Sigma J Eng Nat Sci, Vol. 43, No. 5, pp. 1607–1627, October, 2025

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.00154

Research Article

Examining dimensionality reduction effect of principal component analysis via hierarchical clustering techniques

Yağmur ŞAN[1,*], Tolunay GÖÇKEN[2]

[1]*Department of Industrial Engineering, Faculty of Engineering, Erciyes University, Kayseri, 38039, Türkiye*
[2]*Department of Industrial Engineering, Faculty of Engineering, Adana Alparslan Türkeş Science and Technology University, Adana, 01250, Türkiye*

ARTICLE INFO

Article history
Received: 27 June 2024
Revised: 09 August 2024
Accepted: 03 October 2024

Keywords:
Baker's Gamma Correlation Coefficient; Cophenetic Correlation Coefficient Dimension Reduction; F_M Index; Hierarchical Cluster Analysis, Principal Component Analysis

ABSTRACT

This study purposes examine the effect of the Principal Component Analysis method on Hierarchical Clustering techniques in terms of dimension 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 variables of the dataset1 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 Africa countries. The variables of the dataset2 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 Europe countries. The variables of the dataset3 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 Asia countries. For dataset1, the mean absolute correlation value is 0.2426, and dimension reduction with PCA is decreased 22 variables to 8 variables. For dataset2, the mean absolute correlation value is 0.2346, and dimension reduction with PCA is decreased 38 variables to 10 variables. For dataset3 the mean absolute correlation value is 0.2265 and dimension reduction with PCA is decreased 46 variables to 11 variables. 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. In this study, despite the outlier and noise problems of high-dimensional datasets, the facilitating role of PCA in clustering analysis is investigated.

Cite this article as: Şan Y, Göçken T. Examining dimensionality reduction effect of Principal component analysis via hierarchical clustering techniques. Sigma J Eng Nat Sci 2025;43(5):1607–1627.

***Corresponding author.**
*E-mail address: yagmursan01@gmail.com

This paper was recommended for publication in revised form by Editor-in-Chief Ahmet Selim Dalkilic

Published by Yıldız Technical University Press, İstanbul, Turkey
© Author. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

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. 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 [1]. This vast amount and variety of data consist of bits of information that are meaningless on their own. The scientific discipline that includes the processes and methods of transforming data into information is data mining. As a general definition," Data mining is the process of discovering interesting patterns and knowledge from large amounts of data." [2]. 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. 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 uses three different approaches to determine the number of components and it is possible to obtain different results with these approaches.

PCA and clustering have been used in some agricultural studies to assess species variation or product quality [3-7]. Some studies used clustering and PCA methods to classify and evaluate data from different sectors [8-11,12]. Carried out customer segmentation using PCA, hierarchical clustering, and k-means clustering in his study. On the other hand, [13] compared the dimension reduction effect of PCA and non-negative matrix factorization methods on clustering. In another study that proposes a new method for graph-based dimensionality reduction, a hybrid method that is a combination of NPE and PCA linear dimensionality reduction methods is presented. The presented hybrid method produces a transformation matrix for the generalized eigenvalue problem. According to the analysis results of the study, the hybrid method showed the best performance among PCA, NPE and the presented hybrid method (HDR) [14]. In another study conducted by the same authors, according to the results obtained through empirical analysis using graphic data sets, in linear methods, the principal component analysis, singular value decomposition, and neighborhood preserving embedding methods have been showed better performance than other methods of the statistical information category, dictionary methods, and embedding methods, respectively [15]. In a study, the effectiveness of PCA dimension reduction and SVM classification techniques was examined for anomaly detection over network data. In the study, the positive results of the dimensionality reduction effect of PCA on classification quality and processing time are shown [16]. In another study, PCA and LTP methods and the BAT algorithm were used to reduce the difficulties and processing time in face recognition systems. The dimensionality reduction effects of PCA and LTP were comparatively examined by integrating them with the feature selection function of the Bat algorithm [17]. In another study, the prediction performance of machine learning methods was investigated using multidimensional data obtained in cancer cases. In the research, the effect of PCA and Kernel PCA's dimensionality reduction function on the prediction performance was evaluated comparatively [18]. In the study, which focuses on improving automatic intrusion detection with the aim of minimum fault and correct classification, the system performance was evaluated with the dimensionality reduction effect of random projection and PCA techniques. The effect of the two methods on the results and their accuracy rates were compared [19]. In a study conducted in the field of fluid mechanics, the performances of linear and nonlinear dimension reduction techniques were compared. The methods examined in the study; PCA, Independent Component Analysis, Isometric Mapping and Local Linear Embedding dimensionality reduction techniques. The performance of each method and their suitability according to the characteristics of the flow fields were evaluated separately [20]. PCA is a very effective technique that can be used for many different purposes such as clustering, data reduction and dimensionality reduction. We encounter the use of PCA in different ways in the literature, but the dimensionality reduction ability of PCA has never been examined in detail before. In this study, the dimensionality reduction ability of PCA is investigated and demonstrated in detail with graphics and numerical analysis using high-dimensional data sets. Despite the outlier and noise problems of high-dimensional datasets, the facilitating role of PCA in clustering analysis is searched.

The rest of the study; in section 2, hierarchical clustering methods and their properties are examined in detail. In section 3, PCA and the approaches used to determine the number of components are handled. In section 4, some coefficients used for comparing hierarchical clustering results are given with formulas. In section 5, the datasets and the methodology used in the study are explained in detail. Section 6 consists of the evaluations of the results for each dataset separately. In the conclusion section, the outputs of the study are interpreted in a general framework.

HIERARCHICAL CLUSTERING

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 [2] Hierarchical methods reveal natural clusters in the dataset without the need to specify the number of cluster. This is known as one of the important advantages of hierarchical methods [21]. One of the weaknesses of the hierarchical methods is that they have high computational complexity. The time complexity of hierarchical methods is expressed by $O(n^2)$ and they are not suitable for large datasets [22]. In hierarchical methods, the tree-shaped diagram that presents the clustering results of the data in a hierarchical structure is called a dendrogram [23]. 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 [24]. In hierarchical clustering, the cutting point of the dendrogram has a significant role in determining the clusters. A different number of various cluster combinations can be obtained by cutting the dendrogram at different points [22]. 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) [25]. 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 [26]. There are many different linkage methods according to different distance preferences such as minimum, maximum, average, median, centroid and minimum variance [27].

Single Linkage

In the single linkage method, also called the ***nearest neighbour***, the minimum distance between two clusters is accepted as the criterion for merging [21,25,28]. In each iteration, the two closest clusters are merged and the operations are repeated until the clustering is complete [29]. This method is faster than other hierarchical techniques but, because of its local behaviour, it is sensitive to noises and outliers [30, 31]. 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 [26,30]. This case is known as the ***chaining effect***.

Complete Linkage

In the complete linkage method, unlike the single linkage method, the merging criterion is the furthest distance between two clusters [27]. Due to its non-local behaviour, smaller and tighter clusters are obtained [33]. However, also in this method, the presence of outliers can have negative effects on clustering because of its sensitivity to noise and outliers [26,30,31].

Average Linkage

In the average linkage method, the merging criterion is the average distances of all the pairs of points of the clusters [21-23]. In each iteration, the procedure is repeated according to this criterion until the clustering process is completed. This method has the advantage of outlier insensitivity and is difficult to use for categorical data, but also a very effective method for numerical data [2].

Centroid Linkage

This method performs the merging operation according to the proximity between the centroids of the clusters [21, 32, 33]. Here centroid denotes the center point of the cluster.

Median Linkage

In the merge of two clusters with different sizes, the median value is used to prevent the centroid of the newly formed cluster from shifting predominantly towards the larger cluster [25, 23]. The median value is the midpoint of the distance between the two centroids. It is designed to overcome the disadvantage of the centroid method.

Ward's Criterion Linkage

The main purpose of Ward's criterion linkage method is to obtain homogeneous clusters by keeping the sum of squares of the error of the distances within the cluster minimum [22,25,34]. Contrary to some methods, cluster size is effective on this method. There are two versions of this method; ward1 and ward2. The main difference between these two versions is; Squared Euclidean distance is used in ward1, while Euclidean distance is used in ward2 [35,36]. Significant fomulas used for Hierarchical clustering are shown in Table 1 and Table 2.

Table 1. The time complexity functions of the hierarchical methods for dataset with n objects

Divisive	$O(2^n)$	***Average linkage***	$O(n^2 \log n)$
Agglomerative	$O(n^2)$	***Centroid linkage***	$O(n^2 \log n)$
Single linkage	$O(n^2)$	***Median linkage***	$O(n^2 \log n)$
Complete linkage	$O(n^2 \log n)$	***Ward's linkage***	$O(n^2)$

Principal Component Analysis

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 is defined by Hotelling in 1933 [37]. 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. These new orthogonal and uncorrelated variables are the linear combinations of the original variables in the data set. PCA is a fast running and computationally easy method. The areas and applications where PCA is used are; pattern recognition, dimension reduction, computer vision, image compression, signal processing, video surveillance, face recognition,

Table 2. The merging criteria formulas of the hierarchical linkage methods

Single linkage	$\boldsymbol{D(A,B) = min\{d(y_i, y_j),\quad for\ y_i\ \ in\ A\ \ and\ y_j\ \ in\ B\}}$
Complete linkage	$D(A,B) = max\{d(y_i, y_j),\quad for\ y_i\ \ in\ A\ \ and\ y_j\ \ in\ B\}$
Average linkage	$D(A,B) = \frac{1}{n_A n_B}\sum_{i=1}^{n_A}\sum_{j=1}^{n_B} d(y_i, y_j)$
Centroid linkage	$D(A,B) = d(\overline{y_A}, \overline{y_B})$
Median linkage	$M_{AB} = \frac{1}{2}(\overline{y_A} + \overline{y_B})$
Ward's linkage	$SSE_A = \sum_{i=1}^{n_A}(y_i - \bar{y}_A)'(y_i - \bar{y}_A),\ \ SSE_B = \sum_{i=1}^{n_B}(y_i - \bar{y}_B)'(y_i - \bar{y}_B),$ $SSE_{AB} = \sum_{i=1}^{n_{AB}}(y_i - \bar{y}_{AB})'(y_i - \bar{y}_{AB});\ \ \boldsymbol{I_{AB} = SSE_{AB} - (SSE_A + SSE_B)}$

latent semantic indexing, ranking, and collaborative filtering [38,39]. 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, ……, y_n are the observation vectors in p dimensinoal space, and $\bar{y}$ is the mean vector of the observation vectors. y_1, y_2, ……, y_p are the swarm of points for the variables. The first operation in PCA is the transformation. In the transformation operation, the origin 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, ……, 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.

$$A = \begin{pmatrix} a'_1 \\ \vdots \\ a'_p \end{pmatrix};\quad z_i = A * y_i,\quad i = 1, \ldots\ldots, p$$

$z_1 = a'_1 y,\ z_2 = a'_2 y,\ \ldots\ldots\ldots,\ z_p = a'_p y \rightarrow$ Principal Components

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

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

$\vdots$

$z_P = a_{P1}y_1 + a_{P2}y_2 + \ldots + a_{Pp}y_p \rightarrow$ 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, ……, y_n (S).

$$S_z = ASA' = \begin{pmatrix} S^2_{z_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & S^2_{z_p} \end{pmatrix}$$

Here $S^2_{z_i} = \lambda_i$, and λ_i, $i = 1, \ldots, p$ are the eigenvalues of S which equals to variance values of the principal components. In PCA, the variance values of the components have an arrangement as $\lambda_1 > \lambda_2 > \ldots > \lambda_p$. 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. 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 [40].
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 [25, 37].
4. 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 [25].

THE INDEXES FOR COMPARING HIERARCHICAL CLUSTERING

CPCC (Cophenetic Correlation Coefficient)

CPCC is expressed as the correlation value between the cophenetic matrix created according to the height values in

dendrogram and similarity matrix [41-43]. CPCC takes the value in the range of [-1,1]. 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. CPCC aims to measure how well hierarchical clustering is performed [44-46].

FM-Index (Fowlkes and Mallows Index)

FM_index (Fowlkes and Mallows index) measures the similarity between the clustering results [43]. 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] [47]. The values close to 1 express the high similarity and the values close to 0 express the low similarity between the clustering results.

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 [48]. In generally γ coefficient is used for ordinal variables and it takes the values between the range of [-1,1] [48, 49]. Here is the values close to 0 express the low relationship and the values close to -1 and 1 express the negative and pozitive relationship respectively. The formulations of the indexes are given in Table 3.

MATERIALS AND METHODS

In this study, 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. The most important factor in choosing hierarchical clustering methods in this study is that natural groups in the dataset spontaneously emerge with the application of the technique. That is, there is no need to determine the number of clusters with this technique. Clustering operations are performed both with PCA and without PCA.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, to observe the change in dendrograms, ***tanglegrams*** are used for each method. Tanglegrams are plots that show two hierarchical clustering or two dendrograms comparatively. The data sets used in the analysis are compiled from the United Nations data platform [50]. The missing observation values in the data sets are completed from the relevant international data platforms [51, 52, 53, 54]. 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 in the appendices in the relevant tables. All the methods and stages used in the analysis are applied to each dataset in the same way. 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.

Datasets

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) as Pop. density (per km2, 2020), sex ratio(male/female) and surface area variables and the economic indicator (EI) as growth rates, sectoral employment rates and international trade data variables of the countries and the objects of the data set consist of Africa countries. To evaluate the correlation level

Table 3. The indexes for comparing hierarchical clustering methods

CPCC (Cophenetic Correlation Coefficient)	$\frac{\sum_{i<j}(d_{ij}-\bar{d})(d^*_{ij}-\bar{d}^*)}{\sqrt{\sum_{i<j}(d_{ij}-\bar{d})^2 \sum_{i<j}(d^*_{ij}-\bar{d}^*)^2}}$	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
FM-INDEX (Fowlkes and Mallows Index)	$\sqrt{\frac{TP}{TP+FP} * \frac{TP}{TP+FN}}$	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.
Baker's Gamma Coefficient (Goodman and Kruskal's Gamma Coefficient)	$\gamma = \frac{\pi_c - \pi_d}{\pi_c + \pi_d}$	π_c: The probability of concordant pairs of observations. π_d: The probability of disconcordant pairs of observations.

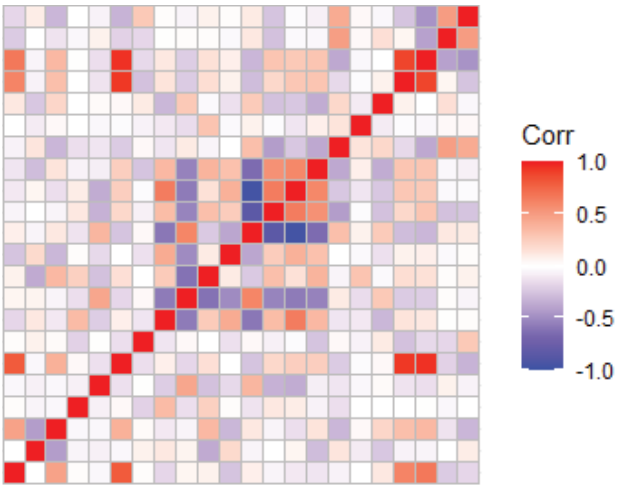

Figure 1. Colored correlation matrix of the variables for dataset1.

between the variables in the data set, the colored correlation matrix for the variables of the dataset1 is given below in Figure 1.

The mean absolute correlation between variables for the dataset1 is 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. Below 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 in Figure 2.

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. 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 in Figure 3.

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) as population growth rate, international migration stock, education and health data variables of the countries and the objects of the data set consist of Europe countries.

The mean absolute correlation between variables for the dataset2 is 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 4.

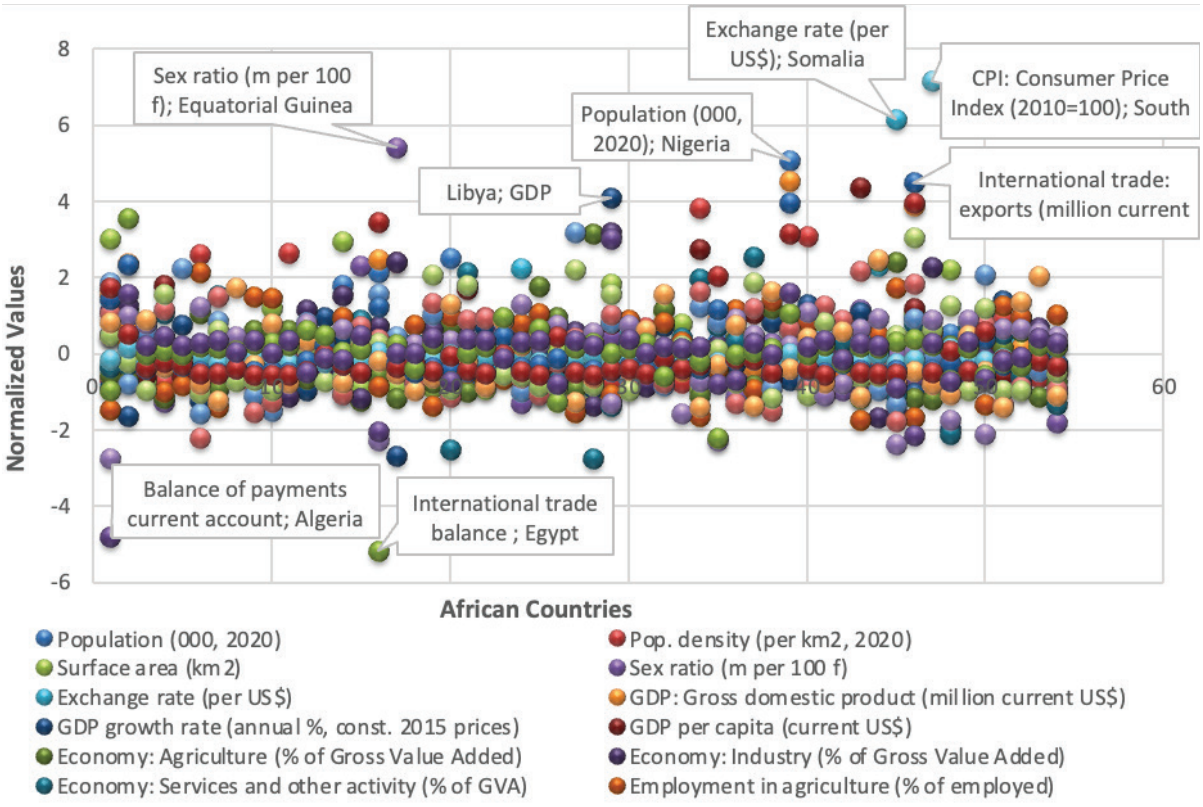

Figure 2. Scatterplot of the normalized dataset1 for the economic indicator variables.

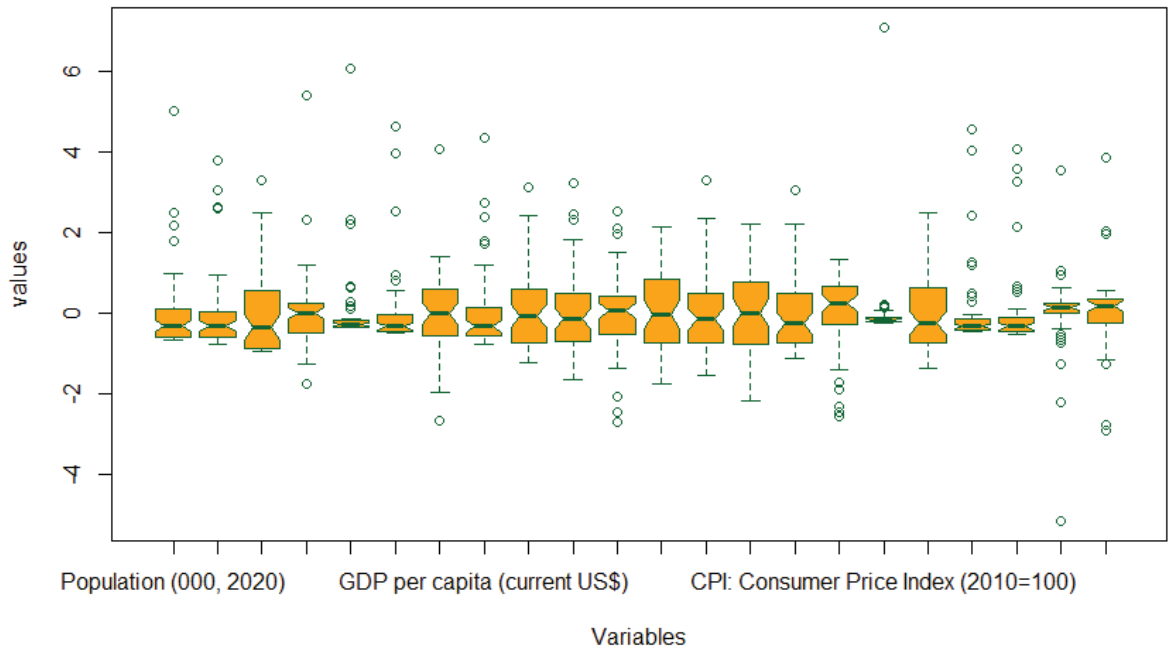

Figure 3. Boxplots for dataset1.

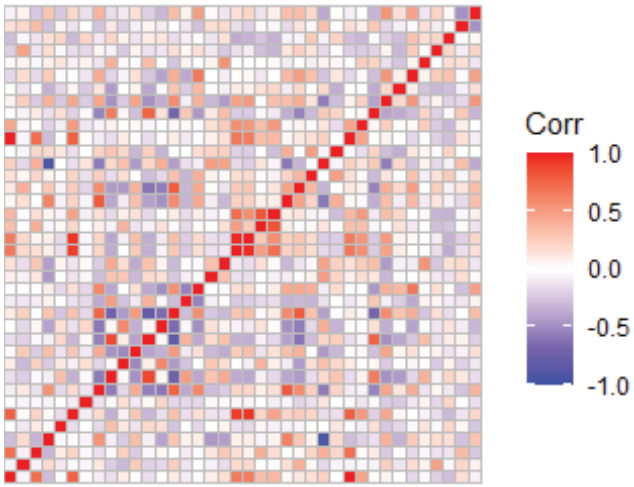

Figure 4. Colored correlation matrix of the variables for dataset2.

Figure 5 is a scatterplot showing how Europe countries are distributed according to the economic indicator and the social indicator variables. It seems to be so many outliers here as in the first dataset. The existence of many outliers can be seen clearly in the boxplots for dataset2 in Figure 6. 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.

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) as individual internet use, CO2 emission estimates and energy production amounts variables of the countries and the objects of the data set consist of Asia countries. Colored correlation matrix for Dataset3 is shown in Figure 7 and the scatterplot is shown in Figure 8.

Figure 9 is a scatterplot showing how Asia 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. Object 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.

RESULTS AND DISCUSSION

In this section, 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.

Results for Dataset 1

In principal component analysis, three approaches can be used to decide the number of components. For the elbow

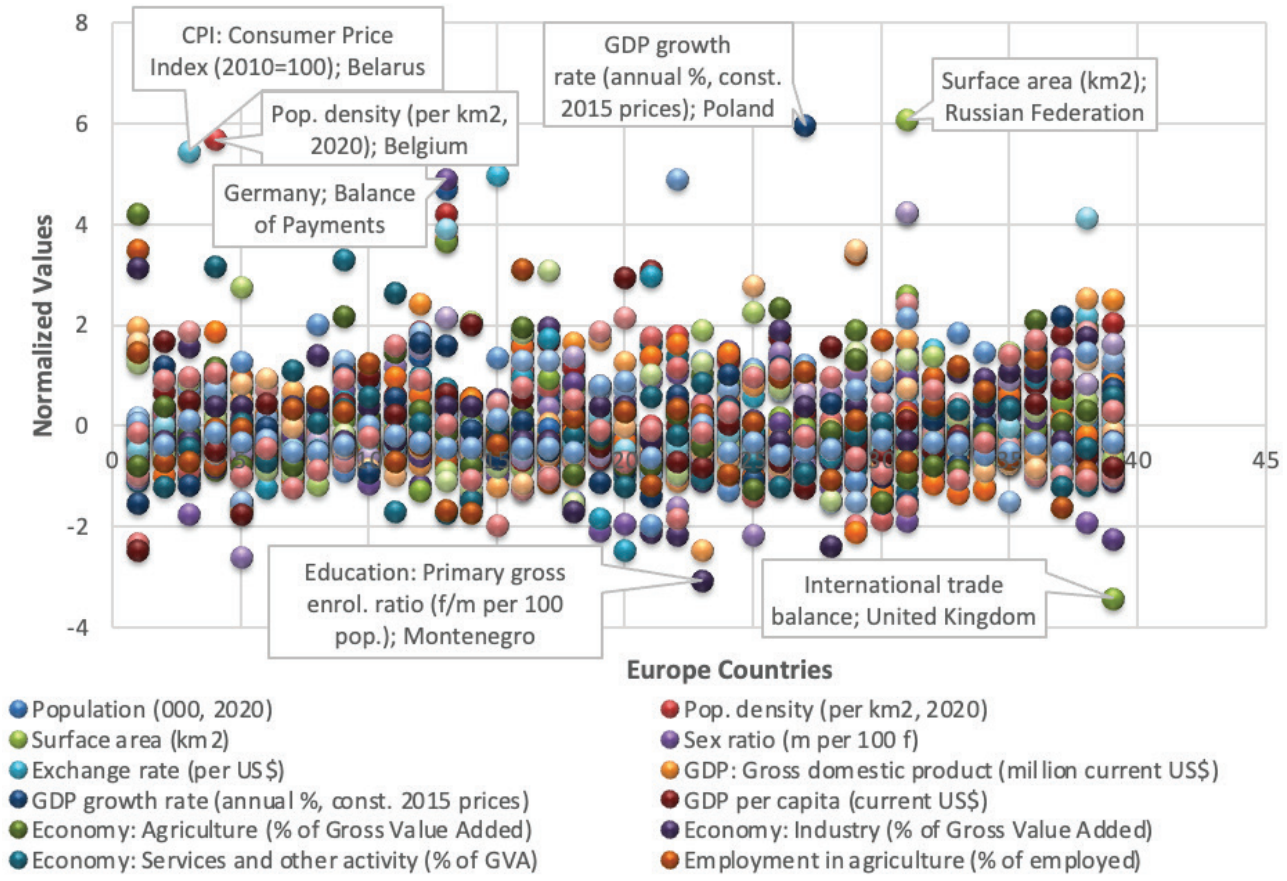

Figure 5. Scatterplot of the dataset2 for the economic indicator and social indicator variables.

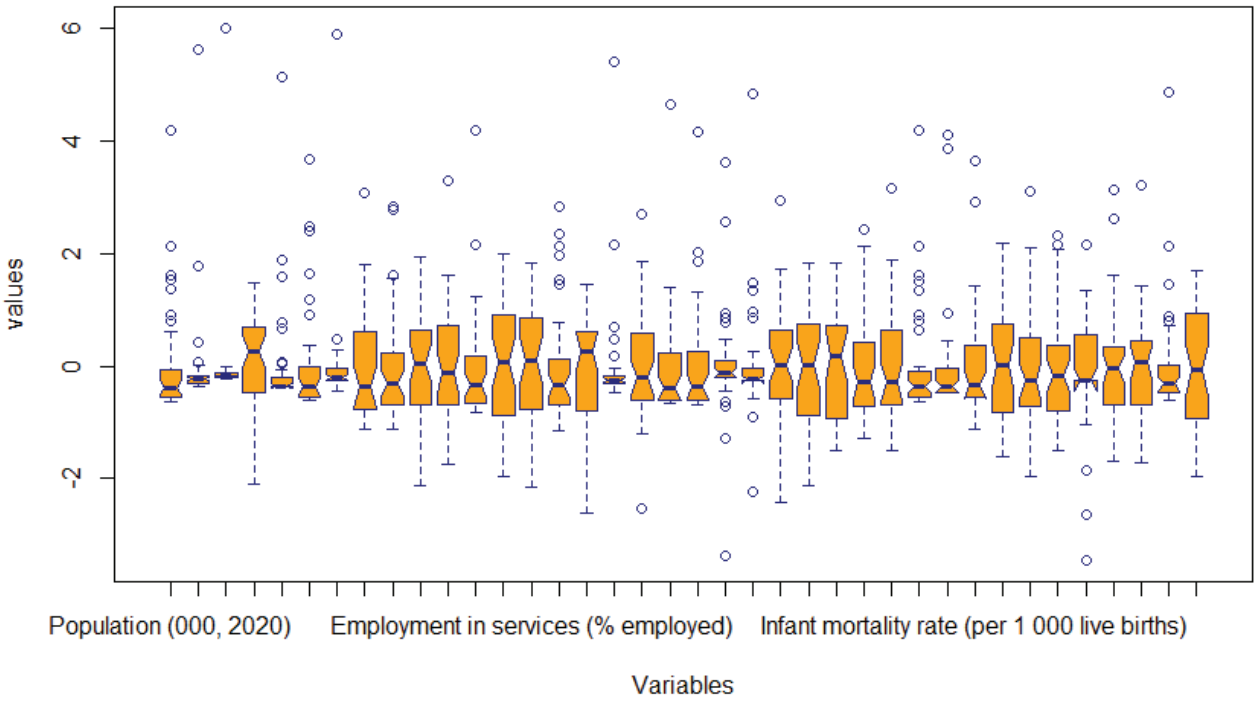

Figure 6. Boxplots for dataset2.

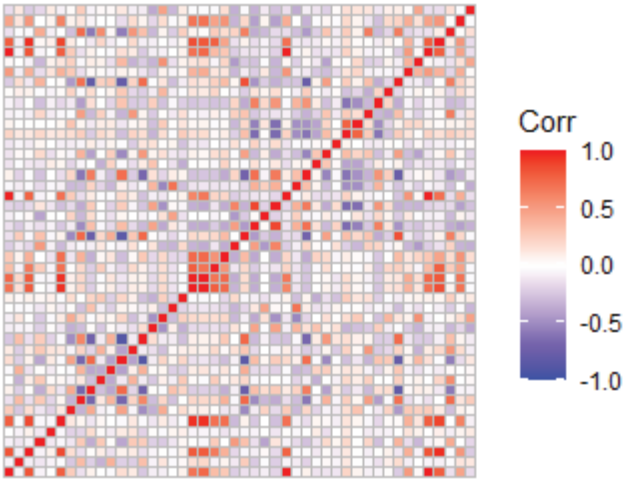

Figure 7. Colored correlation matrix of the variables for dataset3.

test, which is one of these approaches, when the scree plot which includes the variance values of the principal components given in Figure 10 is examined, it is seen that the natural breakpoint in the graph coincides with the 3rd principal component.

According toTable 4 and Table 5, 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 dataset1, the mean-variance value ($\bar{\lambda}$) of the principal components is calculated as 0.9815 and the variance values of the first 7 principal components are above this value. The first 7 components can explain %76.35 of the total variance in the data set, as seen in table 4.

The third approach is to select components whose sum of variance explained percentages exceeds the 80% specific ratio. The first 8 principal components should be selected to achieve this specific ratio because the first 8 components can explain %80.81 of the total variance in the data set. As a result, the number of components for a robust analysis process is determined as 8.

The CPCC (dend1) values show the fit and the similarity ratio of the cophenetic matrix and the distance matrix of the dendrogram obtained without the PCA method. Also, the CPCC (dend2) values show the fit and the similarity ratio of the cophenetic matrix and the distance matrix of the dendrogram obtained after applying the PCA method. According to the CPCC values in the Table 6, it is obvious that the results of PCA applied clustering analysis have higher similarity with the distance matrix in almost all methods. For CPCC(dend1) and CPCC(dend2), while the best clustering results belong to PCA+average and PCA+centroid methods, the worst clustering belongs to the variance-based Ward's method due to the outliers and low correlation in the dataset. Unlike the CPCC (dend1 and dend2), the CPCC (tanglegram) shows the fit and the similarity ratio between the cophenetic matrices of these two dendrograms. Consistent with the tanglegrams, the highest similarity rate belongs to average linkage and centroid

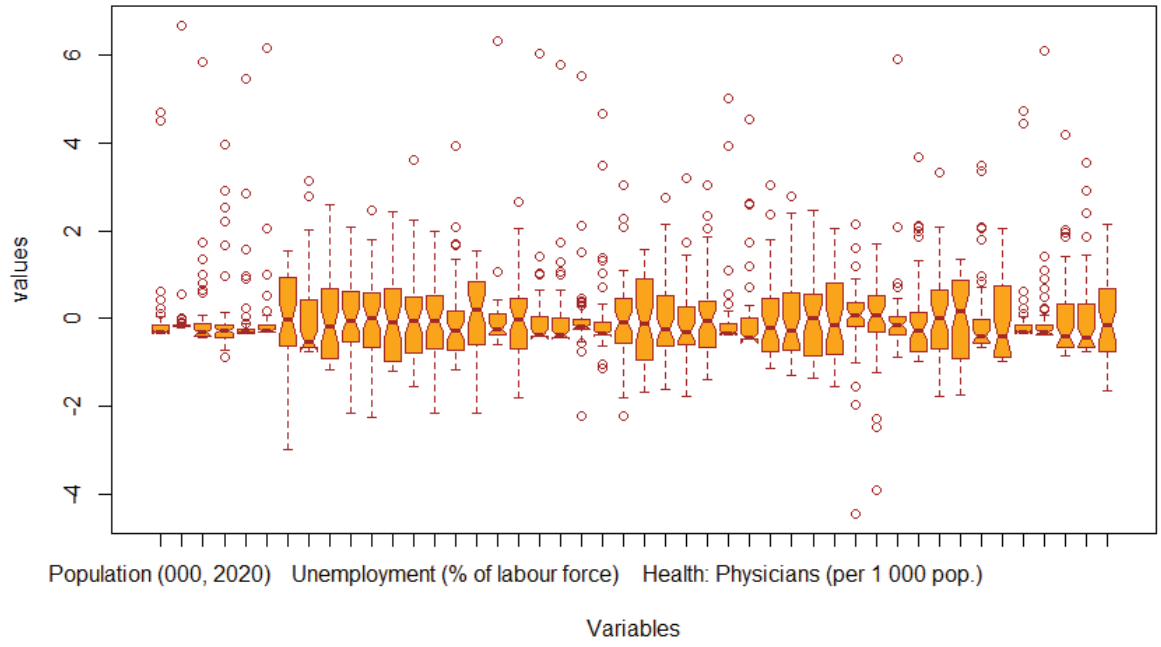

Figure 8. Boxplots for dataset3.

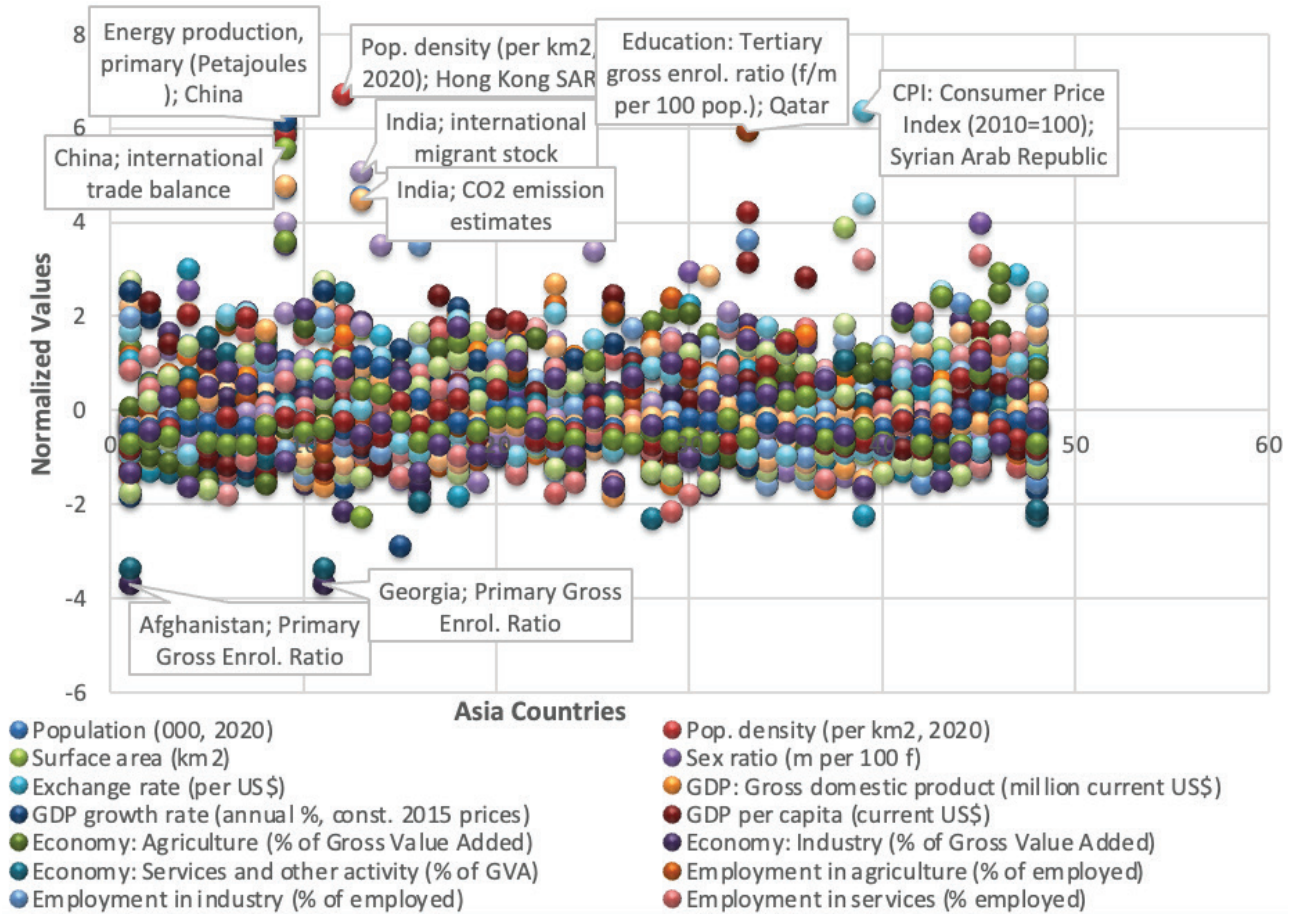

Figure 9. Scatterplot of the dataset3 for the economic indicator, social indicator and environmental and infrastructural variables.

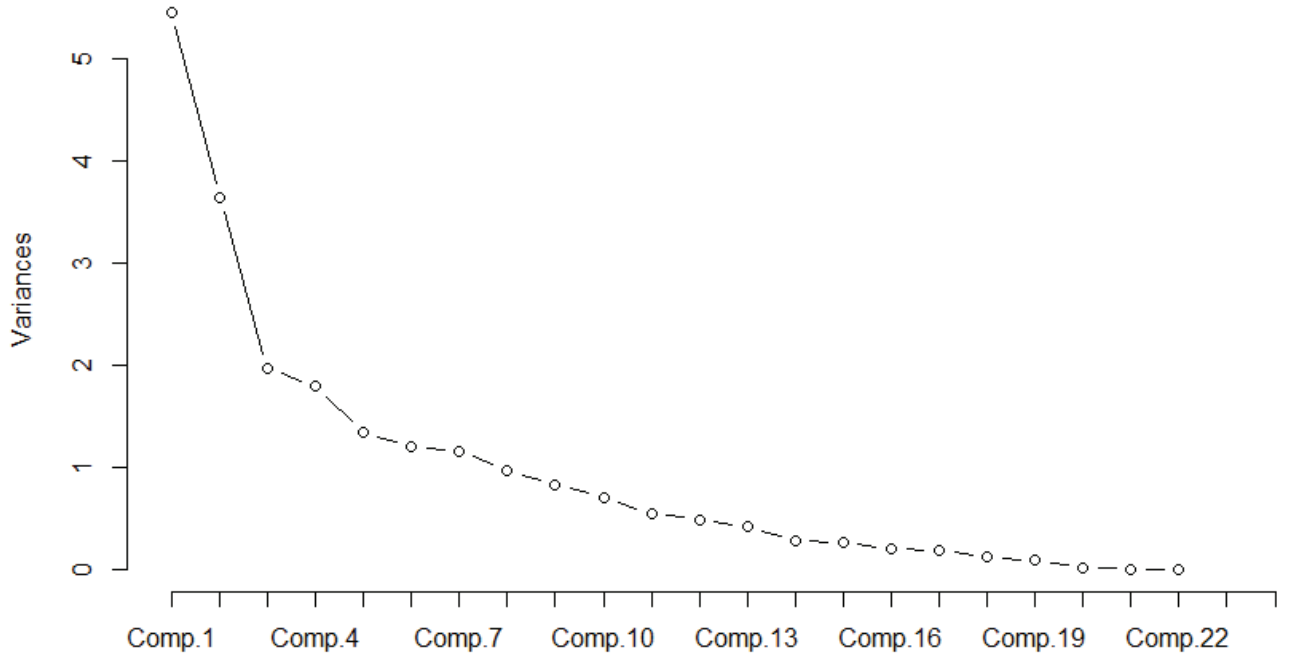

Figure 10. Scree plot of the variance values of the principal components for dataset1.

Table 4. The variance values of the principal components for dataset1

comp.1	comp.2	comp.3	comp.4	comp.5	comp.6	comp.7	comp.8	comp.9
5,44E+0	3,64E+0	1,97E+0	1,79E+00	1,34E+00	1,20E+00	1,15E+00	9,63E-01	8,28E-01

Table 5. The variance explained percentages of the principal components for dataset1

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+0	8,28E+0	6,22E+0	5,55E+0	5,34E+00	4,46E+00	3,83E+00

Table 6. Coefficient and index results used to measure the similarity of dendrograms for dataset1

	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

linkage methods. BGCC (Baker's gamma correlation coefficient) measures the similarity of dendrograms using the odds ratios of concordant and disconcordant pairs of objects. However, when evaluated together with the tanglegrams in Figure 11, it is seen that this coefficient gives inconsistent results for the centroid and median methods. The FM_index determines the similarity ratio of the dendrograms by comparing the cluster contents. Unlike the other coefficients, FM_index uses the number of clusters as a parameter in the calculations. In this study, the number of clusters parameter is determined as 3 for all data sets. If the cluster contents of the two compared dendrograms are the same, the FM_index value will be equal to 1. By evaluating the graphics and all the calculated coefficients together, it is observed that the best clustering result belonged to the PCA+Centroid method, and the cluster contents for the 3 clusters are given below.

Cluster 1: Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cabo Verde, Cameroon, Central African Republic, Chad, Comoros, Congo, Cote d'lvoire, Dem. Rep. Of the Congo, Djbouti, 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 Cluster 2: Nigeria Cluster 3: South Africa

Results for Dataset 2

The variance values of the principal components for dataset2 are given in Figure 12 and Table 7. For the elbow test, the natural breakpoint in the graph coincides with the 3rd principal component. But the first 3 components could explain only %46.44 of the total variance in the dataset according to Table 8.

For the second approach, the mean-variance value ($\bar{\lambda}$) of the principal components is calculated as 0.9743 for dataset2. It is observed that the variance values of the first 10 principal components are above this value, and can explain %80.50 of the total variance in the data set.

For the 80% specific ratio, the first 10 principal components should be selected to achieve this specific ratio, and can explain %80.50 of the total variance in the data set.

According to the CPCC (dend1 and dend2) values in Table 9, it is obvious that the results of PCA applied clustering analysis show higher similarity with the distance matrix in almost all methods. For CPCC(dend1) and CPCC(dend2), while the best clustering results belong to PCA+average and PCA+centroid methods, the worst clustering belongs to the variance-based Ward's method due to the outliers and low correlation in the dataset. Consistent with the tanglegrams, the highest similarity rates for the CPCC (tanglegram) belong to the single linkage, centroid linkage and ward's criterion linkage methods. When evaluated together with

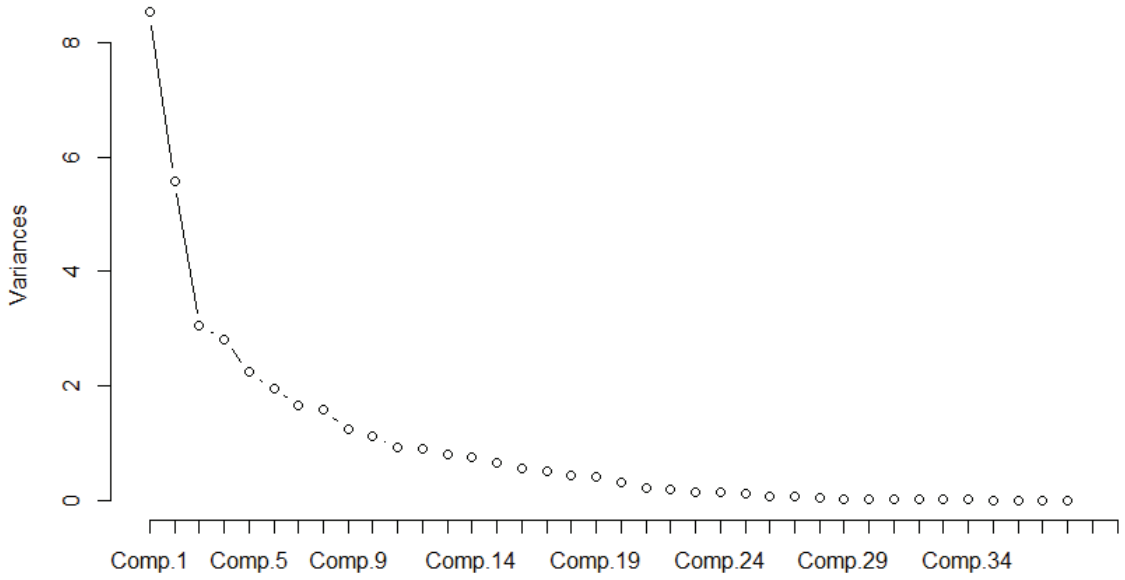

Figure 11. Scree plot of the variance values of the principal components for dataset2.

the tanglegrams in Figure 13, it is seen that BGCC (Baker's gamma correlation coefficient) gives inconsistent results for the centroid and median methods as well in dataset2. The FM_indexes in the table are equal to 1 for all methods. This means that in each method, the contents of the clusters formed as a result of clustering analysis with and without PCA are mutually identical. By evaluating the graphics and all the calculated coefficients together, it is observed that the best clustering result belonged to the PCA+Centroid method, and the cluster contents for the 3 number of clusters are given below.

Cluster 1: 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 Cluster 2: Germany Cluster 3: Russian Federation

Results for Dataset 3

The variance values of the principal components for dataset3 are given in Table 10 and Table 11. For the elbow test, when the scree plot which includes the variance values of the principal components given in Figure 12 is examined, it is seen that the natural breakpoint in the graph coincides with the 6th principal component. But 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.

Table 7. The variance values of the principal components for dataset2

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. The variance explained percentages of the principal components for dataset2

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

Table 9. Coefficient and index results used to measure the similarity of dendrograms for dataset2

	Methods	CPCC (dend 1)	CPCC (dend 2)	CPCC (tanglegram)	BGCC	FM_index
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**

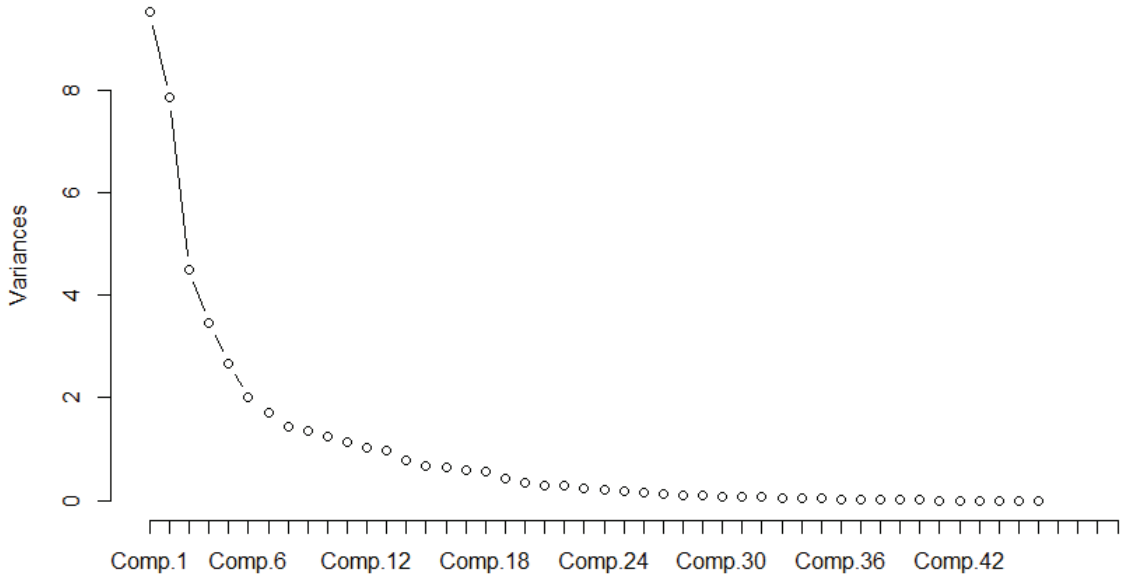

Figure 12. Scree plot of the variance values of the principal components for dataset3.

For the second approach, the mean-variance value ($\bar{\lambda}$) of the principal components is calculated as 0.9791 for dataset3, and it is observed that the variance values of the first 12 principal components are above this value. The first 12 components can explain %84.29 of the total variance in the data set.

For the 80% specific ratio, the first 11 principal components should be selected to achieve this specific ratio. The first 11 components can explain %82.00 of the total variance in the data set.

If the results of these three approaches are evaluated, it is clear that the mean-variance and 80% specific ratio approaches give very close results. As a result of all these evaluations, the number of components to be selected for a robust analysis process is determined as 11.

According to the CPCC (dend1 and dend2) values in Table 12, different from the results of dataset1 and dataset2, for dataset3, the results of PCA applied clustering analysis have close or lower similarity rates with the distance matrix in almost all methods except centroid linkage method. For

Table 10. The variance values of the principal components for dataset3

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**

Table 11. The variance explained percentages of the principal components for dataset3

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

Table 12. Coefficient and index results used to measure the similarity of dendrograms for dataset3

	Methods	CPCC (dend 1)	CPCC (dend 2)	CPCC (tanglegram)	BGCC	FM_index
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

CPCC(dend1) and CPCC(dend2), while the best clustering results belong to PCA+centroid and PCA+average methods, the worst clustering belongs to the variance-based Ward's method due to the outliers and low correlation in the dataset similar to dataset1 and dataset2. Consistent with the tanglegrams, the highest similarity rate belongs to single linkage and centroid linkage methods for the CPCC (tanglegram). When evaluated together with the tanglegrams above, it is seen that BGCC (Baker's gamma correlation coefficient) gives inconsistent results for the centroid and median methods as well in dataset3, and the highest value belongs to single linkage method. The FM_indexes in t table 12 are equal to 1 for all methods except complete linkage and ward's criterion linkage method. This means that in each method except for complete and ward's linkage, the contents of the clusters formed as a result of clustering analysis with and without PCA are mutually identical. By evaluating the graphics and all the calculated coefficients together, it is observed that the best clustering result belonged to the PCA+Centroid method and the cluster contents for the 3 number of clusters are given below.

Cluster 1: 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 Cluster 2: India Cluster 3: China

CONCLUSION

PCA is one of the most important and frequently used methods to perform dimension reduction and to cope with the difficulties such as the increase in the amount and dimension of the data in the data pre-processing stages. The main purpose of this study is to examine the effect of PCA on hierarchical methods in dimension reduction for high-dimensional datasets. In the light of the observed results, it is seen that, 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. Low correlation between variables reduces the dimension reduction efficiency of PCA [26]. Despite 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 with PCA is decreased 22 variables to 8 variables. For dataset2, the mean absolute correlation value is 0.2346, and dimension reduction with PCA is decreased 38 variables to 10 variables. For dataset3 the mean absolute correlation value is 0.2265 and dimension reduction with PCA is decreased 46 variables to 11 variables. When the CPCC values of the dendrograms are examined, it is noteworthy that the values of the Ward method are quite low for three datasets. The main reason explaining this situation is that the Ward method includes a completely variance-oriented merging process and the datasets in our study contain many outliers affecting the total variance. Considering the coefficient values, 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 methods have low outlier sensitivity. For these methods, it is seen that the FM_index values are equal to 1. Therefore, the clustering results are the same and the number of clusters is three except for average and average+PCA in dataset1. Baker's Gamma Correlation Coefficients of centroid linkage and median linkage methods, shown in red in the table 6, table 9 and, table 12 are given misleading results, incompatible with tanglegrams and other coefficients. For each dataset, all tanglegrams are carefully examined and it is observed that almost all the dendrograms obtained after the use of PCA formed a much more ordered hierarchical structure. The reason for this more ordered hierarchical structure can be thought of as the reduction of the negative effects of outliers in the dataset when PCA is used. 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 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 three datasets, the Centroid+PCA method with the highest coefficients and congruent graphical results is more robust and reliable compared to other methods. Despite the low correlation and outlier disadvantages of datasets, it is observed that PCA allows hierarchical methods to work more comfortably in less dimensional space and with less negative variance effect. High-dimensional data are data types used in analysis in many fields or sectors today. Working with high-dimensional data is quite difficult and has disadvantages such as longer processing times and lower quality of results. However, the general result of this study showed that the use of the PCA dimensionality reduction method together with clustering has a positive effect on the analysis process and results. As stated in the introduction of the study, there are many studies in the literature that support the results obtained. For example, it has been shown that PCA positively affects classification time and performance in the Support vector machine method [16]. Likewise, the positive effects of different dimension reduction techniques along with PCA have been shown in the analysis processes carried out in facial recognition systems and automatic intrusion detection systems [17, 19]. In addition, the positive effect of PCA in prediction with high-dimensional cancer data has been stated, and similarly, the positive effects of linear and non-linear dimension reduction methods have been evaluated in the study on fluid mechanics [18, 20]. In this study, the dimensionality reduction ability of PCA was evaluated only through hierarchical clustering techniques. Three different high-dimensional datasets were used in the study, but all three are low-correlation datasets. Positive and beneficial results were obtained in the study carried out under these restrictions. As suggestions for future studies, the dimensionality reduction performance of PCA can be investigated by using different clustering techniques or classification techniques. Additionally, the dimensionality reduction effect of PCA can also be investigated on deep learning methods. Another study that may be useful is to evaluate the performance of PCA on different correlated data sets. All parameters contained in the datasets and tanglegrams comparing the tanglegrams are included in the appendix section as Table A1**,** Figure A1, Figure A2, Figure A3.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

STATEMENT ON THE USE OF ARTIFICIAL INTELLIGENCE

Artificial intelligence was not used in the preparation of the article.

REFERENCES

[1] Statista. Worldwide data created. 2021. Available at: https://www.statista.com/statistics/871513/worldwide-data-created/ Accessed on September 10, 2025.

[2] Han J, Kamber M, Pei J. Data mining concepts and techniques. 3rd ed. San Francisco (CA): Morgan Kaufmann; 2012.

[3] Koij FS, Saba J. Using cluster analysis and principal component analysis to group lines and determine important traits in white bean. Procedia Environ Sci 2015;29:38–40. [CrossRef]

[4] Öten M, Albayrak S. Determination of Variation Between Some Alfalfa (Medicago sativa L.) Genotypes by Principal Component and Clustering Analysis. Turk Tarimsal Arastirma Derg 2018;5:222–228. [Turkish] [CrossRef]

[5] Shoba D, Vijayan R, Robin S, Manivannan N, Iyanar K, Arunachalam P, Nadarajan N, Arumugam Pillai M, Geetha S. Electronic J Plant Breed 2019;10:1095–1104. [CrossRef]
[6] Ioele G, De Luca M, Dinç E, Oliverio F, Ragno G. Artificial neural network combined with principal component analysis for resolution of complex pharmaceutical formulations. Chem Pharm Bull (Tokyo) 2011;59:35–40. [CrossRef]
[7] Kim YI, Boogert ST, Honda Y, Lyapin A, Park H, Terunuma N, Tauchi T, Urakawa J. Principal component analysis of cavity beam position monitor signals. J Instrum 2014;9:P02007. [CrossRef]
[8] Gaitani N, Lehmann C, Santamouris M, Mihalakakou G, Patargias P. Using principal component and cluster analysis in the heating evaluation of the school building sector. Appl Energy 2010;87:2079–2086. [CrossRef]
[9] Nwangburuka CC, Kehinde OB, Ojo DK, Denton OA, Popoola AR. Morphological classification of genetic diversity in cultivated okra, Abelmoschus esculentus (L.) Moench using principal component analysis and single linkage cluster analysis. Afr J Biotechnol 2011;10:11165–11172. [CrossRef]
[10] Margaritis A, Soenen H, Fransen E, Pipintakos G, Jacobs G, Blom J, Van den Bergh W. Identification of ageing state clusters of reclaimed asphalt binders using principal component analysis and hierarchical cluster analysis based on chemo-rheological parameters. Constr Build Mater 2020;244:118276. [CrossRef]
[11] Penkova T. Principal component analysis and cluster analysis for evaluating the natural and anthropogenic territory safety. Procedia Comput Sci 2017;112:99–108. [CrossRef]
[12] Abdulhafedh A. Incorporating k-means, hierarchical clustering and PCA in customer segmentation. J City Dev 2021;3:12–30.
[13] Yıldız K, Çamurcu AY, Doğan B. A comparative analysis of principal component analysis and non-negative matrix factorization techniques in data mining. In: Akademik Bilişim'10; Turkey. 2010. p. 12.
[14] Al-Omairi L, Abawajy J, Chowdhury MU, Al-Quraishi T. High-dimensionality graph data reduction based on a proposed new algorithm. EPiC Ser Comput 2019;63:1–10. [CrossRef]
[15] Al-Omairi L, Abawajy J, Chowdhury MU, Al-Quraishi T. An empirical analysis of graph-based linear dimensionality reduction techniques. Concurr Comput 2021;33:e5990. [CrossRef]
[16] Bharadiya JP. A tutorial on principal component analysis for dimensionality reduction in machine learning. Int J Innov Sci Res Technol 2023;8:2028–2032.
[17] Mousavi A, Arefanjazi H, Sadeghi M, Ghahfarokhi AM, Beheshtinejad F, Masouleh MM. Comparison of feature extraction with PCA and LTP methods and investigating the effect of dimensionality reduction in the bat algorithm for face recognition. Int J Robot Control Syst 2023;3:500–509. [CrossRef]
[18] Kabir MF, Chen T, Ludwig SA. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. Healthc Anal 2023;3:100125. [CrossRef]
[19] Nabi F, Zhou X. Enhancing intrusion detection systems through dimensionality reduction: a comparative study of machine learning techniques for cyber security. Cyber Secur Appl 2024;2:100033. [CrossRef]
[20] Wang Z, Zhang G, Xing X, Xu X, Sun T. Comparison of dimensionality reduction techniques for multi-variable spatiotemporal flow fields. Ocean Eng 2024;291:116421. [CrossRef]
[21] Reddy CK, Vinzamuri B. A survey of partitional and hierarchical clustering algorithms. In: Data clustering algorithms and applications. Boca Raton (FL): CRC Press; 2014. p. 87–110. [CrossRef]
[22] Xu R, Wunsch DC. Clustering. Hoboken (NJ): John Wiley & Sons; 2009.
[23] Everitt BS, Landau S, Leese M, Stahl D. Cluster analysis. 5th ed. Hoboken (NJ): John Wiley & Sons; 2011. [CrossRef]
[24] Forina M, Armanino C, Raggio V. Clustering with dendrograms on interpretation variables. Anal Chim Acta 2002;454:13–19. [CrossRef]
[25] Rencher AC. Methods of multivariate analysis. 2nd ed. Hoboken (NJ): John Wiley & Sons; 2002. [CrossRef]
[26] Yim O, Ramdeen KT. Hierarchical cluster analysis: comparison of three linkage measures and application to psychological data. Quant Methods Psychol 2015;11:8–21. [CrossRef]
[27] Tokuda EK, Comin CH, Costa LD. Revisiting agglomerative clustering. Physica A 2021;574:126433. [CrossRef]
[28] Mohbey KK, Thakur GS. An experimental survey on single linkage clustering. Int J Comput Appl 2013;76:1–6. [CrossRef]
[29] El-Hamdouchi A, Willett P. Techniques for the measurement of clustering tendency in document retrieval systems. J Inf Sci 1987;13:361–366. [CrossRef]
[30] Jarman AM. Hierarchical cluster analysis: comparison of single linkage, complete linkage, average linkage and centroid linkage method. 2020. Preprint. doi: 10.13140/RG.2.2.11388.90240
[31] Emmendorfer LR, Canuto AMP. A generalized average linkage criterion for hierarchical agglomerative clustering. Appl Soft Comput 2021;101:106990. [CrossRef]
[32] Bu J, Liu W, Pan Z, Ling K. Comparative study of hydrochemical classification based on different hierarchical cluster analysis methods. Int J Environ Res Public Health 2020;17:9515. [CrossRef]
[33] Kaufman L, Rousseeuw PJ. Finding groups in data: an introduction to cluster analysis. Hoboken (NJ): John Wiley & Sons; 2005.

[34] Majerova I, Nevima J. The measurement of human development using the Ward method of cluster analysis. J Int Stud 2017;10:239–257. [CrossRef]
[35] Szekely GJ, Rizzo ML. Hierarchical clustering via joint between-within distances: extending Ward's minimum variance method. J Classif 2005;22:151–183. [CrossRef]
[36] Murtagh F, Legendre P. Ward's hierarchical clustering method: clustering criterion and agglomerative algorithm. J Classif 2014;31:274–295. [CrossRef]
[37] Abdi H, Williams LJ. Principal component analysis. WIREs Comput Stat 2010;2:433–459. [CrossRef]
[38] Candes EJ, Li X, Ma Y, Wright J. Robust principal component analysis. J ACM 2011;58:11. [CrossRef]
[39] Mudrova M, Prochazka A. Principal component analysis in image processing. Res Gate Publ 2005. Preprint. Available at: https://www.semanticscholar.org/paper/PRINCIPAL-COMPONENT-ANALYSIS-IN-IMAGE-PROCESSING-Mudrov%C3%A1-Proch%C3%A1zka/76a7fc9d87736c8383576865cf50403e53e74848 Accessed on September 10, 2025.
[40] Gemperline PJ. Principal component analysis. In: Practical guide to chemometrics. 2nd ed. Boca Raton (FL): CRC Press; 2006. p. 211–235. [CrossRef]
[41] Triayudi A, Fitri I. Comparison of parameter-free agglomerative hierarchical clustering methods. ICIC Express Lett 2018;12:973–980.
[42] Silva AR, Dias CTS. A cophenetic correlation coefficient for Tocher's method. Pesqui Agropecu Bras 2013;48:589–596. [CrossRef]
[43] Nino JO, Berzal F. Evaluation metrics for unsupervised learning algorithms. arXiv:1905.05667. Preprint 2019. doi: 10.48550/arXiv.1905.05667
[44] Kumar S, Toshniwal D. Analysis of hourly road accident counts using hierarchical clustering and cophenetic correlation coefficient. J Big Data 2016;3:1–11. [CrossRef]
[45] Rao AR, Srinivas VV. Regionalization of watersheds by hybrid-cluster analysis. J Hydrol 2006;318:37–56. [CrossRef]
[46] Saraçli S, Doğan N, Doğan I. Comparison of hierarchical cluster analysis methods by cophenetic correlation. J Inequal Appl 2013;2013:203. [CrossRef]
[47] Fowlkes EB, Mallows CL. A method for comparing two hierarchical clusterings. J Am Stat Assoc 1983;78:553–569. [CrossRef]
[48] Hryniewicz O. Goodman–Kruskal measure of dependence for fuzzy ordered categorical data. Comput Stat Data Anal 2006;51:323–334. [CrossRef]
[49] Kılıç AF. Kategorik veride faktör analizi için kullanılabilecek alternatif bir korelasyon matrisi: Goodman-Kruskal gamma. Marmara Univ Ataturk Egitim Fak Egitim Bilim Derg 2021;54:151–168. [CrossRef]
[50] United Nations. UN Data. 2021. Available at: https://data.un.org/ Accessed on September 10, 2025.
[51] UNICEF. UNICEF data. 2021. Available at: https://data.unicef.org/ Accessed on September 10, 2025.
[52] World Bank. World Bank open data. 2021. Available at: https://data.worldbank.org/ Accessed on September 10, 2025.
[53] Our World in Data. Our World in Data. 2021. Available at: https://ourworldindata.org/ Accessed on September 10, 2025.
[54] Worldometer. Real time world statistics. 2021. Available at: https://www.worldometers.info/ Accessed on September 10, 2025.

APPENDICES

Table A1. The variables used in datasets

General Information = (GI)	Social Indicators = (SI)
1. Population (000, 2020)	**23.** Population growth rate (average annual %)
2. Pop. density (per km2, 2020)	**24.** Urban population (% of total population)
3. Surface area (km2)	**25.** Fertility rate, total (live births per woman)
4. Sex ratio (m per 100 f)	**26.** Life expectancy at birth (females/males, years)
5. Exchange rate (per US$)	**27.** Population age distribution (0-14/60+ years old, %)
Economic Indicators = (EI)	**28.** International migrant stock (000/% of total pop.)
6. GDP: Gross domestic product (million current US$)	**29**. Refugees and others of concern to UNHCR (000)
7. GDP growth rate (annual %, const. 2015 prices)	**30.** Infant mortality rate (per 1 000 live births)
8. GDP per capita (current US$)	**31.** Health: Current expenditure (% of GDP)
9. Economy: Agriculture (% of Gross Value Added)	**32.** Health: Physicians (per 1 000 pop.)
10. Economy: Industry (% of Gross Value Added)	**33.** Education: Government expenditure (% of GDP)
11. Economy: Services and other activity (% of GVA)	**34.** Education: Primary gross enrol. ratio (f/m per 100 pop.)
12. Employment in agriculture (% of employed)	**35.** Education: Secondary gross enrol. ratio (f/m per 100 pop.)
13. Employment in industry (% of employed)	**36.** Education: Tertiary gross enrol. ratio (f/m per 100 pop.)
14. Employment in services (% employed)	**37.** Intentional homicide rate (per 100 000 pop.)
15. Unemployment (% of labour force)	**38.** Seats held by women in national parliaments (%)
16. Labour force participation rate (female/male pop. %)	**Environment and Infrastructure Indicators = (ENI)**
17. CPI: Consumer Price Index (2010=100)	**39.** Individuals using the Internet (per 100 inhabitants)
18. Agricultural production index (2004-2006=100)	**40.** Threatened species (number)
19. International trade: exports (million current US$)	**41.** Forested area (% of land area)
20. International trade: imports (million current US$)	**42.** CO2 emission estimates (million tons/tons per capita)
21. International trade: balance (million current US$)	**43.** Energy production, primary (Petajoules)
22. Balance of payments, current account (million US$)	**44.** Energy supply per capita (Gigajoules)
	45. Tourist/visitor arrivals at national borders (000)
	46. Important sites for terrestrial biodiversity protected (%)

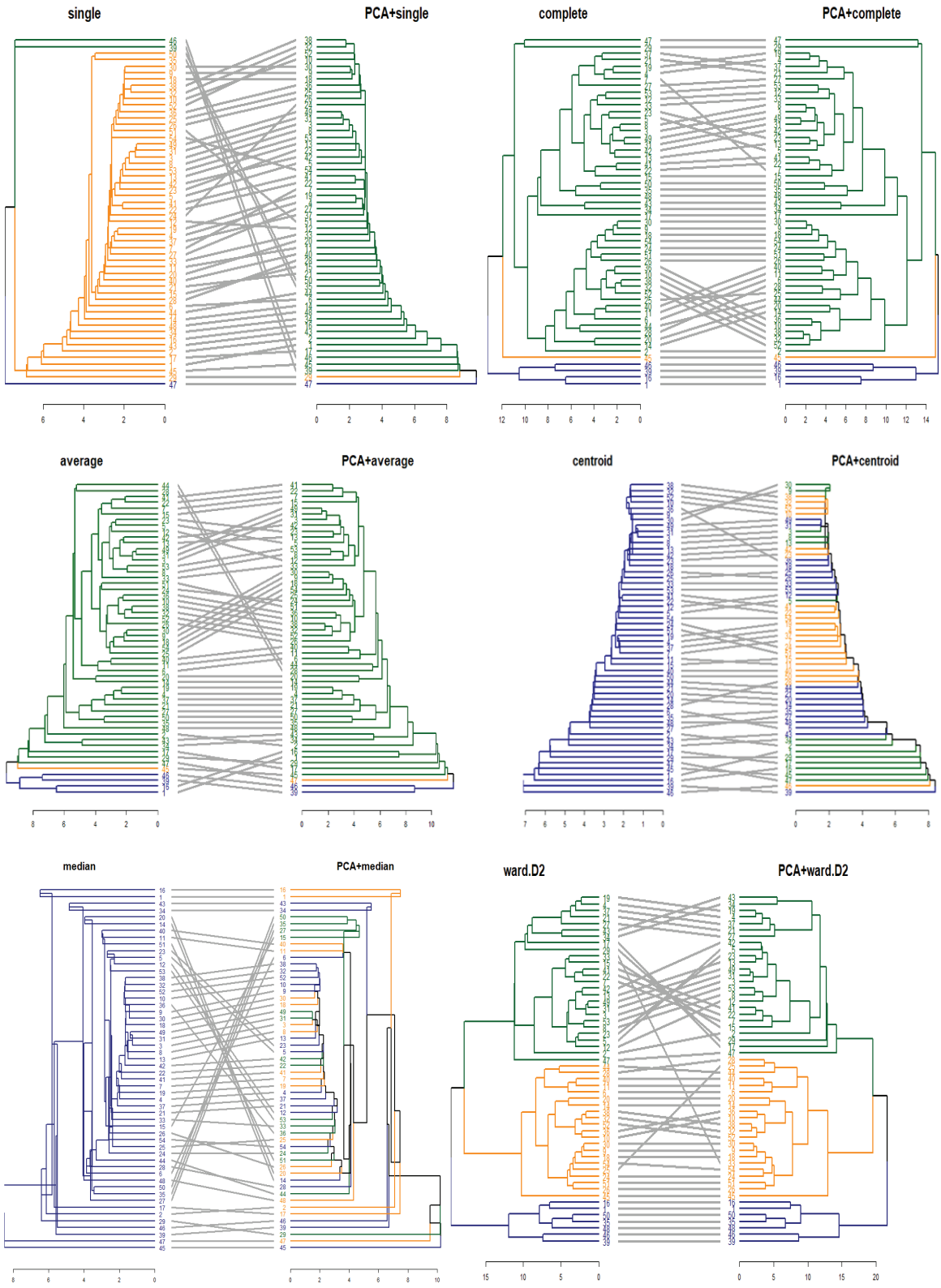

Figure A1. The tanglegrams for dataset1.

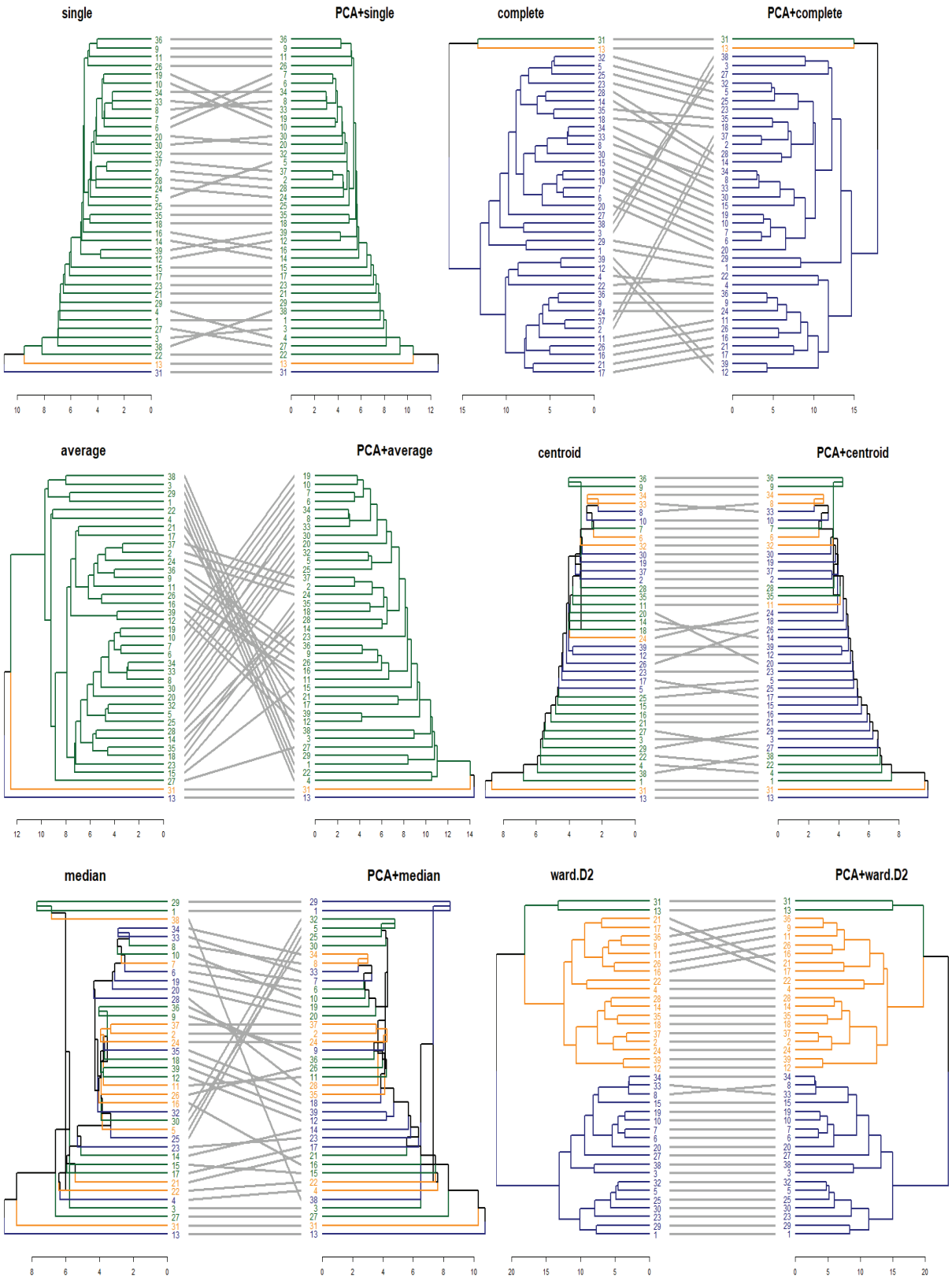

Figure A2. The tanglegrams for dataset2.

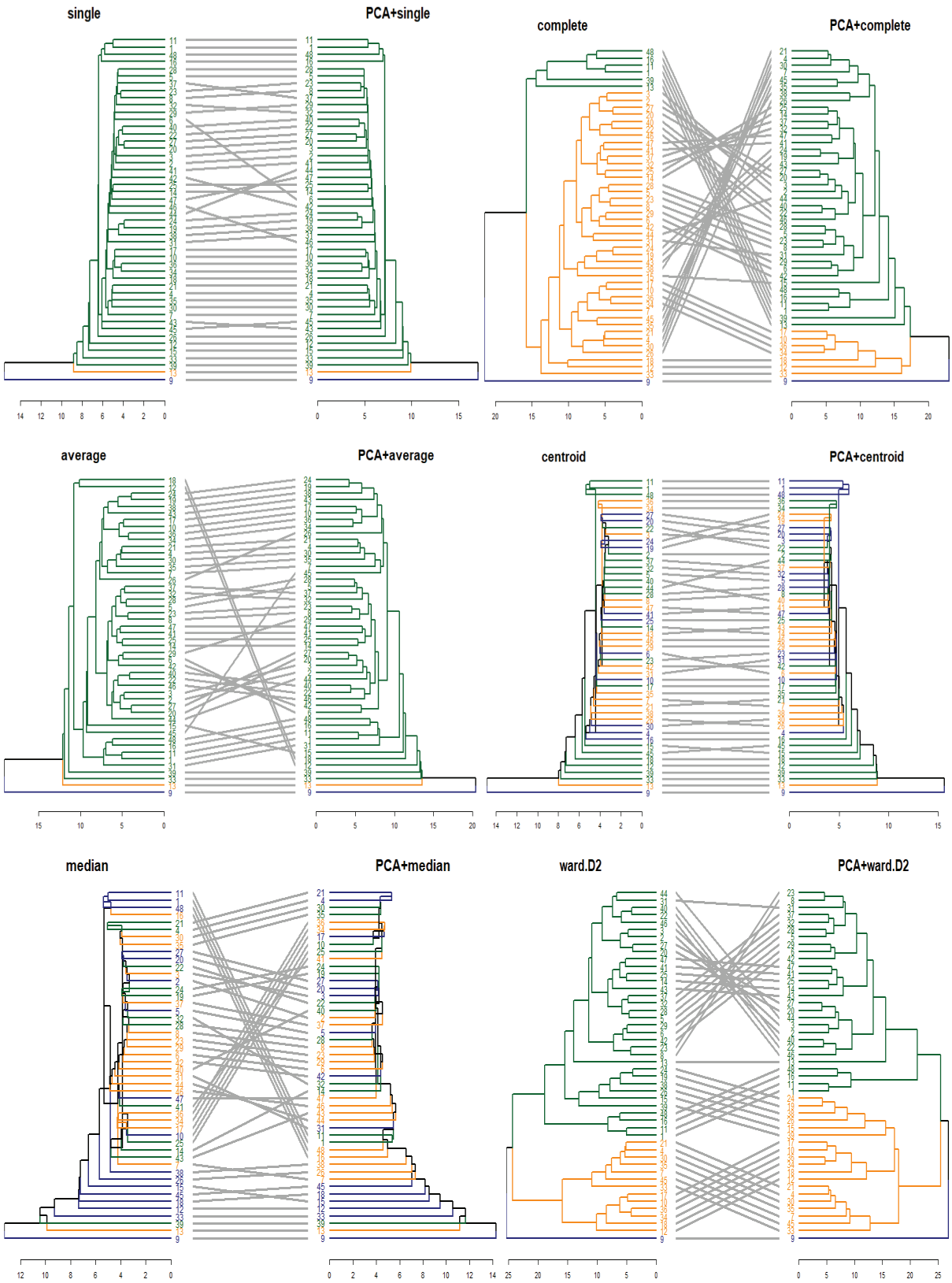

Figure A3. The tanglegrams for dataset3.