Mini Review

Open Access

The use of multivariate statistics in breeding: discussions of common mistakes

Abstract

Predictive methods of genetic diversity have been widely used in recent years mainly through the use of multivariate statistical methods, since with these methods the multiple information of each cultivar is expressed in measures of dissimilarity, which represent the existing diversity in the set of analyzed accesses. However, there are several errors and mistakes that can be made in the wrong choice of one of the methods that cover multivariate analysis. Two of them are: wrong choice of the hierarchical grouping method and the error in deciding on the ideal number of groups in the final stage of cluster analysis.In this context, the main objective of this article is to discuss the main questions that occur in the analysis of the groupings obtained, suggesting some solutions.

Keywords: cluster, hierarchical methods, genetic diversity

Volume 4 Issue 3 - 2020

Priscila Neves Faria
Faculdade de Matemática, Núcleo de Estatística, Universidade Federal de Uberlândia, Brazil

Correspondence: Priscila Neves Faria, Faculdade de Matemática, Núcleo de Estatística, Universidade Federal de Uberlândia, Brazil, Email priscilanves@ufu.br

Received: May 11, 2020 | **Published:** June 18, 2020

Introduction

The knowledge of genetic variability in a species is crucial for establishing protection of cultivars in a breeding program. According to Carvalho et al.[1] for the cultivar protection process to be effective, the responsible breeder must know the behavior of the species, groups and varieties.

For example, to determine how far genetically one population or genotype is from another, biometric methods are used, which are analyzed using multivariate statistics, allowing the unification of multiple information from a set of characteristics. Several methods can be used, among them is cluster analysis.

For reasonable accuracy and unbiased estimates of genetic diversity, adequate attention should be paid to sampling strategies; use of various data sets based on an understanding of their dimensions and constraints; choice of genetic distance measures, clustering procedures and other multivariate methods in data analysis.[2]

There are several clustering techniques proposed in the literature to analyze genetic diversity, however, the most used are those that involve so-called agglomerative hierarchical methods.

The hierarchical methods consist of a series of successive groupings or successive divisions of individuals, where they are aggregated or disaggregated until the formation of the graph, in the form of a classification tree, called dendrogram, which presents an objective synthesis of the results and each branch being represents an individual.

The concept of hierarchical data representation was first developed in biology.[3] Currently, hierarchical methods have been used in several areas of agronomy and bring excellent results in horticulture research. Research in this area is increasingly common in assessing the diversity between genotypes, in order to determine the most similar ones through morpho-agronomic characters, such as green beans,[4] tomatoes,[5] lettuce[6] and peppers.[7]

The delimitations can be established through a visual examination of this graph, in which high-level points of change are evaluated, considering them in general as delimiters of the number of genotypes for a given group.[8]

However, many of the published works do not justify the reason for choosing the methods used, and several times different coefficients were used for the same purpose. The consequence of this is that the different methodologies for calculating data obtained from agronomic and morphological characters can be a source of divergence in the relationship between the measures of evaluation of genetic diversity.

In assessing the diversity among a group of genotypes, the analysis of graphic dispersion, usually using two-dimensional space, brings a large number of important information.

Materials and methods

Common mistakes when applying cluster analysis

A problem in analyzes based on graphical dispersion is the difficulty of establishing similarity groups in a less subjective way, based on simple visual inspection of the dispersion, but, however, this can be overcome with the use of the Cophenetic Correlation Coefficient, widely used in evaluation of the degree of distortion of the dendrogram graph (tree diagram).

Sokal and Rohlf,[9] created the Cophenetic Correlation Coefficient using Pearson's correlation idea. This coefficient calculates measures of the degree of adjustment between the Phenetic Matrix (distance matrix) and the Copenetic Matrix (matrix obtained through the Dendrogram). This degree of adjustment is given in percentage, and the higher this percentage, the greater the consistency of the groupings for the applied hierarchical method. According to Sokal and Rohlf,[9] correlation values equal to or greater than 0.8 are considered good. Otherwise, the ideal is to choose another hierarchical method.

This information, being applied correctly in the study, is an important tool to detect genetic variability.

Most of the times when multivariate analysis is applied to genetic diversity data, the question arises as to which of the existing hierarchical methods would be the most appropriate. The reason is that each of these methods generates a different type of grouping, that is, a different graph of dendrogram, with distortions between the dissimilarity patterns of the individuals studied and a high simplification of the original information of the data.

Mistakes in choosing the ideal number of groups

Another difficulty in the final stage of studies using hierarchical grouping methods is the lack of objective criteria to identify the ideal number of groups formed, because in practice this number is given simply by a visual graphic inspection or established at points of high change in dendrogram level.[10] The wrong decision on the ideal number of groups can cause several errors of conclusion in the results obtained. Milligan and Cooper[11] tested a total of 30 cluster validation measures in order to determine the ideal number of groups for each of the measures used in the hierarchical clustering process and showed that different criteria can lead to very different results.

According to Milligan & Cooper,[11] when applying an index to determine the ideal number of groups, two types of decision errors can occur: the first type of error occurs when the index indicates the selection of g groups, and in fact there are less than g groups in the data set. The second type of error occurs when the index indicates fewer groups in the data set than the real one. Even though the severity of the two types of errors may change according to the context of the problem, the occurrence of the second type of error is considered more serious in most analyzes, as there will be loss of information by merging different groups. According to Everitt,[12] the two best performances in the study by Milligan & Cooper[11] were the techniques introduced by Calinski & Harabasz[13] and Duda & Hart[14] for use in continuous data.

Results

In horticulture, it has been common to find works using comparisons between clustering results have been performed with the cophenetic correlation.[15–17] Cargnelutti filho et al.[16] mentioned in their study that the hierarchical grouping method UPGMA shows greater consistency of the cluster pattern when evaluating cluster of bean cultivars, obtained from the dissimilarity matrix.

Santos et al.[18] studied 11 intraspecific accessions of peanuts, belonging to the subspecies hypogaea and fastigiata and used the dendrogram to obtain of the genetic similarity matrix. For the grouping of lineages, the hierarchical method of the arithmetic mean between unweighted pairs (UPGMA) was used. The dendrogram allowed the formation of three groups among the evaluated genotypes. From these results, it was possible to choose the most divergent genotypes for hybridization work.

Faria et al.[7] compared methods of grouping in forty-nine accessions of pepper of the species Capsicum chinense, defining an ideal number of groups in the dendrogram obtained at the end of the study. within each group and maximum heterogeneity between groups.

According to Yorinori and Kiihl,[19] genetic diversity is the greatest guarantee of the stability of production, productivity and the survival of humanity. However, making mistakes in this process of analyzing genetic diversity has serious consequences for research, impairing the process of perceiving the genetic variability of populations.

Ravanfar et al.[20] studied the genetic diversity for transgenic and non-transgenic broccoli plants. The similarity values found were used to generate a dendrogramvia method UPGMA. The correlation between the similarity matrix and the cophenetic matrix for the clusters was computed.

Acknowledgments

None.

Conflicts of interest

Authors declare no conflict of interest exists.

References

1. Carvalho LP, Lanza MA Fallieri J, et al. Analysis of the genetic diversity among accessions of cotton germplasm collection. *Pesq agropec. bras* 2003;38(Pt 10):1149–1155.
2. Mohammadi SA, Prasanna BM. Analysis of genetic diversity in crop plants – Salient statistical tools and considerations. *Crop Sci.* 2003;(43):1235–1248.
3. Mardia KV, Kent JT, Bibby JM. *Multivariate analysis.* London: Academic Press, 1979;521p.
4. Abreu FB. Divergência genética entre acessos de feijãode-vagem de hábito de crescimento indeterminado. *Horticultura Brasileira* Brasília, 2004;22(Pt3):547–552.
5. Karasawa M. et al. Cluster analysis in quantifying genetic divergence in tomato accessions. *Horticultura Brasileira* Brasília, 2005;23(Pt4):1000–1005.
6. Azevedo AM. Seleção de genótipos de alface para cultivo protegido: divergência genética e importância de caracteres. *Horticultura Brasileira, Brasília*. 2013;31(Pt 2):260–265.
7. Faria PN, Paulo R, Anderson R, et al. Métodos de agrupamento em estudo de divergência genética de pimentas. *Horticultura Brasileira.* 2012;(30):428–432.
8. Cruz CD, Carneiro PCS. Modelos biométricos aplicados ao melhoramento genético. Editora UFV, Viçosa, 2003:623.
9. Sokal RR, Rohlf FJ. The comparison of dendrograms by objective methods. *Taxon*, Berlin, 1962;11(Pt1):30–40.
10. Milligan GW. A monte carlo study of thirty internal criterion measures for cluster analysis. *Psychometrika*, 1981a; (46):187–199.
11. Milligan GW, Cooper MC. An examination of procedures for determining the number of clusters in a data set. *Psychometrika.* 1985;(50):159–179.
12. Everitt BS, Sabine L, Morven L. *Cluster Analysis.*5th Edition.John Wiley& Sons, Ltd, Chichester, UK. 2011.
13. Calinski T, Harabasz JA. Dendrite method for cluster analysis, communications in statistics – Theoryand Methods. 1974;3(1).
14. Duda RO, Hart PE. *Pattern classification and scene analysis.*New York: Wiley. 1973.
15. Gonçalves LSA, R Rodrigues, AT Amaral, et al. Comparison of multivariate statistical algorithms to cluster tomato heirloom accessions. *Genetics and Molecular Research*, 2008;(7):1289–1297.
16. Cargnelutti F, Nerinéia D, Cláudia B. et al. Consistência do padrão de agrupamento de cultivares de feijão conforme medidas de dissimilaridade e métodos de agrupamento. *Pesquisa Agropecuária Brasileira.* 2010;(45):236–243.
17. Cargnelutti filho A, Guadagnin JP. Consistência do padrão de agrupamento de cultivares de milho. *Ciência Rural*. 2011; (41):1503–1508.
18. Santos RC, Godoy IJ, Fávero PA. Melhoramento do amendoim e cultivares comerciais. In: Santos RC. dos; Freire RMM, Lima LM. O Agronegócio do amendoim no Brasil. Embrapa, 2ª ed; 2013;116–184.
19. Yorinori JT, Kiihl RA. Melhoramento de plantas visando resistência a doenças. In: Nass LL, Valois ACC, editors. IS. *Recursos genéticos e melhoramento: plantas.*Rondonópolis: Fundação MT, 2001;715–736.
20. Ravanfar SA, Aziz MA, Shabanimofrad M, et al. Greenhouse evaluation on the performance of heat tolerant transgenic broccoli and genetic diversity analysis using inter simple sequence repeat (ISSR) markers. *Electronic Journal of Biotechnology,* 2013;(16):1–10.

Citation: Faria PN. The use of multivariate statistics in breeding: discussions of common mistakes. *Horticult Int J.* 2020;4(3):91–92.
DOI: 10.15406/hij.2020.04.00164