Mathematical Analysis of the Clustering of Ostracoda Concerning Their Habitat Preferences

MEHMET CEVRI
Mathematics Department, Faculty of Science,
Istanbul University,
34134 Vezneciler, Fatih, Istanbul,
TURKEY

Abstract: - The analysis of data, while interesting when a single variable is involved, becomes truly fascinating and challenging when several variables are present. There are various multivariate analysis methods available for examining the relationships among multiple variables simultaneously. Principal component analysis and cluster analysis are two commonly used techniques that are valuable tools in many scientific fields. Principal component analysis is employed to reduce the dimensionality of correlated measurements, whereas cluster analysis is utilized to classify objects or cases into relatively homogeneous groups. On the other hand, Ostracods can be utilized as bioindicators of the surrounding physical and chemical conditions. This paper presents a methodology for employing principal component analysis to cluster Ostracods based on their habitat preferences. Simulation results obtained using Mathematica software, demonstrate that anthropogenic water sources significantly influence the distribution of non-marine Ostracods.

Key-Words: - Multivariate analysis, principal component analysis, cluster analysis, Ostracoda, dendrogram, optimization.

Received: April 7, 2024. Revised: September 9, 2024. Accepted: October 11, 2024. Published: November 25, 2024.

1 Introduction

In today's world, we are living in the information age, where computational technology and modern facilities are rapidly developing. We frequently encounter large data sets generated by experiments and computer simulations, [1], [2]. Multivariate statistical methods, [3], [4] are employed to identify patterns within a set of variables. One such method is principal component analysis, [5], [6], [7], which employs mathematical procedures to simplify interrelated measures within the data. Principal component analysis (PCA) is the most popular multivariate statistical technique, used by almost all scientific disciplines. PCA is a powerful tool that can be applied to a wide variety of problems in behavioral and social sciences, engineering [8], genetics [9], [10], neuroscience [11] and geography [12]. The advent of computing technology has enabled the application of PCA in a variety of fields. It is also probable that this technique is the oldest multivariate technique. It was initially introduced by [13] and subsequently developed by [14], who also created the terminology "principal component.". Currently, it is one of the most frequently employed tools for exploratory data analysis and the creation of predictive models.

Cluster analysis (CA) is an unsupervised learning technique that aims to divide a set of data into groups or clusters. The observations within the same group tend to exhibit similarities, whereas those in different groups display differences. Further detailed information regarding clustering methods can be found in the references [15], [16], [17], [18], [19].

The main objective of principal component analysis is to reduce the dimensionality of a data set containing a high number of related variables while preserving as much of the variation in the data set as possible. This is achieved by transforming the original variables into a new set of uncorrelated variables known as principal components (PCs). They are ordered in such a way that the first few retain most of the variation present in all the original variables.

Ostracods are a type of small bivalved crustacean that have been around for 500 million years, [20]. Despite their relatively small size, these organisms play a crucial role in a variety of important ecological processes, including sedimentation, mineralization and biochemical cycling. They are distributed worldwide and can live in a wide range of habitats, from hot springs to Maar lakes and salt marshes, [21], [22], [23].

Ostracods are of significant value in the field of biostratigraphy due to their rapid evolutionary rates and extensive geographic distribution. The faunal pattern of an area is highly related to the habitat variety, which depends on ecological variations. The variety of habitat types, such as lakes, streams, and lagoons, plays a critical role in the diversity of species. It is therefore crucial to consider the impact of habitat types on species diversity. Several studies, [24], [25] have been conducted to elucidate the relationship between habitat characteristics and faunal patterns. These studies demonstrate that species richness is frequently associated with habitat diversity, which is in turn related to the size of the study area. Furthermore, they contribute to our comprehension of past environmental conditions and climate change. However, the classification of Ostracods based on their habitat preferences using principal component analysis was not presented. Consequently, this paper will focus on this problem.

The objective of this paper is to identify the relationship between the distribution of Ostracoda species and their habitat preferences. To this end, an effective classification algorithm has been developed based on principal component analysis and cluster analysis in Mathematica software on a set of Ostracoda data. The region of Thrace was selected as the study area, which comprises five provinces with a multitude of water sources, both natural and artificial.

2 Materials and Methods

2.1 Principal Component Analysis

Principal component analysis is a statistical technique that identifies patterns in data and expresses the data in a way that highlights similarities and differences. PCA is an invaluable tool for data analysis, particularly in the context of large data sets that are challenging to represent graphically. It effectively identifies patterns in data that would otherwise be difficult to discern. It is a widely employed technique for extracting the maximum variance from a dataset, which results in a reduction of the number of variables into a smaller number of components, [26], [27]. The objective of PCA is to identify a new set of uncorrelated variables (principal components) that can explain the greatest possible proportion of the total variation. In other words, PCA is designed to reduce the number of variables that need to be considered to a small number of indices, which are called principal components. These are linear combinations of the original variables. In essence, PCA is a method of simplifying data by reducing the number of variables.

Suppose that $\mathbf{X}$ is a vector of p random variables $X_1, X_2, ..., X_p$. To simplify the description of these variables, we subtract the mean of each dataset from each observation. This produces a dataset with a mean of zero, thus,

$$x_j = X_j - \overline{X_j}, \quad j = 1, 2, ..., p \tag{1}$$

Let $\mathbf{x}^T = \left(x_1, x_2, ..., x_p\right)$ be a random vector with covariance matrix $\mathbf{\Sigma}$. Consider forming new variables $Z_1, Z_2, ..., Z_k$ ($k \leq p$) linear combination of x-variables:

$$\begin{aligned} Z_1 &= \boldsymbol{\alpha}_1^T \mathbf{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + ... + \alpha_{1p}x_p \\ Z_2 &= \boldsymbol{\alpha}_2^T \mathbf{x} = \alpha_{21}x_1 + \alpha_{22}x_2 + ... + \alpha_{2p}x_p \\ &\vdots \\ Z_k &= \boldsymbol{\alpha}_k^T \mathbf{x} = \alpha_{k1}x_1 + \alpha_{k2}x_2 + ... + \alpha_{kp}x_p \end{aligned} \tag{2}$$

PCA is a technique for dimensionality reduction from p dimensions to $k < p$ dimensions. It tries to find, in order, the most informative k linear combinations of set variables $Z_1, Z_2, ..., Z_k$.

Having defined PCs, we need to know how to find them. We consider the vector of random variables $\mathbf{x}$ has a known covariance matrix $\mathbf{\Sigma}$ but the more realistic case, where $\mathbf{\Sigma}$ is unknown, follows by replacing $\mathbf{\Sigma}$ by a sample covariance matrix $\mathbf{S}$. To derive the form of the PCs, consider first $\mathbf{Z}_1 = \boldsymbol{\alpha}_1^T \mathbf{x}$ where the vector $\boldsymbol{\alpha}_1$ maximizes $Var\left(\mathbf{Z}_1\right) = \boldsymbol{\alpha}_1^T \mathbf{\Sigma} \boldsymbol{\alpha}_1$ subject to $\boldsymbol{\alpha}_1^T . \boldsymbol{\alpha}_1 = 1$. In this case, the standard approach is to use the technique of Lagrange multipliers $\left(\lambda_i\right)$ that are frequently used when maximizing functions subject to some constraints. To maximize $Var\left(\mathbf{Z}_1\right)$, Lagrange function $L\left(\boldsymbol{\alpha}_1, \lambda_1\right)$,

$$L\left(\boldsymbol{\alpha}_1, \lambda_1\right) = \boldsymbol{\alpha}_1^T \mathbf{\Sigma} \boldsymbol{\alpha}_1 - \lambda_1 \left(\boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 - 1\right), \tag{3}$$

where λ_1 is a Lagrange multiplier. Differentiation with respect to $\boldsymbol{\alpha}_1$ gives,

$$\left(\mathbf{\Sigma} - \lambda_1 \mathbf{I}_p\right) \boldsymbol{\alpha}_1 = 0 \tag{4}$$

where $\mathbf{I}_p$ is the $(p \times p)$ identity matrix. Hence, λ_1 is an eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_1$ is the corresponding eigenvector or the weight. Note that the quantity to be maximized is

$$Var(\mathbf{Z}_1) = \lambda_1 \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = \lambda_1 \ . \quad (5)$$

So λ_1 must be as large as possible. In this case $\boldsymbol{\alpha}_1$ is the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{\Sigma}$ and λ_1 is the largest eigenvalue.

In order to obtain second PC, we want to $Var(\mathbf{Z}_2)$ subject to $\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1 = 0$ and $\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_2 = 1$. Thus, Lagrange function is,

$$L(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \lambda_2, \beta) = \boldsymbol{\alpha}_2^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda_2 \left(\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_2 - 1 \right) - \beta \left(\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1 \right) \quad (6)$$

where λ_2 and β are Lagrange multipliers.

Differentiation with respect to $\boldsymbol{\alpha}_2$ gives,

$$\boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda_2 \boldsymbol{\alpha}_2 - \beta \boldsymbol{\alpha}_1 = 0, \quad (7)$$

and multiplication of Eq. (7) on the left by $\boldsymbol{\alpha}_1^T$ gives

$$\boldsymbol{\alpha}_1^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_2 - \lambda_2 \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_2 - \beta \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = 0, \quad (8)$$

which, since $\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1 = 0$ and $\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_2 = 1$, Eq. (8) reduces to $\beta = 0$. Thus Eq. (7) becomes,

$$\left(\boldsymbol{\Sigma} - \lambda_2 \mathbf{I}_p \right) \boldsymbol{\alpha}_2 = 0 \ . \quad (9)$$

Hence λ_2 once more eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}_2$ the corresponding eigenvector. Again, $Var(\mathbf{Z}_2) = \lambda_2$, so λ_2 is to be as large as possible.

In general, the kth principal component of $\mathbf{x}$ is $\mathbf{Z}_k = \boldsymbol{\alpha}_k^T \mathbf{x}$ and $Var(\mathbf{Z}_k) = Var\left(\boldsymbol{\alpha}_2^T \mathbf{x} \right) = \lambda_k$, where λ_k is the kth largest eigenvalue of $\boldsymbol{\Sigma}$, and $\boldsymbol{\alpha}_k$ is the corresponding eigenvector. Therefore $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k \geq 0$ condition holds for the eigenvalues. Namely, the obtained principal components are in decreasing order of variance, $Var(\mathbf{Z}_1) \geq Var(\mathbf{Z}_2) \geq ... \geq Var(\mathbf{Z}_k)$. In this case $\mathbf{Z}_1$ explains as much variance as possible and $\mathbf{Z}_2$ explains as much of the remaining variance as possible. The kth PC, $\mathbf{Z}_k = \boldsymbol{\alpha}_k^T \mathbf{x}$ maximizes $Var(\mathbf{Z}_k) = \boldsymbol{\alpha}_k^T \boldsymbol{\Sigma} \boldsymbol{\alpha}_k$ subject to $\boldsymbol{\alpha}_k^T . \boldsymbol{\alpha}_k = 1$ and $Cov(\mathbf{Z}_i, \mathbf{Z}_k) = 0, (i \neq k)$. It can be shown that for the third, fourth, ..., pth PCs, the vectors of coefficients $\boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4, ..., \boldsymbol{\alpha}_p$ are the eigenvectors of $\boldsymbol{\Sigma}$ corresponding to $\lambda_3, \lambda_4, ..., \lambda_p$, the third and fourth largest,...,and the smallest eigenvalue, respectively.

It is important to note that on occasion, the vectors $\boldsymbol{\alpha}_k$ are referred to as *principal components*. Although this usage is occasionally defended, it is nonetheless confusing. It is therefore preferable to reserve the term '*principal components* or *principal components scores* **P**' for the derived variables,

$$\mathbf{P} = \mathbf{x}.\boldsymbol{\alpha}, \quad (10)$$

and refer to $\boldsymbol{\alpha}$ as the *eigenvectors* or *loadings matrix*. Consequently, in a PCA model, each eigenvalue represents the degree of variation in the original features that can be explained by the associated principal components. For a more comprehensive understanding, please refer to the information provided in reference, [5]. The degree to which the selected principal components "explain" the variance of each of the variables is quantified by a statistic known as *communality*. The commonalities for the kth variable are computed by taking the sum of the squared loadings for that variable, [28]. This is expressed by:

$$h_k^2 = \sum_{i=1}^{p} a_{ki}^2 , \quad (11)$$

where a_{ki} represents the loadings of variables Z_k.

2.2 Cluster Analysis

The term "cluster analysis" is used to describe a wide range of techniques employed in the construction of classifications. A number of these techniques are discussed in detail by [29] and [30]. Cluster analysis is a method of identifying groups of individuals who are similar to one another. This concept of similarity is of great importance in all scientific fields. The utilization of appropriate measures not only enhances the quality of information selection but also minimizes the time and processing costs. There are numerous similarity measures [31] available for use in a variety of applications and contexts. The Euclidean distance is a metric that is widely known to be the most commonly used to determine the distance between two vectors $\mathbf{x} = (x_1, x_2, ..., x_p)$ and $\mathbf{y} = (y_1, y_2, ..., y_p)$ defined as,

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})(\mathbf{x} - \mathbf{y})^T} = \sqrt{\sum_{j=1}^{p} (x_j - y_j)^2} \quad (12)$$

Once you have selected the distance or similarity measure, you must choose the appropriate clustering algorithm. Ward's method, also known as the incremental sum of squares method, is a commonly used approach in hierarchical clustering. This method utilizes both the within-cluster (squared) distances and the between-cluster (squared) distances, [32], [33]. The objective is to generate clusters that minimize the within-cluster variance and maximize the between-cluster variance. This approach differs from the conventional approach in that it does not combine the two most similar objects successively. Instead, those objects whose merger results in the smallest possible increase in the within-cluster variance are combined. If AB is the cluster obtained by combining clusters A and B, then the sum of within-cluster distances (equivalent to within-cluster sums of squares (SSE)) are:

$$SSE_r = \sum_{i=1}^{n_r} (\mathbf{y}_i - \bar{\mathbf{y}}_r)(\mathbf{y}_i - \bar{\mathbf{y}}_r)^T, \quad (r = A, B \text{ and } AB) \quad (13)$$

where $\bar{\mathbf{y}}_{AB} = (n_A \bar{\mathbf{y}}_A + n_B \bar{\mathbf{y}}_B)/(n_A + n_B)$, n_A, n_B and $n_{AB} = n_A + n_B$ are the numbers of points in A, B and AB respectively. In this case Ward's method joins the two clusters A and B that minimize the increase in SSE, defined as

$$I_{AB} = SSE_{AB} - (SSE_A + SSE_B) = \frac{n_A n_B}{n_A + n_B} (\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)(\bar{\mathbf{y}}_A - \bar{\mathbf{y}}_B)^T. \quad (14)$$

In this paper, we utilize the Ward's method, which is based on Euclidean distances.

3 Simulations and Results

This study definitively identified 27 Ostracoda species in only 60 of the 95 stations in the Thrace region, as shown in Table 1. The full names of the abbreviated codes for the Ostracoda taxa are provided in [34]. The abbreviations employed for these species are presented in Table 1. The similarities of species were analyzed based on their habitat distribution. The analyses were performed using Mathematica software. The differentiation of habitat preferences in the Ostracoda genus represents the most useful taxonomic characterization. The distribution and numbers of these organisms in their habitat preferences play a significant role in the infrageneric classification of their genus. Table 1 provides a clear illustration of the distribution of species and their habitat preferences, presented in a data matrix comprising 27 rows and 5 columns.

Table 1. Species distribution and habitat preferences with number of stations

Species Abbreviations	Lagoons	Reservoirs	Lakes	Streams	Troughs
lehi	1	0	0	0	0
liin	0	6	0	0	0
list	0	1	0	0	0
pare	0	0	1	0	0
cyto	2	0	0	1	0
poel	1	0	0	0	0
potu	0	0	0	1	0
tyam	0	0	1	0	0
llgi	0	1	0	2	0
llbi	0	1	0	0	4
llde	0	2	0	0	1
llmo	0	1	2	0	0
llbr	0	0	0	5	9
cane	0	0	1	3	0
psha	0	0	2	0	0
cyov	0	0	0	1	0
phkr	0	0	3	0	0
cyop	0	0	2	0	0
euin	0	0	0	1	1
prze	0	0	0	2	3
hesa	0	1	0	6	11
hein	1	0	0	9	16
hech	0	0	0	0	4
psol	0	0	0	2	8
cyvi	0	4	2	4	1
pova	0	0	0	0	1
povi	0	0	0	0	2

A confident classification of Ostracoda species will be achieved by assessing their habitat preference variability using the data matrix presented in Table 1. Principal component analysis was used to identify groups among the 27 individuals in the sample. Traditional taxonomic methods are not suitable for species identification, so this method was employed. In a similar manner, the distance between all other pairs of objects can be computed and expressed in a distance matrix.

The non-diagonal elements of the matrix represent the distances between pairs of objects. The diagonal elements are all zero, as the distance from an object to itself is always zero.

In order to ascertain the optimal number of clusters for the data, it is possible to utilize the dendrogram, which illustrates the distance level at

which a combination of objects and clusters was formed. Figure 1 depicts the dendrogram resulting from the application of Ward's clustering method to the complete Ostracoda taxa data set presented in Table 1.

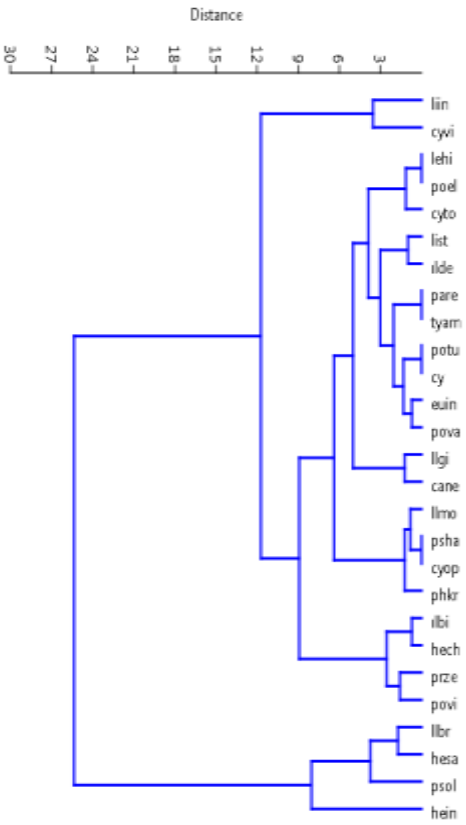

Fig. 1: Similarity analysis of species of Ostracods, calculated using Euclidean distance index and Ward's method

The dendrogram is read from left to right in order to ascertain the distance at which objects have been combined. For example, according to our calculations above, objects "liin", and "cyvi" are combined at a distance level of 5.0. By looking at the dendrogram, we could justify an eight-cluster solution ([llbr, hesa, psol, hein], [prze, povi,, hech, ılbi],…etc), as well as a lot of cluster subgroups ([liin, cyvi], [lehi, poel, cyto], [list, ılde], …etc]).

In principal component analysis, factor loadings, also known as component loadings, represent the correlation coefficients between the variables and observations. The total proportion of the variance in the sentence that is explained by the two factors is simply the sum of their squared factor loadings. This is referred to as the *communality* of the variable sentence. Table 2 provides the first two factor loadings and communalities for variables. The third column of Table 2 shows coefficients of linear combination that define loading 1 or PC1, and the fourth column shows coefficients for loading 2 or PC2. The loadings of the variable *reservoirs* on PC1 and PC2 are (-0.03) and (-0.91), respectively. The minus sign indicates an inverse or negative relationship; the absence of a sign is meant to imply a plus sign indicating a direct or positive relationship. Table 2 shows that variables *lagoons* have a very small role in explaining the variation on PC1, whereas r*eservoirs* are highly correlated with PC2, but negligibly correlated with PC1.

Table 2. Principal component solution, first two factor loadings and communalities for Ostracoda data

Lagoons	0.23	0.01	0.05	0.002	0.98
Reservoirs	1.93	-0.03	-0.91	0.83	42.94
Lakes	0.80	-0.06	-0.08	0.009	1.16
Streams	5.09	0.44	-0.37	0.336	6.61
Troughs	16.7	0.89	0.15	0.821	4.92

On the other hand, the *troughs* variable has the highest loading (0.89) on PC 1. Namely, *troughs* play a big role in explaining the variation on PC1, and *reservoirs* play a big role in explaining the variation on PC2. The commonality, or the proportion of the variance in each variable accounted for by two components shown in Table 2. For instance, 42.94 % of the variance in r*eservoirs* is accounted for in Table 2. We also observed that the factor model explains nearly 1%, 43%, 1 %, 7%, and 5%, respectively of the observed variance of *lagoons, reservoirs, lakes, streams, and troughs*.

How many principal components are needed to reproduce the observed covariance matrix to a satisfactory level of accuracy? It can be posited that a more straightforward determination of the optimal number of principal components, *m*, may be achieved through the utilization of graphical approaches, as proposed by [35], namely, the *scree plot*, which is a plot of eigenvalues versus component numbers, [36]. With regard to the Ostracod data, Figure 2 illustrates an optimal pattern in the plotted data.

Figure 2 demonstrates a sharp increase in the first two eigenvalues, followed by a bend, and then a gradual decrease. The elbow is clearly located at the second principal component. This means that the first two components should be kept for the analysis. Thus, we decide that the optimal number of components *m* should be 2. Once the number of principal components to be included has been

determined, the next step is to calculate the component scores. Table 3 provides the first two component scores for 27 Ostracoda species in the columns labeled PC Score 1 and PC Score 2.

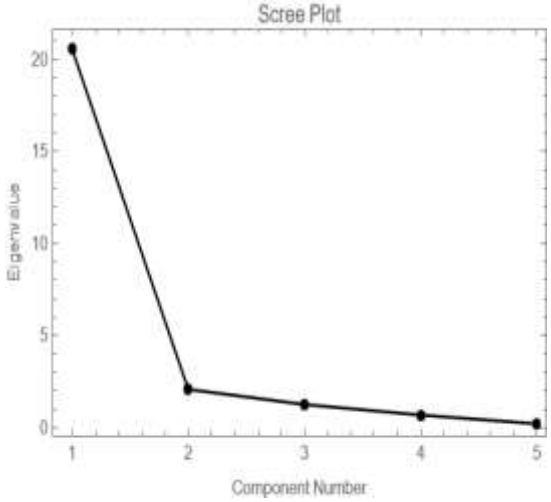

Fig. 2: Scree graph for the covariance matrix for data that most likely have 2 underlying factors

Table 3. The first two principal component scores of Ostracoda data

Species	PC Score1	PC Score 2
lehi	-2.57	0.83
liin	-2.77	-4.68
list	-2.61	-0.13
pare	-2.64	0.71
cyto	-2.11	0.50
poel	-2.57	0.83
potu	-2.14	0.41
tyam	-2.63	0.71
llgi	-1.72	-0.88
llbi	0.97	0.46
llde	-1.74	-0.89
llmo	-2.72	-0.28
llbr	7.68	0.24
cane	-1.30	-0.42
psha	-2.70	0.63
cyov	-2.13	0.41
phkr	-2.76	0.56
cyop	-2.70	0.63
euin	-1.24	0.56
prze	0.99	0.48
hesa	9.88	-0.75
hein	15.72	-0.18
hech	0.99	1.38
psol	5.46	1.22
cyvi	-0.16	-4.37
pova	-1.69	0.93
povi	-0.79	1.08

Table 3 clearly shows that the PC Score 1 and PC Score 2 values represent the coordinates for each of the 27 Ostracoda species in the original axis system. Thus, we are now ready to plot the principal component scores in a 2D graph.

Figure 3 displays the corresponding map, which shows a set of observations plotted concerning the first two principal component scores. We can say that the Ostracoda species are confidently grouped into eighth functional classes. The first cluster specifically includes *hein* and *hesa*, while the eight cluster is comprised *of liin* and *cyvi*.

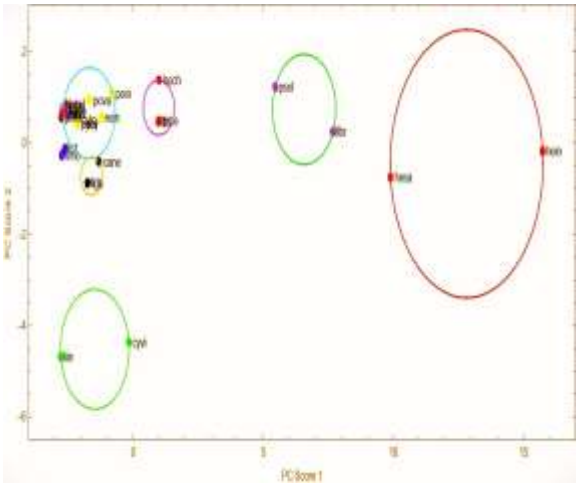

Fig. 3: Plot of the first and second principal component scores. The symbols in the figure legend correspond to the eight functional classes.

4 Conclusions

This paper presents a method that directly applies principal component analysis to correlated multivariate data. The utility of clustering analysis using PCA in identifying the habitat preferences of Ostracoda species is demonstrated through an illustrative example. Ostracods are extremely sensitive to environmental parameters, including temperature, salinity, and oxygen levels. Our analysis used a hierarchical clustering algorithm, but other clustering algorithms such as fuzzy clustering and density-based clustering can also enhance classification. The paper argues that PCA is an advantageous approach for clustering Ostracoda species based on their habitat preferences, particularly when analyzing multivariate data. An increase in the number of artificial water sources may result in an increased likelihood of encountering certain species of Ostracoda, which could potentially give rise to a false sense of richness. It seems likely that new technologies will provide a range of novel

applications for PCA in the years to come. In future research, we intend to expand our classification of many more Ostracoda species to encompass a greater diversity of habitats, employing factor analysis and the silhouette coefficient.

Acknowledgement:
I would like to express my gratitude to Assoc. Prof. Oya Özulug of the Biology Department of Istanbul University for kindly providing the Ostracoda data.

Declaration of Generative AI and AI-assisted Technologies in the Writing Process

During the preparation of this work the author used DeepL tool in order to refine certain expressions to align with academic standards. After using this tool/service, the author reviewed and edited the content as needed and takes full responsibility for the content of the publication.

References:

[1] N. Kereselidze, Mathematical and Computer Modelling of a Dynamic System for Effectively Combating Disinformation, WSEAS *Transactions on Systems*, Vol. 23, 2024, pp.66-72. https://doi.org/10.37394/23202.2024.23.7.

[2] H. Li, Corporate Accounting Management Risks Integrating Improved Association Rules and Data Mining, *WSEAS Transactions on Computer Research*, Vol. 12, 2024, pp.348- 358. https://doi.org/10.37394/232018.2024.12.34.

[3] B.F.J. Manly, *Multivariate Statistical Methods: A primer*, Chapman and Hall, 4th Edn. USA 2017. DOI: 10.1201/9781315382135.

[4] A. C. Rencher, *Methods of Multivariate Analysis*, Wiley, New Jersey, 2002. DOI: 10.1002/0471271357.

[5] I.T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, Springer Verlag, New York, 2002.

[6] G. Saporta and N. Niang, Principal *Component Analysis: Application to Statistical Process Control*, In: Govaert G, ed. Data Analysis. London: John Wiley & Sons, 2009, pp.1–23. http://dx.doi.org/10.1002/9780470611777.ch1.

[7] Y. Sun, S. Zhou, S. Meng, M. Wang and H. Mu, Principal Component Analysis–Artificial Neural Network-Based Model for Predicting the Static Strength of Seasonally Frozen Soils, Scientific *Reports* 13, Article number: 16085, 2023. http://dx.doi.org/10.1038/s41598-023-43462-7.

[8] F.A. Almeida, G.F. Gomes, P.P. Balestrassi and G. Belinato, *Principal Component Analysis: An Overview and Applications in Multivariate Engineering Problems*, Uncertainty Modeling: Fundamental Concepts and Models. Editora Cubo, 2022, pp. 172-194. DOI: 10.4322/978-65-86503-88-3.c06.

[9] E. Elhaik, Principal Component Analysis - Based Findings in Population Genetic Studies are Highly Biased and Must Be Reevaluated, *Scientific Reports* 12(1):14683, 2022. https://doi.org/10.1038/s41598-022-14395-4.

[10] D. Zhang, R. Day, S. Lee, Fast and Robust Ancestry Prediction Using Principal Component Analysis, *Bioinformatics* 36**,** 2020, pp. 3439-3446. https://doi.org/10.1093/bioinformatics/btaa152.

[11] X. Di and B.B. Biswal, Principal Component Analysis Reveals Multiple Consistent Responses to Naturalistic Stimuli in Children and Adults, *Human Brain Mapping* 43, 2022, pp. 3332-3345. https://doi.org/10.1002/hbm.25568.

[12] A. Cartone and P. Postiglione, Principal Component Analysis for Geographical Data: The Role of Spatial Effects in the Definition of Composite Indicators, *Spatial Economic Analysis* 16, 2021, pp. 126-147. DOI: 10.1080/17421772.2020.1775876.

[13] K. Pearson, LIII. On Lines and Planes of Closest Fit to Systems of Points in Space, Philosophical *Magazine Series* 6**,** 1901, pp. 559-572. DOI: 10.1080/14786440109462720.

[14] H. Hotelling, Analysis of A Complex of Statistical Variables into Principal Components, *J.Educ.Psychol*.25, 1933, pp. 417–441. DOI: 10.1037/H0071325.

[15] M.R. Anderberg, Cluster *Analysis for Applications*, Academic Press, New York, 1973.

[16] A.K. Jain and R.C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, NJ., 1988.

[17] R.C. Tryon and D.E. Bailey, Cluster *Analysis*, McGraw-Hill, New York, 1973.

[18] D. Parochial, Mathematics of Classifications, Chu spaces and the Continuum, *WSEAS Transactions on Information Science and Applications*, Vol. 20, 2023, pp.119-130. https://doi.org/10.37394/23209.2023.20.14.

[19] R. Eryigit, Y. Ar, B. Tuğrul, Classification of Trifolium Seeds by Computer Vision Methods, WSEAS *Transactions on Systems*, Vol. 22, 2023, pp.313-320. http://dx.doi.org/10.37394/23202.2023.22.34.

[20] H. Zwair, *Perspective Chapter: Ostracoda*, Formation and Evolution of Earth's Crust, *IntechOpen*, 2023. DOI: 10.5772/intechopen.112211.

[21] O. Özuluğ, S.N. Kubanç, C. Kubanç, and G.İ. Demirci, Checklist of Quaternary and Recent Ostracoda (Crustacea) Species from Turkey with Information on Habitat Preferences, *Turkish Journal of Bioscience and Collections* 2, 2018, pp. 51-100.

[22] M. Yavuzatmaca, Diversity Analyses of Nonmarine Ostracods (Crustacea, Ostracoda) in Streams and Lakes in Turkey, *Turkish Journal of Zoology* 44**,** 2020, pp. 519-530. http://dx.doi.org/10.3906/zoo-2005-20.

[23] D.J. Horne and I. Boomer, *The Role of Ostracoda in Saltmarsh Meiofaunal Communities*, In: Sherwood, B.R., Gardiner, *B.G.*, Harris, T.(eds.) British Saltmarshes, 2000, pp.182-202.

[24] O. Külköylüoğlu, On the Usage of Ostracods (Crustacea) as Bioindicator Species in Different Aquatic Habitats in the Bolu Region (Turkey), *Ecological Indicators* 4, 2004, pp. 139-147. http://dx.doi.org/10.1016/j.ecolind.2004.01.004.

[25] O. Külköylüoğlu, and N. Sari, Ecological Characteristics of The Freshwater Ostracoda in Bolu Region (Turkey), *Hydrobiologia* Vol. 688, 2012, pp. 37-46. http://dx.doi.org/10.1007/s10750-010-0585-0.

[26] B.G. Tabachnick and L.S. Fidell, *Using Multivariate Statistics*, Needham Heights., MA: Pearson, (4th ed.) USA, 2001.

[27] J.E. Jackson, *A User's Guide to Principal Components*, John Wiley & Sons, New York, 1991.

[28] K. McGarigal, S. Cushman, S. Stafford, *Multivariate Statistics for Wildlife and Ecology Research*. Springer Science & Business Media, 2013. DOI: 10.1007/978-1-4612-1288-1.

[29] H.C. Romesburg, *Cluster Analysis for Researchers*, Lifetime Learning Publications, Belmont, California, 1984, p.334.

[30] B.S. Everitt, S. Landau, M. Leese and D. Stahl, *Cluster Analysis.* Wiley, New York. 5th ed. 2011, p.728.

[31] P.H.A. Sneath and R.R. Sokal, *Numerical Taxonomy: The Principles and Practice of Numerical Classification*. W. H. Freeman and Company, San Francisco, 1973.

[32] J.H. Ward, Hierarchical Grouping to Optimize an Objective Function, *Journal of the American Statistical Association* 58, 1963, pp. 236–244. DOI: 10.1080/01621459.1963.10500845.

[33] D. Wishart, An Algorithm for Hierarchical Classifications, *Biometrics* 25**,** 1969, pp.165–170. DOI: 10.2307/2528688.

[34] O. Özuluğ, *Trakya Bölgesi Ostrakod (Crustacea) Faunası*. Istanbul University, Institute of Science, PhD. Thesis, 2000, 70p.

[35] R.B. Cattell, The Scree Test for the Number of Factors, *Multiv.Behav.Res*.1, 1966, pp. 245- 276. DOI: 10.1207/s15327906mbr0102_10.

[36] R.D. Ledesma, P.V. Mora, G. Macbeth, The Scree Test and the Number of Factors: A Dynamic Graphics Approach, *Spanish Journal of Psychology* 18, 2015, pp.1–10. DOI: 10.1017/sjp.2015.13.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The author contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of funding for research presented in a scientific article or scientific article itself

No funding was received for conducting this study.

Conflict of Interest

The author has no conflicts of interest to declare.