Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi
Pamukkale University Journal of Engineering Sciences

The determination of optimal cluster number by Silhouette index at clustering of the European Union member countries and candidate Turkey by waste indicators

Avrupa Birliği üye ülkeler ve aday olan Türkiye'nin atık indikatörlerine göre kümelenmesinde optimum küme sayısının Silhouette indeksi ile belirlenmesi

Tuğba Söküt AÇAR[1], Nilgün AYMAN ÖZ[2*]*

[1]Department of Statistics, Faculty of Arts And Sciences, Çanakkale Onsekiz Mart University, Çanakkale, Turkey.
t.sokut@comu.edu.tr
[2]Department and/or Faculty, University/Institution, City, Country.
nilgunayman@comu.edu.tr

Received/Geliş Tarihi: 13.07.2018, Accepted/Kabul Tarihi: 15.04.2019 doi: 10.5505/pajes.2019.49932
* Corresponding author/Yazışılan Yazar Research Article/Araştırma Makalesi

Abstract

This study aims to identify cluster structure of European Union (EU) Member countries and Candidate Turkey in terms of environmental waste indicators and to determine the other member countries which are classified in the same cluster with Turkey. Hierarchical and non-hierarchical clustering methods were used to determine clusters of 28 member countries and Turkey according to the total 8 environmental waste indicators. The optimal cluster number and the best method were identified with the silhouette index which is a cluster validity index. The results from the cluster analysis using the hierarchical and non-hierarchical methods showed that there are six clusters according to the environmental waste indicators of EU countries and Turkey. The average Silhouette index shows that the k-means gives more valid results than the ward. According to the Silhouette index obtained by k-means method, Turkey has been found to be classified in the same cluster with %50 of the EU countries such as Poland, Hungary, and Latvia etc.

Keywords: Hierarchical and Non-Hierarchical Clustering, K-Means, Ward, Silhouette Index, Waste, Environmental Indicator

Öz

Bu çalışma Avrupa Birliği üye ülkeleri ve aday olan Türkiye'nin çevresel atık göstergeleri açısından kümelenme yapısını tanımlamayı ve Türkiye ile aynı kümede sınıflandırılan diğer üye ülkeleri belirlemeyi amaçlamaktadır. 28 üye ülkenin ve Türkiye'nin toplam 8 çevresel atık göstergesine göre kümelenmesinin belirlenmesinde hiyerarşik ve hiyerarşik olmayan kümelenme yöntemleri kullanılmıştır. En iyi kümeleme metodu ve optimum küme sayısını belirlemek için küme geçerlilik indeksi olan Silhouette indeksi kullanılmıştır. Hiyerarşik ve hiyerarşik olmayan yöntemler kullanılarak elde edilen küme analizi sonuçları göstermiştir ki, AB üye ülkeleri ve Türkiye çevresel atık göstergelerine göre altı kümeden oluşmaktadır. Ortalama Silhouette indeksi göstermiştir ki, K-ortalamalar yöntemi Ward yöntemine göre daha geçerli sonuç vermiştir. K-ortalamalar yöntemi ile elde edilen Silhouette indeksine göre Türkiye; Polonya, Macaristan ve Letonya gibi AB ülkelerinin% 50'si ile aynı kümede sınıflandırmıştır.

Anahtar kelimeler: Hiyerarşik ve Hiyerarşik Olmayan Kümeleme, K-Ortalamalar, Ward, Silhouette İndeksi, Atık, Çevresel İndikatörler

1 Introduction

The main source of a strong economy depends on a clean environment and a healthy society. The rapid increase in the world population, raw material and energy consumption for production of goods and services generates high amount of wastes causing environmental problems such as air pollution, water pollution, soil pollution, subsequently depleting natural resources. The management strategies to deal with these wastes have an effect directly on our health and also our future [1]. The definition of 'waste' was originally described as 'any substance or object which the holder discards or intends to discard' (Waste Framework Directive 75/442/EEC 1975). Wastes are generated by human activities such as industrial production, agricultural activities and consumption and they can generally be classified as solid wastes, liquid and gaseous wastes. An adequate waste management program which does not cause environmental pollution should be developed in order to protect public health. It has been accepted in the literature that the way to be followed in the waste management is to prevent waste formation, to reduce or recover if not prevented, and to dispose of the waste appropriately.

When European Union (EU) has been founded, environmental pollution was not considered a serious problem and therefore, environmental issues were not adopted in the declaration at the first place. The Paris Summit has been accepted as the beginning of environmental policy of the Europe and after 1972, environmental issues have become EU's most significant area of competence [2]. The EU strategy depends on the different categories of waste management and measures of policy success in recycling, waste minimization, etc. [3]. The strategy is based on sustainable waste management and encourages strategies such as waste minimization, reducing, reusing, recycling and recovering of wastes before treating and disposing it as explained in the hierarchy of waste management options. It has been reported that economic cost of the waste management alternatives may not be the only consideration for the local authority decision makers in order to apply the principles of best practicable environmental option [3]. Turkey as a candidate to EU is trying to set goals in order to adapt its waste management strategy to the standards. During the process of adaptation of Turkey to EU, environmental performance has been progressed; new environmental laws and regulations have been issued. Although the progress, it has

been stated in the EU reports that the performance should be improved through the various legislative regulations on the environmental issues. Also according to this report, various environmental standards are not consistent with EU border values.

Various statistical methods are used in the evaluation of environmental indicators. If more than one indicator is used, the data can be summarized with multivariate statistical methods. One of the popular statistical methods based on the classification of the indicators is the clustering analysis. Anderberg [4] noted that the results of a cluster analysis can contribute directly to the development of the classification schemes. Clustering analysis which is a group of multivariate techniques are performed through some determined objects (units) based on their characteristic features. Clustering analysis is a method that the unit-based objects are classified according to their similarities or dissimilarities [5]. Neil [6] indicated that the clustering analysis separates a set of objects into different clusters. In this context, all objects within a group can be classified based on their similarity. Thus, objects outside the group are not similar. Thus, a homogenous classification within the cluster and a heterogeneous classification between the clusters can be carried out. Cluster analysis has been employed as an effective tool in different fields such as biology [7], economics [8], education [9]), health [10]. The application of the clustering methods to environmental field has been reported in the literature as following. Brono et al. [11] analyzed six European countries (France, Italy, the Netherlands, Spain, Switzerland and Slovenia), exploring decision-making processes for the establishment and disposal of waste facilities. Ukpatu et al. [12] used the agglomerative hierarchical cluster analysis to Assessment of water quality of Okoro River Estuary, Southeastern Nigeria. Du et al. [13] used hierarchical and non-hierarchical clustering methods for determining the water quality depending on the distance of water. Arbolino et al. [14] used a hierarchical clustering analysis method to classify regions in Italy in order to determine the efficiency of EU greenhouse gas emission regulation. Williams [3] stated that the classification of waste can be complex due to the inconsistent waste parameters, data collection and reporting methods for different countries. Therefore, it is important to define the waste parameters accurately, to decide the best method to classify the waste management methods to determine the similarities of the countries. In this respect, different clustering methods such as hierarchal and nonhierarchical) should be tested in order to determine the best classification structure. However, the cluster number in the nonhierarchical clustering analysis is decided in advance. Therefore, the researches prefer to determine the cluster number with dendrogram or different determinants using hierarchical methods. Lack of classification according to waste indicators is a problem in terms of not knowing the cluster structure beforehand.

The main objective of this paper is to classify the environmental waste indicators of the 28 EU countries and the candidate Turkey using hierarchical and nonhierarchical clustering methods and to determine the situation of Turkey in EU in terms of waste management indicators. Since there is no preliminary information about the cluster structure in doing this classification, it is aimed to determine optimal cluster number with the Silhouette index. Thus, the present paper will contribute to the literature by providing preliminary information about the cluster structure of EU countries according to the waste indicators. The rest of the paper organized as follows: the rest of paper is organized as follows: the distance measurement and the clustering methods used in the paper are given in the methodology section. Outputs of analysis are given in the results section. The paper ends with conclusion section.

2 Method

Waste Indicators:

In the world, different approaches are applied for the development of environmental indicators and indicator sets are formed within different conceptual frameworks or models. One of these models is the DPSIR framework developed by the EEA (European Environment Agency) in 2004 to develop a PSR framework to identify the relationship between society and the environment. This model consists of five elements as driving force, pressure, state, impact, response. The Ministry of Environment and Urbanization locate the municipal waste disposal, waste disposal and recovery response indicators in response elements. In this study we focused on waste and recovery data from the response element.

The important point in the classification of waste management by hierarchical and non-hierarchical methods for EU member states and Turkey, which is a EU candidate country, was to determine the variables to be used. The study is restricted to 2014 data since the complete data for all variables can only be obtained for 2014. In this paper, it is aimed to classify waste production and management methods according to the following variables. The variables are obtained from Eurostat which is one of the institutions that play the most important role in the international environmental indicators. We received data from this site:

"http://ec.europa.eu/eurostat/data/database" in December 2017.

x_1 is the municipal waste generation and treatment by type of treatment method,

x_2 is total waste (Chemical and medical wastes, Recyclable wastes, Animal and vegetable wastes, Mixed ordinary wastes, Mineral and solidified etc.),

x_3 is the deposit onto or into land,

x_4 is the land treatment and release into water bodies, x_5 is incineration /disposal (D10),

x_6 is incineration/energy recovery (R1),

x_7 is recovery other than energy recovery-backfilling, and x_8 is recovery other than energy recovery-except backfilling. All of the variables units are tonne and are hazardous and nonhazardous total. Factors such as being indicators of environmental waste in selecting variables and reaching the data of all countries with these variables have been influential. To determine the status of Turkey in the average of the 28 EU countries, the descriptive statistics of the used variable are given in Table 1.

Table 1: Descriptive statistics of EU countries according to each indicator with unit tonne.

	x_{Mean}	S_x
x_1	469.607	123.089

x_2	89096381,320	104986456.100
x_3	33748961.140	47112846.650
x_4	5544637.500	15472142.230
x_5	1236050.250	2531409.707
x_6	3855509.321	7655912.563
x_7	8460476.714	19468538.710
x_8	30034841.570	46883318.530

Here x_{Mean} present the mean of the corresponding variable, S_x present the standard deviation of the corresponding variable. All values except x_3 of Turkey's is below the EU average.

Cluster Analysis:

In clustering analysis, it is aimed to classify the objects which are defined according to their characteristic features based on the similarities or dissimilarities. In our study, the objects are defined as countries in EU and candidate Turkey.

The first step of the clustering analysis is to determine the variables and create the data matrix. The variables must be defined consistently in terms of their characteristics, quantity, quality, and other such properties. Anderberg [4] stated that the choice of variables that has the greatest impact on the final results of a cluster analysis. The data matrix is formed after the variables are determined. A data matrix is expressed as follows

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix}_{n \times p} \quad (1)$$

where n represent the number of country, p represent the variable number (waste indicator). Each entry $(x_{i,j})$ in the X refers to the value of i-th country according to the j-th variable. In this study, the data matrix for 29 countries and 8 variables is as follows:

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,8} \\ x_{2,1} & x_{2,2} & \dots & x_{2,8} \\ \vdots & \vdots & \ddots & \vdots \\ x_{29,1} & x_{29,2} & \dots & x_{29,8} \end{bmatrix}_{29 \times 8}$$

If the data measurement types are not the same, all variables must be reduced to the standard form (zero mean and unit variance) by converting $z_{ij} = \frac{X_{ij} - \overline{X_j}}{S_{X_j}}$. Here $X_{i,j}$ denotes each row element in the column vector X_j, $\overline{X_j}$ denotes the average of the j-th variable and S_{X_j} denotes the standard deviation of the j-th variable. The main operations of the cluster analysis can be done on the data matrix or the standardized data matrix [15].

The second step is the creation of distance measure. In clustering analysis, while the n country divide homogeneous within cluster whereas inter cluster heterogeneous structure, different distances measured based on similarities or differences between variables are used. The distances matrices are emphasized because of the variables in this study are quantitative. The Euclidean distance representing the distance between i-th and j-th unit or objects in an n × p dimensional data matrix is the most preferred distance measure in the literature (Everitt et al., 2011). Reasons for using it as popular are that it is not necessary to know a priori information about the data and to minimize the average error squares between the groups. This distance is expressed as an application of the Pythagorean Theorem in two-dimensional space. Euclidean distance between country A and country B with regard to the p variables is calculated as follows

$$d(A,B) = \sqrt{\sum_{k=1}^{p} \left(x_{A,k} - x_{B,k}\right)^2} \quad (2)$$

The third step is based on the method to be used in the clustering analysis. Clustering techniques made by [4] and accepted in the statistical literature are classified as hierarchical and non-hierarchical clustering techniques. In the hierarchical clustering method, the cluster number is determined by analyzing without prior knowledge whereas in the non-hierarchical clustering method, the analysis is performed under the preliminary knowledge of cluster number. Cornish [16] stated that another difference between the two techniques is based on the sample size. According to the study, hierarchical clustering method must be used when small data set are involved while non-hierarchical cluster analysis tends to be used when large data sets are involved.

Hierarchical clustering techniques can be applied by either a series of agglomerative or sequential division procedures. In this study, agglomerative procedure which is mostly preferred is used. In the procedure, primarily, each object (country) is assigned to a single cluster. Two objects that have the most similarity create the first subgroup. In the next step, another pair of clusters is merged and this process continues as hierarchically. Consequently, the procedure continues until the similarity decreases between the groups. The methods frequently used in agglomerative hierarchical clustering analysis are: single linkage, complete linkage, average linkage and ward methods [17].

One of the nonhierarchical clustering methods is the partitioning method. There are many nonhierarchical algorithms for partitioning a set of objects into k clusters, such as k-means and k-median methods.

Since evaluation of the data in this study is done in the SPSS program, hierarchical ward method and non-hierarchical k-means method which are menus in SPSS have been used and details of these methods are included.

Ward [18] has proposed the ward method, which aims to minimize the loss of information between the two groups. Minimize the loss of information is equivalent to minimizing the error sum of square (ESS) within the cluster. So that clusters that maximize homogeneity within themselves are created.

For a given cluster m, ESS_m is the sum of the squared deviations of every item in the cluster from the cluster mean. If there are currently K clusters, define ESS as the sum of the ESS_m as

$$ESS = \sum_{m=1}^{K} ESS_m \quad (3)$$

where $ESS_m = \sum_{l=1}^{n_m} \sum_{k=1}^{p_k} \left(x_{ml,k} - \bar{x}_{m,k}\right)^2$ in which $\bar{x}_{m,k} = \frac{1}{n_m} \sum_{l=1}^{n_m} x_{ml,k}$ (the mean of the m-th cluster for the k-th variable), $x_{ml,k}$ being the score on the k-th variable (k=1, . . ., p)

for the lth object (l=1, . . ., nm) in the m-th cluster (m=1, . . ., K) [19].

The association of every possible pair of clusters is considered in each step, then the two clusters whose combination results in the smallest increase in ESS (minimum loss information) are merged. At first, each cluster occurs of a single item, and, if there are N items, $ESS_k = 0$, k = 1,2, ... , N, so ESS = 0.

.

It is not possible to reassign an object that may have been incorrectly grouped during the clustering at an early phase to see whether it is sensible, the final shape of clusters should be examined carefully at the end of analysis. For this, it is a good way to try out several cluster methods and compare the results using different distance measures [17].

We have considered the k-means method which is the more popular procedure in hierarchical technique. This method firstly proposed by [18]. The purpose of k-means clustering is to separate n objects into k clusters and thus each object is assigned to the cluster with the closest mean.

Sarıman [20] stated that the k-means method is based on a sharp set algorithm since it allows each object belongs to only one cluster. Consider each p-dimensional $(x_1\ x_2\ \ldots . x_n)$ dataset.

In the k-means method which aims to divide n objects to k cluster such as $S = (S_1, S_2, \ldots, S_k)$ the best result is the set with the smallest ESS value.

The sum of the squares of the distances to the center points of the coops where the objects are made is as follows [21]:

$$ESS = \sum_{k=1}^{K} \sum_{i \in S_k}^{n} d^2(x_i, \mu_k) \qquad (4)$$

where ESS_k sum of error squares of objects in the k-th cluster, x_i refers to the value of the i-th object in S_i. μ_k refers the central point of the k-th cluster.

It is crucial to decide the cluster number in the dataset accurately since researchers can use different cluster numbers performing a cluster analysis depending on several methods [22].

When nonhierarchical method is used, it can be decided with graphical representation known as dendrogram. Because the interpreting the dendogram vary from analyst to analyst, it may be used as a preliminary information. In this case, other cluster number determination methods are needed.

The simplest and most popular formula of cluster number is as follows

$$k = \sqrt{\frac{n}{2}} \qquad (5)$$

However, the increase in the number of objects causes the number of clusters to increase meaninglessly. Therefore, Everitt [23] stated that it is more appropriate to use this method in determining the number of clusters in small sample surveys.

Another way is to determine the number of clusters by looking at the cluster validity. The statistics used in cluster validity can be listed as Silhouette, Calinski and Harabazs, Krzanowski and Lai and Lewis and Thomas.

The Silhouette index has performed well in many comparative analyses [24, 25]. In addition, the Silhouette index is used in this study because of it can be workable with many distance measurements including the Euclidian. The Silhouette (SIL) index developed by Rousseuw [26] with the aim of determining the suitability of each unit for the cluster is as follows

$$Sil(x_i) = \frac{b(x_i) - a(x_i)}{maks(a(x_i), b(x_i))} \qquad (6)$$

where $a(x_i)$ is defined as the average dissimilarity of object i to all object s in the same cluster and $b(x_i)$ as the minimum across all other clusters of average dissimilarity of object i to all objects in each cluster. In addition, $Sil(x_i)$ provides the inequality of $-1 \leq Sil(x_i) \leq 1$ and when $Sil(x_i)$ is close to 1 it indicates that the object is well classified. If $Sil(x_i)$ is approximately zero then we can say that the i-th object is between two clusters. When $Sil(x_i)$ is close to -1 it indicates that the i-th object is misclassified. The average of the $Sil(x_i)$ obtained for the relevant cluster number indicates the cluster validity, and Rousseeuw [24] stated that the cluster number corresponding to the maximum average Silhouette value is taken as the optimum. But in general, if the average Silhouette value is above 0.50, then it is accepted that the appropriate clusters are reached.

In the classification of the waste indicators used statistics as follows: the Euclidean distance was used in determining the distance matrix. For clustering of countries ward method was used in hierarchical clustering analysis and k-means method was used in non-hierarchical clustering analysis. The optimum cluster number was determined by Silhouette index. All of statistical analysis was performed through the IBM SPSS 24.0 package program.

3 Results

Each variable is standardized with mean zero and standard deviation of one in order to treat them as having equal importance in determining the structure and cluster analysis were applied after standardization.

It rendered a dendrogram as shown in Figure 3 by using Ward linkage. In Figure 3, the horizontal axis of the dendrogram indicates the distance between the groups that are cluster together and the vertical axis represent the countries. It can be seen that on the left of the dendrogram, countries are close together when they cluster and on the right of the dendrogram there is a only small number of groups and the distance between those groups is larger. Determining the optimal number of clusters by dendrograms may result in confusion due to the analyst's interpretation. For example according to the waste indicators, the analyst may not be classified Denmark in the same class with Cyprus, Malta, Ireland and Luxembourg.

Figure 3: Dendrogram with using ward method

Since the number of clusters is not known in advance for waste classification, it is aimed to determine for an optimal partition which consists of clusters that are as compact and relatively strongly separated as possible in this paper. This can be done by using a well-known index SIL and the results compared with

the $\boldsymbol{k} = \sqrt{\frac{n}{2}} \cong \mathbf{4}$. In Table 2 we see clustering results and SIL loads for $\boldsymbol{k} = \mathbf{4}$. As it can be seen from Table 2 negative loads in the SIL are greater when the ward method is used. For k = 4, the average SIL value for the ward method is 0.403, while for the k-means method it is 0.382. It is determined that $\boldsymbol{k} = \mathbf{4}$ is not a valid cluster number since these two average SIL values are below 0.5.

Table 2: Cluster memberships and SIL statistics for $\boldsymbol{k} = \mathbf{4}$.

Country	Ward	$\boldsymbol{Sil_{Ward}}$	K-means	$\boldsymbol{Sil_{K-means}}$
Belgium	1	0.391	2	0.171
Czech Republic	1	0.581	2	0.428
Estonia	1	0.617	2	0.427
Croatia	1	0.574	2	0.339
Latvia	1	0.585	2	0.375
Lithuanian	1	0.476	2	0.177
Hungary	1	0.602	2	0.370
Portugal	1	0.390	2	0.070
Slovenia	1	0.476	2	0.174
Slovakia	1	0.585	2	0.420
Bulgaria	2	-0.054	2	0.220
Romania	2	-0.045	2	0.322
Denmark	3	0.619	4	0.630
Ireland	3	0.765	4	0.779
Cyprus	3	0.785	4	0.781
Luxembourg	3	0.794	4	0.792
Malta	3	0.749	4	0.758
Austria	1	-0.200	4	0.491
Germany	4	1.000	3	1.000
Greece	1	0.299	2	0.080
Spain	1	0.382	2	0.225
Poland	2	-0.077	2	0.242
Finland	1	0.307	2	0.168
Turkey	**1**	**0.493**	**2**	**0.354**
France	2	0.158	1	0.258
Italy	2	-0.015	1	0.154
United Kingdom	2	0.269	1	0.435
Netherlands	2	0.074	1	0.264
Sweden	2	0.101	1	0.181

Figure 1 and Figure 2 show the average of SIL values according to the different cluster numbers for ward and k-means methods, respectively.

According to the these graphs it is obvious that when the cluster number approaches to sample size (n), the average of SIL value approaches to 1. It can be shown that the optimal cluster number is 6 because the highest increase after the decrease was at k=6 for the two used method.

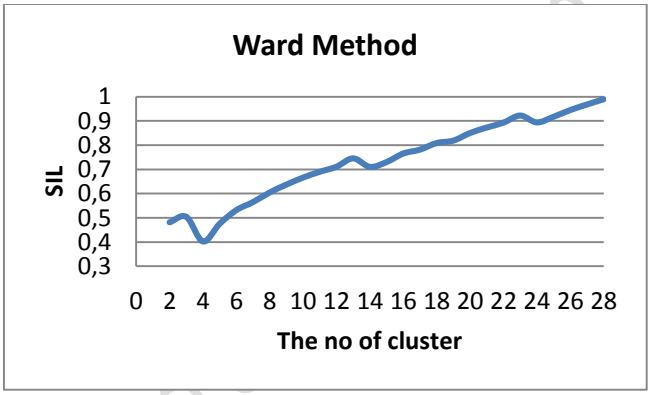

Figure 1: Average Silhouette index on hierarchical clustering with ward method.

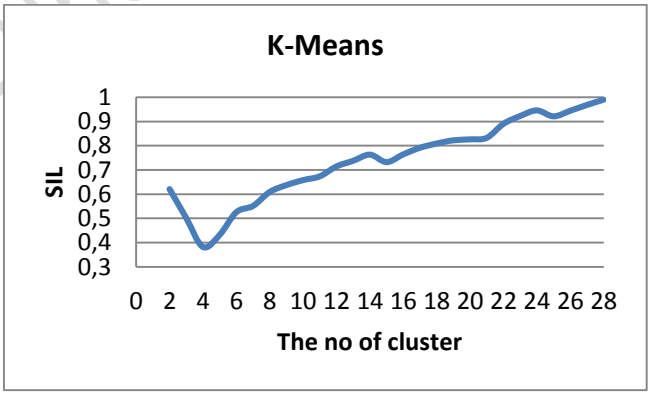

Figure 2: Average Silhouette index on non-hierarchical clustering with k-means method.

Since the optimal cluster number according to SIL index is 6, the results for $k = 6$ are given in Table 3. For k = 6, the average SIL value for the ward method is 0.532, while for the k-means method it is 0.526. The results obtained from Table 3 are compared with Table 2, we can say that k=6 gives much better results according to SIL loads. In addition, it can be said that from Table 3 that the k-means clustering of Austria's will be more accurate than ward method($Sil_{Ward}(x_i) = -0{,}2 < Sil_{K-means}(x_i) = 0{,}42$). Although the average SIL values are closer for two methods, it can be said that the k-means method has better clustering for $k = 6$ then the ward method, since the clustering of Austria is more accurate in k-means method.

Table 3: Cluster memberships and SIL statistics for $k = 6$.

Country	Ward	$\boldsymbol{Sil_{Ward}}$	K-means	$\boldsymbol{Sil_{K-means}}$
Belgium	1	0.391	5	0.309
Czech Republic	1	0.581	5	0.543

Estonia	1	0.617	5	0.561
Croatia	1	0.574	5	0.504
Latvia	1	0.585	5	0.526
Lithuanian	1	0.476	5	0.375
Hungary	1	0.602	5	0.532
Portugal	1	0.39	5	0.276
Slovenia	1	0.476	5	0.376
Slovakia	1	0.585	5	0.542
Bulgaria	2	0.775	3	0.777
Romania	2	0.777	3	0.775
Denmark	3	0.619	4	0.606
Ireland	3	0.765	4	0.750
Cyprus	3	0.785	4	0.753
Luxembourg	3	0.794	4	0.766
Malta	3	0.749	4	0.724
Austria	1	-0.200	4	0.420
Germany	4	1.000	1	1.000
Greece	1	0.299	5	0.172
Spain	1	0.382	5	0.299
Poland	5	0.121	5	0.206
Finland	1	0.307	5	0.210
Turkey	**1**	**0.493**	**5**	**0.426**
France	5	0.449	2	0.529
Italy	5	0.300	2	0.411
United Kingdom	5	0.281	2	0.422
Netherlands	6	0.739	6	0.740
Sweden	6	0.719	6	0.719

Table 4 gives the ANOVA results that indicate which waste variables contribute the most to the cluster solution. Since all Sig. values are smaller than 0.05, it was determined that all the variables used vary according to the clusters. In ANOVA results in cluster analysis, it is normal for variables to differ according to clusters. Because, clustering analysis makes the difference between clusters maximum. According to the ANOVA results for k-means technique, it can be seen that x_5 variable is the variable that contributes most to clustering solution

Table 4: ANOVA analysis of k-means cluster analysis

ANOVA						
	Cluster		Error		F	Sig.
	Mean Square	Degree of freedom	Mean Square	Degree of freedom		
x_1	4,099	5	0.326	23	12.568	0.000
x_2	4,595	5	0.219	23	21.023	0.000
x_3	4,129	5	0.320	23	12.917	0.000
x_4	4,754	5	0.184	23	25.865	0.000
x_5	5,413	5	0.041	23	133.316	0.000
x_6	5,019	5	0.126	23	39.747	0.000
x_7	4,472	5	0.245	23	18.246	0.000
x_8	4,441	5	0.252	23	17.629	0.000

Finally, the results of the accepted k-means obtained from clustering analysis were processed on the European map which is created on ArcGIS (10.3) program (Figure 4).

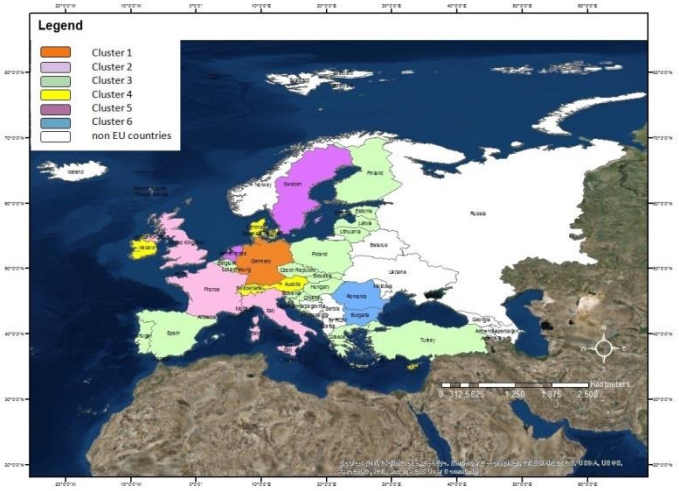

Figure 4: EU map according to the k-means.

4 Conclusion

This study identified the cluster structure of the EU Member countries and Candidate Turkey in terms of waste production and management methods with the hierarchical and nonhierarchical methods. For classification of waste management methods applied in EU, nonhierarchical methods gave better results compared to hierarchical methods. The optimal cluster number was determined as 6 with the Silhouette index Further, the determination of the cluster number with the Silhouette index yielded more valid results than the number of $k = \sqrt{\frac{n}{2}} \cong 4$ cluster number which is calculated based on the number of samples and is frequently referred to in the literature. According to the Silhouette index, Turkey has been found to be classified in the same cluster with %50 of the EU countries such as Poland, Hungary, Latvia etc.

For future studies, Silhouette index can be used as an alternative to $k = \sqrt{\frac{n}{2}} \cong 4$ in different fields such as environment, psychology, biology, economics, and health. In this way, the optimal number of clusters can be obtained based on cluster validity. In the next studies, the cluster structure can be also discussed according to the different environmental indicators such as environmental taxes, environmental protection expenditure, etc. Or the EU position of the other candidate countries can be determined according to the environmental indicators.

5 5 References

[1] Johnstone H. *Facts on Domestic Waste and Industrial Pollutants*. Franklin Watts, New York, 1990.

[2] Jordan AJ, Liefferink D. *Environmental Policy in Europe: the Europeanization of National Environmental Policy*. Routledge, 2004.
[3] Williams PT. *Waste Treatment and Disposal*, Second Edition, John Wiley & Sons, 2005.
[4] Anderberg MR. *Cluster Analysis for Applications*. Academic Press, INC, London, 1973.
[5] Kaufman L, Rousseeuw PJ. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York, John Wiley and Sons, 1990.
[6] Neil HT. *Applied Multivariate Analysis*. Springer-Verlag, New York, 2002.
[7] Türkmen G, Kazancı N. "Assessment of benthic macroinvertebrate communities of some sites at Kelkit Stream and its tributaries (Yeşilırmak river basin, Turkey) with the application of cluster analysis". *Review of Hydrobiology*. 4(1), 29-45, 2011.
[8] Turanlı M, Özden ÜH, Türedi S. "Avrupa Birliğine aday ve üye ülkelerin ekonomik benzerliklerinin kümeleme analiziyle incelenmesi". *İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi,* 5(9), 95-108, 2006.
[9] Akın HB, Eren Ö. "OECD Ülkelerinin eğitim göstergelerinin kümeleme analizi ve çok boyutlu ölçekleme analizi ile karşılaştırmalı analizi". *Oneri.C.10.S.37*, 175-181, 2012.
[10] Tekin B. "Temel sağlık göstergeleri açısından Türkiye'deki illerin gruplandırılması: bir kümeleme analizi uygulaması". *Çankırı Karatekin Üniversitesi İktisadi İdari Bilimler Fakültesi Dergisi*, 5 (2), 389-416, 2015.
[11] Brono D, Fareri P, Ligteringen J. *The Waste and the Backyard the Creation of Waste Facilities*: Success Stories in Six European Countries, 1998.
[12] Ukpatu J, Udoinyang E, Udoh JP. "The use of agglomerative hierarchical cluster analysis for the assessment of mangrove water quality of Okoro River Estuary, Southeastern Nigeria". *International Journal of Geology, Agriculture and Environmental Sciences*. 3(6), 17-24, 2015.
[13] Du X, Shao F, Wu S. "Water quality assessment with hierarchical cluster analysis based on mahalanobis distance". *Environ Monit Assess*, 189(7), 335, 2017.
[14] Arbolino R, Carlucci F, Cira A, Loppolo G, Yiğitcanlar T. "Efficiency of the EU regulation on greenhouse gas emissions in Italy: The hierarchical cluster analysis approach". *Ecological Indicators*. 81, 115-123, 2017.
[15] Romesburg HC. *Cluster Analysis for Researchers*. Lulu Press, North Carolina, 2004.
[16] Cornish R. "Mathematics learninig support centre", http://www.statstutor.ac.uk/resources/uploaded/clusteranalysis.pdf, 2007.
[17] Johnson RA, Wichern DW. *Applied Multivariate Statistical Analysis*. Pearson Prentice, HaWM, 2007.
[18] Ward JRJH. "hierarchical grouping to optimize an objective function". *Journal of the American Statistical Association, 58*, 236-244, 1963.
[19] Everitt BS, Landau S, Leese M, Stahl D. *Cluster Analysis*.5th Edit., John Wiley & Sons, 2011.
[20] Sarıman G. "Veri madenciliğinde kümeleme teknikleri üzerine bir çalışma: k-means ve k- medoids kümeleme algoritmalarının karşılaştırılması". *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 15*(3), 192-202, 2011.
[21] Yürük F, Erdoğmuş P. "Düzce ilinin hayvansal atıklardan üretilebilecek biyogaz potansiyeli ve k-means kümelemeile optimum tesis konumunun belirlenmesi". *İleri Teknoloji Bilimleri Dergisi, 4*(1), 47-56, 2015.
[22] Everitt B. "Unresolved problems in cluster analysis", *Biometrics*. 35(1), 169-181, 1979.
[23] Everitt B. *Cluster Analysis*. Heinmann, London, 1974.
[24] Arbelaitz O, Gurrutxaga I, Muguerza J, M. P´erez J, Perona I. "An extensive comparative study of cluster validity indices". Pattern Recognition, 46 (1), 243–256, 2012.
[25] De Amorim RC, Hennig C. "Recovering the number of clusters in data sets with noise features using feature rescaling factors". Information Sciences, 126-145, 2015
[26] Rousseeuw PJ. "Silhouettes: A graphical aid to the interpretation and validation of clusters analysis". *Journal of Computational and Applied Mathematics*, 20, 53-65, 1987.