Application of hierarchical cluster techniques in the taxonomic analysis of plant germplasm

Osmany Molina Concepción1 Marilys Milián Jiménez2 Carmen C Pons Pérez1

1Bioinformatics Department, Research Institute of Tropical Root and Tuber Crops, Bananas, and Plantains (INIVIT), Cuba.

2Science Technologyand Innovation Department, Central University “Marta Abreu”, Las Villas (UCLV), Cuba.

Author for correspondence: Osmany Molina Concepción ([taxonumeric@inivit.cu)](mailto:fernan@us.es)16)

Orcid: <https://orcid.org/0000-0003-4199-6652>

**Abstract**

Nowadays, there is a limited knowledge of different analysis types, techniques and statistical tools in the numerical taxonomy, available for a more reliable processing in the genotype classification. To solve this problem partially, a study was conducted on four combinatorial data strategies, distance measures, agglomeration and validation methods to determine those clusters that best fit the characteristics of cocoyam (*Xanthosoma* spp.) germplasm. Functions implemented on the basis of the R programming language were used for the analyses. A detailed knowledge of unsupervised hierarchical clustering methods, which best influence the classification results of cocoyam germplasm individually and in its combination is available for the first time. In addition to, the strength of the combination of statistical techniques used in the conformation of consistent structures, as well as in the validation of these ones was shown. Because of its flexibility, this analysis can be applied to other classification studies in plant germplasm banks.

**Keywords:** classification, cocoyam, germplasm, numerical taxonomy, statistical technique

**Introduction**

The study of phytogenetic resources has become a scientific priority, especially those ones with a little study and commercial potential, which makes important the study of this diversity through quantitative methods that help to group populations of the same genus or species by means of similarities of homologous characters (Núñez et al. 2004).

Numerical methods are mainly used to study and calculate the genetic distance among taxonomic entities. For this purpose, different coefficients of similarity or dissimilarity are used, depending on the considered character and the correspondence degree between the phenotype and the genotype. Numerical methods are mathematical procedures that make it possible to quantify the resemblance among taxonomic units, based on the levels of the observed characters (Ruiz De Galarreta 1998).

In general terms, the numerical taxonomy attempts to construct "natural" classifications, based on the phenotypic similarity of the individuals being evaluated, starting from an appropriate choice of a similarity coefficient (Cuadras 1981).

Several unsupervised classification algorithms capable of performing such a task have been developed in recent years: hierarchical and non-hierarchical, which provide solutions to a considerable number of problems in the biological sciences, although they often do not produce objective or stable classifications, since different groupings can be obtained from the same data matrix if different algorithms are used, as well as if new characters are found, the distances will change and the previous classification will change in parallel.

To solve this problem partially, the idea of Clustering Ensemble has been suggested (Vega-Pons and Ruiz 2010a, b; Vega-Pons et al. 2011). This responds to the intuitive idea that, if the quality of certain individual results is not known, the option of combining them may be higher to selecting a single result.

Despite in recent years, important efforts have been made to record and store morphoagronomic and molecular characterization data of different accessions of the main cultivated species, currently, there is a limited knowledge of different types of analyses, techniques and statistical tools, available for a more reliable processing of genotype classification. This is of vital importance for its application in agricultural processes in the fulfillment of the objectives of researchers, geneticists and botanists.

Several existing taxonomic studies were carried out, especially at the Inivit, based on geneticists´ experiences and the use of isolated and not integrated methods in the analysis of the object as a whole. It is necessary to perform studies that integrate the different variables that make up the object, the combination of different statistical classification techniques, which is indispensable to obtain results with higher levels of accuracy and precision to the performance of these techniques separately, to find a better taxonomic classification of the accessions present in cocoyam germplasm (Milián et al. 2018).

It is an experimentally demonstrated fact and a generally accepted philosophical consideration that there is no an only one computational tool that is optimal in solving any type of classification. Particularly, the classification of some plant species is considered ambiguous at the international level and nationally there are limitations in the correct classification of the new accessions that are introduced to the plant germplasm bank every year (Milián et al. 2018).

In this context, taking into account the fact that statistics is a science allied to scientific research, an analysis structure is traced, based on a set of techniques to determine those clusters that best fit the characteristics of cocoyam germplasm (*Xanthosoma* spp.) from its study on four combinatorial strategies of data, distance measures, agglomeration and validation methods.

**Materials and methods**

Four combinatorial variants were carried out with the available data to determine the clusters that best fit the characteristics of the germplasm under study. In the first variant, a study of Gower's metric (Gower 1971) was carried out for mixed variables, where qualitative variables (nominal and ordinal) and quantitative variables are integrated. In the second variant, qualitative variables were integrated with coded quantitative variables (according to Minimum Descriptor Systems) treated as ordinals.

In variant three, the transformed qualitative variables were integrated, through the multiple correspondence análisis (Tenenhaus and Young 1985) with the *MCA*() function described in the *FactoMineR* package, with the normalized quantitative variables, for which the *data.Normalization*() function of the *clusterSim* package was used with transformation by Z (*z-score*). Later, both results were merged into a single continuous data matrix. In the fourth variant, the qualitative variables received a similar procedure to the one used in variant three and, after standardization, the quantitative variables received an analysis of main components, using the *precomp*() function of the *stats* package. The components with contributions greater than one were taken (Kaiser 1974; Piggot and Sharman 1986; Jackson 1993). With the union of both results, a continuous data matrix was formed.

In this grouping analysis, a group of basic procedures were defined to carry out the study on the four combinatorial variants of data, which are described below:

***Databases***

To conduct the research, data from a study of accessions of a cocoyam (*Xanthosoma* spp.) test collection (71 accessions, where 20 qualitative variables and 16 quantitative variables were evaluated (Milián 2008; Milián et al. 2018) from the germplasm bank, stored at the Research Institute of Tropical Roots and Tuber Crops (Inivit).

***Data treatment***

It was worked with the minimum descriptors established internationally for the species under study, therefore none variable selection method was used, taking into consideration that they are the best taxonomic characters.

There are several standardization forms, but in the present work, only two are taken into account, the standardization carried out by Gower's metrics (Crossa and Franco 2004; Mason et al. 2005; Podani and Schmera 2006) when applied on variant one; and one implemented in the *data.Normalization* (n1) function of the clusterSim package (Walesiak and Dudek 2017) with which two distance matrixes are formed, from the Euclidean distance, taking into account the quantitative variables for variants three and four.

***Distance measure***

As described above, continuous and categorical variables were evaluated in each accession or cultivar from the germplasm banks, making difficult the elaboration of numerical scales that integrate all these variable types; actually, there is no a universal solution to the problem of how to combine them to build a distance measure. As it was already described, among the alternatives to deal with this problem is the Gower´s metric, which allows the analysis of mixed variables, which is implemented in the *gowdis*() function that includes the extension for ordinal variables (Podani 1999), described in the *FD* package (Laliberté and Legendre 2010; Laliberté et al. 2014).

As measures to evaluate the differences and similarities among objects for quantitative variables on variants three and four, the Euclidean distance of the *dist* function (*stats* package) was used.

***Cluster analysis***

Hierarchical clusters create a cluster hierarchy that are represented in tree structure, called dendrogram (Sembiring et al. 2010). In this research, the hierarchical clustering methods of Ward or Minimum Variance (Ward 1963), Average or UPGMA (*Unweighted Pair-Group Method using Arithmetic Averages*) (Sneath and Sokal 1973) were used, Single Linkage Agglomerative Clustering (Gower 1967), Complete Linkage Agglomerative Clustering (Sorensen 1948) with the *hclust*() function in the *stats* package which is part of the basic R installed by default. There is also the *cluster* package (Maechler et al. 2021) which extends the range of cluster analysis, because it includes partitional methods too.

***Clustering validation***

Once the result of the agglomerative method in the taxonomic analysis is obtained, it is important to check if the data set shows a tendency to form groups, which is determined through the agglomerative coefficient (AC) (Maechler et al. 2005) with the *coef.hclust*()function in the *cluster* package. On the other hand, hierarchical methods impose certain structure on the data and it is often necessary to consider whether this is acceptable or whether unacceptable distortions are introduced in the original relationships. The most commonly used method to verify this fact is the cophenetic correlation coefficient, proposed by Sokal and Rohlf (1962) which is defined as the correlation (which can be calculated through the *cor*() function *stats* package) among the initial distances, taken from the original data and the final distances, with which the individuals have got together during the method development and it is obtained with the *cophenetic*() function of the *stats* package, which returns the cophenetic distance of the dendrogram. High values of this coefficient show that during the process has not happened a great distortion, regarding with the original structure of the data.

Additionally, a statistical significance can be obtained by applying the Mantel´s test (Mantel 1967) implemented in the *mantel*() function of the *vegan* package (Oksanen et al. 2011), which is widespread in the ecological community.

Mantel´s test finds Mantel´s statistics as a correlation matrix between two dissimilarity matrixes, in this case the tree ultrametric and the original distance matrix of the data. However, the significance cannot be evaluated directly, because there are N(N-1)/2 entries for only N observations. For this, Mantel developed an asymptotic test that uses the permutation of the rows and columns of the first dissimilarity matrix repeatedly.

***Optimal number of clusters***

Once the results have been obtained, it is advisable, whenever possible to validate them with the aim of avoiding three error types: to force the clustering structure of the data set, look for a very different cluster number from the one that exists in reality, or accept an erroneous result delivered by the method.

When hierarchical clustering techniques are used, the researcher is not always interested only in the complete hierarchy, but also in a subset of partitions, obtained from it, resulting from a given data structure. The partitions are obtained by cutting the dendrogram or selecting one of the solutions in the succession of clusters which make up the hierarchy.

Unfortunately, this fundamental step is among the problems that are not fully solved with a statistical basis yet, among other reasons, because of the lack of an appropriate null hypothesis that is easily verifiable and the complex nature of the multivariate sampling distributions. There are very specific answers: for example, if the variables had a normal distribution and as a dissimilarity measure, the Euclidean quadratic distance will be used and at every step of the agglomeration process, it is written the distance value among the clusters that are merging, the results can be expected to have a Chi-square distribution and then statistically determine when clusters that are too far apart are merging (Grau 1994). But, it is evident that this criterion is very particular and not suitable for the problema as such.

The proposed solutions have been diverse and have been focused on the separation of clusters and their cohesion. For this reason, several validation indexes have emerged, but in practice there is no one that can determine the optimum cut-off level of the dendrogram. Some of the most widely used are Calinski-Harabasz´s (G1) index (G1) (Caliński and Harabasz 1974), Baker & Hubert´s (G2) index, which is Gamma de Goodman y Kruskal's adaptation (Everitt et al. 2011) and Hubert and Levine´s (G3) index (Gatnar and Walesiak 2004). Another measure that allows to establish a criterion on the optimal number of clusters is the Silhouette Width (Kaufman and Rousseeuw 1990) all implemented in the *clusterSim* package; and finally Dunn's index (Dunn 1974) which is implemented in the *clValid* package.

On the other hand, in these data there is not any available a priori knowledge of the classes that should be formed, which makes more difficult to decide when a clustering obtained from a given algorithm is better than another ("No free lunch" theorem (Wolpert and Macready 1997)). Therefore, the research should be focused on the analysis of the results based on statistical tests and the empirical experience of genebank curators.

***Combination of clusters and consensus***

To combine the results of the different clustering algorithms, the *cl\_ensemble* function of the *clue* package was used to achieve a better quality of the results achieved by the individual algorithms and compensate possible errors made by an individual algorithm. The consensus tree which is able to combine all the existing information from different results into a final tree was obtained with the *cl\_consensus* function (*clue* package) using the three available methods in this function (*Euclidean, Manhattan* and *Majority*) to calculate the consensus clusters in the hierarchical algorithms.

As stated, the classification is unsupervised, but in the case of the cocoyam germplasm under study (Milián et al. 2018) it is possible to have an idea of the conformation of some clusters, due to the already established identification of some species.

***The R statistical programming language***

In recent years, a programming language especially suitable for statistical analysis, called R (R Development Core Team 2019a) has strongly emerged as a free software alternative in many different teaching and research situations.

R is a programming language, mainly oriented to the statistical analysis and visualization of quantitative and qualitative information. It was officially introduced in 1997 and it is a free software governed by the General Public License ("General Public License" or GPL) of the Free Software Foundation ("Free Software Foundation" or GNU). Through R commands, it is possible to access all procedures and execute functions; however, it is also possible to design graphical interfaces to facilitate the interaction with unfamiliar users with their console (Salas 2008).

R allows the interconnection with other open code softwares, such as Java and Weka (Waikato Environment for Knowledge Analysis) which increases its potential in an impressive way.

The R language has been enriched by experienced researchers, statisticians and developers from all over the world, so it has manifested a vertiginous extension to multiple science fields, to such an extent that some of the most recent methods are only available through R packages, which extends its functionalities. Currently, there are more than 15266 published packages with free and available licenses in a general repository (R Development Core Team 2019b).

**Results and discussion**

The Cuban collection of cocoyam includes 57 accessions of the species *Xanthosoma sagittifolium* (L.) Schott (SAG) and 10 accessions that form a group that does not have a defined species, or whose specific classification is not defined for *Xanthosoma* sp. (SP). Moreover, the species *Xanthosoma brasilense* Engl. (BRA), whose the only one representative is the clone 'Belembe'; the species *Xanthosoma atrovirens* Koch & Bouché (ATR), constituted by the clone 'Amarilla Trinidad'; while the accession 'Jardín' represents the species *X. nigrum* (Vell.) Mansf. (NIG) (Milián 2008; Milián et al. 2018).

In variant one, Gower's metric for mixed variables is used on the data matrix of qualitative (nominal (18) and ordinal (2)) and quantitative (16) variables from cocoyam (*Xanthosoma* spp.) collection.

A second distance matrix is obtained by applying Gower's metric to the quantitative characteristics (16), taken to established ranges for cocoyam (*Xanthosoma* spp.) collection, these new coded variables (according to Minimum Descriptor Systems) are treated as ordinal and they are integrated to the nominal (18) and ordinal (2) variables already established.

In the third variant, qualitative variables (nominal and ordinal) are transformed through the multiple correspondence analysis, where 31 dimensions are taken, representing 91.176 % of the total variability of the data. The 16 quantitative variables were standardized and later, both results emerged in a single continuous data matrix, to which the Euclidean distance was applied. Sorensen (1948) and Gatnar and Walesiak (2004) have shown that at least 70 % of the total explained inertia is the most appropriate cut-off value.

The procedure on qualitative variables in the fourth variant is similar to the one used in variant three and with quantitative variables, an analysis of main components was performed, where the first six dimensions are taken, according to Kaiser´s criterion (eigenvalues > 1.0) which constitute 71.271 % of the total variance, observed in the data and from it the increase is not substantial. Both results are merged into a single continuous data matrix to which the Euclidean distance was applied.

To the four distance matrixes obtained for the four variants, the four agglomeration methods selected for the study were applied.

As it is shown in Table 1 for the four variants, the Ward´s method has the highest agglomerative coefficient, followed by the complete linkage, average and single linkages, respectively. High values of this coefficient basically measure the level of clustering structure found by the method on the data.

The average method was higher in the coefficient of cophenetic correlation with respect to the other methods, followed by single linkage, complete linkage and Ward, in that order for the four variants. The high values of the average method indicate that the indexed hierarchy well reproduces the initial dissimilarity, similar conclusions were stated by Blackburn et al. (2005), so it is the one that best fits the original distance matrix (Sokal and Rohlf 1962; Mohammadi and Prasanna 2003; Podani and Schmera 2006). Analogous responses were obtained by (Molina et al. 2018) for yam (*Dioscorea* spp.) and plantain (or banana) (*Musa* spp.) germplasm databases.

Table 1. Summary of the agglomerative and cophenetic correlation coefficient for the four agglomerative methods under study on the four variants. Cocoyam collection (*Xanthosoma* spp.)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Databases | Agglomeration methods | Variant 1 | | | | Variant 2 | | | | Variant 3 | | | | Variant 4 | | | |
|  | AC |  | CCC |  | AC |  | CCC |  | AC |  | CCC |  | AC |  | CCC |
| Cocoyam (*Xanthosoma* spp,) | Ward | 1 | 0,885 | 4 | 0,578 | 1 | 0,899 | 4 | 0,475 | 1 | 0.887 | 4 | 0.461 | 1 | 0,916 | 4 | 0,456 |
| Average | 3 | 0,558 | 1 | 0,762 | 3 | 0,606 | 1 | 0,717 | 3 | 0.522 | 1 | 0.781 | 3 | 0,595 | 1 | 0,80 |
| Single Linkage | 4 | 0,466 | 2 | 0,701 | 4 | 0,487 | 2 | 0,633 | 4 | 0.429 | 2 | 0.763 | 4 | 0,512 | 2 | 0,725 |
| Complete Linkage | 2 | 0,632 | 3 | 0,628 | 2 | 0,691 | 3 | 0,619 | 2 | 0.643 | 3 | 0.517 | 2 | 0,702 | 3 | 0,625 |

AC: agglomerative coefficient, CCC: coefficient of cophenetic correlation

The validation indexes (Table 2) do not have a homogeneous behavior for any of the agglomeration methods within each variant and, on the other hand the Silhouette width index is the only one that is constant for the four agglomeration methods in all variants.

As it can be seen, it is not an easy task to determine the optimal number of clusters, since more than one possible solution may appear. Generally, the study of the solution by crop specialists, plus the interpretation of the validation indexes are the key to establish a group number that gives answer to the analysis.

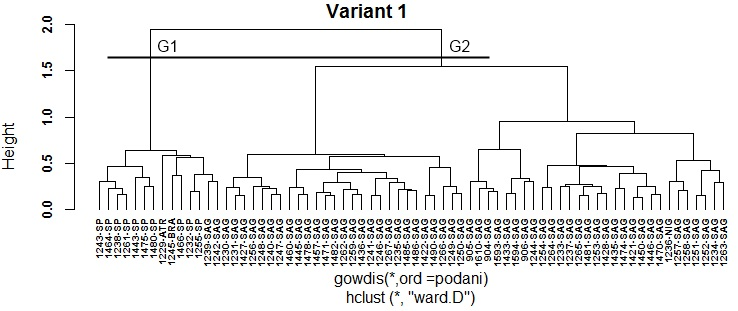
Table 2. Results of five validation indexes applied to four agglomerative hierarchical methods on cocoyam (*Xanthosoma* spp.) collection database for the four variants of data structure

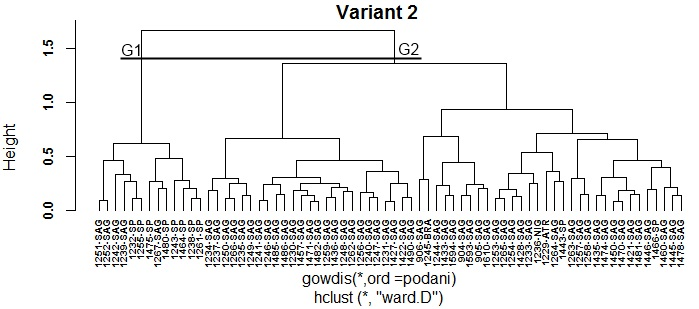
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variants of data structure | Agglomeration methods | Validation indexes | | | | |
| Dunn  BC/ Value | Silhouette  BC/ Value | G1  BC/ Value | G2  BC/ Value | G3  BC/ Value |
| Variant1 | Ward | 2/0,428 | 2/0,197 | 2/9,397 | 7/0,741 | 2/0,463 |
| Average | 2/0,626 | 2/0,273 | 4/4,555 | 3/0,909 | 2/0,259 |
| Single Linkage | 2/0,626 | 2/0,273 | 3/2,114 | 3/0,909 | 3/0,259 |
| Complete Linkage | 6/0,403 | 2/0,163 | 2/7,976 | 10/0,681 | 20/0,500 |
| Variant 2 | Ward | 2/0,318 | 2/0,177 | 2/7,069 | 10/0,755 | 6/0,446 |
| Average | 2/0,527 | 2/0,268 | 4/4,457 | 2/0,807 | 2/0,239 |
| Single Linkage | 2/0,527 | 2/0,268 | 3/2,114 | 2/0,807 | 2/0,239 |
| Complete Linkage | 3/0,373 | 2/0,208 | 2/8,043 | 10/0,725 | 3/0,479 |
| Variant 3 | Ward | 9/0,400 | 2/0,176 | 2/12,714 | 9/0,638 | 9/0,561 |
| Average | 2/0,541 | 2/0,290 | 6/6,079 | 3/0,859 | 10/0,610 |
| Single Linkage | 4/0,592 | 2/0,265 | 4/3,214 | 4/0,919 | 4/0,509 |
| Complete Linkage | 2/0,494 | 2/0,269 | 3/7,236 | 2/0,732 | 10/0,648 |
| Variant 4 | Ward | 5/0,287 | 2/0,207 | 2/9,590 | 10/0,678 | 9/0,381 |
| Average | 3/0,495 | 2/0,335 | 6/4,120 | 5/0,905 | 10/0,542 |
| Single Linkage | 2/0,530 | 2/0,330 | 5/2,363 | 5/0,905 | 5/0,441 |
| Complete Linkage | 2/0,476 | 2/0,333 | 3/6,400 | 2/0,840 | 8/0,561 |

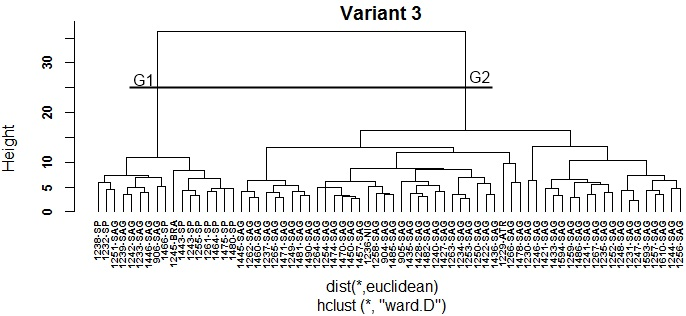
BC: best cutoff index, Value: value of the best cutoff index

The dendrograms in Figure 1, obtained for the four variants with Ward's method, show that they have compact and defined structures, due to the high agglomerative coefficient index.

The distance established between groups one and two of variants one and two is smaller than the distance established for these two defined groups in variants three and four which may be, due to the dimensionality reduction established in these two latter variants.







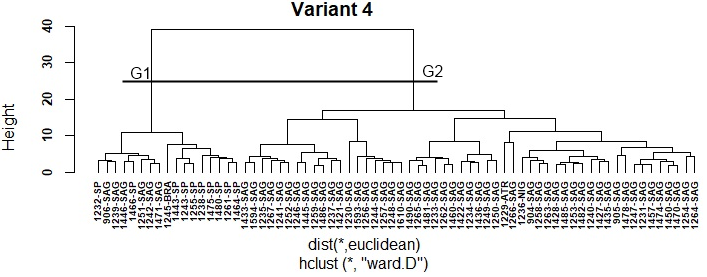


Figure 1. Dendrograms of the Ward´s clustering method. Four variants. Cocoyam (*Xanthosoma* spp.) collection

In an individual analysis of each dendrogram in Figure 1, it is observed that in variant one, two groups (G1 and G2) are clearly established at the height of 1.5, which is not valid from the point of view of analysis based on the four species present in cocoyam (*Xanthosoma* spp.) collection, according to specialists´ biological criterio (Milián et al. 2018). According to the species distribution, group one has 10 accessions (yellow and orange yellow pulp accessions in their entirety) that do not have a defined species (SP), two of the species *X. sagittifolium* (a creamy white (1239-SAG) and a light pink (1242-SAG) one *X. atrovirens*, constituted by the accession ‘Amarilla Trinidad’ and one of the species *X. brasilense* ('Belembe' accession).

Group two has the rest of the accessions, 57 of the species *X. sagittifolium* and one of the species *X. nigrum*, represented by the accession 'Jardín'. The left branch groups most of the accessions with white and creamy white pulp color, six light pink (1256-SAG, 1460-SAG, 1478-SAG, 1235-SAG, 1266-SAG, 1250-SAG) and one purple (1485-SAG). The right branch of group two is composed of purple and light purple, pink and light pink, white (1244-SAG) and one yellow accession (906-SAG).

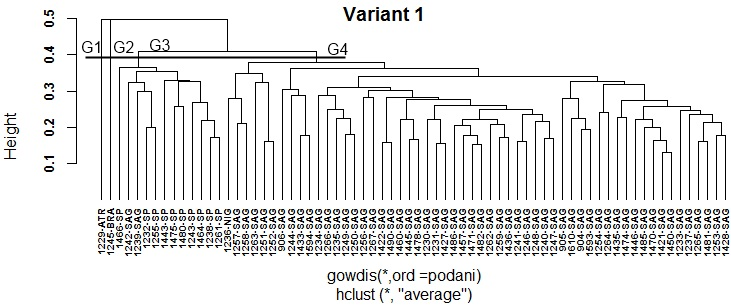
In variant two, group one keeps eight out of ten accessions of the species SP and incorporates three accessions of the species *X. sagittifolium*. The rest of the accessions are kept in group two, 54 of the species *X. sagittifolium*, two of the species SP and one of each of the species *X. nigrum*, *X. atrovirens* and *X. brasilense*.

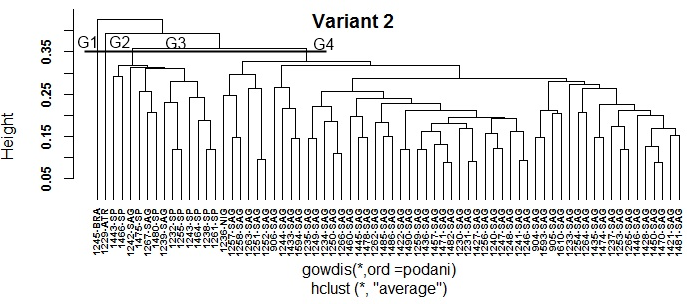
Group one of variant three has ten accessions of the species SP, it incorporates the accession 906-SAG to complete six of the species *X. sagittifolium* and one of the species *X. brasilense*, the other accessions remain in group two, 53 of the species *X. sagittifolium*, one of the species *X. atrovirens* and *X. nigrum*.

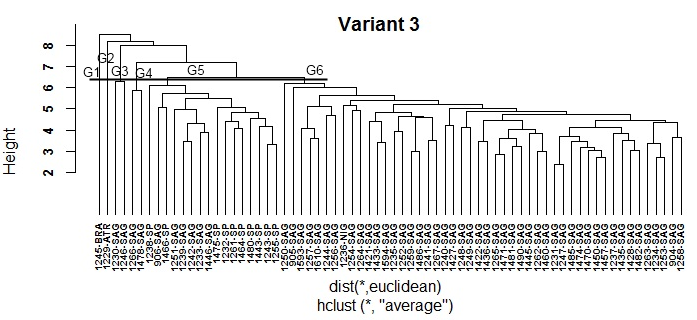
In variant four, group one has the ten accessions of the species SP, six of the species *X. sagittifolium*, one of the *X. brasilense*. Group two has the rest of the accessions, 53 of the species *X. sagittifolium* and one of the species *X. atrovirens* and *X. nigrum*.

Ward's method in variant one is higher in the classification, based on species and pulp color, with respect to the rest of the variants.

The dendrograms in Figure 2 represent the average agglomeration method for the four treated variants in the present study and the estimation of the optimal cut-off indexes are established in two groups, as it has already been indicated it is not consistent from the point of view of distribution of the present species in the database (*X. sagittifolium*, *SP X*., *X. nigrum*, *X. atrovirens* and *X. brasilense*). However, a cut into four groups for the first and second variants and six for the third and fourth variants, establishes the best distributions according to the species and the pulp color. These groups are established by Calinski-Harabasz´s Index (G1).







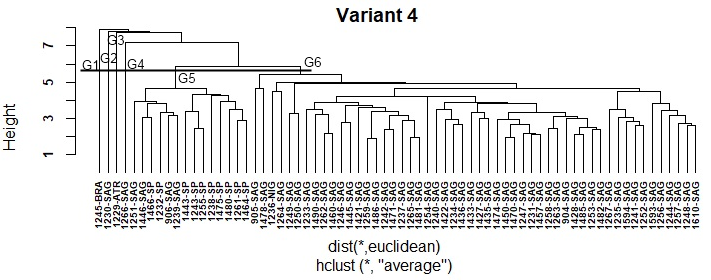


Figure 2. Dendrograms of the average clustering method. Variants 1, 2, 3 and 4. Cocoyam (*Xanthosoma* spp.) collection

Variant one with the average agglomeration method achieves a lower species mixture compared to the other variants. Group one (G1) is formed by the only one accession of the species *X. atrovirens*, the second one (G2) by the species *X. brasilense*. Group three (G3) is composed of the ten accessions of the species SP and two of the species *X. sagittifolium* (1242-SAG and 1239-SAG) which always appear together with the species SP. In variants two, three and four, the group that is formed by accessions of the SP species incorporates others of the species *X. sagittifolium*, increasing the mix of accessions in this group.

In the single linkage method, the optimal cut-off indexes are in two and three groups for variant one. In two and three for variant two. In the third one, they are defined in two and four, and in variant four they are set at two and five (Table 2). The dendrograms show a difficult structure to interpret (Figure 3). This confirms what has been stated by several authors, that this method does not form a dendrogram with well identified partitions, but rather an elongated shape that is difficult to interpret.

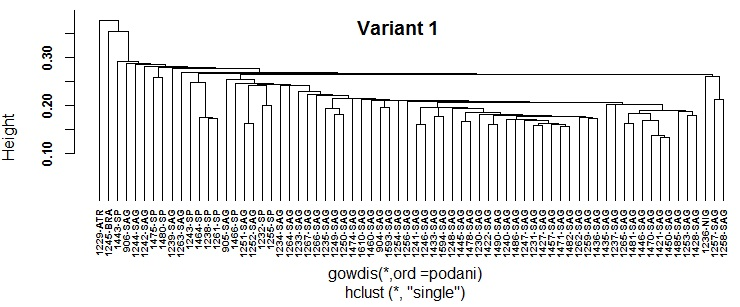
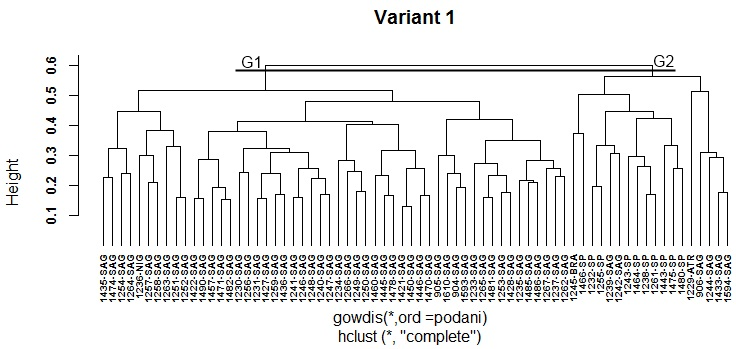
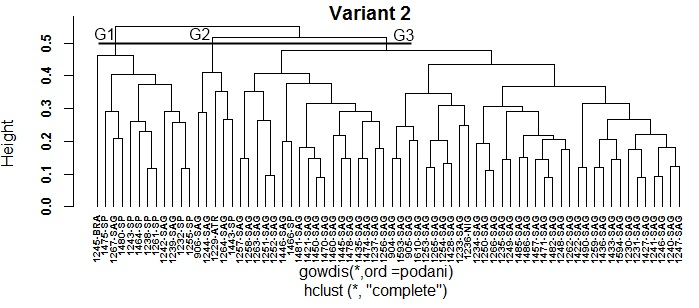
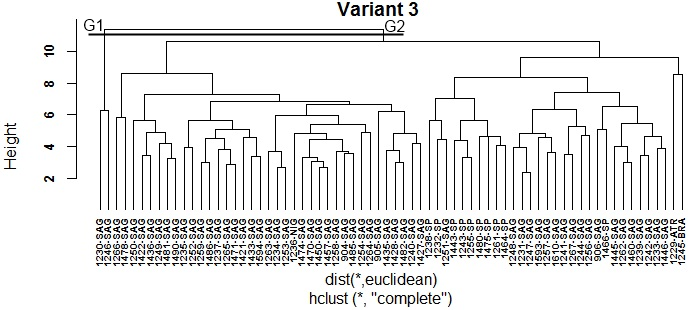


Figure 3. Dendrograms of the single linkage clustering method. First variant. Cocoyam (*Xanthosoma* spp.) collection

On the other hand, the optimal cutoff indexes for the complete linkage method are defined as two, six, ten and 20 groups for variant one. In two, three and ten for variants two and three. Variant four in two, three and eight (Table 2). As in the other variants, none of the cut indexes establish that the groups can be clearly defined, according to species and the pulp color (Figure 4). There is a greater mixture with the species *X. sagittifolium* in the group, where the SP species are in the majority in all variants, with respect to Average and Ward ´s methods in variant one.







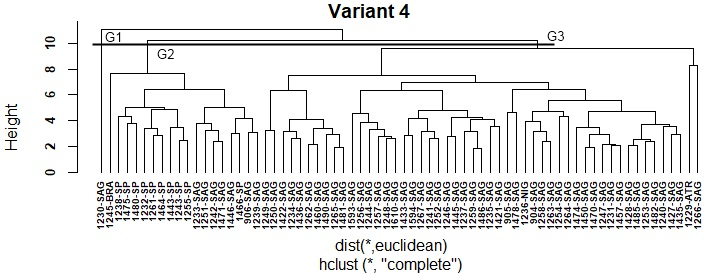


Figure 4. Dendrograms of the complete linkage clustering method. Variants 1, 2, 3 and 4. Cocoyam (*Xanthosoma* spp.) collection

From the above analyses it is derived that the average agglomeration method in variant one is the one that best represents the structure of cocoyam database. In studies performed by Milligan (1980, 1989), Milligan and Cooper (1986, 1987) provide evidence of the usefulness of the average agglomeration method. Belbin and McDonald (1993) in studies of three ecological data structures, concludes that the average method is the one that best recovers the data

structure.

Considering the different results obtained by the clustering methods, it is convenient to agglomerate them to achieve a better quality of the results achieved by the individual algorithms and compensate the possible errors made in the performance of each one. In addition to, if there is no a complete information about the problem, making a set of diverse partitions is advisable, since more varied the set of partitions is, greater is the available information at the combination stage.

The combination of groupings to achieve a consensus dendrogram was performed with the four agglomerative methods under study. As it was observed in Table 1 above, Ward's cophenetic correlation coefficient value is low, however it has a high agglomerative coefficient value and presents a dendrogram with a high group structure, according to the species distribution. The complete linkage method, although it has a value around 0.6 can be considered with a reasonable structure (Rousseeuw 1986). The average and single linkage methods have acceptable cophenetic correlation coefficient values.

From the combination of groupings, three consensus trees are obtained for the four variants by the Euclidean, Manhattan and Majority methods; as it is shown in Table 3, the cophenetic correlation coefficients for the Euclidean method are lower than those ones established by the Manhattan method, and in turn, the Majority method has very poor values, which is a reflection of the own dendrogram structure.

Table 3. Summary of the cophenetic correlation coefficient indexes with the consensus methods Euclidean, Manhattan and Majority. Cocoyam (*Xanthosoma* spp.) collection

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Database | Consensus methods | Cophenetic correlation coefficient | | | |
| Variant 1 | Variant 2 | Variant 3 | Variant 4 |
| Cocoyam (*Xanthosoma* spp.) | Euclidean | 0,626 | 0,533 | 0,535 | 0,523 |
| Manhattan | 0,674 | 0,716 | 0,755 | 0,745 |
| Majority | 0,281 | 0,247 | 0,205 | 0,238 |

Figure 5 represents the consensus dendrogram with the Euclidean distance method, in which a clear formation of two groups (G1 and G2) can be observed for variant one. Within group one, the species *X. atrovirens* and *X. brasilense*. form separate groups, which at the same time, they join the rest of the 12 accessions, ten of the species SP and two accessions of the species *X. sagittifolium*. Group two has the rest of the accessions, 57 of the species *X. sagittifolium* and one of the species *X. nigrum*.

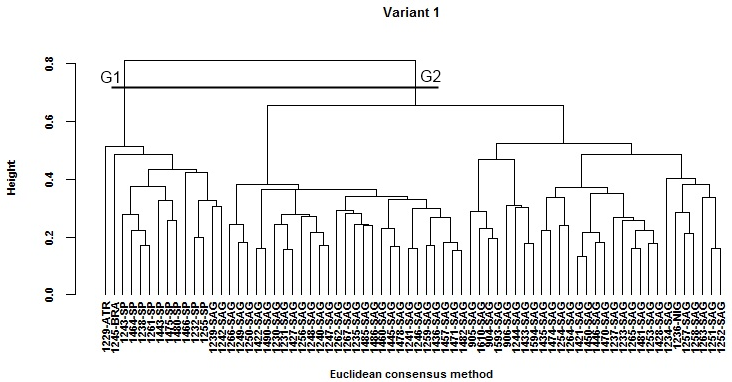


Figure 5. Consensus dendrogram with the Euclidean method for the first variant. Cocoyam (*Xanthosoma* spp.) collection

The consensus dendrogram obtained by the Manhattan method (Figure 6) for variant one is the one that best represents the species groupings, according to the structure of cocoyam database, a clear biologic description of the database is achieved, according to the present species.

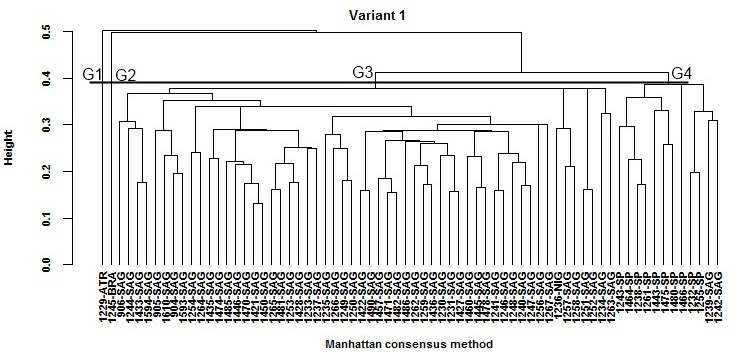


Figure 6. Consensus dendrogram with the Manhattan method for the first variant. Cocoyam (*Xanthosoma* spp.) collection

In multiple iterations for the Euclidean and Manhattan methods, dendrograms are obtained with small variations in the structure of the lower groups, in terms of height and order of the accessions within the two main groups, but not between them. The process is not deterministic; it may happen that running the same algorithm, with the same parameters, on the same data, slightly different results are obtained.

On the other hand, the Majority method (Figure 7) has a low cophenetic correlation coefficient index, which is reflected in the dendrogram structure, not achieving a clear biologic description of the database for all variants. In multiple iterations with this method, equal dendrograms are obtained without variations in the structure of the groups, height and order of the accessions.

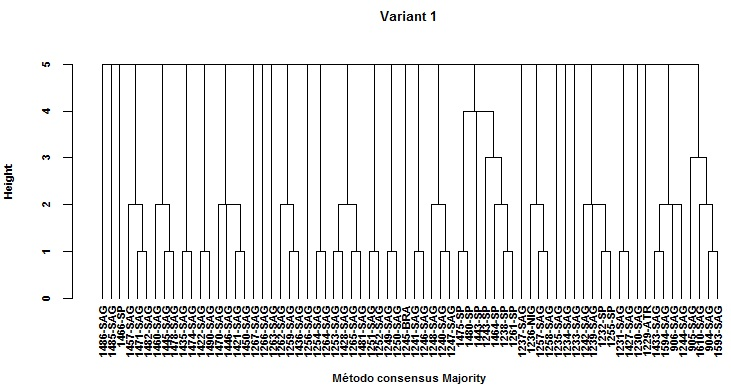


Figure 7. Consensus dendrogram with the Majority method for the first variant. Cocoyam (*Xanthosoma* spp.) collection

Interpreting what characterizes each cluster is the most important part of the analysis, because it provides a means of evaluating the correspondence of the clusters of those ones proposed by a priori theory, or by practical experience.

Exact clustering is not a simple task and it is difficult to make general recommendations. It is always advisable to try to with more than one method. If several methods give similar results, then it can be assumed that natural groupings do indeed exist.

Comparing the relative merits of each method is difficult, because when we apply clustering algorithms to a data set, the results differ.

The authors agree that there is no a clustering method that is capable of finding all possible cluster types at any data set (Jain et al. 1999). This could be the unsupervised version of the "No free lunch" principle that has been established, as a theorem of the non-existence of a universally best supervised classifier for all problems.

Kleinberg (2002) showed the impossibility of designing a perfect clustering algorithm and different algorithms will be suitable to find different types of clustering in the data. Sembiring et al. (2010) also concluded that different methods can give different structures.

Some of the present limitations in the different dealt variants are:

Hierarchical cluster analyses give the possibility of using different measure types, to estimate the distance among the cases, the possibility of changing the original metrics of the variables and selecting among a great variety of agglomeration methods. But, there is no any combination of these possibilities that optimizes the obtained solution for these data matrixes. Therefore, it is recommended to evaluate different solutions to choose the most consistent one.

The main problem lies in choosing the true number of clusters, since more than one possible solution may appear and the indexes may give different results with the same data and the same method used.

Despite the limitations of the analysis of clusters, this technique is widely used in many branches of science, because it is eminently exploratory, since most of the time it does not use any type of statistical model to carry out the classification process. These methods are very suitable for extracting information from a set of data, without imposing prior restrictions in the form of statistical models, at least explicitly, and therefore they can be very useful as a tool of hypotheses elaboration about the problem under consideration, without imposing patterns or previously established theories.

Therefore, an analysis of the dendrogram structure by the cultivar specialists, together with the studied indexes is necessary to determine those clusters that best fit the characteristics of the database under study and make a decision on the best supported partitioning, also based on the available priori classes. The results obtained allow confirm the contribution of statistical techniques to achieve a better germplasm classification individually and in its combination, in this case accessions of plantain, cocoyam and yam, as well as validation methods which determine to a great extent the obtaining of more consistent and reliable structures.

Gower's metric for mixed variables where qualitative variables (nominal and ordinal) and the quantitative ones are integrated, together with the Average and Ward's agglomeration methods, resulted the best variant used for germplasm classification, given by its closer approach to the biological vision of the database under study; similar results are shown for yam (*Dioscorea* spp.) and plantain (*Musa* spp.) collections in studies performed by Molina et al. (2018). Furthermore, it can be stated that the consensus tree with the Manhattan method better represents cocoyam (*Xanthosoma* spp.) database structure in variant one, with respect to all variants and the Euclidean and Majority methods.

Taking into account that *Xanthosoma* sp. (SP) accessions keep a tendency to be together in the same cluster in variant one of consensus by Manhattan and Euclidean methods, it would be favorable to carry out molecular analyses in future characterization studies within this genus to determine if they represent a new species.

**Conclusions**

A detailed knowledge of unsupervised hierarchical clustering methods that best influence the results of *Xanthosoma* cocoyam germplasm classification individually and in its combination, as well as validation methods that greatly determine more consistent structures for the first time, which expands the knowledge of analysis types, techniques and statistical tools that are available for a more reliable processing in the classification of plant germplasm banks.

The best result for a more reliable processing in the classification of genotypes was achieved with the strategy one. Moreover, the combination of groupings to achieve a consensus dendrogram is an effective technique to get a better yield and quality of the results achieved by individual algorithms.

Because of its flexibility, this analysis can be applied to other classification studies in plant germplasm banks.

**Bibliographic references**

Belbin L, McDonald C (1993) Comparing three classification strategies for use in ecology. J Veg Sci 4:341–348

Blackburn TM, Petchey OL, Cassey P, Gaston KJ (2005) Functional Diversity of Mammalian Predators and Extinction in Island Birds. Ecology 86:2916–2923

Caliński T, Harabasz J (1974) A dendrite method for cluster analysis. Commun Stat 3:1–27. https://doi.org/10.1080/03610927408827101

Crossa J, Franco J (2004) Statistical methods for classifying genotypes. Euphytica 137:19–37

Cuadras CM (1981) Métodos de Análisis Multivariante. Eunibar DL XII:642

Dunn JC (1974) Well separated clusters and optimal fuzzy partitions. J Cybern 4:95–104

Everitt BS, Landau S, Leese M, Stahl D (2011) Cluster Analysis. 5th Edition. 2011 John Wiley & Sons, Lid. ISBN: 978-0-470-74991-3. pp 346

Gatnar E, Walesiak M (2004) Metody statystycznej analizy wielowymiarowej w badaniach marketingowych [Multivariate statistical analysis methods in marketing esearch]. Wroclaw

Gower JC (1967) A comparison of some methods of cluster analysis. Biometrics 23:623–628

Gower JC (1971) A general coefficient of similarity and some of its properties. Biometrics 27:857–871

Grau R (1994) Estadística aplicada con ayuda de paquetes de software. Editorial universitaria, Universidad de Guadalajara, Jalisco, México.

Jackson DA (1993) Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. Ecology 74:2204–2214. https://doi.org/10.2307/1939574

Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31(3):264–323

Kaiser HF (1974) An index of factorial simplicity. Psychometrika 39:31–36. https://doi.org/10.1007/BF02291575

Kaufman L, Rousseeuw P (1990) Finding Groups in Data: An Introduction to Cluster Analysis. A John Wiley & Sons, INC., Publication, New York.

Kleinberg J (2002) An impossibility theorem for clustering. In Proc. of the 16th conference on Neural Information Processing S**ystems**, 15. pp 446–453

Laliberté E, Legendre P (2010) A distance-based framework for measuring functional diversity from multiple traits. Ecology 91:299–305

Laliberté E, Legendre P, Shipley B (2014) FD: measuring functional diversity from multiple traits, and other tools for functional ecology. R package version 1.0-12

Maechler M, Rousseeuw P, Struyf A, Hubert M, Hornik K (2021). *cluster: Cluster Analysis Basics and Extensions*. R package version 2.1.1. https://CRAN.R-project.org/package=cluster. Consulted: 17 enero 2021.

Maechler M, Rousseeuw P, Struyf A, Hubert M (2005) Cluster Analysis Basics and Extensions. R package version 1.14.0

Mantel N (1967) The detection of disease clustering and a generalized regression approach. Cancer Res 27:209–220

Mason N, Mouillot D, Lee W, Wilson J (2005) Functional richness, functional evenness and functional divergence: The primary components of functional diversity. Oikos 111:112–118. https://doi.org/10.1111/j.0030-1299.2005.13886.x

Milián Jiménez M (2008) Caracterización de la variabilidad de los cultivares de la colección cubana de germoplasma del género Xanthosoma (Araceae). Tesis para aspirar al grado de Doctor en Ciencias Biológicas, Ciudad de la Habana, pp 123

Milián JM; Molina CO and Figueroa Y (2018) Integrated Characterization of Cuban Germplasm of Cocoyam (Xanthosoma Sagittifolium (L.) Schott). J Plant Genet Crop Res 1(1):1–18

Milligan GW (1989) A Study of the Beta-Flexible Clustering Method. Multivar Behav Res 24:163–176. https://doi.org/10.1207/s15327906mbr2402\_2

Milligan GW (1980) An examination of the effect of six types of error perturbation on fifteen clustering algorithms. Psychometrika 45:325–342. https://doi.org/10.1007/BF02293907

Milligan GW, Cooper MC (1986) A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. Multivar Behav Res 21:441–458. https://doi.org/10.1207/s15327906mbr2104\_5

Milligan GW, Cooper MC (1987) Methodology Review: Clustering Methods. Appl Psychol Meas 11:329–354. https://doi.org/10.1177/014662168701100401

Mohammadi SA, Prasanna BM (2003) Analysis of genetic diversity in crop plants - salient statistical tools and considerations. Crop Sci 43:1235–1248

Molina Concepción O, García Rodríguez R, Milián Jiménez M, et al (2018) Evaluación de técnicas para la clasificación del genofondo de raíces, rizomas y tubérculos tropicales, plátanos y bananos en Cuba. Publisher: Informe del Proyecto PAH P131LH001027. Instituto de Investigaciones de Viandas Tropicales (Inivit), MINAG, Cuba.

Núñez CA, Rodríguez JE, Ángel RN, Barrientos AF (2004) Construcción de dendrogramas de taxonomía numérica mediante el coeficiente de distancia χ2: una revisión. Chapingo Ser Hortic 10:229–237

Oksanen J, Blanchet FG, Kindt R, et al (2011) vegan: Community Ecology Package. R package version 1.17-9

Piggot JR, Sharman K (1986) Methods to aid interpretation of multidimensional data. In: Piggott J.R. (ed.), Statistical Procedures in Food Research. London: Elsevier Applied Science Publishers Ltd. pp 181-232

Podani J (1999) Extending Gower’s general coefficient of similarity to ordinal characters. Taxon 48:331–340

Podani J, Schmera D (2006) On dendrogram-based measures of functional diversity. Oikos 115:179–185

R Development Core Team (2019a) R: A language and environment for statistical computing (version 3.6.1). R Foundation for Statistical Computing. Vienna, Austria., Vienna, Austria. http://www.r-project.org/. Consulted: 7 April 2019.

R Development Core Team (2019b) R: A language and environment for statistical computing. R Foundation for Statistical Computing

Rousseeuw PJ (1986) A visual display for hierarchical classification. Data Anal Inform 4:743–748

Ruiz De Galarreta JI (1998) Agrupación de poblaciones locales de maíz (Zea mays L.) mediante caracteres morfológicos y parámetros ambientales. Serv Publicaciones Univ Lleida 161

Salas C (2008) ¿Por qué comprar un programa estadístico si existe R? Ecol Austral Vol. 18:223–231

Sembiring RW, Mohamad Zain J, Embong A (2010) A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course. J Comput 2:33–38

Sneath PHA, Sokal RR (1973) Numerical taxonomy. The principles and practice of numerical classification. W. H. Freeman and Co, San Francisco, California, USA

Sokal RR, Rohlf FJ (1962) The Comparison of Dendrograms by Objective Methods. Taxon 11:33–40. https://doi.org/10.2307/1217208

Sorensen TA (1948) A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to alalysies of vegetation on Danish commons. Biol Skr 5:1–34

Tenenhaus M, Young F (1985) An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. Psychometrika 50:91–119. https://doi.org/10.1007/BF02294151

Vega-Pons, Ruiz JS (2010a) Combinación de agrupamiento: un estado del arte. CENATAV, La Habana, Cuba

Vega-Pons S, Correa-Morris J, Ruiz-Shulcloper J (2010b) Weighted Partition Consensus via Kernels. Pattern Recognit 43:2712–2724. https://doi.org/10.1016/j.patcog.2010.03.001

Vega-Pons, Ruiz JS (2011) A survey of clustering ensemble algorithms. International Journal of Pattern Recognition and Artificial Intelligence, 25(3):337-372. DOI: [10.1142/S0218001411008683](https://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1142%2FS0218001411008683)

Walesiak M, Dudek A (2017) clusterSim: Searching for optimal clustering procedure for a data set. <http://cran.fhcrc.org/web/packages/clusterSim/clusterSim.pdf>. Consulted: 12 december 2017.

Ward JHJr (1963) Hierarchical grouping to optimize an objective function. J Am Stat Assoc 58:236–244

Wolpert DH, Macready WG (1997) No free lunch theorems for optimization. IEEE Trans Evol Comput 1:67–82. **DOI:**[10.1109/4235.585893](https://doi.org/10.1109/4235.585893)