Available online at www.alphanumericjournal.com

alphanumeric journal

The Journal of Operations Research, Statistics, Econometrics and Management Information Systems

Volume 9, Issue 1, 2021

Received: June 09, 2021
Accepted: June 28, 2021
Published Online: June 30, 2021

AJ ID: 2021.09.01. STAT.02
DOI: 10.17093/alphanumeric.949958
Research Article

Classification of Cancer Types by Cluster Analysis Methods

Aynur İncekırık, Ph.D. *

Assist. Prof, Department of Econometrics, Faculty of Economics and Administrative Sciences, Manisa Celal Bayar University, Manisa, Turkey, aynur.incekirik@cbu.edu.tr

Öznur İşçi Güneri, Ph.D.

Prof., Department of Statistics, Faculty of Science, Muğla Sıtkı Koçman University, Muğla, Turkey, oznur.isci@mu.edu.tr

Burcu Durmuş

Lecturer, Departments of Strategy Development, Muğla Sıtkı Koçman University, Istanbul, Turkey, burcudurmus@mu.edu.tr

* Manisa Celal Bayar Üniversitesi, İktisadi ve İdari Bilimler Fakültesi, Ekonometri Bölümü, Şehit Prof. Dr. İlhan Varank Yerleşkesi 45140 – Yunusemre, Manisa, Türkiye

ABSTRACT

Cluster analysis can be defined as the group of methods that aim to classify multivariate observations by using similarity/dissimilarity measures between observations. The clusters obtained as a result of the analysis are required to be homogeneous within themselves and heterogeneous among themselves. This study aims to cluster cancer types in datasets created by considering age group characteristics according to gender. In the study, clustering analysis was applied to four different datasets created from the data registered between 1982 and 2016 for 57 cancer types in men and women according to age groups at the Australian Institute of Health and Welfare, and the analysis results were evaluated and interpreted. In addition, in determining the clustering method and the number of clusters, Cophenetic correlation coefficients and 26 cluster validity indices were used, respectively. The distribution of cancer types in age groups determined by gender was observed in 4 different datasets created with 3 different age group characteristics that led to the best separation of cancer groups, and the clustering tendencies of cancers in the relevant age groups were investigated. R-3.5.1 package program was used for analyses. In this study, the analysis results of the k-means method and the average linkage method, which was decided to be the most successful method due to the high cophenetic correlation coefficient value, were evaluated and interpreted. The number of clusters was determined as 3 with the help of cluster validity indices. When the results obtained are examined, it is seen that breast cancer in women and prostate cancer in men is the most common type of cancer in the age group of 40 and above, and that these cancers are alone in a cluster. In addition, it is seen that the 0-14 age group characteristic fails to separate the clusters.

Keywords: Cluster Analysis, Cancer Types, Cluster Validity İndex, k-Means, Cophenetic Correlation Coefficient

1. Introduction

The concept of cluster evokes the concepts of similarity and distance. Clustering is defined as the classification of similar objects using the data obtained from the objects following a useful summary of the data. In other words, clustering occurs when observations close to each other come together in a multidimensional space (Seber, 1990). Cluster analysis is one of the multivariate analysis methods that aims to gather objects together according to the variables they have, that is, to classify objects or individuals. Here, each object is similar to other objects according to a specified criterion. The main purpose of this analysis is to reveal the natural grouping of variables or objects (Johnson & Wichern, 1998). Each data here is characterized by a representative value vector. Another purpose of the analysis is to divide a population in such a way as to gather the data with similar representative values in the same cluster and to collect the data with different representative values in different clusters (Na et al., 2002).

Cluster analysis is a method used to explore the characteristics of grouping a collection of objects based on distance measures. The aim here is to reveal the similarities between the objects (Alhamed & Lakshmivarahan, 2002). Therefore, the concept of similarity is very important in cluster analysis and forms the basis of the analysis. This concept is determined by a measure of the fit between the objects under analysis. Here, by choosing a measure suitable for the data structure, the distance between two objects, in other words, between the observation clusters, is measured (Sharma, 1996). In general, various distance measures can be calculated for each pair of objects. However, the similarity or the distribution of the relationship between objects is most striking in the measure of Euclidean distance. Therefore, it is the most preferred distance measure in cluster analysis applications (Grimm and Yarnold, 2000).

The definition of similarity and homogeneity differs from analysis to analysis, depending on the objects used in the application. Cluster analysis is based on purely numerical data and the number of clusters is not known beforehand. In addition, it is a method that includes most of the classification methods. The results of this analysis have to be interpreted for a specific purpose and a specific situation. Therefore, the choice of analysis methods is determined depending on the purpose and content of the analysis. It is seen that some clustering methods give better results than other methods in some cases (Jackson, 1983).

The main purpose of this study is to cluster 57 cancer types selected for certain age groups between 1982 and 2016 by age groups and gender at the Australian Institute of Health and Welfare, based on a multivariate data structure consisting of the number of occurrences. The existence of the cluster structure was revealed with the help of the component scores graph and the number of clusters was decided by using the cluster validity indices. Clustering methods are divided into two as hierarchical and non-hierarchical methods.

2. Literature Review

Today, cluster analysis has a wide range of applications in the natural sciences, medical and social sciences. Cluster analysis is a research method based on Linnaeus' work in 1753 on the classification of animals and plants. To apply this analysis, the data type and similarity measure to be used should be defined in detail (Hofman & Jarvis, 1998).

Cormack (1971) determined in his study that the number of articles published on clustering and classification is over a thousand per year. The analysis, in which the science of classification is handled, has been applied in many disciplines, including archaeology, anthropology, agriculture, economics, education, geography, geology, linguistics, market research, genetics, medicine, political science, psychology, psychiatry, and sociology. Larson (1992) found that a company called Clarite's applied cluster analysis on 40 different groups that included zip codes of districts, population density, and census information such as income and age. A classification scheme called PRIZM (potential rating index for the postcode market) was created by giving different names to the mentioned groups. This scheme has been considered very useful for mailing advertisements, radio station formats, and store positioning decisions (Stockburger, 2003).

Greenawalt et al. (2007) collected biopsy specimens from patients who wanted to be treated at, Peter MacCallum Cancer Center and St. Vincent Hospital in Melbourne, Australia during diagnostic endoscopy. The gene expression profile of esophageal cancer was examined by Hierarchical cluster analysis and the full link method was applied using a Pearson similarity measure and Gene Cluster 3.0 and it was visualized with TreeView. Cluster analyzes were performed using the tools included in the R statistical package. The similarities and differences in gene expression observed between each of these groups appear to lead to a clearer understanding of tumorigenesis of the esophagus.

Sonğur (2016) made some evaluations on the countries by revealing how the Organization for Economic Development and Cooperation (OECD) countries are clustered according to health indicators and which OECD countries Turkey is similar to. Data on health indicators used in the study were taken from the “World Bank Database” and hierarchical clustering analysis was applied to these data through the SPSS 20 package program. Turkey is in the fourth cluster with Israel, Mexico, and Chile, whose socio-economic status is relatively lower than other countries.

In the study conducteed by Stundzaite-Barsauskiene et al. (2019), asnwers were sought for questions about the participants′ face perception (FP), self-esteem (SE) and psychosocial wellbeing (PW) by applying facial anthropometry to a total of 90 adult patients and 30 people in the control group after nasal surgery in Vilnius University Hospital “Santaros Klinikos” Plastic and Reconstructive Surgery Department and National Cancer Institute between 2007 and 2017. Here, clustering analysis was applied with the help of the SPSS 25 package program to compare the relationships between facial perception, self-esteem, and psychosocial well-being in patients after rhinoplasty due to trauma, cancer, and aesthetic needs. It was observed that the face measurements were not related to the perception of the whole face in the individuals examined and that there were certain relations between FP, SE, and PW of the patients after aesthetic nose surgery.

Drachenberg et al. (2019) analyzed CTC data from 65 untreated patients with biopsy-confirmed D'Amico-defined intermediate-risk prostate cancer with the center-of-gravity method by comparing the pathology results of radical prostatectomy. As a result of the analysis, patients were placed in three subgroups with different potential risks of aggressive disease.

Demircioğlu and Eşiyok (2019) evaluated the recent health data of 36 countries, which are members of the OECD and EU and identified the countries that showed similarities and tried to determine the place of Turkey among these countries. The k-means method, which is one of the clustering algorithms, was used in the analysis of the data and the values were analyzed with WEKA software. As a result of the analysis, countries were divided into binary, triple, and quadruple clusters. In this study, in terms of the fight against the epidemic, the countries of the world were clustered according to their similarities of health indicators. The positions of 36 countries, including Turkey, relative to each other were evaluated.

İlkin et al. (2020) performed the segmentation of lesion areas in skin images with a k-means clustering algorithm using a dataset consisting of 70 macroscopic melanoma skin cancer images to increase the accuracy of diagnoses made by doctors. Diagnosis of melanoma skin cancer in the early stages is vital because of its impact on recovery prognosis and is largely made by visual assessment of the skin. When the metric results were examined, it was observed that the results obtained were better when the number of centers was selected as 4 in the k-means clustering algorithm. The regions obtained in the study and the regions without lesions were clustered according to their color values.

In the study by Tekin (2020), the effects of the COVID-19 epidemic on countries were comparatively examined with cluster analysis. The variables used in the study are the number of cases, tests, and deaths per million population, the rate of change of some financial indicators, and the level of pre-epidemic health indicators of countries. The variables used in the study were obtained from the World Health Organization, the World Bank, and the OECD. Hierarchical clustering and Ward's method, which are cluster analysis methods, were used in the study. In the study, the similarities and differences of the countries in the context of the mentioned indicators were tried to be revealed and four different clustering analyzes were carried out using the data consisting of four different datasets with the hierarchical clustering method. The seven, five, four, and triple cluster structures that emerged as a result of the study were compared and interpreted.

Yılmaz and Söyük (2020) aimed to group the member countries of the World Bank homogeneously in terms of health risk factors and to rank the countries in these groups in terms of health status indicators. Accordingly, clustering analysis was conducted with the k-means method in R program based on a total of seven risk factors; PM2.5 air pollution, use of basic drinking water services, malnutrition prevalence, smoking prevalence, total alcohol consumption per capita, insufficient physical activity prevalence in adults, and obesity prevalence in adults. As a result of the clustering analysis using the k-means algorithm, 38 of the 122 countries were clustered in the first cluster and 84 in the second cluster.

3. Materials and Method

In the study, single linkage, complete linkage, average linkage, centroid, Ward, and median methods among the hierarchical methods and k-means method, which is the most popular non-hierarchical method, were used.

Hierarchical clustering methods are represented in dendrograms with a tree diagram structure and are applied to a matrix of D=(dij) distances between objects x1, x2, ..., xn, not the objects themselves. Hierarchical cluster analysis methods are descriptive (Mardia et al., 1979). Here, the clusters are in the form of a tree structure called a dendrogram, which represents different degrees of data distribution. The horizontal axis of the dendrogram represents "objects" and the vertical axis represents "distances". The branches of the tree give n-1 connections. Here, the first fork shows the first link, the second fork shows the second link, and this process continues until all the links are joined in the trunk of the tree. Also known as the aggregation approach, dendrograms are formed by leaves at the top of the root of the tree, where each data point acts as a single cluster and these clusters are grouped into a large cluster based on similarity measures at different stages. It can also be formed from the root of the tree to the leaves. In short, the dendrogram reveals successive splits or mergers. Although all objects in such a process will end up in the same group, the grouping process itself that the dendrogram shows is interesting. For example, the width of the edges connecting the branches gives information about the degree of difference between clusters (Dahl & Naes, 2004).

Agglomerative hierarchical clustering algorithms consider each object separately, then combine the two clusters by the distance of the measure between the clusters at each step and continue this process until only one cluster remains (Mathieu, 1991). In other words, these methods start with individual objects. That is, each object or observation creates its clusters. Thus, at the initial stage, there are as many clusters as there are objects. First, the most similar objects are grouped, and these initial groups are combined according to their similarity. In other words, the two closest clusters of individuals are agglomerated (aggregated) into a new cluster. Thus, the number of clusters decreases at each step. In some cases, a third object is combined with the first two objects in a set. As a result, the similarity measure value of the objects decreases, and all individuals are classified into a large cluster (Mardia et al., 1979).

In this study, the average linkage method from hierarchical methods and the k-means method from non-hierarchical methods are briefly explained. In addition, the cophenetic correlation coefficient and cluster validity index are mentioned.

3.1. Average Linkage Method

The average linkage method follows the steps of integrative hierarchical clustering algorithms. This method is used to calculate distances both within groups (within-groups) and between groups (between-groups). First, in the U and V sets, the D={ dik } distance matrix is searched to find the closest (most similar) objects. The closest objects are combined to form the (UV) set. In the third step, the numbers of objects in the (UV) and W clusters are shown as N(UV) and NW, respectively and when the distance between the k number of objects in the W cluster and the i number of objects in the UV cluster is given as dik, the distance between the W and (UV) clusters is given as follows;

$$d_{(UV)W} = \frac{\sum_i \sum_k d_{ik}}{N_{(UV)} N_w} \quad (1)$$

The average linkage method treats the distance between clusters as the average distance between pairs of observations (Mathieu, 1991). This method groups clusters so that the average distance between all objects in the result set is minimal (Toms et al., 2001).

Non-hierarchical methods create a simple distribution of the objects within a set of non-overlapping clusters. The data is distributed into k groups, where each group must contain at least one data object and each object must belong to exactly one group. The number of clusters must be determined by the user. However, not every k value leads to natural clustering, so it is recommended to create the algorithm with different k values many times and choose the k that gives the most meaningful result. It is also possible to automatically decide for k by allowing the computer program to try different values of k and choose the best value in the subject that will meet several optimization criteria. The basic idea in most non-hierarchical methods is to select some different distributions of the data and improve cluster memberships to obtain a better distribution (Hofman & Jarvis, 1998).

The biggest problem encountered in all non-hierarchical clustering processes is how to select cluster seeds. For example, the initial and final sets created by a succession threshold selection depend on the order of the objects and are modified according to the data. However, when cluster seeds are randomly selected, different results are obtained for each cluster of randomly determined seed points. Thus, the researcher has to consider the effect of the process by which the cluster seed is selected on the final results (Hair et al., 1998).

3.2. K-Means Method

The most commonly used non-hierarchical clustering method is the k-means algorithm. In this algorithm, the number of clusters must be known beforehand. The algorithm aims to obtain a clustering structure that is homogeneous within clusters and heterogeneous between clusters, depending on the initial seed. The algorithm steps can be defined as follows;

Step 1: Determine k cluster seed.

Step 2: Observations are assigned to the cluster with the closest seed.

Step 3: Cluster seed is updated by calculating the average vector of the elements assigned to the cluster. If there are seeds closer to observation than the seeds of its cluster, the observation is transferred to the nearest cluster.

Step 4: Repeat step (3) until all transitions stop.

The results of the algorithm depend on the number of clusters and the initial cluster seeds. Cluster validity indices can be used to select the number of clusters. The following approaches have been proposed to determine the initial cluster seed; a. random determination of the k seed b. Taking the first k observation as the seed c.

Taking the k mutually distant observation as the seed d. By using one of the hierarchical clustering algorithms, the obtained cluster centers are taken as the k seeds (Bulut, 2019).

3.3. Cophenetic Correlation Coefficient

The cophenetic correlation coefficient, which was introduced by Sokal and Rohlf in 1962, is used as a criterion to evaluate the degree of fit of the database in classification and the efficiency of various clustering analysis methods. In the studies in the literature, it is seen that this coefficient generally gives the best results in the average linkage method (Carvalho et al., 2019). When researchers have to compare different dendrograms, they may want to know which method causes the least distortion of the information contained in the original similarity matrix. Sokal and Rohlf suggested calculating the correlation between the original similarity matrix and the cophenetic values in the dendrogram, based on the aforementioned distortion, for cluster analysis (Lessig, 1972). The cophenetic value is the height at the left side of a dendrogram and expresses the measure of dissimilarity or distance between two clusters. As this value decreases, the similarity in clusters increases (Bulut, 2018). The cophenetic correlation coefficient is widely used in studies to evaluate the efficiency of hierarchical cluster analysis methods. The fact that this coefficient is high indicates that the dataset used in the application is more successful in clustering analysis. It can be formulated as follows.

$$\mathrm{c} = \frac{\sum_{i<j}(x(i,j)-x)(t(i,j)-t)}{\sqrt{[\sum_{i<j}(x(i,j)-x)^2][\sum_{i<j}(t(i,j)-t)^2]}} \qquad (2)$$

3.4. Cluster Validity Index

For the cluster analysis results to give realistic and reliable interpretations, the number of clusters must be determined well. One of the ways to decide the number of clusters visually is to examinine the results of the dendrogram, which is a hierarchical tree diagram. There are many cluster validity indices (Silhoutte, Dunn etc.) proposed in the literature to determine the number of clusters and the clustering method (Altın, 2021). Although there is no specific method used by researchers to decide on the number of clusters, there are many methods applied. Among these methods, there are also cluster validity indices. Cluster validity indices are classified into two main categories as internal and external indices. The main difference between these index categories is whether external information is used to detect cluster validity (Aggarwal & Reddy, 2014).

4. Application

4.1. Datasets and Characteristics

The registered data of cancer types selected for certain age groups from the Australian Institute of Health and Welfare, according to age groups and gender, between 1982 and 2016 were taken from the website (AİHW, 2021). For 57 determined cancer types, with 33 years of data from 1982 to 2016, the number of occurrences of the variables (characteristics), that is, the frequency of occurrence was examined for the datasets I, II, III, and IV. First dataset was determined to include 0-39, 40-69 and 70 and over age groups in women; the second dataset was determined to include 0-14, 15-39, 40 and over age groups in women; the third dataset was determined to include 0-39, 40-69 and 70 and over age groups in men and the fourth dataset was determined to include 0-14, 15-39, 40 and over age groups in men. In this study, each age group was taken as a characteristic, and cluster analysis applications were made for all the three characteristics. The values determined in terms of the characteristics of the 0-39, 40-69, and 70 and over age groups and the 0-14, 15-39, 40 and age groups constitute the X_1, X_2, X_3 variables, respectively. Considering the sum of the number of appearances of these variables between 1982 and 2016, the values of women and men during these 33 years form the (57x3) dimensional data matrix with a chance sample of N=57 volume and three variables.

Characteristics used in the application; 0-39 and 0-14 Age Group Characteristics (X_1): It shows the frequency of occurrence of a cancer type in males and females for both the 0-39 age group and the 0-14 age group from 1982 to 2016. 40-69 and 15-39 Age Group Characteristics (X_2): This shows the frequency of occurrence of a cancer type in males and females for both the 40-69 age group and the 15-39 age group from 1982 to 2016. Age Group Characteristics 70 and Over and 40 and Over (X_3): This shows the frequency of occurrence of a type of cancer in males and females from 1982 to 2016, for both the 70 and older age group and the 40 and older age group. The datasets used in the analysis; Dataset I: It shows the frequency of occurrence of cancer types in women aged 0-39, 40-69, and 70 and over. Dataset II: This shows the frequency of occurrence of cancer types in women aged 0-14, 15-39, and 40 and over. Dataset III: It shows the frequency of occurrence of cancer types in men in the age groups of 0-39, 40-69, and 70 and over. Dataset IV: It shows the frequency of occurrence of cancer types in 0-14, 15-39 and 40 and over age group men (İncekırık, 2005).

In this study, it was aimed to investigate whether cancer types cluster based on the frequency of occurrence according to age groups and whether there are apparent clusterings. In other words, it was tried to determine what kind of clustering structure the cancer types showed according to the characteristics used. In the study, it was aimed to determine according to which original variables the clusters differ from each other significantly. Cluster analysis methods were applied to the multivariate chance samples obtained for different characteristics in men and women, using the R-3.5.1 package programming language and factoextra, readxl, scales, NbClust, moments, scatterplot3d, psych, cluster, and stats packages, and the results were interpreted (Charrad et al., 2015).

4.2. Analysis of Data with Hierarchical and Non-Hierarchical Cluster Analysis Methods

The datasets used in the study were analyzed according to the single linkage, complete linkage, average linkage, centroid, Ward, median and k-means method. Firstly, the principal component analysis was applied to visually examine the clustering structure in the data and two and three-dimensional principal component score graphs were obtained for all datasets to see the clustering structure in the principal component space. It was desired to see whether these principal component scores formed certain clusters in men and women. Since clusters can be seen visually in these graphs, it can be said that the principal component scores are significantly successful.

To choose the best method among hierarchical clustering analysis methods, the cophenetic correlation coefficient values obtained by applying six different methods to the datasets are given in Table 1. In this table, it is senn that the value of the cophenetic correlation coefficient calculated using the average linkage method is higher than the others. Therefore, it can be said that the most successful method of clustering the datasets used in the analysis is the average linkage method. During the evaluation and interpretation of the study results, the findings of the average linkage and k-means methods were mainly taken into account.

Methods		**Gender**		
	Female		**Male**	
	Dataset I	Dataset II	Dataset **III**	Dataset **IV**
Single Linkage	0.9591942	0.9645086	0.9608571	0.9450074
Complate Linkage	0.9598888	0.9526376	0.9657261	0.9641139
Average Linkage	***0.9796056***	***0.9731935***	***0.9733951***	***0.9733661***
Ward Linkage	0.7437232	0.7597301	0.7879299	0.7785894
Centroid	0.9707725	0.9691959	0.9678371	0.9696039
Median	0.977374	0.9701301	0.968819	0.9652075

Table 1. Cophenetic Correlation Coefficient for Hierarchical Methods

In the study, two- and three-dimensional graphics created with cluster validity indices were used to determine the number of clusters. The 26 internal cluster validity index values included in the NbClust package in the R 3.5.1 program are calculated for all datasets and given in Table 2 and Table 3. When the values in the table are examined, it is seen that the number of clusters suitable for both methods is mostly 3. Therefore, it was decided that the appropriate number of clusters should be 3.

Average Linkage Method					**k--Means Method**			
İndexes	Dataset I		Dataset II		Dataset I		Dataset II	
	Number of Clusters	Value İndex	Number of Clusters	Value İndex	Number of Clusters	Value İndex	Number of Clusters	Value İndex
KL	*3*	*78,5361*	*3*	*74.5774*	4	78.708	4	8.2384
CH	11	445,3301	11	1518.296	4	278.858	4	417.9046
Hartigan	*3*	*137,5229*	*3*	*183.0866*	*3*	109.8437	*3*	73.4982
CCC	11	9,9372	11	7.1418	4	6.5051	4	2.3235
Scott	*3*	*90,8339*	*3*	*92.2853*	*3*	150.0353	*3*	76.0276
Marriot	*3*	*6,23*	*3*	*2.249694*	*3*	2.205222E+29	*3*	6.528529
TrCovW	*3*	*1,71*	*3*	*1.73215*	*3*	2.666605E+20	*3*	1.024476
TraceW	*3*	*282E+10*	*3*	*5687289*	*3*	25537223136	*3*	33705491292
Friedman	12	3025486	9	272.4953	*3*	70.4985	*3*	21.117
Rubin	9	-352496	9	-111.3126	4	-5.4735	4	-7.2376
Cindex	*3*	*0.1904*	2	0.1843	9	0.0561	9	0.0477
DB	2	0.1057	2	0.1118	*3*	0.3787	*3*	0.3053
Silhouette	2	0.8956	2	0.887	*3*	0.8062	*3*	0.8258
Duda	*3*	*17082*	*3*	*5.0539*	2	0.8999	2	1.7606
PseudoT2	5	505743	6	0	4	-16.7864	4	-3.2025
Beale	*3*	*-0.5294*	*3*	*-1.0242*	2	0.1516	2	-0.5884
Ratkowsky	*3*	*0.4996*	*3*	*0.34*	2	0.5372	2	0.3642
Ball	*3*	*17E+10*	*3*	*3,294735*	3	1,69876	*3*	2,4325
PtBiserial	2	0.8234	2	0.8013	3	0.7806	*3*	0.794
Frey	NA	NA	NA	NA	NA	NA	NA	NA
McClain	2	0.0035	2	0.0037	*3*	0.0276	*3*	0.0267
Dunn	2	13474	2	0.9718	*3*	0.5155	*3*	0.584
Hubert	0	0	0	0	0	0	0	0
SDindex	2	0,0001	4	1e-04	4	5e-04	4	4e-04
Dindex	0	0	0	0	0	0	0	0
SDbw	12	0.0025	12	0.0011	9	0. 1235	9	0.1073

Table 2. Average Linkage and k-Means Method Cluster Validity Index Values (Female)

Average Linkage Method					k--Means Method			
İndexes	Dataset III		Dataset IV		Dataset III		Dataset IV	
	Number of Clusters	Value İndex	Number of Clusters	Value İndex	Number of Clusters	Value İndex	Number of Clusters	Value İndex
KL	8	80.0312	3	39,4999	4	354,013	4	54,0034
CH	8	445.0442	11	1657,691	4	348,035	9	1060,375
Hartigan	5	138.3532	4	190,4617	4	158,8728	4	222,9823
CCC	8	4.6071	11	6,5411	4	3,2762	4	4,0513
Scott	6	105.8093	5	96,2678	4	100,6247	4	102,2918
Marriot	3	97.03291	3	3,74E+27	4	1,18E+29	4	4,80E+27
TrCovW	3	4.890673	3	4,33E+21	3	2,05E+20	3	1,53E+21
TraceW	3	4.250814	3	8,52E+10	4	2,04E+10	4	4,12E+10
Friedman	12	97.6063	10	402,7591	9	79,1122	9	156,5588
Rubin	6	-13.2178	10	-155,64	4	-15,5251	9	-109,071
Cindex	6	0.1882	2	0,1926	3	0,0602	3	0,0534
DB	2	0.1168	2	0,1148	4	0,4566	4	0,3771
Silhouette	2	0.8793	2	0,8836	2	0,8173	2	0,8271
Duda	3	0.2602	3	0,102	2	1,3257	2	1,4364
PseudoT2	4	0	5	13,5263	2	-10,319	2	-12,7609
Beale	3	2.0161	5	2,0934	2	-0,4101	2	-0,5071
Ratkowsky	3	0.4381	3	0,2824	2	0,4675	2	0,3018
Ball	3	2,533894	3	4,99	3	1,30	3	2,51E+10
PtBiserial	3	0.8134	3	0,8131	2	0,8019	2	0,8015
Frey	NA	NA	NA	NA	6	1,4954	6	2,2588
McClain	2	0.0039	2	0,0037	2	0,0244	2	0,023
Dunn	2	1.0124	2	1,0128	2	0,1816	2	0,1614
Hubert	0	0	0	0	0	0	0	0
SDindex	3	1e-04	4	0,0001	4	0,0003	4	0,0002
Dindex	0	0	0	0	0	0	0	0
SDbw	12	0.0034	12	0,0008	9	0,0527	9	0,024

Table 3. Average Linkage and k-Means Method Cluster Validity Index Values (Male)

The results obtained when the average linkage method was applied to all datasets by taking the number of clusters as 3 are given in Figure 1, Table 4, and Table 5.

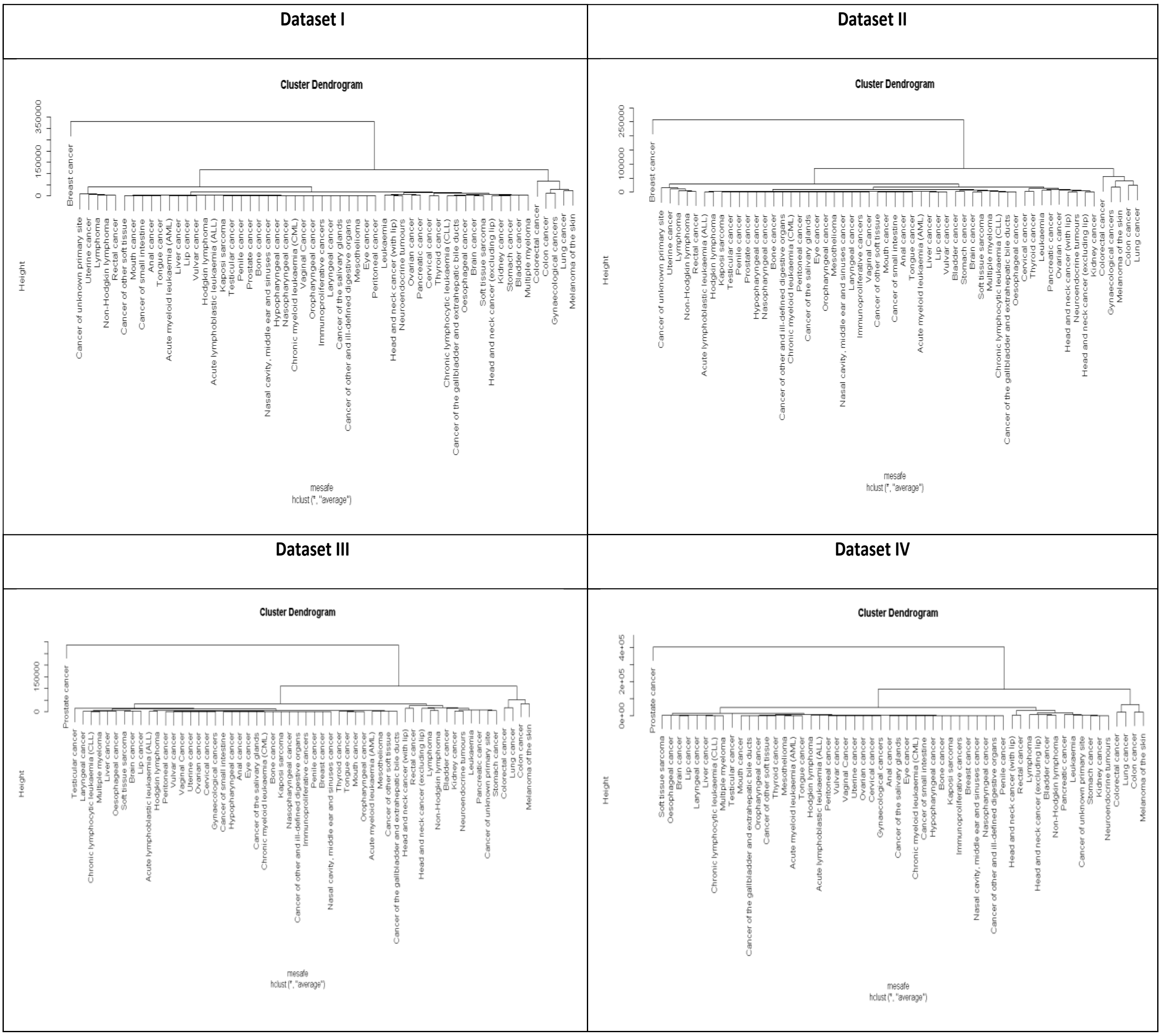
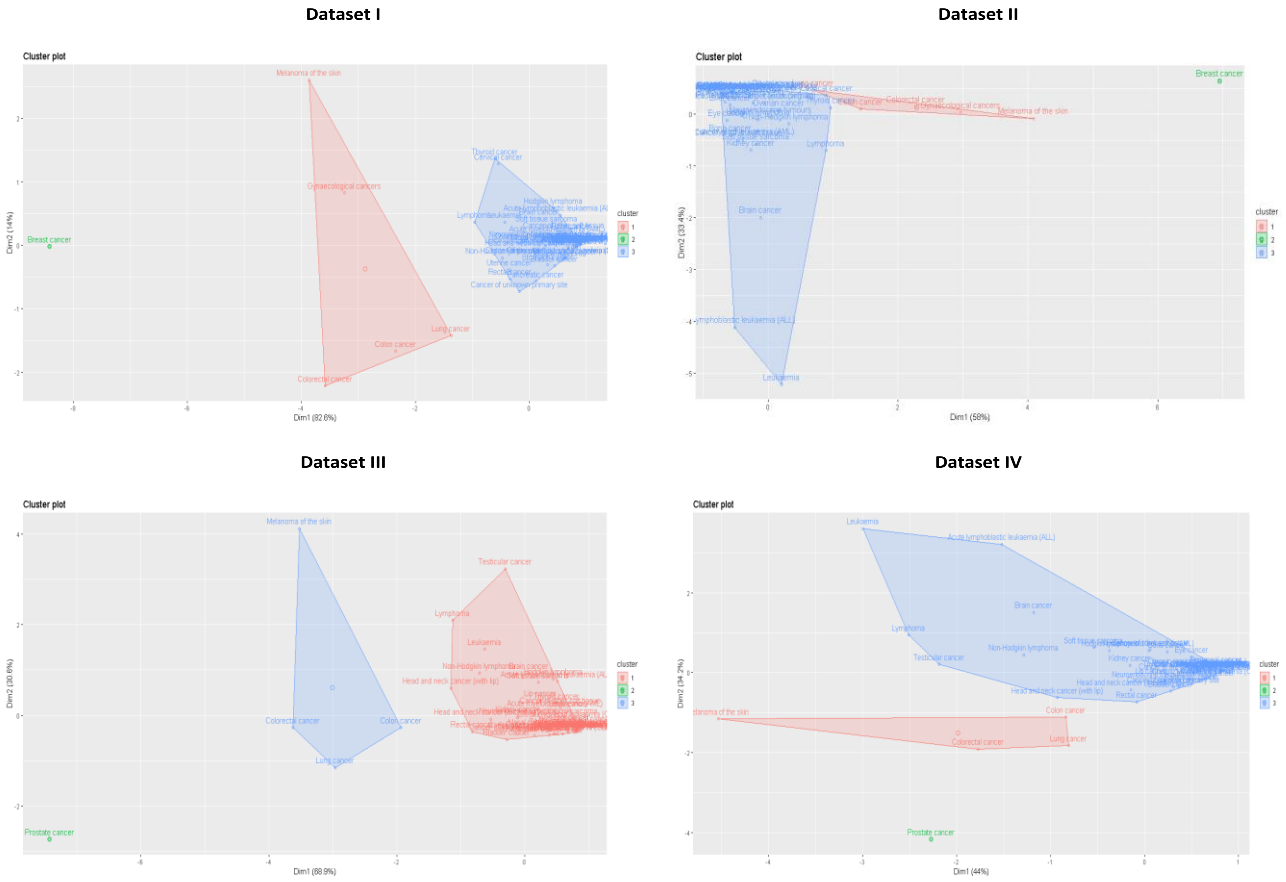


Figure 1. Average Linkage Method Dendrograms

When the dendrograms in Figure 1 are examined, it is seen that most of the cancer types are gathered in the same cluster. In addition, it is seen that breast cancer in women and prostate cancer in men are far from the other cancers. It is known that the cancer types mentioned are the most common cancers in women and men. Another remarkable point is that the incidence of gynecological cancer is high in women, but low in men. It is seen that colon, colorectal, gynecological, lung, and skin melanoma cancers, which have a very high incidence, have the same degree of membership in the same cluster.

When the clustering results for women's datasets (I. and II.) are examined in Table 4, it is seen that breast cancer alone has a membership in the second cluster, 5 cancer types in the first cluster, and the other 51 cancer types in the third cluster. Similarly, when the clustering results created for the male datasets (III. and IV.) by both methods are examined in Table 5, is is seen that prostate cancer in men is in a separate cluster.

The results of the analysis of the k-means method applied to all datasets by taking k=3 are shown in Figure 2, Table 4, and Table 5.


Dataset I

Dataset II

Dataset III

Dataset IV

Figure 2. k-Means Method Principal Component Scores Plots

When the principal component scores graphs in Figure 2 are examined, it is seen that breast cancer in women behaves as an observation that deviates quite significantly from other cancers and far from other cancers, and in the second cluster alone, 5 cancer types such as colon, lung, melanoma of the skin, gynecological, colorectal cancers are also quite far from the others. They have membership in the first cluster, acting like a slingshot and clearly distinguishing themselves from breast cancer. It is observed that the remaining 51 cancer types are in the third cluster. In men, prostate cancer alone is in the second cluster, 5 cancer types are in the third cluster, and 52 other cancer types are in the first cluster. Table 1 and Table 2 show that most of the cancer types are in cluster 3 for women and cluster 1 for men. When the clusters in these tables are examined, it is seen that there are similar results in both methods.

In the study, four different datasets are divided into three clusters considering age groups for women and men. Cancers in these clusters are prominently displayed in the principal component score graphs in Figure 2 and the dendrograms in Figure 1.

Clusters	Cancer Types
Cluster 1	Colon Cancer, Colorectal Cancer, Gynaecological Cancers, Lung Cancer, Melanoma Of The Skin
Cluster 2	Breast Cancer
Cluster 3	Acute Lymphoblastic Leukaemia (ALL), Acute Myeloid Leukaemia (AML), Anal Cancer, Bladder Cancer, Bone Cancer, Brain Cancer, Cancer Of Other And İll-Defined Digestive Organs, Cancer Of Other Soft Tissue, Cancer Of Small İntestine, Cancer Of The Gallbladder And Extrahepatic Bile Ducts, Cancer Of The Salivary Glands, Cancer Of Unknown Primary Site, Cervical Cancer, Chronic Lymphocytic Leukaemia (CLL), Chronic Myeloid Leukaemia (CML), Eye Cancer, Head And Neck Cancer (Excluding Lip), Head And Neck Cancer (With Lip), Hodgkin Lymphoma, Hypopharyngeal Cancer, Immunoproliferative Cancers, Kaposi Sarcoma, Kidney Cancer, Laryngeal Cancer, Leukaemia, Lip Cancer, Liver Cancer, Lymphoma, Melanoma Of The Skin, Mesothelioma, Mouth Cancer, Multiple Myeloma, Nasal Cavity, Middle Ear And Sinuses Cancer, Nasopharyngeal Cancer, Neuroendocrine Tumours, Non-Hodgkin Lymphoma, Oesophageal Cancer, Oropharyngeal Cancer, Ovarian Cancer, Pancreatic Cancer, Penile Cancer, Peritoneal Cancer, Prostate Cancer, Rectal Cancer, Soft Tissue Sarcoma, Stomach Cancer, Testicular Cancer, Thyroid Cancer, Tongue Cancer, Uterine Cancer, Vaginal Cancer, Vulvar Cancer

Table 4. Clusters Obtained by Average linkage and k-Means Method (Female)

Clusters	Cancer Types
Cluster 1	Acute Lymphoblastic Leukaemia (ALL), Acute Myeloid Leukaemia (AML), Anal Cancer, Bladder Cancer, Bone Cancer, Brain Cancer, Breast Cancer, Cancer Of Other And İll-Defined Digestive Organs, Cancer Of Other Soft Tissue, Cancer Of Small İntestine, Cancer Of The Gallbladder And Extrahepatic Bile Ducts, Cancer Of The Salivary Glands, Cancer Of Unknown Primary Site, Cervical Cancer, Chronic Lymphocytic Leukaemia (CLL), Chronic Myeloid Leukaemia (CML), Eye Cancer, Gynaecological Cancers, Head And Neck Cancer (Excluding Lip), Head And Neck Cancer (With Lip), Hodgkin Lymphoma, Hypopharyngeal Cancer, Immunoproliferative Cancers, Kaposi Sarcoma, Kidney Cancer, Laryngeal Cancer, Leukaemia, Lip Cancer, Liver Cancer, Lymphoma, Mesothelioma, Mouth Cancer, Multiple Myeloma, Nasal Cavity, Middle Ear And Sinuses Cancer, Nasopharyngeal Cancer, Neuroendocrine Tumours, Non-Hodgkin Lymphoma, Oesophageal Cancer, Oropharyngeal Cancer, Ovarian Cancer, Pancreatic Cancer, Penile Cancer, Peritoneal Cancer, Rectal Cancer, Soft Tissue Sarcoma, Stomach Cancer, Testicular Cancer, Thyroid Cancer, Tongue Cancer, Uterine Cancer, Vaginal Cancer, Vulvar Cancer
Cluster 2	Prostate Cancer
Cluster 3	Colon Cancer, Colorectal Cancer, Lung Cancer, Melanoma Of The Skin

Table 5. Clusters Obtained by Average linkage and k-Means Method (Male)

According to the common results obtained by using k-means and average linkage methods in Table 4 and Table 5, the deviating value of breast cancer in women creates a separate single-member cluster and shows a very different characteristic compared to other cancer types in terms of age groups. Colon cancer, Colorectal cancer, Lung cancer, Melanoma of the skin, Gynaecological cancers observations are similar in terms of age groups. The remianing 51 cancer type observations formed a separate cluster. In males, on the other hand, prostate cancer, as a deviating value, forms a separate single-member cluster and shows a very different characteristic compared to the other cancer types in terms of age groups. Colon cancer, Colorectal cancer, Lung cancer, Melanoma of the skin observations are similar in terms of age groups. The remaining 52 cancer-type observations formed a separate cluster.

To determine whether the age group characteristics are normally distributed in all datasets, skewness, and kurtosis values were obtained by using the "skewness" and "kurtosis" functions in the "moments" package in the R 3.5.1 program (Altın, 2021). When these values were examined, it was seen that the variables did not show a normal distribution. According to the cluster analysis, considering the age groups, cancer types were gathered in three separate clusters. However, the "Kruskal-Wallis Test" was conducted both to increase the reliability of the cluster analysis and to reveal over which age groups the clusters differ. The test results obtained by using the Kruskall-Walls test for the age group characteristics that do not meet the normality assumption are given in Table 7. Here, the fact that there is a significant difference between the clusters in the majority of age groups reveals the reliability of the clustering.

When Table 6 is examined to determine whether there is a difference between the mean values of the cancers in clusters 1, 2, and 3 in terms of age group characteristics, it can be said that there is a significant difference between clusters in terms of all age group characteristics. The breast cancer having a membership in cluster 2 in the datasets I andI I in women and the prostat cancer having a membership in cluster 2 in the datasets III andI V in men, can be said to appear more in the age group of 40 years and older. The clusters obtained differ more clearly from each other in terms of the age groups of 40-69 and 70 years and older and the age group of 40 years and older. The most striking finding here is the high incidence of cancer in individuals aged 40 and over. In addition, it is seen that the incidence of cancer in individuals aged 0-14 is quite low.

When Table 7 is examined, it is seen that there is no statistically significant difference between the 3 clusters in terms of only the X1 Variable (0-39 and 0-14 age group characteristics). According to these results, it was determined that the other age group variables successfully separated at least two clusters in all of the datasets. When the frequency values in the datasets are examined, and given the low number of cancer diseases in the 0-14 age group, and the narrow age range, this result is quite striking. Since the probability value of the 0-14 age group variable is high in the datasets II and IV, it can be said that this age group failed to separate at least two clusters.

Clusters	**Female**					
	Dataset I			Dataset II		
	0-39	40-69	70+	0-14	15-39	40+
Cluster 1	23737	252028	103660	6	23731	355688
Cluster 2	10409	66396,2	62449,2	195,6	10213,4	128845,4
Cluster 3	1557,412	7268,333	6866,922	242,3137	1315,098	14135,25
Clusters	**Male**					
	Dataset III			Dataset IV		
	0-39	40-69	70+	0-14	15-39	40+
Cluster 2	171	218820	215322	7	164	434142
Cluster 3	6618,6	84931,6	80802	120,4	6498,2	165733,6
Cluster 1	1802,231	10657,37	8054,596	303,8846	1498,346	18711,96

Table 6. Averages of Frequencies of Age Group Variables

Clusters	**Female**					
	Dataset I			Dataset II		
	0-39	40-69	70+	0-14	15-39	40+
Chi-Square	10,893	15,857	15,857	2,472	11,721	15,857
df	2	2	2	2	2	2
Asymp. Sig.	,004	,000	,000	,291	,003	,000
Clusters	**Male**					
	Dataset III			**Dataset IV**		
	0-39	**40-69**	**70+**	**0-14**	**15-39**	**40+**
Chi-Square	5,625	13,482	13,482	1,293	5,982	13,482
df	2	2	2	2	2	2
Asymp. Sig.	,060	,001	,001	,524	,050	,001

Table 7. Kruskal Wallis Test Results

5. Conclusion and Suggestions

In this study, in which the cophenetic correlation coefficient that we use in choosing the best method with cluster analysis methods and the cluster validity indices that we take into account in deciding the number of clusters were used, it was tried to reveal how useful it could be to determine the clustering structure of cancer types according to age groups in women and men through the applied analysis.

In the study, the variables that best distinguish the clusters of cancer types selected for certain age groups at the Australian Institute of Health and Welfare, that is, the variables with the highest discriminating power, were revealed. At the same time, attention was paid to the fact that the variables contributed significantly to the cluster analysis and were significant. With this study, it was examined in terms of which original variables the clusters differed from each other more clearly. It was investigated whether the determined cancer types clustered based on the frequency of occurrence according to age groups and whether there were significant clusters. In other words, it was revealed how cancer types exhibited a clustering structure according to age group characteristics. In the study, 57 cancer types were clustered in a way that they were similar to each other by using the determined age group characteristics and 3 different clustering results were obtained.

First of all, principal component score graphs were used to see the clustering structure in the datasets. When we examine these graphs for all datasets, the first thing that draws attention is the cancer types that are out of the cluster and can be seen to be quite far from other cancer types. As a result, it was visually revealed that there are 3 clustering structures in all of the datasets created according to three different characteristics. In addition, another important point to note is that this analysis shows that cancer types that occur frequently in women and men are clustered in these three different age groups. Here, it can be said that the dimension reduction was successful and it was concluded that the graphical representations of the groups in the two-dimensional principle component space were meaningful. Following the principal component analysis, the results obtained using average linkage and k-means methods in terms of 3 age group characteristics on 57 cancer types were examined and interpreted.

The number of clusters for women and men was taken as k=3. Cluster validity indices were used to determine the number of clusters. After determining the clusters formed by cancer types by applying the average linkage method to all datasets, finally, 57 cancer types were brought together in terms of these datasets to form clusters. Accordingly, when the cluster membership results, which show in which cluster the cancer types whose distances are calculated are found, are examined, it is seen that for the datasets (I and II), there are 5 cancer types in the first cluster, 1 cancer type in the second cluster, and 51 cancer types in the third cluster; for the datasets (III and IV), there are 52 cancer types in the first cluster, 1 cancer type in the second cluster, and 4 cancer types in the third cluster. Obtained non-hierarchical clustering method results can be compared with the results of hierarchical clustering methods. In this study, the results of the average linkage method and the k-means method were found to be similar.

As a result of both clustering methods, cancer types were divided into 3 clusters in women and men. In the second cluster of these 3 clusters the breast cancer in women and prostate cancer in men take place with a single membership degree. In the first cluster, there are colon, colorectal, gynecological, lung cancers Melanoma of The Skin in women and in the third cluster, there is Colon Cancer, Colorectal Cancer, Lung Cancer, Melanoma of The Skin in men. The striking point in this cluster is that gynecological cancers are seen much more frequently in women than in men. There are 51 types of cancer in cluster 3 in women and 52 in cluster 1 in men. The most important result of the analysis is that the most common cancer types in women and men are breast and prostate, respectively, and these cancers occur in the age group of forty years and above. It can be said that these cancers are very much affected by old age, depending on gender. Similarly, it can be said that cancers in cluster 1 in women and cluster 3 in men occur in elderly age groups and gynecological cancer is common in women. After breast and prostate cancers, the most common cancers in men and women alike are colon, colorectal, lung, and skin melanoma. In addition, it can be seen that gynecological cancers occur frequently in women.

In this cluster analysis study conducted according to the determined age groups, it was determined that cancer types were collected in 3 clusters. In addition, the "Kruskal-Wallis Test" was applied to reveal the age groups that provide the difference in the clustering of cancers and to reflect the difference between clusters. According to the test results, it was found that the clusters differed significantly in terms of age groups. In all the datasets, the age group with no significant difference between clusters was determined as "0-14". In addition, in the third dataset for males, the age group with no significant difference between clusters is "0-39". The fact that other age group variables show a significant difference between clusters shows that these variables are a very important determinant in the clustering of cancer types. We can say that the 0-14 age group variable does not have a determining role in the clustering of cancer types. It can be said that this situation is due to the remarkably low incidence of cancer in the mentioned age group.

As a result, it can be said that the study in which cancer types in men and women were clustered according to age groups yielded the expected clusters. While prostate cancer is the most common disease especially in men, it is breast cancer in women, and these cancers were clustered significantly in the results of the study. With this study, individuals can be made aware of cancer diseases that occur in age groups depending on gender. Cancer, which is known as the most common disease in the world today, is an increasing health problem worldwide. In this regard, it is of great importance for individuals to recognize the types of cancer and shape their lifestyles in all aspects to be protected from this disease. On the other hand, the age groups and clusters created in this study can be used in research on "early diagnosis", which is a very important issue in cancer. In addition, by determining the characteristics of different age groups in men and women, comparisons can be made by obtaining different clusters on more datasets.

Cluster analysis studies can be carried out with separate datasets regarding cancer types in Turkey according to regions. Thus, by comparing the classifications obtained for each region with other regions, cancers that occur in Turkey can be interpreted depending on gender and age group. In this way, more support can be given to studies in the field of medicine in Turkey. It can be said that the study made an important contribution to the use of multivariate techniques in the area of medicine.

References

Aggarwal, C. C., & Reddy, C. K. (2014). Data Clustering: Algorithms and Applications. Chapman&Hall/CRC Data Mining and Knowledge Discovery Series, London.

Alhamed, A., Lakshmivarahan, S., & Stensrud, D. J. (2002). Cluster analysis of multimodel ensemble data from SAMEX. Monthly weather review, 130(2), 226-256.

Altın, E. (2021). Türkiye′ de İller Bazında Ulaşım Faaliyetlerinin Gelişim Durumunun Kümeleme Analizi ile Belirlenmesi (Yüksek Lisans Tezi), Manisa Celal Bayar Üniversitesi Sosyal Bilimler Enstitüsü, Manisa.

Australian Institute of Health and Walfare (2021). Cancer Data in Avustralia. Retrieved from https://www.aihw.gov.au/reports/cancer/cancer-data-in-australia/data?page=1, (Accessed Date:: 12.03.2021).

Bulut, H. (2018). R Uygulamaları ile Çok Değişkenli İstatistiksel Yöntemler, Ankara: Nobel Akademik Yayıncılık Eğitim Danışmanlık Tic. Ltd. Şti.

Bulut, H. (2019). Türkiye'deki İllerin Yaşam Endekslerine Göre Kümelenmesi. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi, 23(1), 74-82.

Carvalho, P. R., Munita, C. S., & Lapolli, A. L. (2019). Validity Studies Among Hierarchical Methods of Cluster Analysis Using Cophenetic Correlation Coefficient. Brazilian Journal of Radiation Sciences, 7(2A).

Charrad, M., Ghazzali, N., Boiteau, V., & Niknafs, A. (2015). NbClust: Determining the Best Number of Clusters in a Data Set. R Package, Version 3.0.

Cormack, R. M. (1971). A Review of Classification. Journal of the Royal Statistical Society: Series A (General), 134(3), 321-353.

Dahl, T., & Næs, T. (2004). Outlier and Group Detection in Sensory Panels Using Hierarchical Cluster Analysis with the Procrustes Distance. Food Quality and Preference, 15(3), 195-208.

Demircioğlu, M., & Eşiyok, S. (2020). Covid–19 Salgını ile Mücadelede Kümeleme Analizi ile Ülkelerin Sınıflandırılması. İstanbul Ticaret Üniversitesi Sosyal Bilimler Dergisi, 19(37), 369-389.

Drachenberg, D., Awe, J. A., Rangel Pozzo, A., Saranchuk, J., & Mai, S. (2019). Advancing Risk Assessment of İntermediate Risk Prostate Cancer Patients. Cancers, 11(6), 855.

Greenawalt, D. M., Duong, C., Smyth, G. K., Ciavarella, M. L., Thompson, N. J., Tiang, T., ... & Phillips, W. A. (2007). Gene Expression Profiling of Esophageal Cancer: Comparative Analysis of Barrett's Esophagus, Adenocarcinoma, and Squamous Cell Carcinoma. International Journal Of Cancer, 120(9), 1914-1921.

Grimm, L. G., & Yarnold, P. R. (2000). Reading and Understanding More Multivariate Statistics. American Psychological Association.

Hair, J. F., Anderson, R. E., Tatham, R. L., & Black, W. C. (1999). Multivariate Data Analysis, Fifth Edition, Prentice Hall International Editions, New Jersey.

Hofman, I., & Jarvis, R. (1998). Robust and Efficient Cluster Analysis Using a Shared Near Neighbours Approach. In Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No. 98EX170) (Vol. 1, pp. 243-247). IEEE.

İlkin, S., Aytar, O., Gençtürk, T. H., & Şahin, S. (2020). Dermoskopik Görüntülerde Lezyon Bölütleme İşlemlerinde K-Ortalama Kümeleme Algoritmasının Kullanımı. Gazi Üniversitesi Fen Bilimleri Dergisi Part C: Tasarım Ve Teknoloji, 8(1), 182-191.

İncekırık, A. (2005). Çok Değişkenli İstatistiksel Bir Boyut İndirgeme Yöntemi Olarak Kümeleme Analizi ve Bir Uygulama. (Yüksek Lisans Tezi). Dokuz Eylül Üniversitesi Sosyal Bilimler Enstitüsü, İzmir.

Jackson, B.B. (1983). Multivariate Data Analysis, Richard D. Irwın, Inc., Homewood, Illinois.

Johnson, R. A., & Wichern, D. W. (1998). Applied Multivariate Statistical Analysis (Vol. 4, No. 8). Upper Saddle River, Nj: Prentice Hall.

Kent, J. T., Bibby, J., & Mardia, K. V. (1979). Multivariate Analysis. Academic Press, London.

Lessig, V. P. (1972). Comparing Cluster Analyses with Cophenetic Correlation. Journal of Marketing Research, 9(1), 82-84.

Mathieu, Richard G. (1991). Proceedings of the Portland. International Conference on Management of Engineering and Technology, Portland.

Na, L., Wensheng, G., Kexiong, T., & Xiaoning, W. (2002). Application of a Combinatorial Neural Network Model Based On Cluster Analysis in Transformer Fault Diagnosis. In 2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering. TENCOM'02. Proceedings. (Vol. 3, pp. 1873-1876). IEEE.

Seber, G. A. (2009). Multivariate Observations (Vol. 252). John Wiley & Sons.

Sharma, S. (1996). Applied Multivariate Techniques. John Wiley and Sons, Inc., New York.

Sonğur, C. (2016). Sağlık Göstergelerine Göre Ekonomik Kalkınma ve İşbirliği Örgütü Ülkelerinin Kümeleme Analizi. SGD-Sosyal Güvenlik Dergisi, 6(1), 197-224.

Stockburger, D. W. (1998). Multivariate Statistics: Concepts, Models, and Applications. David W. Stockburger.

Stundzaite-Barsauskiene, G., Tutkuviene, J., Barkus, A., Jakimaviciene, E. M., Gibaviciene, J., Jakutis, N., ... & Dadoniene, J. (2019). Facial perception, Self-Esteem and Psychosocial Well-Being in Patients After Nasal Surgery Due to Trauma, Cancer and Aesthetic Needs (Cluster Analysis of Multiple İnterrelations). Annals of Human Biology, 46(7-8), 537-552.

Tekin, B. (2020). Covid-19 Pandemisi Döneminde Ülkelerin Covid-19, Sağlık ve Finansal Göstergeler Bağlamında Sınıflandırılması: Hiyerarşik Kümeleme Analizi Yöntemi. Finans Ekonomi ve Sosyal Araştırmalar Dergisi, 5(2), 336-349.

Toms, M. L., Cummings-Hill, M. A., Curry, D. G., & Cone, S. M. (2001). Using Cluster Analysis for Deriving Menu Structures for Automotive Mobile Multimedia Applications. SAE transactions, 265-271.

Yılmaz, F., & Söyük, S. (2020). Sağlık Risk Faktörlerine Göre Ülkelerin Kümelenmesi ve Çok Kriterli Karar Verme Teknikleriyle Sağlık Durumu Göstergelerinin Analizi. Sosyal Güvence, (17), 283-320.