COMPARAÇÃO ENTRE COEFICIENTES SIMILARIDADE UM APLICAÇÃO EM CIÊNCIAS FLORESTAIS

Macio Augusto Albuquerque

Universidade Estadual da Paraíba

macioaa@uol.com.br

Edwirde Luiz Silva

Universidade Estadual da Paraíba

edwirde@uepb.edu.br

Kleber Napoleão N. Oliveira Barros

Universidade Estadual da Paraíba

kleberbarros@cct.uepb.edu.br

Silvio Fernando Alves Xavier Junior

Universidade Estadual da Paraíba

silvioxj@gmail.com

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 também 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 validação 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.

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 co-phenetic 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.

Palavras-chave: Análise de agrupamento; validação; dados florestais..

1 INTRODUÇÃO

As técnicas de estatística multivariada têm sido amplamente empregadas em estudos florestais envolvendo simultaneamente variáveis de clima, de solo, de relevo e de vegetação. Essas técnicas são utilizadas, a fim de ordenar, visando a determinar a influência de fatores do meio na composição e na produtividade do local, e de agrupar, com o propósito de classificação [1]. Quando o objetivo é a classificação de grupos, um grande número de coeficientes de similaridade e dissimilaridade que são encontrados na literatura como [2]; [3] e [4] concordância simples, [5] e [6], assim sendo é possível observar diferentes coeficientes utilizados com o mesmos ou diferentes propósitos. No entanto, nem todos os autores justificam a razão da escolha de determinado coeficiente, ou seja a escolha é subjetiva e pode comprometer a natureza da análise. As medidas de semelhança são grandezas numéricas que quantificam o grau de associação entre os pares de objetos, indivíduos, itens, etc. Uma medida s_{ij} é considerada de similaridade se, para todo x_i e x_j obedece as seguintes propriedades: para $i \neq j$ a similaridade está entre 0 e 1, enquanto que s_{ij} é igual a 1 para $i = j$. A obtenção dos cálculos e da estrutura da resultante da análise numérica é a da matriz de associação, que não necessariamente reflete todas as informações originalmente contidas na matriz de dados, pois os objetos ou os descritores são representados em espaço reduzidos. Isso ressalta a importância da escolha de uma adequada medida de associação e determina o tema da análise. Por isso deve-se levar em conta as seguintes considerações:

1. A natureza do estudo (ou seja, a questão inicial e a hipótese) determina o tipo de estrutura que deve ser evidenciado por meio de uma matriz de associação e, consequentemente, o tipo de medida de (dis)similaridade a ser utilizado.

2. As medidas são representadas por diferentes equações matemáticas e, na análise da matriz de associação, muitas vezes são exigidas coeficientes com propriedades matemáticas específicas.

3. É preciso considerar também o aspecto computacional, e, portanto, a escolha do coeficiente, muitas vezes depende da sua disponibilidade no pacote computacional ou da facilidade do usuário em programá-lo.

Considere a comparação de um par de elementos (i e j) a partir dos resultados de q variáveis binárias, cada uma codificada de tal forma que possa assumir valores 0 ou 1 (por exemplo, 0 no caso de ausência de determinado espécie e 1 em sua presença). Dessa forma, para cada variável, uma das seguintes configurações deve ser observada: $0-0$, $0-1$, $1-0$ ou $1-1$, sendo o primeiro valor relativo à observação i e o segundo a observação j.

Os coeficientes dessas variáveis normalmente se concentram em medir a (dis)similaridade, baseados na contagem das concordâncias (positivas ou negativas), existentes entre os elementos. Existem alguns coeficientes que utilizam como principal elemento de sua medida o número de discordâncias (positiva ou negativa).

De um modo geral, as medidas de (dis)similaridade são interrelacionadas e facilmente transformáveis entre si. Há um grande número de coeficientes de similaridade e/ou de dissimilaridade para caracteres binários disponíveis na literatura. Se a similaridade for denominada s_{ij}, a medida de dissimilaridade será o seu complementar entre cada par de grupos: as distâncias podem, por exemplo, ser escolhidas como $d_{ij} = 1 - s_{ij}$, $d_{ij} = \sqrt{1 - s_{ij}}$, $d_{ij} = 1 - s_{ij}^2$ e $d_{ij} = \sqrt{1 - s_{ij}^2}$. A maioria dos métodos de análise de agrupamento requer uma medida de (dis)similaridade entre os elementos a serem agrupados, que normalmente expressam uma função de distância ou uma métrica. A função só pode ser considerada uma similaridade ou uma dissimilaridade se satisfizer determinadas propriedades ou axiomas.

As distâncias são usadas, como semelhanças, a fim de medir a associação entre objetos. Os coeficientes de distâncias podem ser subdivididos em três grupos. O primeiro grupo tende a todas as propriedades:

1. $d_{ij} = d_{ji}$ (simétrica)
2. $d_{ij} > 0$, se $i \neq j$; (positividade)
3. $d_{ij} = 0$, se e somente se, $i = j$; (Reflexiva)
4. $d_{ij} \leq d_{iz} + d_{zj}$ essa é conhecida como a desigualdade triangular

O segundo grupo de distâncias é o simétrico. Esses coeficientes não seguem o axioma de desigualdade de triângulo. O terceiro grupo que contém não métricas pode levar a valores negativos, violando a propriedade de positividade. Pesquisadores são, em princípio, livres para definir e utilizar qualquer medida de associação adequada para o fenômeno em estudo, todavia a matemática impõe algumas restrições a esta escolha, por isso que os coeficientes de associação são frequentemente encontrados na literatura. Alguns deles são de grande aplicabilidade, enquanto outros foram criados para necessidades específicas. Sucessivos coeficientes foram redescobertos por vários autores e podem ser conhecidos sob diversos nomes.

O objetivo deste trabalho foi pesquisar a influência da escolha de diferentes coeficientes de similaridades sobre a subsequente análise de agrupamento aplicando o método a dados sobre a densidade de 17 espécies de mata de silvicultura da Universidade Federal de Viçosa, Viçosa - MG.

1.1 AUSÊNCIA CONJUNTA

Os coeficientes de similaridades podem ser divididos em dois grupos: os que consideram a ausência conjunta (coeficientes simétricos) e os que não consideram a ausência conjunta (coeficientes assimétricos). Alguns coeficientes de similaridade que consideram a ausência conjunta são apresentados abaixo, ressaltando que ela é indicada pela letra d (duplo zero) nas expressões.

Um atributo é simétrico se ambos os seus estados são igualmente importante e têm a mesma ponderação. Nestes casos, as correspondências zero-zero e um-um são completamente equivalentes e devem ser ambas incluídas no coeficiente de similaridade. A

similaridade, que é baseada em atributos simétricos, é chamada de similaridade invariante, pois o resultado não muda quando alguns ou todos os atributos são codificados de forma diferente. Para similaridade invariante, o coeficiente mais conhecido para avaliar as similaridades entre os objetos x_i e x_j é o coeficiente de Simple matching [7]. Os coeficientes de [6], [5] e [8] são outros exemplos de coeficientes de similaridade simétricos que tratam as correspondências positivas e negativas da mesma forma. Os coeficientes diferem nas ponderações que atribuem às correspondências e às ausências de correspondências.

O coeficiente de Sokal e Michener S_{SM} [7] também chamado Concordância Simples (Simple Matching) é uma medida estatística muito utilizada para comparar a similaridade ou diversidade dos dados. Ele utiliza o número total de concordâncias dos atributos da amostra. Dados dois elementos quaisquer da amostra A e B, cada qual contendo n categorias binárias, o S_{SM} é definido por

$$S_{SM} = \frac{a+d}{a+b+c+d}$$

em que, conforme Tabela 1,

$a :=$ corresponde à proporção de categorias em que A e B tem valor 1.
$b :=$ corresponde à proporção de categorias em que ambos $A = 0$ e $B = 1$.
$c :=$ corresponde à proporção de categorias em que $A = 1$ e $B = 0$.
$d :=$ corresponde à proporção de categorias em que ambos A e B tem valor 0.

TABELA 1: Notação de proporções binárias para variáveis binárias

		B		
	Proporções	0	1	Total
A	0	a	b	p_A
	1	c	d	q_A
	Total	p_B	q_B	$n = p_A + p_B + q_A + q_B$

O índice de concordância simples é um caso especial de proporções de concordâncias para duas variáveis nominais. É membro do parâmetro família $S = (a+d)/[a+\theta(b+c)+d]$, cujos membros são intercambiáveis com relação a uma comparação ordinal. Simple matching pode ser interpretado como o número de 1s e 0s compartilhado pelas variáveis nas mesmas posições, divididas pelo comprimento total das variáveis. Por comparação de dois algoritmos de agrupamento, para medir o acordo de dois psicologistas que classificam pessoas em categorias não definidas. Sokal e Sneath [9] propuseram o coeficiente $S_{SS} = 2(a+b)/(2a+b+c+2d)$, que dá duas vezes mais peso a quantidade $(a+d)$ quando comparadas com $(b+c)$ [6, 7].

Além disso, estes mesmos autores propuseram os coeficientes

$$\frac{1}{4}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{c+d}+\frac{d}{b+d}\right), \quad \frac{ad}{\sqrt{p_1p_2q_1q_2}} \text{ e } \frac{a+d}{b+c}.$$

Esse último coeficiente não funciona bem quando a soma de a + d for maior que b + c, pois os mesmos fogem dos axiomas dos coeficientes que é ser maior que 1. Como alternativas (que não incluem a quantidade d) existem os coeficientes $\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$ [10] e $\frac{a}{\sqrt{p_1p_2}}$ [11]. O coeficiente por [5] é chamado híbrido por [12], uma vez que inclui a quantidade d no denominador, mas não no numerador. Russel e Rao também chamado de Concordâncias Positivas $\frac{a}{a+b+c+d}$.

1.2 DESCONSIDERAM A AUSÊNCIA CONJUNTA

Dados dois atributos binários assimétricos, a concordância de dois $1's$ (correspondência positiva) é considerado mais significativo que a concordância de dois $0's$ (uma correspondência negativa). A similaridade baseada em tais atributos é chamada similaridade

não-invariante, para ela, o coeficiente mais conhecido é o coeficiente de Jaccard [3], onde o número de correspondências negativas, d, não é considerado importante e por isso é ignorado no cálculo. O índice de similaridade Jaccard indica a semelhança entre duas comunidades, comparando o número de espécies entre as áreas utilizadas em seu cálculo e os números de espécies exclusivas para cada área e o número de espécies comuns entre elas.

O coeficiente de similaridade de Jaccard, também conhecido como coeficiente de comunidade. Esta medida de similaridade é definida pela razão entre o número de elementos da intersecção e o número de elementos da união:

$$S_{Jac} = \frac{|A \bigcap B|}{|A \bigcup B|} = \frac{P(A \bigcap B)}{P(A) + P(B) - P(A \bigcap B)} = \frac{p_1 p_2}{p_1 + p_2 - p_1 p_2} = \frac{a}{a + b + c} \quad (1)$$

com coeficiente de dissimilaridade $d = (b + c)/(a + b + c)$.

O índice S_{Jac} é membro da família de coeficientes $S_{GL1} = a/[a + \theta(b + c)]$. Membros desta família são intercambiáveis com respeito a uma comparação ordinal. S_{Jac} tem limite inferior determinado pela proporção $a^2/(p_1 p_2)$. Satisfaz a desigualdade triangular $d_j = 1 - s_j$. Uma generalização multivariada satisfaz uma generalização forte da inequalidade do triângulo. Alguns coeficientes de similaridade desconsideram a ausência conjunta e são uma métrica: Jaccard, Sorense-Dice, Andeberg, coeficiente de similaridade que desconsidera a ausência conjunta e que é uma semimétrica.

O coeficiente de Sorensen [4], é similar ao Jaccard, porém permitindo peso 2 para concordâncias. Caso as variáveis envolvidas sejam nulas e indefinidas, a aplicação do índice é não recomendada. Sua

$$S_{Sor} = \frac{2|A \bigcap B|}{|A| + |B|} = \frac{2P(A \bigcap B)}{P(A) + P(B)} = \frac{2p_1 p_2}{p_1 + p_2} = \frac{2a}{2a + b + c} \quad (2)$$

Este coeficiente é membro da família de coeficientes $S_{GL1} = a/[a+\theta(b+c)]$. Membros desta família são intercambiáveis com respeito a uma comparação ordinal. Também é um caso especial de um coeficiente de Czekanowski [13]. É limitado abaixo por $S_{BB} = \frac{a}{max(p_1 p_2)}$ e acima por $S_{DK} = \frac{a}{\sqrt{(p_1 p_2)}}$.

1.3 COEFICIENTES DE ASSOCIAÇÃO

Tais coeficientes mostram como os pares de indivíduos estão associados. Geralmente variam de -1, quando a mudança em uma variável é acompanhada por alteração de igual magnitude na outra, porém em sentido contrário, $a+1$ quando a mudança em uma variável é acompanhada por mudança de igual magnitude na outra. Esse coeficiente mede a força das concordâncias em relação às discordâncias, quanto mais próximo de 1, maior será a similaridade entre os elementos quando da concordância. Quanto mais próximo de menos 1, maior semelhança em relação à discordância. Há também coeficientes de associação que variam no intervalo $[-\infty, \infty]$ qui-quadrado. O coeficiente de correlação tem sido utilizado com sucesso, precisamente quando se pretende que os resultados da classificação que não sejam afetados por diferenças de dispersão e de escala das variáveis.

Trabalhos foram feitos, para se comparar os coeficientes de similaridade em estudos de espécies, que podem auxiliar na escolha destes. Duarte, Santos e Melo [14] comparam diferentes coeficientes de similaridade em estudos com feijão e determina que o coeficiente de Sorensen apresenta-se como o mais adequado para estudo de divergência genética para essa espécie, quando utilizados marcadores RAPD. Os marcadores de Russel e Rao não se adequam para essa espécie e para marcadores dominantes, enquanto os coeficientes de Sorensen [15], Ochiai e Kulczynski são os mais adequados para o estudo de divergência genética em feijão por meio de marcadores RAPD. Utilizando marcadores RAPD e AFLP

em milho para comparação de coeficientes de similaridade [16], é demonstrado que, para esta situação, pode-se utilizar os coeficientes de Jaccard, Sorensen, Anderberg e Ochiai, já que os resultados para estes coeficientes apresentaram pouca variação. Os mesmos autores salientam, ainda, que esse resultado vem corroborar a maior utilização do índice de Jaccard nas análises de divergência genética, apesar de não ser o mais indicado para todas as espécies.

Em todos os coeficientes utilizados nesses artigos, foram aplicados um único método, UPGMA, além dos coeficientes encontrados no programa NTSYS (Numerical Taxonomy and Multivariate System) versão 1.7 [17] e [18]. Entre esses 10 coeficientes, Concordância simples, Rogers e Tanimoto, Russel e Rao e Jaccard são métricas e Sorense, Ochiai e Kulczynski são semimetricas, existem também três coeficientes de associação Hamann, Yule e Phi. É importante observar os coeficientes de associação, pois os seus valores variam de -1 a $+1$.

Por conseguinte é fundamental verificar os valores das correlações, pois, quando os mesmos assumem valores negativos, não representam uma métrica, apesar disso é possível também associar ao coeficiente de correlação uma função de distância. [19] definem uma "métrica" de correlação a partir da seguinte transformação $d_{ij} = [0,5(1 - r_{ij})]^{1/2}$ que todos os coeficientes similaridade foram transformados e analisados como medida de distância e não foram verificadas as propriedades deles, para analisar se os mesmos representam uma "métrica". Deveriam ser aplicados outros métodos, a fim de que estes tenham seus comportamentos analisados e verificados. A escolha dos métodos de agrupamento também deve ser criteriosa. Os diferentes métodos podem produzir diferentes resultados nos mesmos dados, uma vez que os autores dos artigos fazem da seguinte maneira: encontram os coeficientes de similaridade e os transformam em dissimilaridade, sem observar as propriedades de similaridade e dissimilaridade, todavia alguns desses coeficientes, como Sorensen, não atendem a propriedade da desigualdade triangular (não sendo uma métrica, e sim uma semimétrica) e os coeficientes de correlação que variam de -1 a $+1$ não podem ser feitos pelas transformações que são utilizadas nos coeficientes.

1.4 RELAÇÕES ENTRE COEFICIENTES DE CONCORDÂNCIA

Alguns coeficientes de concordância estão relacionados de tal forma que a partir de um, pode-se obter outro e vice-versa. As principais relações são entre tais índices é mostrada a seguir.

O coeficiente de Hamann S_H [20] é definido por

$$S_H = 2S_{SM} - 1 = \frac{a - b - c + d}{a + b + c + d}.$$

em que S_{SM} é o coeficiente de similaridade de Sokal e Michener. O coeficiente de similaridade de Hamann foi proposto inicialmente para quantificar dados taxonômicos sem associação genética.

O coeficiente de McConnaughey S_{Mcc} [21] e [22] foi determinado num contexto filogenético para discriminar comunidades de plâncton. O coeficiente de Kulczynski S_{Kul} [23] é um índice apropriado para medir a similaridade entre matrizes de abundância de espécies. O coeficiente S_{Kul} foi utilizado com sucesso para medir a similaridade em genótipos de feijão mediante marcadores RAPD [24]. Estes dois coeficientes se relacionam pela seguinte igualdade:

$$S_{Mcc} = 2S_{Kul} - 1 = \frac{a^2 - bc}{(a + b)(a + c)}.$$

1.5 FAMÍLIAS DE PARÂMETRO

Gower e Legendre definiram duas famílias de parâmetros na qual todos os membros são lineares no numerador e no denominador [8]. Eles fazem uma distinção entre coeficientes

que não incluem a quantidade d. A primeira família para dados de presença e ausência é determinada por:

$$S_{GL} = \frac{a}{a + \theta(b + c)} = \frac{a}{\theta(p_1 + p_2) + (1 - 2\theta)a},$$

em que $\theta > 0$. Alguns membros da família são:

$$S_{GL}(\theta = 1) = S_{Jac} = \frac{a}{p_1 + p_2 - a}, \tag{3}$$

$$S_{GL}(\theta = \frac{1}{2}) = S_{Sor} = \frac{2a}{p_1 + p_2}, \tag{4}$$

$$S_{GL}(\theta = 2) = S_{SS} = \frac{a}{a + 2(b + c)}. \tag{5}$$

O parâmetro θ confere peso a medida de concordância a. Se $0 < \theta < 1$ um maior peso será dado para a em relação a b, c e d.

1.6 A FAMÍLIA DOS COEFICIENTES

Considere uma família λ de coeficiente da forma $S = \lambda + \mu(a + d)$, onde as proposições **a** e **d** são definidas na 1, e onde λ e μ são diferentes para todos os coeficientes, lembre-se que isso depende da probabilidade marginal da 1. Desde Simple matching, todos os membros da família são transformações lineares de S_{SM}, a proporção e observação estão de acordo, dadas as probabilidades marginais. Além disso, os coeficientes Sokal e Michener, Rand, Hamann e Hubert , Sorenson, são da família.

Assim, o coeficiente $S_{Sor} = \frac{2a}{p_1 p_2} = \frac{(a + d) - 1}{p_1 p_2} + 1$ pode ser escrito na forma $S_{Cze} = \lambda + \mu(a + d)$, onde $\lambda = \frac{-1}{p_1 p_2} + 1$ e $\mu = \frac{1}{p_1 p_2}$.

A distância correspondente ao coeficiente de Sorensen foi descrito por [25] sob o nome de coeficiente não métrico, usado para comparar dissimilaridade de duas amostras: “uma variante dá um duplo peso para a presença, porque pode-se considerar que a presença de uma espécie é mais informativa do que a sua ausência”. Ausência pode ser devido a vários fatores, como discutido acima, mas não necessariamente refletem as diferenças no ambiente. Dupla-presença, por outro lado, é um forte indício de semelhança. Note, no entanto, que Sorensen é monótono para Jaccard. Essa propriedade significa que se a semelhança de um par de objetos calculados com Jaccard é superior ao do outro par de objetos, o mesmo será verdade quando se utiliza Sorensen.

Pode-se considerar que a presença de uma espécie é mais informativa do que a sua ausência. A ausência pode ser causada por vários fatores, como discutido acima, o que necessariamente não reflete as diferenças no ambiente. Dupla-presença, ao contrário, é um forte indício de semelhança. Note, no entanto, que Sorensen é monótona para Jaccard,. Essa propriedade significa que, se a semelhança de um par de objetos sumarizada com Jaccard é maior do que outro par de objetos, o mesmo será verdade quando utilizar Sorensen. Em outras palavras, Jaccard e Sorensen só diferem por suas escalas (ponderações).

A escolha dos coeficientes (dis)similaridade, para análise dos resultados de um experimento, deve obedecer a critérios, a fim de que os resultados apresentados sejam confiáveis. Cada coeficiente de (dis)similaridade possui características próprias que devem ser levadas em conta, juntamente com o indivíduo ou variável estudada.

Poucas pesquisas foram realizadas para determinar as vantagens e desvantagens de cada um dos coeficientes de (dis)similaridade. De uma forma, em geral, muito dos trabalhos não justificam a escolha dos coeficientes a serem empregados. Para uma maior fidelidade, os trabalhos deveriam conter uma justificativa para a escolha dos coeficientes de (dis)similaridade e dos métodos de agrupamento utilizados.

2 MATERIAL E MÉTODO

Foram utilizados dados de um levantamento da vegetação da Mata da Silvicultura (2), da Universidade Federal de Viçosa, em Viçosa, MG, retirado de [26].

TABELA 2: Densidade de 17 espécies da mata da silvicultura, em parcelas de $20 \times 50\ m$, Universidade Federal de Viçosa, Viçosa-MG

Espécies	Parcelas												
	1	2	3	4	5	6	7	8	9	10	11	Total	Média
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

Os coeficientes de agrupamento utilizados foram Jaccard, Sorense, Concordância simples, Russel e Rao e Rogers e Tanimoto e o Método das Médias das Distâncias. Esses coeficientes foram utilizados por serem os mais usados na prática e pela facilidade de serem encontrados nos mais diversos programas computacionais.

2.1 DISTÂNCIA MÉDIA (AVERAGE LINKAGE)

Este método consiste em agrupar os dois objetos mais semelhantes e na sequencia utilizar a média aritmética das distâncias dos objetos de cada grupo para confeccionar a nova matriz de distâncias. Utiliza-se a similaridade média dos indivíduos ou grupo que se pretende unir a um grupo já existente.

2.2 COMPARAÇÃO DOS COEFICIENTES

2.2.1 CORRELAÇÃO COFENÉTICA

A correlação cofenética mede o grau de ajuste entre a matriz de dissimilaridade original (matriz D) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz C). No caso, C é aquela obtida após a construção do dendrograma. Tal correlação foi calculada usando [27]:

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

em que

c_{ij} : valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética;

d_{ij} : valor de dissimilaridade entre os indivíduos i e j, obtidos a partir da matriz de dissimilaridade;

$$\bar{c} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij};$$

$$\bar{d} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij}.$$

Nota-se que essa correlação equivale à correlação de Pearson entre a matriz de dissimilaridade original e aquela obtida após a construção do dendrograma. Assim quanto mais próximo de 1, menor será a distorção provocada pelo agrupamento dos indivíduos com os métodos.

2.2.2 COEFICIENTE DE VALIDAÇÃO

O índice de Rand ajustado $(Rand_{aj})$ determina a semelhança entre duas parcelas P_1 e P_2 examinando a qual grupo pares de espécies pertencem nos dois grupos. Isso quer dizer que se duas espécies pertencerem ao mesmo grupo P_1 e P_2 o valor do índice aumenta; por outro lado, se as duas espécies pertencerem, ao mesmo grupo em P_1 mas pertencem a grupo diferentes em P_2 o valor do índice diminui. O índice de Rand ajustado é a versão normalizada do índice Rand, onde: $k_{(P_1)}$ e $k_{(P_2)}$ são o número de grupos das parcelas P_1 e P_2; n é a quantidade de dados do conjunto inicial; n_i é o número de espécies do grupo $C_i \in P_1$ e n_j é o número de espécies do grupo $C_j \in P_2$; n_{ij} é o número de espécies que pertencem aos grupos $C_i \in P_1$ e $C_j \in P_2$, ou seja, o número de espécies comuns a P_1 e P_2.

$$Rand_{aj} = \frac{\sum_{i=1}^{k_{p1}} \binom{n_{ij}}{2} - \binom{n}{2}^{-1} \sum_{i=1}^{k_i} \binom{n_{i.}}{2} \sum_{j=1}^{k_{p21}} \binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{k_{p1}} \binom{n_{i.}}{2} + \sum_{j=1}^{k_{p2}} \binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1} \sum_{i=1}^{k_{p1}} \binom{n_{i.}}{2} \sum_{j=1}^{k_{p2}} \binom{n_{.j}}{2}} \tag{6}$$

Valores próximos a 0 para índice de Rand ajustado indicam parcelas aleatórias, que pouco revelam sobre a relação entre as espécies, enquanto valores próximos a 1 são obtidos por parcelas mais relevantes.

2.3 ESTRESSE (STRESS)

O valor do estresse (S), foi calculado por [28]:

$$S = \sqrt{\frac{\sum_{i=1}^{n-1} \sum_{j=2}^{n} (s_{ij} - c_{ij})^2}{\sum_{i=1}^{n-1} \sum_{j=2}^{n} s_{ij}}}$$

em que: c_{ij} = valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética; e s_{ij} = valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade.

Esta representação estatística de estresse (soma de quadrados de resultados padronizados), é uma parametrização que determina a precisão de ajuste da projeção gráfica. Neste caso, foi usada para determinar a precisão do ajuste obtido com a da projeção da matriz de similaridade no dendrograma. O estresse foi classificado com os critérios apresentados na Tabela 3.

Esta representação estatística do estresse (padronizado soma residual dos quadrados) foi proposta por [28]. É um parâmetro que mede a distorção entre a matriz original e aquela obtida após a construção do dendrograma.

TABELA 3: Classificação do estresse

Nível de estresse (%)	Ajuste
40	Insatisfatório
20	Regular
10	Bom
5	Excelente
0	Perfeito

3 RESULTADOS E DISCUSSÃO

Dendrogramas para diversos coeficientes de similaridade, utilizando ligação mediana, nos dados das parcelas podem ser visualizados nas Figuras 1, 2 e 3. De forma geral, os dendrogramas apresentaram estruturas de agrupamento similares. Para os coeficientes de concordância simples e Rogers e Tanimoto, observa-se a mesma formação de grupos com a correlação cofenética semelhante e o coeficiente de Rand ajustado idêntico e coeficiente de assimetria de Russel e Rao foi o que apresentou ordenamento dos grupos diferente dos demais coeficientes de assimetria e também o valor coefenética diferente dos outros coeficientes e com o valor de Rand ajustado equivalente, aos demais coeficientes se apresentaram com estresse insatisfatório (Tabela 4).

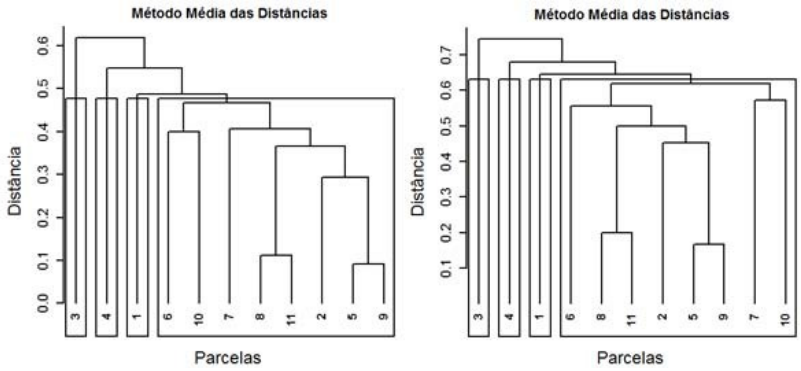

FIGURA 1: Dendrogramas representando as sequências das fusões das parcelas, obtidos pelo emprego do método das médias da distância, com base nos Coeficiente de assimetria Jaccard (esquerda) e Sorense (à direita)

Embora a estrutura geral dos agrupamentos seja bastante parecida, pode-se observar que há pequenas alterações nos níveis em que as parcelas são agrupadas, ou seja, as parcelas que estão dentro de um mesmo grupo podem ser agrupadas em outra ordem, quando se mudam os coeficientes. Entretanto, isso causa poucos problemas práticos. É importante destacar que o fato desse tipo de análise não apresentar um critério objetivo para identificação dos grupos dificulta muito a interpretação dos resultados.

Os coeficientes de correlação de cofenética entre os cinco coeficientes de similaridade, para ambos as parcelas, foram todos moderado, demonstrando que há uma razoável associação entre dados original e os dendrogramas e que isso independe do coeficiente usado e do número de grupos, com poucas alterações (Tabela 4). O Jaccard com correlação a $0,55$, Sorensen-Dice a $0,57$, concordância simples $0,50$, Russel e Rao a $0,67$ e Rogers e Tanimoto a $0,45$, o que indica que há mudança na fileiras usando qualquer um destes coeficientes, isto é, que classificam a similaridade entre as parcelas exatamente na mesma ordem. Observa-se que a correlação cofenética não permite fazer uma clara distinção entre os coeficientes,

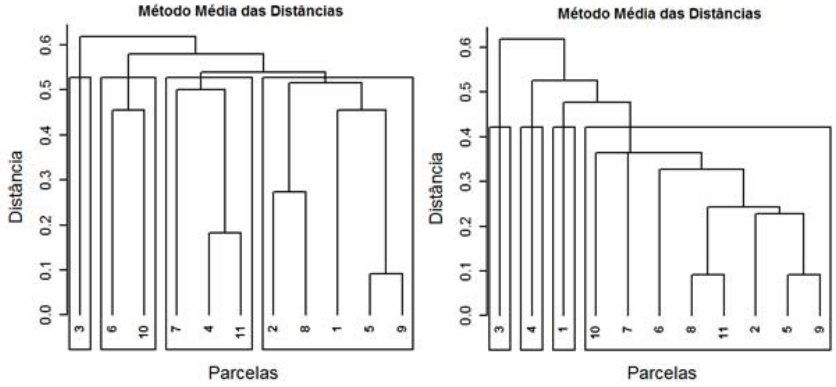

FIGURA 2: Dendrogramas representando as sequencias das fusões das parcelas, obtidos pelo emprego do método das médias da distância, com base nos Coeficiente de assimetria Concordância simples (esquerda) e Russel Rao (à direita)

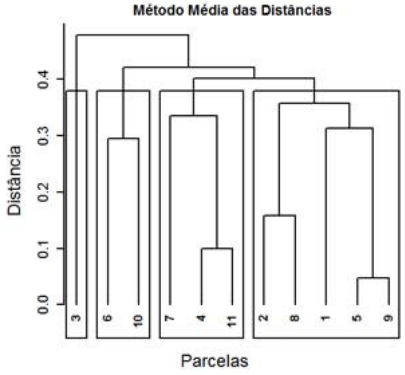

FIGURA 3: Dendrogramas representando as sequencias das fusões das parcelas, obtidos pelo emprego do método das médias da distância, com base no Coeficiente de assimetria Rogers e Tanimoto

TABELA 4: . Estresses, coeficiente cofenética, Rand e Rand ajustado gerado entre os coeficientes de similaridade

Coeficiente de assimetria	Coeficiente cofenético	Estresse %	Rand	Rand ajustado
Jaccard	0,55	34	0,91	0,76
Sorensen	0,57	25	0,98	0,95
concordância.simples	0,50	29	0,96	0,84
Russel e Rao	0,67	32	0,90	0,76
Rogers e Tanimoto	0,45	38	0,96	0,84

quanto aos dendrogramas obtidos. Os níveis de estresses apresentados para os cinco coeficientes (Tabela 4), para ambos as parcelas, foram de baixa magnitude. O nível de estresse variou 33% para Jaccard, 25% para o coeficiente de Sorrense-Dice, 29% para o coeficiente de Concordância simples, 32% para o coeficiente de Russel e Rao e o nível de estresse variou de 38% para o coeficiente de Rogers e Tanimoto. O valor máximo $(Rand = 1)$ corresponde a uma situação onde as duas classificações coincidem, não havendo pares que estejam numa mesmo grupo num caso, e em grupos diferentes no outro. Uma vez que neste caso é conhecida o agrupamento das 11 parcelas em 17 espécies, é possível comparar os agrupamentos obtidos através das análises agrupamentos com esta divisão por parcelas. No caso do agrupamento resultante do coeficiente Jaccard obtém-se um valor de $0,91$, Sorensen obtém-se um valor de $0,96$, concordância.simples obtém-se um valor de $0,84$, Russel e Rao obtém-se um valor de 0.76, Rogers e Tanimoto obtém-se um valor de 0,84 do índice de Rand ajustado, enquanto que o índice de Rand os valores foram semelhantes para todos os coeficientes de similaridade, observando que se pode utiliza-se de qualquer

coeficiente de similaridade ao realizar a comparação entre os coeficientes e pelo método de ligação média, dado que os valores de Rand são mais altos do que os valores de Rand ajustado. Considerando que realizou-se os resultado de forma independente para cada coeficiente.

4 CONCLUSÃO

A conclusão prática é que, na maioria das aplicações de dados, devem ser observadas se as propriedades dos coeficientes de similaridade e dissimilaridade são atendidas e a escolha dos coeficientes corretos, para as variáveis, pode provavelmente ser limitada aos cinco coeficientes seguintes: Jaccard, Sorensen, Russel Rao, Sokal Michener e Rogers Tanimoto.

REFERÊNCIAS

[1] A. L. Souza and D. R. Souza, "Análise multivariada para estratificação volumétrica de uma floresta ombrófila densa de terra firme, amazônia oriental," *Revista Árvore*, vol. 30, pp. 49–54, 2006.

[2] L. R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, pp. 297–302, 1945.

[3] P. Jaccard, "Etude comparative de la distribuition florale dans une portion des alpes et des jura," *Bulletin de la Societe Voudoise des Sciences Natureller, Payot*, vol. 37, pp. 547–579, 1901.

[4] T. Sorensen, "A 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*, vol. 5, pp. 1–34, 1948.

[5] P. F. Russel and T. R. Rao, "On habitat and association of species of anophelinae larvae in south-eastern madras," *Management Science. Journal Malaria Inst India*, vol. 3, pp. 153–178, 1940.

[6] D. Rogers and T. Tanimoto, "A computer program for classifying plants," *Science*, vol. 132, pp. 1115–1118, 1960.

[7] R. R. Sokal and C. D. Michener, "A statistical method for evaluating systematic relationships," *University of Kansas Science Bulletin*, vol. 38, pp. 1409–1438, 1958.

[8] J. Gower and P. Legendre, "Metric and euclidean properties of dissimilarity coefficients," *Jounal of Classification*, vol. 3, pp. 5–48, 1986.

[9] R. R. Sokal and P. H. A. Sneath, *Numeric taxonomy: the principles and practice of numerical classification*. San Francisco: W.H. Freeman, 1963.

[10] S. Kulczynski, "Die pflanzenassociationen der pienenen," *Bulletin International de LAcademie Polonaise des Sciences et des Letters, Classe des Sciences Mathematiques et Naturelles*, vol. 2(B), pp. 57–203, 1927.

[11] A. Ochiai, "Zoogeographic studies on the soleoid fishes found in japan and its neighbouring regions," *Bulletin of the Japanese Society of Scientific Fisheries*, vol. 22, pp. 526–530, 1957.

[12] P. H. A. Sneath and R. R. Sokal, *Numeric taxonomy: the principles and practice of numerical classification*. San Francisco: W.H. Freeman, 1973.

[13] A. R. Edwards, S. R. Mortimer, C. S. Lawson, D. B. Westbury, S. J. Harris, B. A. Woodcock, and V. K. Brown, "Hay strewing, brush harvesting of seed and soil disturbance as tools for the enhancement of botanical diversity in grasslands," *Biological Conservation*, vol. 134, no. 3, pp. 372–382, 2007.

[14] J. M. Duarte, J. B. dos Santos, and L. C. Melo, "Comparison of similarity coefficients based on rapd markers in the common bean," *Sociedade Brasileira de Genetica*, vol. 22(3), pp. 427–32, 1999.

[15] E. C. Beatriz Marti Emygdio, Irajá Ferreira Antunes and J. L. Nedel, "Eficiância de coeficientes de similaridade em genótipos de feijão mediante marcadores rapd," *Pesquisa Agropecuária Brasileira*, vol. 38(2), pp. 243–50, 2003.

[16] A. da Silva Meyer and et al, "Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (zea mays l)," *Genetics and Molecular Biology*, vol. 27, pp. 83–91, 2004.

[17] F. J. Rohlf, "Program numerical taxonomy and multivariate analysis system. version 1.70," 1992.

[18] C. D. Cruz, "Programa genes: aplicativo computacional em genética e estatística," 2001.

[19] J. M. Mulvey and H. P. Crowder, "Cluster analysis: an application of lagrangian relaxation," *Management Science*, vol. 25, pp. 329–340, 1979.

[20] A. H. Cheetham and J. E. Hazel, "Binary (presence-absence) similarity coefficients," *Journal of Paleontology*, pp. 1130–1136, 1969.

[21] B. H. McConnaughey and L. P. Laut, *The determination and analysis of plankton communities*. Lembaga Penelitian Laut, 1964.

[22] J. D. Holliday, C. Hu, and P. Willett, "Grouping of coefficients for the calculation of inter-molecular similarity and dissimilarity using 2d fragment bit-strings," *Combinatorial chemistry & high throughput screening*, vol. 5, no. 2, pp. 155–166, 2002.

[23] D. P. Vázquez and D. Simberloff, "Changes in interaction biodiversity induced by an introduced ungulate," *Ecology Letters*, vol. 6, no. 12, pp. 1077–1083, 2003.

[24] B. M. Emygdio, I. F. Antunes, E. Choer, and J. L. Nedel, "Eficiência de coeficientes de similaridade em genótipos de feijão mediante marcadores rapd," *Pesquisa Agropecuária Brasileira*, vol. 38, no. 2, pp. 243–250, 2003.

[25] D. H. Watson and et al, "Virus specific antigens in mammalian cells infected with herpes simplex virus," *Immunology*, vol. 11, pp. 399–408, 1966.

[26] M. A. de Albuquerque, "Estabilidade em análise de agrupamento: estudo de caso em ciência florestal," *Revista Árvore- Vicosa-MG*, vol. 2(30), pp. 257–265, 2006.

[27] W. O. Bussab and P. A. Morettin, *Estatística básica*. Saraiva, 2010.

[28] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis," *Psychometrika*, vol. 29, pp. 1–27, 1964.