
BELGIUM	37 400,00	105,90	1,10	7,80	1,68	3,20	2,10
DENMARK	48 400,00	37,90	4,70	6,20	1,79	3,10	3,00
GERMANY	38 400,00	68,20	5,30	4,10	1,60	3,40	2,00
FRANCE	33 300,00	96,60	1,10	10,10	1,92	3,70	1,90
LUXEMBOURG	90 700,00	20,80	5,90	6,30	1,41	3,80	2,10
NETHERLANDS	41 600,00	61,80	4,40	6,00	1,66	3,50	2,00
AUSTRIA	40 800,00	83,60	7,00	6,00	1,53	3,10	1,80
FINLAND	39 300,00	63,00	-0,30	8,80	1,57	1,90	2,50
SWEDEN	46 600,00	42,10	7,60	6,90	1,85	2,50	2,40
IRELAND	57 500,00	72,80	6,60	16,80	1,81	3,00	0,70
UK	36 600,00	88,20	5,40	13,00	1,79	3,80	1,80

Table 2: Division of the EU countries to the second cluster

CLUSTER=2							
COUNTRY	**GDP per capita (EUR)**	**General gross debt (% of GDP)**	**Inflation rate (%)**	**Unemployment rate (%)**	**Total fertility rate (‰)**	**Infant mortality rate (‰)**	**Crude divorce rate (‰)**
BULGARIA	6 800,00	29,00	7,10	7,60	1,54	6,50	1,50
MALTA	22 300,00	56,20	4,80	4,70	1,37	7,40	0,80
ROMANIA	8 700,00	37,40	5,00	5,90	1,64	7,00	1,50
SLOVAKIA	15 000,00	51,80	7,00	9,70	1,48	5,40	1,70

Table 3: Division of the EU countries to the third cluster

CLUSTER=3							
COUNTRY	**GDP per capita (EUR)**	**General gross debt (% of GDP)**	**Inflation rate (%)**	**Unemployment rate (%)**	**Total fertility rate (‰)**	**Infant mortality rate (‰)**	**Crude divorce rate (‰)**
CZECH REPUBLIC	16 700,00	36,80	6,70	4,00	1,63	2,80	2,40
ESTONIA	16 500,00	9,40	3,80	6,80	1,60	2,30	2,50
LITHUANIA	12 800,00	40,50	7,30	9,60	1,74	3,70	3,10
LATVIA	13 500,00	40,10	4,50	7,90	1,69	4,50	3,10
HUNGARY	11 600,00	76,00	13,60	5,10	1,53	3,90	2,00

Table 4: Division of the EU countries to the fourth cluster

CLUSTER=4							
COUNTRY	**GDP per capita (EUR)**	**General gross debt (% of GDP)**	**Inflation rate (%)**	**Unemployment rate (%)**	**Total fertility rate (/‰)**	**Infant mortality rate (‰)**	**Crude divorce rate (‰)**
GREECE	16 200,00	180,80	-1,50	23,60	1,38	4,20	1,00
SPAIN	24 100,00	99,00	4,60	19,60	1,34	2,70	2,10
CROATIA	11 200,00	80,60	2,10	13,40	1,42	4,30	1,70
ITALY	27 900,00	132,00	-0,20	11,70	1,34	2,80	1,60
CYPRUS	21 700,00	106,60	1,70	13,00	1,37	2,60	2,30
POLAND	11 100,00	54,20	2,30	6,20	1,39	4,00	1,70
PORTUGAL	18 100,00	129,90	6,10	11,20	1,36	3,20	2,20
SLOVENIA	19 500,00	78,60	3,80	8,00	1,58	2,00	1,20

Table 5: Cluster centroids of selected indicators

CLUSTER	VARIABLE	MEAN
1	GDP per capita (EUR)	46 418,180
	General gross debt (% of GDP)	67,355
	Inflation rate (%)	4,436
	Unemployment rate (%)	8,364
	Total fertility rate (‰)	1,692
	Infant mortality rate (‰)	3,182
	Crude divorce rate (‰)	2,027

2	GDP per capita (EUR)	13 200,000
	General gross debt (% of GDP)	43,600
	Inflation rate (%)	5,975
	Unemployment rate (%)	6,975
	Total fertility rate (‰)	1,508
	Infant mortality rate (‰)	6,575
	Crude divorce rate (‰)	1,375
3	GDP per capita (EUR)	14 220,000
	General gross debt (% of GDP)	40,560
	Inflation rate (%)	7,180
	Unemployment rate (%)	6,680
	Total fertility rate (‰)	1,638
	Infant mortality rate (‰)	3,440
	Crude divorce rate (‰)	2,620
4	GDP per capita (EUR)	18 725,000
	General gross debt (% of GDP)	107,713
	Inflation rate (%)	2,363
	Unemployment rate (%)	13,338
	Total fertility rate (‰)	1,398
	Infant mortality rate (‰)	3,225
	Crude divorce rate (‰)	1,725

5. CONCLUSION

We compared the EU countries on the ground of selected socio-economic and demographic indicators (GDP per capita, government gross debt as a percentage of GDP, inflation rate, unemployment rate, total fertility rate, infant mortality rate and crude divorce rate). The analysis was based on data from the year 2016. The Eurostat website served as a source of the data. The indicators were selected subjectively, but we tried to select indicators which influence the actual functioning of the country economy. With the use of cluster analysis, we grouped the EU countries on the basis of selected socio-economic and demographic indicators into four clusters. Countries that form one cluster are similar according to selected indicators. The cluster analysis was made using SAS Enterprise Guide statistical software.

Belgium, Denmark, Germany, France, Luxembourg, Netherlands, Austria, Finland, Sweden, Ireland and United Kingdom create the first cluster which has the highest average GDP and the highest average fertility rate. One of the factors that may have affected the inclusion of these countries into a common cluster may be the location of the countries, as they are located relatively close to each other.

The second cluster consists of countries like Bulgaria, Malta, Romania and Slovakia. This cluster reaches the highest average infant mortality rate and the lowest average crude divorce rate. The common factor that brought these countries together could be the year of joining the EU, since Malta and Slovakia joined the EU in 2004 and Romania and Bulgaria three years later.

The Czech Republic, Estonia, Lithuania, Latvia and Hungary are the four countries forming the third cluster. This cluster is characterized by the lowest average government gross debt and the lowest average unemployment rate. On the other hand, this cluster reaches the highest average inflation rate and crude divorce rate.

The last, fourth cluster consists of Greece, Spain, Croatia, Italy, Cyprus, Poland, Portugal and Slovenia. This cluster is characterized by the highest average government gross debt and the highest average unemployment rate. By contrast, the average inflation rate is the lowest.

REFERENCES

Gola, P. (2009, April 27). *Veřejné dluhy ve světě – Česko si zatím stojí dobře.* Retrieved October 10, 2018 from http://www.finance.cz/zpravy/finance/217869-verejne-dluhy-ve-svete-cesko-si-zatim-stoji-dobre.

http://appsso.eurostat.ec.europa.eu/nui/show.do?dataset=demo_minfind&lang=en.

https://ec.europa.eu/eurostat/tgm/refreshTableAction.do?tab=table&plugin=1&pcode=tps00206&language=en.

https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&language=en&pcode=tec00001&plugin=1.

https://ec.europa.eu/eurostat/tgm/table.do?tab=table&init=1&plugin=1&pcode=tepsr_ wc170&language=en.

Chajdiak, J., Komorník, J. & Komorníková, M. (1999). *Štatistické metódy*. Bratislava: STATIS, 275.

Jurčová, D. (2002). *Krátky slovník základných demografický pojmov*. Bratislava: Výskumné demografické centrum, 37.

Kubanová, J. (2003). *Statistické metody pro ekonomickou a technickou praxi*. Bratislava: Statistic, 246.

Meloun, M. & Militký, J. (2004). *Statistická analýza experimentálníc dat*. Praha: ACADEMIA, 953.

Řezánková, H., Húsek, D. & Snášel, V. (2009). *Shluková analýza dat (druhé rozšírené vydání)*. Praha: Professional Publishing, 220.

Stankovičová, I. & Vojtková, M. (2007). *Viacrozmerné štatistické metódy s aplikáciami*. Bratislava: Iura Edition, 261.

Šenkýřová, L. (2010, September 20). *Čo znamená často skloňovaný pojem inflácia?* Retrieved October 10, 2018 from http://www.finance.sk/spravy/finance/35273-co-znamena-casto-sklonovany-pojem-inflacia/.

Wilks, D. (2011). *Statistical Methods in the Atmospheric Sciences*. Academic Press. 603-616.