APPLICATION OF MULTIVARIATE STATISTICAL ANALYSIS IN ECOLOGICAL ENVIRONMENT RESEARCH

Nguyen Thi Hai Ly[1*], Lu Ngoc Tram Anh[2], and Nguyen Ho[1]

[1]Department of Agriculture and Environmental Resources, Dong Thap University, Vietnam
[2]Department of Natural Sciences Teacher Education, Dong Thap University, Vietnam
[*]Corresponding author: Nguyen Thi Hai Ly, Email: nthly@dthu.edu.vn

Article history

Received: 14/09/2020; Received in revised form: 23/12/2020; Accepted: 12/01/2021

Abstract

Multivariate statistics has proven many outstanding advantages and has been used extensively in various studies in the ecological environment field. They supported ecologists to discover the structure and previous relatively objective summary of the primary features of the data. In this paper, some important statistical techniques, including principal component analysis (PCA), canonical correspondence analysis (CCA) and cluster analysis, are explained briefly. Each of them is also examined by a corresponding case-study. The PCA is applied to identify and analyze the relationship between mangrove plant communities and soil factors. Meanwhile, the CCA is put in an application to analyze the relationship between the two sets of species and soil data, from which to determine the effect of soil on the distribution of dominant species. Finally, cluster analysis is examined to analyze the similarities among species in the studied area.

Keywords: *Canonical correlation analysis, cluster analysis, data analysis, ecology, environment, principal component analysis.*

ỨNG DỤNG PHÂN TÍCH THỐNG KÊ ĐA BIẾN TRONG NGHIÊN CỨU SINH THÁI MÔI TRƯỜNG

Nguyễn Thị Hải Lý[1*], Lư Ngọc Trâm Anh[2] và Nguyễn Hồ[1]

[1]Khoa Nông nghiệp và Tài nguyên môi trường, Trường Đại học Đồng Tháp, Việt Nam
[2]Khoa Sư phạm Khoa học tự nhiên, Trường Đại học Đồng Tháp, Việt Nam
[*]Tác giả liên hệ: Nguyễn Thị Hải Lý, Email: nthly@dthu.edu.vn

Lịch sử bài báo

Ngày nhận: 14/9/2020; Ngày nhận chỉnh sửa: 23/12/2020; Ngày duyệt đăng: 12/01/2021

Tóm tắt

Thống kê đa biến có những ưu điểm vượt trội và được ứng dụng trong các nghiên cứu về sinh thái môi trường. Phương pháp này hỗ trợ các nhà sinh thái học tìm hiểu cấu trúc và mô tả một cách tương đối khách quan về các đặc điểm cơ bản của dữ liệu. Trong bài báo này, một số kỹ thuật thống kê quan trọng như phân tích thành phần chính (PCA), phân tích tương quan chính tắc (CCA), phân tích cụm được giải thích tóm tắt. Mỗi kỹ thuật phân tích được khảo sát bởi những nghiên cứu ứng dụng điển hình. PCA áp dụng để xác định và phân tích mối quan hệ giữa quần xã thực vật ngập mặn và các đặc tính thổ nhưỡng. CCA ứng dụng phân tích quan hệ giữa loài và đất nhằm xác định ảnh hưởng của đất đến sự phân bố các loài ưu thế. Phân tích cụm vận dụng để phân tích sự tương đồng của các loài trong khu vực nghiên cứu.

Từ khóa: *Phân tích tương quan chính tắc, phân tích cụm, phân tích dữ liệu, sinh thái học, môi trường, phân tích thành phần chính.*

DOI: https://doi.org/10.52714/dthu.10.5.2021.902

Cite: Nguyen, T. H. L., Lu, N. T. A., & Nguyen, H. (2021). Application of multivariate statistical analysis in ecological environment research. *Dong Thap University Journal of Science, 10*(5), 115-120. https://doi.org/10.52714/dthu.10.5.2021.902.

1. Introduction

The multivariate analysis is well-known as a comprehensive and structured explanation of how to analyze and interpret data observed on many variables (Bui, 2018). However, the application of these methods in the field of ecological environment is still limited. From the ecological point of view, an organism is synthetically affected by a complex set of combination of many environmental factors. Among them, the relationship between species and environmental factors follow the Shelford's law of tolerance and it is not completely linear relationship (Pausas & Austin, 2001) (Figure 1). Therefore, the survey data in natural ecosystems shows both the presence (quantified by the number of individuals) and the absence (number of individuals equals 0) in the surveyed standard plots (Jan Lepˇs & Petrˇ Smilauer, 2003). Accordingly, using traditional univariate linear analysis to discover the relationship between environmental factors and the distribution of species in the ecological studies is not applicable. Based on these views, the paper presents multivariate statistical methods applied in the study of environmental ecology with the support of Canoco ver. 4.5 and Primer ver. 6.0.

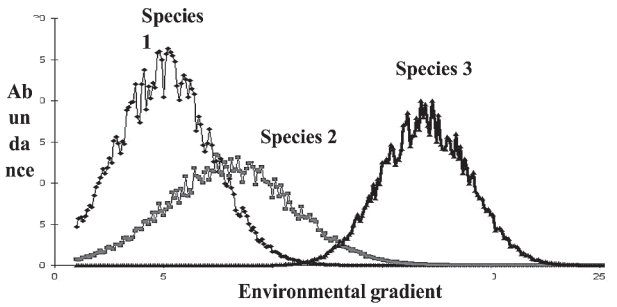

Figure 1. An example of the ability of three species to adapt various environmental gradients (Michael, 2020)

2. Multivariate statistical analysis methods and case studies

2.1. Principal Component Analysis (PCA)

Principal component analysis (PCA) is a dimensionality-reduction method often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one but at the same time minimizing information loss (Steven, 2019). This method groups the analysis objects and helps find out the main factors that will contribute greatly to the fluctuation of the data set. PCA finds a new space in which the coordinate axes in the new space are constructed so that the variance of the data on each axis is greatest. The principle of this technique is quite simple. Firstly, PCA will find out which direction has the most fluctuations in the data set. Then, the horizontal axis will be rotated following that direction and the vertical axis in the perpendicular direction. This aimed to reduce variables that are unnecessary or unimportant factors in the dataset (Bui, 2018). The PCA method analyzes the main components, but the two main ones (PC1 and PC2) are usually selected and will form a model of new plane in space. This plane is a multi-dimensional spatial window (Figure 2) and each observation can be projected onto this plane corresponding to each point. According to Clarke & Gorley (2006), PCA in PRIMER is an ordination, in which the dimensionality of a dataset was reduced, while preserving as much 'variability' (i.e. statistical information) as possible. The samples are regarded as points in multidimensional variable space projected onto the most appropriate plane selected. The researchers can select the number of principal components (new axes), and 2-dimensional or 3-dimensional plots of any combination of these PC's will be presented. PCA has many applications, but the common application in ecological environment studies is to analyze and describe the relationship among environmental factors, the impact of environmental factors on different communities, as well as relationships among species in the natural ecosystem. Besides, this method can be classified into antagonistic organism groups, low antagonists and strong antagonists (Bui, 2018; Jan Lepˇs & Petrˇ Smilauer, 2003).

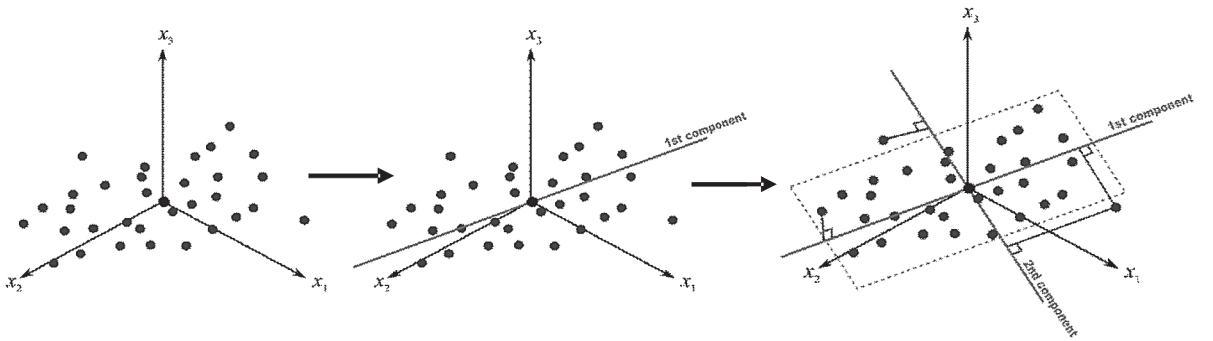

Figure 2. The graph for the new plane model in space by PCA (Kevin, 2020)

As a case study, we applied the PCA technique to identify and analyze the relationship between mangrove plant communities and soil factors in Con Trong, Ngoc Hien district, Ca Mau province. The data sets included mangrove species components recorded in 43 plots; along with environmental variables such as pH, salinity, nitrogen, phosphor and potassium in soil. The results have determined the correlation coefficients in two axes PC1 and PC2 in PCA (Figure 3). In particular, soil pH on the 1st layer (0-20 cm) and the 2nd layer (20-60 cm) were important factors affecting the PC1 axis (with coefficients of -0.443 and -0.475) followed by nitrogen and salinity in the 2nd layer (coefficients are -0.373, -0.424, and 0.366). Phosphor and potassium in the 2nd layer and salinity the 1st soil layer affected the PC2 axis with coefficients of -0.580 and -0.499; 0.341; 0.329, respectively. Taking into account these results, the mangrove communities in Con Trong were divided into 2 groups according to the influence of the soil properties. The 1st group consists of communities with the dominant species of *Rhizophora apiculata* Blume, *Avicennia alba* Blume, *Bruguiera parviflora* (Roxb.) Wight and Arn. ex Griff., was mainly influenced by soil pH, nitrogen, and salinity in the 2nd soil layer. The 2nd group included the mixed communities R. apiculata and *A. alba*, and the community in which *R. apiculata* was the dominant species. These communities were affected by some factors such as the content of phosphor, potassium in the 2nd soil layer and salinity in the 1st soil layer (Lu et al., 2018).

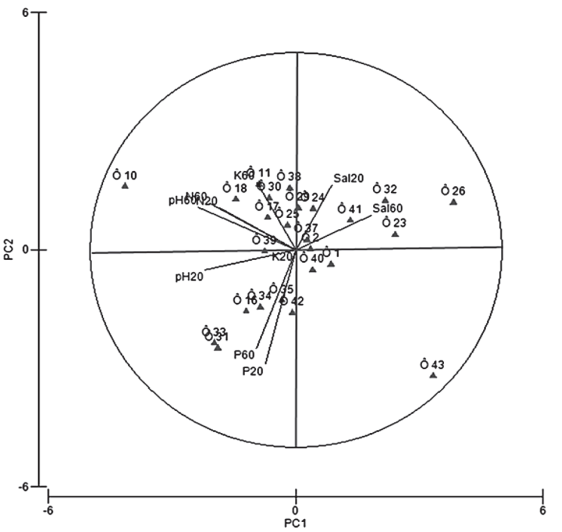

Figure 3. The PCA graph shows the relationship of mangrove plant communities and soil properties. pH: soil acidity; P: phosphor; K: potassium and Sal: salinity; 20: the 1st soil layer (0-20 cm); 60: the 2nd soil layer (20-60 cm)
(Lu et al., 2018)

2.2. Canonical Correlation Analysis (CCA)

Canonical correlation analysis is a multivariate statistical model. It is used to identify and measure the associations among two sets of variables X and Y. This method formulates a set of canonical variables and does not distinguish between independent and dependent variables. From X and Y, the CCA will generate the first two canonical variables W1 and V1, respectively. The results of the CCA will prove the closed or non-closed relationship between the two sets of variables X and Y thanks to the square correlation coefficient of W1 and V1 (Bui, 2018). Besides, CCA also shows the relationship between

variables in the same group of variables and between groups of variables together. Currently, in studies on ecology and biodiversity, CCA is applied in statistical analysis to identify and describe the relationships between species associations and their environmental factors. This method is designed to extract synthetic environmental gradients from environment data sets. The advantages of CCA graphs provide sufficient information about the three objects, namely environmental factors, species composition, and sampling points (Jan Lepˇs & Petr˘Smilauer, 2003).

Another case study aimed to identify environmental factors that affected the distribution and diversity of vascular plants in the opened depression floodplain regions in An Giang province. The research questions were: (1) Do the distribution and diversity of vascular plants vary according to the soil types in the ecological region of An Giang province? (2) Which soil properties determine the distribution and diversity of vascular plants in the ecological region? The CCA method was applied to analyze the relationship between the two sets of species variables *(species.dta)* and soil environment factors *(soil.dta)* to determine which soil environment variables that would most affect the distribution of dominant species on each soil type (Figure 4). Canoco software version 4.5 was used to extract and visualize the influences of soil factors on the dominant species in the studied area (Nguyen, 2020).

Due to the low topography and upstream position in the Vietnam Mekong Delta, the opened depression of floodplain is flooded annually for 3 to 4 months with a depth of inundation over 0.5 m and is characterized by heavy acid sulfate soils. This area consists of three types of soils as acid sulfidic peat soil, active acid sulfate soil with sulfuric materials present topsoil layer from 0 to 50 cm (Near acid sulfate soil), and depth in soil over 50 cm (Deep acid sulfate soil) (Figure 4). Axis 1 describes the characteristics of near acid sulfate soils and deep acid sulfate soils. The deep acid sulfate soil is positively correlated with pHKCl, the amount of silt and sand but inversely correlated with the amount of clay, while near acid sulfate soils have the opposite characteristics. The correlation scores of soil factors with Axis 1 were -0.817 (clay), 0.774 (sand), 0.956 (silt) and 0.999 (pHKCl). On Axis 2, the representation for acid sulfidic peat soil is positively correlated with porosity (correlation score of 0.933). The soil properties of high pHKCl, silt and sand affected the predominant distribution of *Melastoma affine* in deep acid sulfate soil. The soil characteristics of low pHKCl, silt, sand and high clay affected the abundance of *Melaleuca* and *Elaeocarpus hygrophilus*, so they appeared predominantly in near acid sulfate soil. Correlation scores were -0.964 for *Melaleuca cajuputi*, -0.907 for *Melaleuca leucadendra* and -0.897 for E. hygrophilus. The habitat of *Eleocharis* genus was affected by pHKCl. *Eleocharis dulcis* positively correlated pHKCl (0.981), so it dominated in a deep acid sulfate soil. *Eleocharis ochrostachys* positively correlated pHKCl (-0.906), so it dominated in a near acid sulfate soil.

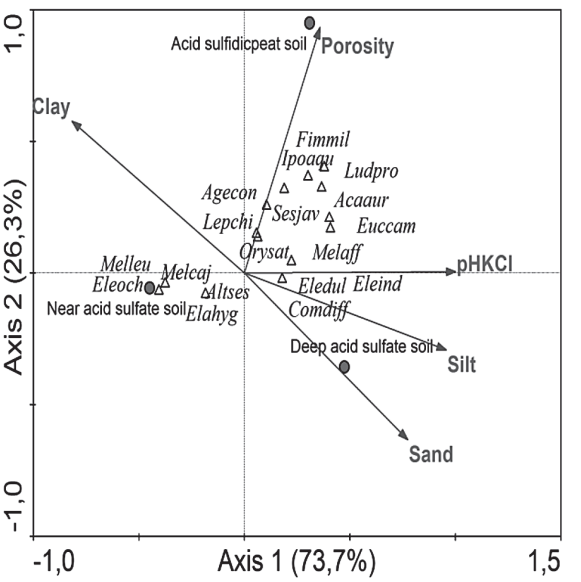

Figure 4. The effect of some soil properties on predominant woody and herbaceous plants in the opened depression floodplain area

(Nguyen, 2020)

Notes: The species component:

Melcaj= Melaleuca cajuputi; Melleu= Melaleuca leucadendra; Elahyg= Elaeocarpus hygrophilus; Sesjav= Sesbania javanica; Melaff= Melastoma affine; Eledul=Eleocharis dulcis; Eleoch=Eleocharis ochrostachys; Ludpro=Ludwigia prostrata; Fimmil= Fimbristylis miliacea; Eleind=Eleusine indica; Ipoaqu=Ipomoea aquatica; Altses=Alternanthera sessilis; Lepchi=Leptochloa chinensis; Agecon=Ageratum conyzoides; Comdiff=Commelina diffusa.

Table 1 shows the eigenvalue decreasing from Axis 1 to Axis 2, of which 73.7% of explanatory variables for Axis 1 and 26.3% for Axis 2. The Pearson correlation coefficients between dominant species and some soil properties in Axis 1 and Axis 2 are 0.940 and 0.607 ($p<0.05$), respectively. The Monte Carlo test results showed that the factors of sand, silt, clay and pHKCl have significantly affected the distribution of predominant woody and herbaceous species in acid sulfate soils (Nguyen Thi Hai Ly, 2020).

Table 1. The results of CCA on the relationship between plant and soil

	Axis 1	Axis 2
Eigenvalue	0.633	0.226
Cumulative variance of species-soil relation (%)	73.7	26.3
Pearson correlation, species-soil relation	0.940	0.607
Monte Carlo test (P-value)	0.002	0.003

2.3. Cluster analysis

To convert the raw data into scientific information, the researchers need to apply the methodology for simplifying data. In statistics, there are two common methods to simplify data: factor analysis and cluster analysis. The factor analysis involves aggregating relevant variables into factors. In contrast, cluster analysis classifies groups of related objects into a representative group of an environmental variable. This analysis method will be effective when objects in the same cluster are closely related and different from other clusters. In the ecology field, cluster analysis is commonly applied to analyze the relationship between species that present in the same ecological environment. Scientifically, the cluster technique will classify species that appear together and have a relatively equal number of individuals into the same group. Based on the individual data of each species in the survey plots, this method will create a distance matrix. Species' medium distance is smaller than that of other species and is classified into one group. The species with a large average distance will be split into other groups (Bui, 2018). Cluster analysis results in a tree diagram that shows the sample groups at different similarities when using Primer software (Clarke & Gorley, 2006). Figure 5 clearly reveals the division of species groups at different levels of similarity in Mui Ca Mau National Park by applying cluster technique to analyze the number of individuals and mangrove species composition. The similarity coefficient between *A. alba* and *R. apiculata* was 63.25, indicating a close correlation between these two species in the studied area. At the 40% similarity, the branching diagram has a group of two species of *X. granatum* and *B. cylindrica* and the group of three species *R. apiculata*, *A. alba* and *B. parviflora*. At the 20% similarity, only two species appeared independently *S. alba* and *X. granatum*. Cluster analysis results showed the distribution of some groups of species or the tendency of random occurrence of some other species in the same environmental conditions in the studied area.

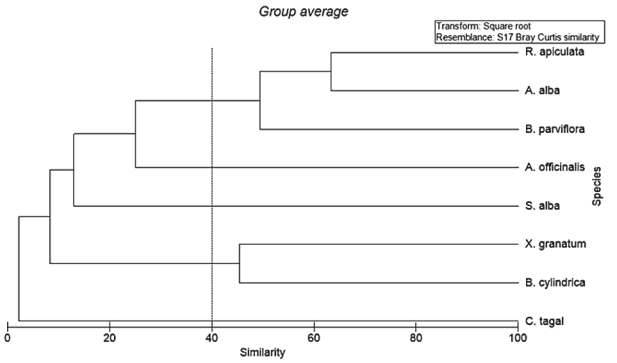

Figure 5. Cluster analysis of mangrove species in Mui Ca Mau National Park

3. Conclusion

The multivariate analysis techniques including PCA, CCA and Cluster analysis show many advantages such as thorough exploitation of data, comprehensive and objective analysis results. Therefore, the application of these into data analysis would help statistical data processing be fast, efficient and accurate. The reliable results from the case studies have demonstrated the effectiveness of the multivariate analysis techniques applied in ecological environment field. These results might be considered as a scientific basis for researchers to make the right and rational judgments and thereby proposing appropriate solutions in the use and management of the environment as well as biological resources.

References

Bui, M. H. (2018). Multivariate analysis methods for forestry research data, using SAS. *Journal of Forestry Science and Technology*, *1*(2018), 43-52.

Clarke, K. R., & Gorley, R. N. (2006). *Primer V6: User Manual/Tutorial*. UK: Primer-E Ltd.

Jan Lepˇs, & Petrˇ Smilauer. (2003). *Multivariate analysis of ecological data using CANOCO*. UK: Cambridge University Press.

Kevin D. (2020). *Principal Component Analysis (PCA)*. Process improvement using data (325 – 370). Ontario: McMaster University. Retrieved from https://learnche.org/pid/.

Lu, N. T. A., Vien, N. N., Nguyen, T. P. T., & Nguyen, T. H. L.. The effects of soil characteristics on mangrove species distribution at Con Trong, Ong Trang estuary, Ngoc Hien district, Ca Mau province. *Can Tho University Journal of Science*, *(54)*, 75-80.

Michael, W. P. (2020). *Ordination methods - An overview*. http://ordination.okstate.edu.

Nguyễn, T. H. L. (2020). *Nghiên cứu sự phân bố và đa dạng thực vật bậc cao trên các vùng sinh thái khác nhau tại tỉnh An Giang*. Trường Đại học Cần Thơ, Việt Nam.

Pausas, J. G., & Austin, M. P. (2001). Patterns of plant species richness in relation to different environments: An appraisal. *Journal of Vegetation Science*, *(12)*, 153-166.

Steven, M. H. (2019). *Principal component analysis (PCA)*. Athens: University of Georgia.