



**REPUBLIC OF TURKEY
ADANA ALPARSLAN TÜRKER SCIENCE AND TECHNOLOGY
UNIVERSITY**

**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF INDUSTRIAL ENGINEERING**

**EXAMINATION OF PRINCIPAL COMPONENT
ANALYSIS ON HIERARCHICAL CLUSTERING
METHODS IN TERMS OF DIMENSION REDUCTION**

**YAĞMUR SONAY
MASTER OF SCIENCE**



**REPUBLIC OF TURKEY
ADANA ALPARSLAN TÜRKESİ SCIENCE AND TECHNOLOGY
UNIVERSITY**

**GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
DEPARTMENT OF INDUSTRIAL ENGINEERING**

**EXAMINATION OF PRINCIPAL COMPONENT ANALYSIS
ON HIERARCHICAL CLUSTERING METHODS IN TERMS
OF DIMENSION REDUCTION**

**YAĞMUR SONAY
MASTER OF SCIENCE**

**SUPERVISOR
PROF. DR. TOLUNAY GÖÇKEN**

ADANA 2022

**EXAMINATION OF PRINCIPAL COMPONENT ANALYSIS ON HIERARCHICAL
CLUSTERING METHODS IN TERMS OF DIMENSION REDUCTION**

submitted by **Yağmur SONAY** in partial fulfillment of the requirements for the degree of
**Master of Science in Department of Industrial Engineering, Adana Alparslan Türkeş
Science and Technology University** by,

Assoc. Prof. Dr. Ahmet
BEYÇİOĞLU
Graduate School Director

Prof. Dr. Neşe YALÇIN
Department Chair

Prof. Dr. Tolunay GÖÇKEN
Supervisor

Thesis Committee

Prof. Dr. Tolunay GÖÇKEN
Adana Alparslan Türkeş Science and Technology University

Assoc.Prof. Dr. Gülsün
NAKİBOĞLU
Çukurova Üniversitesi

Assoc. Prof. Dr. Ayşe Tuğba
DOSDOĞRU
Adana Alparslan Türkeş Science
and Technology University

Thesis Defense Date

30/06/2022



I hereby declare that all information in this thesis has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all information that is not original to this work.

[Signature]

Yağmur SONAY

ABSTRACT

EXAMINATION OF PRINCIPAL COMPONENT ANALYSIS ON HIERARCHICAL CLUSTERING METHODS IN TERMS OF DIMENSION REDUCTION

Yağmur SONAY

Department of Industrial Engineering

Supervisor: Prof. Dr. Tolunay GÖÇKEN

June 2022, 118 pages

In the light of technological and scientific developments, with the introduction of concepts such as artificial intelligence and machine learning into our lives, data production in the world is increasing exponentially. Parallel to this, data has become a precious mine that can be bought, processed, converted into meaningful and important information, and sold. This rapid and huge increase causes many difficulties in the process of transforming data into information. One of these difficulties is that the data to be analyzed is too large or too dimensional. These features of the data make it difficult for the applied analysis technique to work correctly and comfortably. To cope with these problems, there are many methods applied in the pre-processing stages of data. This thesis study purposes examine the effect of the Principal Component Analysis method on Hierarchical Clustering techniques in terms of dimensionality reduction in high-dimensional data sets. The study was carried out using Principal Component Analysis and Hierarchical Clustering methods on the data sets with 22, 38, and 46 variables, created with 2020 data from the United Nations data platform. The results obtained from the analysis were compared and interpreted using tanglegrams and some similarity coefficients. The results of the study showed that the Principal Component Analysis method had positive effects on hierarchical clustering results and dendrograms despite low correlation and outliers.

Keywords: hierarchical cluster analysis, principal component analysis, cophenetic correlation coefficient, F_M index, Baker's gamma correlation coefficient, tanglegram.

ÖZET

TEMEL BİLEŞEN ANALİZİNİN HİYERARŞİK KÜMELEME YÖNTEMLERİ ÜZERİNDE BOYUT İNDİRGEME BAKIMINDAN İNCELENMESİ

Yağmur SONAY

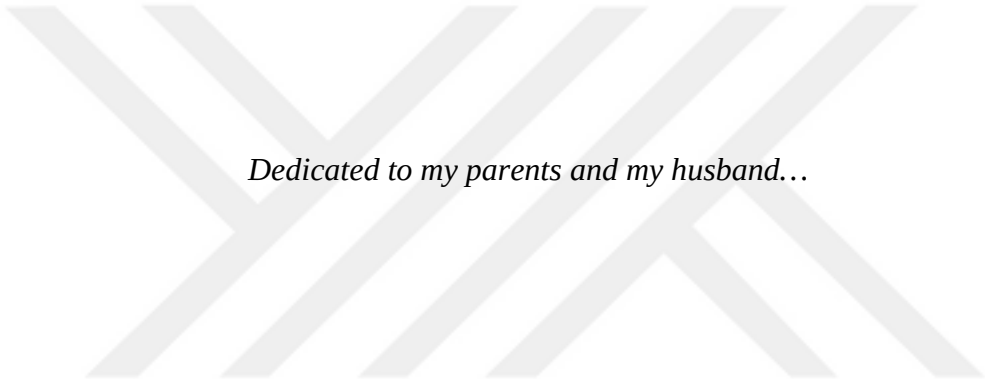
Endüstri Mühendisliği Anabilim Dalı

Danışman: Prof. Dr. Tolunay GÖÇKEN

Haziran 2022, 118 sayfa

Teknolojik ve bilimsel gelişmeler ışığında, yapay zeka ve makine öğrenimi gibi kavramların da hayatımıza girmesiyle birlikte, dünyada veri üretimi ciddi bir şekilde katlanarak artmaktadır. Buna paralel olarak veri; alınıp, işlenip, anlamlı ve önemli bilgiye çevrilerek satılabilen değerli bir maden haline gelmiştir. Verinin, bu denli hızlı ve çok fazla artışı, bilgiye dönüştürülme sürecinde pek çok zorluğa neden olmaktadır. Bu zorlukların başında, analiz edilecek verilerin çok yüksek boyutlarda ya da çok büyük miktarlarda olması gelmektedir. Verinin bu özellikleri uygulanan analiz tekniğinin doğru ve rahat çalışmasını zorlaştırır. Bu olumsuzluklarla başa çıkabilmek için verinin işlenme öncesi aşamalarında uygulanan pek çok yöntem mevcuttur. Bu tez çalışmasının amacı, yüksek boyutlu veri setlerinde Temel Bileşen Analizi yönteminin Hiyerarşik Kümeleme teknikleri üzerindeki boyut indirgeme etkisinin incelenmesidir. Çalışma, Birleşmiş Milletler veri platformu üzerinden 2020 verileriyle oluşturulan, 22, 38 ve 46 değişkenli veri setleri üzerinde, Temel Bileşen Analizi ve Hiyerarşik Kümeleme yöntemleri kullanılarak gerçekleştirilmiştir. Analizlerden elde edilen sonuçlar tanglegramlar ve bazı benzerlik katsayıları kullanılarak karşılaştırılmış ve yorumlanmıştır. Çalışmanın sonuçları Temel Bileşen Analizi yönteminin düşük korelasyon ve aykırı değerlere rağmen hiyerarşik kümeleme sonuçları ve dendrogramlar üzerinde olumlu etkiler yarattığını göstermiştir.

Anahtar Kelimeler: hiyerarşik kümeleme analizi, temel bileşen analizi, kofenetik korelasyon katsayısı, F_M indeks, Baker gamma korelasyon katsayısı, tanglegram.



Dedicated to my parents and my husband...

ACKNOWLEDGEMENTS

First of all, I would like to express my gratitude and endless thanks to my very valuable advisor Prof. Dr. Tolunay GÖÇKEN who guides me in the best way with her knowledge, experience, and interest, trusts me and motivates me.

Secondly, I would like to thank my family so much, especially my mother Perihan ŞAN and my father Ali ŞAN, who have been with me financially and morally since the day I was born.

Finally, I would like to express my endless thanks to my husband Muhsin Emin SONAY, who has been with me in every moment of my life and whom I love very much.

And lastly, I believe that the knowledge and experience that a person has is her best guide. For this reason, I would like to express my sincere thanks to all my esteemed professors who shed the light of science in my life.

TABLE OF CONTENTS

ABSTRACT	IV
ÖZET.....	V
ACKNOWLEDGEMENTS.....	VII
TABLE OF CONTENTS.....	VIII
LIST OF TABLES	XV
NOMENCLATURE	XVI
1. INTRODUCTION.....	1
2. LITERATURE REVIEW	4
3. CLUSTER ANALYSIS.....	7
4. CLUSTER ANALYSIS METHODS	10
4.1. PARTITIONAL CLUSTERING TECHNIQUES :	10
4.2. DENSITY-BASED CLUSTERING TECHNIQUES :	12
4.3. GRID-BASED CLUSTERING TECHNIQUES :	13
4.4. HIERARCHICAL CLUSTERING TECHNIQUES :	15
4.4.1. AGGLOMERATIVE CLUSTERING (AGNES):.....	17
4.4.1.1. Single (Minimum) Linkage:	19
4.4.1.2. Complete (Maximum) Linkage:.....	19
4.4.1.3. Average (Mean) Linkage:.....	20
4.4.1.4. Centroid Linkage:	20
4.4.1.5. Median Linkage:	21
4.4.1.6. Ward's Criterion (Minimum Variance):	21
4.4.1.7. Advanced Hierarchical Methods:	22
4.4.2. DIVISIVE CLUSTERING (DIANA):	22
5. PRINCIPAL COMPONENT ANALYSIS	23
5.1. MATHEMATICAL BACKGROUND FOR PRINCIPAL COMPONENT ANALYSIS.....	24
5.1.1. How to decide the number of the principal components?	26
6. THE INDEXES FOR COMPARING HIERARCHICAL CLUSTERING	27
6.1. CPCC (COPHENETIC CORRELATION COEFFICIENT):	27
6.2. FM_INDEX (FOWLKES AND MALLOWS INDEX):	28

6.3. BAKER'S GAMMA COEFFICIENT (GOODMAN AND KRUSKAL'S GAMMA COEFFICIENT):.....	29
7. MATERIALS AND METHODS.....	30
7.1. R STATISTICAL PROGRAMMING:	30
7.2. INFORMATION ABOUT DATASETS:	31
7.2.1. Dataset 1. (African Dataset):	33
7.2.2. Dataset 2. (European Dataset):.....	33
7.2.3. Dataset 3. (Asian Dataset):.....	34
7.3. THE STEPS OF THE ANALYSIS:	34
8. RESULTS AND DISCUSSIONS	37
8.1. APPLYING PRINCIPAL COMPONENT ANALYSIS TO THE AFRICAN DATASET:.....	37
8.2. CLUSTERING RESULTS OF AFRICAN COUNTRIES TO THE ECONOMIC INDICATORS:	40
8.2.1. Comparative Results for Single Linkage Method:.....	44
8.2.2. Comparative Results for Complete Linkage Method:.....	46
8.2.3. Comparative Results for Average Linkage Method:.....	48
8.2.4. Comparative Results for Centroid Linkage Method:	50
8.2.5. Comparative Results for Median Linkage Method:	52
8.2.6. Comparative Results for Ward's Criterion Linkage Method:	54
8.3. APPLYING PRINCIPAL COMPONENT ANALYSIS TO THE EUROPE DATASET:	56
8.4. CLUSTERING RESULTS OF EUROPE COUNTRIES TO THE ECONOMIC INDICATORS AND SOCIAL INDICATORS:	59
8.4.1. Comparative Results for Single Linkage Method:.....	62
8.4.2. Comparative Results for Complete Linkage Method:.....	65
8.4.3. Comparative Results for Average Linkage Method:.....	66
8.4.4. Comparative Results for Centroid Linkage Method:	67
8.4.5. Comparative Results for Median Linkage Method:	68
8.4.6. Comparative Results for Ward's Criterion Linkage Method:	69
8.5. APPLYING PRINCIPAL COMPONENT ANALYSIS TO THE ASIAN DATASET:	71
8.6. CLUSTERING RESULTS OF ASIAN COUNTRIES TO THE ECONOMIC INDICATORS, SOCIAL INDICATORS AND ENVIRONMENT AND INFRASTRUCTURE INDICATORS:	74
8.6.1. Comparative Results for Single Linkage Method:.....	77
8.6.2. Comparative Results for Complete Linkage Method:.....	79

8.6.3. Comparative Results for Average Linkage Method:.....	81
8.6.4. Comparative Results for Centroid Linkage Method:	82
8.6.5. Comparative Results for Median Linkage Method:	83
8.6.6. Comparative Results for Ward’s Criterion Linkage Method:	84
9. CONCLUSIONS.....	86
10. RECOMMENDATIONS.....	90
REFERENCES.....	91
APPENDICES	101



LIST OF FIGURES

Figure 3.1.	The euclidean distance and matrix with representative numbers.....	9
Figure 4.1.	The k-means partitioning algorithm	11
Figure 4.2.	Partitional clustering	11
Figure 4.3.	Density-based clustering	13
Figure 4.4.	Grid-based clustering	14
Figure 4.5.	Hierarchical clustering with a dendrogram.	16
Figure 5.1.	Illustration of the principal component analysis (PCA)	24
Figure 8.1.	Colored correlation matrix of the variables for dataset 1.....	37
Figure 8.2.	Scatterplot of the African dataset for the economic indicator variables.	41
Figure 8.3.	Colored correlation matrix of the variables for dataset 2.....	56
Figure 8.4.	Scatterplot of the European dataset for the economic indicator and social indicator variables.	59
Figure 8.5.	Colored correlation matrix of the variables for dataset 3.....	71
Figure 8.6.	Scatterplot of the European dataset for the economic indicator and social indicator variables.	74
Figure 8.1.1.	Scree plot of the variance values of the principal components for dataset 1. ..	38
Figure 8.1.2.	Scree plot of the variance explained rates of the principal components for dataset 1.....	39
Figure 8.2.1.	Tanglegram for comparison dendrograms of single linkage method and PCA+single linkage method.	44
Figure 8.2.2.	Tanglegram for comparison dendrograms of complete linkage method and PCA+complete linkage method.	46
Figure 8.2.3.	Tanglegram for comparison dendrograms of average linkage method and PCA+average linkage method.....	48
Figure 8.2.4.	Tanglegram for comparison dendrograms of centroid linkage method and PCA+centroid linkage method.....	50
Figure 8.2.5.	Tanglegram for comparison dendrograms of median linkage method and PCA+median linkage method	52
Figure 8.2.6.	Tanglegram for comparison dendrograms of ward.D2 linkage method and PCA+ward.D2 linkage method.	54
Figure 8.3.1.	Scree plot of the variance values of the principal components for dataset 2. ..	57
Figure 8.3.2.	Scree plot of the variance explained rates of the principal components for dataset 2.....	58
Figure 8.4.1.	Tanglegram for comparison dendrograms of single linkage method and PCA+single linkage method.	63
Figure 8.4.2.	Tanglegram for comparison dendrograms of complete linkage method and PCA+complete linkage method.	65
Figure 8.4.3.	Tanglegram for comparison dendrograms of average linkage method and PCA+average linkage method.....	66

Figure 8.4.4.	Tanglegram for comparison dendrograms of centroid linkage method and PCA+centroid linkage method.	67
Figure 8.4.5.	Tanglegram for comparison dendrograms of median linkage method and PCA+median linkage method.	68
Figure 8.4.6.	Tanglegram for comparison dendrograms of ward.D2 linkage method and PCA+ward.D2 linkage method.	69
Figure 8.5.1.	Scree plot of the variance values of the principal components for dataset 3. ...	72
Figure 8.5.2.	Scree plot of the variance explained rates of the principal components for dataset 3.	73
Figure 8.6.1.	Tanglegram for comparison dendrograms of single linkage method and PCA+single linkage method.	77
Figure 8.6.2.	Tanglegram for comparison dendrograms of complete linkage method and PCA+complete linkage method.	79
Figure 8.6.3.	Tanglegram for comparison dendrograms of average linkage method and PCA+average linkage method.	81
Figure 8.6.4.	Tanglegram for comparison dendrograms of centroid linkage method and PCA+centroid linkage method.	82
Figure 8.6.5.	Tanglegram for comparison dendrograms of median linkage method and PCA+median linkage method.	83
Figure 8.6.6.	Tanglegram for comparison dendrograms of ward.D2 linkage method and PCA+ward.D2 linkage method.	84

LIST OF FIGURES FOR APPENDICES

Figure A1.	Single linkage clustering results for the African countries.....	101
Figure A2.	Single linkage clustering results after applying PCA for the African countries.....	101
Figure A3.	Complete linkage clustering results for the African countries.	102
Figure A4.	Complete linkage clustering results after applying PCA for the African countries.....	102
Figure A5.	Average linkage clustering results for the African countries.	103
Figure A6.	Average linkage clustering results after applying PCA for the African countries.....	103
Figure A7.	Centroid linkage clustering results for the African countries.....	104
Figure A8.	Centroid linkage clustering results after applying PCA for the African countries.....	104
Figure A9.	Median linkage clustering results for the African countries.....	105
Figure A10.	Median linkage clustering results after applying PCA for the African countries.....	105
Figure A11.	Ward's Criterion linkage clustering results for the African countries.....	106
Figure A12.	Ward's Criterion linkage clustering results after applying PCA for the African countries.	106
Figure A13.	Single linkage clustering results for the Europe countries.	107
Figure A14.	Single linkage clustering results after applying PCA for the Europe countries.....	107
Figure A15.	Complete linkage clustering results for the Europe countries.....	108
Figure A16.	Complete linkage clustering results after applying PCA for the Europe countries.....	108
Figure A17.	Average linkage clustering results for the Europe countries.....	109
Figure A18.	Average linkage clustering results after applying PCA for the Europe countries.....	109
Figure A19.	Centroid linkage clustering results for the Europe countries.	110
Figure A20.	Centroid linkage clustering results after applying PCA for the Europe countries.....	110
Figure A21.	Median linkage clustering results for the Europe countries.	111
Figure A22.	Median linkage clustering results after applying PCA for the Europe countries.....	111
Figure A23.	Ward's Criterion linkage clustering results for the Europe countries.	112
Figure A24.	Ward's Criterion linkage clustering results after applying PCA for the Europe countries.....	112
Figure A25.	Single linkage clustering results for the Asian countries.	113
Figure A26.	Single linkage clustering results after applying PCA for the Asian countries.	113
Figure A27.	Complete linkage clustering results for the Asian countries.	114
Figure A28.	Complete linkage clustering results after applying PCA for the Asian countries.....	114

Figure A29. Average linkage clustering results for the Asian countries.	115
Figure A30. Average linkage clustering results after applying PCA for the Asian countries.....	115
Figure A31. Centroid linkage clustering results for the Asian countries.....	116
Figure A32. Centroid linkage clustering results after applying PCA for the Asian countries.....	116
Figure A33. Median linkage clustering results for the Asian countries.	117
Figure A34. Median linkage clustering results after applying PCA for the Asian countries.	117
Figure A35. Ward's Criterion linkage clustering results for the Asian countries.	118
Figure A36. Ward's Criterion linkage clustering results after applying PCA for the Asian countries.....	118



LIST OF TABLES

Table 3.1.	Distance measures and formulations using for cluster analysis.....	8
Table 4.1.	The Lance and Williams update formulations and parameters.....	18
Table 7.1.	The variables which include general information about the countries.	31
Table 7.2.	The variables which include economical information about the countries.....	31
Table 7.3.	The variables which include social information about the countries.....	32
Table 7.4.	The variables which include environmental and infrastructural information about the countries.	32
Table 8.1.1.	The variance values of the principal components for dataset 1.	38
Table 8.1.2.	The variance explained percentages of the principal components for dataset 1.	39
Table 8.2.1.	Clustering results for single linkage and PCA+single linkage method.	45
Table 8.2.2.	Clustering results for complete linkage and PCA+complete linkage method. ...	47
Table 8.2.3.	Clustering results for average linkage and PCA+average linkage method.....	49
Table 8.2.4.	Clustering results for centroid linkage and PCA+centroid linkage method.	51
Table 8.2.5.	Clustering results for median linkage and PCA+median linkage method.....	53
Table 8.2.6.	Clustering results for ward.D2 linkage and PCA+ward.D2 linkage method.....	55
Table 8.3.1.	The variance values of the principal components for dataset 2.	56
Table 8.3.2.	The variance explained percentages of the principal components for dataset 2.	58
Table 8.4.1.	Clustering results for single linkage and PCA+single linkage method.	64
Table 8.4.2.	Clustering results for ward.D2 linkage and PCA+ward.D2 linkage method.....	70
Table 8.5.1.	The variance values of the principal components for dataset 3.	71
Table 8.5.2.	The variance explained percentages of the principal components for dataset 1.	73
Table 8.6.1.	Clustering results for single linkage and PCA+single linkage method.	78
Table 8.6.2.	Clustering results for complete linkage and PCA+complete linkage method. ...	80
Table 8.6.3.	Clustering results for ward.D2 linkage and PCA+ward.D2 linkage method.....	85
Table 9.1.	Coefficient and index results used to measure the similarity of dendrograms for all datasets and hierarchical clustering methods.	87

NOMENCLATURE

PCA	: Principal Component Analysis
CPCC	: Cophenetic Correlation Coefficient
FM_index	: Fowlkes and Mallows index
BGCC	: Baker's Gamma Correlation Coefficient
ST_DBSCAN with Noise	: Spatial and Temporal_Density-Based Spatial Clustering of Applications
SEGE	: Socio - Economic Development
G-10	: Group of Ten
EU	: European Union
GI	: General Informations
EI	: Economic Indicators
SI	: Social Indicators
EII	: Environmental and Infrastructural Indicators
γ	: Gamma
Π	: Pi
B	: Beta
λ	: Lamda
α	: Alpha

1. INTRODUCTION

In the digitalized world, due to scientific and technological developments, every transaction we do in our daily life and even every step we take appears as data. A huge amount of data is produced every day and this amount is increasing day by day. According to Statista's research and forecasts, by 2025, the volume of data/information created, captured, copied, and consumed will be more than 180 zettabytes (Statista.com).

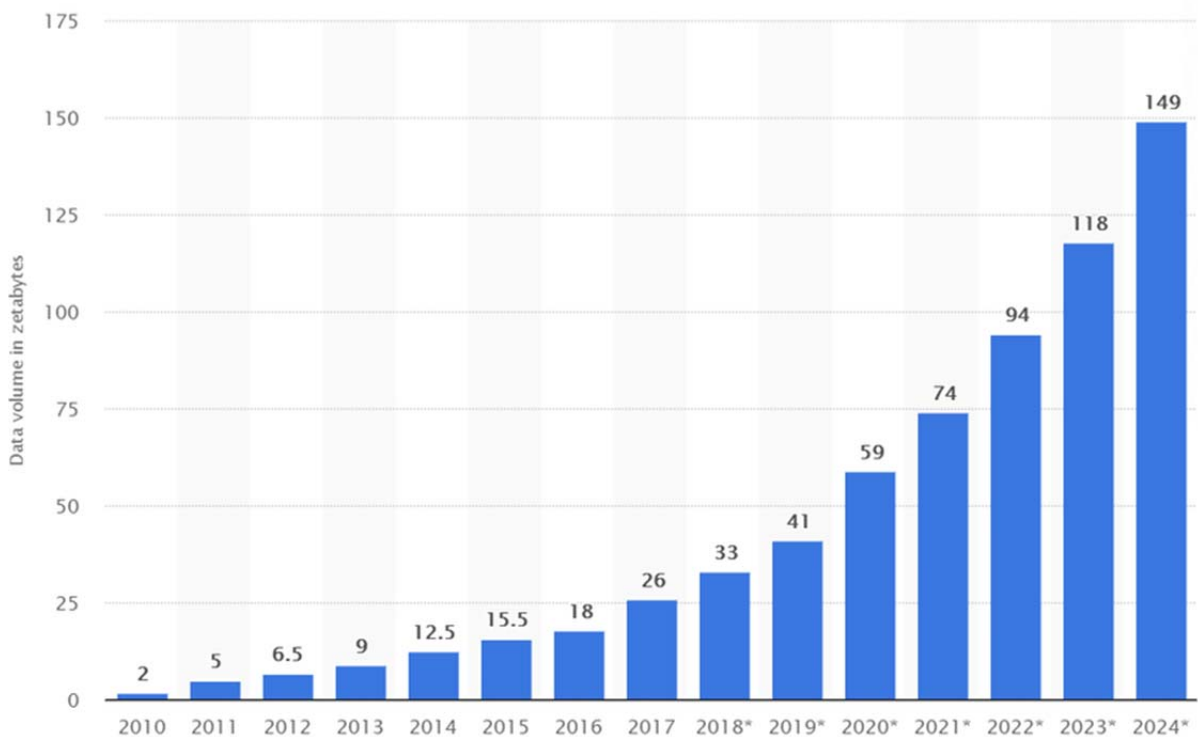


Figure 1. Volume of data/information created, captured, copied, and consumed worldwide from 2010 to 2024 (in zettabytes).

This vast amount and variety of data consists of bits of information that are meaningless on their own. However, with the development of technology and science, many methods have been developed and continue to be developed to process this data and transform it into important information that has meaning. The scientific discipline that includes these processes and methods is data mining. As a general definition, "Data mining is the process of discovering interesting patterns and knowledge from large amounts of data." (Han et al.,

2012). There are many methods for extracting useful information from data in data mining: Classification, Clustering, Association rules, Decision trees, etc. The method used varies according to the type of data, structure, and the information desired to be obtained. Data mining applications are associated with many applications, systems, and algorithms.



Figure 2. Data Mining steps (Source:bigdatapath.wordpress.com).

An important feature of data mining is that it includes techniques that discover hidden groups, patterns and information in complex and meaningless data. This process is named knowledge discovery, the most widely used technique in cluster analysis. Cluster analysis, in its simplest definition, is the process of grouping observations in a data set according to their similarities and differences. The main purpose of cluster analysis is to discover the natural groups in the data set by collecting the closest observations in the same cluster. However, sometimes these natural groups or clusters can be difficult to observe due to the noise of outliers and unrelated features in the dataset. This is especially true for high-dimensional datasets. In such cases, what needs to be done is to minimize the unnecessary and irrelevant variables and data in the data set. Dimension reduction techniques are used to reduce the number of unnecessary features in the data set and to perform operations with fewer variables more easily in matters

such as noisy data and outliers. Principal Component Analysis (PCA) is the most widely used dimension reduction method especially in cluster analysis. PCA reveals the basic pattern and characteristic information in the dataset by compressing the data according to the variance values. PCA is one of the most frequently used methods with cluster analysis for several aims.

In this thesis , despite the outlier and noise problems of high-dimensional datasets, the facilitating role of PCA in clustering analysis is investigated. The economic, social, environmental, and infrastructure data of the countries for the year 2020, compiled from the United Nations data platform, are used in the creation of datasets. In the data sets created, African, European, and Asian countries are grouped by hierarchical methods according to 22, 38, and 46 variables, respectively. To observe the effect of PCA, the tanglegram and coefficients of the method results with and without dimension reduction are compared and interpreted.

In the next parts of the thesis, related studies are reviewed, and clustering analysis and PCA topics are detailed with sub-titles. Datasets, programming language and packages used in the analysis, and information about the analysis stages are given in the materials and methods section. Both the graphical and numerical results of the analysis process are discussed in the results and discussion part and examined in detail. A general evaluation is made in the conclusion section with the outputs of the analysis results.

2. LITERATURE REVIEW

Cluster analysis and PCA, although they are different techniques, can be used for the same purpose in many studies. The most used and known of these purposes are grouping and determining the most important factors in the data set. Article studies in the literature are generally carried out using data from different fields for these two purposes. Some of these studies are as follows.

PCA and clustering have been used in some agricultural studies to assess species variation or product quality (Koj & Saba, 2015; Öten & Albayrak, 2018; Shoba et al., 2019; De Luca, 2011; Kim et al., 2014). Some studies have used clustering and PCA methods to classify and evaluate data from different sectors (Gaitani et al., 2010; Nwangburuka et al., 2011; Margaritis et al., 2020; Penkova, 2017). Abdulhafedh (2021) carried out customer segmentation by using PCA, hierarchical clustering and k-means clustering in his study. On the other hand, Yıldız et al. (2010) compared the dimension reduction effect of PCA and non-negative matrix factorization methods on clustering.

The thesis studies on cluster analysis in the literature generally used clustering methods for the same purpose, and objects are grouped with data from different fields and sectors. The different thesis studies are as follows. In one of the studies in which PCA and clustering analysis are used together, cancer-related data is used, and in the other study, it is aimed to classify the provinces in Turkey in terms of modernity-postmodernity with SEGE data (İncekırık, 2005; Kaya, 2020). In one of the thesis studies in which model-based clustering analysis is discussed, remotely sensed satellite image data is classified, and in another study, a new method based on multi-criteria decision making is presented to determine the number of clusters in model-based clustering (Servi, 2009; Birbiçer, 2017; Akoğul, 2018).

Cluster validity indices are compared in some studies, also the artificial neural network method is used to test the results differently (Hacıoğlu, 2016; Malkoç, 2018; Alkın, 2019). Erilli (2012), in the doctoral thesis, studied fuzzy clustering analysis on fuzzy numbers with

the data of chess game scores. Firat (2016), considered PCA, cluster analysis and factor analysis methods as classification methods in his thesis study. Özgür (2017) used fuzzy Vikor, fuzzy Topsis and cluster analysis methods to determine the hospital establishment location, and compared the results. Kaya conducted a thesis study in 2018 that included the comparison of fuzzy clustering algorithms.

Bulut (2017) introduced and implemented the ROBEM hybrid algorithm for dimension reduction and classification. Akay (2018) examined clustering analysis on mixed-variable panel data and proposed a new distance measure. Özeltindis (2018) introduced temporal-spatial data mining methods and made an application on Turkey's temperature and precipitation data with the ST-DBSCAN algorithm in his thesis. Taşatan (2018) performed clustering on highway distance data with a new method based on the single linkage and complete linkage methods.

Customer grouping of a company is carried out using particle swarm optimization together with k-means algorithm, and G-10 countries are clustered in terms of economic freedoms using Edas method together with k-means (Cenger, 2018; Levent, 2019). Özgen (2019) tested cluster quality with discriminant analysis using different clustering methods and distance measures. Young isolated neutron stars clustered in the thesis studied by Ay (2019). Çizmeci (2019) handled students' attitudes towards mathematics and geometry lessons by using exploratory factor analysis and clustering analysis.

Yelmen, in the thesis studied in 2020, used different word representation methods and PCA together on the question documents and analyzed them with different clustering methods. Engeloğlu (2010), in his doctoral thesis, grouped EU member countries in terms of producer and consumer confidence indices with the asymmetric causality test and cluster analysis.

There are a small number of thesis studies on principal component analysis. In 2009, Börü examined the main factors affecting the development of the financial sector in Turkey with

PCA. Similarly, Keskin (2016) tried to determine the factors affecting reproduction for sheep with PCA. Gümüş (2015) classified olive oil products from different regions with PCA according to certain parameters. Doğanlarlioğlu (2019) proposed an advanced PCA method for remotely sensed multispectral data. İnal (2020) evaluated different basis functions using smoothed functional PCA in datasets.



3. CLUSTER ANALYSIS

Cluster analysis or simply clustering is an unsupervised learning method that contains many techniques aiming to discover knowledge and to find different groups or hidden patterns in a data set (Kassambara, 2017). According to some sources, cluster analysis is also considered as automatic classification or numerical classification method (Everitt et al., 2011). In some sources, it is also called unsupervised classification and exploratory data analysis, because clustering uses unlabeled data unlike classification (Xu & Wunsch, 2009). As a result of cluster analysis, observations are divided into groups according to their similarities and each of these groups is called a cluster. We can simply formulate the clustering as follows; here, C_i , $i = 1, \dots, k$, represents the clusters formed at the end of the clustering process, and S represents all the objects in the data.

$$C = C_1 \dots \dots C_k ; \quad S = \bigcup_{i=1}^k C_i ; \quad C_i \cap C_j = \emptyset \quad \text{for } i \neq j \quad (3.1)$$

Cluster analysis is used in many field such as; medicine, psychiatry, sociology, criminology, anthropology, archaeology, geology, geography, remote sensing, market research, economics, and engineering (Rencher, 2002). To give examples from daily life; in business for segmentation of customers or products, in psychiatry grouping patients, diseases or medicaments, in astronomy clustering astronomical objects, in bioinformatics clustering gene and DNA data, in archaeology grouping fossils. In addition, cluster analysis is also used for data summarization, data reduction and outlier detection. Examples of more advanced systems using cluster analysis are; image segmentation in computer science, student's academic performance detection in education, travel package recommender system in tourism, growth pattern identification in medical, and crime pattern detection in social science (D. & Vidyavathi, 2015).

Cluster analysis performs the clustering process by looking at the similarities and differences of the data. Similar observations are grouped into the same cluster as much as possible, while different ones are grouped into different clusters whenever possible. In clustering, many different distance measures are used to identify similarities and differences between observations. Although Euclidean distance is most widely used, different clustering results may be obtained with different distance measures. Because the measure of similarity and the difference of all observations are determined according to the distance formula used, the

distance measure directly affects clustering and cluster quality. The distance measures used in clustering analysis were given in the table below.

Table 3. 1. Distance measures and formulations using for cluster analysis.

Euclidean distance :	$d_{\text{euc}} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
Manhattan distance:	$d_{\text{man}} = \sum_{i=1}^n x_i - y_i $
Minkowski distance:	$d_{\text{mink}} = (\sum_{i=1}^n x_i - y_i ^p)^{1/p}$
Pearson correlation distance:	$d_{\text{cor}} = 1 - \frac{ \sum_{i=1}^n x_i y_i }{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}$
Eisen cosine correlation distance:	$d_{\text{eisen}} = 1 - \frac{ \sum_{i=1}^n x_i y_i }{\sqrt{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i^2}}$
Spearman correlation distance:	$d_{\text{spear}} = 1 - \frac{\sum_{i=1}^n (x'_i - \bar{x}') (y'_i - \bar{y}')}{\sqrt{\sum_{i=1}^n (x'_i - \bar{x}')^2 \sum_{i=1}^n (y'_i - \bar{y}')^2}}$
Kendall correlation distance:	$d_{\text{kend}} = 1 - \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$
Squared euclidean distance :	$d_{\text{squ}} = \sum_{i=1}^n (x_i - y_i)^2$
Mahalanobis distance :	$d_{\text{mah}} = \sqrt{(x - y)^T S^{-1} (x - y)}$
Chebyshev distance :	$d_{\text{cheby}} = \max(x_1 - x_2 , y_1 - y_2)$

It is important to determine the most appropriate distance measure according to the purpose of the clustering and the data set. For example, correlation-based distance measures such as Spearman and Pearson are highly sensitive to intervariable relationships and outliers and therefore they are often used in gene data sets (Kassambara, 2017). The distance of each observation to other observation points is calculated according to the selected distance measure and a matrix with all distance values is created. This matrix is called *distance matrix* or *similarity matrix*. For example, for a k-dimensional data set with n observations, the distance matrix would be in the form shown below.

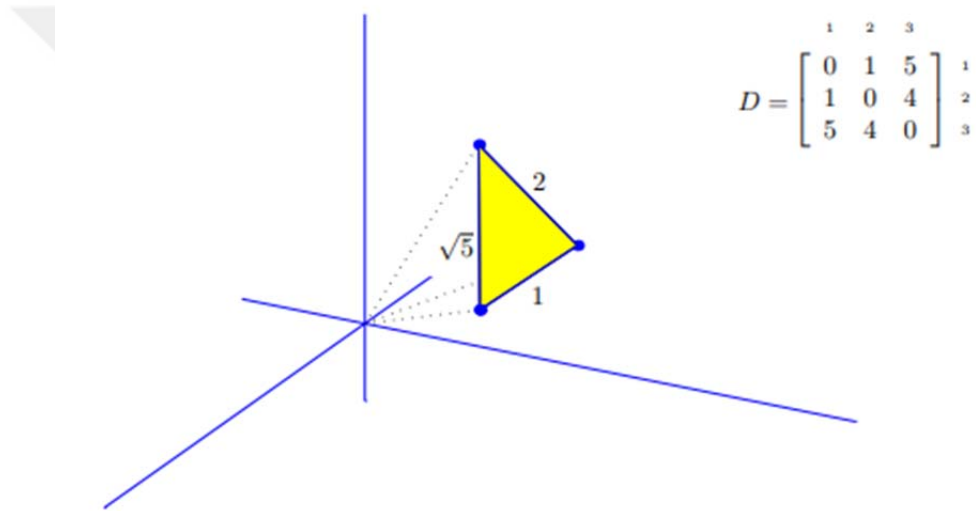


Figure 3. 1. The euclidean distance and matrix with representative numbers (Dattorro, 2008).

The distance matrix is calculated in all techniques used in cluster analysis. But how to use the distance matrix depends on the technique. For example, some techniques use the shortest distance, while some techniques use the farthest distance or average distance.

4. CLUSTER ANALYSIS METHODS

There are several cluster analysis techniques, and still, new techniques are added to the literature with the improvement and development studies on existing methods. Just as the choice of distance measure, different clustering techniques also affect clustering results and cluster quality. The purpose of all clustering methods is to divide the observations into clusters in the most appropriate way. However, the process and calculations they follow show differences. Each of the techniques has strengths and weaknesses, advantages and disadvantages. Most of the techniques are used for numerical data sets. However, the data in many forms can be clustered with clustering analysis. For example, text, graphics, pictures, multimedia, mixed data and categorical data. The basic clustering methods are detailed below with general features. Only the hierarchical clustering method has been discussed more comprehensively because it is used as the method in the thesis. Before explaining the clustering methods, some terms related to clustering are given below.

Centroid: Center of cluster.

$$C_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_i \quad (3.2)$$

Radius: Square root of the squared mean distance between observation and centroid.

$$R = \sqrt{\frac{1}{n} \sum_{i=1}^n (X_i - C)^2} \quad (3.3)$$

Diameter: Square root of the distance mean of the pairs of observations in the cluster.

$$DM = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n \sum_{j=1}^n (X_i - X_j)^2} ; \quad (i \neq j) \quad (3.4)$$

Medoid: Observation near the cluster center (Altunkaynak, 2019).

4.1. PARTITIONAL CLUSTERING TECHNIQUES :

Partitional clustering method which is prominent with simplicity and competitive computational complexity features, partitions the data into k clusters with some determined criteria (Hajkacem et al., 2019). In the partition method, each observation belongs to a single group or cluster and the clusters are separated in this way (Han et al., 2012). The most popular

partitional clustering method is the K-means algorithm because it is simple, practical and it has effective results (Al Hasan et al., 2009). In addition, it has the good reasons to be chosen; firstly, its time complexity is $O(nkl)$ (Here is n = pattern number, k = cluster number, l = iteration number) and its space complexity is $O(k+n)$ for the data matrix (Madhulatha, 2012). Although the K-means algorithm has some advantages in terms of implementation, it has significant disadvantages such as the optimum number of clusters problem and sensitivity for outliers and noises (Han et al., 2012). The number of clusters can be determined randomly, as well as the elbow method and the silhouette method are also used as analytical methods. The K-means algorithm follows the steps below in order as the working process;

Algorithm: k -means.

Input:

k : the number of clusters,

D : a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects from D as the initial cluster centers;

(2) **repeat**

(3) (re)assign each object to the cluster to which the object is the most similar;

(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;

(5) **until** no change;

Figure 4. 1. The k -means partitioning algorithm (Han et al., 2012).

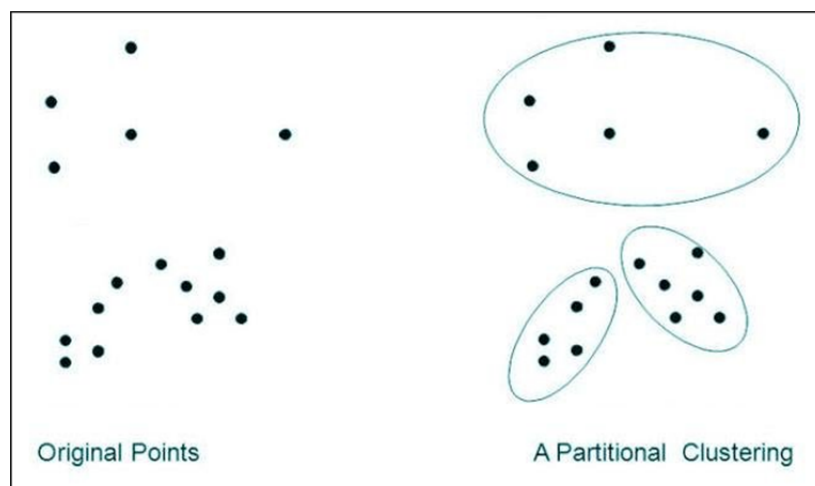


Figure 4. 2. Partitional clustering (Alhenaki, 2019).

Although the k-medoids algorithm is not as popular as the k-means, it is less affected by outliers and noises, so it is more robust than k-means. CLARA and PAM algorithms are also advanced versions of the k-medoids algorithm applied for large and small data sets (Han et al., 2012). The other techniques of the partitioning clustering method are; K-Medians Clustering, K-Modes Clustering, K-Prototypes, Fuzzy K-Means Clustering, Fuzzy C-Means Clustering, X-Means Clustering, Intelligent K-Means Clustering, Bisecting K-Means Clustering, Kernel K-Means Clustering, Mean Shift Clustering, Weighted K-Means Clustering, Genetic K-Means Clustering, CLARA (Clustering for Large Applications), CLARANS (A Clustering Algorithm based on Randomized Search), PAM (Partitioning Around Medoids).

4.2. DENSITY-BASED CLUSTERING TECHNIQUES :

Density-based clustering algorithms are used to discover clusters with a certain shape and density in the data (Madhulatha, 2012). The main goal of this method is to find the clusters which have arbitrary shapes (especially nonconvex) and their distribution parameters (Rokach & Maimon, 2014). Density-based clustering is generally preferred for geographical, geological, and space data. The most familiar algorithms of this method are; DBSCAN (Density-Based Spatial Clustering of Applications with Noise), DENCLUE (DENSITY-based CLUstEring), OPTICS (Ordering Points To Identify the Clustering Structure) (Ayed et al., 2014). The most well-known among them is the DBSCAN algorithm. Although DBSCAN works very well in outlier detection, it does not work well in high-dimensional data (Everitt et al., 2011). In general, the time complexity of DBSCAN for a data set with n observations is $O(n \log n)$, but this may differ depending on the dimension of the data (Ester, 2014). There are some challenges of density-based clustering methods. For example; most of the internal validation indexes or quality measures are not suitable for the evaluation of density-based clustering results and the other challenge is clustering on streaming data (Campello et al., 2020).

The basic steps defining how the algorithm works are given below (Everitt et al., 2011; www.analyticsvidhya.com, 2021; www.towardsdatascience.com, 2019);

1. Objects are classified according to their densities.
2. The number of objects (M) and the Radius value (R) in the clusters are determined by the user.
3. Each cluster has a core and the remaining within the cluster objects are called non-core.
4. Objects that can reach density are called boundary objects.
5. Objects in regions of low density are defined as noise.
6. A set with M objects and radius R is chosen as an initial point.
7. Respectively; first, densities are directly connected to the core and then neighbouring objects are included in the set (except the noises).
8. The processes are repeated for other core points until clustering is complete.

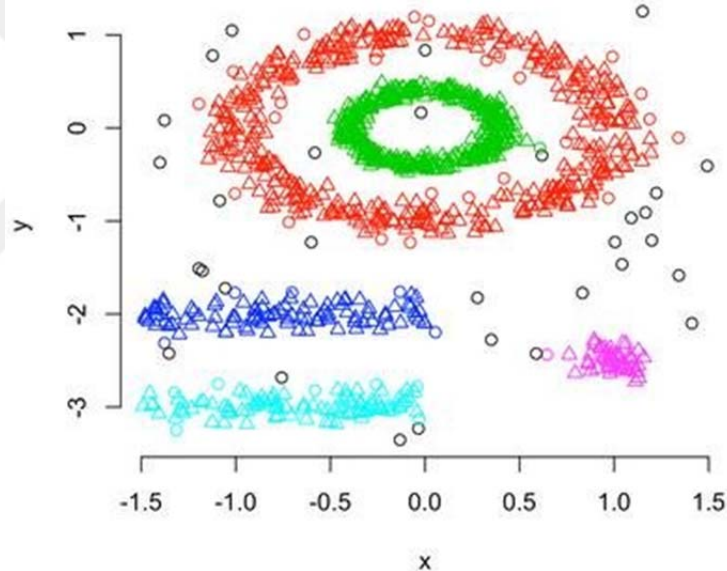


Figure 4. 3. Density-based clustering (V & H A, 2017).

4.3. GRID-BASED CLUSTERING TECHNIQUES :

Grid-based clustering methods partition the data space into the cells by using grid structures and the cells exist in a hierarchical structure from low level to high level according to the information they carry like the number of points, mean, standard deviation, minimum and maximum values and distribution types (Xu & Wunsch, 2009). Grid-based algorithms have some advantages as being query-independent, easy parallelization and fast time-complexity.

However, they also have the disadvantage; all the clusters that are found are horizontal or vertical and there is no diagonal cluster (Madhulatha, 2012). STING algorithm (STatistical INformation Grids) is the most popular grid-based algorithm. STING's time complexity is $O(L)$ (Here L is the number of lowest level cells) and STING shows better performance than other algorithms in simulation studies, because it works faster (Xu & Wunsch, 2009). The other algorithms of the grid-based method are; CLIQUE (CLustering In QUEst) and, MAFIA (Merging of Adaptive Finite Intervals). One of the most important features of grid-based clustering methods is that while other methods work best with numerical data, grid-based algorithms can work for different types of data (Berkhin, 2002).

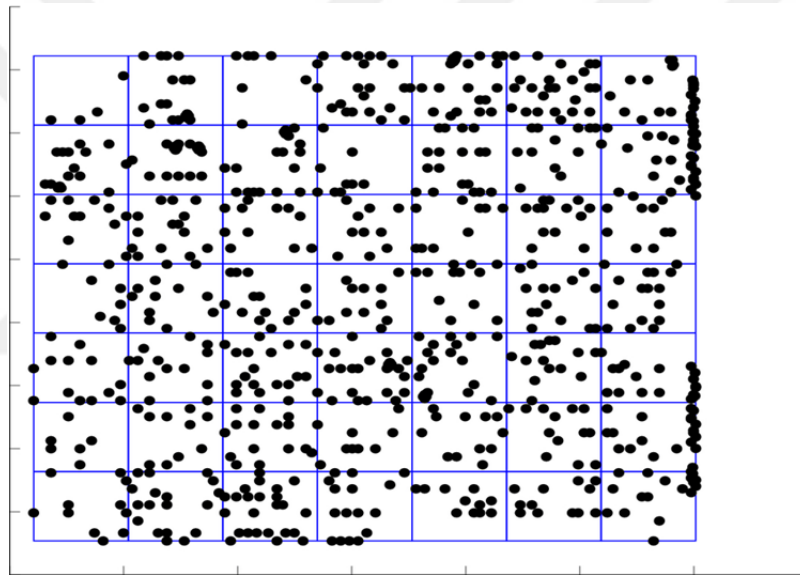


Figure 4. 4. Grid-based clustering (Zojaji et al., 2020)

As with all algorithms, grid-based algorithms have some challenges; non-uniformity, locality and dimensionality. Grid-based algorithms basically follow the steps below (Cheng et al., 2014).

1. Form grid structure to define the cells,
2. Find densities for the cells,
3. Classify the cells by their densities,
4. Determine the center of clusters,
5. Passage of neighbouring cells.

4.4. HIERARCHICAL CLUSTERING TECHNIQUES :

Hierarchical clustering methods are the methods that perform clustering operations by following a hierarchical structure. These methods form clusters by grouping the objects in the data set according to their similarities, and as a result of the operations, a tree-like hierarchical structure form emerges (Han et al., 2012). Although partitional clustering methods can give effective results in terms of cluster quality, they have a significant disadvantage in that the optimal number of clusters is uncertain. On the other hand, hierarchical methods, perform an unambiguous clustering without such a disadvantage.

Hierarchical methods reveal natural clusters in the dataset without the need to specify the number of cluster. Unlike partitional methods, in hierarchical methods, determining the number of clusters works as a stopping criterion. In this way, it is possible to end the process by obtaining the desired number of clusters in hierarchical clustering or to interrupt the clustering process at the desired stage. This is known as one of the important advantages of hierarchical methods (Reddy & Vinzamuri, 2014). Hierarchical methods are generally used for hierarchical classification of data, summarizing datasets, and visualizing data. Hierarchical methods are used quite frequently, especially in biology science for animal and plant classifications, as they give effective results (Kaufman & Rousseeuw, 1990). One of the weaknesses of the hierarchical methods is that they have high computational complexity. The time complexity they have is expressed by $O(n^2)$ and they are not suitable for large datasets (Xu & Wunsch, 2009).

In hierarchical methods, the tree-shaped diagram that presents the clustering results of the data in a hierarchical structure is called a **dendrogram** (Everitt et al., 2011). A dendrogram is a clear representation form that is easy to interpret and understand. The vertical axis of the dendrogram shows the distance values between clusters and data points, while the horizontal axis shows the data points (Forina et al., 2002). In hierarchical clustering, the cutting point of the dendrogram has a significant role in determining the clusters. A different number of different cluster combinations can be obtained by cutting the dendrogram at different points

(Xu & Wunsch, 2009). Some criteria can be used to cut a dendrogram from the optimal point. These criteria are as follows:

- An appropriate cutting point can be used for any criteria such as a specified number of clusters or a specified height.
- A high cut-off point may be determined for the minimum similarity between clusters, or a low-level point may be preferred for maximum similarity.
- The largest gap between groups in a dendrogram can generally be considered the most ideal cut-off point at which natural clusters will emerge (Manning et al., 2009).



Figure 4. 5. Hierarchical clustering with a dendrogram.
(Source:<https://harshsharma1091996.medium.com/hierarchical-clustering-996745fe656b>)

In a dendrogram, the merging and splitting points of the clusters can also be observed (Rani & Rohil, 2013). Another weakness of hierarchical methods is that the points of merging and splitting of the clusters cannot be changed and there cannot be any exchange of objects between the clusters. This feature turns into a disadvantage when optimal points are not selected and this directly affects the cluster quality (Kaufman & Rousseeuw, 1990). Advanced hierarchical methods have been offered to reduce these and the other disadvantages of hierarchical methods; BIRCH, Chameleon, ROCK and CURE (Rani & Rohil, 2013; Han et al., 2012). Hierarchical clustering methods are divided into two groups in terms of the

strategic way they follow. These two groups operate in opposite directions, bottom-up (agglomerative) and top-down (divisive) (Rencher, 2002).

4.4.1. AGGLOMERATIVE CLUSTERING (AGNES):

Agglomerative hierarchical clustering is the iterative process that starts with pairing the clusters n single-object in p-dimensional (p = number of the attributes) space according to the determined similarity criterion and ends with the cluster containing all objects (Sembiring et al., 2010; Day & Edelsbrunner, 1984; Orloci, 1967). Agglomerative hierarchical clustering includes clustering methods that are strategically bottom-up and work slowly but are very flexible (Kettani et al., 2014). Agglomerative hierarchical clustering methods have space and time complexity expressed by $O(n^2)$ (Day & Edelsbrunner, 1984). Agglomerative methods are more popular than divisive methods because they are easier to implement and understand (Tokuda et al., 2020).

The basic steps of the Agglomerative hierarchical clustering are given below ((Reddy & Vinzamuri, 2014);

1. The dissimilarity matrix is created.
2. Objects and clusters closest to each other are combined.
3. The dissimilarity matrix is updated after each merge (Old cluster information is deleted and new cluster information is added).
4. Merging the clusters is repeated until a single cluster containing all objects is formed.

Many similarities and distance metrics can be used in Agglomerative hierarchical clustering, and the difference between the methods is based on the distance criteria between groups (Tokuda et al., 2020). ***The Lance and Williams Update formulations*** which present as an alternative to all similarity metrics contain the dissimilarity matrix update formulas that are suitable for all hierarchical methods (Reddy & Vinzamuri, 2014). According to the Lance and Williams updates formulation, the distance between combined clusters i and j with a cluster k is calculated using the formula below (Everitt et al., 2011).

$$d_{k(ij)} = \alpha_i d_{ki} + \alpha_j d_{kj} + \beta d_{ij} + \gamma |d_{ki} - d_{kj}| ; \quad (4.1)$$

$$\alpha_i + \alpha_j + \beta = 1 , \quad \alpha_i = \alpha_j , \quad \beta < 1 , \quad \gamma = 0$$

The update process is performed according to the parameter values determined for each method (Everitt et al., 2011; Reddy & Vinzamuri, 2014). The parameter values and the formulas are given below in table 4.1.

Table 4. 1. The Lance and Williams update formulations and parameters (Here n_k , n_i and n_j represent the cluster sizes).

Lance and Williams Parameters			
Methods	α_i	β	γ
Single Linkage	1/2	0	-1/2
Complete Linkage	1/2	0	1/2
Average Linkage	$n_i/(n_i + n_j)$	0	0
Centroid Linkage	$n_i/(n_i + n_j)$	$-n_i n_j / (n_i + n_j)^2$	0
Median Linkage	1/2	-1/4	0
Ward's Method	$(n_k + n_i)/(n_k + n_i + n_j)$	$-n_k/(n_k + n_i + n_j)$	0

The Lance and Williams formulations are suitable for all of the graphical methods represented in euclidean space using pairwise dissimilarity matrix and also with the geometric methods which represented in euclidean space with the prototype vectors for the dissimilarity such as centroid, median and ward methods and provide ease of computation (Ah-Pine & Wang, 2016).

The method in which clusters are brought together using the lines and distance matrix based on attribute vectors and distance criteria is called the linkage technique (Yim & Ramdeen, 2015). There are many different linkage methods according to different distance preferences such as minimum, maximum, average, median, centroid and minimum variance (Tokuda et

al., 2020). All of these linkage methods and other advanced agglomerative hierarchical clustering methods are detailed below respectively.

4.4.1.1. Single (Minimum) Linkage:

In the single linkage method, also called the **nearest neighbour**, the minimum distance between two clusters is accepted as the criterion for the merging (Mohbey & Thakur, 2013; Rencher, 2002; Reddy & Vinzamuri, 2014). In each iteration, the two closest clusters are merged and the operations are repeated until the clustering is complete (El-Hamdouchi & Willett, 1987). A and B represent the clusters and the distance criterion for the method is given below;

$$D(A, B) = \min\{d(y_i, y_j) \mid \text{for } y_i \text{ in } A \text{ and } y_j \text{ in } B\} , \quad (4.2)$$

This method is faster than other hierarchical techniques with the time complexity expressed by $O(n^2)$ but, because of its local behaviour, it is sensitive to noises and outliers (Jarman, 2020; Emmendorfer & Canuto, 2021). For example, the clusters which are not closest in real can be merged for the closest objects in the clusters and this fault also affects the following iterations of the process (Jarman, 2020; Yim & Ramdeen, 2015; Mansoori, 2013). This case is known as the **chaining effect**.

4.4.1.2. Complete (Maximum) Linkage:

Unlike the single linkage method, in the complete linkage method, the furthest distance between two clusters is considered as the merging criterion (Rokach & Maimon, 2005; Yim & Ramdeen, 2015; Manning et al., 2009). The distance criterion for the method is as follow;

$$D(A, B) = \max\{d(y_i, y_j) \mid \text{for } y_i \text{ in } A \text{ and } y_j \text{ in } B\} , \quad (4.3)$$

Due to its non-local behaviour, smaller and tighter clusters are obtained (Kaufman & Rousseeuw, 1990; Rokach & Maimon, 2005; Manning et al., 2009). However, also in this method, the presence of outliers can have negative effects on clustering because of its sensitivity to noise and outliers (Yim & Ramdeen, 2015; Emmendorfer & Canuto, 2021; Jarman, 2020). The time complexity of the method can be expressed by $O(n^2 \log n)$ (Jarman, 2020).

4.4.1.3. Average (Mean) Linkage:

In this method, the average of the distances of all the pairs of points of the clusters is accepted as the criterion in the merging of two clusters (Xu & Wunsch, 2009; Everitt et al., 2011; Reddy & Vinzamuri, 2014). In each iteration, the procedure is repeated according to this criterion until the clustering process is completed. The distance criterion for the method is as follow;

$$D(A, B) = \frac{1}{n_A n_B} \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} d(y_i, y_j)$$

(4.4)

This method, which has the advantage of outlier insensitivity, is difficult to use for categorical data but is also a very effective method for numerical data (Han et al., 2012). The average linkage method has the time complexity expressed by $O(n^2 \log n)$ (Jarman, 2020).

4.4.1.4. Centroid Linkage:

This method performs the merging operation according to the proximity between the centroids of the clusters (Reddy & Vinzamuri, 2014; Bu et al., 2020; Kaufman & Rousseeuw, 1990). Here centroid denotes the centre point of the cluster and, the centroids of A and B clusters expressed by $\bar{y}_A$ and $\bar{y}_B$ respectively (Rencher, 2002);

$$D(A, B) = d(\bar{y}_A, \bar{y}_B) \quad (4.5)$$

The time complexity of the centroid linkage method can be expressed by $O(n^2 \log n)$ (Jarman, 2020).

4.4.1.5. Median Linkage:

In the merging of two clusters of different sizes, the midpoint of the distance between the two centroids, that is, the median value is used to prevent the centroid of the newly formed cluster from shifting predominantly towards the larger cluster (Rencher, 2002; Everitt et al., 2011). It is designed to overcome the disadvantage of the centroid method. The median value between A and B clusters can be expressed as follow;

$$M_{AB} = \frac{1}{2}(\bar{y}_A + \bar{y}_B) \quad (4.6)$$

4.4.1.6. Ward's Criterion (Minimum Variance):

The main purpose of this method is to obtain homogeneous clusters by keeping minimum the sum of squares of the error of the distances within the cluster (Rencher, 2002; Majerova & Nevima, 2017; Xu & Wunsch, 2009). Contrary to some methods, cluster size is effective on this method (Rencher, 2002).

$$SSE_A = \sum_{i=1}^{n_A} (y_i - \bar{y}_A)'(y_i - \bar{y}_A), \quad SSE_B = \sum_{i=1}^{n_B} (y_i - \bar{y}_B)'(y_i - \bar{y}_B), \quad (4.7)$$

$$SSE_{AB} = \sum_{i=1}^{n_{AB}} (y_i - \bar{y}_{AB})'(y_i - \bar{y}_{AB}); \quad I_{AB} = SSE_{AB} - (SSE_A + SSE_B)$$

There are two versions of this method; **ward1** and **ward2**. The main difference between these two versions; Squared Euclidean distance is used in ward1, while Euclidean distance is used in ward2 (Szekely & Rizzo, 2005; Murtagh & Legendre, 2014).

4.4.1.7. Advanced Hierarchical Methods:

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) aims to cluster large amounts of data using a hierarchical structure and without outlier sensitivity (Han et al., 2012; Xu & Wunsch, 2009). It uses a special hierarchical structure called a clustering feature. The method, expressed as with the time complexity $O(n)$, does not include the weaknesses such as the scalability and inability to return to previous step which agglomerative methods have (Han et al., 2012; Xu & Wunsch, 2009). **CHAMELEON** follows a dynamic and graphic-based hierarchical clustering process. It accepts the strength of the connection and the level of closeness between objects as merging criteria (Han et al., 2012; Xu & Wunsch, 2009; Reddy & Vinzamuri, 2014). **CURE** (Clustering Using REpresentatives) and **ROCK** (RObust Clustering using linKs) are both carried out the same clustering process and both aim to get clusters which include minimum distance and well-distributed objects. However, the main difference of these two approaches is; ROCK is used for categorical data, while CURE is used for numerical data (Xu & Wunsch, 2009; Reddy & Vinzamuri, 2014).

4.4.2. DIVISIVE CLUSTERING (DIANA):

Divisive clustering method works with the (top-down) strategy opposite to agglomerative method (Chavent, 1998; Emmendorfer & Canuto, 2021; Everitt et al., 2011). It optimally and progressively divides the largest cluster containing all objects into smaller clusters. The clustering process continues until n clusters, each with a single object, are obtained. The number of possible clustering, which is $2^{n-1} - 1$ in the first stage, increases exponentially and fast at each stage (Chavent, 1998; Zhong et al., 2008; Kaufman & Rousseeuw, 1990). Therefore, divisive methods are very time-consuming and computationally expensive methods (Zhong et al., 2008). This method, which has a time complexity that can be expressed by $O(2^n)$, is not preferred because of this significant disadvantage (Emmendorfer & Canuto, 2021; Xu & Wunsch, 2009). This method uses the same criteria as the ward method. In other words, it aims to produce homogeneous clusters with a low sum of squares of error (Reddy & Vinzamuri, 2014; Rencher, 2002). This method has two different approaches as *monothetic* and *polythetic* (Everitt et al., 2011; Xu & Wunsch, 2009). While the monothetic approach is based on division over a single variable, the polythetic approach is based on division, using all the variables at each stage (Rencher, 2002).

5. PRINCIPAL COMPONENT ANALYSIS

In real life, most of the data analyzed for information purposes is data with a large number of variables and called high dimensional (Maaten et al, 2009). The high dimensionality of the data can cause many problems such as noisy variables, highly correlated variables (multicollinearity), depending on the analysis technique used. Such problems may prevent us from obtaining robust analysis results. Two types of techniques are used to reduce dimensions; feature selection and feature extraction methods. Feature selection methods are the process of eliminating the least relevant variables in the data set. Feature extraction is the process of generating new variables that represent all data with appropriate combinations of observations and variables. Data and information loss are minimal in feature extraction methods, and therefore it is more advantageous than feature selection methods. ***The principal component analysis (PCA)*** is the most well known and widely used unsupervised method among feature extraction methods.

PCA, also known as Karhunen Loeve expansion or the Hotelling transformation, is one of the oldest multivariate methods frequently used in many scientific fields. Although its history back to Pearson (1901), the form used today was defined by Hotelling in 1933 (Abdi & Williams, 2010). PCA is the process of creating new variables, called the ***principal components***, to obtain important information by looking at the spread of observation values over variables (Karamizadeh et al, 2013). These new orthogonal and uncorrelated variables, i.e. principal components, are the linear combinations of the original variables in the data set. PCA is a fast running and computationally easy method (Karamizadeh et al., 2013).

The areas and applications where PCA is used are; pattern recognition, dimension reduction, computer vision, image compression, signal processing, video surveillance, face recognition, latent semantic indexing, ranking, and collaborative filtering (Karamizadehet al., 2013; Candes et al., 2011; Mudrova & Prochazka, 2005). The main purposes of the PCA are as follows (Tharwat, 2016; Abdi & Williams, 2010);

1. Extracting the most important information from the available data.
2. Reducing the number of dimensions by compressing the dataset.
3. Obtaining a simpler form that can describe the entire dataset.
4. Examining the structure and the relationships of observations and variables
5. Detecting outliers in the dataset.

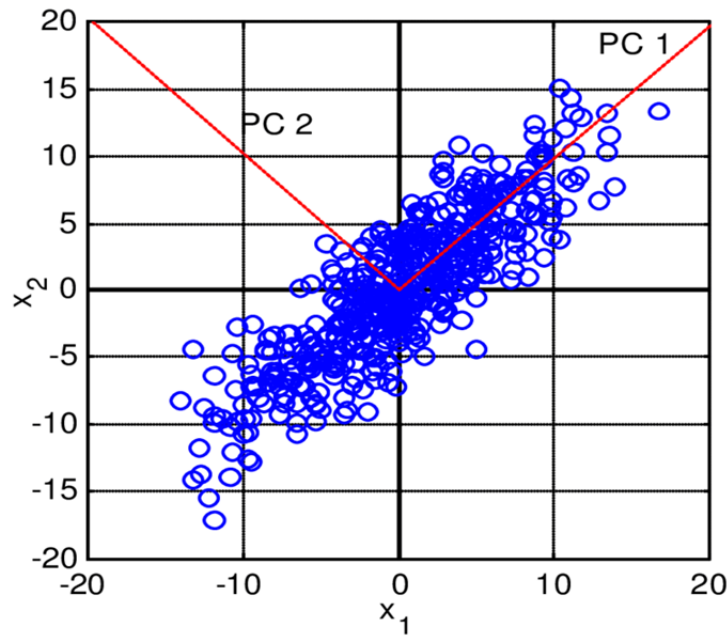


Figure 5. 1. Illustration of the principal component analysis (PCA) (Upadhyaya et al., 2002).

5.1. Mathematical Background For Principal Component Analysis

In PCA, the first principal component (PC1) is the linear combination of the observations which have maximum variance and the second principal component (PC2) is the linear combination of the observations which have maximum variance as orthogonal to the first principal component (PC1). $y_1, y_2, \dots, y_n$ are the observations vectors in p dimensional space, and $\bar{y}$ is the mean vector of the observation vectors. $y_1, y_2, \dots, y_p$ are the swarm of points for the variables. The first operations in the principal component analysis are the transformation and the rotation. In the transformation operation, the origins of each y_i are translated to $\bar{y}$ and in certain cases the transformation of $y_i - \bar{y}$ is applied. But it is generally assumed that each y_i is centralized. The main purpose of this procedure is finding the optimal

axes for the points. The second important procedure of the PCA is rotation. In the rotation phase, each of y_i is multiplied with the orthogonal matrix A and the new variables (PCs) $z_1, z_2, \dots, z_p$ are obtained.

The orthogonal matrix transforms each point y_i into a point z_i that is the same distance from the origin, and these new variables are uncorrelated.

$$A = \begin{pmatrix} a'_{11} \\ \vdots \\ a'_{1p} \end{pmatrix} ; \quad z_i = A * y_i , \quad i = 1, \dots, p \quad (5.1)$$

$$z_1 = a'_{11}y_1, z_2 = a'_{12}y_2, \dots, z_p = a'_{1p}y_p \rightarrow \text{Principal Components}$$

$$z_1 = a_{11}y_1 + a_{12}y_2 + \dots + a_{1p}y_p \rightarrow \text{PC1.}$$

$$z_2 = a_{21}y_1 + a_{22}y_2 + \dots + a_{2p}y_p \rightarrow \text{PC2.}$$

$\vdots$

$$z_p = a_{p1}y_1 + a_{p2}y_2 + \dots + a_{pp}y_p \rightarrow \text{PCp.}$$

If we handle the other details of PCA; the sample covariance matrix of z (S_z) is calculated by using the orthogonal matrix (A) and the sample covariance matrix of $y_1, y_2, \dots, y_n$ (S).

$$S_z = ASA' = \begin{pmatrix} S_{z_1}^2 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & S_{z_p}^2 \end{pmatrix} \quad (5.2)$$

Here $S_{z_i}^2 = \lambda_i$, and λ_i , $i = 1, \dots, p$ are the eigenvalues of S which equals to variance values of the principal components. In PCA, the variance values of the components has the arrangement as $\lambda_1 > \lambda_2 > \dots > \lambda_p$. Therefore, PC1. has the largest variance value and PCp. has the smallest variance value respectively. The fact that the variance value of the first principal component is the largest is since it represents most of the variables proportionally.

The proportion of variance explained = $\frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^p \lambda_i}$; k = number of selected component (5.3)

5.1.1. How to decide the number of the principal components?

1. Starting from the first component, the components whose sum of the explained variance percentages exceed %80 are selected. The ratio of the variance explained by each of the principal components is expressed by the formula below (Gempertine, 2006);

$$\% Var z_i = \frac{\lambda_i}{\sum_{i=1}^p \lambda_i} * 100\% \quad (5.4)$$

2. Starting from the first component, the number of components whose λ value is greater than the average $\bar{\lambda} = \sum_{i=1}^p \lambda_i / p$ value is selected.
3. In the graph in which the λ values of the components are included (Scree graph), the number of components can be decided by looking at the point where the natural break (Elbow test) between the large and small values occurs (Rencher, 2002; Abdi & Williams, 2010)

If the correlation between variables in the data set is high, the k value (number of components selected) will be much smaller than the p value. But otherwise, if the correlation between variables is low, the value of k will be close to p. This case will reduce the effectiveness of the principal component analysis in terms of dimension reduction (Rencher 2002).

6. THE INDEXES FOR COMPARING HIERARCHICAL CLUSTERING

The measures used to evaluate clustering results are called cluster validity indices. In general, these indices may be based on different approaches; internal criteria, external criteria and relative criteria (Halkidi et al., 2001). External criteria are based on a priori knowledge for clustering, while internal criteria are based on intra-cluster distance calculations. Relative criteria are used if there is a parameter difference for the same method (Altunkaynak, 2019). Here, CPCC is an internal criteria and, FM_index and Baker's Gamma Coefficient are external criteria. The validity indices for comparing hierarchical clustering results mainly depend on the topology of the dendrogram, the labels on nodes, and the heights (Fowlkes & Mallows, 1983). The most well known metrics for comparing hierarchical clustering results are CPCC (Cophenetic Correlation Coefficient), FM_index (Fowlkes and Mallows index) and Cor Baker's Gamma Coefficient (Goodman and Kruskal's Gamma Coefficient).

6.1.CPCC (Cophenetic Correlation Coefficient):

Cophenetic correlation coefficient was firstly proposed by Sokal and Rohlf in 1962. CPCC is expressed as the correlation value between the cophenetic matrix created according to the height values in dendrogram and similarity matrix (Triayudi & Fitri, 2018; Silva & Dias, 2013; Nino & Berzal, 2019). In measuring correlation while applying the CPCC method, Pearson, Spaerman, and Kendall correlation techniques can be used for calculations. CPCC takes the value in the range of $[-1,1]$ (Triayudi & Fitri, 2018). The high values of the coefficient express the high similarities between the cophenetic matrix and the distance matrix, while the low values close to zero express the low similarities (Nino & Berzal, 2019). CPCC aims to measure how well hierarchical clustering is performed (Kumar & Toshniwal, 2016; Rao & Srinivas, 2006; Saraçlı et al., 2013).

$$CPCC = \frac{\sum_{i < j} (d_{ij} - \bar{d})(d_{ij}^* - \bar{d}^*)}{\sqrt{\sum_{i < j} (d_{ij} - \bar{d})^2 \sum_{i < j} (d_{ij}^* - \bar{d}^*)^2}} \quad (6.1)$$

d_{ij} ; distance between the pairs (i,j)

d_{ij}^* ; cophenetic distance between the pairs (i,j)

$\bar{d}$; average distance for similarity matrix

$\bar{d}^*$; average distance for cophenetic matrix

6.2. FM_index (Fowlkes and Mallows index):

FM_index (Fowlkes and Mallows index) measures the similarity between the clustering results (Nino & Berzal, 2019). It uses the number of cluster as the parameter in the calculations. Therefore, different results are obtained with different cluster number input. FM_index takes the value between the range of [0,1] (Fowlkes & Mallows, 1983). The values close to 1 express the high similarity and the values close to 0 express the low similarity between the clustering results.

$$FM = \sqrt{\frac{TP}{TP+FP} * \frac{TP}{TP+FN}} \quad (6.2)$$

TP : The count of pairs which are in the same cluster both in C_1 and C_2 .

FP : The count of pairs which are in the same cluster in C_1 but not in C_2 .

FN : The count of pairs which are in the same cluster in C_2 but not in C_1 .

6.3.Baker's Gamma Coefficient (Goodman and Kruskal's Gamma Coefficient):

The Goodman and Kruskal's Gamma Coefficient was proposed in 1954 by Goodman & Kruskal to measure the relationship depending on the probabilities π_c and π_d (Hryniewicz, 2006). In generally γ coefficient is used for ordinal variables and it takes the values between the range of $[-1,1]$ (Kılıç, 2021; Hryniewicz, 2006). Here is the values close to 0 express the low relationship and the values close to -1 and 1 express the negative and positive relationship respectively.

$$\gamma = \frac{\pi_c - \pi_d}{\pi_c + \pi_d} \quad (6.3)$$

π_c : The probability of concordant pairs of observations.

π_d : The probability of discordant pairs of observations.

7. MATERIALS AND METHODS

Hierarchical clustering methods automatically group objects according to certain criteria without requiring the number of clusters parameter. Principal component analysis, on the other hand, is one of the most frequently used feature extraction methods, which performs dimension reduction with the least information loss. In this thesis, hierarchical linkage techniques (Single, Complete, Average, Centroid, Median and Ward's Criterion) are used as clustering methods and clustering processes are applied to the high dimensional data sets separately for each method. In another analysis, the same clustering procedures are applied to the same data sets again. However, this time, PCA is applied to all data sets as a dimension reduction method before the clustering process. The main purpose here is to evaluate the effects of PCA on the hierarchical clustering process and its results. FM_index, CPCC and Baker's gamma coefficient metrics are used to measure similarities between dendrograms and to evaluate the results. In addition, *tanglegrams* are used for each method to observe the change in dendrograms. Tanglegrams are the plots that show two hierarchical clustering or two dendrograms comparatively. The information about the statistical programming language R and the packages used in the analysis is detailed in the next section.

7.1. R Statistical Programming:

R is an open-source code language where statistical calculations can be made and graphical techniques can be applied (r-project.org, 2021). R language was designed by Ross Ihaka and Robert Gentleman in 1993 (wikipedia.org, 2021). R, which is generally used for statistical software development and data analysis, can receive new updates continuously with the advantage of being an open-source language. Thanks to the packages and libraries it contains, a wide variety of methods can be applied and rich graphical representations can be created in R Programming. Denextend and cluster packages in R used at the analysis stages of the thesis. Cluster package is generally used for classical clustering analysis. Denextend package, on the other hand, is used for advanced clustering processes, dendrogram and tanglegram representations. In addition, the coefficients used to evaluate the clustering results are also included in the denextend package.

7.2. Information About Datasets:

The data sets used in the analysis are compiled from the United Nations data platform (data.un.org, 2021). The missing observation values in the data sets are completed from the relevant international data platforms (data.unicef.org, 2021; data.worldbank.org, 2021; ourworldindata.org, 2021; worldometers.info, 2021). While the objects of the datasets are created from the relevant countries, the variables in the datasets are created from the relevant indicator variables. The indicators and the variables used for the datasets are given below in the relevant tables.

Table 7. 1. The variables which include general information about the countries.

General Information = (GI)
1. Population (000, 2020)
2. Pop. density (per km2, 2020)
3. Surface area (km2)
4. Sex ratio (m per 100 f)
5. Exchange rate (per US\$)

Table 7. 2. The variables which include economical information about the countries.

Economic Indicators = (EI)
6. GDP: Gross domestic product (million current US\$)
7. GDP growth rate (annual %, const. 2015 prices)
8. GDP per capita (current US\$)
9. Economy: Agriculture (% of Gross Value Added)
10. Economy: Industry (% of Gross Value Added)
11. Economy: Services and other activity (% of GVA)
12. Employment in agriculture (% of employed)
13. Employment in industry (% of employed)
14. Employment in services (% employed)
15. Unemployment (% of labour force)
16. Labour force participation rate (female/male pop. %)
17. CPI: Consumer Price Index (2010=100)
18. Agricultural production index (2004-2006=100)
19. International trade: exports (million current US\$)
20. International trade: imports (million current US\$)
21. International trade: balance (million current US\$)
22. Balance of payments, current account (million US\$)

Table 7. 3. The variables which include social information about the countries.

Social Indicators = (SI)
23. Population growth rate (average annual %)
24. Urban population (% of total population)
25. Fertility rate, total (live births per woman)
26. Life expectancy at birth (females/males, years)
27. Population age distribution (0-14/60+ years old, %)
28. International migrant stock (000/% of total pop.)
29. Refugees and others of concern to UNHCR (000)
30. Infant mortality rate (per 1 000 live births)
31. Health: Current expenditure (% of GDP)
32. Health: Physicians (per 1 000 pop.)
33. Education: Government expenditure (% of GDP)
34. Education: Primary gross enrol. ratio (f/m per 100 pop.)
35. Education: Secondary gross enrol. ratio (f/m per 100 pop.)
36. Education: Tertiary gross enrol. ratio (f/m per 100 pop.)
37. Intentional homicide rate (per 100 000 pop.)
38. Seats held by women in national parliaments (%)

Table 7. 4. The variables which include environmental and infrastructural information about the countries.

Environment and Infrastructure Indicators = (ENI)
39. Individuals using the Internet (per 100 inhabitants)
40. Threatened species (number)
41. Forested area (% of land area)
42. CO2 emission estimates (million tons/tons per capita)
43. Energy production, primary (Petajoules)
44. Energy supply per capita (Gigajoules)
45. Tourist/visitor arrivals at national borders (000)
46. Important sites for terrestrial biodiversity protected (%)

While creating the data sets, the observation values of the relevant variables of the countries are transferred to the excel environment. The number of variables and the number of objects each data set has are different from each other. Informations belong to the datasets are detailed below.

7.2.1. Dataset 1. (African Dataset):

Dataset1 consists of **22 variables** (dimension), **54 objects** and a total of **1188 observations**. The variables of the data set consist of the general information (**GI**) variables and the economic indicator (**EI**) variables of the countries and the objects of the data set consist of **African countries**. African countries in the dataset are listed in the order below.

African Countries:

Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Nigeria, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, South Africa, South Sudan, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe.

7.2.2. Dataset 2. (European Dataset):

Dataset2 consists of **38 variables** (dimension), **39 objects** and a total of **1482 observations**. The variables of the data set consist of the general information (**GI**) variables, the economic indicator (**EI**) variables and the social indicator (**SI**) variables of the countries and the objects of the data set consist of **European countries**. European countries in the dataset are listed in the order below.

European Countries:

Albania, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, North Macedonia, Norway,

Poland, Portugal, Republic of Moldova, Romania, Russian Federation, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Kingdom.

7.2.3. Dataset 3. (Asian Dataset):

Dataset3 consists of **46 variables** (dimension), **48 objects** and a total of **2208 observations**. The variables of the data set consist of the general information (**GI**) variables, the economic indicator (**EI**) variables, the social indicator (**SI**) variables and the environmental and infrastructural indicator (**EII**) variables of the countries and the objects of the data set consist of **Asian countries**. Asian countries in the dataset are listed in the order below.

Asian Countries:

Afghanistan, Armenia, Azerbaijan, Bahrain, Bangladesh, Bhutan, Brunei Darussalam, Cambodia, China, Cyprus, Georgia, Hong Kong SAR, India, Indonesia, Iran, Iraq, Israel, Japan, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lao People's Dem. Rep., Lebanon, Malaysia, Maldives, Mongolia, Myanmar, Nepal, Oman, Pakistan, Philippines, Qatar, Republic of Korea, Saudi Arabia, Singapore, Sri Lanka, State of Palestine, Syrian Arab Republic, Tajikistan, Thailand, Timor-Leste, Turkey, Turkmenistan, United Arab Emirates, Uzbekistan, Viet Nam, Yemen.

7.3. The Steps of the Analysis:

The information of the analysis phases performed for each method is given in order below. All the methods and stages used in the analysis are applied in the same way for each dataset. All the phases of the analysis are carried out in R programming, and the Cluster package and Denextend package are used for the clustering analysis procedure.

- **Normalization:** To eliminate the negativities caused by the measurement and unit differences in the data sets, the normalization process is applied to the entire data set.
- **Determining the distance matrix:** The distance matrix is created according to the determined distance measurement method with the normalized dataset.
- **Applying the clustering algorithms:** The operation is performed with the determined clustering algorithm by using the calculated distance matrix.
- **Calculating the principal components:** The principal component analysis is applied to the normalized dataset and principal components are obtained as much as the number of variables.
- **Obtaining variance values and variance explanation values of the principal components:** The variance values and variance explanation rates of the principal components are calculated. These values can be plotted using a scree plot.
- **Deciding the number of components to add to the dataset:** According to the criteria suggested in principal component analysis; the number of components to be selected is decided by looking at the variances of the principal components, the explained variance ratios and the scree plots.
- **Normalization after applying PCA:** The determined principal components are added to the data set and normalization is performed.
- **Determining the distance matrix after applying PCA:** The distance matrix is created using the determined distance metric with the PCA applied and normalized data set.
- **Applying the clustering algorithms after applying PCA:** The clustering process is carried out with a determined clustering algorithm after applying PCA.

- ***Creating the tanglegrams for each method:*** The tanglegrams are created for comparison of the dendrograms for each hierarchical clustering method.
- ***Obtaining coefficient values to evaluate the results:*** The coefficients determined to evaluate the similarity of the results and dendrograms obtained without the application of PCA and after the application of PCA are calculated for each method.



8. RESULTS AND DISCUSSIONS

Analysis results and graphical representations for each dataset are given below under the relevant headings. Before clustering analysis results and dendrogram comparisons, principal component analysis results are evaluated in terms of dimension reduction. Afterward, clustering analysis results are comparatively evaluated for each method separately.

8.1. Applying Principal Component Analysis to the African Dataset:

The African dataset consists of **22** variables and **54** objects. To evaluate the relationship between the number of components added to the data set in principal component analysis and the correlation level between the variables in the data set, the colored correlation matrix for the variables of the African dataset is given below.

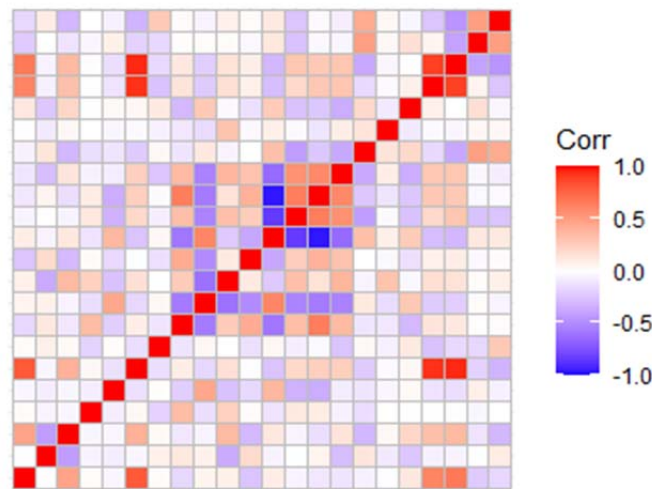


Figure 8. 1. Colored correlation matrix of the variables for dataset 1.

The mean absolute correlation between variables for the African dataset was calculated as **0.2426**. Due to the low level of correlation between the variables in the data set, it is seen that light colors are dominant in the colored correlation matrix in **figure 8.1**. The variance values of the principal components for the African dataset are given below with the scree plot and the table.

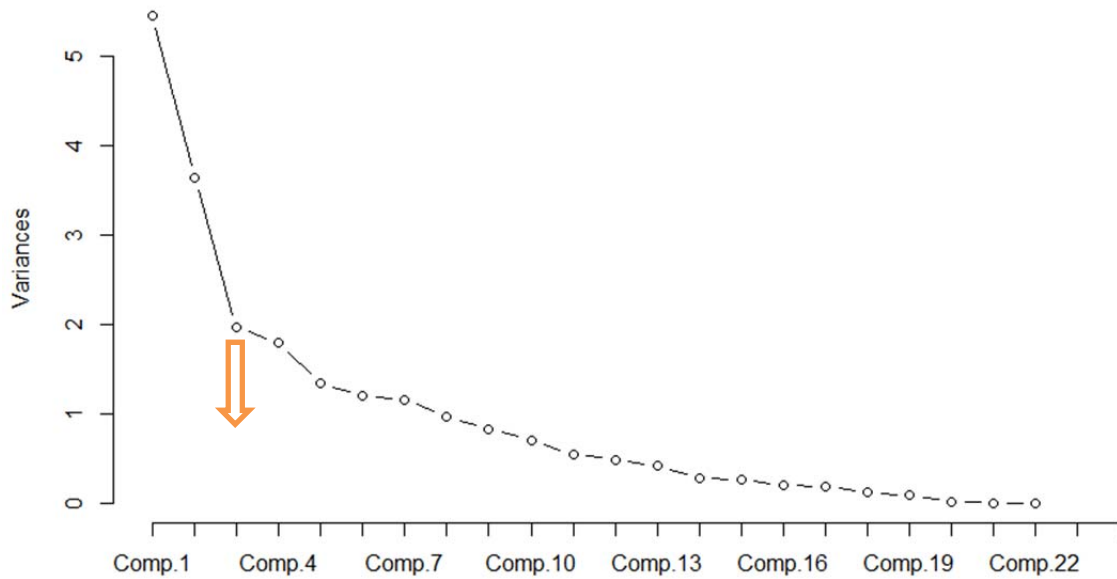


Figure 8.1.1. Scree plot of the variance values of the principal components for dataset 1.

Table 8.1.1. The variance values of the principal components for dataset 1.

Variance Values of the Principle Components								
comp.1	comp.2	comp.3	comp.4	comp.5	comp.6	comp.7	comp.8	comp.9
5,44E+00	3,64E+00	1,97E+00	1,79E+00	1,34E+00	1,20E+00	1,15E+00	9,63E-01	8,28E-01

In principal component analysis, three approaches that can be used to decide the number of components are described in **chapter 5**. For the elbow test, which is one of these approaches, when the scree plot which includes the variance values of the principal components given above is examined, it is seen that the **natural breakpoint** in the graph coincides with the 3rd principal component. However, in **table 8.1.1**, when the variance explained percentages of the principal components are examined, it is seen that the **first 3 components** can explain only **%51.11** of the total variance in the dataset. This ratio is not sufficient to represent the entire data set.

Another approach that can be used to decide on the number of components is to select components with variance above the mean-variance value. For the African data set, the mean-variance value ($\bar{\lambda}$) of the principal components is calculated as **0.9815**, and it is observed that the variance values of the first 7 principal components are above this value. When evaluated according to the variance explained percentages of the components, it is observed that the **first 7 components** can explain **%76.35** of the total variance in the data set.

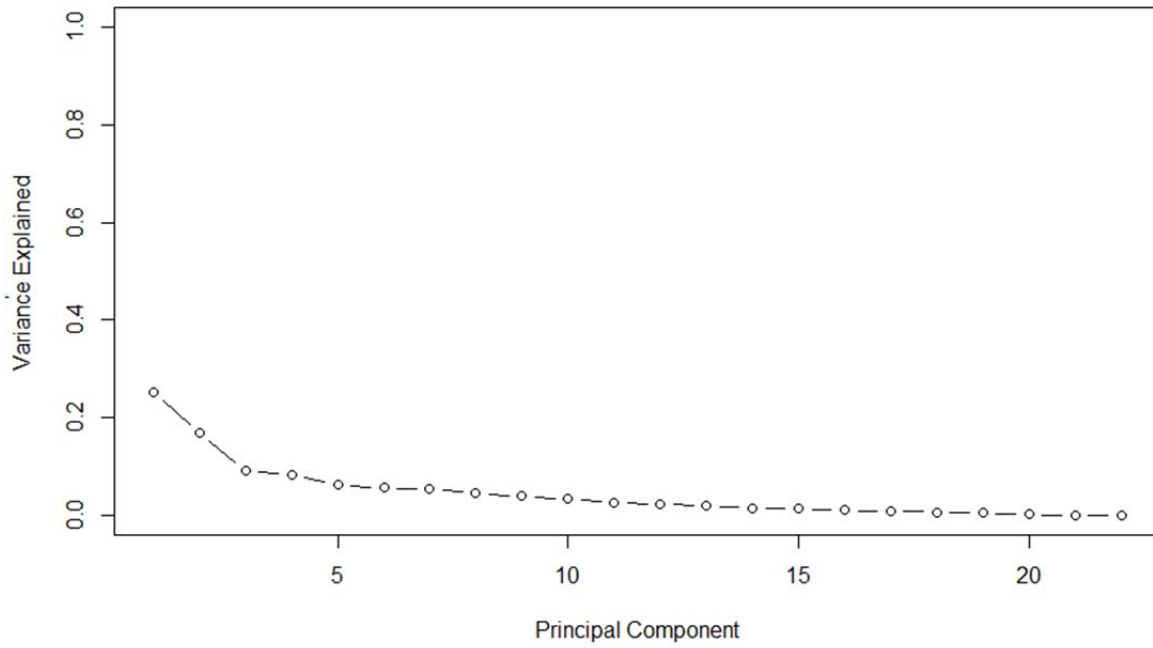


Figure 8.1. 2. Scree plot of the variance explained rates of the principal components for dataset 1.

Table 8.1. 2. The variance explained percentages of the principal components for dataset 1.

Variance Explained Percentages of the Principle Components								
comp.1	comp.2	comp.3	comp.4	comp.5	comp.6	comp.7	comp.8	comp.9
2,52E+01	1,68E+01	9,11E+00	8,28E+00	6,22E+00	5,55E+00	5,34E+00	4,46E+00	3,83E+00

The third approach is to select components whose sum of variance explained percentages exceeds the **80%** specific ratio. Above, the scree plot and the table containing the variance explained percentages of the principal components for the African dataset are given. When the values in the **table 8.1.2.** are examined, it is seen that the first 8 principal components should be selected to achieve this specific ratio. It is observed that the **first 8 components** can explain **%80.81** of the total variance in the data set. If the results of these three approaches

applied in deciding on the number of components to be selected are evaluated, it is clear that the average variance and 80% specific ratio approaches yield consistent results. However, it has been observed that the elbow test does not give reliable and robust results, and it gives different results than the other two approaches. As a result of all these evaluations, the number of components to be selected for a robust analysis process is determined as 8.

8.2. Clustering Results of African Countries to the Economic Indicators:

The comparisons of clustering results with PCA and without PCA for each method are given below, with tanglegrams and coefficients for the African dataset. The integer labels in the tanglegrams represent African countries in alphabetical order. Additionally, hierarchical clustering results, in which African countries are grouped according to the economic indicator variables, are also included in the appendices as dendrograms.

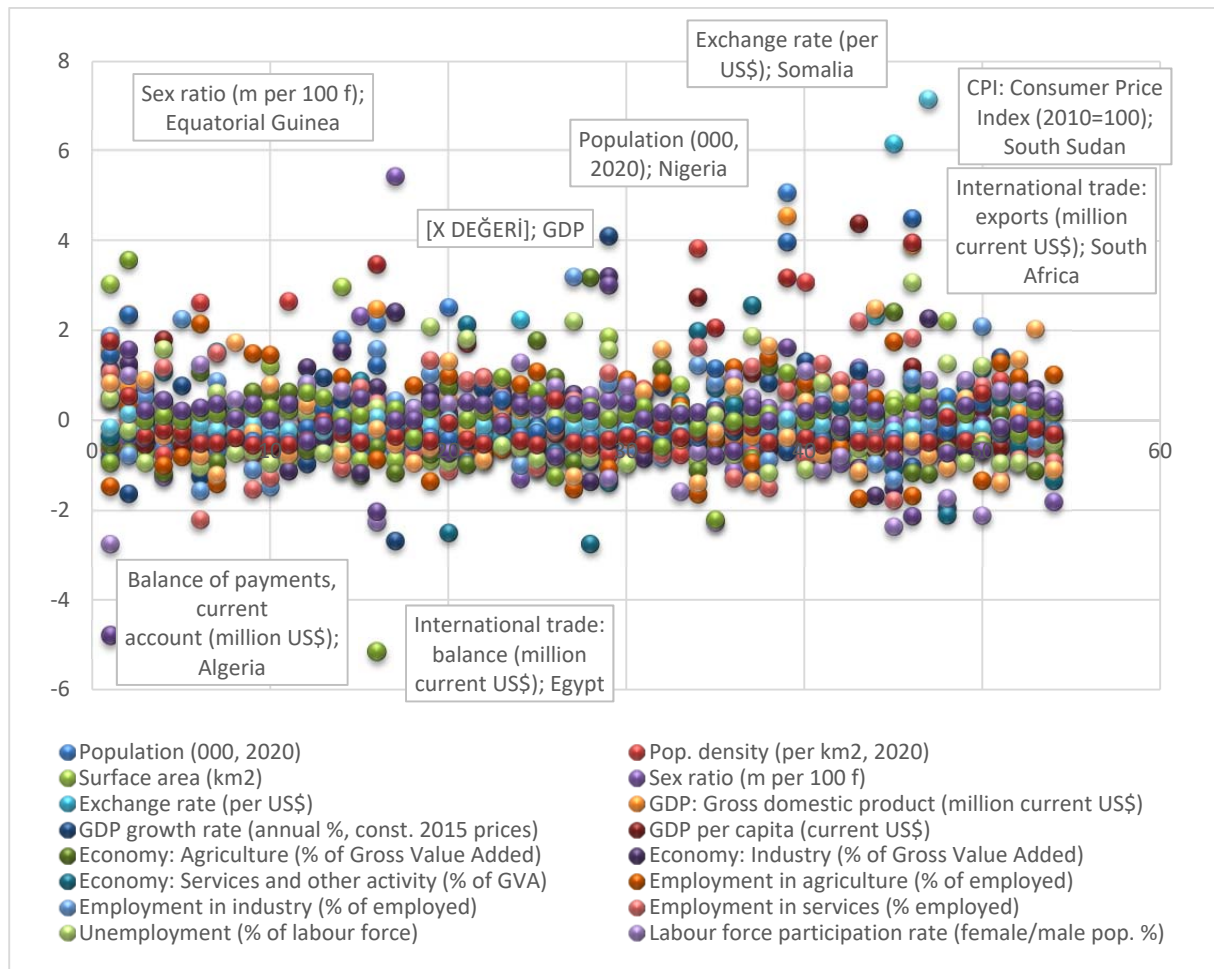


Figure 8. 2. Scatterplot of the normalized African dataset for the economic indicator variables.

Above is a scatterplot showing how African countries are distributed according to the economic indicator variables. When the distribution of the normalized data in the graph is examined, it can be seen that there are so many outliers in the data set. Some of these outliers are labeled on the scatterplot.

The variance value of a data point expresses how far that data point is from the arithmetic mean of all observation values. That is, the further away a data point is from the mean value, the higher the variance value it will have. Because of this fundamental reason, outliers have high variance in the data set. Outliers in the data set, depending on their number, directly affect the correlation between variables negatively. Correlation shows the strength and direction of the relationship between two variables. Because of outliers with high variance, even variables that have high correlation may appear to be low correlated. This situation creates negative effects on many methods, especially the techniques in which variance and correlation are effective. Here, the reason for the existence of outliers should be well

investigated. This situation can have many reasons, such as measurement error or real values. Depending on these reasons, the actions to be taken vary, but if the observation values are real, no change is made and this situation is taken into account in the evaluations.

More deterministic methods other than scatterplot should be used to express the high values of the objects on the variables as outliers. One of the most robust ways to identify outliers in a data set is to plot a Box and Whisker plot for each variable. Potential outlier values are determined based on the quartiles and median values in the boxplot. A large number of outliers in the data set can be clearly observed in Figure 8.3.



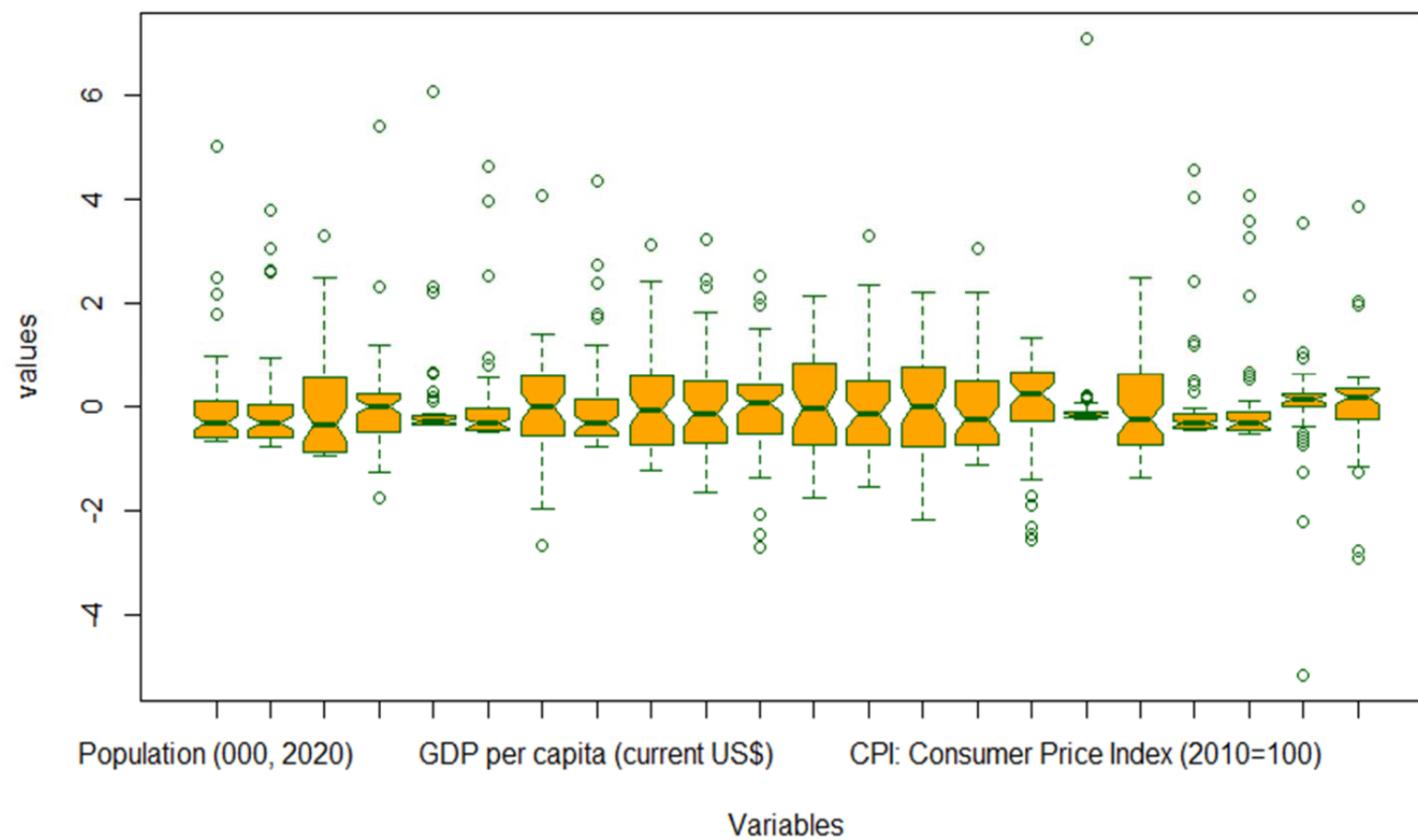


Figure 8.3. Boxplots for African dataset.

8.2.1. Comparative Results for Single Linkage Method:

For the single linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.75**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.98**, and Fowlkes and Mallows index (**FM_index**) value is **0.94**. For the African dataset, the tanglegram that includes the comparison of the dendrograms of the single linkage method is shown below in **Figure 8.2.1**. While most of the objects in the dendrograms have small shifts, some objects have large shifts, such as 46 and 39 (South Africa, Nigeria). The objects 47 and 29 (South Sudan, Libya) remained without change, and 29 is considered as a separate cluster. In addition, it is seen that the dendrogram formed after PCA has a more neat hierarchical structure.

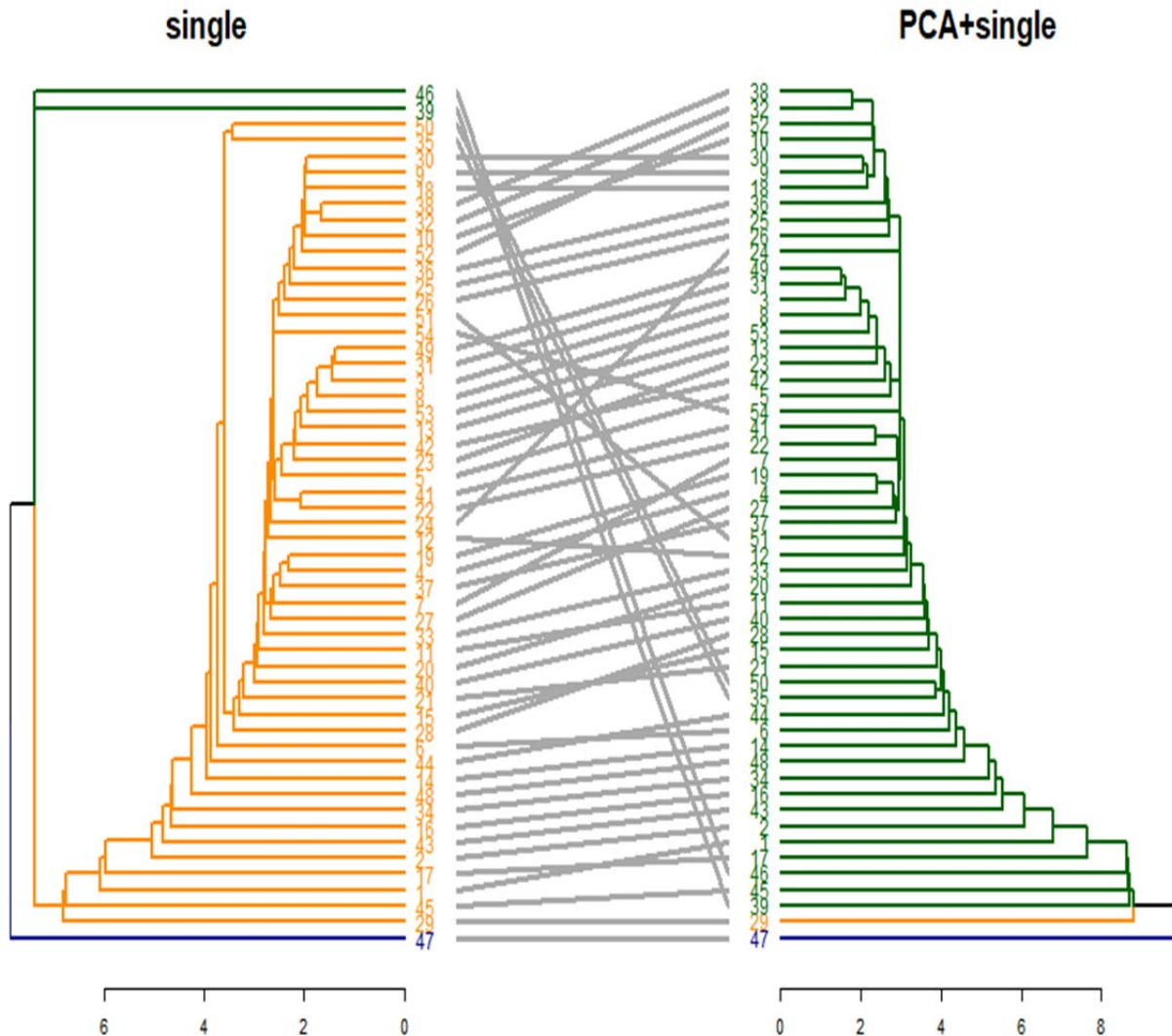


Figure 8.2. 1. Tanglegram for comparison dendrograms of single linkage method and PCA+single linkage method.

Table 8.2. 1. Clustering results for single linkage and PCA+single linkage method.

Single linkage		
Cluster 1	Cluster 2	Cluster 3
Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	Nigeria, South Africa	South Sudan

PCA+Single linkage		
Cluster 1	Cluster 2	Cluster 3
Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Nigeria, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, South Africa, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	Libya	South Sudan

The CPCC value above is calculated by considering the matrices created with the branch lengths of the two dendrograms. While the CPCC value for tanglegrams expresses the similarity between the two dendrograms, the CPCC values for dendrograms show the relationship and similarity between the distance matrix and the cophenetic matrices for each dendrogram. For the first dendrogram (single) the CPCC value is **0,8519** and for the second dendrogram (PCA+single) the CPCC value is **0,8704**. Compared to other countries South Sudan has very high value for the consumer price index; Libya has high values for GDP growth rate, balance of payments, economy industry; Nigeria has high values for population, GDP, international trade exports and imports; South Africa has high values for international trade exports and imports, unemployment, employment in services variables.

8.2.2. Comparative Results for Complete Linkage Method:

For the complete linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.96**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.98**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the African dataset, the tanglegram that includes the comparison of the dendrograms of the complete linkage method is shown below in **Figure 8.2.2**. While no displacement is observed in almost half of the objects, low and moderate shifts are observed in other objects. For both dendrograms, it can be said that the cluster contents of the dendrograms will be the same when the number of clusters is not high.

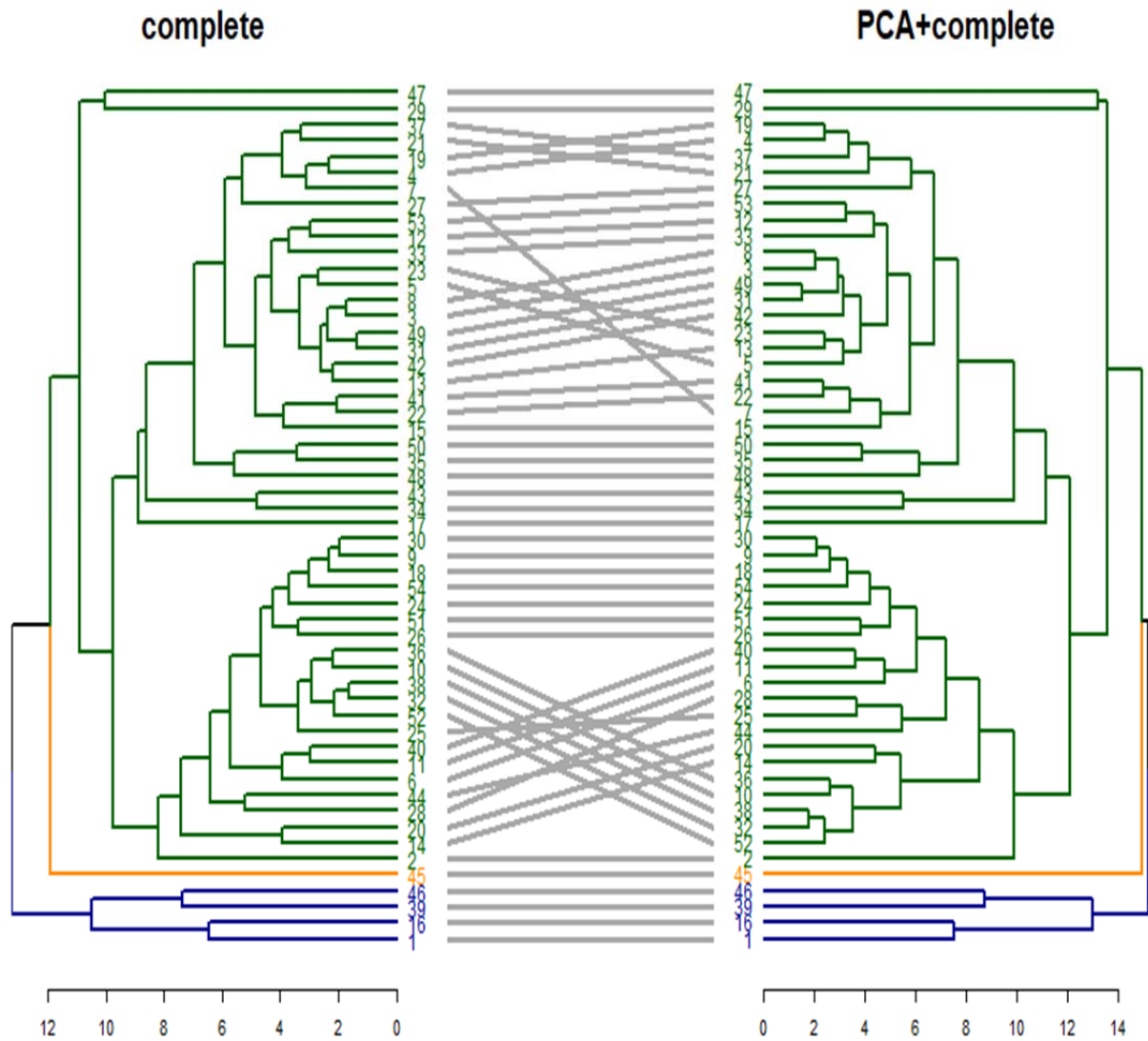


Figure 8.2. 2. Tanglegram for comparison dendrograms of complete linkage method and PCA+complete linkage method.

Table 8.2. 2. Clustering results for complete linkage and PCA+complete linkage method.

Complete linkage		
Cluster 1	Cluster 2	Cluster 3
Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, South Sudan, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	Algeria, Egypt, Nigeria, South Africa	Somalia

PCA+Complete linkage		
Cluster 1	Cluster 2	Cluster 3
Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, South Sudan, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	Algeria, Egypt, Nigeria, South Africa	Somalia

For the first dendrogram (complete) the CPCC value is **0,7724** and for the second dendrogram (PCA+complete) the CPCC value is **0,7261**. It is seen above that the cluster contents of both dendrograms are the same. If evaluated together with the scatterplot, it is known that the countries in cluster 2 and cluster 3 have higher or lower values than other countries in terms of some variables: Algeria has very low balance of payments and high surface area; Egypt has very low international trade balance and high international trade imports and GDP; Somalia has very high exchange rate.

8.2.3. Comparative Results for Average Linkage Method:

For the average linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.97**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.98**, and Fowlkes and Mallows index (**FM_index**) value is **0.92**. For the African dataset, the tanglegram that includes the comparison of the dendrograms of the average linkage method is shown below in **Figure 8.2.3**. No displacement is observed in a small group of objects, but most of the objects have low-level shifts. The objects 44 and 28 (Sierra Leone, Liberia) shifted from the end to the middle in the dendrograms.

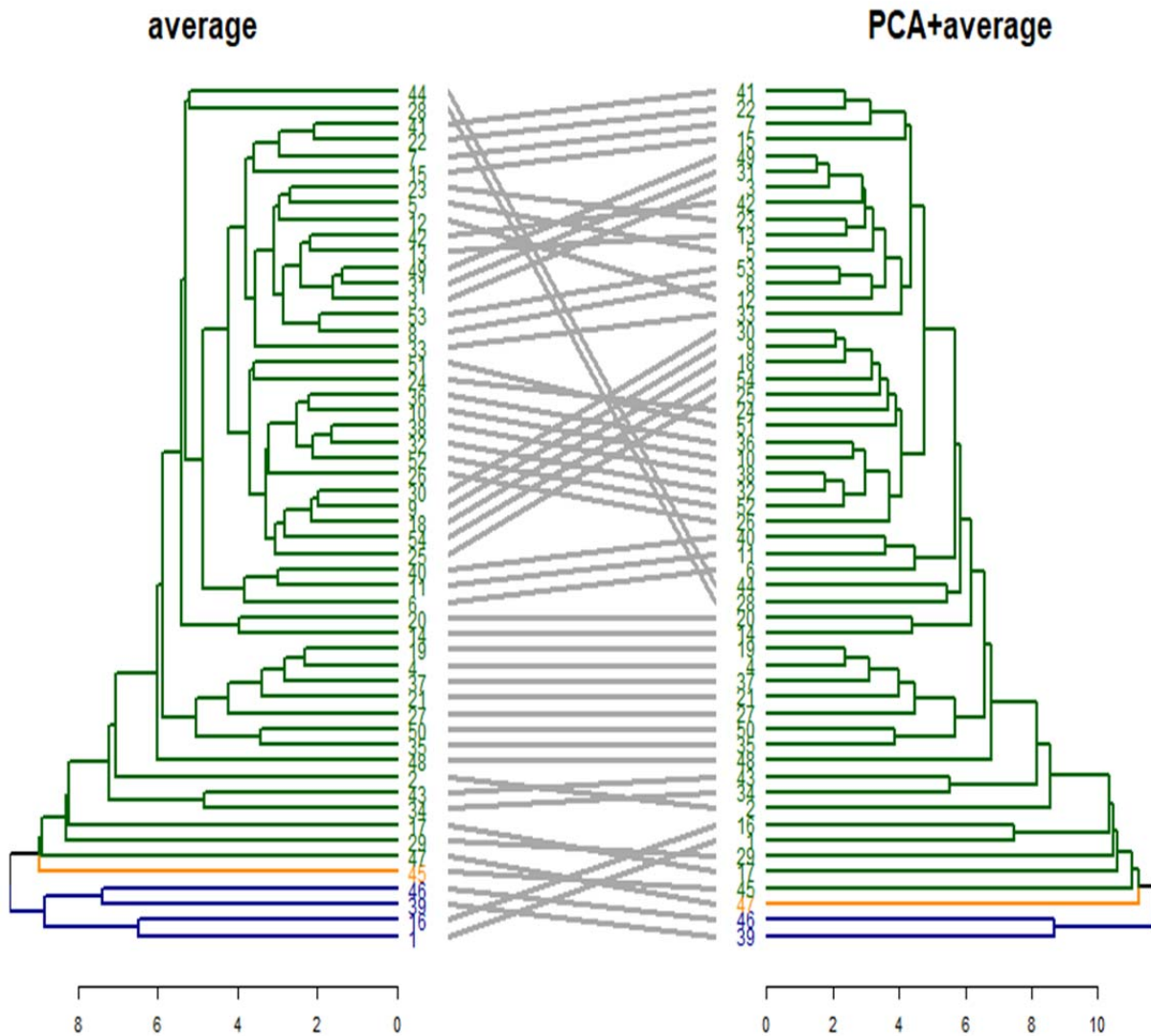


Figure 8.2. 3. Tanglegram for comparison dendrograms of average linkage method and PCA+average linkage method.

Table 8.2. 3. Clustering results for average linkage and PCA+average linkage method.

Average linkage		
Cluster 1	Cluster 2	Cluster 3
Algeria, Egypt, Nigeria, South Africa	Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	South Sudan

PCA+Average linkage		
Cluster 1	Cluster 2	Cluster 3
Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	Nigeria, South Africa	South Sudan

For the first dendrogram (average) the CPCC value is **0,9089** and for the second dendrogram (PCA+average) the CPCC value is **0,9211**. The CPCC values are quite high, this meaning that the distance matrix and cophenetic matrices are highly congruent and similar. It is seen above that the cluster contents of both dendrograms are not the same.

8.2.4. Comparative Results for Centroid Linkage Method:

For the centroid linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.99**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.39**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the African dataset, the tanglegram that includes the comparison of the dendrograms of the centroid linkage method is shown below in **Figure 8.2.4**. Very small shifts are observed for almost all of the objects, except for a few stable. It is noteworthy here that the BGCC calculated according to the probabilities of concordant pairs and discordant pairs of the objects is quite low.

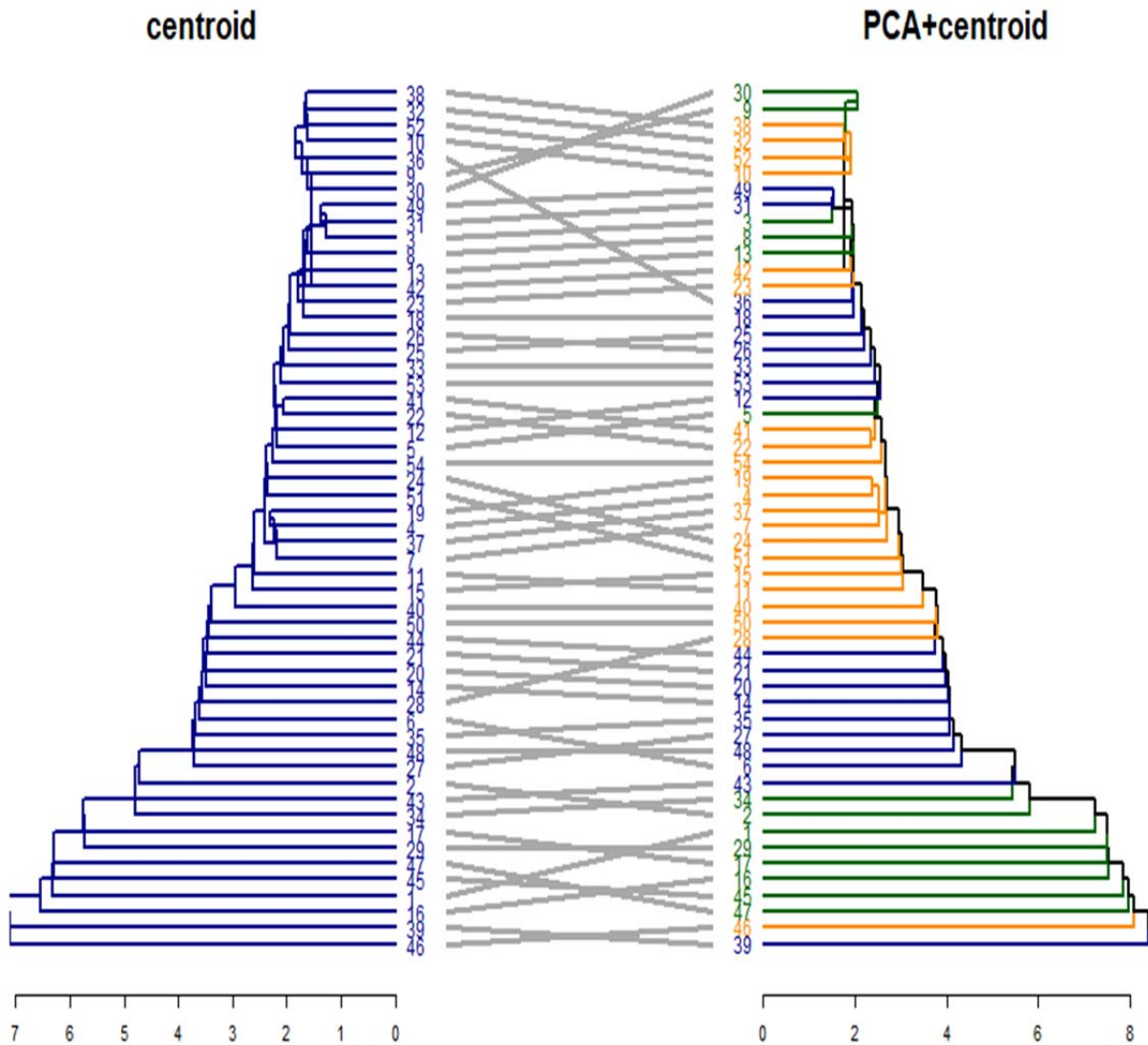


Figure 8.2. 4. Tanglegram for comparison dendrograms of centroid linkage method and PCA+centroid linkage method.

Table 8.2. 4. Clustering results for centroid linkage and PCA+centroid linkage method.

Centroid linkage		
Cluster 1	Cluster 2	Cluster 3
Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, South Sudan, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	Nigeria	South Africa

PCA+Centroid linkage		
Cluster 1	Cluster 2	Cluster 3
Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Somalia, South Sudan, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	Nigeria	South Africa

For the first dendrogram (centroid) the CPCC value is **0,8977** and for the second dendrogram (PCA+centroid) the CPCC value is **0,9102**. The CPCC values of the both dendrograms are quite high and the cluster contents are the same. When we evaluated the clustering results, tanglegram and scatterplot, centroid linkage method is less sensitive to outliers compared to other methods. Because in the other methods, some countries like Somalia and South Sudan are evaluated separated clusters as they have just one very high value. It can be seen from the results that this is not the case in the centroid linkage method. This feature is very important to obtain more robust results.

8.2.5. Comparative Results for Median Linkage Method:

For the median linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.92**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.42**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the African dataset, the tanglegram that includes the comparison of the dendrograms of the median linkage method is shown below in **Figure 8.2.5**. While small shifts are observed in most of the objects, large shifts are observed in some objects. As with the centroid method, the BGCC value is low for the median method.

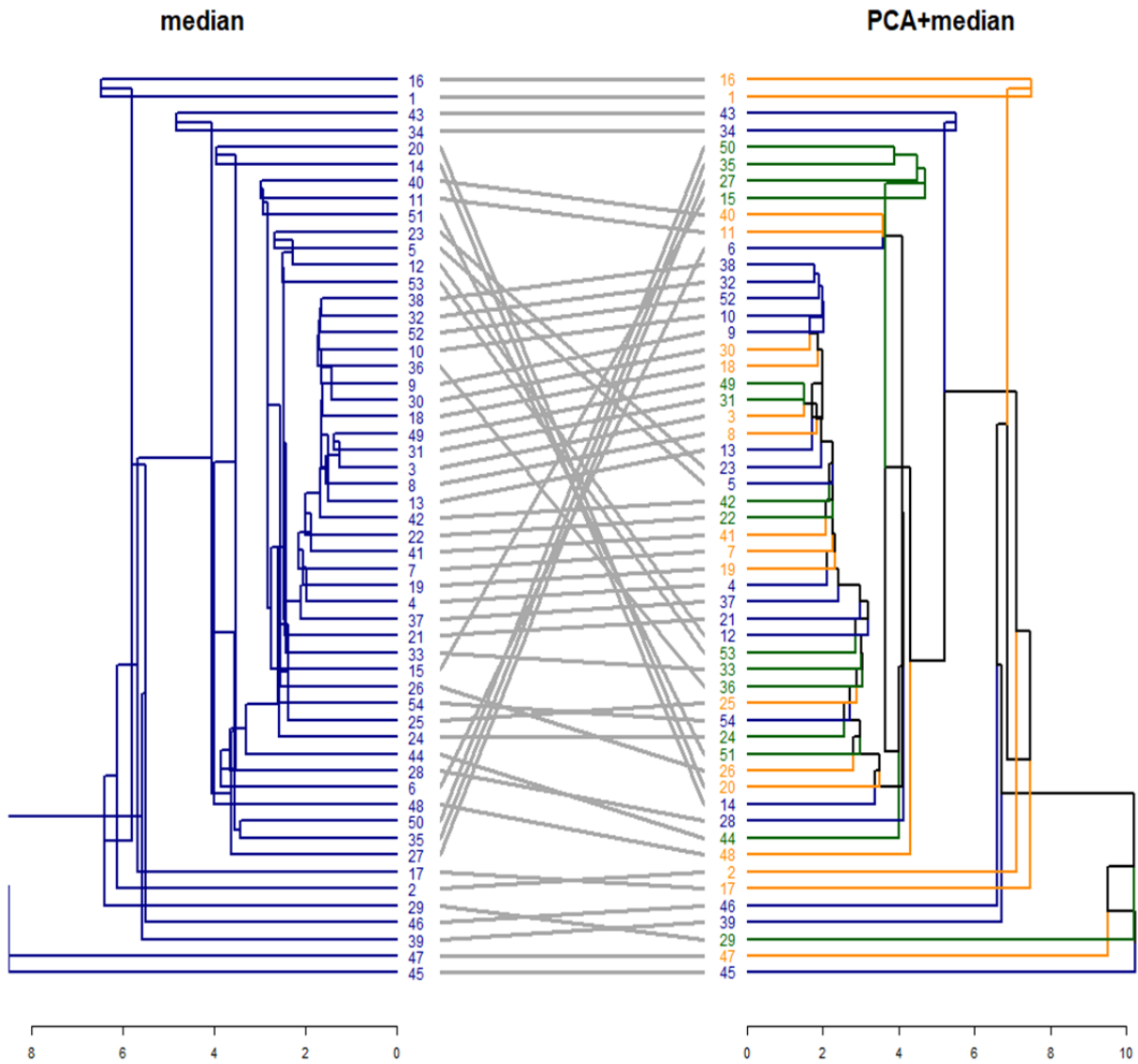


Figure 8.2. 5. Tanglegram for comparison dendrograms of median linkage method and PCA+median linkage method.

Table 8.2. 5. Clustering results for median linkage and PCA+median linkage method.

Median linkage		
Cluster 1	Cluster 2	Cluster 3
Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Nigeria, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, South Africa, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	Somalia	South Sudan

PCA+Median linkage		
Cluster 1	Cluster 2	Cluster 3
Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'Ivoire, Dem. Rep. Of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Nigeria, Rwanda, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, South Africa, Sudan, Togo, Tunisia, Uganda, United Rep. Of Tanzania, Zambia, Zimbabwe	Somalia	South Sudan

For the first dendrogram (median) the CPCC value is **0,7948** and for the second dendrogram (PCA+median) the CPCC value is **0,8365**. The CPCC values of both dendrograms are not very high, but the cluster contents are the same. Here Somalia and South Sudan both have just one extreme high value for one variable and so they are evaluated as separated clusters.

8.2.6. Comparative Results for Ward's Criterion Linkage Method:

For the ward's criterion linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.92**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.95**, and Fowlkes and Mallows index (**FM_index**) value is **0.95**. For the African dataset, the tanglegram that includes the comparison of the dendrograms of the ward's criterion linkage method is shown below in **Figure 8.2.6**. While almost half of the objects are stable, moderate and low shifts are observed in the remaining objects. It is seen that object 33 (Mauritania) has a large shift in the dendrograms.

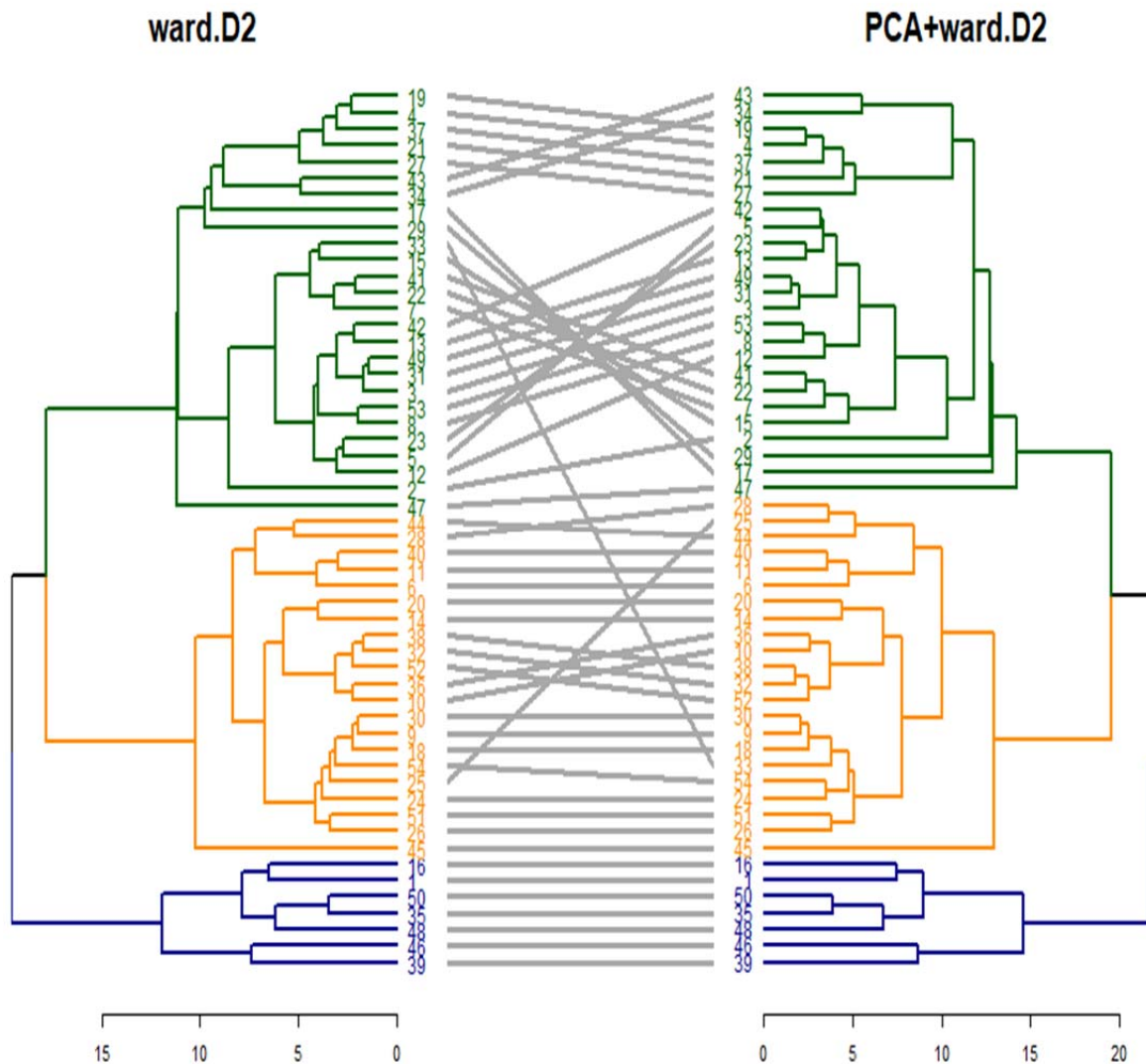


Figure 8.2. 6. Tanglegram for comparison dendrograms of ward.D2 linkage method and PCA+ward.D2 linkage method.

Table 8.2. 6. Clustering results for ward.D2 linkage and PCA+ward.D2 linkage method.

Ward.D2 linkage		
Cluster 1	Cluster 2	Cluster 3
Nigeria, South Africa, Sudan, Morocco, Tunisia, Algeria, Egypt	South Sudan, Angola, Congo, Burkina Faso, Ghana, Cameroon, Zambia, Benin, Malawi, Togo, Cote d'Ivoire, Senegal, Cabo Verde, Gambia, Sao Tome and Principe, Djibouti, Mauritania, Libya, Equatorial Guinea, Mauritius, Seychelles, Lesotho, Gabon, Namibia, Botswana, Eswatini	Burundi, Central African Republic, Chad, Comoros, Dem. Rep. Of the Congo, Eritrea, Ethiopia, Guinea, Guinea-Bissau, Kenya, Liberia, Madagascar, Mali, Mozambique, Niger, Rwanda, Sierra Leone, Somalia, Uganda, United Rep. Of Tanzania, Zimbabwe

PCA+Ward.D2 linkage		
Cluster 1	Cluster 2	Cluster 3
Nigeria, South Africa, Sudan, Morocco, Tunisia, Algeria, Egypt	South Sudan, Angola, Congo, Burkina Faso, Ghana, Cameroon, Zambia, Benin, Malawi, Togo, Cote d'Ivoire, Senegal, Cabo Verde, Gambia, Sao Tome and Principe, Djibouti, Libya, Equatorial Guinea, Mauritius, Seychelles, Lesotho, Gabon, Namibia, Botswana, Eswatini	Burundi, Central African Republic, Chad, Comoros, Dem. Rep. Of the Congo, Eritrea, Ethiopia, Guinea, Guinea-Bissau, Kenya, Liberia, Madagascar, Mali, Mauritania, Mozambique, Niger, Rwanda, Sierra Leone, Somalia, Uganda, United Rep. Of Tanzania, Zimbabwe

For the first dendrogram (ward.D2) the CPCC value is **0,5129** and for the second dendrogram (PCA+ward.D2) the CPCC value is **0,4961**. The CPCC values of both dendrograms are quite low and the cluster contents are not the same. That is, the relationship and similarity between the distance matrix and the cophenetic matrices of the dendrograms are quite low. As can be seen, the clustering results and CPCC values of the ward method are quite different from other methods. The main reason for explaining this is to aim to obtain homogeneous clusters in the ward method, and the cluster size is effective in this method, and the dataset used has a lot of outliers with high variance.

8.3. Applying Principal Component Analysis to the Europe Dataset:

The European dataset consists of **38** variables and **39** objects. The colored correlation matrix for the variables of the European dataset is given below.

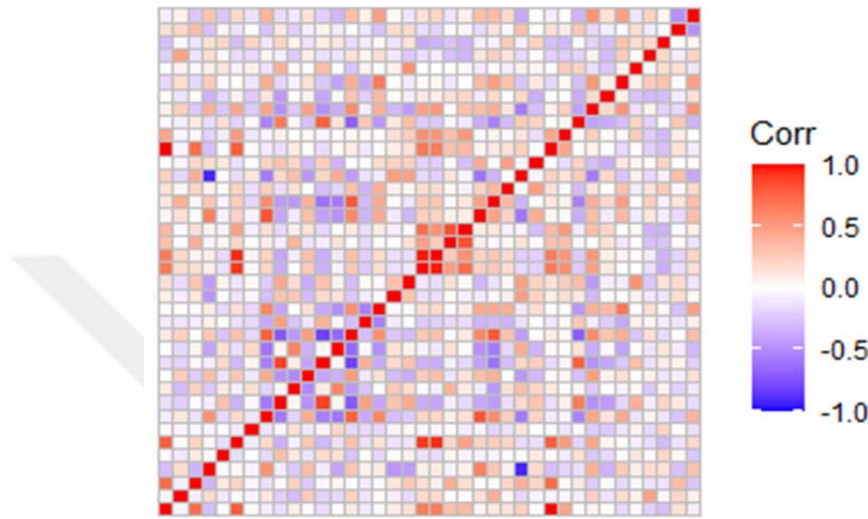


Figure 8. 3. Colored correlation matrix of the variables for dataset 2.

The mean absolute correlation between variables for the European dataset are calculated as **0.2346**. Due to the low level of correlation between the variables in the data set, it is seen that light colors are dominant in the colored correlation matrix in **figure 8.3**. The variance values of the principal components for the European dataset are given below with the scree plot and the table.

Variance Values of the Principle Components					
comp.1	comp.2	comp.3	comp.4	comp.5	comp.6
8,54E+00	5,58E+00	3,05E+00	2,82E+00	2,24E+00	1,96E+00
comp.7	comp.8	comp.9	comp.10	comp.11	comp.12
1,66E+00	1,59E+00	1,23E+00	1,11E+00	9,27E-01	8,87E-01

Table 8.3. 1. The variance values of the principal components for dataset 2.

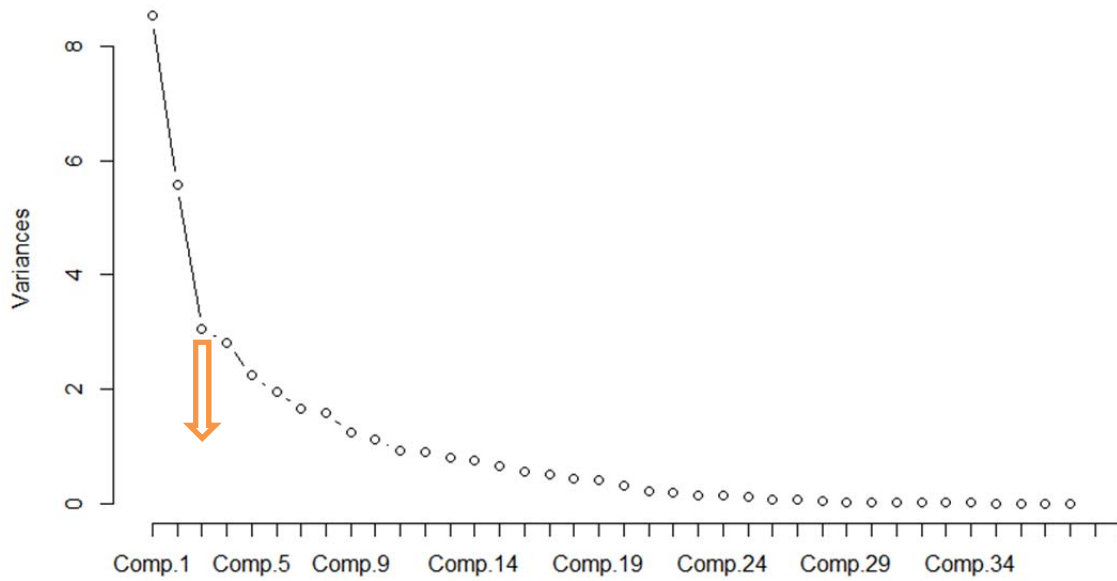


Figure 8.3 1. Scree plot of the variance values of the principal components for dataset 2.

For the elbow test, when the scree plot which includes the variance values of the principal components is given above is examined, it is seen that the **natural breakpoint** in the graph coincides with the 3rd principal component. However, in **table 8.3.2.**, when the variance explained percentages of the principal components are examined, it is seen that the **first 3 components** could explain only **%46.44** of the total variance in the dataset. This ratio is not sufficient to represent the entire data set.

For the second approach, the mean-variance value ($\bar{\lambda}$) of the principal components is calculated as **0.9743** for the European data set, and it is observed that the variance values of the first 10 principal components are above this value. When evaluated according to the variance explained percentages of the components, it is observed that the **first 10 components** can explain **%80.50** of the total variance in the data set.

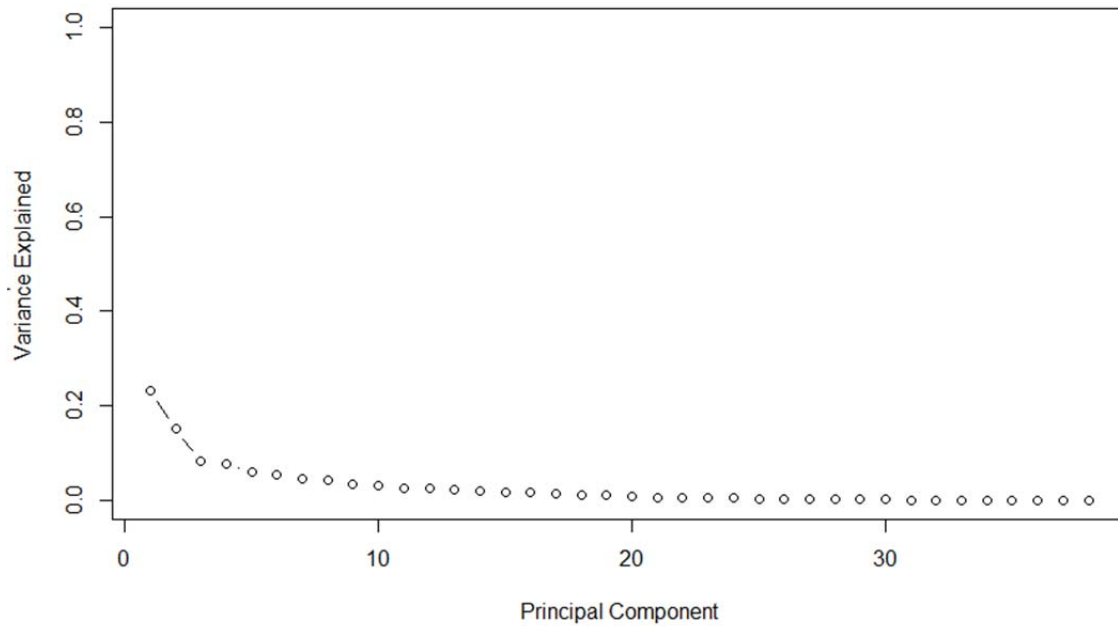


Figure 8.3.2. Scree plot of the variance explained rates of the principal components for dataset 2.

Table 8.3. 2 . The variance explained percentages of the principal components for dataset 2.

Variance Exp. Percentages of the Principle Components					
comp.1	comp.2	comp.3	comp.4	comp.5	comp.6
2,31E+01	1,51E+01	8,24E+00	7,61E+00	6,04E+00	5,29E+00
comp.7	comp.8	comp.9	comp.10	comp.11	comp.12
4,49E+00	4,31E+00	3,32E+00	3,00E+00	2,50E+00	2,40E+00

For the **80%** specific ratio, in figure 8.3.2 and table 8.3.2, the scree plot and the table containing the variance explained percentages of the principal components for the European dataset are given. When the values in **table 8.3.2.** are examined, it is seen that the first 10 principal components should be selected to achieve this specific ratio. It is observed that the **first 10 components** can explain **%80.50** of the total variance in the data set. If the results of these three approaches applied in deciding on the number of components to be selected are evaluated, it is clear that the average variance and 80% specific ratio approaches have the same results. However, it has been observed that the elbow test does not give reliable and robust results, and it gives different results than the other two approaches. As a result of all these evaluations, the number of components to be selected for a robust analysis process is determined as **10**.

8.4. Clustering Results of Europe Countries to the Economic Indicators and Social Indicators:

The comparisons of clustering results with PCA and without PCA for each method are given below, with tanglegrams and coefficients for the European dataset. The integer labels in the tanglegrams represent European countries in alphabetical order. Additionally, hierarchical clustering results, in which European countries are grouped according to the economic indicator and the social indicator variables, are also included in the appendices as dendrograms.

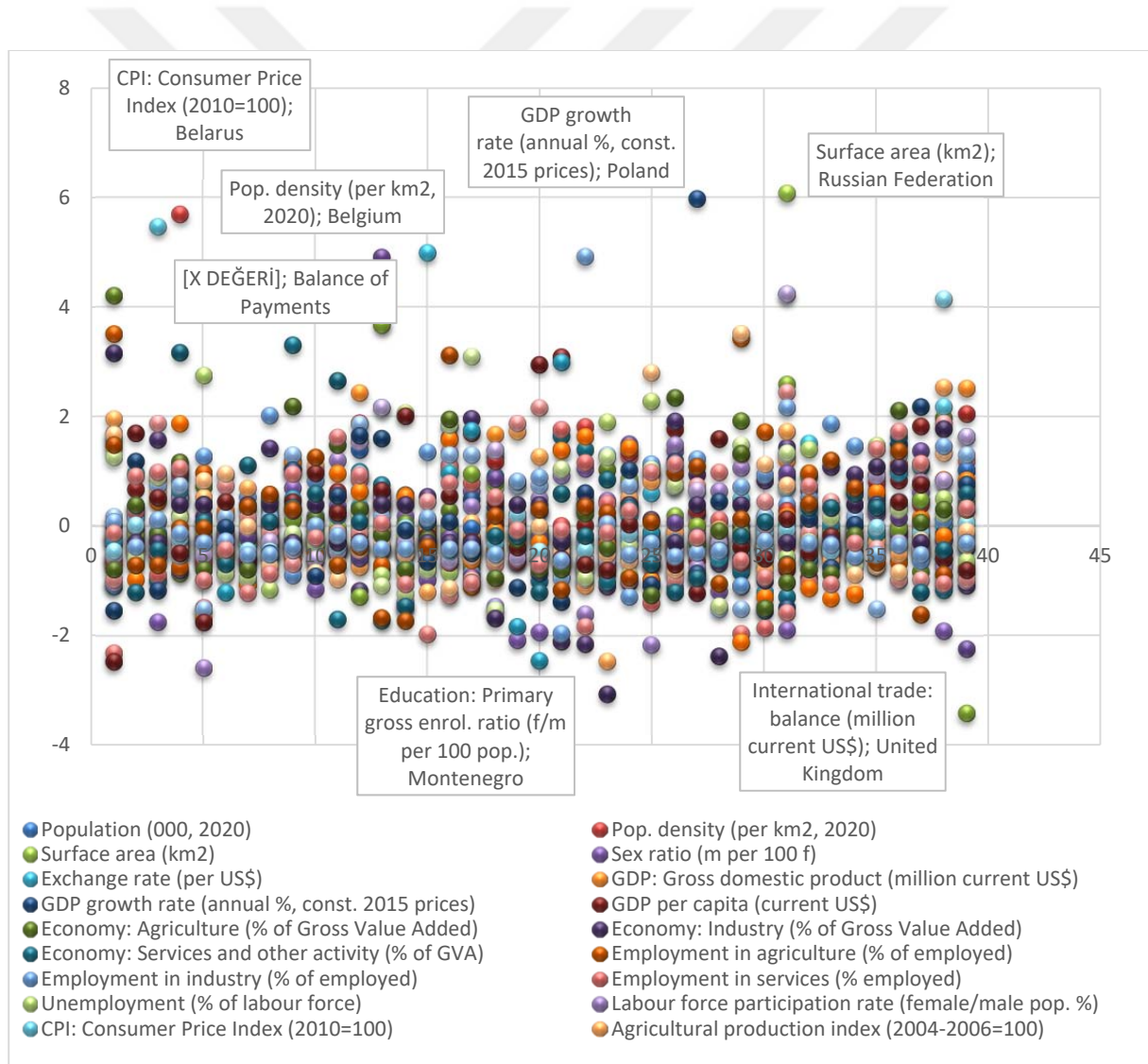


Figure 8. 4. Scatterplot of the European dataset for the economic indicator and social indicator variables.

Above is a scatterplot showing how European countries are distributed according to the economic indicator and the social indicator variables. It is also seems to be so many outliers here as in the first dataset. With the boxplots plotted for each variable in the dataset, the existence of many outliers is clearly seen in Figure 8.5. Several situations cause outliers to exist: measurement errors, data entry errors, or the actual value of the data. Since the observation values in the data set are real values, it is not possible and accurate to clear outliers. However, it is a fact that this situation will negatively affect the analysis process and results. This is a natural situation that can be encountered in high-dimensional real datasets.



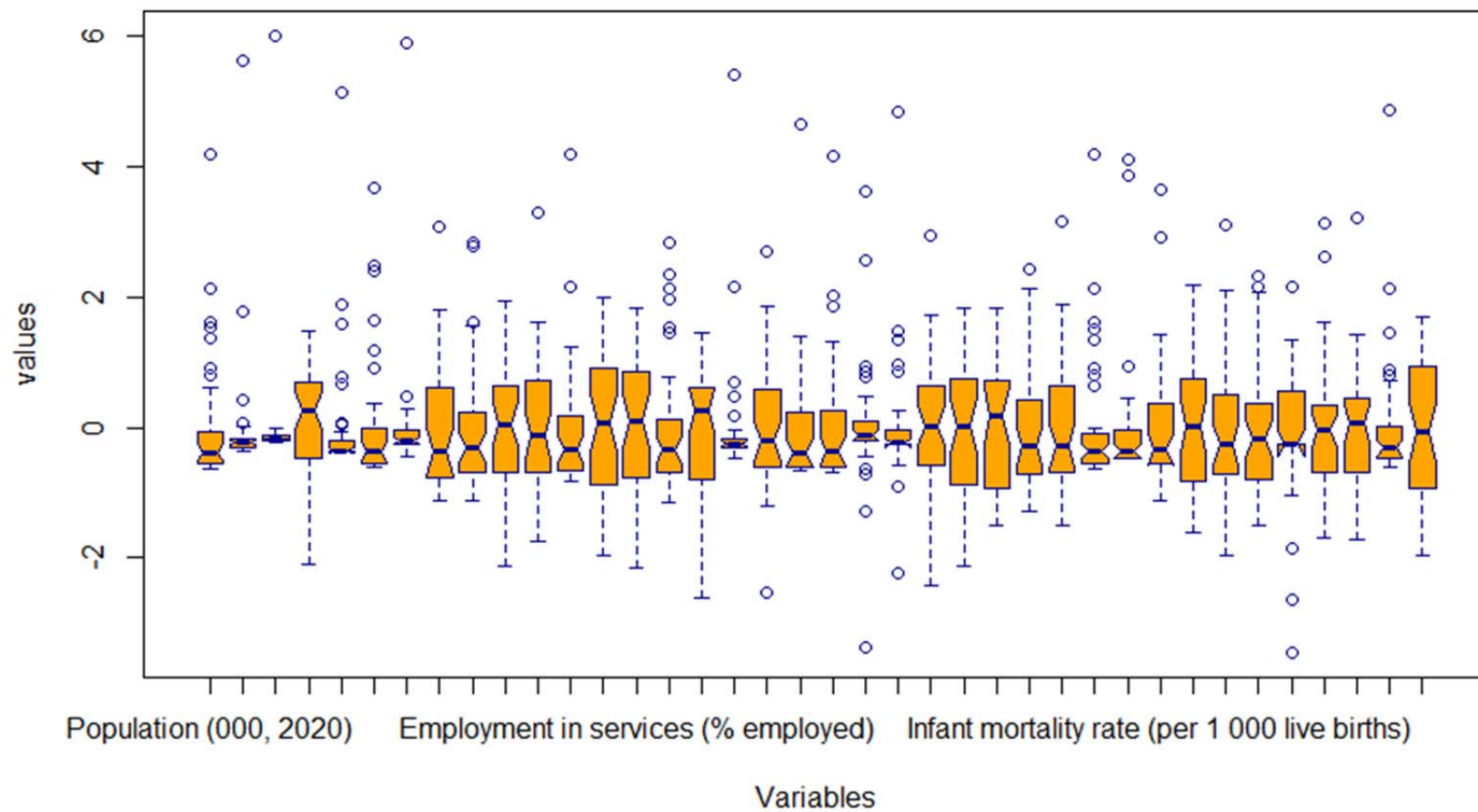


Figure 8.5. Boxplots for Europe dataset.

8.4.1. Comparative Results for Single Linkage Method:

For the single linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.99**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.98**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Europe dataset, the tanglegram that includes the comparison of the dendrograms of the single linkage method is shown below in **Figure 8.4.1**. Minor shifts are observed in almost half of the objects. Here, 13 (Germany) and 31 (Russian Federation) are considered as separate clusters because Germany has higher values (GDP, Imports, Exports) compared to other countries, and Russia has an extremely high value for the surface area.



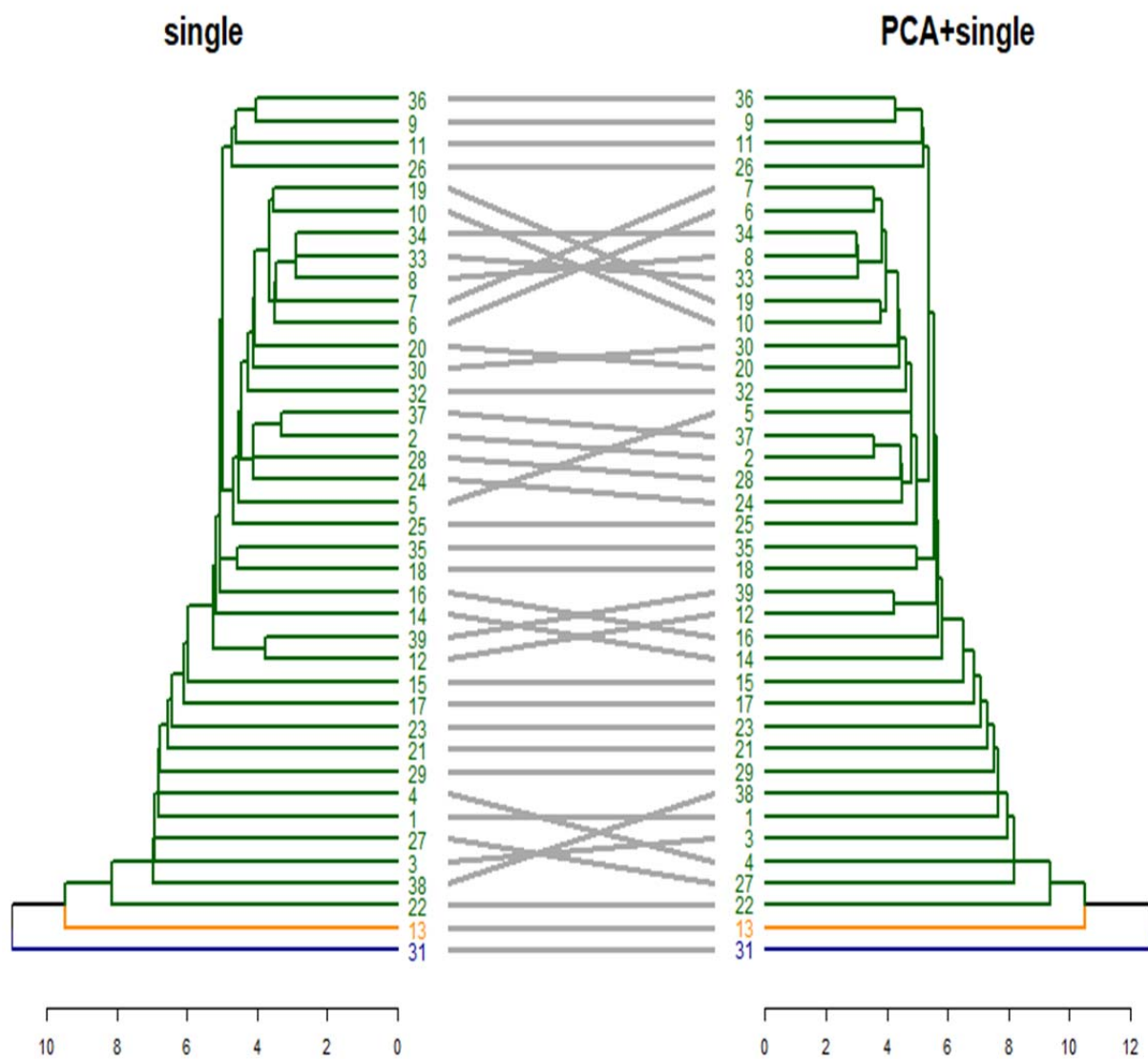


Figure 8.4. 1. Tanglegram for comparison dendrograms of single linkage method and PCA+single linkage method.

Table 8.4. 1. Clustering results for single linkage and PCA+single linkage method.

Single linkage		
Cluster 1	Cluster 2	Cluster 3
Albania, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Denmark, Estonia, Finland, France, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, North Macedonia, Norway, Poland, Portugal, Republic of Moldova, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Kingdom	Germany	Russian Federation

PCA+Single linkage		
Cluster 1	Cluster 2	Cluster 3
Albania, Austria, Belarus, Belgium, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Denmark, Estonia, Finland, France, Greece, Hungary, Iceland, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Montenegro, Netherlands, North Macedonia, Norway, Poland, Portugal, Republic of Moldova, Romania, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Kingdom	Germany	Russian Federation

For the first dendrogram (single) the CPCC value is **0,8013** and for the second dendrogram (PCA+single) the CPCC value is **0,8092**. The CPCC values are not very high, this means that the distance matrix and cophenetic matrices are not very similar. It is seen above that the cluster contents of both dendrograms are the same. Even, the results of all methods for this data set are the same except for the Ward's criterion method. When the clustering results and scatterplot are evaluated together, we see that Russia and Germany form separated clusters. Compared to other countries Germany has high values for the balance of payments, international trade imports and exports, international trade balance, refugees and others of concern and international migrant stock variables. Similarly, Russia Federation has high values compared to other countries for the surface area, international migrant stock, international trade balance, life expectancy at birth and intentional homicide rate variables.

8.4.2. Comparative Results for Complete Linkage Method:

For the complete linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.84**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.84**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Europe dataset, the tanglegram that includes the comparison of the dendrograms of the complete linkage method is shown below in **Figure 8.4.2**. Moderate shifts are observed in all objects except 13 and 31 (Germany, Russian Federation), while large shifts are observed in some objects. For the first dendrogram (single) the CPCC value is **0,7480** and for the second dendrogram (PCA+single) the CPCC value is **0,7351**. The CPCC values are not high, this meaning that the distance matrix and cophenetic matrices are not very similar.

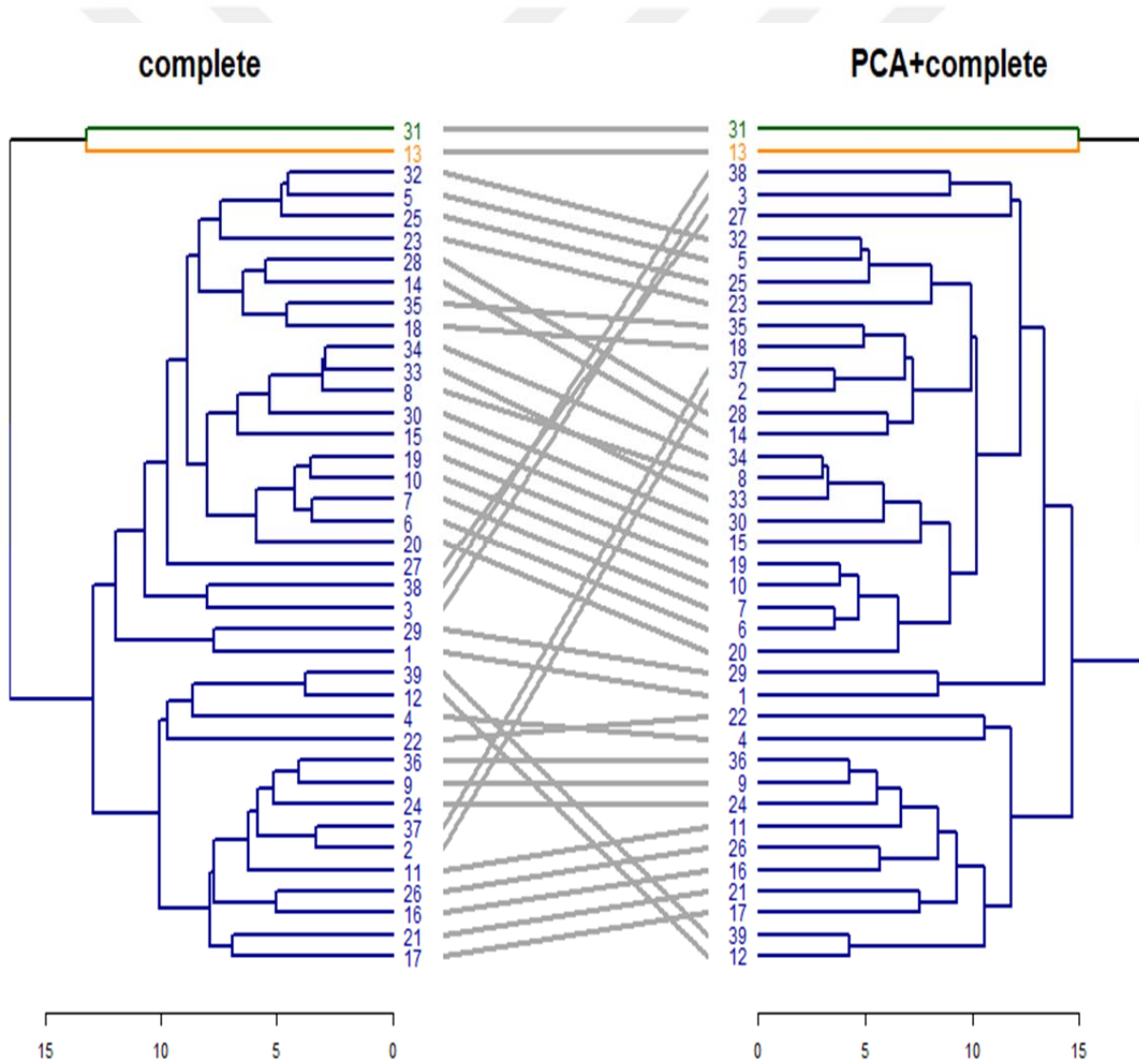


Figure 8.4. 2. Tanglegram for comparison dendrograms of complete linkage method and PCA+complete linkage method.

8.4.3. Comparative Results for Average Linkage Method:

For the average linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.50**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.88**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Europe dataset, the tanglegram that includes the comparison of the dendrograms of the average linkage method is shown below in **Figure 8.4.3**. Large shifts are observed in all objects except 13 and 31 (Germany, Russian Federation) objects. It is seen that the dendrogram created after PCA has a more neat hierarchical structure. For the first dendrogram (average) the CPCC value is **0,8588** and for the second dendrogram (PCA+average) the CPCC value is **0,8662**. The CPCC values are not very high, this meaning that the distance matrix and cophenetic matrices are not very similar.

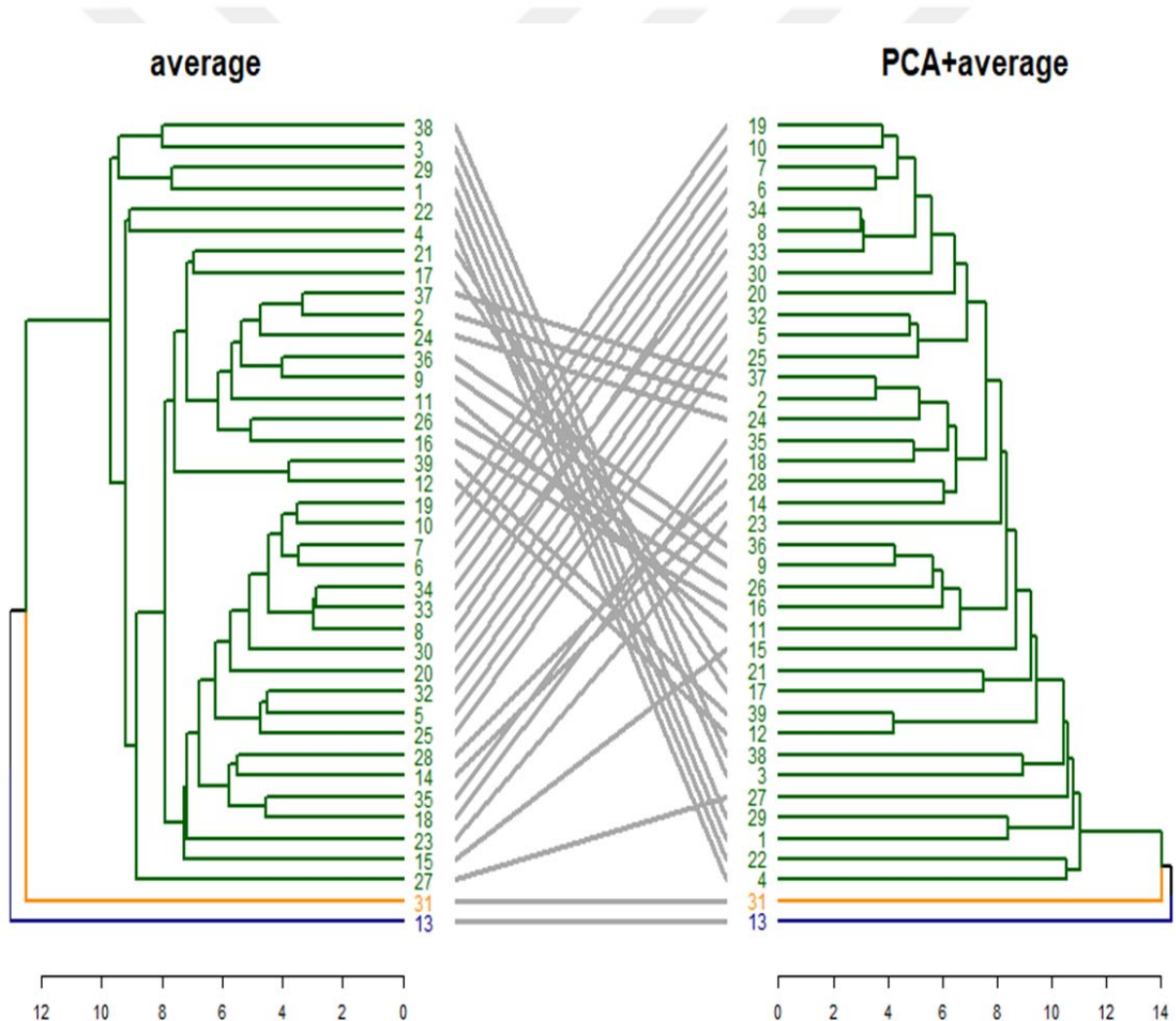


Figure 8.4. 3. Tanglegram for comparison dendrograms of average linkage method and PCA+average linkage method.

8.4.4. Comparative Results for Centroid Linkage Method:

For the centroid linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.99**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.33**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Europe dataset, the tanglegram that includes the comparison of the dendrograms of the centroid linkage method is shown below in **Figure 8.4.4**. Small shifts are observed in a few of the objects. However, as in the 1st data set, it is noteworthy that the BGCC value for the centroid method is low in this data set. For the first dendrogram (centroid) the CPCC value is **0,8370** and for the second dendrogram (PCA+centroid) the CPCC value is **0,8491**. The CPCC values are not very high, this meaning that the distance matrix and cophenetic matrices are not very similar.

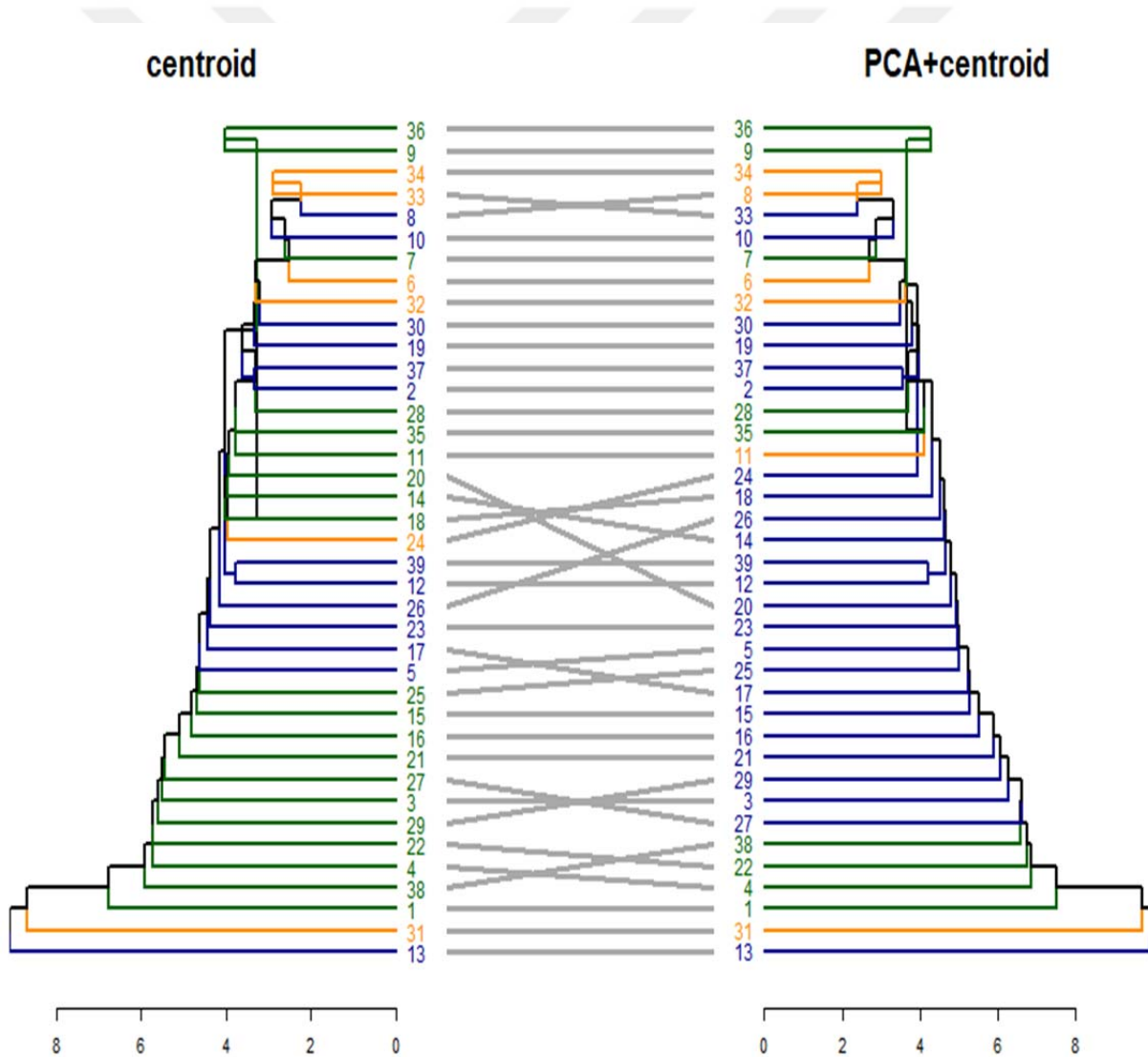


Figure 8.4. 4. Tanglegram for comparison dendrograms of centroid linkage method and PCA+centroid linkage method.

8.4.5. Comparative Results for Median Linkage Method:

For the median linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.90**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.42**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Europe dataset, the tanglegram that includes the comparison of the dendrograms of the median linkage method is shown below in **Figure 8.4.5**. Moderate and large shifts are observed in most of the objects. As in the first dataset, the BGCC value for the median method is low in this dataset. For the first dendrogram (median) the CPCC value is **0,7814** and for the second dendrogram (PCA+median) the CPCC value is **0,8022**. The CPCC values are not high, this meaning that the distance matrix and cophenetic matrices are not very similar.

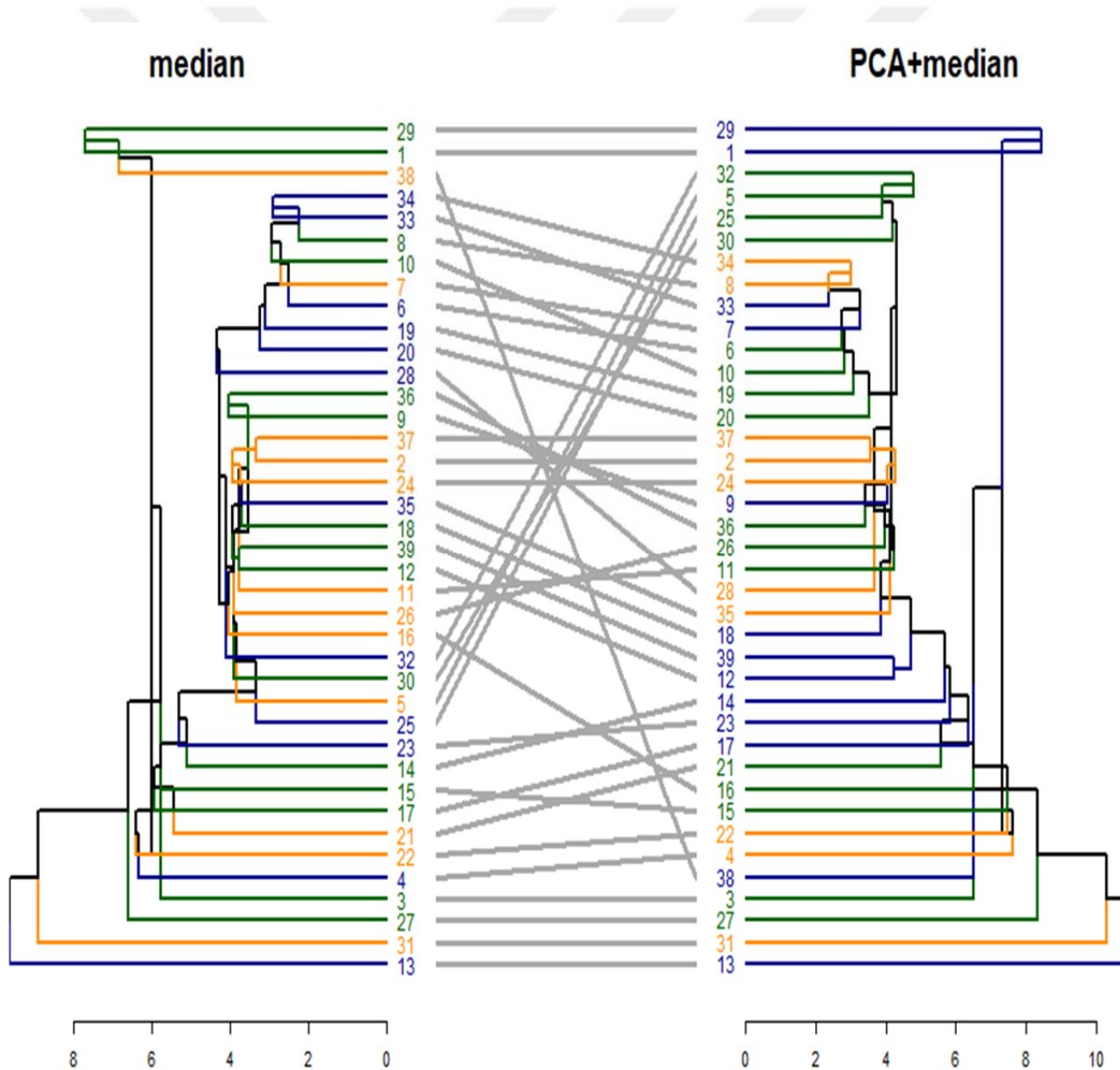


Figure 8.4. 5. Tanglegram for comparison dendrograms of median linkage method and PCA+median linkage method.

8.4.6. Comparative Results for Ward's Criterion Linkage Method:

For the ward's criterion linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.99**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.99**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Europe dataset, the tanglegram that includes the comparison of the dendrograms of the ward's criterion linkage method is shown below in **Figure 8.4.6**. Both dendrograms are almost identical, except for minor shifts that are insignificant for a few objects. Here objects 31 and 13 (Russia, Germany) are both in the same cluster, because ward method aims to create homogenous clusters.

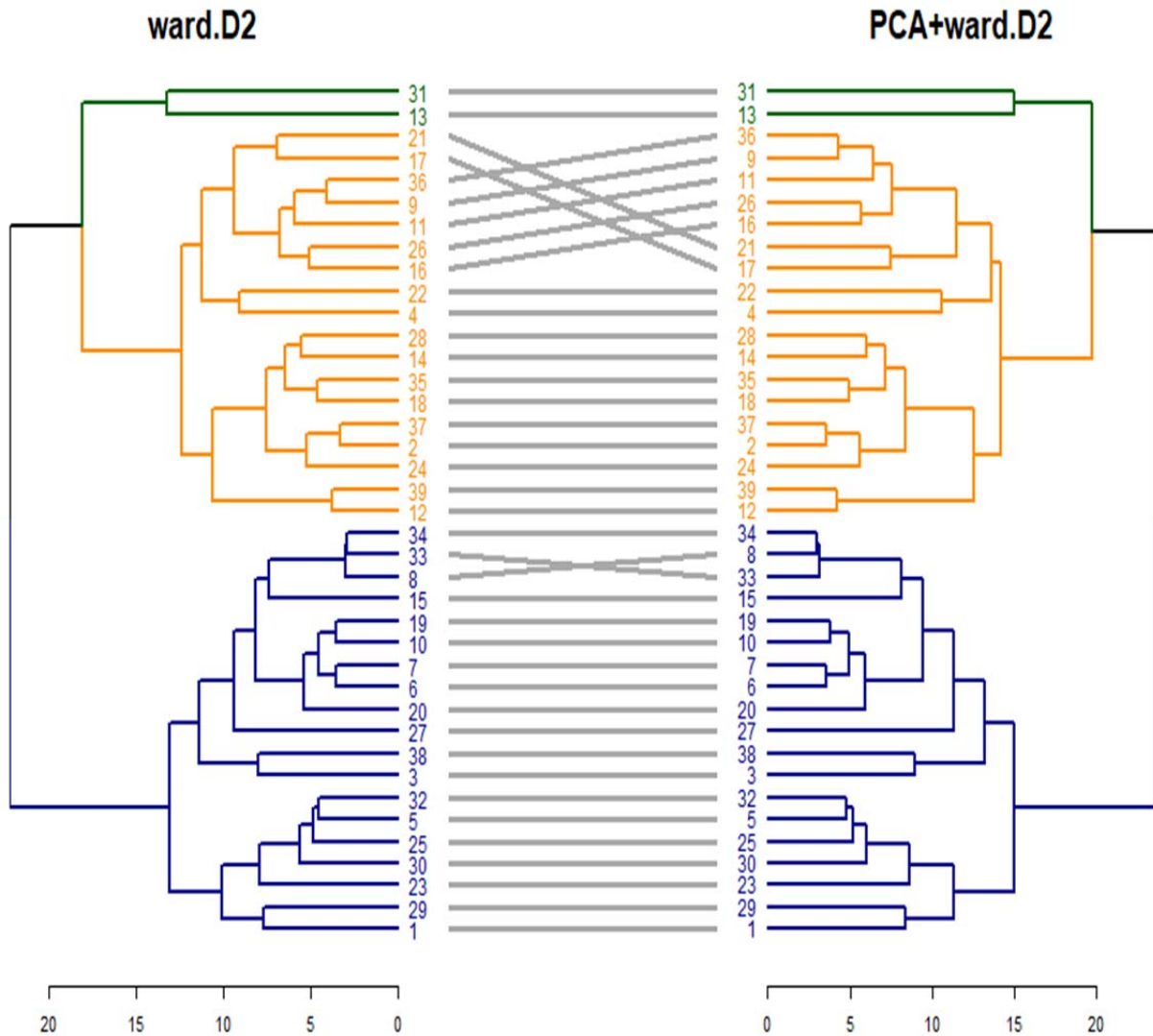


Figure 8.4. 6. Tanglegram for comparison dendrograms of ward.D2 linkage method and PCA+ward.D2 linkage method.

Table 8.4. 2. Clustering results for ward.D2 linkage and PCA+ward.D2 linkage method.

Ward.D2 linkage		
Cluster 1	Cluster 2	Cluster 3
Albania, Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Estonia, Hungary, Latvia, Lithuania, Montenegro, North Macedonia, Poland, Republic of Moldova, Romania, Serbia, Slovakia, Slovenia, Ukraine	France, United Kingdom, Netherlands, Austria, Switzerland, Italy, Spain, Greece, Portugal, Belgium, Malta, Iceland, Norway, Finland, Denmark, Sweden, Ireland, Luxembourg	Germany, Russian Federation

PCA+Ward.D2 linkage		
Cluster 1	Cluster 2	Cluster 3
Albania, Belarus, Bosnia and Herzegovina, Bulgaria, Croatia, Czechia, Estonia, Hungary, Latvia, Lithuania, Montenegro, North Macedonia, Poland, Republic of Moldova, Romania, Serbia, Slovakia, Slovenia, Ukraine	France, United Kingdom, Netherlands, Austria, Switzerland, Italy, Spain, Greece, Portugal, Belgium, Malta, Iceland, Norway, Finland, Denmark, Sweden, Ireland, Luxembourg	Germany, Russian Federation

For the first dendrogram (ward.D2) the CPCC value is **0,4782** and for the second dendrogram (PCA+ward.D2) the CPCC value is **0,4646**. The CPCC values of both dendrograms are quite low but the cluster contents are the same. That is, the relationship and similarity between the distance matrix and the cophenetic matrices of the dendrograms are quite low. As can be seen, the clustering results and CPCC values of the ward method are quite different from other methods. The main reason explaining this is to aim to obtain homogeneous clusters by keeping within cluster variance minimum in the ward method and the cluster size is effective in this method and the dataset used has a lot of outliers with high variance.

8.5. Applying Principal Component Analysis to the Asian Dataset:

The Asian dataset consists of **46** variables and **48** objects. The colored correlation matrix for the variables of the Asian dataset is given below.

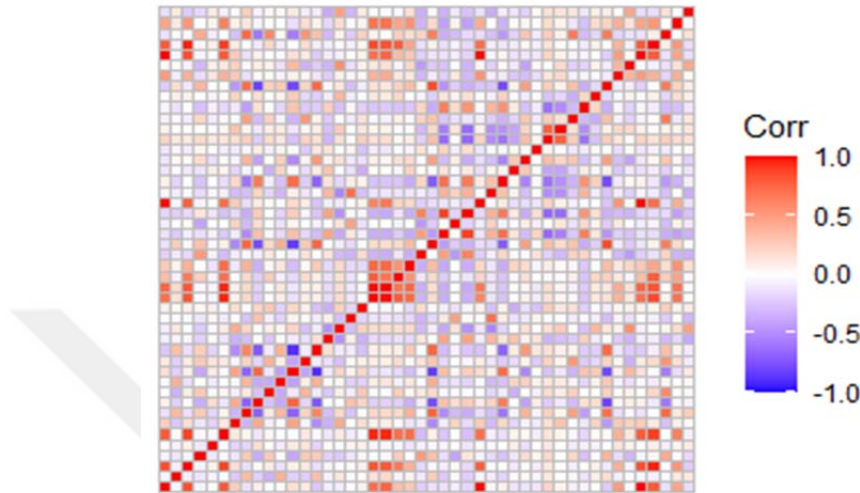


Figure 8. 5. Colored correlation matrix of the variables for dataset 3.

The mean absolute correlation between variables for the Asian dataset was calculated as **0.2265**. Due to the low level of correlation between the variables in the data set, it is seen that light colors are dominant in the colored correlation matrix in **figure 8.5**. The variance values of the principal components for the Asian dataset are given below with the scree plot and the table.

Table 8.5. 1. The variance values of the principal components for dataset 3.

Variance Values of the Principle Components					
comp.1	comp.2	comp.3	comp.4	comp.5	comp.6
9,53E+00	7,86E+00	4,51E+00	3,45E+00	2,68E+00	2,02E+00
comp.7	comp.8	comp.9	comp.10	comp.11	comp.12
1,72E+00	1,43E+00	1,35E+00	1,25E+00	1,15E+00	1,03E+00

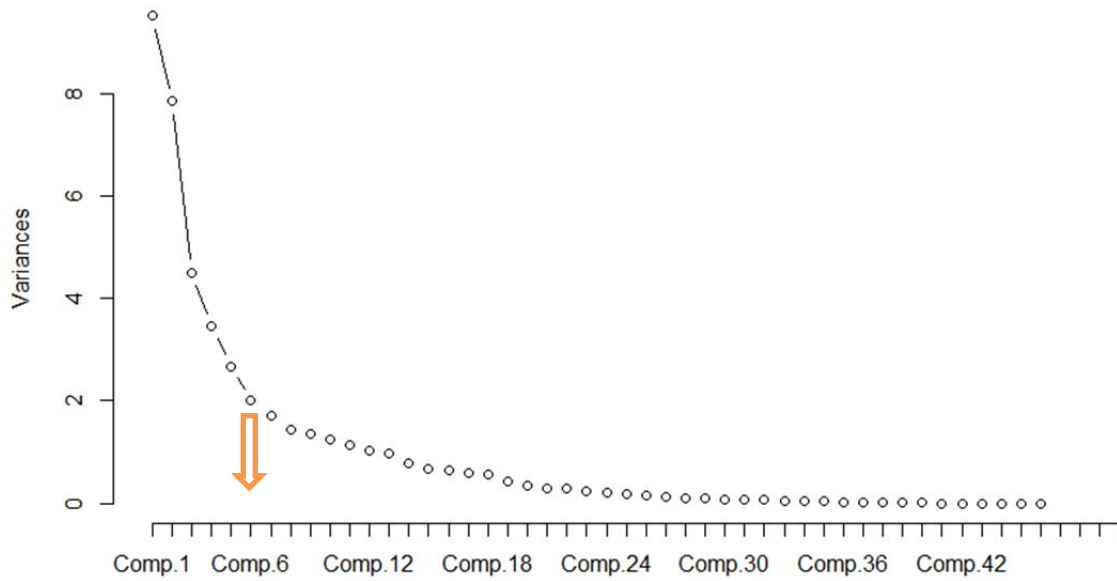


Figure 8.5. 1. Scree plot of the variance values of the principal components for dataset 3.

For the elbow test, when the scree plot which includes the variance values of the principal components given above is examined, it is seen that the **natural breakpoint** in the graph coincides with the 6th principal component. However, in **table 8.5.2**, when the variance explained percentages of the principal components are examined, it is seen that the **first 6 components** can explain only **%66.68** of the total variance in the dataset. This ratio is not sufficient to represent the entire data set.

For the second approach, the mean-variance value ($\bar{\lambda}$) of the principal components is calculated as **0.9791** for the Asian data set, and it is observed that the variance values of the first 12 principal components are above this value. When evaluated according to the variance explained percentages of the components, it is observed that the **first 12 components** can explain **%84.29** of the total variance in the data set.

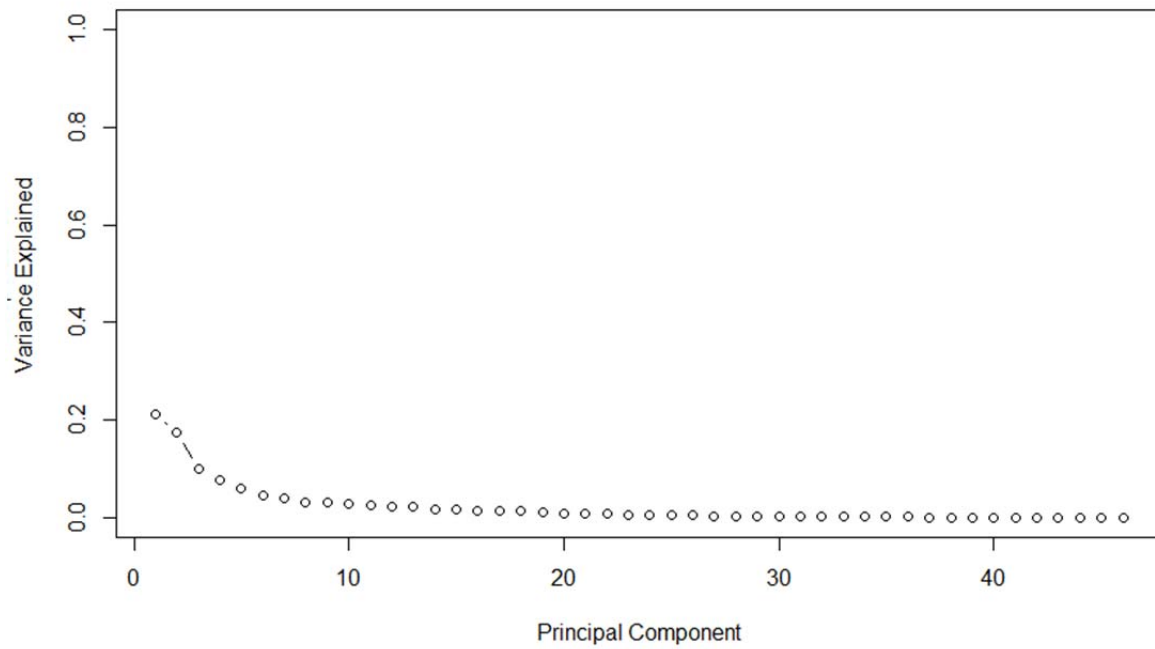


Figure 8.5. 2. Scree plot of the variance explained rates of the principal components for dataset 3.

Table 8.5. 2. The variance explained percentages of the principal components for dataset 1.

Variance Exp. Percentages of the Principle Components					
comp.1	comp.2	comp.3	comp.4	comp.5	comp.6
2,12E+01	1,74E+01	1,00E+01	7,66E+00	5,94E+00	4,48E+00
comp.7	comp.8	comp.9	comp.10	comp.11	comp.12
3,82E+00	3,18E+00	3,00E+00	2,78E+00	2,54E+00	2,29E+00

For the **80%** specific ratio, the scree plot and the table containing the variance explained percentages of the principal components for the Asian dataset are given. When the values in the **table 8.5.2** are examined, it is seen that the first 11 principal components should be selected to achieve this specific ratio. It is observed that the **first 11 components** can explain **%82.00** of the total variance in the data set. If the results of these three approaches applied in deciding on the number of components to be selected are evaluated, it is clear that the average variance and 80% specific ratio approaches give very close results. However, it has been observed that the elbow test does not give reliable and robust results, and it gives different results than the other two approaches. As a result of all these evaluations, the number of components to be selected for a robust analysis process is determined as **11**.

8.6. Clustering Results of Asian Countries to the Economic Indicators, Social Indicators and Environment and Infrastructure Indicators:

The comparisons of clustering results with PCA and without PCA for each method are given below with tanglegrams and coefficients for the Asian dataset. The integer labels in the tanglegrams represent Asian countries in alphabetical order. Additionally hierarchical clustering results, in which Asian countries are grouped according to the economic indicator, the social indicator, and the environment and infrastructure indicator variables, are also included in the appendices as dendrograms.

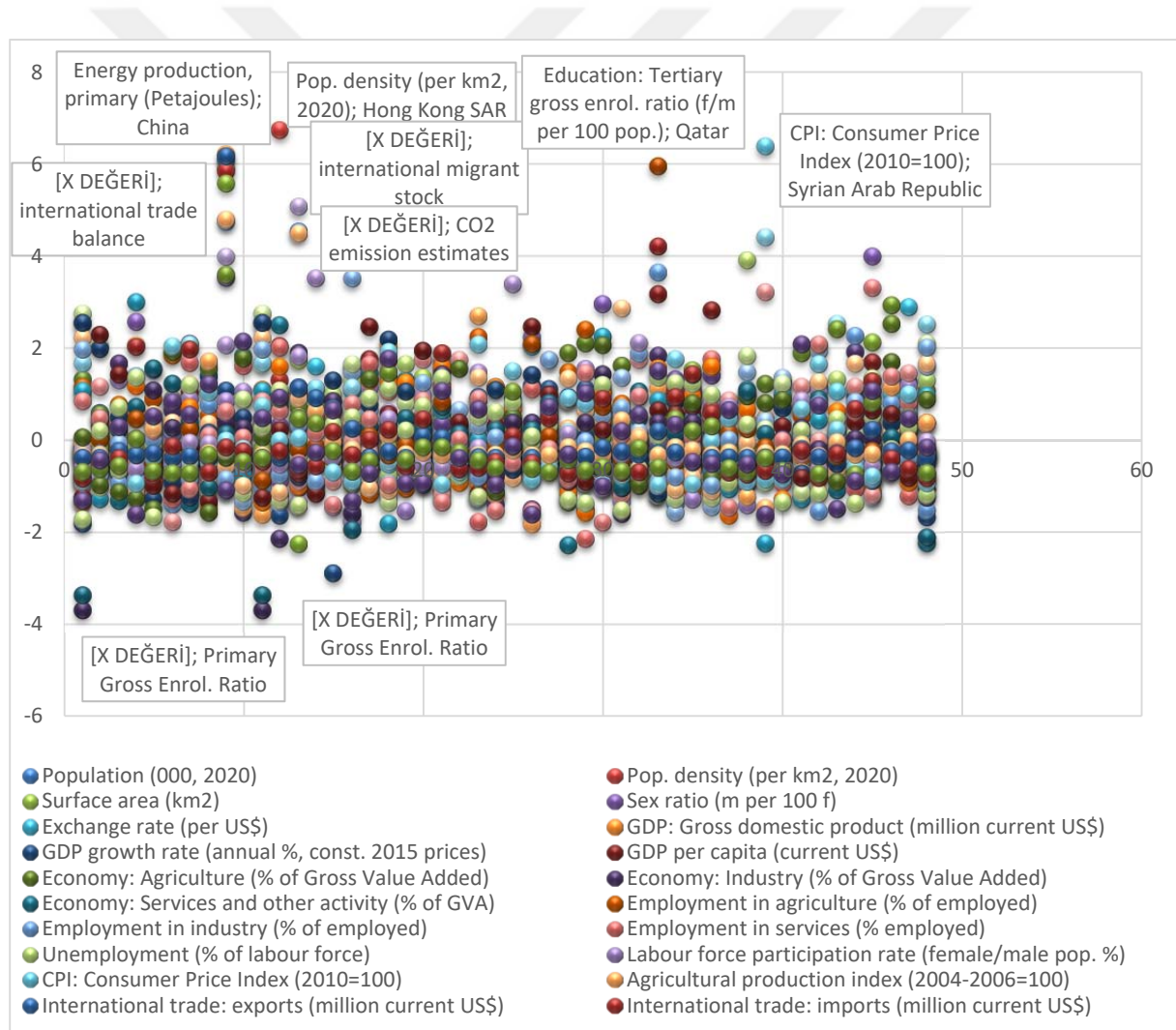


Figure 8. 6. Scatterplot of the Asia dataset for the economic indicator, social indicator and environmental and infrastructural variables.

Above is a scatterplot showing how Asian countries are distributed according to economic indicator, social indicator and, environment and infrastructure indicator variables. When the distribution of the normalized data in the graph is examined, it can be seen that there are so many outliers in the data set as the first and second datasets. Here China, for many variables, has extreme values compared to other countries. The boxplots plotted with normalized values show that there are many outlier values in almost all variables in a way that supports the scatterplot.



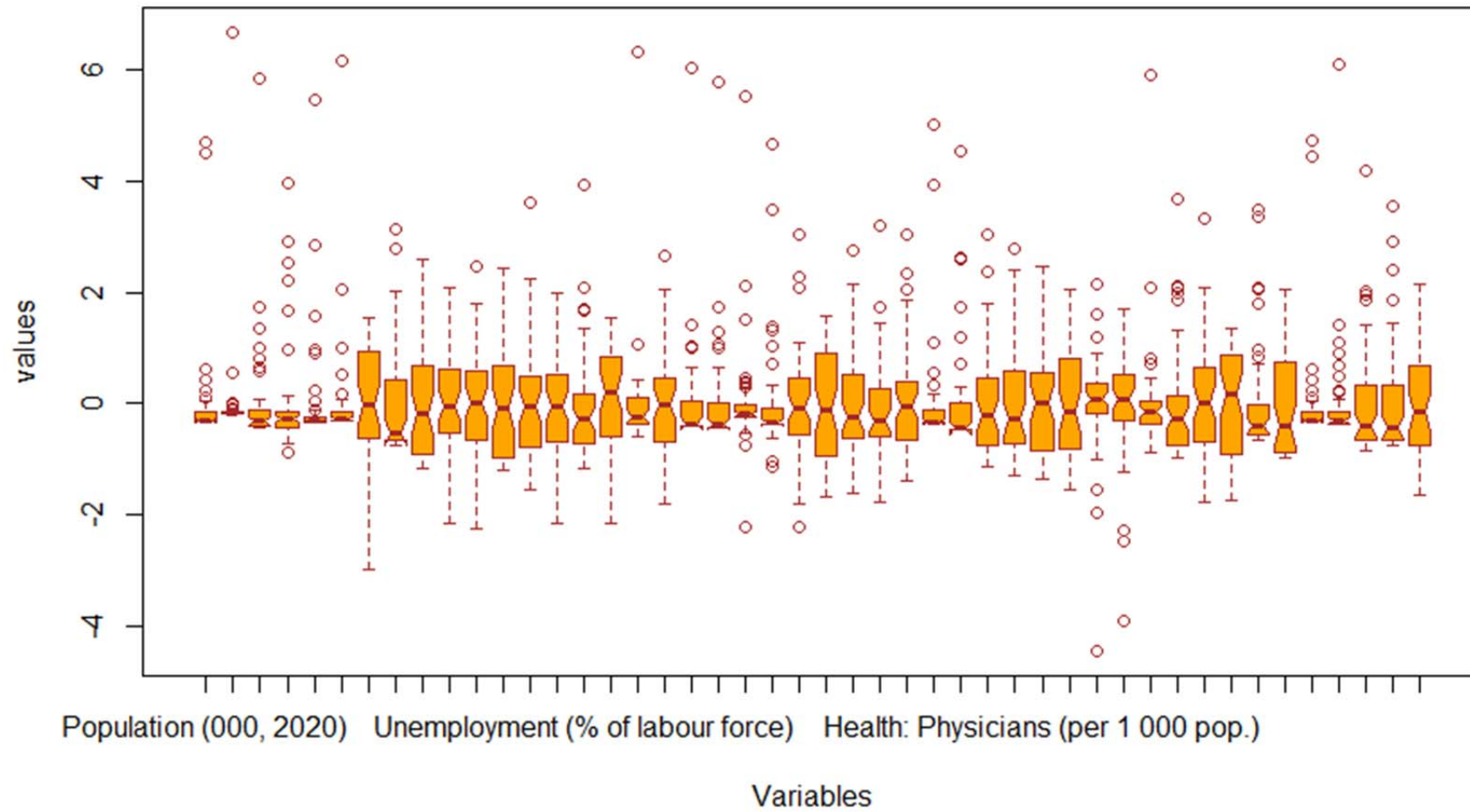


Figure 8.7. Boxplots for Asia dataset.

8.6.1. Comparative Results for Single Linkage Method:

For the single linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.99**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.99**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Asian dataset, the tanglegram that includes the comparison of the dendrograms of the single linkage method is shown below in **Figure 8.6.1**. The two dendrograms are almost identical, except for minor shifts in very few objects. Here, 9 (China) is treated as a separate cluster with a significant difference, as it has extreme high values for GDP and surface area variables. However, the fact that China's values such as population, international trade exports and imports are higher than other countries have also been effective in this situation.

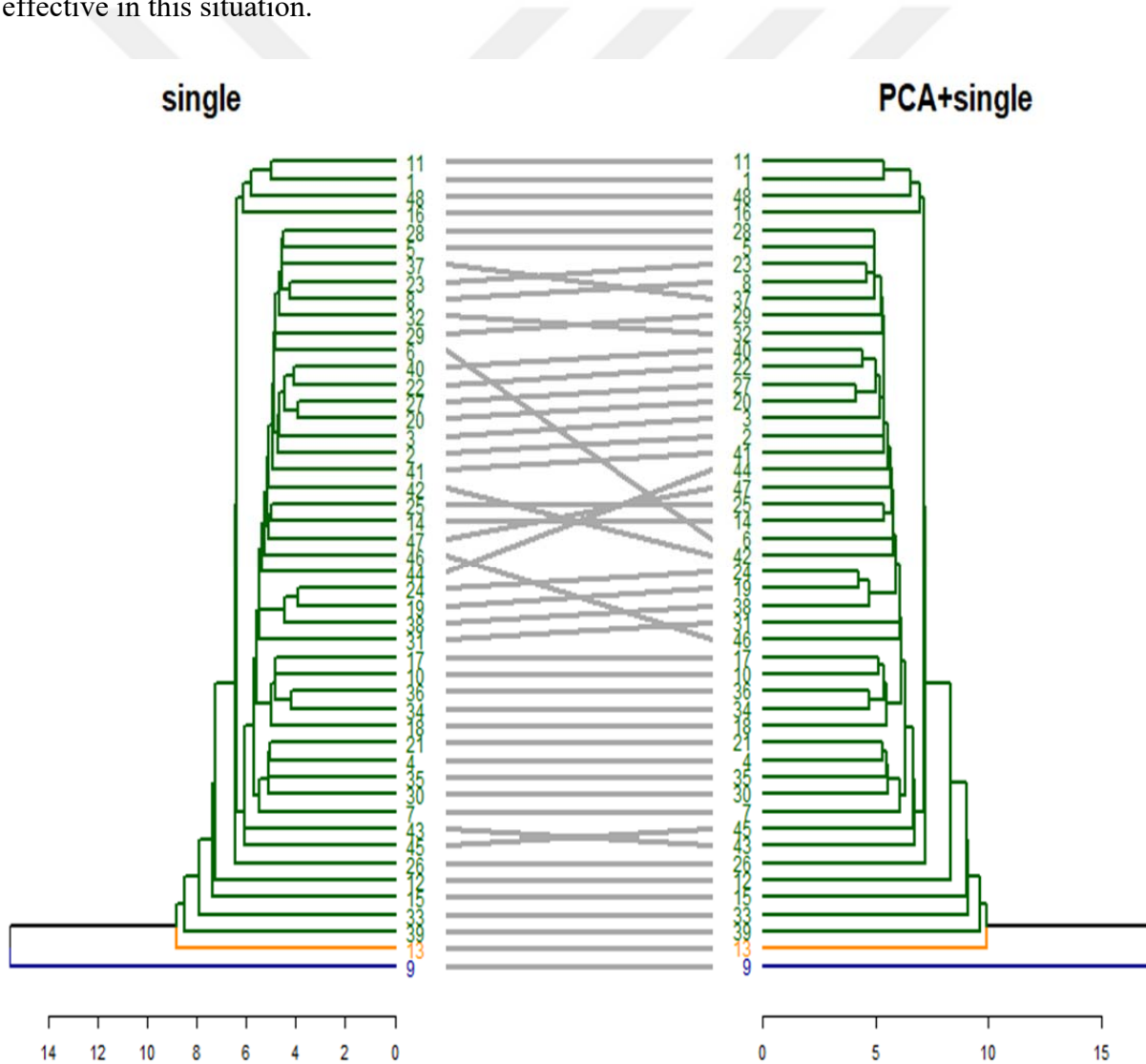


Figure 8.6. 1. Tanglegram for comparison dendrograms of single linkage method and PCA+single linkage method.

Table 8.6. 1. Clustering results for single linkage and PCA+single linkage method.

Single linkage		
Cluster 1	Cluster 2	Cluster 3
Afghanistan, Armenia, Azerbaijan, Bahrain, Bangladesh, Bhutan, Brunei Darussalam, Cambodia, Cyprus, Georgia, Hong Kong SAR, Indonesia, Iran, Iraq, Israel, Japan, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lao People's Dem. Rep., Lebanon, Malaysia, Maldives, Mongolia, Myanmar, Nepal, Oman, Pakistan, Philippines, Qatar, Republic of Korea, Saudi Arabia, Singapore, Sri Lanka, State of Palestine, Syrian Arab Republic, Tajikistan, Thailand, Timor-Leste, Turkey, Turkmenistan, United Arab Emirates, Uzbekistan, Viet Nam, Yemen	China	India

PCA+Single linkage		
Cluster 1	Cluster 2	Cluster 3
Afghanistan, Armenia, Azerbaijan, Bahrain, Bangladesh, Bhutan, Brunei Darussalam, Cambodia, Cyprus, Georgia, Hong Kong SAR, Indonesia, Iran, Iraq, Israel, Japan, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lao People's Dem. Rep., Lebanon, Malaysia, Maldives, Mongolia, Myanmar, Nepal, Oman, Pakistan, Philippines, Qatar, Republic of Korea, Saudi Arabia, Singapore, Sri Lanka, State of Palestine, Syrian Arab Republic, Tajikistan, Thailand, Timor-Leste, Turkey, Turkmenistan, United Arab Emirates, Uzbekistan, Viet Nam, Yemen	China	India

For the first dendrogram (single) the CPCC value is **0,8718** and for the second dendrogram (PCA+single) the CPCC value is **0,8681**. The cpcc values of both dendrograms are high and the cluster contents are the same. That is, the relationship and similarity between the distance matrix and the cophenetic matrices of the dendrograms are high. When the clustering results and scatterplot are evaluated together, we see that China and India form separated clusters. Compared to other countries China has high values for the energy production primary, energy supply, international trade balance, CO2 emission estimates, tourist/visitor arrivals and international migrant stock variables. Similarly, India has high values compared to other countries for the international migrant stock, CO2 emission estimates variables and has low value for the international trade balance variable.

For the complete linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.59**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.74**, and Fowlkes and Mallows index (**FM_index**) value is **0.72**. For the Asian dataset, the tanglegram that includes the comparison of the dendrograms of the complete linkage method is shown below in **Figure 8.6.2**. Large shifts are observed for almost all objects except a few objects. Dependently on this, it is seen that the coefficient values are also low.

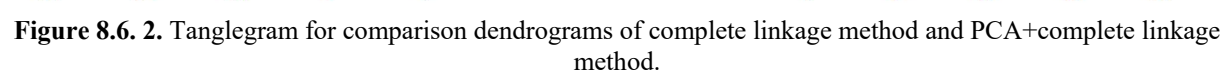


Table 8.6. 2. Clustering results for complete linkage and PCA+complete linkage method.

Complete linkage		
Cluster 1	Cluster 2	Cluster 3
Afghanistan, Georgia, India, Iraq, Syrian Arab Republic, Yemen	Armenia, Azerbaijan, Bahrain, Bangladesh, Bhutan, Brunei Darussalam, Cambodia, Cyprus, Hong Kong SAR, Indonesia, Iran, Israel, Japan, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lao People's Dem. Rep., Lebanon, Malaysia, Maldives, Mongolia, Myanmar, Nepal, Oman, Pakistan, Philippines, Qatar, Republic of Korea, Saudi Arabia, Singapore, Sri Lanka, State of Palestine, Tajikistan, Thailand, Timor-Leste, Turkey, Turkmenistan, United Arab Emirates, Uzbekistan, Viet Nam	China

PCA+Complete linkage		
Cluster 1	Cluster 2	Cluster 3
Afghanistan, Armenia, Azerbaijan, Bahrain, Bangladesh, Bhutan, Brunei Darussalam, Cambodia, Georgia, India, Indonesia, Iran, Iraq, Jordan, Kazakhstan, Kuwait, Kyrgyzstan, Lao People's Dem. Rep., Lebanon, Malaysia, Maldives, Mongolia, Myanmar, Nepal, Oman, Pakistan, Philippines, Saudi Arabia, Sri Lanka, State of Palestine, Syrian Arab Republic, Tajikistan, Thailand, Timor-Leste, Turkey, Turkmenistan, United Arab Emirates, Uzbekistan, Viet Nam, Yemen	China	Cyprus, Hong Kong SAR, Israel, Japan, Qatar, Republic of Korea, Singapore

For the first dendrogram (single) the CPCC value is **0,8458** and for the second dendrogram (PCA+single) the CPCC value is **0,7562**. The CPCC value of first dendrogram are higher than the second dendrogram and the cluster contents are not the same. That is, the relationship and similarity between the distance matrix and the cophenetic matrix of the first dendrogram is higher compared to the second dendrogram. Here compared to other countries Afghanistan and Georgia have low values for the primary and secondary gross enroll. Ratio and Georgia has also high values for pop. Age distribution and current expenditure in health; Iraq has high value for intentional homicide rate; Syria has high value for consumer price index and refugees and others of concern to UNHCR; Yemen has high value for refugees and others of concern and low value for secondary gross enrol. Ratio; Hong Kong has high values for pop. Density and services and other activity in economy; Israel has high value for physicians in health; Singapore has high value for GDP per capita; Qatar has high values for tertiary gross enroll ratio, energy supply, employment in industry and GDP per capita.

8.6.3. Comparative Results for Average Linkage Method:

For the average linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.90**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.97**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Asian dataset, the tanglegram that includes the comparison of the dendrograms of the average linkage method is shown below in **Figure 8.6.3**. Major shifts are observed in a few objects, while minor shifts are observed in general. For the first dendrogram (average) the CPCC value is **0,9129** and for the second dendrogram (PCA+average) the CPCC value is **0,9108**. The CPCC values of both dendrograms are quite high and the cluster contents are the same. That is, the relationship and similarity between the distance matrix and the cophenetic matrices of the dendrograms are very high. When the clustering results and scatterplot are evaluated together, we see that China and India form separated clusters as in the single linkage method.

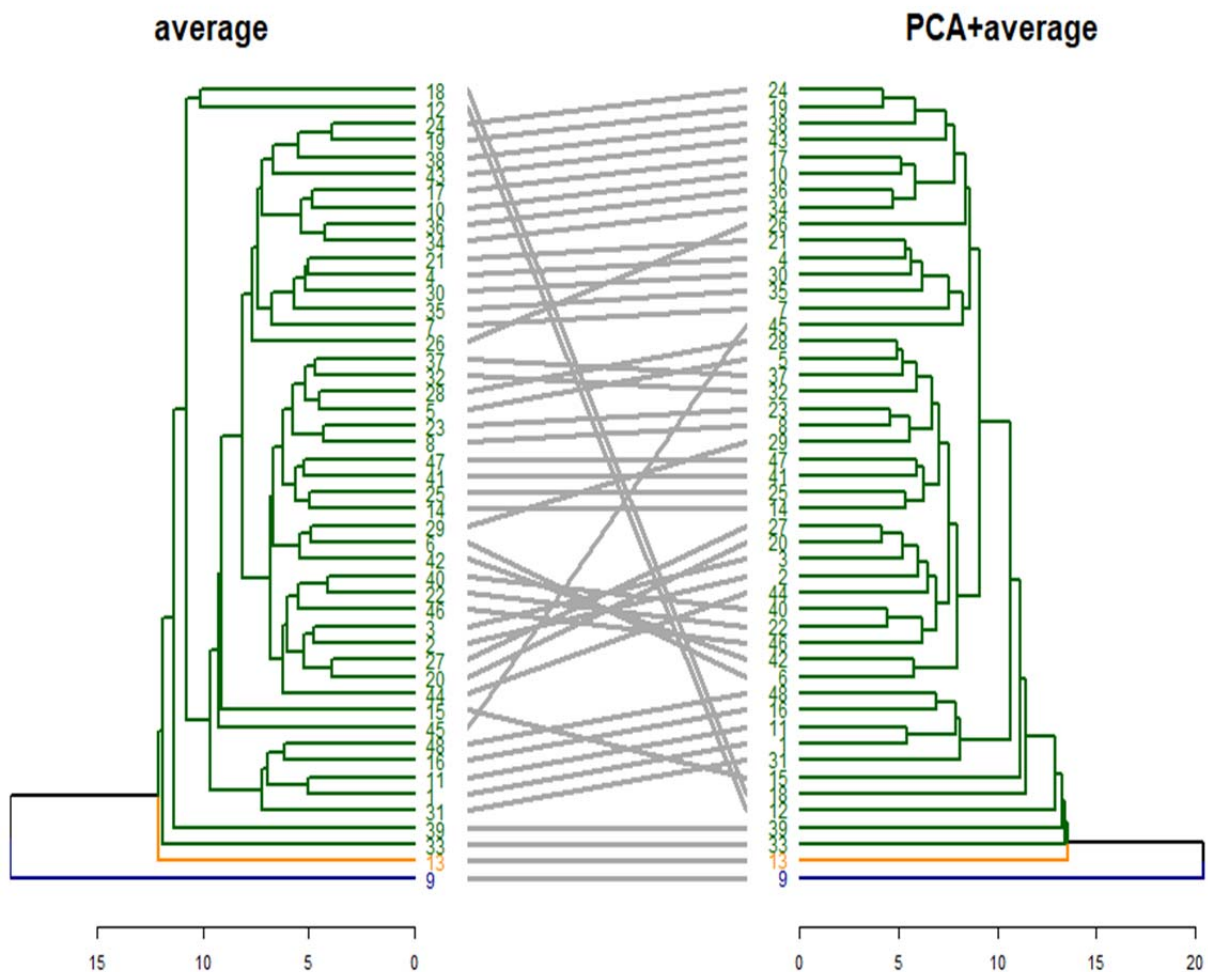


Figure 8.6. 3. Tanglegram for comparison dendrograms of average linkage method and PCA+average linkage method.

8.6.4. Comparative Results for Centroid Linkage Method:

For the centroid linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.99**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.59**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Asian dataset, the tanglegram that includes the comparison of the dendrograms of the centroid linkage method is shown below in **Figure 8.6.4**. While no displacement is observed in almost half of the objects, small shifts are observed in the remaining objects. Similar to the other two datasets, the BGCC value is calculated very low for the centroid method. For the first dendrogram (centroid) the CPCC value is **0,8714** and for the second dendrogram (PCA+centroid) the CPCC value is **0,8709**. The CPCC values of both dendrograms are quite high and the cluster contents are the same. When the clustering results and scatterplot are evaluated together, it is seen that China and India form separated clusters as in the single linkage and average linkage methods.

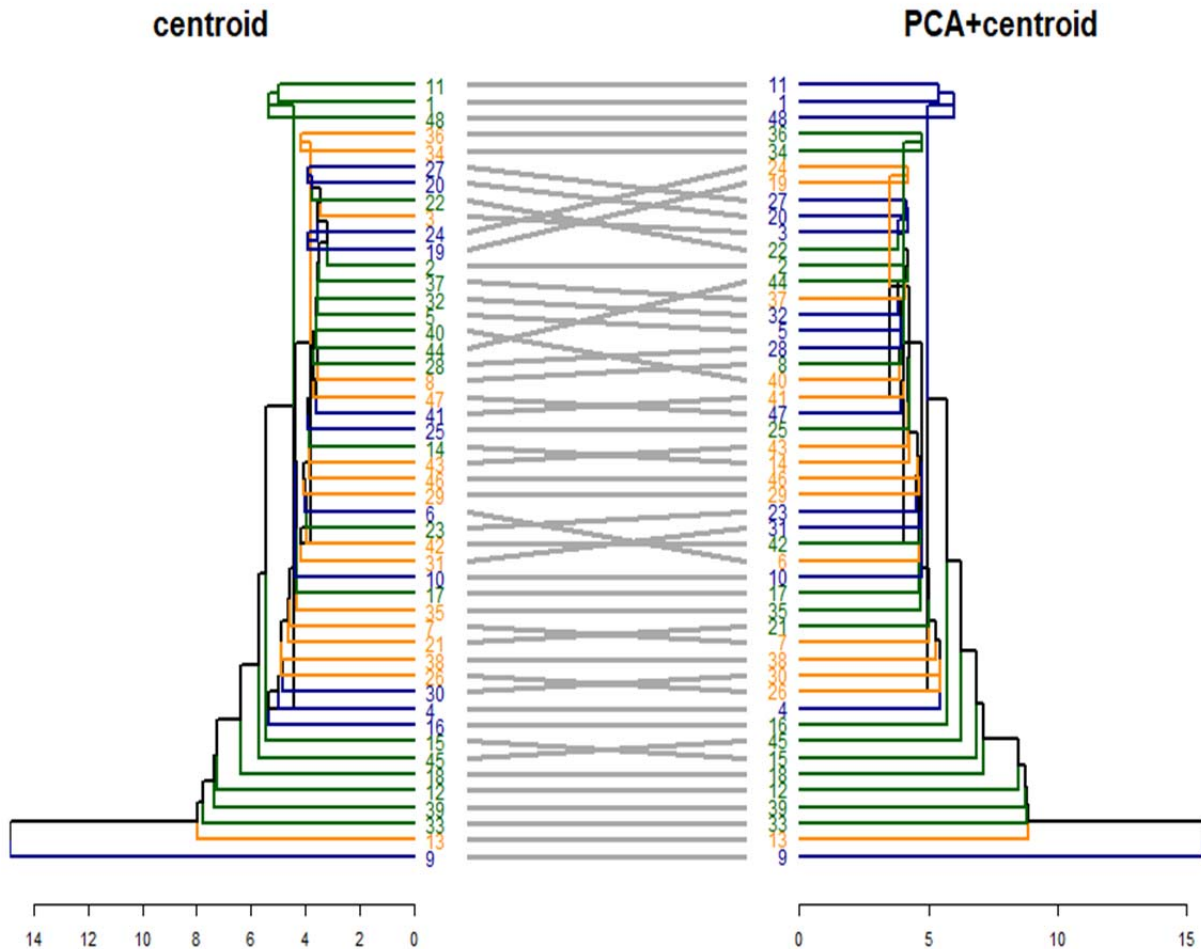


Figure 8.6. 4. Tanglegram for comparison dendrograms of centroid linkage method and PCA+centroid linkage method.

8.6.5. Comparative Results for Median Linkage Method:

For the median linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.94**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.43**, and Fowlkes and Mallows index (**FM_index**) value is **1**. For the Asian dataset, the tanglegram that includes the comparison of the dendrograms of the median linkage method is shown below in **Figure 8.6.5**. Moderate to large displacements are observed in most of the objects. Similar to the centroid method, the BGCC value is also low for the median method. For the first dendrogram (average) the CPCC value is **0,8325** and for the second dendrogram (PCA+average) the CPCC value is **0,8174**. The cpcc values of both dendrograms are high and the cluster contents are the same. That is, the relationship and similarity between the distance matrix and the cophenetic matrices of the dendrograms are high. When the clustering results and scatterplot are evaluated together, we see that China and India form separated clusters as in the single linkage, average linkage and centroid linkage methods.

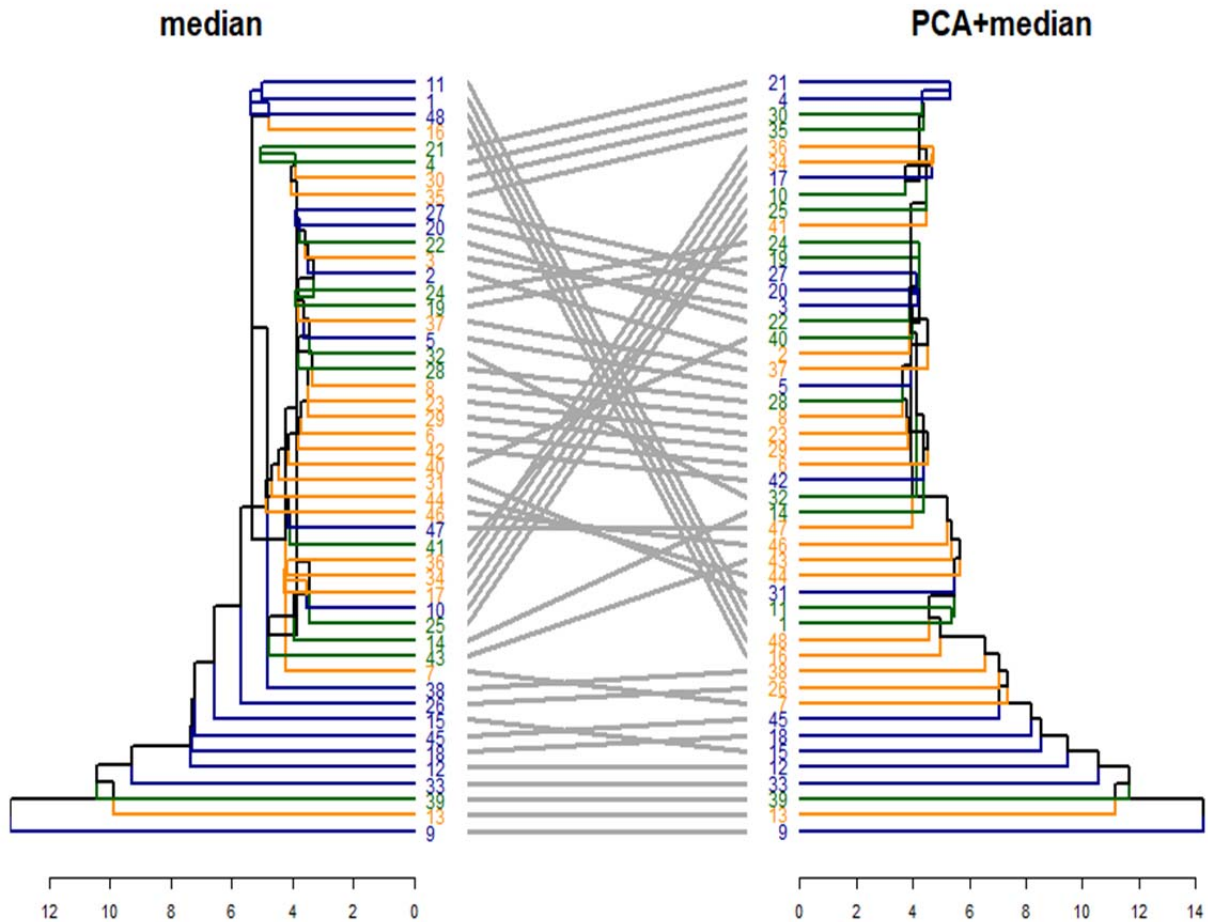


Figure 8.6. 5. Tanglegram for comparison dendrograms of median linkage method and PCA+median linkage method.

8.6.6. Comparative Results for Ward's Criterion Linkage Method:

For the ward's criterion linkage method, the evaluation coefficients and indexes are calculated as follows, respectively: Cophenetic Correlation Coefficient (**CPCC**) value is **0.80**, Baker's Gamma Correlation Coefficient (**BGCC**) value is **0.78**, and Fowlkes and Mallows index (**FM_index**) value is **0.79**. For the Asian dataset, the tanglegram that includes the comparison of the dendrograms of the ward's criterion linkage method is shown below in **Figure 8.6.6**. Except for 9 and 13 (China, India), moderate shifts are observed for all objects. The reason why 13 (India) is considered as a separate cluster for this method and other methods is that its population, GDP, surface area and import values are higher than many other Asian countries. It is noteworthy that the coefficient values for this method are low in concordant with the tanglegram.

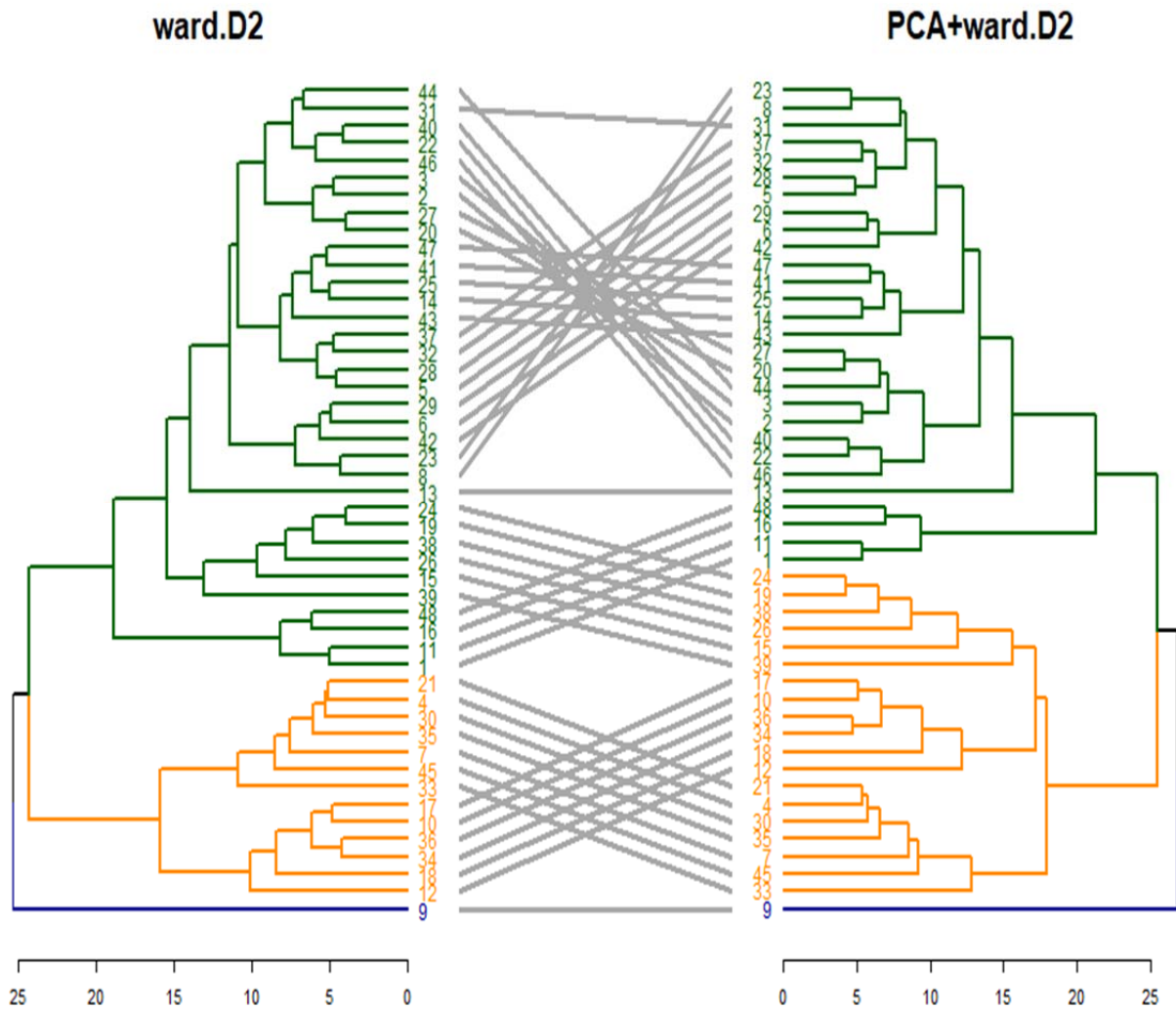


Figure 8.6. 6. Tanglegram for comparison dendrograms of ward.D2 linkage method and PCA+ward.D2 linkage method.

Table 8.6. 3. Clustering results for ward.D2 linkage and PCA+ward.D2 linkage method.

Ward.D2 linkage		
Cluster 1	Cluster 2	Cluster 3
Afghanistan, Armenia, Azerbaijan, Bangladesh, Bhutan, Cambodia, Georgia, India, Indonesia, Iran, Iraq, Jordan, Kazakhstan, Kyrgyzstan, Lao People's Dem. Rep., Lebanon, Malaysia, Maldives, Mongolia, Myanmar, Nepal, Pakistan, Philippines, Sri Lanka, State of Palestine, Syrian Arab Republic, Tajikistan, Thailand, Timor-Leste, Turkey, Turkmenistan, Uzbekistan, Viet Nam, Yemen	Hong Kong SAR, Japan, Republic of Korea, Singapore, Cyprus, Israel, Qatar, United Arab Emirates, Bahrain, Brunei Darussalam, Kuwait, Oman, Saudi Arabia	China

PCA+Ward.D2 linkage		
Cluster 1	Cluster 2	Cluster 3
Afghanistan, Armenia, Azerbaijan, Bangladesh, Bhutan, Cambodia, Georgia, India, Indonesia, Iraq, Kazakhstan, Kyrgyzstan, Lao People's Dem. Rep., Malaysia, Mongolia, Myanmar, Nepal, Pakistan, Philippines, Sri Lanka, Tajikistan, Thailand, Timor-Leste, Turkey, Turkmenistan, Uzbekistan, Viet Nam, Yemen	Hong Kong SAR, Japan, Republic of Korea, Singapore, Cyprus, Israel, Qatar, United Arab Emirates, Bahrain, Brunei Darussalam, Kuwait, Oman, Saudi Arabia, Iran, Syrian Arab Republic, Maldives, State of Palestine, Jordan, Lebanon	China

For the first dendrogram (ward.D2) the CPCC value is **0,5453** and for the second dendrogram (PCA+ward.D2) the CPCC value is **0,5148**. The CPCC values of both dendrograms are quite low and the cluster contents are not the same. That is, the relationship and similarity between the distance matrix and the cophenetic matrices of the dendrograms are quite low. As can be seen, the clustering results and cpcc values of the ward method are quite different from other methods as in the previous datasets. The main reason explaining this is to aim to obtain homogeneous clusters by keeping within cluster variance minimum in the ward method and the cluster size is effective in this method and the dataset used has a lot of outliers with high variance.

9. CONCLUSIONS

In today's world, where virtualization and data production is increasing rapidly, every stage of the data analysis process has become very important. At the point where almost everything turns into data, the increase in the amount and dimension of the data is one of the most important difficulties in the applicability of the existing methods. To cope with these difficulties and to alleviate the analysis load of the methods, many methods are used in the data preprocessing stages. The principal component analysis is one of the most important and frequently used methods for dimension reduction. The main purpose of this thesis study is to examine the effect of principal component analysis on hierarchical methods in terms of dimension reduction for high-dimensional datasets.

All of the datasets used in the study have low correlation levels and include many outliers. Outliers affect the variance in the dataset and reduce the correlation between the variables. Some of the hierarchical methods used (single linkage, complete linkage, and ward's criterion) and PCA are sensitive to outliers. One of the purposes of use of these methods is outlier detection and therefore they are among the methods used in Fraud analysis. In particular, the principal components in PCA are created depending on the variance in the dataset and the correlation between the variables. Therefore, the negative effect of outliers on these methods is inevitable.

In the light of the observed results, while the Elbow method, which is one of the approaches used in PCA, gave misleading and insufficient results, the mean-variance and 80% specific rate approach gave consistent and reliable results. As mentioned in Chapter 5, low correlation between variables reduces the dimension reduction efficiency of PCA (Rencher, 2002). In normal clustering, all variables and the total variance in the dataset are used in the analysis process. However, if dimension reduction with PCA is applied, approximately 80% of the total variance in the dataset can be represented with much fewer variables, depending on the correlation and variance values in the dataset. These few principal components carry the most important information in the dataset and information loss is minimal because there is no variable elimination. The presence of outliers in the data sets and their variance values also affect the dimension reduction performance. Considering the outliers and low correlation in the datasets, effective results are obtained in terms of dimension reduction performance, especially in the 2nd and 3rd datasets. For Dataset1, the mean absolute correlation value is

0.2426, and dimension reduction is performed from 22 variables to 8 variables by PCA; for dataset2, the mean absolute correlation value is 0.2346, and dimension reduction is performed from 38 variables to 10 variables by PCA; and for dataset3 the mean absolute correlation value is 0.2265 and dimension reduction is performed from 46 variables to 11 variables by PCA.

Table 9. 1. Coefficient and index results used to measure the similarity of dendrograms for all datasets and hierarchical clustering methods.

	METHODS	CPCC (dend 1)	CPCC (dend 2)	CPCC (tanglegram)	BGCC	FM_index
Data set 1.	single	0,8519	0,8704	0.7465474	0.9856606	0.9425261
	complete	0,7724	0,7261	0.9655519	0.9867451	1
	average	0,9089	0,9211	0.9710973	0.9809861	0.9201198
	centroid	0,8977	0,9102	0.994414	0.3989163	1
	median	0,7948	0,8365	0.9269909	0.428965	1
	ward.D2	0,5129	0,4961	0.921376	0.95822	0.95849
Data set 2.	single	0,8013	0,8092	0.9959106	0.9813049	1
	complete	0,7480	0,7351	0.8459703	0.8403557	1
	average	0,8588	0,8662	0.509475	0.8847822	1
	centroid	0,8370	0,8491	0.9947938	0.3375059	1
	median	0,7814	0,8022	0.9027062	0.4272787	1
	ward.D2	0,4782	0,4646	0.9947947	0.9996194	1
Data set 3.	single	0,8718	0,8681	0.9959265	0.9962016	1
	complete	0,8458	0,7562	0.5993731	0.7446837	0.7299865
	average	0,9129	0,9108	0.9060994	0.9764306	1
	centroid	0,8714	0,8709	0.9939731	0.5969392	1
	median	0,8325	0,8174	0.9428428	0.430804	1
	ward.D2	0,5453	0,5148	0.8058044	0.7859016	0.7952145

The CPCC (dend 1) values in Table 9.1 give the fit and similarity ratio of the cophenetic matrix and distance matrix of the dendrogram obtained as a result of the without PCA method and also the CPCC (dend 2) values give the fit and similarity ratio of the cophenetic matrix and distance matrix of the dendrogram obtained as a result of the with applying PCA method. Unlike these, the CPCC (tanglegram) shows the fit and similarity ratio between the cophenetic matrices of these two dendrograms. When the CPCC values of the dendrograms are examined, it is noteworthy that the values of the Ward method are quite low for 3 datasets. The main reason explaining this situation is that the Ward method includes a completely

variance-oriented merging process. There are outliers affecting the variance level in all 3 datasets, and this has a negative impact on the Ward method, which performs the merging operation with the criterion of minimizing the total variance within the cluster. In other words, according to the clustering results of the 3 datasets, the technique most negatively affected by outliers is the Ward method. Considering the coefficient values in the table, the best clustering results for all datasets belong to Average linkage and Centroid linkage methods in terms of CPCC(dend1) and CPCC(dend2) values. Both of these methods are methods with low outlier sensitivity. For the first two datasets, Average+PCA and Centroid+PCA have higher values, while in the third dataset, the values of Average with average+PCA, and centroid with centroid+PCA methods are almost the same. For these methods, it is seen that the FM_index values are equal to 1, and therefore the clustering results are the same for the number of clusters 3 except for average and average+PCA in dataset1.

BGCC refers to the similarity coefficient calculated over the concordance and discordance probabilities of the merged pairs of objects. Baker's Gamma Correlation Coefficients of centroid linkage and median linkage methods, shown with red in the table, was given misleading results, incompatible with tanglegrams and other coefficients.

When tanglegrams are observed in which dendrograms can be examined comparatively, the contents of all formed natural clusters are the same, except for a few dendrograms for each dataset. This result is shown both graphically in tanglegrams and is also shown with FM_index results in Table 9.1. In the methods with an index value of 1 in the table, the content of the cluster is exactly the same. For each dataset, all tanglegrams are carefully examined and it is observed that almost all of the dendrograms obtained after the use of PCA formed a much more ordered hierarchical structure. The reason for this can be thought reducing the negative effects of outliers in the dataset on the methods by PCA. When the CPCC (tanglegram) values, which measure the similarity of the cophenetic matrices of the dendrograms, are examined, it is clearly observed that the highest values for all three datasets belong to Centroid and centroid+PCA, in concordance with the graphical results. This situation can be considered as clear proof that the Centroid linkage method is the most compatible with PCA and gives the most robust results.

The coefficients and graphs used in the study enabled us to consider the results from three different perspectives: the compatibility probability of the merged object pairs, the branch heights changing according to the merging criteria (cophenetic matrix), and the contents of

the clusters created. As an evaluation of all coefficients and graphs, for all 3 datasets, the Centroid+PCA method with the highest coefficients and congruent graphical results was found to be more robust and reliable compared to other methods. In addition, despite the low correlation and outlier disadvantages of datasets, it was observed that PCA allows hierarchical methods to work more comfortably in less dimensional space and with less negative variance effect.



10.RECOMMENDATIONS

This study shows that every step of the analysis process should be carefully evaluated. In particular, all the positive or negative features of the dataset used should be examined, the sensitivities and strengths of the methods used should not be ignored. In the light of the results obtained from the study, the evaluations of all graphical methods should be made not only by looking at the coefficient values but also by comparing them with the graphic outputs.

In future studies, the methods, related to this analysis stage, which is not considered important enough, can be discussed in detail. As the recommendations; Feature selection methods, feature extraction methods, or outlier detection methods can be examined and the effect of these strategies on the different analysis methods can be evaluated according to their performance.

REFERENCES

- Abdi, H., & Williams, L. J. (2010). Principal component analysis. John Wiley & Sons, Inc. WIREs Comp Stat, 2, (pp. 433–459). DOI: 10.1002/wics.101.
- Abdulhafedh, A. (2021). Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation. Journal of City and Development. 2021, 3(1), 12-30. DOI: 10.12691/jcd-3-1-3.
- Ah-pine, J., & Wang, X. (2016). Similarity Based Hierarchical Clustering with an Application to Text Collections. In 15th International Symposium on Intelligent Data Analysis (IDA 2016). Stockholm, Sweden. (pp.320-331).
- Akay, Ö. (2018). Panel verilerde kümeleme analizi (Ph.D. Thesis). Çukurova University, Adana.
- Akoğul, S. (2018). Çok değişkenli verilerin modele dayalı kümeleme analizinde kümeleme etkinliğinin artırılması (Ph.D. Thesis). Selçuk University, Konya.
- Al Hasan, M., Chaoji, V., Salem, S., & Zaki, M. J. (2009). Robust Partitional Clustering by Outlier and Density Insensitive Seeding. Elsevier Publication.
- Alhenaki, L. (2019). A Genetic-Frog Leaping Algorithm For Text Document Clustering. Thesis for: Master in Computer Science. Researchgate Publication.
- Alkın, D. (2019). Kümeleme analizi yöntemlerinin küme geçerlilik indekslerine göre karşılaştırılması (M.Sc. Thesis). Necmettin Erbakan University, Konya.
- Altunkaynak, B. (2019). Veri madenciliği yöntemleri ve R uygulamaları (2.Baskı). Seçkin Yayıncılık San.ve Tic. A.Ş.
- Analytics Vidhya (2021). Retrieved in 2021 from <https://www.analyticsvidhya.com/blog/2020/09/how-dbscan-clustering-works/>
- Ay, F. (2019). Clustering analysis of young isolated neutron stars on p-p parameter space (M.Sc. Thesis). Istanbul Technical University, İstanbul.

Ayed, A. B., Halima, M. B., & Alimi, A. M. (2014). Survey on clustering methods : Towards fuzzy clustering for big data. In International Conference of Soft Computing and Pattern Recognition. IEEE (pp. 331-336).

Berkhin, P., (2002). Survey Of Clustering Data Mining Techniques. Researchgate Publication. DOI:10.1007/3-540-28349-8_2.

Birbiçer, İ. (2017). Veri madenciliğinde modele dayalı kümeleme analizi (M.Sc. Thesis). Çukurova University, Adana.

Börü, M. (2009). The determinants of financial development in Turkey: A principal component analysis (M.Sc. Thesis). Middle East Technical University, Ankara.

Bu, J., Liu, W., Pan, Z., & Ling, K. (2020). Comparative Study of Hydrochemical Classification Based on Different Hierarchical Cluster Analysis Methods. International Journal of Environmental Research and Public Health, 17, 9515; doi:10.3390/ijerph17249515.

Bulut, H. (2017). Çok değişkenli verilerde robust kümeleme analizi ile boyut indirgeme ve birimlerin sınıflandırılması (Ph. D. Thesis). Ondokuz Mayıs University, Samsun.

Campello, R. J. G. B., Kröger, P., Sander, J., & Zimek, A. (2020). Density-based clustering. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. University of Southern Denmark. DOI:10.1002/widm.1343.

Candes, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust Principal Component Analysis. Journal of the ACM, Vol. 58, No. 3, Article 11. DOI: 10.1145/1970392.1970395.

Cenger, A. (2018). Parçacık sürü optimizasyonu ve veri madenciliği yöntemleriyle kümeleme analizi: İzmirde kurulu bir firmanın müşterilerinin kümelenmesi uygulaması (M.Sc. Thesis). Dokuz Eylül University, İzmir.

Chavent, M. (1998). A monothetic clustering method. Pattern Recognition Letters, Volume 19, Issue 11, (pp. 989-996). [https://doi.org/10.1016/S0167-8655\(98\)00087-7](https://doi.org/10.1016/S0167-8655(98)00087-7).

Cheng, W., Wang, W., & Batista, S. (2014). Chapter 6: Grid-Based Clustering. In Data Clustering Algorithms and Applications. CRC Press Taylor & Francis Group.

Çizmeçi, N. (2019). Açıklayıcı faktör analizi ve kümeleme analizi ile Matematik ve Geometri tutum ölçeği üzerine bir çalışma (M.Sc. Thesis). Akdeniz University, Antalya.

D., N., & Vidyavathi, B. M. (2015). A Survey on Applications of Data Mining using Clustering Techniques. *International Journal of Computer Applications* (0975 – 8887) Volume 126 – No.2.

Dattorro, J. (2005). *Convex optimization & Euclidean distance geometry*. Meboo Publishing, USA.

Day, W. H. E., & Edelsbrunner, H. (1984). Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of Classification* volume 1, pages7–24.

De Luca, M., Terouzi, W., Loele, W., Kzaiber, F., Oussama, A., Oliverio, F., Tauler, R., & Ragno, G. (2011). Derivative FTIR spectroscopy for cluster analysis and classification of morocco olive oils. *Food Chemistry* 124 (2011) 1113–1118.

Doğanlarlıoğlu, A. (2009). *Uzaktan algılamada geliştirilmiş temel bileşen analizi ile görüntü birleştirme* (M.Sc. Thesis). İstanbul Teknik Üniversitesi, İstanbul.

El-Hamdouchi, A., & Willett, P. (1987). Techniques for the measurement of clustering tendency in document retrieval systems. *Journal of Information Science*. DOI:10.1177/016555158701300607.

Emmendorfer, L. R., & Canuto, A. M. P. (2021). A generalized average linkage criterion for Hierarchical Agglomerative Clustering. *Applied Soft Computing Journal*. <https://doi.org/10.1016/j.asoc.2020.106990>.

Engeloğlu, Ö. (2020). *AB ülkeleri ve Türkiye'de tüketici ve üretici güven endeksinin belirleyicileri: asimetrik nedensellik testi ve kümeleme analizi* (Ph.D. Thesis). Hacı Bayram Veli University, Ankara.

Erilli, N.A. (2012). *Bulanık sayıların bulanık kümeleme analizinde kullanımı ve satranç oyuncularının sınıflandırılması* (Ph.D. Thesis). Ondokuz Mayıs University, Samsun.

Ester, M. (2014). *In Data Clustering Algorithms and Applications*. CRC Press Taylor & Francis Group.

Everitt, B. S., Landau, S., Leese, M. & Stahl, D. (2011). *Cluster Analysis* (5. Edition). John Wiley & Sons, Ltd. Publication.

Fırat, F. (2016). *Benzerlik ölçüleri ve kümeleme analizi* (M.Sc. Thesis). İstanbul University, İstanbul.

- Forina, M., Armanino, C., & Raggio, V. (2002). Clustering with dendrograms on interpretation variables. Elsevier, *Analytica Chimica Acta* 454 (2002) 13–19.
- Fowlkes, E. B., & Mallows, C. L. (1983). A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, Vol. 78, No. 383, (pp. 553-569). Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association.
- Gaitani, N., Lehmann, C., Santamouris, M., Mihalakakou, G., & Patargias, P. (2010). Using principal component and cluster analysis in the heating evaluation of the school building sector. *Applied Energy* 87 (2010) 2079–2086.
- Gemperline, P. J. (2006). Chapter 4: Principal Component Analysis. In *Practical Guide to Chemometrics* (2nd Edition). Published by CRC Press, Taylor & Francis Group.
- Gümüş, Z.P. (2015). Differentiation of olive oil produced in different areas in Turkey based on major parameters using clusters/PCA (Principal Component Analysis) methods (Ph.D. Thesis). Ege University, İzmir.
- Hacıoğlu, H.K. (2016). Kümeleme analizinde kullanılan bazı benzerlik indekslerinin karşılaştırılması (M.Sc. Thesis). Gazi University, Ankara.
- Hajkacem, M. A. B., Ben n'cir, C., & Nadia, E. (2019). Clustering Methods for Big Data Analytics, Unsupervised and Semi-Supervised Learning. Researchgate.net publication.
- Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On Clustering Validation Techniques. *Journal of Intelligent Information Systems*, 17:2/3, (pp. 107–145). Kluwer Academic Publishers. Manufactured in The Netherlands.
- Han, J., Kamber, M., & Pei, J. (2012). *Data Mining Concepts and Techniques* (3rd Edition). Morgan Kaufmann Publication, USA.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6), 417–441. <https://doi.org/10.1037/h0071325>.
- Hryniewicz, O. (2006). Goodman–Kruskal measure of dependence for fuzzy ordered categorical data. *Computational Statistics & Data Analysis* 51, (pp. 323 – 334). DOI: 10.1016/j.csda.2006.04.014.
- İnal, S. (2020). Fonksiyonel temel bileşen analizi (M.Sc. Thesis). Yıldız Teknik University, İstanbul.

İncekırık, A. (2005). Çok değişkenli istatistiksel bir boyut indirgeme yöntemi olarak kümeleme analizi ve bir uygulama (M.Sc. Thesis). Dokuz Eylül University, İzmir.

Jarman, A. M. (2020). Hierarchical Cluster Analysis: Comparison of Single linkage, Complete linkage, Average linkage and Centroid Linkage Method. Researchgate publication, DOI: 10.13140/RG.2.2.11388.90240.

Karamizadeh, S., Abdullah, S. M., Manaf, A. A., Zamani, M., & Hooman, A. (2013). An Overview of Principal Component Analysis. Journal of Signal and Information Processing, 4, (pp. 173-175). DOI: 10.4236/jsip.2013.43B031.

Kassambara, A. (2017). Multivariate Analysis I (1.Edition). STHDA.

Kaufman, L., & Rousseeuw, P. J. (1990, 2005). Finding Groups in Data, An Introduction to Cluster Analysis. Published by John Wiley & Sons, Inc., Hoboken, New Jersey.

Kaya, A. (2018). Bulanık kümeleme analizinde bulanık kümeleme algoritmalarının karşılaştırılması (M.Sc. Thesis). Anadolu University, Eskişehir.

Kaya, A.B. (2020). Türkiye kentlerinin modernlik ve postmodernlik kademelenmesinin SEGE göstergeleri ile temel bileşenler ve kümeleme analizi (M.Sc. Thesis). Gazi University, Ankara.

Keskin, E. (2016). Koyunlarda üreme performansına etki eden faktörler üzerine temel bileşen analizi (M.Sc. Thesis). Atatürk University, Erzurum.

Kettani, O., Ramdani, F., & Tadili, B. (2014). An Agglomerative Clustering Method for Large Data Sets. International Journal of Computer Applications (0975 – 8887) Volume 92 – No.14.

Kılıç, A. F. (2021). Kategorik Veride Faktör Analizi İçin Kullanılabilecek Alternatif Bir Korelasyon Matrisi: Goodman-Kruskal Gamma. Marmara Üniversitesi Atatürk Eğitim Fakültesi Eğitim Bilimleri Dergisi, 54, (pp. 151-168). DOI: 10.15285/maruaebd.853905.

Kim, S.J., Ha, T.J., Kim, J., Chang, D.C., & Kim, K.S. (2014). Classification of Korean Chrysanthemum Species based on Volatile Compounds Using Cluster Analysis and Principal Component Analysis. J Korean Soc Appl Biol Chem (2014) 57(6), 789–796. DOI 10.1007/s13765-014-4162-5.

Koij, F.S. & Saba, J. (2015). Using Cluster Analysis and Principal Component Analysis to Group Lines and Determine Important Traits in White Bean. Agriculture and Climate Change - Adapting Crops to Increased Uncertainty (AGRI 2015). Procedia Environmental Sciences 29 (2015) 38 – 40.

Kumar, S., & Toshniwal, D. (2016). Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient (CPCC). The Journal of Big Data, (3:13). DOI 10.1186/s40537-016-0046-3.

Levent, M. (2019). Edas yöntemi ve kümeleme analizi ile G-10 ülkelerinin ekonomik özgürlük kriterleri ile değerlendirilmesi (M.Sc. Thesis). İstanbul Aydın University, İstanbul.

Madhulatha, T. S. (2012). An Overview on Clustering Methods. IOSR Journal of Engineering. Vol. 2(4) (pp: 719-725).

Majerova, I., & Nevima, J. (2017). The measurement of human development using the Ward method of cluster analysis. Journal of International Studies, 10(2), 239-257. doi:10.14254/2071-8330.2017/10-2/17.

Malkoç, H. (2018). Kümeleme analizinde yapay sinir ağlarının küme geçerlilik indeksi olarak kullanılması (M.Sc. Thesis). Cumhuriyet University, Sivas.

Manning, C. D., Raghavan, P., & Schütze, H. (2009). Hierarchical Clustering, In Introduction to Information Retrieval. Publisher: Cambridge University Press. DOI: <https://doi.org/10.1017/CBO9780511809071.018> (pp. 346-368).

Margaritis, A., Soenen, H., Fransen, E., Pipintakos, G., Jacobs, G., Bloom, J., & Den Bergh, W.W. (2020). Identification of ageing state clusters of reclaimed asphalt binders using principal component analysis (PCA) and hierarchical cluster analysis (HCA) based on chemo-rheological parameters. Article in Construction and Building Materials, May 2020. DOI: 10.1016/j.conbuildmat.2020.118276.

Mohbey, K. K., & Thakur, G. S. (2013). An Experimental Survey on Single Linkage Clustering. International Journal of Computer Applications (0975 – 8887), Volume 76–No.17.

Mudrova, M., & Prochazka, A. (2005). Principal Component Analysis In Image Processing. PDF Available on Researchgate.net. Institute of Chemical Technology, Prague.

Murtagh, F., & Legendre, P. (2014). Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. *Journal of Classification*, 31 (3), 274-295. DOI: 10.1007/s00357-014-9161-z.

Nino, J. O. P., & Berzal, F. (2019). Evaluation Metrics for Unsupervised Learning Algorithms. PDF Available on ArXiv.

Nwangburuka, C.C., Kehinde, O.B., Ojo, D.K., Denton, O.A., & Popoola, A.R. (2011). Morphological classification of genetic diversity in cultivated okra, *Abelmoschus esculentus* (L) Moench using principal component analysis (PCA) and single linkage cluster analysis (SLCA). *African Journal of Biotechnology* Vol. 10(54), pp. 11165-11172.

Orloci, L. (1967). An Agglomerative Method for Classification of Plant Communities. *The Journal of Ecology*, Vol. 55, No. 1, (pp. 193-206). Published by: British Ecological Society, JSTOR.

Our World in Data (2021). Retrieved in 2021 from <https://ourworldindata.org/>

Öten, M. & Albayrak, S. (2018). Bazı Yonca (*Medicago sativa* L.) Genotipleri Arasındaki Varyasyonun Kümeleme ve Temel Bileşen Analizi Metodu ile Belirlenmesi. *Turk J Agric Res* 2018, 5(3): 222-228.

Özaltındış, T. (2018). Mekansal-zamansal veri madenciliğinde kümeleme analizi (M.Sc. Thesis). Mimar Sinan Güzel Sanatlar University, İstanbul.

Özgen, H.G. (2019). İç göç verilerini kullanarak kümeleme analizinde elde edilen kümelerin geçerliliğinin sınanması (M.Sc. Thesis). Gazi University, Ankara.

Özgür, Ç. (2017). Kümeleme analizi ile bütünleşik bulanık topsis ve bulanık vikor yöntemlerine dayanarak hastane kuruluş yer seçimi (M.Sc. Thesis). Yıldız Teknik University, İstanbul.

Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2, (pp. 559-572). <https://doi.org/10.1080/14786440109462720>.

Penkova, T.G. (2017). Principal component analysis and cluster analysis for evaluating the natural and anthropogenic territory safety. *International Conference on Knowledge Based and Intelligent Information and Engineering Systems, KES2017*, 6-8 September 2017, Marseille, France. *Procedia Computer Science* 112 (2017) 99–108.

- Rani, Y., & Rohil, H. (2013). A Study of Hierarchical Clustering Algorithm. *International Journal of Information and Computation Technology*, Volume 3, (pp. 1115-1122). International Research Publications House.
- Rao, A. R., & Srinivas, V. V. (2006). Regionalization of watersheds by hybrid-cluster analysis. *Journal of Hydrology* 318 (pp. 37–56). DOI: 10.1016/j.jhydrol.2005.06.003.
- Reddy, C. K., & Vinzamuri, B. (2014). A Survey of Partitional and Hierarchical Clustering Algorithms. In *Data Clustering Algorithms and Applications*. CRC Press Taylor & Francis Group.
- Rencher, A. C. (2002). *Methods of Multivariate Analysis* (2.Edition). John Wiley & Sons, Inc. Publication. Brigham Young University.
- Rokach, L., & Malmon, O. (2005). *Clustering Methods*. Data Mining And Knowledge Discovery Handbook. Researchgate.net publication, DOI: 10.1007/0-387-25465-X_15.
- Saraçlı, S., Doğan, N., & Doğan, I. (2013). Comparison of hierarchical cluster analysis methods by cophenetic correlation. Article In *Journal of Inequalities and Applications*. DOI: 10.1186/1029-242X-2013-203.
- Sembiring, R. W., Zain, J. M., & Embong, A. (2010). A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course. *Journal Of Computing*, Volume 2, Issue 12.
- Servi, T. (2009). Çok değişkenli karma dağılım modeline dayalı kümeleme analizi (Ph.D. Thesis). Çukurova University, Adana.
- Shoba, D., Vijayan, R., Robin, S., Manivannan, N., Iyanar, K., Arunachalam, P., Nadarajan, N., Pillai, P.A., & Geetha, S. (2019). Assessment of genetic diversity in aromatic rice (*Oryza sativa* L.) germplasm using PCA and cluster analysis. *Electronic Journal of Plant Breeding*, 10 (3): 1095 – 1104.
- Silva, A. R., & Dias, C. T. S. (2013). A cophenetic correlation coefficient for Tocher's method. *Pesq. agropec. bras.*, Brasília, v.48, n.6, (pp.589-596). DOI: 10.1590/S0100-204X2013000600003.

Sokal, R. R., & Rohlf, F. J. (1962). The Comparison of Dendrograms by Objective Methods. *Taxon*, Vol. 11, No. 2, (pp. 33-40). Published by: International Association for Plant Taxonomy (IAPT).

Statista (2021). Retrieved in 2021 from <https://www.statista.com/statistics/871513/worldwide-data-created/>

Szekely, G. J., & Rizzo, M. L.(2005). Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. *Journal of Classification* 22:151-183. DOI: 10.1007/s00357-005-0012-9.

Taşatan, A. (2018). Hiyerarşik kümeleme tekniklerinde küme eleman sayısının eşitlenmesine yönelik bir yaklaşım önerisi ve gerçek karayolu uzaklık verilerine dayalı kümeleme analizi (M.Sc. Thesis). Kocaeli University, Kocaeli.

Tharwat, A. (2016). Principal component analysis - a tutorial. *International Journal of Applied Pattern Recognition* 3(3):197. DOI:10.1504/IJAPR.2016.079733.

Tokuda, E. K., Comin, C.H., & Costa, L. D. (2020). Revisiting Agglomerative Clustering. Published on ArXiv, DOI:10.1016/j.physa.2021.126433.

Towards Data Science (2021). Retrieved in 2021 from <https://towardsdatascience.com/dbscan-clustering-explained-97556a2ad556>

Triayudi, A.,& Fitri, I. (2018). Comparison Of Parameter-Free Agglomerative Hierarchical Clustering Methods. *ICIC Express Letters*, Volume 12, Number 10, (pp. 973-980). DOI: 10.24507/icicel.12.10.973.

Unicef Data (2021). Retrieved in 2021 from <https://data.unicef.org/>

United Nations Data (2021). Retrieved in 2021 from <https://data.un.org/>

Upadhyaya, B. R., Lu, B., Zhao, K., Mullens, J. A., & Leuba, J. M. (2002). Neri Project 99-119. Task 2. Data-Driven Prediction Of Process Variables. Final Report. Researchgate.net, DOI:10.2172/885670.

V, S., & A, G. H. (2017). Appraising Research Direction & Effectiveness of Existing Clustering Algorithm for Medical Data. *International Journal of Advanced Computer Science and Applications*, Vol. 8, No. 3.

Van Der Maaten, L., Postma, E., & Van Den Herik, J. (2009). Dimensionality Reduction: A Comparative Review. *Journal of Machine Learning Research*, 10, 1-41.

World Bank Open Data (2021). Retrieved in 2021 from <https://data.worldbank.org/>

Worldometer – Real Time World Statistics (2021). Retrieved in 2021 from <https://www.worldometers.info/>

Xu, R., & Wunsch, D. C. (2009). *Clustering*. John Wiley & Sons, Inc., Publication. Hoboken, New Jersey.

Yelmen, E. (2020). Soru dokümanlarının anlamsal benzerliklerine dayalı derin öğrenme tabanlı kümeleme analizi (M.Sc. Thesis). Kocaeli University, Kocaeli.

Yıldız, K., Çamurcu, A.Y., & Doğan, B. (2010). A Comparative Analysis of Principal Component Analysis and Non-Negative Matrix Factorization Techniques in Data Mining. *Akademik Bilişim'10 - XII. Akademik Bilişim Konferansı Bildirileri* 10 - 12 Şubat 2010 Muğla Üniversitesi.

Yim, O., & Ramdeen, K. T. (2015). Hierarchical Cluster Analysis: Comparison of Three Linkage Measures and Application to Psychological Data. *The Quantitative Methods for Psychology*, Vol.11, no.1. DOI: 10.20982/tqmp.11.1.p008.

Zhong, C., Miao, D., Wang, R., & Zhou, X. (2008). DIVFRP: An automatic divisive hierarchical clustering method based on the furthest reference points. *Pattern Recognition Letters* 29, (pp. 2067–2077). DOI: 10.1016/j.patrec.2008.07.002.

Zojaji, M., Meybodi, M. R. M., & Mirzaie, K. (2020). A rapid learning automata-based approach for generalized minimum spanning tree problem. *Journal of Combinatorial Optimization*. Springer, DOI:10.1007/s10878-020-00605-0.

APPENDICES

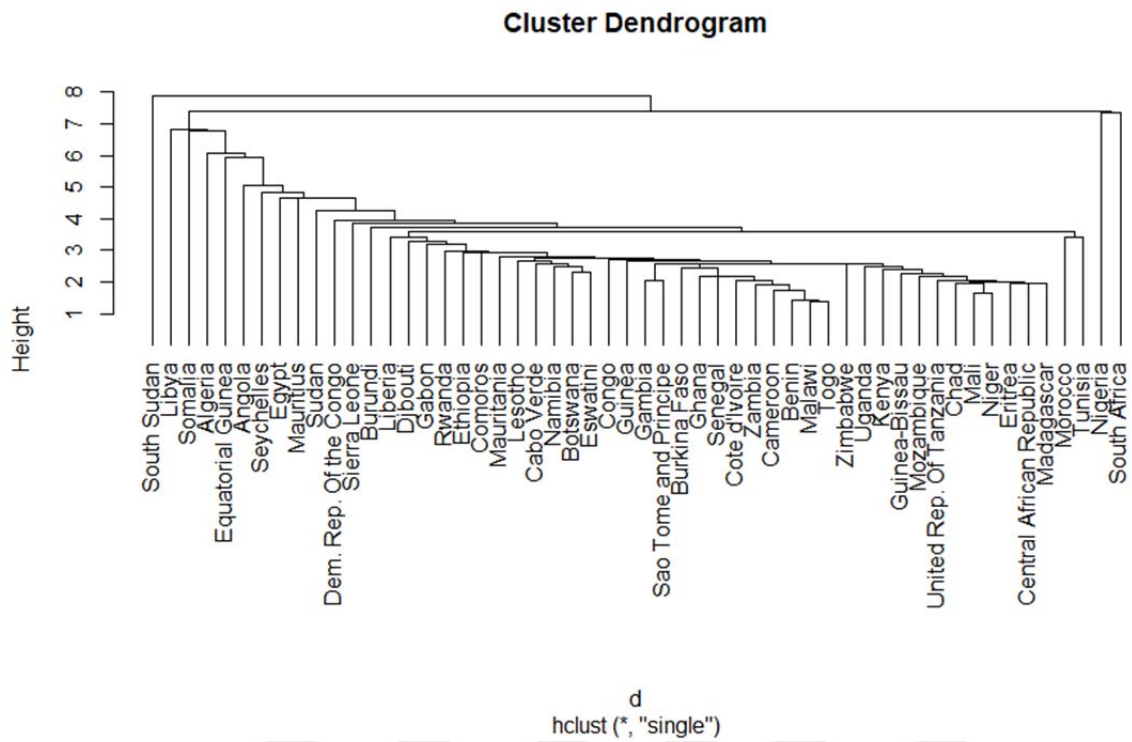


Figure A1. Single linkage clustering results for the African countries.

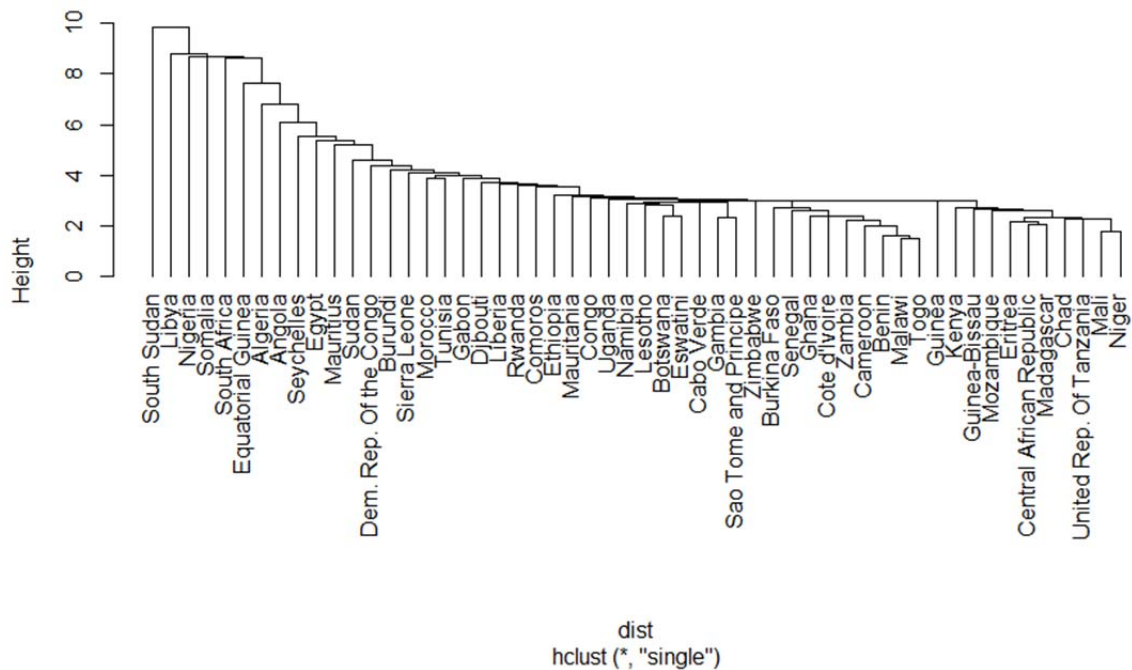


Figure A2. Single linkage clustering results after applying PCA for the African countries.

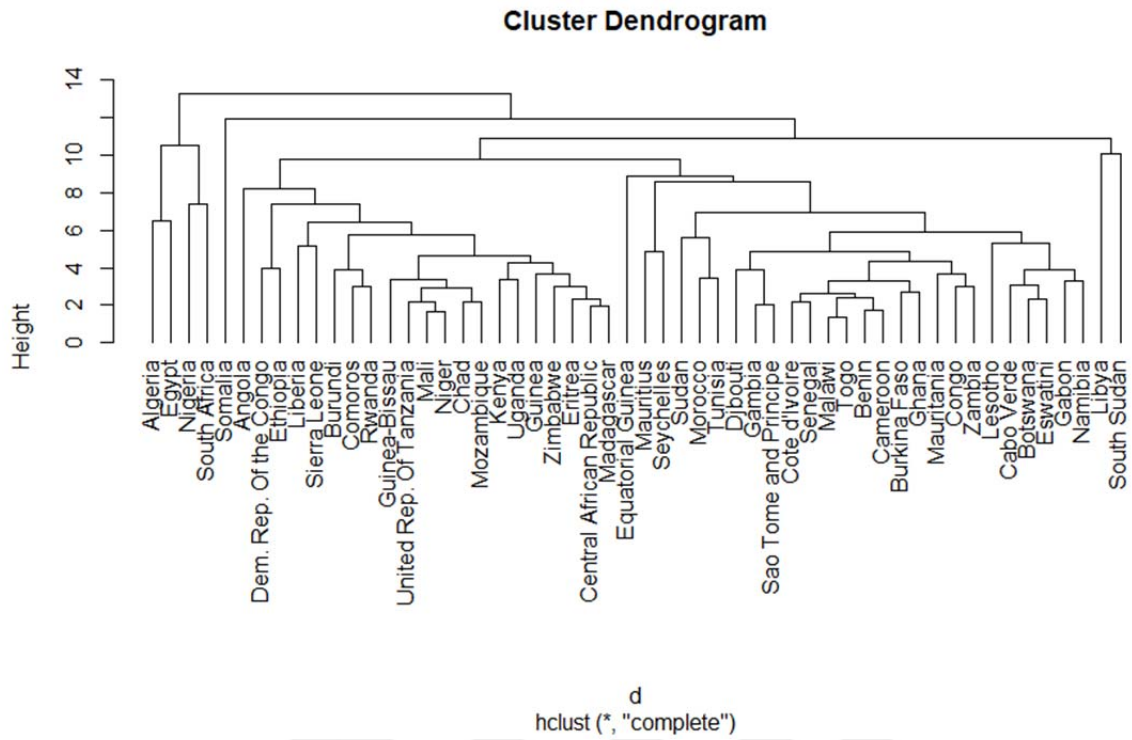


Figure A3. Complete linkage clustering results for the African countries.

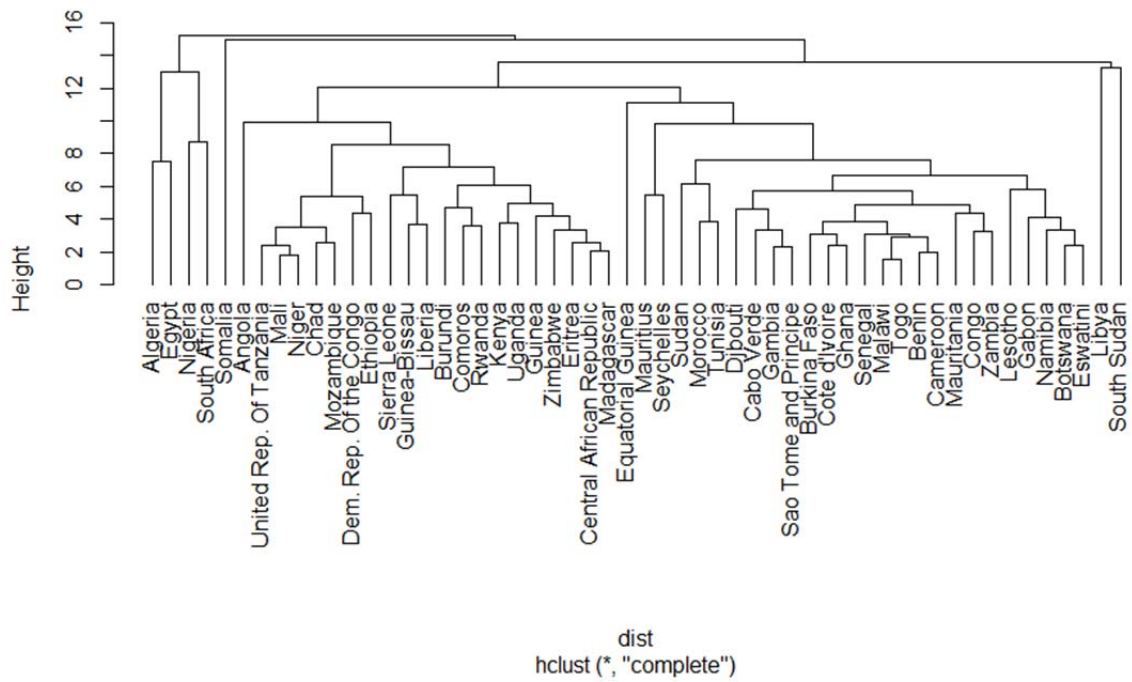


Figure A4. Complete linkage clustering results after applying PCA for the African countries.

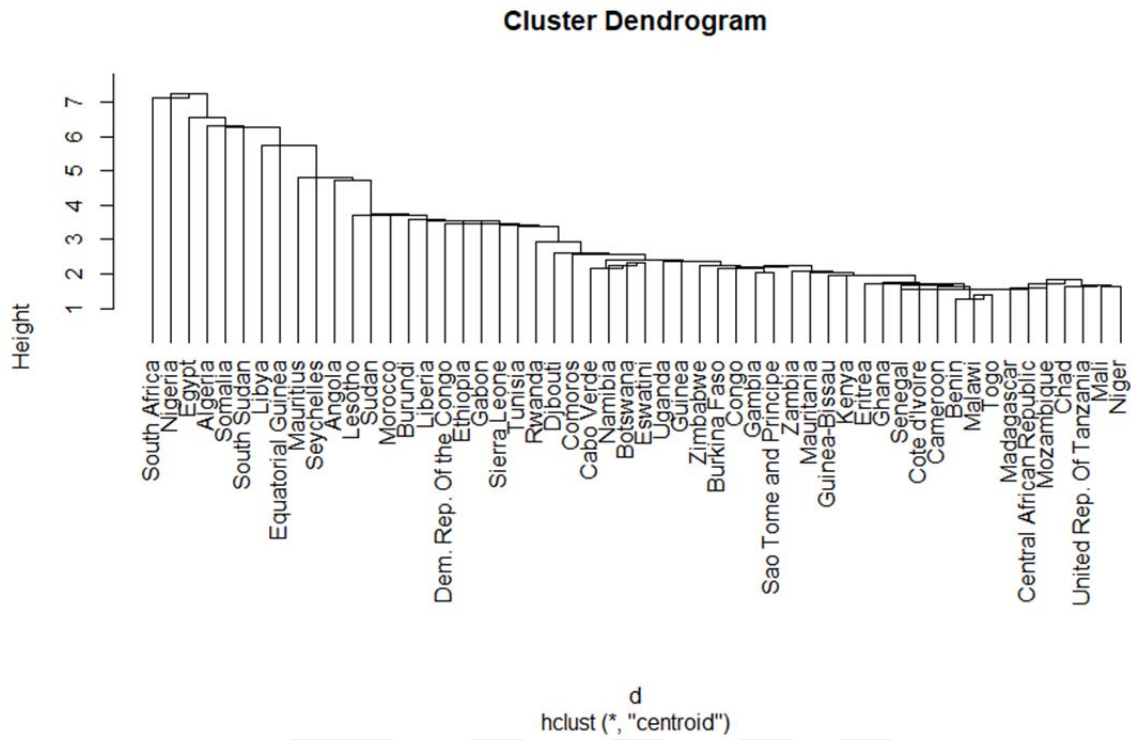


Figure A7. Centroid linkage clustering results for the African countries.

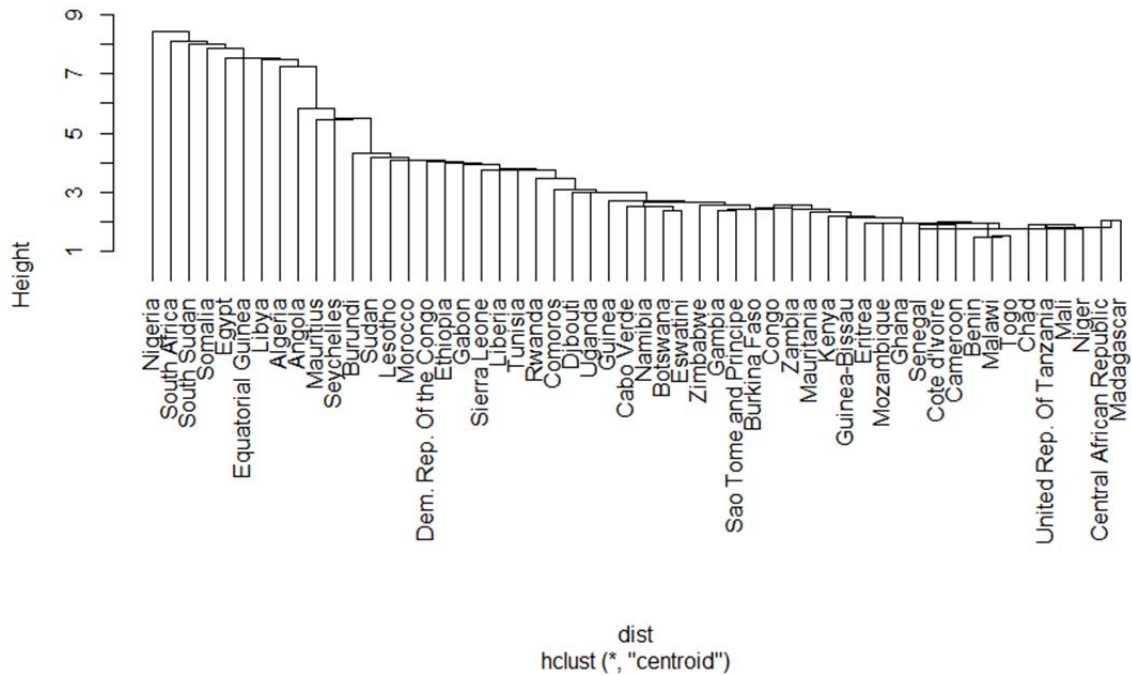


Figure A8. Centroid linkage clustering results after applying PCA for the African countries.

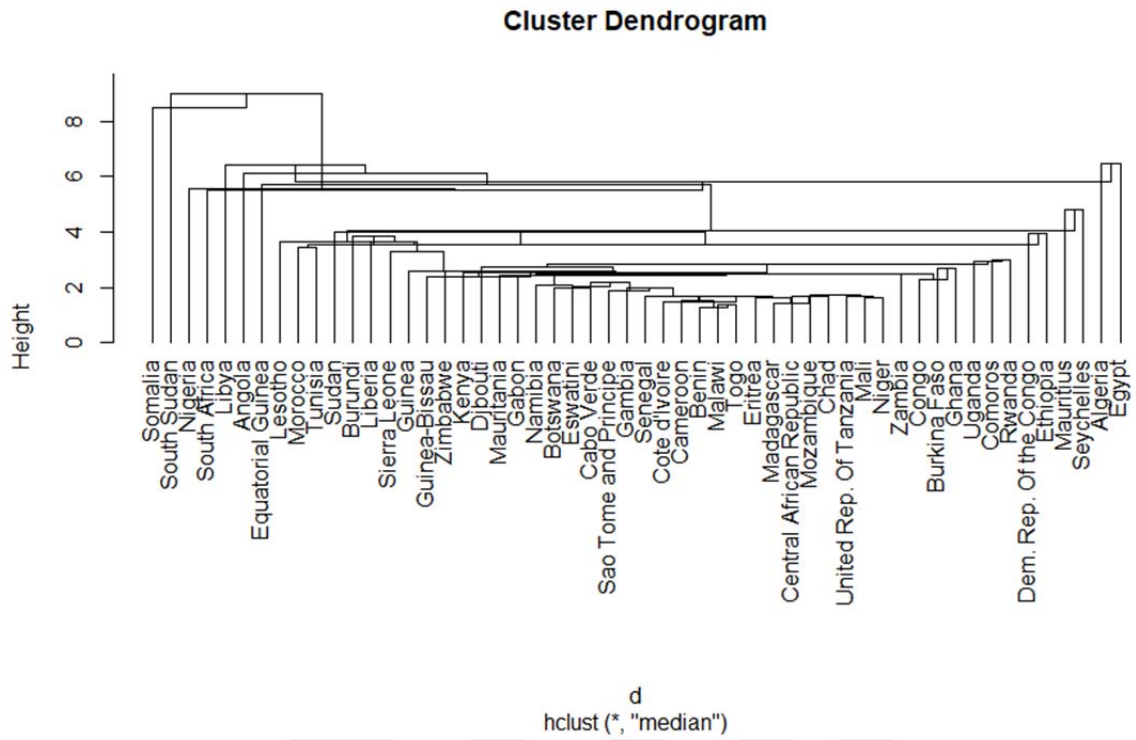


Figure A9. Median linkage clustering results for the African countries.

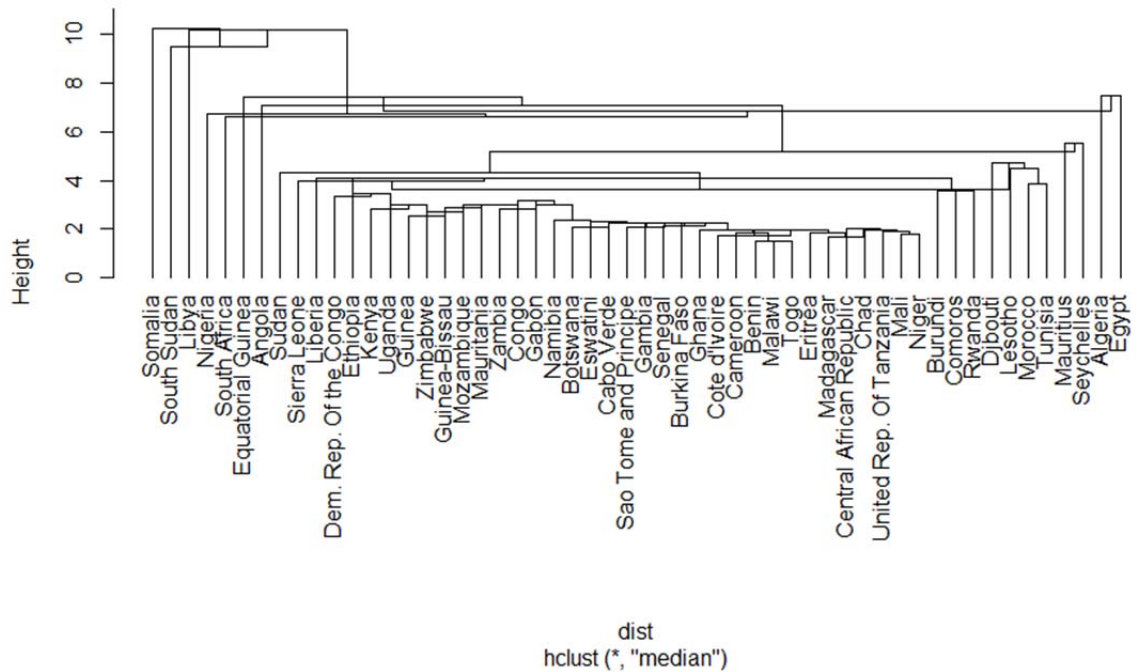


Figure A10. Median linkage clustering results after applying PCA for the African countries.

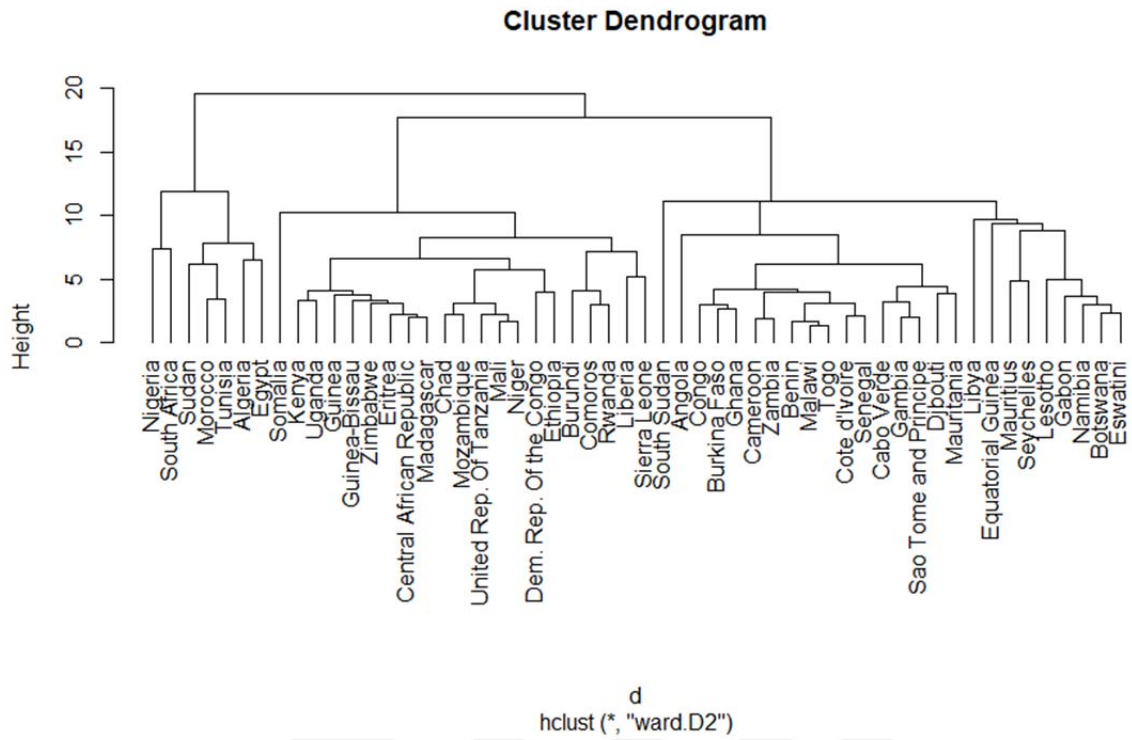


Figure A11. Ward's Criterion linkage clustering results for the African countries.

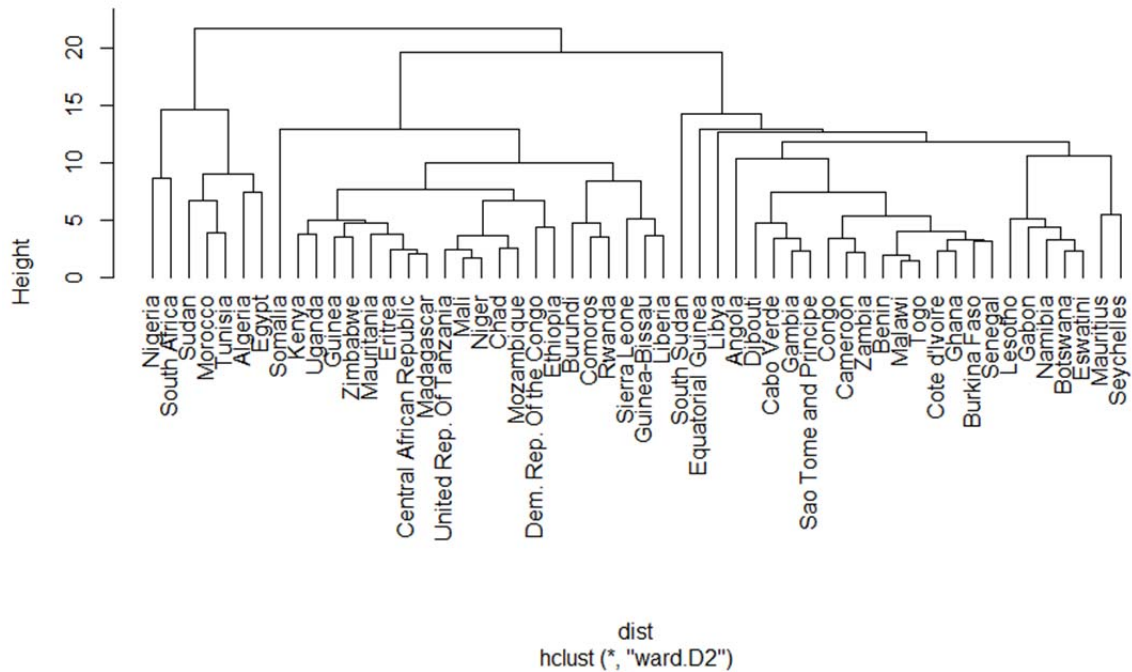


Figure A12. Ward's Criterion linkage clustering results after applying PCA for the African countries.

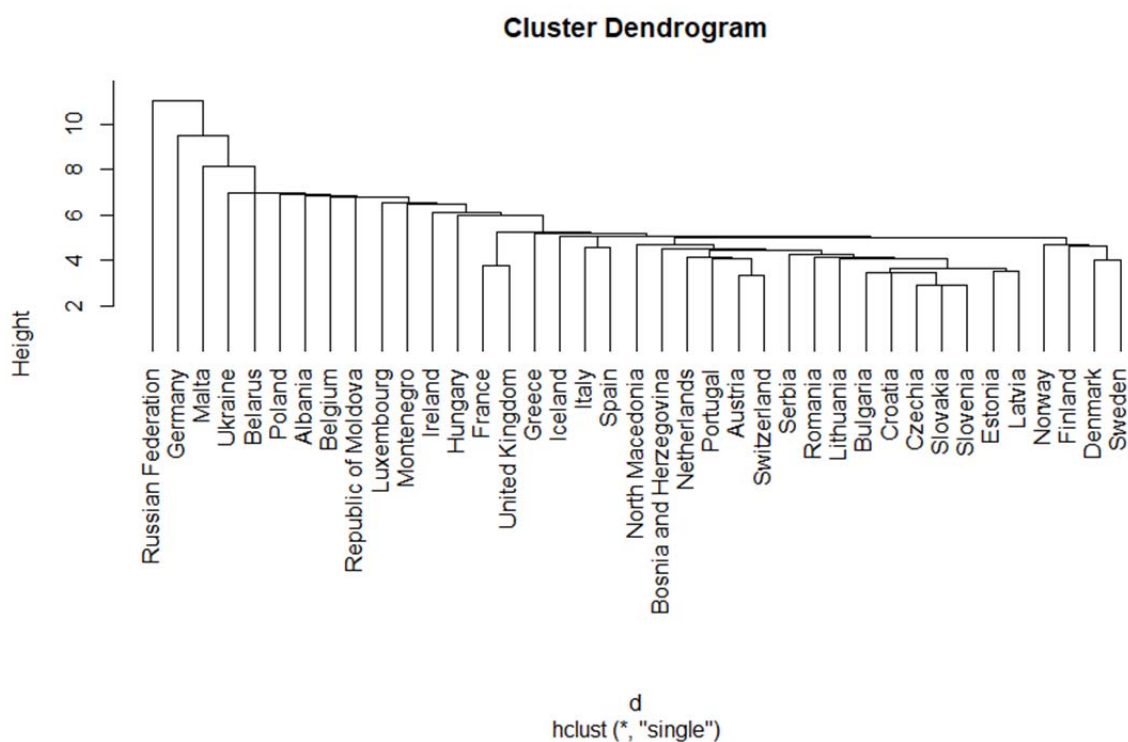


Figure A13. Single linkage clustering results for the Europe countries.

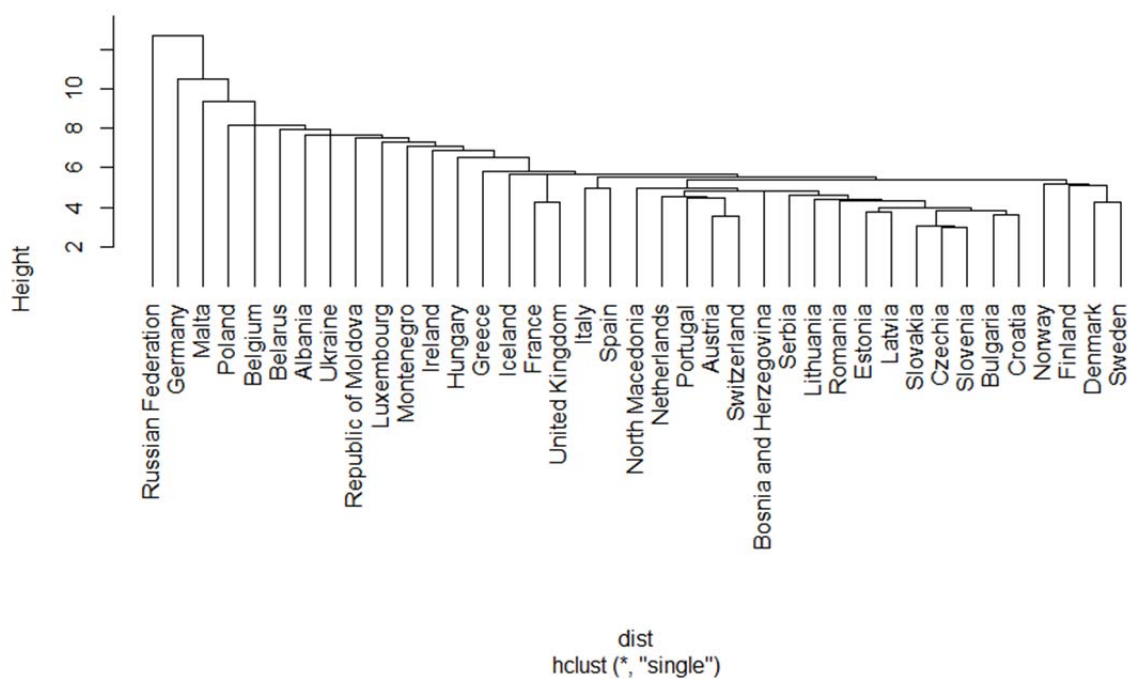


Figure A14. Single linkage clustering results after applying PCA for the Europe countries.

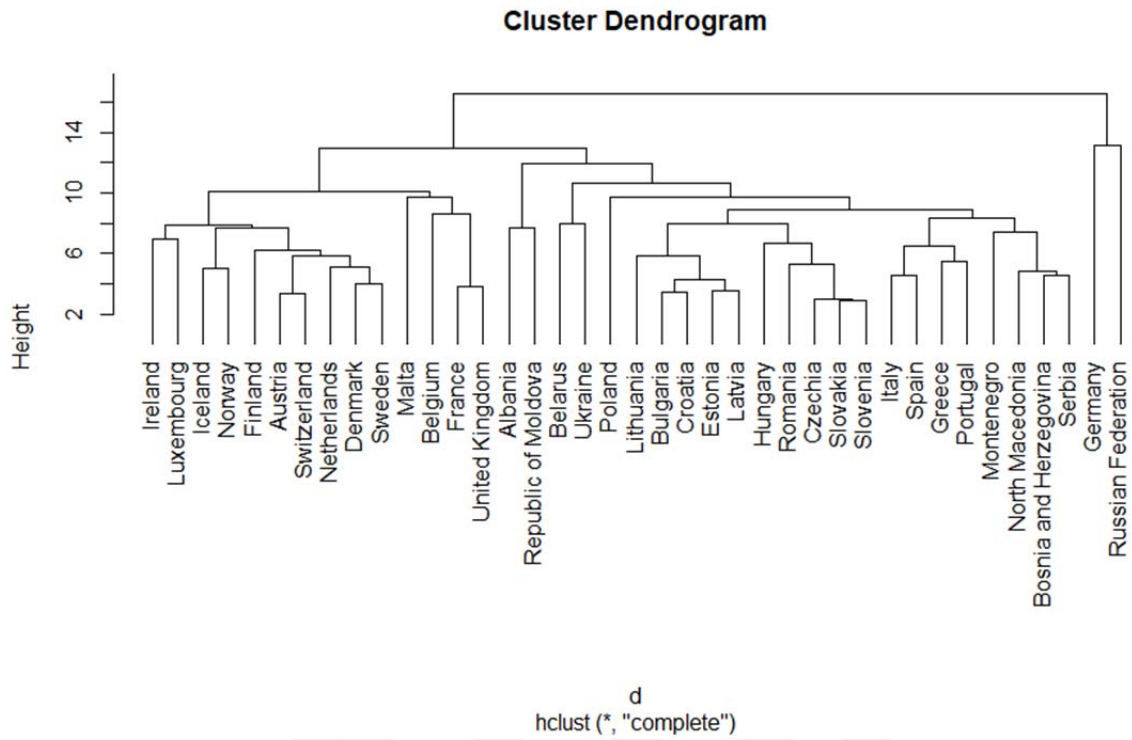


Figure A15. Complete linkage clustering results for the Europe countries.

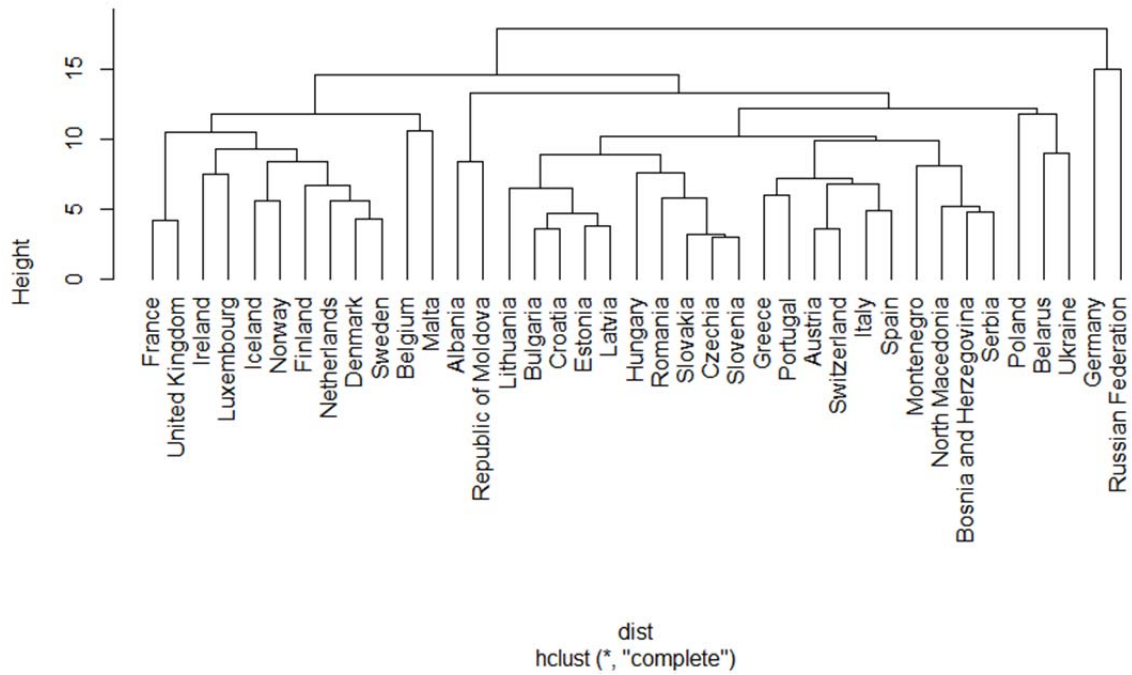


Figure A16. Complete linkage clustering results after applying PCA for the Europe countries.

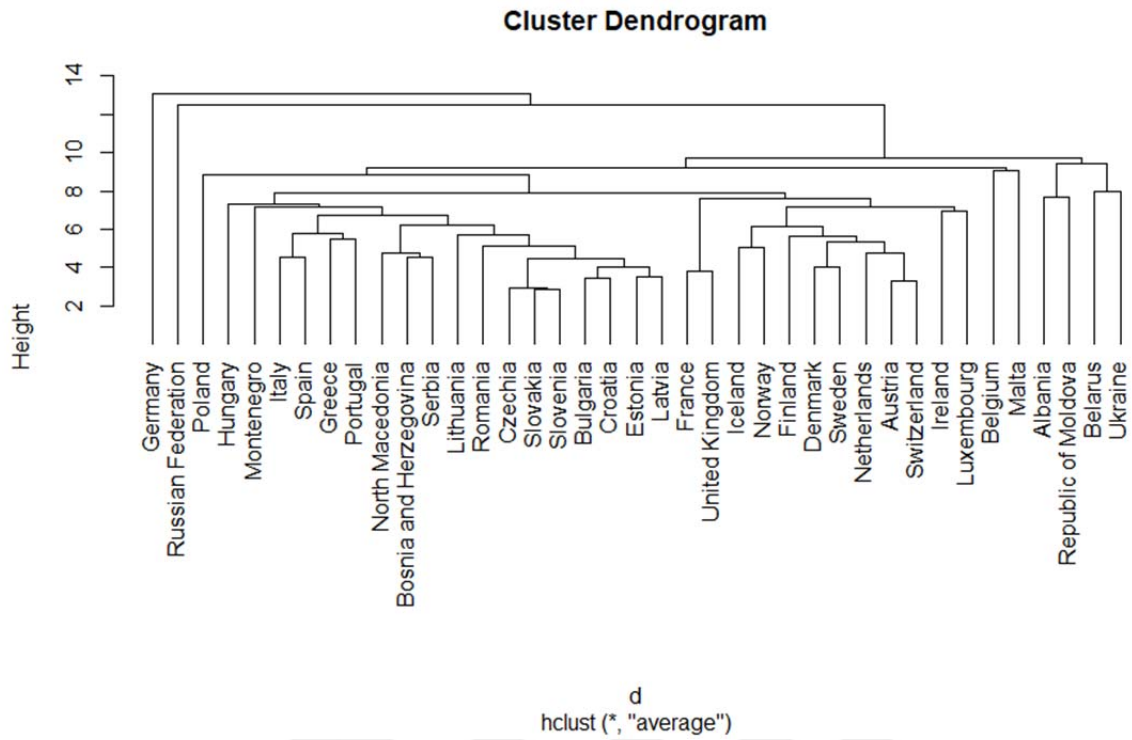


Figure A17. Average linkage clustering results for the Europe countries.

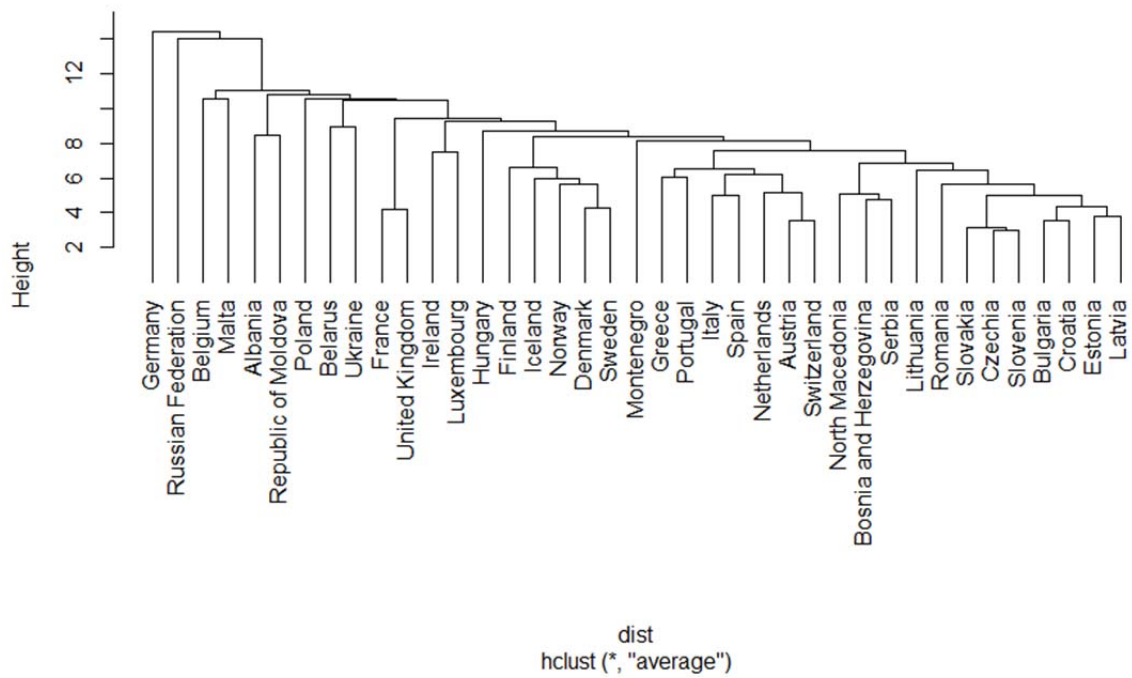


Figure A18. Average linkage clustering results after applying PCA for the Europe countries.

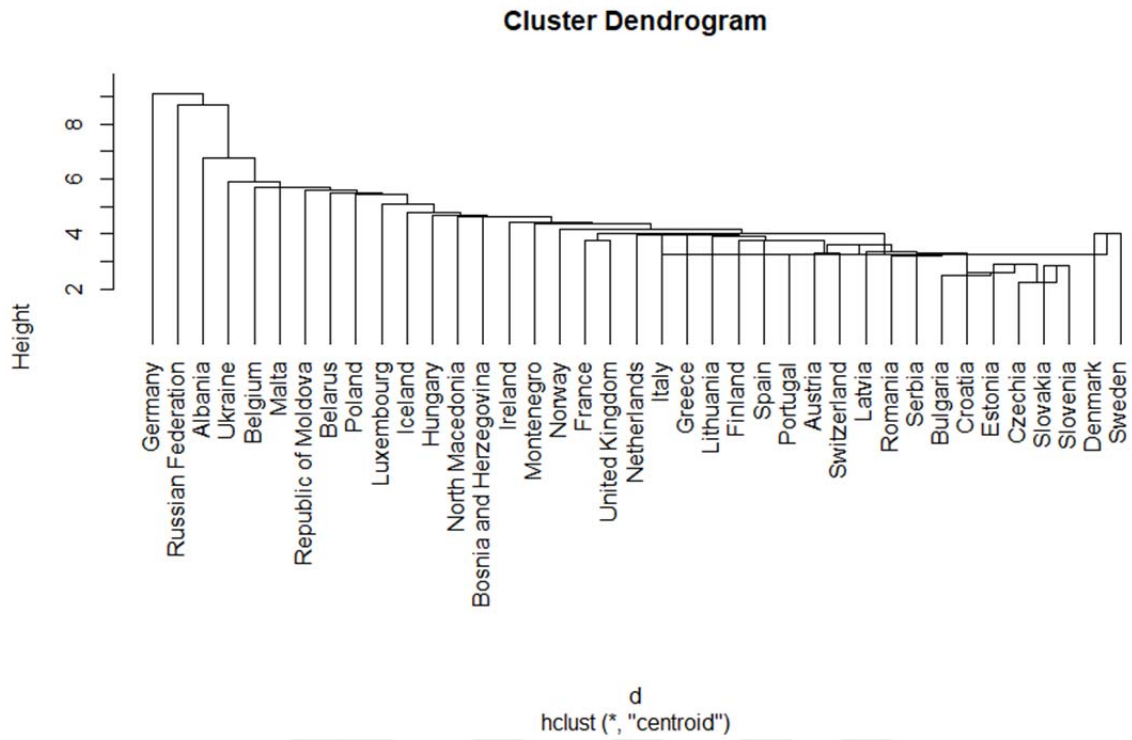


Figure A19. Centroid linkage clustering results for the Europe countries.

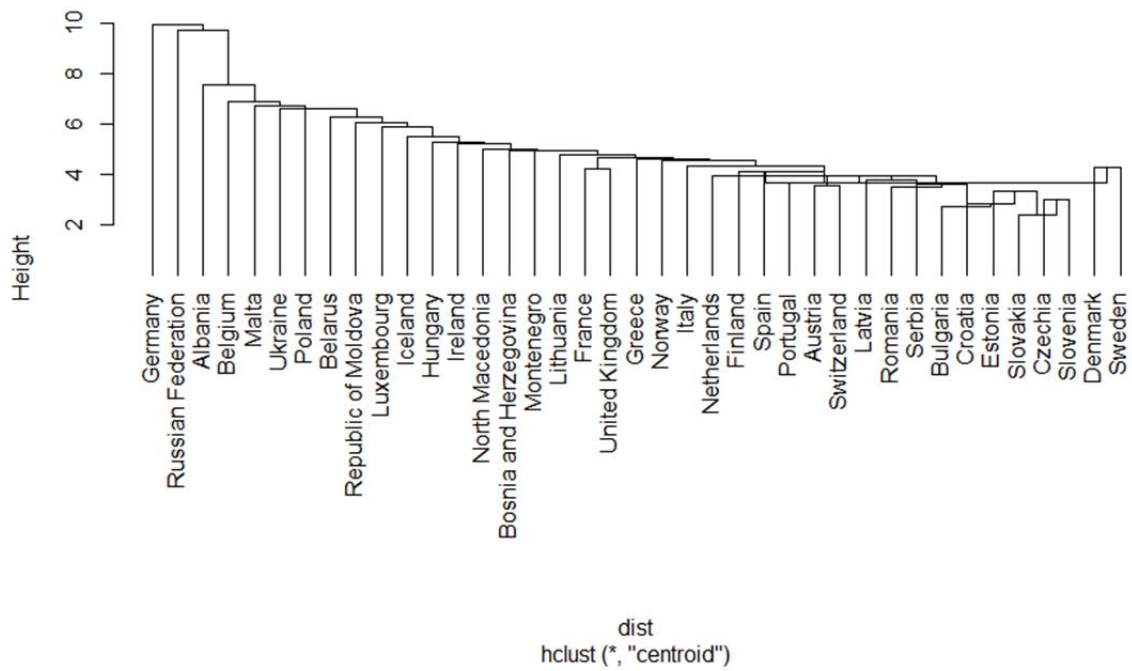


Figure A20. Centroid linkage clustering results after applying PCA for the Europe countries.

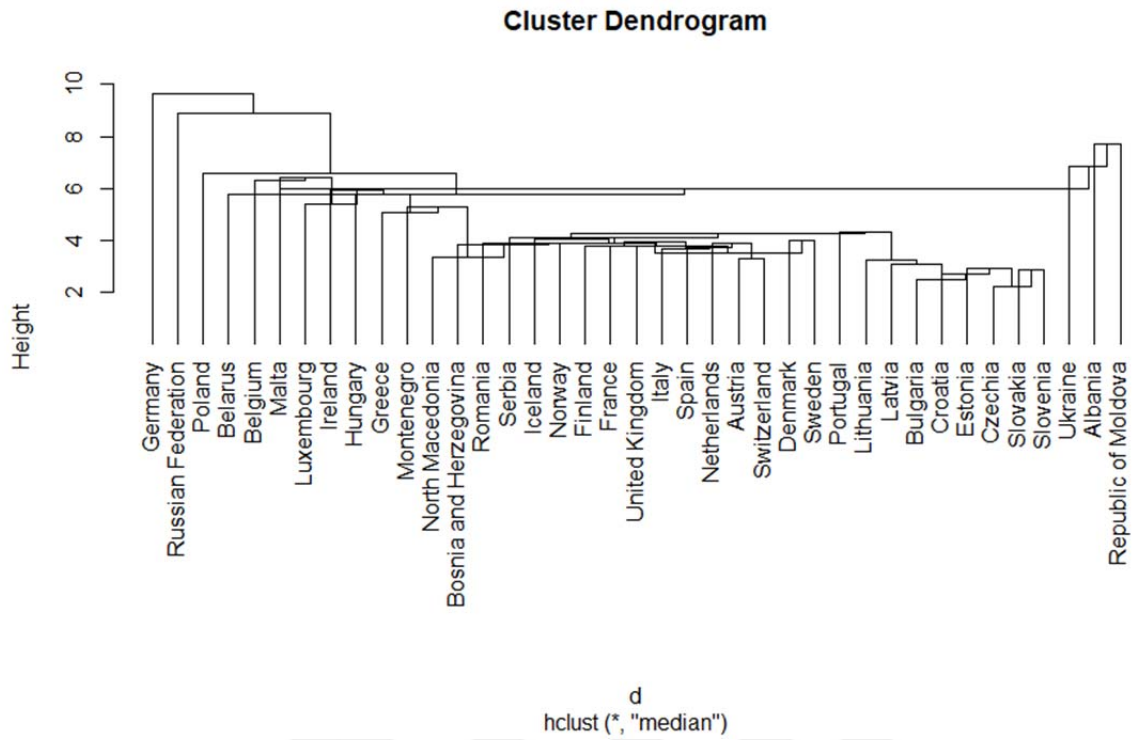


Figure A21. Median linkage clustering results for the Europe countries.

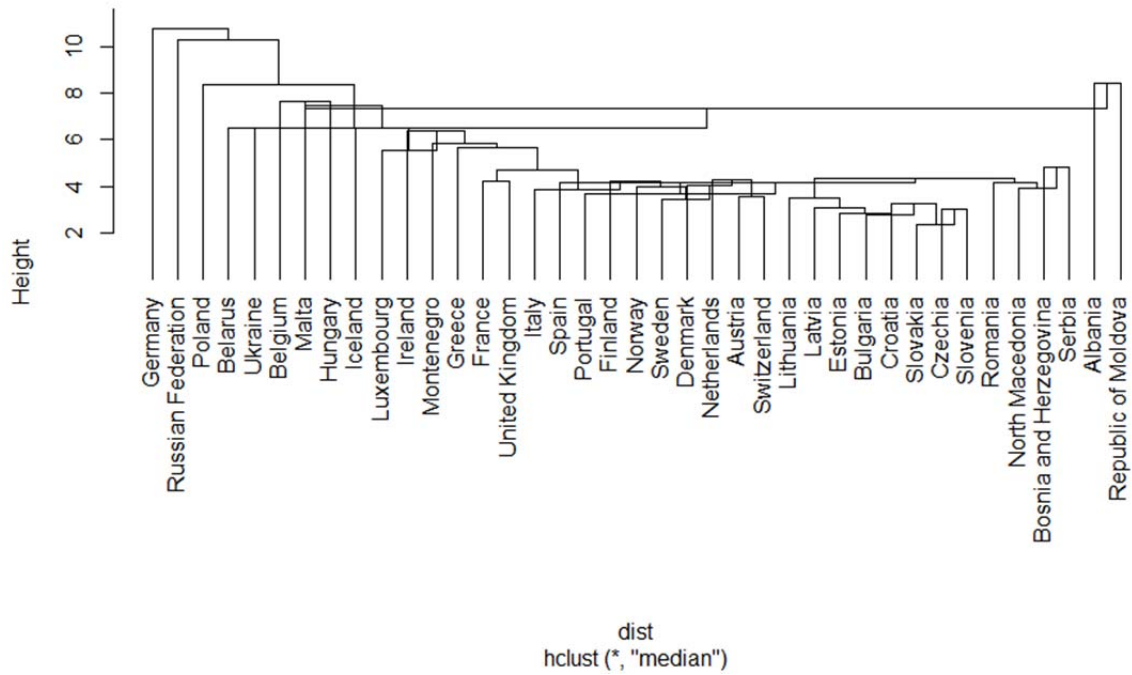


Figure A22. Median linkage clustering results after applying PCA for the Europe countries.

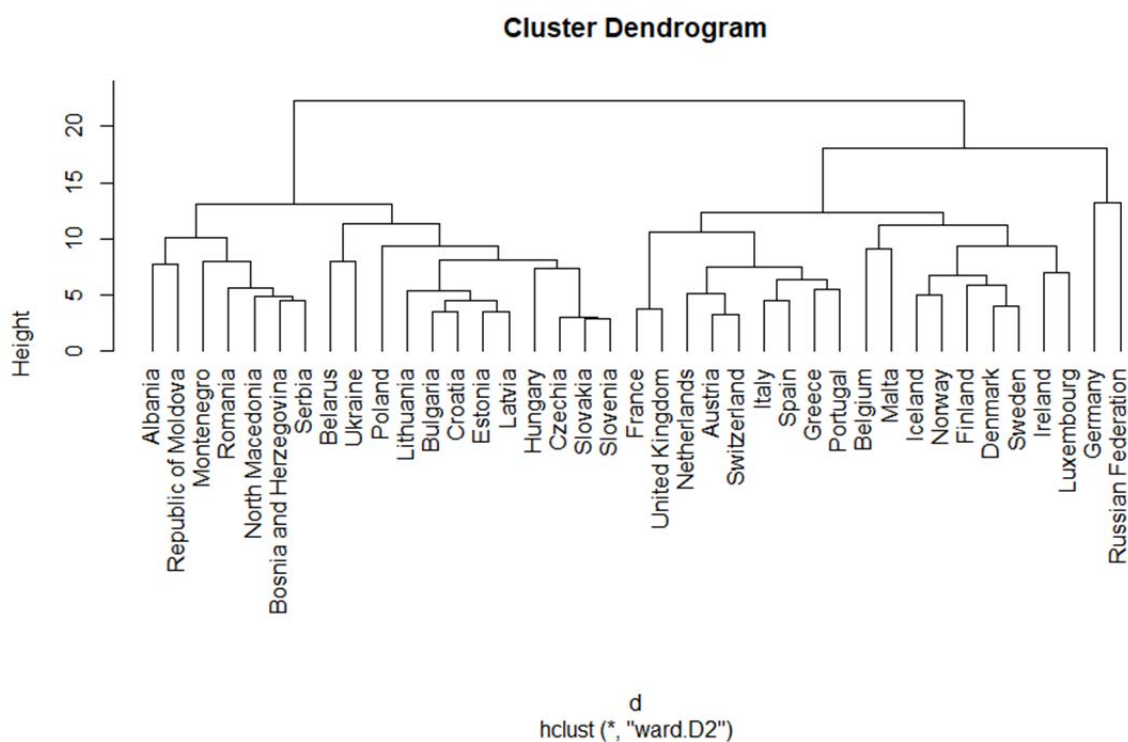


Figure A23. Ward's Criterion linkage clustering results for the Europe countries.

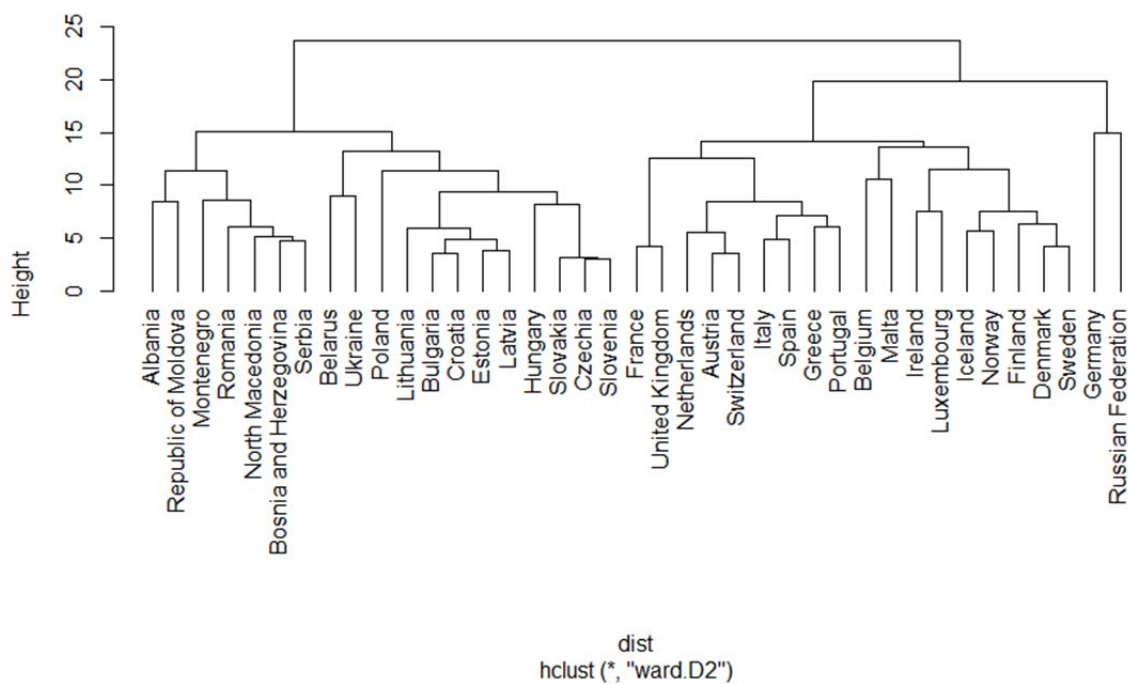


Figure A24. Ward's Criterion linkage clustering results after applying PCA for the Europe countries.

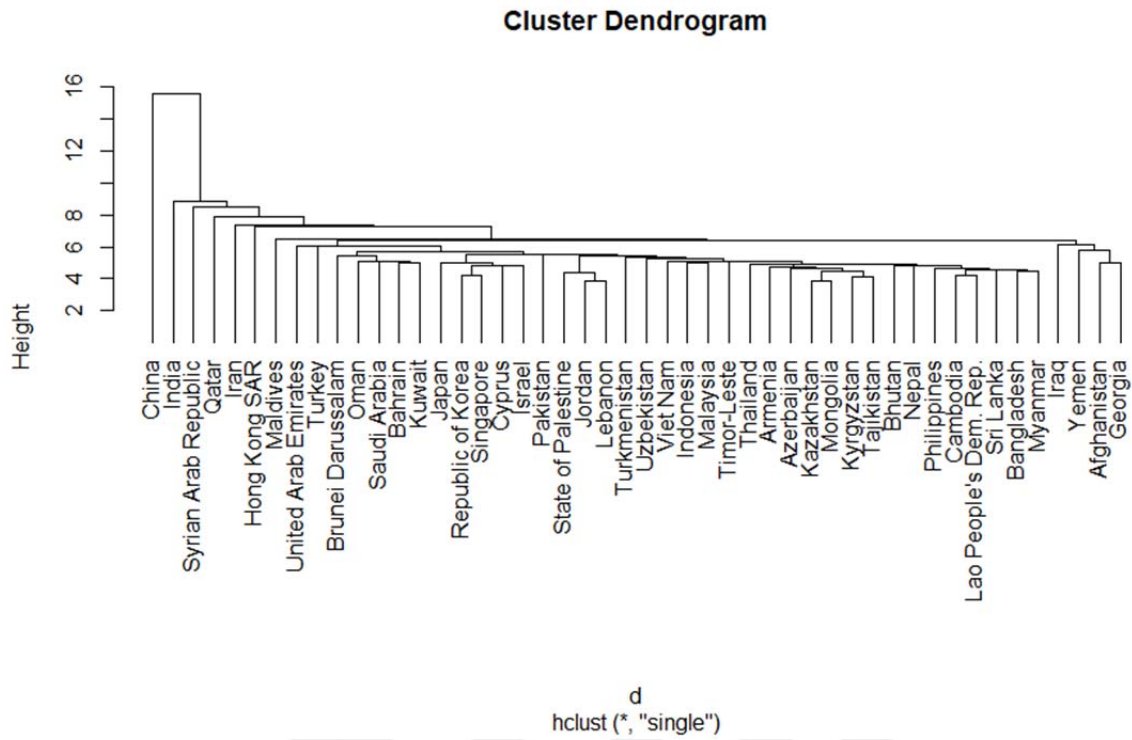


Figure A25. Single linkage clustering results for the Asian countries.

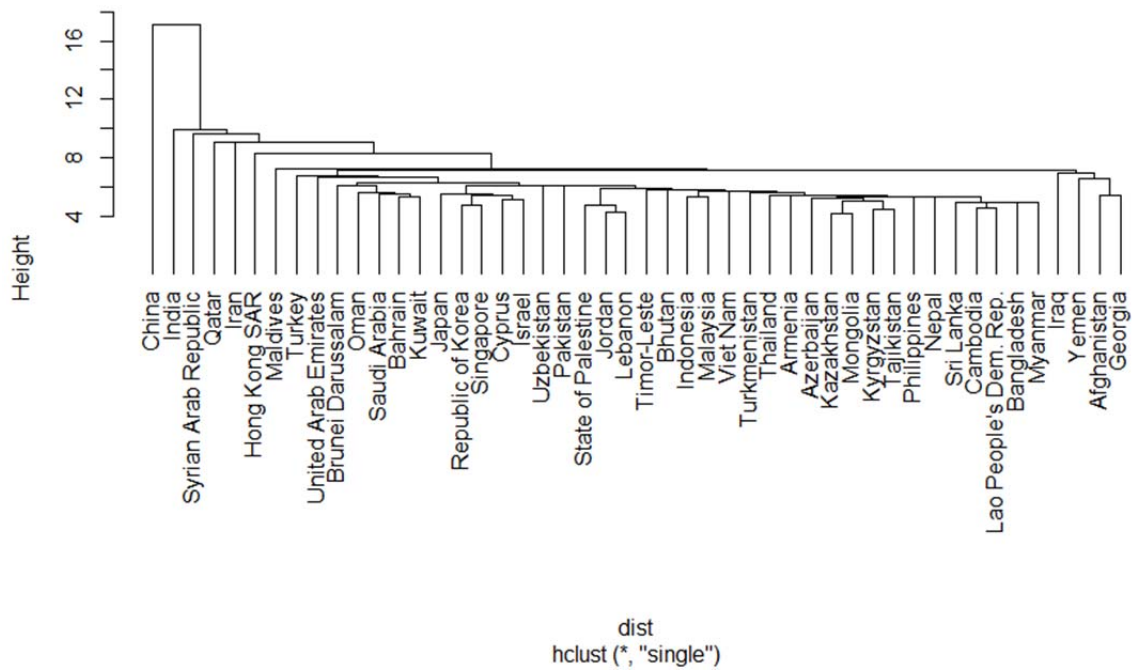


Figure A26. Single linkage clustering results after applying PCA for the Asian countries.

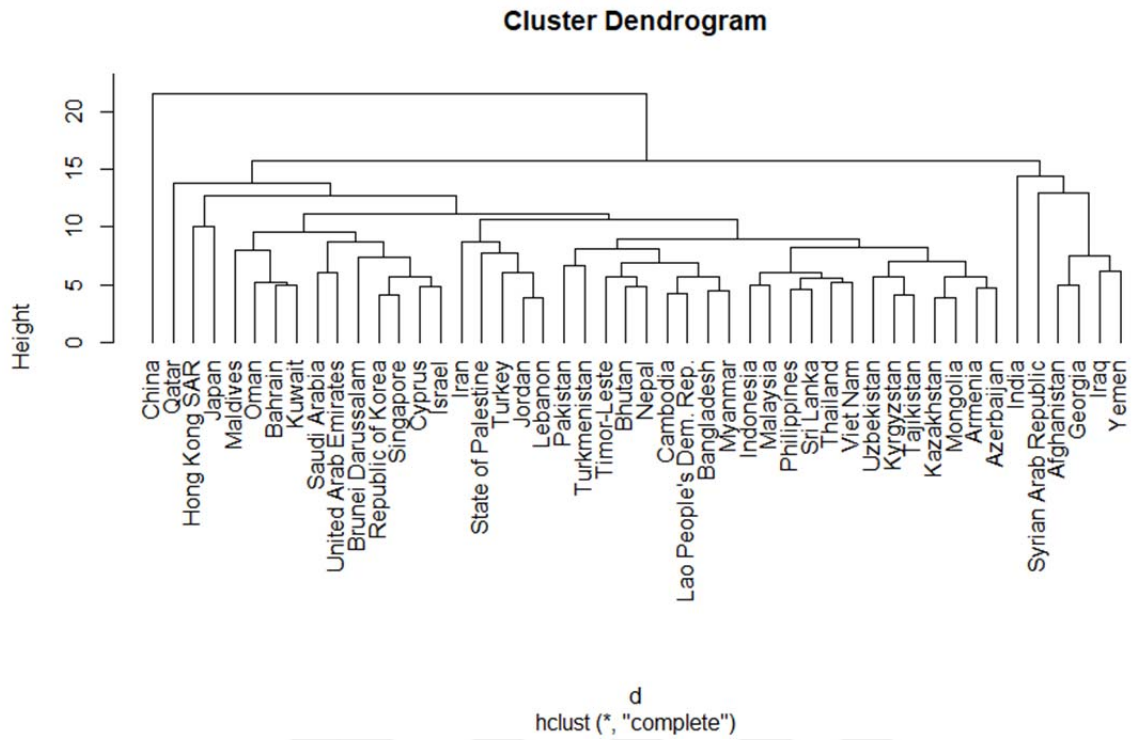


Figure A27. Complete linkage clustering results for the Asian countries.

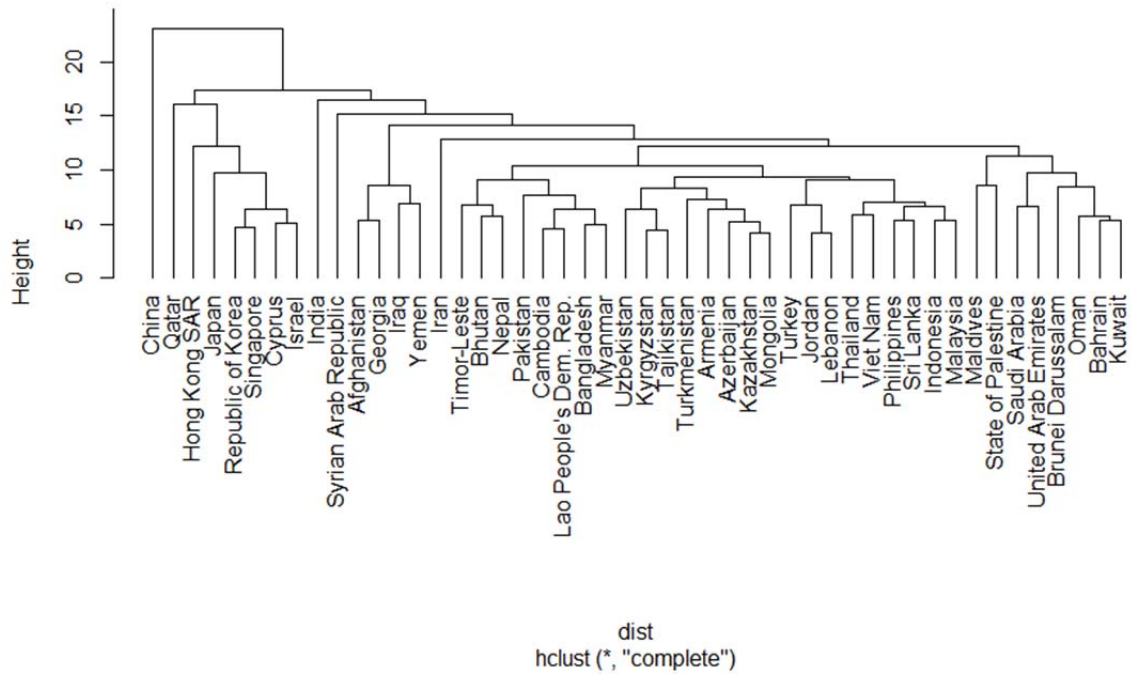


Figure A28. Complete linkage clustering results after applying PCA for the Asian countries.

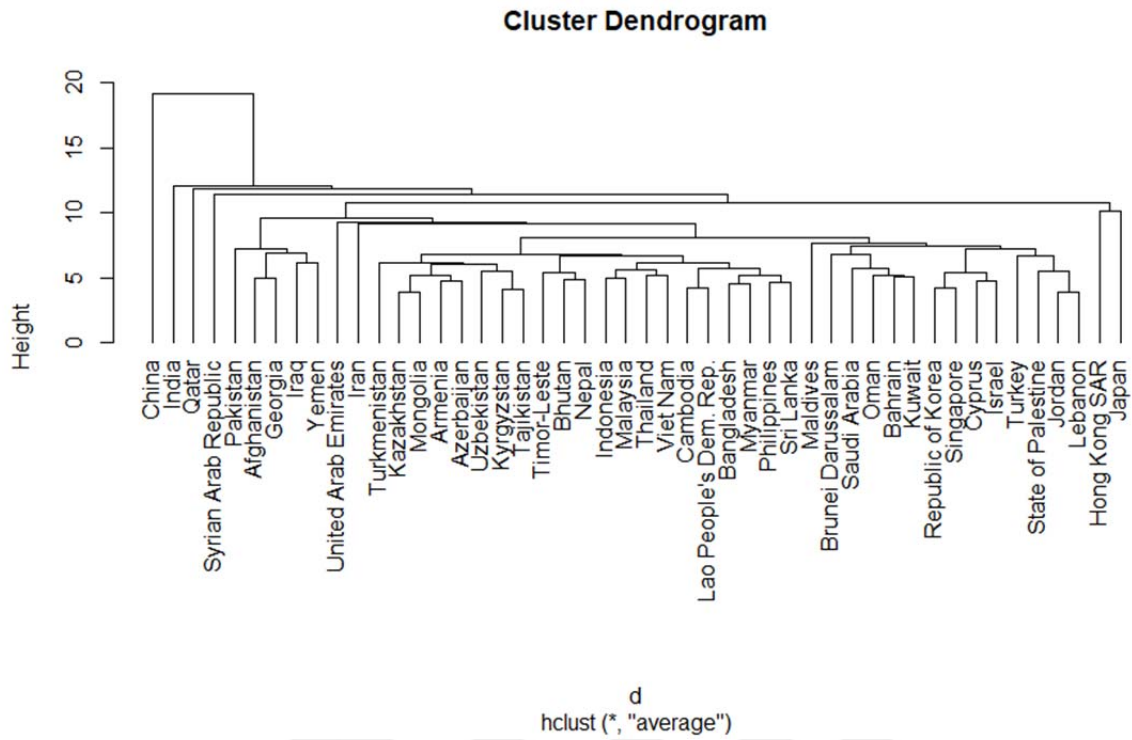


Figure A29. Average linkage clustering results for the Asian countries.

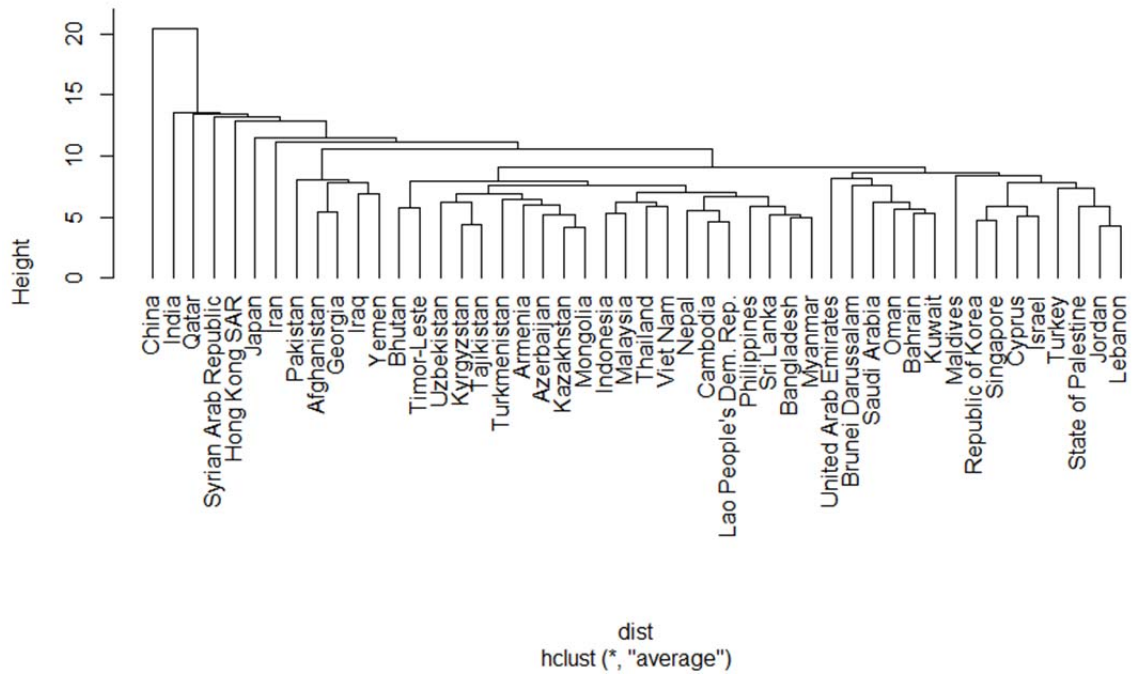


Figure A30. Average linkage clustering results after applying PCA for the Asian countries.

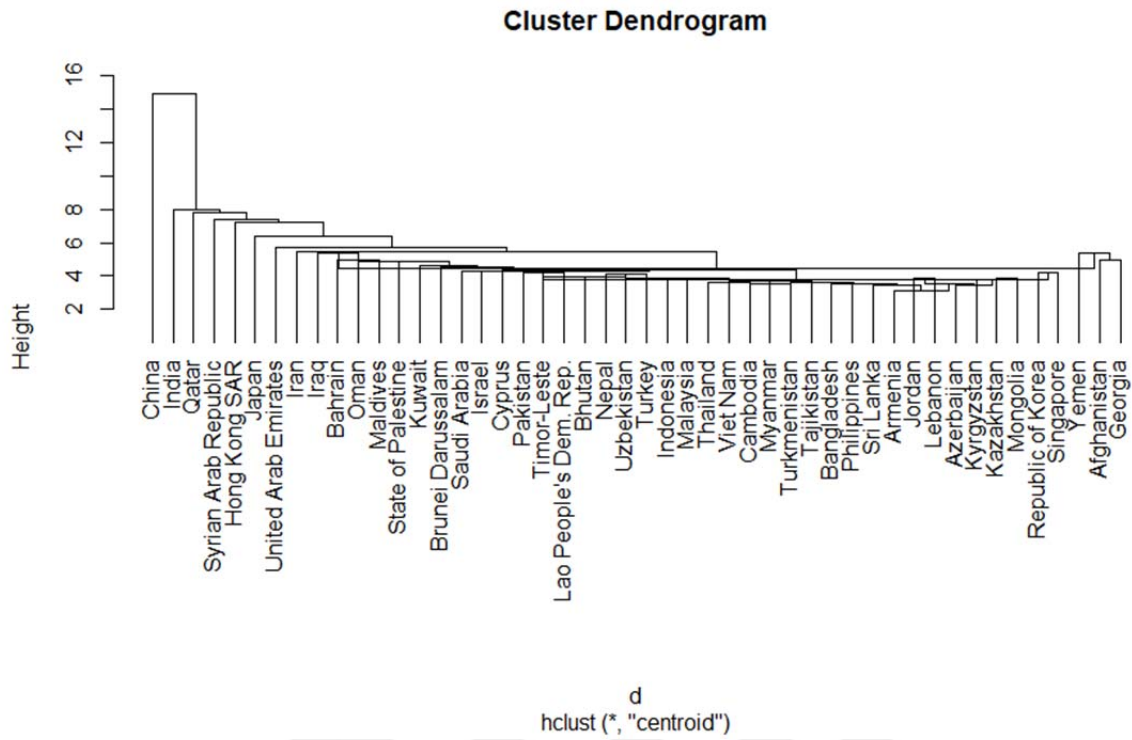


Figure A31. Centroid linkage clustering results for the Asian countries.

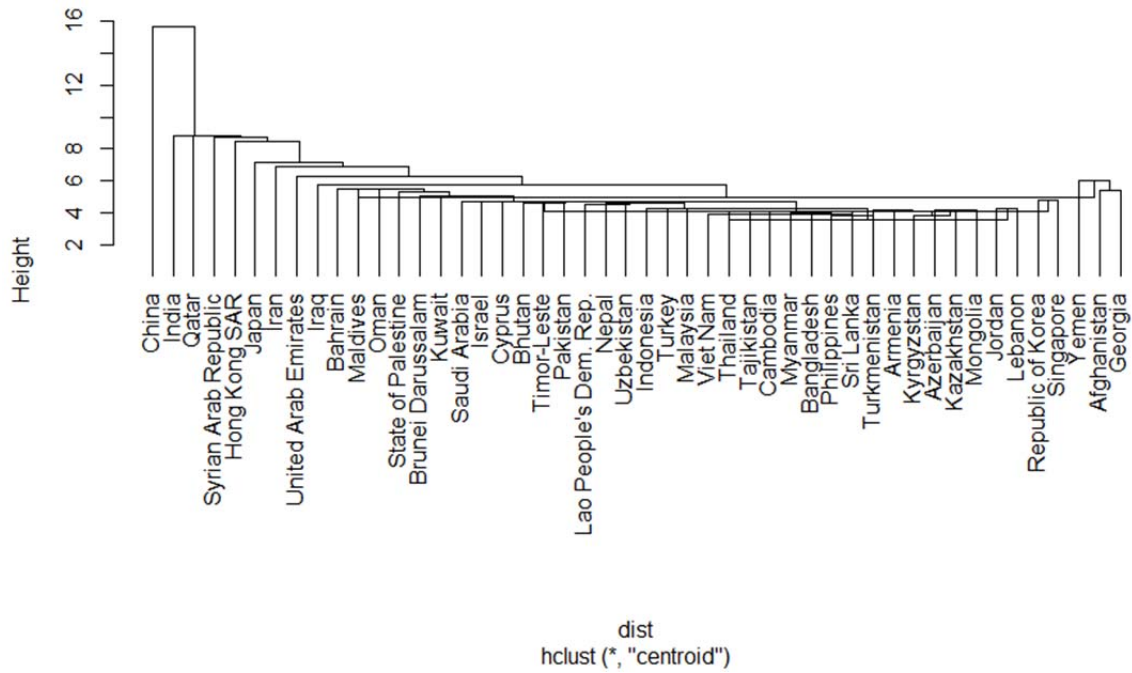


Figure A32. Centroid linkage clustering results after applying PCA for the Asian countries.

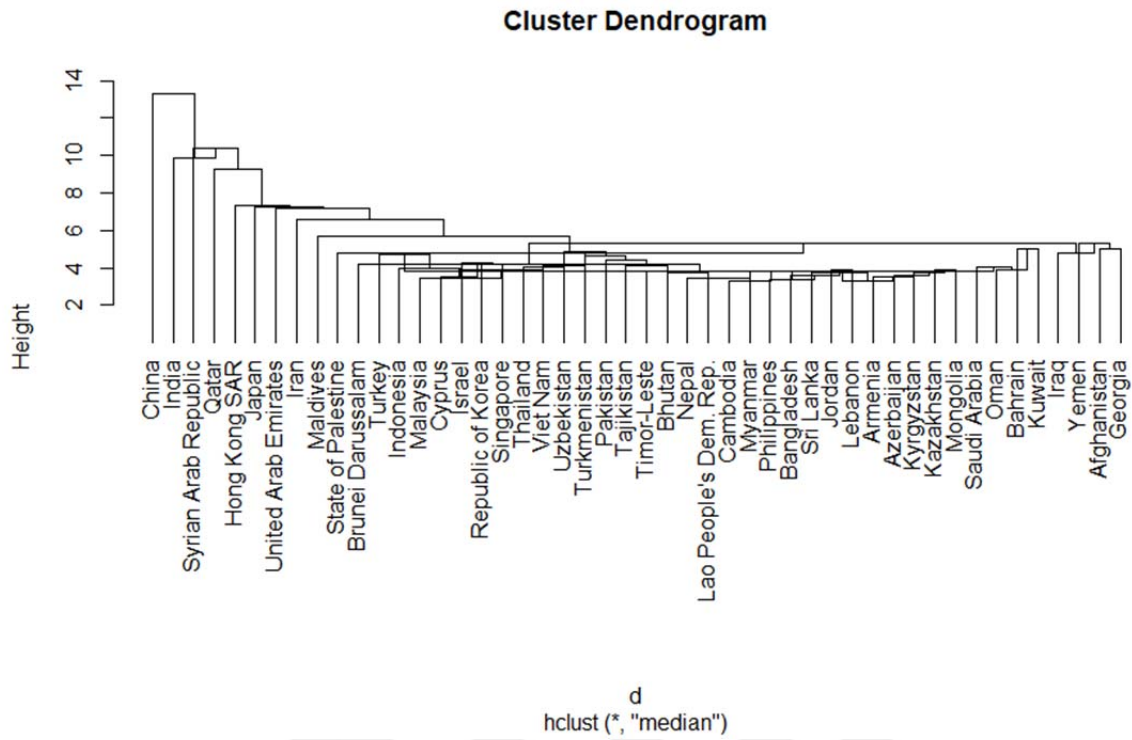


Figure A33. Median linkage clustering results for the Asian countries.

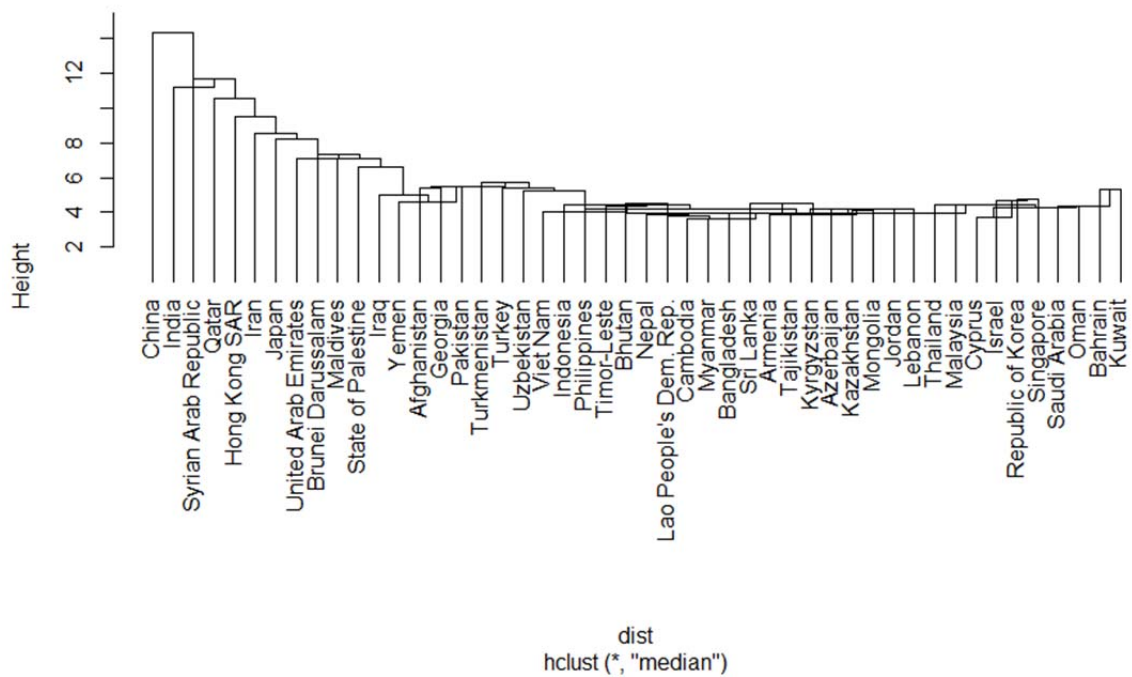


Figure A34. Median linkage clustering results after applying PCA for the Asian countries.

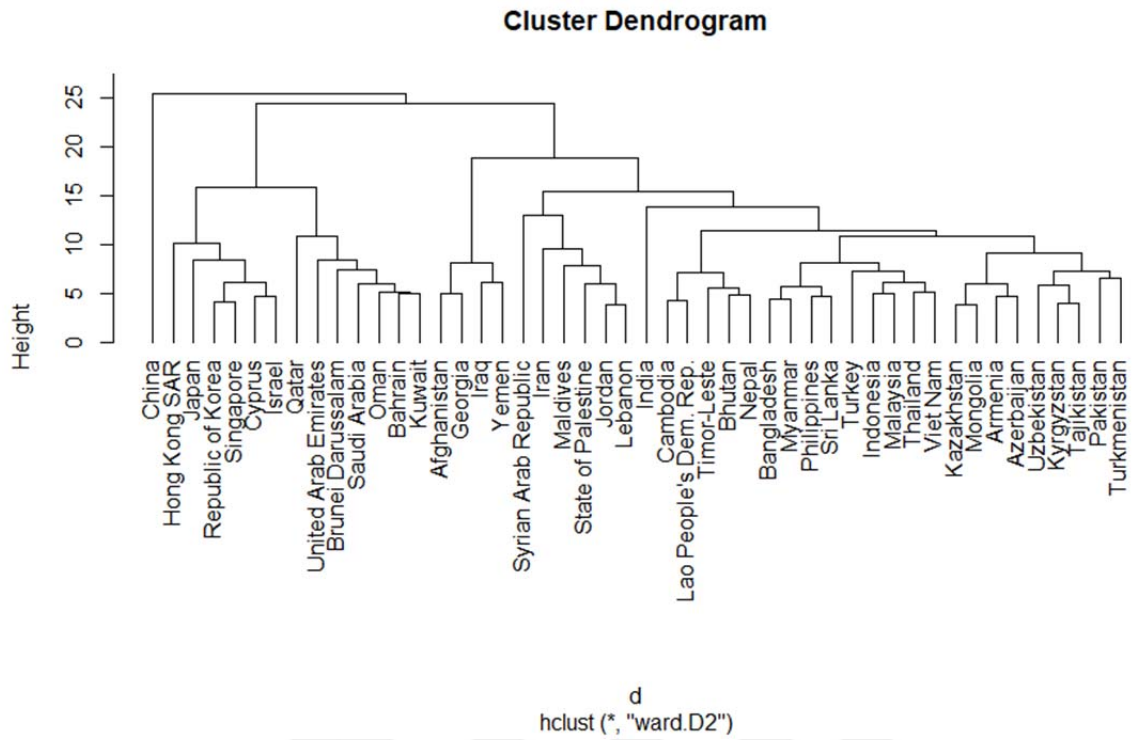


Figure A35. Ward's Criterion linkage clustering results for the Asian countries.

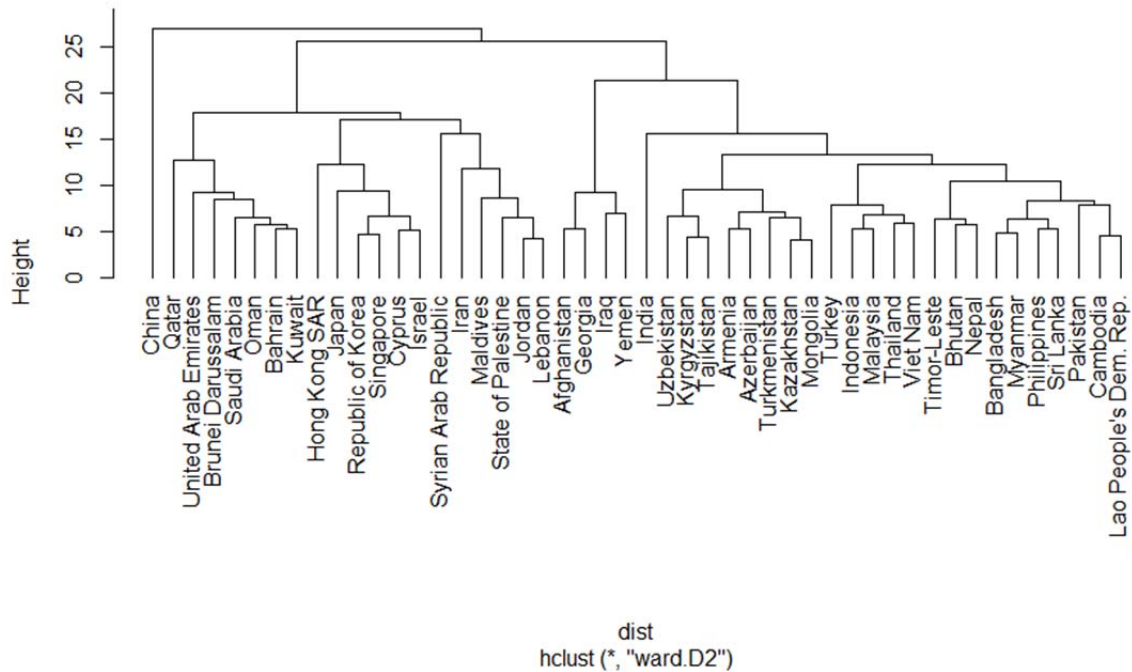


Figure A36. Ward's Criterion linkage clustering results after applying PCA for the Asian countries.