Acta Scientiarum

http://www.periodicos.uem.br/ojs/
ISSN on-line: 1807-8621
Doi: 10.4025/actasciagron.v47i1.74349

MultivariateAnalysis: an R package for multivariate analysis

Ana Luíza Medrado Monteiro* **and Alcinei Místico Azevedo**

Instituto de Ciências Agrárias, Universidade Federal de Minas Gerais, Avenida Universitária, 1000, Bairro Universitário, 39404-547, Montes Claros, Minas Gerais, Brazil. *Author for correspondence. E-mail: analuizamedradom@gmail.com

ABSTRACT. Statistical analysis is essential in research. As modern production processes evolve, the increasing volume of data needing processing has demanded techniques like multivariate analysis for simultaneous data handling. Multivariate analyses are typically complex and often require statistical software. The MultivariateAnalysis package, an R package available on the CRAN platform, was developed to facilitate these analyses. Introduced in 2021 by researcher Alcinei Místico Azevedo, it encompasses techniques such as principal component analysis, principal coordinate analysis, hierarchical clustering, Mantel correlation, dendrograms, canonical variables, dissimilarity measurements, and multivariate variance analysis. This paper aims to detail the MultivariateAnalysis package, offering a practical guide from initial steps to results, enhancing user understanding of the package's functions and potential applications. Its open-source code permits function additions. As of 2024, MultivariateAnalysis has reached version 5.0, featuring enhancements in graphical functions that provide a simple, flexible, and intuitive workspace applicable across various knowledge domains.

Keywords: data analysis; R packages; multivariate analysis.

Received on October 25, 2024.
Accepted on December 4, 2024.

Introduction

Technological advances have significantly transformed processes over recent decades, with agriculture being a standout sector. Many tasks previously performed manually are now automated, such as using satellite imagery to identify weed-infested areas or employing artificial intelligence in agricultural mechanization to aid everything from planting to harvesting. These innovations were propelled by Agriculture 4.0, which has revolutionized Brazilian agriculture. Technological operations extend beyond field activities; they also enhance operational and decision-making processes in production, such as extracting climate-related information in real time (Silva & Cavichioli, 2020).

The evolution of agriculture has also prompted data digitization, leading to a surge in data volumes that need interpretation. This abundance of data impacts decision-making processes. Similarly, the digitization of agriculture has spurred growth in scientific data production. Employing techniques that analyze multiple variables simultaneously has become crucial in decision-making. Multivariate analyses, which examine various measurements on subjects or objects simultaneously, are key statistical techniques (Hair Jr. et al., 2009).

Principal component analysis, discriminant analysis, hierarchical clustering, multivariate variance analysis, and canonical correlations are primary multivariate analysis techniques. These are categorized into five groups: exploring dependencies between variables, prediction, hypothesis construction and testing, data reduction or structural simplification, and grouping of objects or variables (Gouvea et al., 2011). These categories encompass specific techniques that can overlap across categories and can be executed manually or with software. Notable free software includes R (R Core Team, 2019) and its interface, RStudio (Allaire, 2012).

R and RStudio offer significant advantages due to the diversity of available packages on the Comprehensive R Archive Network (CRAN), which aid researchers with various functions and are free of charge. For multivariate analyses, packages such as Vegan (Oksanen et al., 2018), which applies multivariate analyses in ecology, the biotools package (Silva, 2016), with cluster analysis functions, the Candisc package (Friendly & Fox, 2013), featuring multivariate variance analysis, discriminant analysis, and canonical correlation, and the metan package (Olivoto & Dal'Col Lúcio, 2020), for multi-environmental analyses, are available.

While these packages are invaluable for multivariate analyses, they may not encompass all techniques or could have complex routines often requiring multiple packages for a single task. In contrast, the MultivariateAnalysis package, developed in 2021, simplifies multivariate analysis execution. This R package allows for techniques such as hierarchical clustering, principal component analysis, principal coordinate analysis, Mantel correlation, dendrogram, multivariate analysis of variance, distance measures, and canonical variables. Renowned for its simplicity and intuitiveness, it can be applied across various knowledge domains. This paper describes the MultivariateAnalysis package, highlighting its main functions and applicability in Agricultural Sciences.

MultivariateAnalysis package

The core of the MultivariateAnalysis package lies in its ability to implement various multivariate analysis techniques. These include dissimilarity measures, dendrograms, multivariate analysis of variance, principal component analysis, canonical variables, principal coordinate analysis, Tocher clustering, and K-means clustering, among others. Figure 1 below illustrates some graphical results produced using the MultivariateAnalysis package.

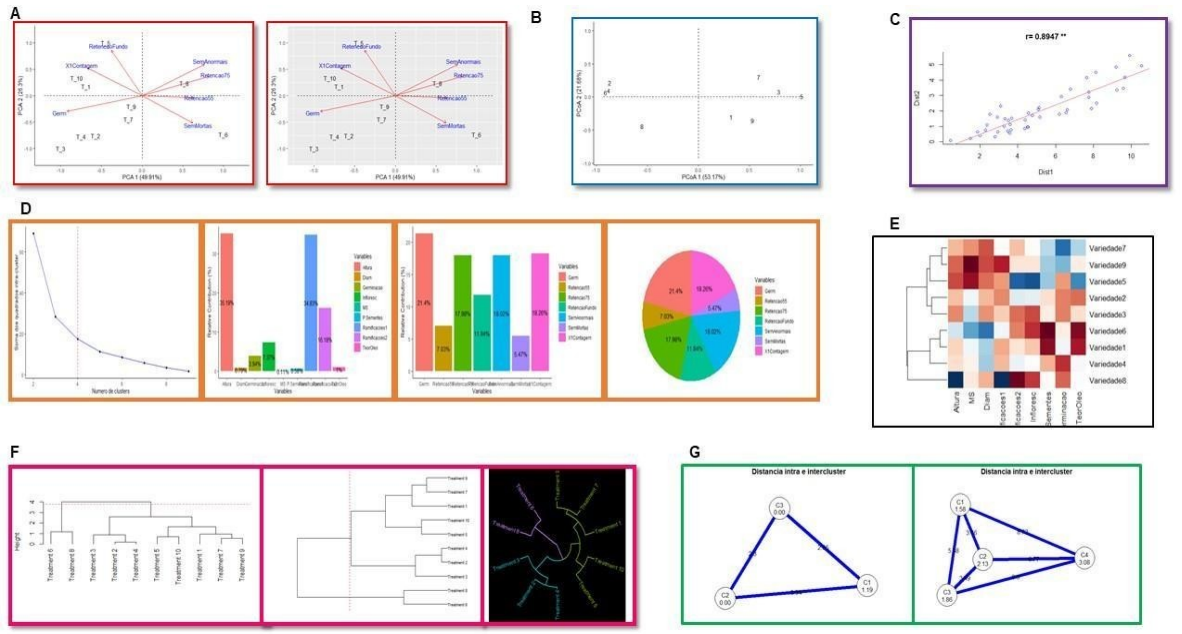

Figure 1. Examples of graphical output from the MultivariateAnalysis package: A) Principal Components, B) Principal Coordinates, C) Cophenetic Correlation of the Tocher Cluster, D) Kmeans Clusters, E) HeatPlot, F) Dendrogram, and G) Intra and inter-cluster distance.

As of 2024, the MultivariateAnalysis package is at version 5.0 and can be downloaded from the CRAN repository at https://cran.r-project.org/web/packages/MultivariateAnalysis. It can be installed directly in R Studio using the command *install.packages* (*MultivariateAnalysis*). To aid users in understanding the package's functions, a playlist of video tutorials is available on the YouTube channel "Alcinei Azevedo – Dicas e Aulas" at https://www.youtube.com/@alcineiazevedo-dicaseaulas9398 and on the website www.expstat.com.

To demonstrate the package's capabilities, the widely utilized Iris dataset Anderson (1935). is employed as an example in software and algorithm applications. To begin operations with MultivariateAnalysis, the package must first be activated using the library(MultivariateAnalysis) command. Subsequently, data for analysis must be imported using the data () function, specifying the dataset. The data tabulation format should be carefully considered depending on the function being executed. For guidance on execution or data formatting, users can refer to examples in the MultivariateAnalysis manual.

Data verification

Demonstrating the capabilities of a software package often involves using previously published data. In this case, the functionalities of the MultivariateAnalysis package are displayed using genetic dissimilarity data between sweet potato genotypes (Andrade et al., 2017). This data set includes molecular data, morphological data, and weighted averages of these matrices.

For dissimilarity measures in MultivariateAnalysis, the molecular data reveal that the closest genotypes are UFVJM-26 and UFVJM-27, while the most distant are Espanhola and UFVJM-42, as indicated in Table 1. This information is also graphically represented in Figure 2. For the morphological data, the closest individuals according to MultivariateAnalysis results are UFVJM-37 and UFVJM-42, and the most distant are T-car 1 and UFVJM-55, as shown in both Table 1 and Figure 2.

Dendrograms derived from these dissimilarity matrices are presented in Figure 2. The molecular data are depicted in Figure 2A, morphological data in Figure 2B, and the weighted average data between the matrices in Figure 2C. Correspondingly, the inter and intra-cluster distances have formed seven groups, illustrated in Figure 3. In MultivariateAnalysis, the formation of seven clusters aligns with the findings of Andrade et al. (2017), confirming the package's precision in generating dendrograms and measuring inter and intra-cluster distances.

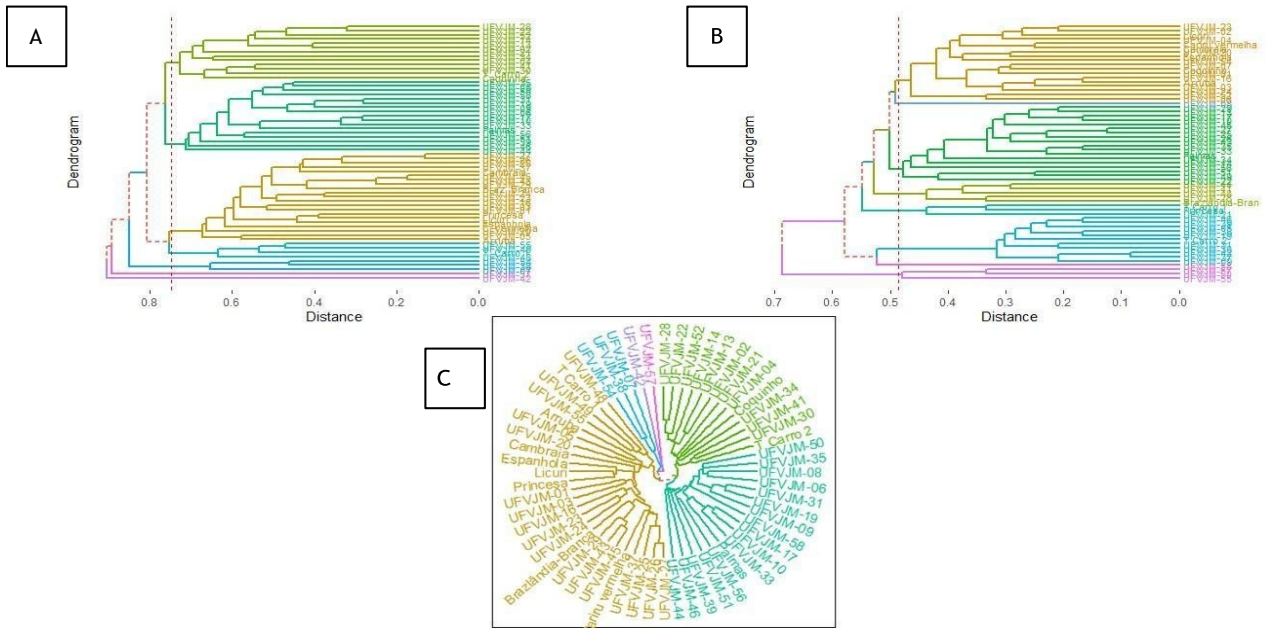

Figure 2. MultivariateAnalysis-derived dendrograms for genetic dissimilarity data of 60 sweet potato genotypes: A) molecular data, B) morphological data, and C) weighted average of the matrices.

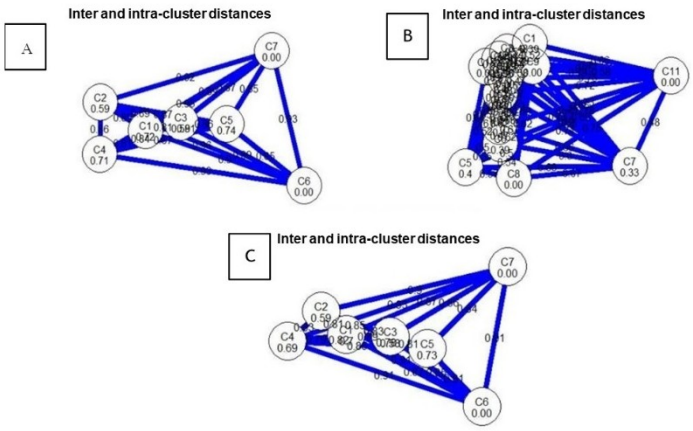

Figure 3. Inter and intra-cluster distances of data from 60 sweet potato genotypes: A)molecular data, B) morphological data, and C) weighted average of the matrices.

Tocher clustering of genetic dissimilarity data between sweet potato genotypes, conducted using MultivariateAnalysis, resulted in the formation of seven clusters. Notably, group one contains the largest number of genotypes within a single group, as detailed in Table 1. This clustering outcome corresponds with findings from the reference article by Andrade et al. (2017). The consistency between the results obtained in this study and those reported by previous authors underscores the precision and accuracy of MultivariateAnalysis in managing such complex datasets.

Table 1. Clustering of 60 sweet potato (*Ipomoea batatas)* genotypes using Tocher's method.

Cluster	Genotypes
1	UFVJM-26, UFVJM-27, UFVJM-25, UFVJM-29, UFVJM-45, UFVJM-43, UFVJM-37 Braz. Branca, UFVJM-24, UFVJM-23, Cambraia, UFVJM-16, UFVJM-03, Princesa, Licuri, Spanish, UFVJM-01, C.vermelha , UFVJM-05, UFVJM-20, Arruba , UFVJM-48, UFVJM-55 UFVJM-49, UFVJM-08, T.Car 1, UFVJM-13, UFVJM-35, UFVJM-58, UFVJM-10, UFVJM-33, UFVJM-17, UFVJM-19, UFVJM-09, UFVJM-06, UFVJM-50, UFVJM-56, Palmas, UFVJM-31, UFVJM-52, UFVJM-22, UFVJM-02, UFVJM-14, UFVJM-28, UFVJM-34, Coquinha, UFVJM-51, UFVJM-39,
2	UFVJM-38, UFVJM-54, UFVJM-07
3	UFVJM-30, UFVJM-41, T. Car 2
4	UFVJM-04, UFVJM-21
5	UFVJM-44, UFVJM-46
6	UFVJM-42
7	UFVJM-57

Dissimilarity measures

MultivariateAnalysis manages dissimilarity measures for qualitative, quantitative, and mixed data, like the GENES software (Cruz, 2016) and functions within packages such as *DisimForMixed* (Pathberiya, 2016), *distantia* (Benito & Birks, 2020), and *TSclust* (Montero & Vilar, 2015). The primary function for calculating distances, called *Distancia*(), takes the data as its first argument and the chosen model as the second, as illustrated in Figure 4.

Results from this function include a dissimilarity matrix and metrics such as the smallest and largest distances, mean and range of distances, standard deviation, and the coefficient of variation of distances. It also identifies the closest and most distant individuals. For further guidance, users can refer to examples with the command*? Distancia*. Regarding data types, for binary or multi-categorical qualitative data, the dataset may include numbers or texts to assess the similarity or dissimilarity among individuals. Binary data consists solely of 0s and 1s, while mixed data incorporates both texts and numbers. The dissimilarity indices available in the MultivariateAnalysis package are depicted in Figure 4.

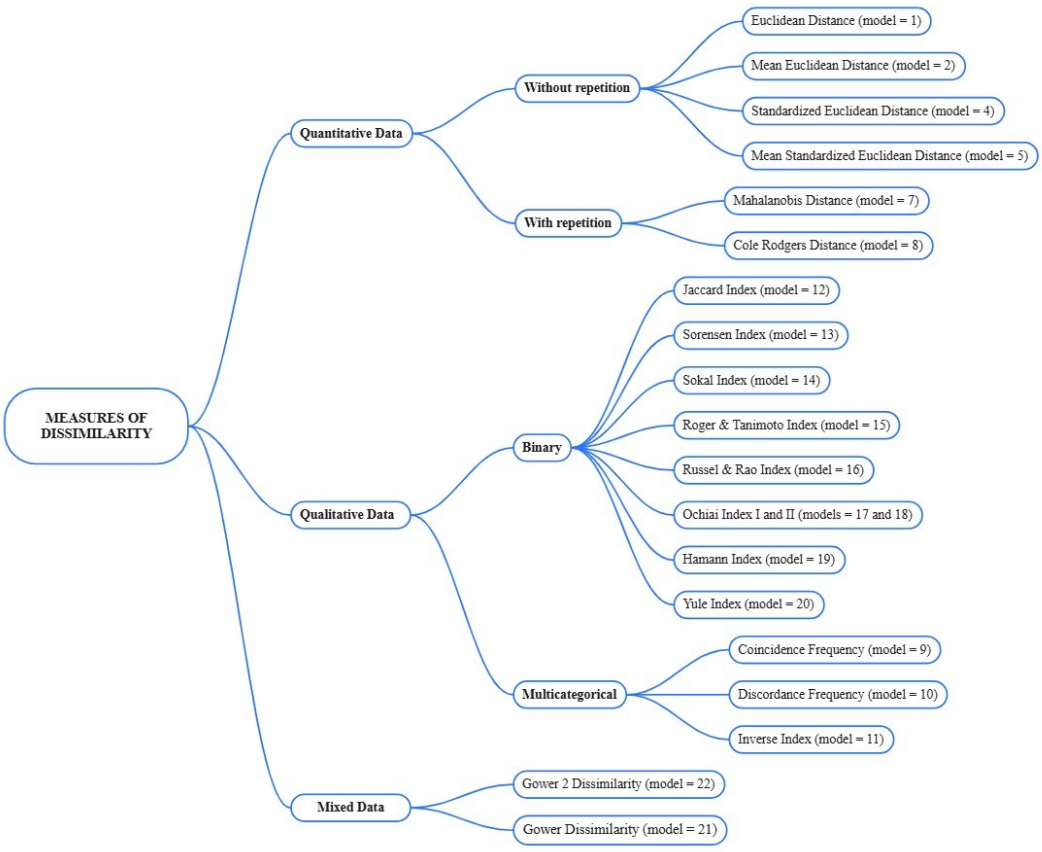

Figure 4. Dissimilarity measure models by the MultivariateAnalysis package

Analysis with dissimilarity matrix

In MultivariateAnalysis, the functions *SummaryDistance()*, *Dendrogram()*, *Tocher()*, and *CoordinatesPrincipal()* utilize information from dissimilarity matrices. The *Dendrogram()* function, specifically, can also be executed using other packages such as *factoextra* (Kassambara & Mundt, 2020), *dendextend* (Galili, 2015), *DendSer* (Hurley & Earle, 2022), and *ggdendro* (Vries et al., 2024). Within MultivariateAnalysis, *Dendrogram()* requires four arguments: first, the data; second, the method; third, the layout; and fourth, the type of cut. The available methods and the types of analysis that can be conducted with dissimilarity matrices are illustrated in Figure 5.

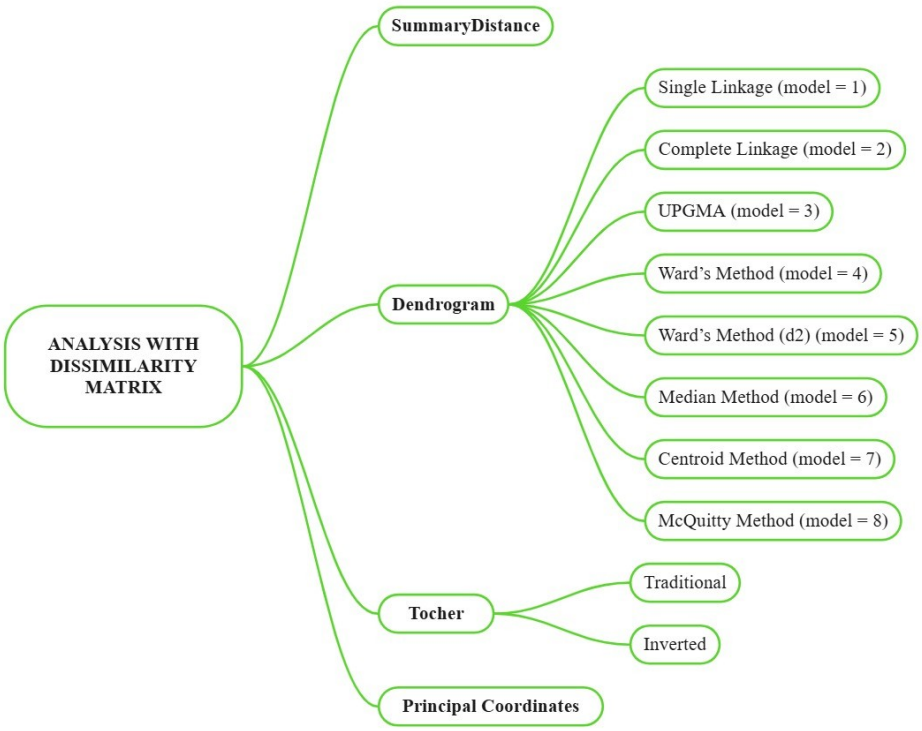

Figure 5. Types of analysis conducted using a dissimilarity matrix using the MultivariateAnalysis package.

The dendrogram results in MultivariateAnalysis include a graphical representation, an estimate of cophenetic correlation, significance testing via the Mantel test, cut-off criteria, and the clusters formed. The package allows for the specification of the dendrogram cut-off point using various methods such as Mojena, Cindex, Frey, McClain, and Dunn.

To execute Tocher's method in MultivariateAnalysis, there are two available versions: the original and the sequential (Figure 5). The Tocher() function requires two main arguments: the first being the dissimilarity matrix and the second specifying the method—either sequential or original. Additional arguments like *xlab*, *ylab*, and *bty* can modify the graphical representation of the *x*-axis, *y*-axis, and plot borders, respectively, in the scatter plot of the cophenetic correlation.

Principal Coordinate Analysis (PCoA) is a generalization of principal component methods that utilizes eigenvalues and eigenvectors derived from a dissimilarity matrix (Jongman et al., 1995). This analysis can be performed using the vegan package (Oksanen et al., 2018) as well as by MultivariateAnalysis. The required arguments for this function include the dissimilarity matrix, the layout style, and a plotting argument (plot=T). Figure 6 displays the results of the principal coordinate analysis for the genetic divergence data of 60 sweet potato genotypes.

In MultivariateAnalysis, dispersion measures include Principal Components Analysis (PCA), Principal Coordinates Analysis (PCoA), and Canonical Variables. Principal component analyses can also be performed using other packages like FactoMineR (Lê et al., 2014), ade4 (Dray & Dufour, 2007), and amap (Lucas, 2014). The execution within MultivariateAnalysis is facilitated by the function *ComponentesPrincipais()*. This package is noted for the ease of executing functions, intuitively named commands, and data accuracy.

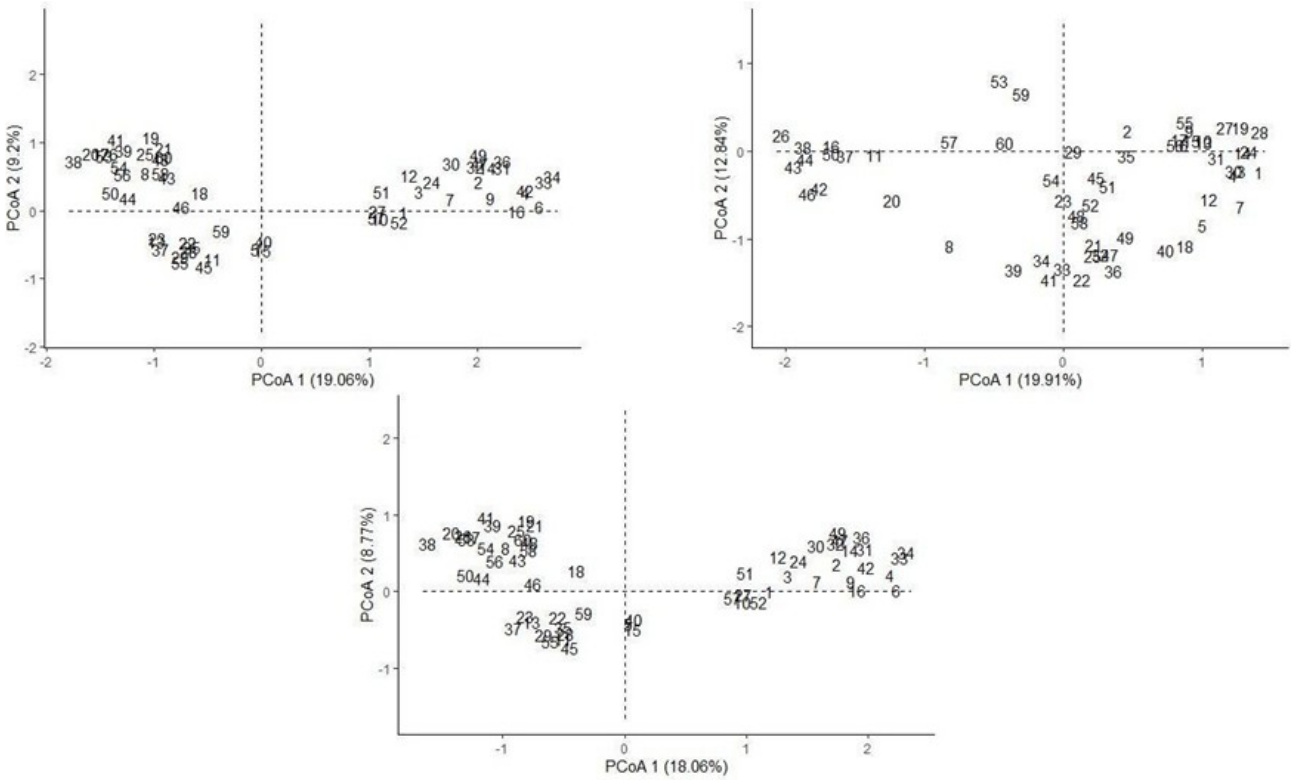

Figure 6. Principal coordinate analysis results for Iris data using MultivariateAnalysis package.

Dispersion methods

Canonical Variables analysis in MultivariateAnalysis accommodates various experimental designs, as illustrated in Figure 7. Other packages, such as CCA (González & Déjan, 2013) and yacca (Butts, 2022), also perform canonical variable analyses, each offering unique features.

The function for executing canonical variables, *CanonicalVariables()*, is straightforward to use. It requires the data and the model related to the experimental design as arguments. Additional optional arguments enhance the graphical quality of the outputs, such as *CorCol*, *CorPlot*, *VarCol*, and *bty*, which adjust the colors, presence of arrows, text colors, and borders, respectively. Besides providing a principal components graph, the results in MultivariateAnalysis also include detailed explanations, scores, and their significance.

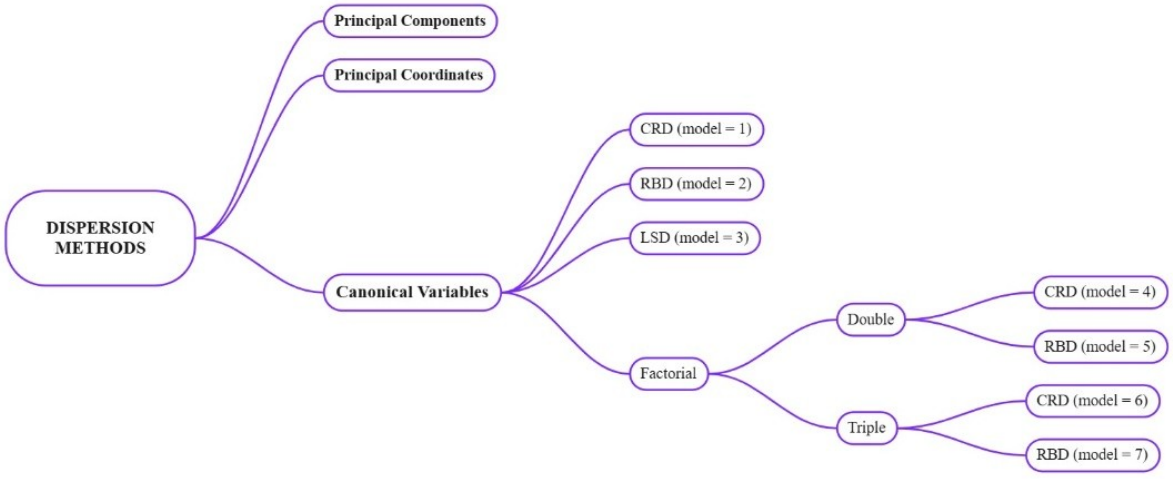

Figure 7. Performance of dispersion methods using MultivariateAnalysis.

Multivariate analysis of variance

Among multivariate data analysis techniques, the multivariate analysis of variance (MANOVA) is particularly notable. According to Fávero et al. (2009), MANOVA can be defined as an extension of the

univariate analysis of variance, incorporating multiple dependent variables. This technique, as Hair Jr. et al. (2009) note, is used to compare groups across these multiple dependent variables.

In MultivariateAnalysis, MANOVA can be conducted with single, double, or triple factors across any design, as shown in Figure 8.

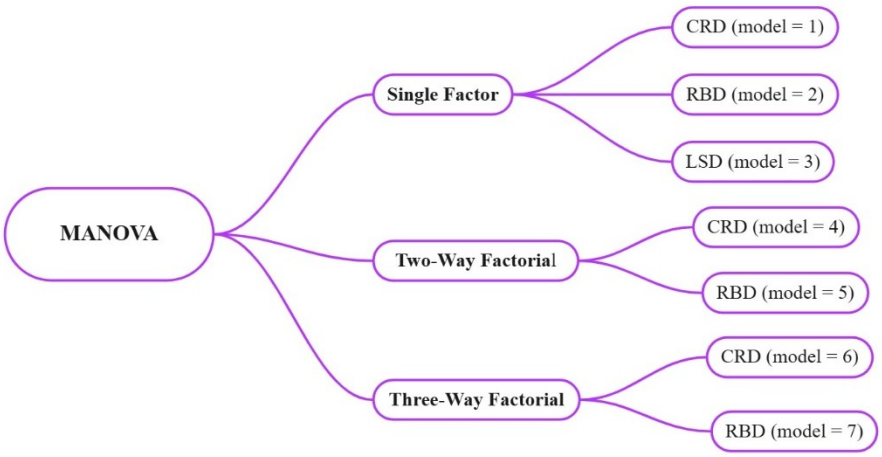

Figure 8. Performance of multivariate analysis of variance using MultivariateAnalysis.

In addition to MultivariateAnalysis, there are several other R packages capable of performing multivariate analysis of variance (MANOVA), including the car package (Fox & Weisberg, 2012) and Candisc (Friendly & Fox, 2013). A key feature of MultivariateAnalysis is its simplicity in executing analyses, which extends to MANOVA. The process typically involves just two main steps: data importation and the execution of the *MANOVA()* function, which requires arguments for data identification along with the selected factorial and design model. Below, is an example of a single factor under a completely randomized design (CRD).

Conclusion

The MultivariateAnalysis package was specifically developed to make performing multivariate analyses easier, more effective, and highly intuitive, thus streamlining statistical analysis tasks for researchers. This package consolidates various functions, previously dispersed across multiple other packages, into a single comprehensive toolkit tailored for MultivariateAnalysis needs.

As of 2024, MultivariateAnalysis has reached version 5.0 and is currently being updated to include English translations, broadening its accessibility to a global audience of researchers. The package is freely available through the CRAN repository, offering transparent and reproducible records that not only facilitate current analytical tasks but also support the integration of new functions soon.

Data availability

The information regarding the package is available on the website: https://www.expstat.com/pacotes-do-r/multivariateanalysis.

References

Allaire, J. J. (2012). *RStudio: integrated development environment for R*. *770*(394), 165-171. https://posit.co/download/rstudio-desktop/.

Anderson, E. (1935). *Iris Data- R*. https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/iris.html

Andrade, E. K. V., Andrade Júnior, V. C., Laia, M. L., Fernandes, J. S. C., Oliveira, A. J. M., & Azevedo, A. M. (2017). Genetic dissimilarity among sweet potato genotypes using morphological and molecular descriptors. *Acta Scientiarum. Agronomy*, *39*(4), 447-455. https://doi.org/10.4025/actasciagron.v39i4.32847

Benito, B. M., & Birks, H. J. B. (2020). Distantia: An open-source toolset to quantify dissimilarity between multivariate ecological time series. *Echography*, *43*(5), 660-667. https://doi.org/10.1111/ecog.04895

Butts, C. T. (2022). *Maintainer Carter T*. Package 'yacca'.

Cruz, C. D. (2016). Genes Software – extended and integrated with the R, Matlab and Selegen. *Acta Scientiarum. Agronomy*, *38*(4), 547-552. https://doi.org/10.4025/actasciagron.v38i4.32629

Silva, A. R. (2016). *Package ' biotools '*. https://CRAN. R-project. org/package=~ biotools

Vries, A., & Ripley, B. D. (2024). *ggdendro: Create dendrograms and tree diagrams using 'ggplot2'*. R package version 0.2.0. https://andrie.github.io/ggdendro/

Dray, S., & Dufour, A.-B. (2007). The ade4 package: implementing theduality diagram for ecologists. *Journal of Statistical Software*, *22*(4), 1-20. https://doi.org/10.18637/jss.v022.i04

Fávero, L. P. L., Belfiore, P. P., Silva, F. L., & Chan, B. L. (2009). *Análise de dados: modelagem multivariada para tomada de decisões*. Elsevier.

Fox, J., Weisberg, S., Adler, D., Bates, D., Baud-Bovy, G., Ellison, S., Firth, D., Friendly, M., Gorjanc, G., & Graves, S. (2012). *Package 'car'*, 16. R Foundation for Statistical Computing. https://cran.r-project.org/package=car/car.pdf

Friendly, M., & Fox, J. (2013). *candisc: Visualizing generalized canonical discriminantand canonical correlation analysis*. R package version 0.6-5. http://CRAN.R-project.org/package=candisc

Galili, T. (2015). endextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), 3718-3720. https://doi.org/10.1093/bioinformatics/btv428

González, I., Dejean, S., Martin, P. G. P., & Baccini, A. (2008). CCA: An R Package to extend canonical correlation analysis. *Journal of Statistical Software*, *23*(12), 1-14. https://doi.org/10.18637/jss.v023.i12.

Gouvea, M. A., Prearo, L. C., & Romeiro, M. C. (2011). Avaliação do emprego da técnica de análise multivariada de variância em teses e dissertações dos programas de pós graduação em administração da Universidade de São Paulo e da Universidade Federal do Grande ABC. *Revista Estudos do CEPE*, *34*, 69-97

Hair Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., Tatham, R. L., Gouvêa, M. A., & Sant'Anna, A. S. (2009). *Análise multivariada de dados*. Bookman.

Hurley, C. B., & Earle, D. (2022). *DendSer: Dendrogram seriation: Ordering for visualisation*. https://doi.org/10.32614/CRAN.package.DendSer.

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1), 1-18. https://doi.org/10.18637/jss.v025.i01

Jongman, R. H. G., Ter Braak, C. J. F., & van Tongeren, O. F. R. (1995). *Data analysis in community and landscape ecology*. Cambridge University Press. https://doi.org/10.1017/CBO9780511525575

Kassambara, A., & Mundt, F. (2020). *Package ' factoextra '. Extract and visualize the results of multivariate data analyses*. R Package Version 1.0.7 https://CRAN.R-project.org/package=factoextra

Lucas, A. (2014). *amap: Another multidimensional analysis package*. http://CRAN. R- project. org/package= amap

Montero, P., & Vilar, J. A. (2014). TSclust: An R package for time series clustering. *Journal of Statistical Software*, *62*(1), 1-43. https://doi.org/10.18637/jss.v062.i01

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O'Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2018). *vegan: Community ecology package*. R package version 2.5-2. https://CRAN.R-project.org/package=vegan

Olivoto, T., & Dal'Col Lúcio, A. (2020). metan: An R package for multi-environment trial analysis. *Methods in Ecology and Evolution*, *11*(6), 783-789. https://doi.org/10.1111/2041-210X.13384

Pathberiya, H. A. (2016). *Calculate dissimilarity matrix for dataset with mixed atributes*. https://cran.r-project.org/web/packages/DisimForMixed/DisimForMixed.pdf

R Core Team. (2019). *A: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-proje ct.org/

Silva, J. M. P., & Cavichioli, F. A. (2020). O uso da agricultura 4.0 como perspectiva do aumento da produtividade no campo. *Interface Tecnológica*, *17*(2), 616-629. https://doi.org/10.31510/ infa.v 17i2.1068