Clustering of Countries by the Factors Affecting Levels of Development and It's Comparison by Years

Coskun Parim[1,*], Batuhan Ozkan[1], Erhan Cene[1]
[1]*Department of Statistics, Yildiz Technical University, Davutpasa Campus, 34220, Istanbul, Turkey*

***Abstract*— In the globalizing world, there are many variables that affect the development levels and economies of countries. A comprehensive analysis of these variables is crucial for the future of countries. In this sense, countries are classified as underdeveloped countries, transition countries, developing countries, and developed countries etc. It is an undeniable fact that the countries classified in this way and in the same class have similar characteristics. In this study, it is aimed to reveal the economic changes of Balkan and former Soviet Union countries over the last 20 years with clustering of these countries by using the factors that affect levels of development. First, socio-economic variables which are considered to affect levels of development were taken according to years, missing data imputation methods were used for identification of missing values of the variables. Later, variables which affect levels of development are determined and with the help of these variables, similar countries are separated into clusters with cluster analysis. Same procedures are made for 1995 and 2015 years, changes of countries over the years are shown.**

***Keywords*— balkan countries, clustering analysis, data analysis, developing countries, missing value**

I. INTRODUCTION

The economic development of countries has been an important concept throughout history. Besides, there are too many variables that show the social and cultural development of countries except the economy. Using these variables, countries can be compared with each other. The level of development in the literature, developing countries, transition countries, developed countries, such as names can be said. In this context, it can be seen that similar countries have the same names in terms of development.

Nowadays as is known, access to data is easier than in the past. However, there are still missing values in the data. Different statistical methods are used to solve the missing data problem because it is important to have complete observations in data analysis. Given the variables that show the development of countries, it is not possible to compare countries if a country has missing data. In this study, 14 different variables were used to compare the development levels of 54 different countries. The focus of the study is to compare the situation of these countries in 1995 and 2015.

In our study, missing observations in the data were completed by using missing value imputation methods. Then, clustering analysis was used to classify the countries' development using variables. These transactions were made for both 1995 and 2015. The aim is to see the changes in the countries in the last two decades.

This work will continue as follows: Section II discusses previous studies, data and variables are defined in Section III, Section IV Methods will be mentioned, Section V will interpret the results. Finally, Section VI will conclude the study.

Manuscript received November 30, 2018; accepted February 11, 2019.
*Corresponding author: cparim@yildiz.edu.tr

II. PREVIOUS WORKS

Carree, M., Van Stel, A., Thurik, R., & Wennekers, S. [1] used data panel of 23 OECD countries. The relationship between economic development and business ownership is examined and the equilibrium points were pointed out.

Maddison, A. [2] made a comparison of levels of GDP Per Capita in Developed and Developing Countries.

Taş, Ç. K., & Özel, S. Ö. [3] compared with the factor analysis according to the development levels of Turkey and the European Union (EU) Countries in terms of the socio-economic indicators.

Saint-Arnaud, S., & Bernard, P. [4] applied hierarchical cluster analysis of the welfare regimes using a set of quantitative social indicators in advanced countries.

Hulten, C. R., & Isaksson, A. [5] researched economic growth in a panel of high and low-income countries and pointed out to have a gap between the rich and poorer countries because of low levels of technological efficiency.

Noorbakhsh, F., Paloni, A., & Youssef, A. [6] detected empirical findings using Human Capital and FDI Inflows in Developing Countries.

Goldberg, L. S., & Klein, M. W. [7] investigated the relationships among trade, foreign direct investment and the real exchange rate between Developing Countries.

Williamson, J. B., & Boehmer, U. [8] investigated the utility of gender stratification theory for national differences in female life expectancy in less developed countries.

Grzebyk, M., & Stec, M. [9] tried to evaluate the progress made by EU countries in the areas of sustainable development by using statistical analysis in 2005 and 2012.

III. DATA AND VARIABLES

A. Data

The data used in the analysis were collected from different websites. The data were determined separately for both 1995 and 2015. The missing values in the data were determined by using the different missing value imputation methods. Then, because the variables were measured with different measurement units, standardization was made for both datasets. After that, clustering analysis was applied according to the k-means method using standardized variables.

B. Variables

The following are the variables that show the development of the countries.

- **FDI:** Logarithm of Foreign Direct Investment (FDI) stocks US Dollars at current prices and currency exchange rates in millions.
- **GDP:** Logarithm of Gross Domestic Product (GDP) at constant 2005 U.S. dollars.
- **Exchange Rate:** Real Effective Exchange Rate
- **Trade Openness:** (Import + Export) / total GDP
- **CL:** Civil Liberty Index: 1: High Civil Liberty, 7: Low Civil Liberty
- **KOF:** Index of Globalization 1: No Globalization, 100: Total globalization
- **Inflation:** Average Consumer Prices
- **PR:** Political Right Index 1: High Political Right 7: Low Political Right
- **Freedom:** Economic Freedom Index 0: No Freedom 100: Complete Freedom
- **Secondary:** Educated percentage of the working-age population
- **Tertiary:** Educated percentage of the working-age population
- **Internet:** Internet users (per 100 people)
- **Pop:** Logarithm of absolute values in thousands
- **Energy:** Energy Productivity
- **Labor Productivity:** GDP per person engaged (constant 1990 US$ at PPP)

IV. METHODS

In the study, it should be stated that have been determined missing values by using the appropriate ones between missing values imputation methods. The data are then standardized because it has different measurements. In the last stage, clustering analysis was applied by using standardized variables.

In our study it was used, "MICE", "Amelia", "missForest", "Hmisc", "mi" packages for missing data imputation and kmeans() function for clustering analysis in r programming language.

A. Missing Values Imputation

It can be said that special focus in the fields of statistics for the imputation of missing values [10]. There are many different missing value imputation methods. Some of these are such as Hot Deck Imputation, Cold Deck Imputation, Multiple Imputation, Regression Imputation, Expectations-Maximization etc. Using the most appropriate method for the structure of the data is important to obtain the right results.

As is known, the main purpose of regression analysis is to estimate dependent variable values by means of one or more independent variables. In this method, the missing variable is determined by the help of simple or multiple linear regression analysis in numerical data type, while in the binary qualitative data type this estimation is made by logistic regression [11]. On the other hand, in our study, the data were used as time series because we know values of years between 1995 and 2015. While the missing value of any observation is found, the variable with the missing value is taken as the dependent variable and the year is taken as an independent variable and the trend equation is found. The missing value imputation is realized using this trend equation.

B. K-Means Clustering

There are various clustering techniques, such as k-means clustering, k-medoids clustering, hierarchical clustering, and density-based clustering etc. The K-Means in these clustering methods mostly used because it is easily applicable. The method has been developed to group a certain number of sets (k).

The method basically can be considered in 3 stages [12].

Step 1: Determine the centroids

Step 2: Determine the distance of each object to centroids.

Step 3: Cluster object for minimum distance.

Following the algorithm is applied to minimize this function [13].

$$J = \sum_{i=1}^{k} \sum_{i=1}^{n} \left\| X_{j}^{i} - c_i \right\|^2 \quad (1)$$

where x_j^i and c_i show the j-th data point and the i-th centroids. $\left\| x_j^i - c_i \right\|$ means the L^2 norm of $(x_j^i - c_i)$.

V. RESULTS

The descriptive statistics of the raw data of the variables affecting the development levels of the countries are given in Table I and Table II. The minimum value, maximum value, standard deviation and means of all countries for both 1995 and 2015 are shown in these tables. In order to better explain the data, these variables are given the raw state. After this stage, the data is standardized before clustering analysis.

TABLE I. DESCRIPTIVE STATISTICS FOR VARIABLES RELATED TO THE LEVEL OF DEVELOPMENT OF COUNTRIES IN 1995

	1995 (N=54)			
Variables	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
FDI	3.47	11.52	7.32	2.08
GDP	7.80	14.20	10.66	1.74
Internet	0.00	2.90	0.24	0.58
ExchangeRate	53.46	407.22	97.33	52.56
PR	1.00	7.00	4.06	2.02
CL	1.00	7.00	4.20	1.66
Freedom	23.30	72.00	51.01	11.71
KOF	27.87	69.48	49.31	10.66
Labor Productivity	6.12	10.64	8.93	1.06
Trade Openness	15.08	177.36	74.18	37.39
Pop	5.94	14.03	9.64	1.63
Inflation	1.06	2672.23	142.66	400.44
Secondary	13.00	91.20	56.32	23.97
Tertiary	1.24	39.13	10.97	7.87
Energy	3.48	7.86	6.05	1.04

TABLE II. DESCRIPTIVE STATISTICS FOR VARIABLES RELATED TO THE LEVEL OF DEVELOPMENT OF COUNTRIES IN 2015

	2015 (N=54)			
Variables	**Minimum**	**Maximum**	**Mean**	**Std. Deviation**
FDI	7.62	14.02	10.53	1.47
GDP	8.71	16.00	11.52	1.70
Internet	11.60	89.65	52.55	22.21
ExchangeRate	62.74	529.72	112.56	61.40
PR	1.00	7.00	3.54	2.10
CL	1.00	7.00	3.50	1.79
Freedom	34.30	76.80	59.85	8.94
KOF	37.43	84.20	65.22	11.81
Labor Productivity	6.97	10.98	9.50	0.94
Trade Openness	23.04	276.23	87.51	48.57
Pop	6.06	14.15	9.82	1.72
Inflation	0.13	910.00	36.81	152.60
Secondary	21.78	98.12	71.70	23.58
Tertiary	2.51	57.71	16.38	11.32
Energy	4.81	7.67	6.37	0.74

The following table shows the results of the analysis according to the k-means method. The principal component analysis was used to select the value of k for the k-means cluster method. As a result of the analysis, 3 factors were obtained for 1995 and 2015, variance explanations were calculated as 67.823 and 66.750 respectively. As a result of this analysis, k values are taken as 3 for both of them. In addition, the analysis results for 1995 and 2015 are shown in Table III and Table IV respectively. 3 clusters of sizes 18, 20, 16 are calculated according to K-means clustering for 1995 and similarly 3 clusters of sizes 13, 23, 18 are given in the table for 2015. ANOVA analysis for the variables in the clusters was found to have a significant difference between the clusters for both two years. It also differs in 3 clusters with regard to 14 variables.

Table III. RESULTS OF THE K-MEANS METHOD FOR 1995

1995		
Cluster 1 (N=18)	Cluster 2 (N=20)	Cluster 3 (N=16)
Algeria	Argentina	Albania
Brazil	Bulgaria	Angola
Cameroon	Croatia	Armenia
China	Cyprus	Azerbaijan
Cote d'Ivoire	Czech Republic	Belarus
Egypt	Estonia	Bosnia and Herzegovina
Ghana	Hungary	Ethiopia
India	Latvia	Georgia
Indonesia	Lebanon	Kazakhstan
Kenya	Lithuania	Kyrgyzstan
Mexico	Malaysia	Macedonia
Morocco	Malta	Moldova
Nigeria	Poland	Tajikistan
Pakistan	Romania	Turkmenistan
Peru	Russia	Ukraine
Senegal	Slovakia	Uzbekistan
Tunisia	Slovenia	
Turkey	South Africa	
	South Korea	
	Venezuela	

TABLE IV. RESULTS OF THE K-MEANS METHOD FOR 2015

2015		
Cluster 1 (N=13)	Cluster 2 (N=23)	Cluster 3 (N=18)
Argentina	Albania	Algeria
Brazil	Armenia	Angola
China	Bosnia and Herzegovina	Azerbaijan
India	Bulgaria	Belarus
Indonesia	Croatia	Cameroon
Kazakhstan	Cyprus	Cote d'Ivoire
Mexico	Czech Republic	Egypt
Morocco	Estonia	Ethiopia
Peru	Georgia	Ghana
Russia	Hungary	Kenya
South Africa	Latvia	Kyrgyzstan
Turkey	Lebanon	Nigeria
Ukraine	Lithuania	Pakistan
	Macedonia	Senegal
	Malaysia	Tajikistan
	Malta	Turkmenistan
	Moldova	Uzbekistan
	Poland	Venezuela
	Romania	
	Slovakia	
	Slovenia	
	South Korea	
	Tunisia	

When we look at all 3 clusters in 1995, it is seen that countries like Brazil, Senegal, and Morocco in the first cluster are developing countries. Similarly, when the second cluster is examined, we can see the developed European countries such as the Czech Republic, Poland, and Romania. Transition countries such as Turkmenistan, Uzbekistan, and Angola can be said to be in the third cluster. When the results are examined it is quite clear that it is very logical and consistent. Having studied the analysis made for 2015, clearly understandable that countries such as Belarus, Turkey, and Malta unchanged. On the other hand, it can be concluded that countries like Pakistan, Nigeria, Ghana, and Egypt are transitioning from the status of developed countries to transition countries. In addition, Argentina and Russia were included in the set of developed countries in 1995. According to the results of 2015, it can be said that it is included in the cluster of developing countries. What is important here is that the countries in the same cluster have similar characteristics. For example, while Venezuela was in the same cluster as Hungry and Estonia in 1995, in 2015 it was in the same cluster as Tajikistan and Turkmenistan. This result shows that the development of Venezuela in the last 20 years is negative.

VI. CONCLUSIONS

In this study, 14 different variables are examined for the development of the countries. 54 countries data composed developed, developing and transition countries were used. Among these countries, the former Soviet Union countries, Balkan countries, and South America countries are included. In addition, data were collected in order to compare for 1995 and 2015. First of all, missing observations are assigned by using missing value imputation methods. The variable was then standardized and included in the analysis. The results were analyzed using the k-means method for 1995 and 2015. ANOVA analysis revealed that the used 14 variables differed within formed clusters. 3 clusters formed as a result of clustering analysis formed a logical classification. The analyzes for 1995 and 2015 were compared. It can be understood from the results that some countries have gone forward or backward in 20 years.

Also, it is normal for countries to progress or recession in terms of development in 20 years. From the results, it can be said that Turkey remains stable in its class. Besides, while Venezuela was in the same cluster with developing countries in 1995, it can be said that in 2015, it was in the same cluster with the transition countries. Of course, based on the results it is also possible to make different comments about the countries.

VII. REFERENCES

[1] Carree, M., Van Stel, A., Thurik, R., & Wennekers, S. (2002). Economic development and business ownership: an analysis using data of 23 OECD countries in the period 1976–1996. Small business economics, 19(3), 271-290.

[2] Maddison, A. (1983). A comparison of levels of GDP per capita in developed and developing countries, 1700–1980. The Journal of Economic History, 43(1), 27-41.

[3] Taş, Ç. K., & Özel, S. Ö. (2017). Faktör analizi yöntemi ile Türkiye ve Avrupa Birliği Üyesi Ülkelerin sosyo-ekonomik göstergeler bakımından gelişmişlik düzeylerinin karşılaştırılması. Journal of the Cukurova University Institute of Social Sciences, 26(3), 60.

[4] Saint-Arnaud, S., & Bernard, P. (2003). Convergence or resilience? A hierarchical cluster analysis of the welfare regimes in advanced countries. Current Sociology, 51(5), 499-527.

[5] Hulten, C. R., & Isaksson, A. (2007). Why development levels differ: The sources of differential economic growth in a panel of high and low-income countries (No. w13469). National Bureau of Economic Research.

[6] Noorbakhsh, F., Paloni, A., & Youssef, A. (2001). Human capital and FDI inflows to developing countries: New empirical evidence. World development, 29(9), 1593-1610.

[7] Goldberg, L. S., & Klein, M. W. (1997). Foreign Direct Investment, Trade and Real Exchange Rate Linkages in Developing Countries (No. w6344). National Bureau of Economic Research.

[8] Williamson, J. B., & Boehmer, U. (1997). Female life expectancy, gender stratification, health status, and level of economic development: A cross-national study of less developed countries. Social Science & Medicine, 45(2), 305-317.

[9] Grzebyk, M., & Stec, M. (2015). Sustainable development in EU countries: concept and rating of levels of development. Sustainable Development, 23(2), 110-123.

[10] Demirhan, H., & Renwick, Z. (2018). Missing value imputation for short to mid-term horizontal solar irradiance data. Applied Energy, 225, 998-1012.

[11] Alpar, R. (2011). Çok Değişkenli İstatistiksel Yöntemler, Ankara: Detay Yayıncılık.

[12] Dalatu, P. I., Fitrianto, A., & Mustapha, A. (2017). Hybrid distance functions for K-Means clustering algorithms. Statistical Journal of the IAOS, 33(4), 989-996.

[13] Wang, Q., Wang, Y., Niu, R., & Peng, L. (2017). Integration of Information Theory, K-Means Cluster Analysis and the Logistic Regression Model for Landslide Susceptibility Mapping in the Three Gorges Area, China. Remote Sensing, 9(9), 938.