International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Clustering of Member and Candidate Countries of the European Union

Hasan Bulut[a]*, Yüksel Öner[b], Çağlar Sözen[c]

[a,b]*Department of Statistics, University of Ondokuz Mayıs, Samsun, Turkey*

[c]*Department of Banking and Finance, University of Giresun, Giresun, Turkey*

[a]*Email: hasan.bulut@omu.edu.tr*

[b]*Email: yoner@omu.edu.tr*

[c]*Email: caglar.sozen@giresun.edu.tr*

Abstract

The clustering analysis aims to classify multivariate observations. For this, it uses any similarity or difference measures. In literature, clustering analysis is used to classify countries in many studies. In this study, we aim to classify the EU Member and Candidate Countries by cluster analysis in terms of some economic variables and to reveal the similarities of candidate and member countries. We have used Ward Algorithm which is a hierarchical cluster method and k-means Algorithm that is a non-hierarchical cluster method. Moreover, we have used clustering validation indexes for comparison of clustering results. To this aim, Dunn, Connectivity and Silhouette indexes are preferred as clustering validation indexes.

Keywords: European Union; K Means; Ward; Cluster Algorithm; Cluster Validation Indexes.

1. Introduction

The European Union (EU) currently has 28 member countries [1], which are generally located in Europe. The basis of the EU is based on granting new duties and powers to the European Economic Community with the Maastricht Treaty. The EU has about 500 million people and 30% of the world's gross domestic product.

---

* Corresponding author.

Furthermore, based on the Schengen Agreement between Member States, there are privileges, freedom of travel etc. The Euro is used in 19 EU member countries as common currency. What are aimed at this is to become a common and single market among the member countries.

Turkey's EU process begins with the signing of a partnership agreement between Turkey and European Economic Community in 1963. With the acceptance of Turkey in full membership negotiations in 2005, the last stage has been reached. This study aims to clustering the member states of the EU and the candidate countries in terms of economic indicators with clustering analysis which is one of the multivariate statistical methods. The main reason why the study area is limited to economic indicators is that the basis of the EU is an economic community. There are many studies in the literature on the cluster of countries. These studies which are similar to our study are Turanlı and his colleagues [2], Ersöz [3], Akın and Ören [4], Aykın and Korkmaz [5], Tekin [6], Atal [7], Turan and his colleagues [8].

2. Methods

Clustering analysis is generally reviewed under two headings. They are hierarchical and non-hierarchical algorithms. The main difference between these approaches is that the number of clusters should be predetermined in non-hierarchical algorithms.

2.1. Hierarchical Clustering- Ward Method

The general structure of the hierarchical algorithms is given below.

i. Firstly, all individuals are taken as a cluster and the distance matrix D is calculated.
ii. The nearest two clusters, which are the smallest d_{AB} value, are merged.
iii. Distance matrix is updated by reducing one the number of cluster.
iv. Steps (ii) and (iii) are repeated $(n-1)$ times and the process ends when all observations are collected in one cluster [9].

What is important here is how the distances between the observations are calculated and how the distances between the clusters and observations or other clusters are calculated. For this reason, there are many approaches. The some of these approaches are single linkage, complete linkage, average linkage, Ward. In this study, only the Ward method has been used.

The Ward method aims to minimize the squared distances within the cluster and to maximize the squared distances between the clusters. Firstly, the squared distances within the cluster for A and B clusters are defined as:

$$SS_A = \sum_{i=1}^{n_A} (x_i - \overline{x}_A)'(x_i - \overline{x}_A) \qquad (1)$$

$$SS_B = \sum_{i=1}^{n_B} (x_i - \overline{x}_B)'(x_i - \overline{x}_B) \quad (2)$$

When A and B clusters are merged, the new AB cluster is obtained. And, the squared distances within the cluster for AB cluster is defined as

$$SS_{AB} = \sum_{i=1}^{n_{AB}} (x_i - \overline{x}_{AB})'(x_i - \overline{x}_{AB}) \quad (3)$$

where n_A,n_B and n_{AB} are the number of observations in A, B and AB clusters, respectively. Moreover $\overline{x}_{AB}$ is calculated as:

$$\overline{x}_{AB} = \frac{n_A\overline{x}_A + n_B\overline{x}_B}{n_A + n_B} \quad (4)$$

In Ward method, when A and B clusters are merged, it is wanted that the increasing in sum of square (SS_{AB}) is low. To measure this increasing, I_{AB} is defined as:

$$I_{AB} = SS_{AB} - (SS_A + SS_B) \quad (5)$$

In all situations, the clusters which have minimum I_{AB} value are merged [10].

2.2. Non-Hierarchical Clustering- K Means Method

The most used non-hierarchical clustering method is k means algorithm. In this algorithm, the number of cluster is predetermined previously. Algorithm steps are below:

i. k cluster seeds are determined.

ii. Observations are assigned to cluster which is the nearest seed itself.

iii. Cluster seeds are updated by calculating mean vector of observations in cluster. If there is a more near seed than seed of itself cluster for any observation, the observation is transferred to the nearby cluster.

iv. Until all transferring between clusters are finished, Step (iii) is replied [11].

The aim of method is to find result homogenous within clusters and heterogeneity between clusters.

3. Cluster Validation Indexes

Because there are many clustering alternative and the results depend on cluster numbers, criteria have needed for the validation of clustering results.

For this purpose, there are clustering validation indexes. The some of these indexes are introduced below.

3.1. Connectivity Index

Connectivity index provides an approach that the index value is increased when the closest observations are in another cluster. $\boldsymbol{nn_{i(j)}}$ is defined as the $\boldsymbol{j_{th}}$ nearest neighbour to $\boldsymbol{i_{th}}$ observation. Connectivity index is calculated as:

$$\boldsymbol{Conn(C) = \sum_{i=1}^{n}\sum_{j=1}^{L} x_{i,nn_{i(j)}}} \quad (\mathbf{6})$$

where L is a parameter value that decides how many neighborhoods to look at and for all observations $\boldsymbol{x_{i,nn_{i(j)}}}$ is calculated as below

$$\boldsymbol{x_{i,nn_{i(j)}} = \begin{cases} 0 & , \quad i \text{ and } nn_{i(j)} \text{ are in same cluster} \\ 1/j & , \quad i \text{ and } nn_{i(j)} \text{ are in different cluster} \end{cases}} \quad (\mathbf{7})$$

Connectivity index takes values from zero to infinity and it is wanted that it is minimum [12].

3.2. Silhouette Width Index

Silhouette width index is mean of silhouette values of each observation. The silhouette value of $\boldsymbol{i_{th}}$ observation is calculated as

$$\boldsymbol{S(i) = \frac{b_i - a_i}{\max(a_i, b_i)}} \quad (\mathbf{8})$$

where $\boldsymbol{a_i}$ is means of distances between $\boldsymbol{i_{th}}$ observation and other observations in same cluster with it. Then, $\boldsymbol{d(x_i, C_j)(j = 1, 2, \dots, k)}$ distances are calculated for all cluster and the smallest of these distances takes as $\boldsymbol{b_i}$.

Silhouette width index takes values from -1 to 1 and it is wanted that it is maximum [13].

3.3. Dunn Index

Dunn index is calculated by dividing the smallest of distances that is between each observation with observations in other clusters by the maximum distance between observations in same cluster. Dunn index takes values from zero to infinity and it wanted that it is maximum [12].

4. Application

In this study, the aim is to cluster 28 members of the European Union and 6 candidate countries according to their similarities in terms of some economic indicators. For this purpose, the data set is provided from official

web site of World Bank [14]. The variables used in study are given in Table 1.

Table 1: Economic indicators used study

Code	Indicator Name
X1	Age dependency ratio (% of working-age population)
X2	Alternative and nuclear energy (% of total energy use)
X3	Consumer price index (2010 = 100)
X4	Energy imports, net (% of energy use)
X5	Foreign direct investment, net inflows (% of GDP)
X6	Foreign direct investment, net (BoP, current US$)
X7	GDP (current US$)
X8	GDP growth (annual %)
X9	GDP per capita (current US$)
x10	GNI (current US$)
x11	GNI per capita (current LCU)
x12	Inflation, consumer prices (annual %)
x13	Population ages 15-64 (% of total)
x14	Population density (people per sq. km of land area)
x15	Rural population (% of total population)
x16	Unemployment, total (% of total labor force) (modeled ILO estimate)
X17	Labor force participation rate, total (% of total population ages 15-64) (modeled ILO estimate)

Because there are different algorithms in cluster analysis, cluster validation index values have be examined to decide which method will use.

In study, Ward method as hierarchical algorithm and k-means method as non-hierarchical algorithm are used. Cluster validation index values have be given for different clustering methods and the number of cluster in Table 2.

According to Table 2, all indexes propose Ward method. Moreover, While Connectivity and Silhouette width indexes determine the number of cluster as three, Dunn index determine as four. According to a traditional method using to determine the number of cluster, number of cluster is almost equal $k \cong \sqrt{n/2}$ [15].

According to this approach, number of cluster can be take $k \cong \sqrt{34/2} = 4.12 \cong 4$. Consequently, clustering results have be investigated by taking as number of cluster $k = 3,4,5$. Dendogram plot which be obtained from Ward method has be given in Figure 1. Optimum clustering results are given in Table 3.

Table 2: Determination of clustering method and number of clusters with cluster validity indexes

		Cluster Validation Index		
Method	**k**	**Connectivity**	**Dunn**	**Silhouette**
Ward	**3**	***7.55****	0.56	**0.80***
	4	10.45	**0.57***	0.66
	5	12.95	**0.57***	0.66
	6	19.47	0.14	0.63
	7	23.01	0.28	0.63
k-means	**3**	9.29	0.34	0.77
	4	14.35	0.10	0.67
	5	16.76	0.12	0.64
	6	19.26	0.19	0.64
	7	23.01	0.28	0.63

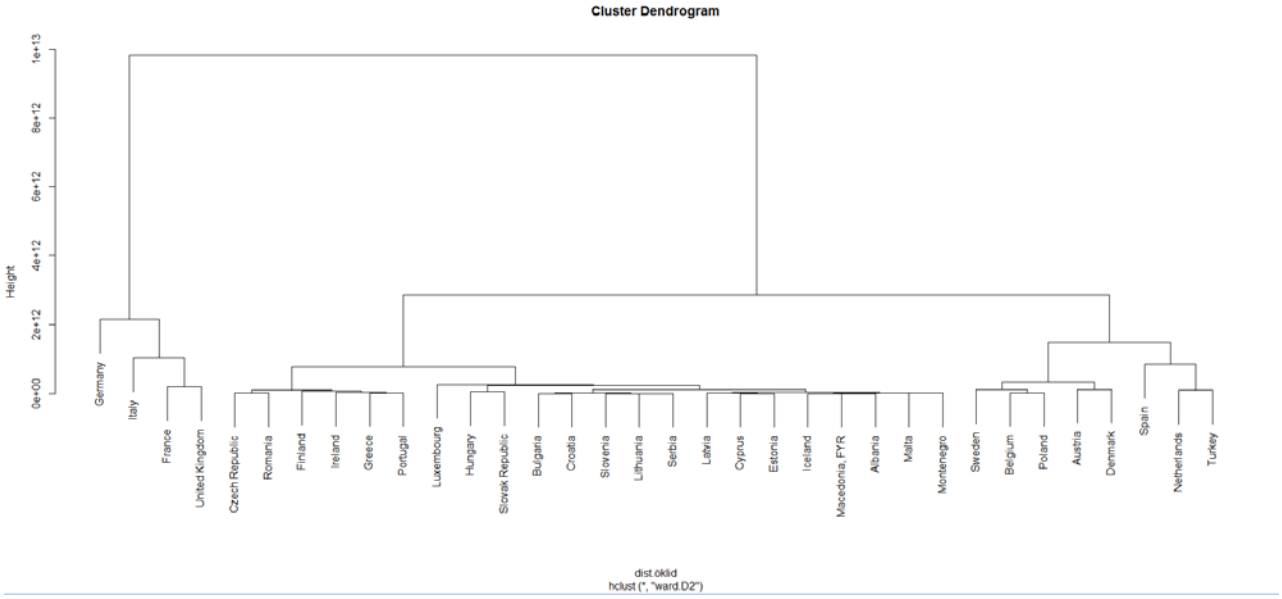

Figure 1: Dendogram Graph related to clustering of countries with Ward method

According to the cluster results given in Table 3, in all cases ($k = 3,4,5$), Germany, Italy, France and the United Kingdom are in a different position with respect to economic indicators from other countries and they have merged as a cluster. When number of cluster is three, the second cluster consists of Sweden, Belgium, Poland, Austria, Denmark, Spain, Netherlands and Turkey. The rest countries have also composed third cluster.

When number of cluster is four, K_{32} cluster is divided to K_{42} (Spain, Netherlands and Turkey) and K_{43} (Sweden, Belgium, Poland, Austria, Denmark) clusters. There has not been any change in the cluster consisted of the other

countries.

Table 3: The clustering of countries according to different number of cluster

Number of Cluster	Cluster Code	Countries
3	K_{31}	Germany, Italy, France, United Kingdom
	K_{32}	Sweden, Belgium, Poland, Austria, Denmark, Spain, Netherlands, ***Turkey****
	K_{33}	Czech Republic, Romania, Finland, Ireland, Greece, Portugal, Luxembourg, Hungary, Slovak Republic, Bulgaria, Croatia, Slovenia, Lithuania, ***Serbia****, Latvia, Cyprus, Estonia, ***Iceland****, ***Macedonia****, ***Albania****, Malta, ***Montenegro****
4	K_{41}	Germany, Italy, France, United Kingdom
	K_{42}	Spain, Netherlands, ***Turkey****
	K_{43}	Sweden, Belgium, Poland, Austria, Denmark
	K_{44}	Czech Republic, Romania, Finland, Ireland, Greece, Portugal, Luxembourg, Hungary, Slovak Republic, Bulgaria, Croatia, Slovenia, Lithuania, ***Serbia****, Latvia, Cyprus, Estonia, ***Iceland****, ***Macedonia****, ***Albania****, Malta, ***Montenegro****
5	K_{51}	Germany, Italy, France, United Kingdom
	K_{52}	Spain, Netherlands, ***Turkey****
	K_{53}	Sweden, Belgium, Poland, Austria, Denmark
	K_{54}	Luxembourg, Hungary, Slovak Republic, Bulgaria, Croatia, Slovenia, Lithuania, ***Serbia****, Latvia, Cyprus, Estonia, ***Iceland****, ***Macedonia****, ***Albania****,Malta, M***ontenegro****
	K_{55}	Czech Republic, Romania, Finland, Ireland, Greece, Portugal
*The candidate countries of EU		

When number of cluster is five, there has not been any change in the first three cluster. In this case, K_{44} cluster is divided to K_{54} (Luxembourg, Hungary, Slovak Republic, Bulgaria, Croatia, Slovenia, Lithuania, Serbia, Latvia, Cyprus, Estonia, Iceland, Macedonia, Albania, Malta, Montenegro) and K_{55} (Czech Republic, Romania, Finland, Ireland, Greece, Portugal) clusters.

5. Conclusions and Recommendations

As mentioned in this study, the basis of the EU is the European Economic Community. In order to enter the Eurozone, the common currency of the EU, countries must meet certain conditions. For this reason, the candidate countries of the EU must be examined economically. According to analyze results, Turkey has shown that it is economically different from all other candidate countries. When cluster neighbors of Turkey are considered, Belgium and the Netherlands are especially noteworthy because these countries are founder member of European Economic Community. Moreover, Brussels which is the capital of EU is in Belgium. Consequently, Turkey is economically most similar country to the main building blocks of the European Union among the candidate countries, and this situation should consider in full membership meetings to EU.

References

[1] https://europa.eu/european-union/index_en

[2] Turanlı, M., Özden, Ü. H. and Türedi, S. "Analysis of the economical similarities of European Union members and candidate countries with cluster analysis". İstanbul Trade University Social Sciences Journal, 5.9:95-109, 2006.

[3] Ersöz, F. "Comparison of the Selected Health Indicators of OECD Member Countries with Cluster and Discriminant Analysis". Journal of Medical Sciences, 29.6:1650-1659, 2009.

[4] Akın, H. B., & Özge, E. "OECD Countries With Education Indicators Comparative Analysis of Cluster Analysis and Multi-Dimensional Scaling Analysis". Proposal Journal, 10.37:175-181, 2012.

[5] Aykın, S. M. and Korkmaz, A. "Clustering Turkey and the Member States in Terms of EU-2020 Strategy Indicators". ESOGÜ Journal of Faculty of Economics and Administrative Sciences, 9.1:7-20, 2014.

[6] Tekin, B. "Grouping of cities in terms of primary health indicators in Turkey: an application of cluster analysis". Journal of Karatekin University Faculty of Economics and Administrative Sciences, 5.2:389-417, 2015.

[7] Atal, S. "Fuzzy Clustering Analyze and clustering OECD Countries in development". ESOGÜ Journal of Institute of Science, 2015.

[8] Turan, K. K., Özarı, Ç., "Comparing Turkey and The Middle East Countries with Cluster Analysis: Economic Perspective". İstanbul Aydın University, 29: 143-165, 2016.

[9] Tatlıdil, H. Applied Multivariate Statistical Analyze. Ankara: Academy Publishing, 1996.

[10] Rencher, A. C. Methods of Multivariate Analysis. A John Wiley & Sons, Inc. Publication, 2002.

[11] Aggarwal, C.C., Reddy, C. K. Data Clustering Algorithms and Applications. Boca Raton: CRC Press, 2014.

[13] Brock, G., Pihır, V., Datta, S., Datta, S. "clValid: An R Package for Cluster Validation". Journal of Statistical Software, 25.4:1-22, 2008.

[13] https://cran.r-project.org/web/packages/cluster/cluster.pdf

[14] http://www.worldbank.org/

[15] Alpar, R. Applied Multivariate Statistical Methods. Ankara: Detay Publishing, 2013.