Research, Society and Development, v. 11, n. 2, e48511226046, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i2.26046

# Comparison between similarity coefficients with application in forest sciences

## Comparação entre coeficientes similaridade com aplicação em ciências florestais

## Comparación entre coeficientes de similitud con aplicación en ciencias forestales

Received: 01/22/2022 | Reviewed: 01/29/2022 | Accept: 02/01/2022 | Published: 02/03/2022

**Mácio Augusto de Albuquerque**
ORCID: https://orcid.org/0000-0002-0113-9130
Universidade Estadual da Paraíba, Brazil
E-mail: marcioaa@uepb.edu.br
**Emanuela Rodrigues do Nascimento**
ORCID: https://orcid.org/0000-0001-9750-734X
Universidade Estadual da Paraíba, Brazil
E-mail: nascimento.manu25@gmail.com
**Kleber Napoleão Nunes de Oliveira Barros**
ORCID: https://orcid.org/0000-0003-2515-3292
Universidade Federal Rural de Pernambuco, Brazil
E-mail: kleber.barros@ufrpe.br
**Patrícia Silva Nascimento Barros**
ORCID: https://orcid.org/0000-0003-0681-2029
Universidade Federal da Paraíba, Brazil
E-mail: patricia@dcx.ufpb.br

**Abstract**
The multivariate statistic has been used in divergence studies concerning plants species. Analysis of the similarity or distance among individuals is an important tool for population. This among aims to present the main show the main coefficients of similarity and dissimilarity and their properties and the importance of axioms for the complement of similarity for the methods in cluster analysis. We evaluated the changes caused by five different similarity coefficients in the group of 11 plots and 17 species. We tested the coefficients of Jaccard, Sorensen-Dice, Simple Agreement, Russel e Rao e Rogers e Tanimoto comparisons being made between them by cophenetic correlations, Rand, adjusted Rand and stress between the distances obtained by the addition of these coefficients, and also by means of dendrograms (visual inspection), projection efficiency in a two-dimensional space and groups formed by the method of average linkage. The results showed that the use of different similarity coefficients caused few changes in the grouping of installments in groups, and the validation obtained between similar plots. Even though few changes in the structure of most different groups, these coefficients changed some relationships between plots with high similarity.
**Keywords:** Clustering analysis; (Dis)similarity coefficient; Validation.

**Resumo**
A estatística multivariada tem sido utilizada em estudos de divergências dentro de espécies vegetais. A análise da similaridade ou dissimilaridade entre objetos é uma ferramenta importante no estudo das populações. Este trabalho visa apresentar os principais coeficientes de similaridade e dissimilaridade, bem como suas propriedades e a importância dos axiomas para o complemento da similaridade e para os métodos em análise de agrupamento. Foram avaliadas as alterações provocadas por cinco diferentes coeficientes de similaridade no agrupamento de 11 parcelas e 17 espécies. Foram testados os coeficientes de Jaccard, Sorensen-Dice, Concordância simples, Russel e Rao e Rogers e Tanimoto sendo as comparações entre eles realizadas pelas correlações cofenéticas, Rand, Rand ajustado e estresse entre as distâncias obtidas pelo complemento destes coeficientes, e pela avaliação dos dendrogramas (inspeção visual), eficiência da projeção no espaço bidimensional e grupos formados pelo método de ligação média. Os resultados evidenciaram que a utilização de diferentes coeficientes de similaridade provocou poucas alterações no agrupamento das parcelas em grupos, sendo as validações obtidas entre as parcelas semelhantes. Mesmo provocando poucas mudanças na estrutura dos grupos mais diferenciados, estes coeficientes alteraram alguns relacionamentos entre parcelas com alta similaridade.
**Palavras-chave:** Análise de agrupamento; Coeficientes de (dis)similaridade; Validação.

**Resumen**
Las estadísticas multivariadas se han utilizado en estudios de divergencias dentro de las especies de plantas. El análisis de similitud o disimilitud entre objetos es una herramienta importante en el estudio de poblaciones. Este trabajo tiene como objetivo presentar los principales coeficientes de similitud y disimilitud, así como sus propiedades y la importancia de los axiomas para el complemento de similitud y para los métodos en análisis de conglomerados. Se evaluaron los cambios causados por cinco coeficientes de similitud diferentes en el agrupamiento de 11 parcelas y

17 especies. Se probaron los coeficientes de Jaccard, Sorensen-Dice, Concordancia simple, Russell y Rao y Rogers y Tanimoto y las comparaciones entre ellos se realizaron mediante las correlaciones cofenéticas, Rand, Rand ajustado y tensión entre las distancias obtenidas por el complemento de estos coeficientes, y por la evaluación de los dendrogramas (inspección visual), eficiencia de proyección en espacio bidimensional y grupos formados por el método de enlace medio. Los resultados mostraron que el uso de diferentes coeficientes de similitud provocó pocos cambios en la agrupación de parcelas en grupos, con validaciones obtenidas entre parcelas similares. Aun provocando pocos cambios en la estructura de los grupos más diferenciados, estos coeficientes alteraron algunas relaciones entre parcelas con alta similitud.
**Palabras clave:** Análisis de conglomerados; Coeficientes de (des)similitud; Validación.

## 1. Introduction

Multivariate statistical techniques have been widely used in forestry studies involving climate, soil, relief and vegetation variables simultaneously. These techniques are used in order to order, in order to determine the influence of environmental factors on the composition and productivity of the site, and to group, for the purpose of classification (Souza and Souza, 2006).

When the objective is to classify groups, a large number of (dis)similarity coefficients are found in the literature Jaccard, Sorensen-Dice, Simple concordance, Russell and Rao and Rogers and Tanimoto, and it is possible to observe different coefficients used with the same or different purposes. However, not all authors justify the reason for choosing a certain coefficient, that is, the choice is subjective and can compromise the nature of the analysis.

Similarity measures are numerical quantities that quantify the degree of association between pairs of objects or individuals, items, etc., and are considered a measure of similarity $s_{ij}$ if, for everyone $x_i$ e $x_j$ that satisfy the following properties: $0 \leq s_{ij} \leq 1$ if $i \neq j$, $s_{ij}=1$, and $s_{ij}=s_{ji}$. The calculation and structure of the numerical analysis result is obtained from the association matrix, which does not necessarily reflect all the information originally contained in the data matrix, as the objects or descriptors are represented in reduced space. This underscores the importance of choosing an appropriate measure of association and determines the topic of analysis. Therefore, the following considerations must be taken into account:

1. The nature of the study (ie, the initial question and the hypothesis) determines the type of structure that must be evidenced through an association matrix and, consequently, the type of (dis)similarity measure to be used.
2. Measures are represented by different mathematical equations and, in association matrix analysis, coefficients with specific mathematical properties are often required.
3. It is also necessary to consider the computational aspect, and, therefore, the choice of coefficient often depends on its availability in the computational package or on the user's ease in programming it.

Consider the comparison of a pair of elements (iej) from the results of q binary variables, each coded in such a way that it can assume values 0 or 1 (for example, 0 in the absence of a certain species and 1 in its presence). Thus, for each variable, one of the following settings must be observed: 0-0, 0-1, 1-0 or 1-1, the first value being relative to observation i and the second to observation j.

The coefficients of these variables normally focus on measuring (dis)similarity, based on counting the agreements (positive or negative) that exist between the elements. There are some coefficients that use the number of disagreements (positive or negative) as the main element of their measurement.

In general, measures of (dis)similarity are interrelated and easily transformable among themselves. There are a large number of similarity and/or dissimilarity coefficients for binary characters available in the literature. Such coefficients can be easily converted to dissimilarity coefficients. If the similarity is called s, the dissimilarity measure will be its complement between each pair of groups: the distances can, for example, be chosen as

$d_{ij} = 1 - s$, $d_{ij} = \sqrt{1-s}$ or $d_{ij} = 1 - s^2$, $d_{ij} = \sqrt{1-s^2}$ when $s_{ij}$ is a similarity coefficient. Most cluster analysis methods require a measure of (dis)similarity between the elements to be clustered, which usually express a distance function or a metric. The function can only be considered a similarity or a dissimilarity if it satisfies certain properties or axioms.

Distances are used, like similarities, in order to measure the association between objects. Distance coefficients can be subdivided into three groups.

The first group tends to all properties:

1.$d_{ij} = d_{j,i}$ (symmetrical) 2. $d_{ij} > 0$, if $i \neq j$; (positivity) 3. $d_{ij} = 0$, if and only if, i = j; (reflective) 4. $d_{ij} \leq d_{iz} + d_{zj}$ this is known as the triangular inequality. The second group of distances is symmetrical. These coefficients do not follow the triangle inequality axiom. The third group that contains non-metrics can lead to negative values, violating the positivity property. Researchers are, in principle, free to define and use any measure of association suitable for the phenomenon under study, however mathematics imposes some restrictions on this choice, which is why association coefficients are frequently found in the literature. Some of them are of great applicability, while others were created for specific needs. Successive coefficients have been rediscovered by several authors and may be known under different names.

## 2. Joint Absence

The similarity coefficients can be divided into two groups: those that consider the joint absence (symmetric coefficients) and those that do not consider the joint absence (asymmetric coefficients). Some similarity coefficients that consider the joint absence are presented below, emphasizing that it is indicated by the letter d (double zero) in the expressions.

An attribute is symmetric if both of its states are equally important and have the same weight. In these cases, the zero-zero and one-one correspondences are completely equivalent and must both be included in the similarity coefficient. Similarity, which is based on symmetrical attributes, is called similarity invariant, as the result does not change when some or all of the attributes are coded differently. For invariant similarity, the best-known coefficient for evaluating similarities between objects $x_i$ and $x_j$ is the coefficient of Sokal and Michener"(1958)" (Simple matching). Rogers and Tanimoto (1950), Russell and Rao (1940), and Gower and Legendre coefficients are other examples of symmetric similarity coefficients that treat positive (a) and negative (d) correspondences in the same way. The coefficients differ in the weights they assign to matches and to no matches.

Coefficient of similarity that considers the joint absence and is a metric, as it has all the properties of the axioms of dissimilarity. By using these coefficients, it assumes that there is no difference between presence (double 1) and absence (double 0).

**Sokal and Michener also called Simple matching and Simple matching** $S = \frac{a+d}{a+b+c+d}$

- No undetermined value
- It is a special case of proportion of agreement for two nominal variables
- Family parameter member $S = \frac{a+d}{a+\theta(b+c)+d}$ members are interchangeable with respect to an ordinal comparison
- It becomes after correction for the chance to use $E(a+d) = p_1 p_2 + q_1 q_2$
- D = 1 – S satisfies the triangle inequality
- Two multivariate generalizations satisfy a strong geneRaoization of the triangular inequality

**Table 1.** Table of binary proportions for binary variables.

| Two variables | | | | |
|---|---|---|---|---|
| a variable | Proportions | value 1 | value 2 | Total |
| | Value | a | b | $p_1$ |
| | Value | c | d | $q_1$ |
| | Total | $p_2$ | $q_2$ | 1 |

Source: Authors.

**2.1 Some coefficients are related.**

Hamann's Coefficient(1961) $S_{Hamann} = 2(S_{Soka\ e\ Michener}) - 1 = 2\left(\frac{a+d}{a+b+c+d}\right) - 1 = \frac{a-b-c+d}{a+b+c+d}$ McConnaughey (1964) = 2(Kulczynski (1927)) – 1= $2\left(\frac{1}{2}[\frac{a}{a+b} + \frac{a}{a+c}]\right) - 1 = \frac{a^2-bc}{(a+b)(a+c)}$. Since a, b, c, and d are proportional, Simple matching found $S_{SM}$= a + d. According to Simple matching, it can be interpreted as the number of 1s and 0s shared by variables in the same positions, divided by the total length of the variables. By comparing two clustering algorithms, to measure the agreement of two psychologists who classify people into undefined categories. For Sokal and Michene, Rogers and Tanimoto, Sokal and Sneath (1963) proposed the coeficiente $\frac{2(a+d)}{2a+b+c+2d}$, which gives twice as much weight to quantity (a + d) as compared to (b + c). Furthermore, Sokal and Sneath proposed the coefficients $\frac{1}{4}\left(\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{c+d} + \frac{d}{b+d}\right)$, $\frac{ad}{\sqrt{p_1p_2q_1q_2}}$ *and* $\frac{a+d}{b+c}$ . This last coefficient does not work well when the sum of a + d is greater than b + c, as they are not based on the axioms of the coefficients, which is to be greater than 1. As alternatives (which do not include the quantity d) there are the coefficients $\frac{1}{2}\left(\frac{a}{a+b} + \frac{a}{a+c}\right)$ Kulczynski (1927) and $\frac{a}{\sqrt{p_1p_2}}$ Ochiai (1957). The coefficient by Russell and Rao is called hybrid by Sokal and Sneath, since it includes the quantity d in the denominator but not in the numerator.

**Russell and Rao also called Positive Concordances** $\frac{a}{a+b+c+d}$

- No undetermined value
- D = 1 – S satisfies the triangle inequality
- Matrix coefficient is totally positive of order 2
- First eigenvector of the coefficient matrix reflects an ordering of a stochastic model.

The similarity coefficient considers the joint absence and is a semimetric, as it does not meet the fourth property of the dissimilarity axioms, Anderberg (1973) $\frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$; Gower 2 (1985), Ochiai II $\frac{ad}{\sqrt{(a+b)(a+c)(b+d)(d+c)}}$; Sneath and Sokal.

**2.2 Disregard the joint absence**

Given two asymmetric binary attributes, the agreement of two 1's (a positive match) is considered more significant than the agreement of two 0's (a negative match). Similarity based on such attributes is called non-invariant similarity, for which the best known coefficient is the Jaccard coefficient (1901), where the number of negative matches, d, is not considered important and is therefore ignored in the calculation. The Jaccard similarity index indicates the similarity between two communities, comparing the number of species between the areas used in its calculation and the numbers of species unique to each area and the number of species common between them.

**Jaccard also known as community coefficient** $s = \frac{a}{a+b+c}$ $d = \frac{b+c}{a+b+c}$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{P(A \cap B)}{P(A) + P(B) - P(A \cap B)} = \frac{p_1 p_2}{p_1 + p_2 - p_1 p_2} = \frac{a}{a + b + c}$$

Properties

- Undetermined value if d = 1
- Member of the family parameter $S_{GL1} = \frac{a}{[a+\theta(b+c)]}$; members are interchangeable with respect to an ordinal comparison.
- Limited by the correlation of the proportion below $S_{Sorgenfrei} = \frac{a^2}{p_1 p_2}$
- Satisfaz a desigualdade triangular $d_j = 1 - s_j$
- A multivariate generalization satisfies a strong generalization of triangle inequality.

Some similarity coefficients disregard joint absence and are a metric: Jaccard, Sorense-Dice, Andeberg, similarity coefficient that disregards joint absence and which is a semimetric.

Sørensen (1948) is similar to Jaccard but allows for a weight of 2 for concordances. If the variables involved are null and undefined, they are not recommended. A method of establishing plant breadth groups on species similarity and of its application in vegetation analysis in common terms:

$$S(A,B) = \frac{2|A \cap B|}{|A| + |B|} = \frac{2P(A \cap B)}{P(A) + P(B)} = \frac{2p_1 p_2}{p_1 + p_2} = \frac{2a}{2a + b + c}$$

or yet

$$S_{Sorense} = \frac{2a}{p_1 + p_2} = \frac{(a+d) - 1}{p_1 + p_2} + 1$$

- Undetermined value if d = 1
- Member of the family parameter $S_{GL1} = \frac{a}{[a+\theta(b+c)]}$; member are interchangeable with respect to an ordinal comparison.
- Special case of a coefficient of Czkanowski
- Limited below $S_{BB} = \frac{a}{\max(p_1 p_2)}$
- Limited above by $S_{DK} = \frac{a}{\sqrt{p_1 p_2}}$
- Becomes $S_{cohem}$ after the possibility of using correction $E(a+d) = p_1 p_2 + q_1 q_2$
- Three simple multivariate genes

The Kulczynski I coefficient is not a metric, and when the positive agreements are greater than the differences, they will not meet the properties. $0 \leq s(s_i, s_j) \leq 1$.

**2.3 Association coefficients**

Such coefficients show how the pairs of individuals are associated. They generally range from -1, when a change in one variable is accompanied by an equal-magnitude change in the other, but in the opposite direction, to +1 when a change in one variable is accompanied by an equal-magnitude change in the other. Hamann, Yule, this coefficient measures the strength of agreements in relation to disagreements, the closer to 1, the greater the similarity between the elements when they agree. The closer to –1, the greater similarity in relation to disagreement. There are also association coefficients that vary over the range [-∞, +∞], Thu – square. The correlation coefficient has been used successfully, precisely when it is intended that the results of the classification are not affected by differences in dispersion and scale of the variables.

Work was carried out to compare the similarity coefficients in species studies, which may help in the choice of these. Duarte et al. (1999), compares different coefficients of similarity in studies with beans and determines that the Sorensen coefficient is the most suitable for the study of genetic divergence for this species, when RAPD markers are used. Work was

carried out to compare the similarity coefficients in species studies, which may help in the choice of these. Duarte et al. (1999), compares different coefficients of similarity in studies with beans and determines that the Sorensen coefficient is the most suitable for the study of genetic divergence for this species, when RAPD markers are used. Meyer et al. (2004), using RAPD and AFLP markers in corn to compare similarity coefficients, demonstrates that, for this situation, the coefficients of Jaccard, Sorensen, Anderberg and Ochiai can be used, since the results for these coefficients showed little variation.

The same authors also point out that this result confirms the greater use of the Jaccard index in the analysis of genetic divergence, although it is not the most suitable for all species.

In all the coefficients used in these articles, a single method, UPGMA, and ten coefficients found in a program, NTSYS (Numerical Taxonomy and Multivariate System) version 1.7 (Rohlf, 1992), GENES (Cruz, 2001) were applied. Among these 10 coefficients, Simple concordance, Rogers and Tanimoto, Russel and Rao and Jaccard are metric and Sorense, Ochiai and Kulczynski are semimetric, there are also three association coefficients Hamann, Yule and Phi(Pearson). It is important to note association coefficients, as their values range from -1 to +1. Therefore, it is essential to verify the correlation values, because when they assume negative values, they do not represent a metric, although it is also possible to associate a distance function to the correlation coefficient. Mulvey and Crowder (1979) define a correlation "metric" based on the following transformation: $d_{ij}=[0{,}5(1-r_{ij})]^{1/2}$ "that all similarity coefficients were transformed and analyzed as a measure of distance and their properties were not verified, to analyze if they represent a metric". Other methods should be applied so that their behavior is analyzed and verified.

The choice of clustering methods must also be judicious. The different methods can produce different results in the same data, since the authors of the articles do the following: they find the similarity coefficients and transform them into dissimilarity, without observing the properties of similarity and dissimilarity, however some of these coefficients, such as the Sorensen, do not meet the property of triangular inequality (not being a metric, but a semimetric) and the correlation coefficients that vary from -1 to +1 cannot be made by the transformations that are used in the coefficients.

## 2.4 Parameter Families

Gower and Legendre (1986) define two parameter families in which all members are linear in the numerator and denominator. They make a distinction between coefficients that do not include the quantity d.A primeira família para dados de presença e ausência é determinada por:

$$S_{\text{Gower e Legendre}}1 = \frac{a}{a+\theta(b+c)} = \frac{a}{\theta(p_1+p_2)+(1-2\theta)a}$$

Where $\theta > 0$ , in order to obtain negative values. Members of the family:

$$S_{\text{Gower e Legendre}}(\theta = 1) = S_{Jaccard} = \frac{a}{p_1+p_2-a}$$

$$S_{\text{Gower e Legendre}}\left(\theta = \frac{1}{2}\right) = S_{Sorence\ e\ Dice} = \frac{2a}{p_1+p_2}$$

$$S_{\text{Gower e Legendre}}(\theta = 2) = S_{Sokal\ e\ Sneath} = \frac{a}{a+2(b+c)}$$

Members with $0 < \theta < 1$ gives more weight to a.

With attendance and absence data, this is regularly done. In the case of the relation where there is only some equality of positive numbers, that is, a is much smaller than (b + c) similar arguments can be used for the opposite case $\theta > 1$. All members $S_{\text{Gower e Legendre}}(\theta)$ $0 < \theta < 1$ and all given by 0 and 1, such that $0 \leq S_{\text{Gower e Legendre}}1\,(\theta) \leq 1$ adding the limbs to every other hop. $0 \leq S_{Sokal\ e\ Sneath} \leq S_{Jaccard} \leq S_{Sorence} \leq 1$

### 2.5 The family of coefficients

Consider a family $\mathcal{L}$ of form coefficient $S = \lambda + \mu(a + d)$, where propositions **a** and **d** are defined in table 1, and where $\lambda$ and $\mu$ are different for all coefficients, remember that this depends on the marginal probability of table 1. Since Simple matching, all family members are linear transformations of $S_{SM}$, the proportion and observation agree, given the marginal probabilities. Also, the Sokal and Michener, Rand, Hamann and Hubert, Sorenson coefficients are in the family.

So, the coeficiente $S_{sorense} = \frac{2a}{p_1+p_2} = \frac{(a+d)-1}{p_1+p_2} + 1$ can be written in the form $S_{cze} = \lambda + \mu(a + d)$, where $\lambda = \frac{-1}{p_1+p_2} + 1$ e $\mu = \frac{1}{p_1+p_2}$. The distance corresponding to the Sorensen coefficient was described by Barros et al. (2020) under the name of non-metric coefficient, used to compare dissimilarity of two samples: "a variant gives a double weight to the presence, because the presence of a species can be considered to be more informative than its absence". Absence may be due to various factors, as discussed above, but do not necessarily reflect differences in the environment. Dual presence, on the other hand, is a strong indication of similarity. Note, however, that Sorensen is monotonous for Jaccard. This property means that if the similarity of a pair of objects calculated with Jaccard is superior to that of the other pair of objects, the same will be true when using Sorensen.

It can be considered that the presence of a species is more informative than its absence. Absence can be caused by a number of factors, as discussed above, which do not necessarily reflect differences in the environment. Double-presence, on the contrary, is a strong indication of similarity. Note, however, that Sorensen is monotonous for Jaccard. This property means that if the similarity of a summarized pair of objects to Jaccard is greater than that of another pair of objects, the same will be true when using Sorensen. In other words, Jaccard and Sørensen only differ in their scales (weightings).

The choice of (dis)similarity coefficients for analyzing the results of an experiment must comply with criteria, so that the results presented are reliable. Each (dis)similarity coefficient has its own characteristics that must be taken into account, together with the individual or variable studied.

Little research has been carried out to determine the advantages and disadvantages of each of the (dis)similarity coefficients. In general, much of the work does not justify the choice of coefficients to be used. For greater fidelity, the works should contain a justification for the choice of (dis)similarity coefficients and the clustering methods used.

## 3. Material and Methods

Data from a survey of the vegetation of the Forestry Forest (Table 2), from the Federal University of Viçosa, in Viçosa, MG, taken from Albuquerque et al. (2006).

**Table 2 -** Density of 17 species from the forestry forest, in plots of 20 x 50 m, municipality of Viçosa-MG.

| Species | Portion | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | Total | Average |
| *Casearia decandra* Jacq. | 8 | 1 | 27 | 0 | 1 | 9 | 2 | 3 | 22 | 15 | 7 | 95 | 8,6 |
| *Anadenanthera peregrina* Speg. | 0 | 0 | 0 | 0 | 0 | 0 | 12 | 1 | 17 | 1 | 9 | 40 | 3,6 |
| *Apuleia leiocarpa* (Vog.) Macbr. | 3 | 9 | 4 | 6 | 22 | 9 | 5 | 2 | 7 | 4 | 4 | 75 | 6,8 |
| *Mabea fistulifera* Mart. | 6 | 3 | 3 | 4 | 29 | 12 | 0 | 4 | 4 | 4 | 4 | 73 | 6,6 |
| *Anadenanthera macrocarpa* (Benth.) Brenan. | 0 | 12 | 0 | 1 | 0 | 0 | 1 | 0 | 2 | 0 | 0 | 16 | 1,5 |
| *Platypodium elegans* Vog. | 0 | 0 | 1 | 1 | 9 | 1 | 0 | 0 | 5 | 11 | 1 | 29 | 2,6 |
| *Machaerium floridum* (Benth.) Ducke | 0 | 0 | 10 | 1 | 9 | 2 | 1 | 0 | 0 | 11 | 5 | 39 | 3,5 |
| *Copaifera lansdorffii* Desf. | 1 | 1 | 0 | 2 | 1 | 13 | 0 | 0 | 0 | 3 | 1 | 22 | 2,0 |
| *Ocotea pretiosa* Mez. | 2 | 0 | 2 | 2 | 2 | 6 | 0 | 5 | 0 | 2 | 2 | 23 | 2,1 |
| *CabRaoea cangerana* Saldanha | 1 | 0 | 0 | 2 | 0 | 0 | 1 | 6 | 2 | 3 | 1 | 16 | 1,5 |
| *Piptadenia gonoacantha* Macbr. | 0 | 0 | 0 | 0 | 0 | 0 | 6 | 0 | 1 | 0 | 5 | 12 | 1,1 |
| *Dalbergia nigra* Allem. ex Benth. | 5 | 0 | 7 | 0 | 5 | 0 | 0 | 0 | 0 | 1 | 0 | 18 | 1,6 |
| *Luehea divaricata* Mart. | 0 | 0 | 1 | 0 | 0 | 0 | 2 | 0 | 0 | 5 | 2 | 10 | 0,9 |
| *Cecropia hololeuca* Miq. | 7 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 9 | 0,8 |
| *Melanoxylon brauna* Schott. | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 2 | 1 | 3 | 0,3 |
| *Cedrela fissilis* Vell. | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0,1 |
| *Croton floribundus* Spreng. | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0,1 |

Source: Albuquerque et al. (2006).

The clustering coefficients used were Jaccard, Sorense, Simple Agreement, Russell and Rao and Rogers and Tanimoto and the Mean Distance Method. These coefficients were used because they are the most used in practice and because they are easy to find in the most diverse computer programs.

### 3.1 Average distance

This method consists of grouping the two most similar objects and then using the arithmetic mean of the distances of the objects in each group to create the new distance matrix.

The average similarity of the individuals or group that is intended to be joined to an existing group is used.

### 3.2 Comparison of coefficients

#### 3.2.1 Cophenetic correlation

For the various clustering coefficients used, the respective cophenetic matrices resulting from the simplification provided by the coefficients were obtained. Based on the original and cophenetic matrices, the cophenetic correlation was obtained, according to the expression (Albuquerque et al., 2006).

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{i=1}^{n}(c_{ij}-\bar{c})(d_{ij}-\bar{d})}{\sqrt{\sum_{i=1}^{n-1}\sum_{i=1}^{n}(c_{ij}-\bar{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{i=1}^{n}(d_{ij}-\bar{d})^2}}$$

on what: $c_{ij}$ = dissimilarity value between individuals i and j, obtained from the cophenetic matrix; and $d_{ij}$ = dissimilarity value between individuals i and j, obtained from the dissimilarity matrix.

## 4. Validation Coefficient

### 4.1 Rand

The adjusted Rand index determines the similarity between two plots $P_1$ and $P_2$ examining to which group pairs of species belong in the two groups. This means that if two species belong to the same group $P_1$ and $P_2$ the index value increases;

on the other hand, if the two species belong to the same group in $P_1$ but belong to a different group in $P_2$ the index value goes down. The adjusted Rand index is the normalized version of the Rand index, where: $k_{P_1}$ and $k_{P_2}$ are the number of parcel groups $P_1$ and $P_2$; n is the amount of data in the initial set; $n_i$ is the number of species in the group $C_i \in P_1$ and $n_j$ is the number of species in the group $C_j \in P_2$; $n_{ij}$ is the number of species that belong to the groups $C_i \in P_1$ and $C_j \in P_2$, that is, the number of species common to $P_1$ and $P_2$.

$$\text{Rand ajustado} = \frac{A - B}{C - D}$$

$$\text{Rand ajustado} = \frac{\sum_{i=1}^{k_{P_1}} \sum_{j=1}^{k_{P_2}} \binom{n_{ij}}{2} - \binom{n}{2}^{-1} \sum_{i=1}^{k_i} \binom{n_{i.}}{2} \sum_{j=1}^{k_{P21}} \binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{k_{P_1}} \binom{n_{i.}}{2} + \sum_{j=1}^{k_{P_2}} \binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1} \sum_{i=1}^{k_{P_1}} \binom{n_{i.}}{2} \sum_{j=1}^{k_{P_2}} \binom{n_{.j}}{2}}$$

Values close to 0 for the adjusted Rand index indicate random plots, which reveal little about the relationship between species, while values close to 1 are obtained by installments most relevant.

### 4.2 Stress

This statistical representation of stress (standardized residual sum of squares) was proposed by Kruskal (1964). It is a parameter that measures the distortion between the original matrix and the one obtained after the construction of the dendrogram.

**Table 3**. Stress rating.

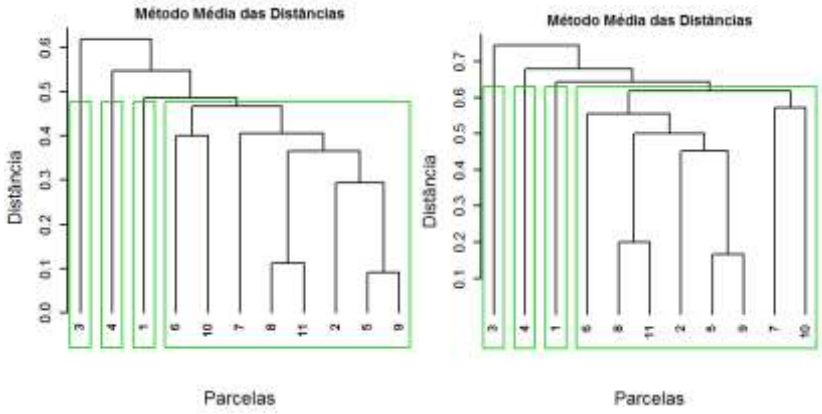
| Stress level (%) | Adjustment |
|---|---|
| 40 | Unsatisfactory |
| 20 | Regular |
| 10 | Good |
| 5 | Great |
| 0 | Perfect |

Source: Authors.

## 5. Results and Discussion

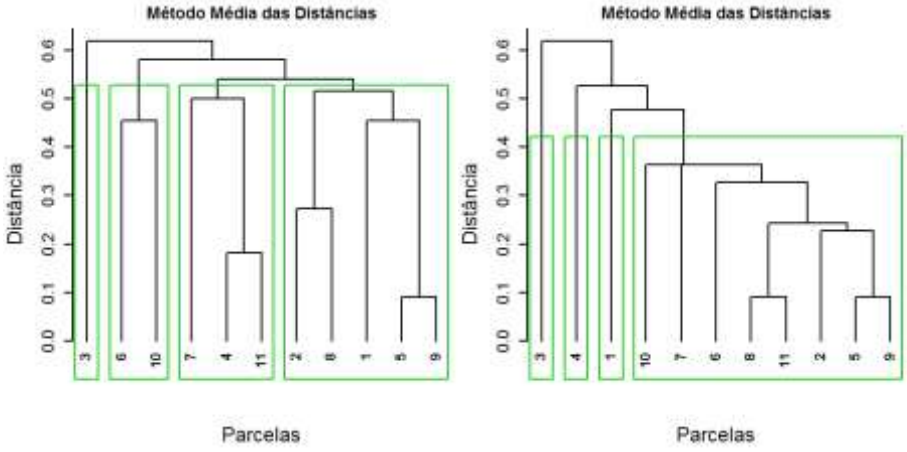
A visual observation of the dendrogram can be made based on the Figures 1. In general, the dendrograms presented similar grouping structures. Simple agreement and Rogers and Tanimoto, the same formation of groups is observed with similar cophenetic correlation and the adjusted Rand coefficient identical and the Russell and Rao asymmetry coefficient was the one that presented the ordering of the groups different from the other asymmetry coefficients and also the coephenetic value different from the other coefficients and with the adjusted Rand value equivalent to the other coefficients and all coefficients showed unsatisfactory stress.

**Figure 1.** Jaccard (left) and Sorense (right) asymmetry coefficient.



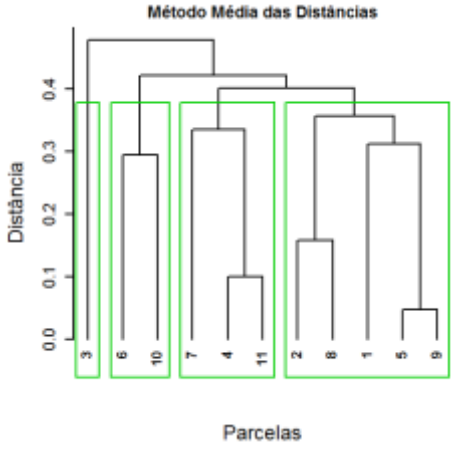

Source: Authors

**Figure 2.** Asymmetry coefficient Simple agreement (left) and Russell Rao (right).




Source: Authors

**Figure 3.** Rogers and Tanimoto asymmetry coefficient.




Source: Authors

Although the general structure of the groupings is very similar, it can be observed that there are small changes in the levels at which the parcels are grouped, that is, parcels that are within the same group can be grouped in another order, when the parcels are changed. coefficients. However, this causes few practical problems. It is important to highlight that the fact that

this type of analysis does not present an objective criterion for identifying the groups makes it very difficult to interpret the results.

**Tabela 4.** Stresses, cophenetic coefficient, Rand and adjusted Rand generated between similarity coefficients.

| Asymmetry coefficient | Coefficients cophenetic | Stresses % | Rand | Adjusted Rand |
|---|---|---|---|---|
| Jaccard | 0,55 | 34 | 0,91 | 0,76 |
| Sorensen | 0,57 | 25 | 0,98 | 0,95 |
| Simple agreement | 0,50 | 29 | 0,96 | 0,84 |
| Russell and Rao | 0,67 | 32 | 0,90 | 0,76 |
| Rogers and Tanimoto | 0,45 | 38 | 0,96 | 0,84 |

Source: Authors

The cophenetic correlation coefficients between the five similarity coefficients, for both plots, were all moderate, demonstrating that there is a reasonable association between the original data and the dendrograms and that this is independent of the coefficient used and the number of groups, with few changes (Table 4). The Jaccard with correlation at 0.54, Sorensen-Dice at 0.57, simple agreement at 0.50, Russell and Rao at 0.67 and Rogers and Tanimoto at 0.44, which indicates that there is change in ranks using either of these coefficients, that is, they rank the similarity between the plots in exactly the same order. It is observed that the cophenetic correlation does not allow making a clear distinction between the coefficients, regarding the dendrograms obtained.

The stress levels presented for the five coefficients (Table 4), for both plots, were of low magnitude. The stress level ranged from 33% for Jaccard 25%, for the Sorrese-Dice coefficient to 29% for the Simple Agreement coefficient, for the Russell and Rao coefficient 32% and the stress level ranged from 38% for the coefficient by Rogers and Tanimoto.

The Rand index takes values in the range [0, 1]. The maximum value (Rand = 1) will correspond to a situation where the two classifications coincide, with no pairs that are in the same group in one case, and in different groups in the other.

Since in this case the grouping of the 11 plots into 17 species is known, it is possible to compare the groupings obtained through the cluster analysis with this division by plots.No caso do agrupamento resultante do coeficiente Jaccard obtains a value of 0.91, Sorensen obtains a value of 0.96, concordance.simple obtains a value of 0.84, Russell and Rao obtains a value of 0.76, Rogers and Tanimoto obtains a value of 0.84 of the adjusted Rand index was used, while the Rand index values were similar for all similarity coefficients, noting that any similarity coefficient can be used when comparing the coefficients and by the average binding method, given that the Rand values are higher than the adjusted Rand values. Considering that the results were performed independently for each coefficient.

## 6. Conclusion

The practical conclusion is that, in most data applications, it must be observed that the properties of the similarity and dissimilarity coefficients are met and the choice of the correct coefficients, for the variables, can probably be limited to the following five coefficients: Jaccard, Sorensen, Russell Rao, Sokal Michener and Rogers Tanimoto.

## References

Albuquerque, M. A., Barros, K. N. N. O., Gouveia, J. F., & Ferreira, R. L. C. (2016). *Determination and validation of group numbers in a cluster analysis: A case study applied to forestry science. Acta Scientiarum.* Technology, 38(3), 339-344.

Albuquerque, M. A. D., Ferreira, R. L. C., Silva, J. A. A. D., Santos, E. D. S., Stosic, B., & Souza, A. L. D. (2006). Estabilidade em análise de agrupamento: estudo de caso em ciência florestal. *Revista árvore*, *30*, 257-265.

Barros, K. N. N., de Albuquerque, M. A., dos Santos Gomes, A., & Dantas, D. R. G. (2020). Análise de agrupamentos exploratória dos usuários do Programa Multidisciplinar de Tratamento do Tabagismo do HUAC, Campina Grande–PB. *Research, Society and Development*, *9*(8), e825986532-e825986532.

Gower, J. C., & Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. Journal of classification, 3(1), 5-48.

Hamann, U. The principal components of scalable attitudes. In P. F. Lazarsfeld (Ed.), (1961). *Mathematical Thinking in the Social Sciences.Glencoe*: free Press. 216-257.

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull Soc Vaudoise Sci Nat*, *37*, 547-579.

Gordeev, S. Y., & Gordeeva, T. V. (2021, November). Species of Orthosia Ochsenheimer, 1816 and Anorthoa Berio, 1980 genera (Lepidoptera, Noctuidae) in Western Transbaikalia. In *IOP Conference Series: Earth and Environmental Science* (Vol. 908, No. 1, p. 012015). IOP Publishing.

Ochiai,( 1957). A. Zoographic studies on the soleoid fishes found in Japan and its neighboring regions. *Bulletin of the Japanese Society of Scientific Fish*eries, v.22: 526-530.

McConnaughey, B. H. (1964). *The determination and analysis of plankton communities*. Lembaga Penelitian Laut..

Meyer, A. D. S., Garcia, A. A. F., Souza, A. P. D., & Souza Jr, C. L. D. (2004). Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L). *Genetics and Molecular Biology*, *27*(1), 83-91.

Mulvey, J. M., & Crowder, H. P. (1979). Cluster analysis: An application of Lagrangian relaxation. *Management Science*, *25*(4), 329-340.

Russell, P. F., & Rao, T. R. (1940). On habitat and association of species of anopheline larvae in south-eastern Madras. *Journal of the Malaria Institute of India*, *3*(1).

Souza, A. L. D., & Souza, D. R. D. (2006). Análise multivariada para estratificação volumétrica de uma floresta ombrófila densa de terra firme, Amazônia Oriental. *Revista Árvore*, *30*, 49-54.

Sneath, P. H., & Sokal, R. R. (1973). *Numerical taxonomy. The principles and practice of numerical classification*.

Sokal, R. R. (1958). A statistical method for evaluating systematic relationships. *Univ. Kansas, Sci. Bull.*, *38*, 1409-1438.

Sorensen T. A. (1948). method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons. *Vidensk Selsk Biol Skr* 5:1-34.

Watson, D. H., Shedden, W. I. H., Elliot, A., Tetsuka, T., Wildy, P., Bourgaux-Ramoisy, D., & Gold, E. (1966). Virus specific antigens in mammalian cells infected with herpes simplex virus. *Immunology*, *11*(4).