THAYNARA APARECIDA DE SOUZA NETO

ESTUDO DE DIVERGÊNCIA GENÉTICA EM ACESSOS DE *Capsicum annuum* L. UTILIZANDO MÉTODOS DE AGRUPAMENTO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística Aplicada e Biometria, para obtenção do título de *Magister Scientiae*.

Orientador: Paulo Roberto Cecon

Coorientador: Sebastião Martins Filho

VIÇOSA - MINAS GERAIS
2022

Ficha catalográfica elaborada pela Biblioteca Central da Universidade Federal de Viçosa - Campus Viçosa

T

S729e
2022

Souza Neto, Thaynara Aparecida de, 1996-
Estudo de divergência genética em acessos de *Capsicum annuum* L., utilizando métodos de agrupamento / Thaynara Aparecida de Souza Neto. – Viçosa, MG, 2022.
1 dissertação eletrônica (94 f.): il. (algumas color.).

Inclui apêndice.
Orientador: Paulo Roberto Cecon.
Dissertação (mestrado) - Universidade Federal de Viçosa, Departamento de Estatística, 2022.
Inclui bibliografia.
DOI: https://doi.org/10.47328/ufvbbt.2022.179
Modo de acesso: World Wide Web.

1. Análise por agrupamento. 2. *Capsicum annuum* - Melhoramento genético - Métodos estatísticos. 3. Diversidade genética. I. Cecon, Paulo Roberto, 1955-. II. Universidade Federal de Viçosa. Departamento de Estatística. Programa de Pós-Graduação em Estatística Aplicada e Biometria. III. Título.

CDD 22. ed. 519.535

Bibliotecário(a) responsável: Alice Regina Pinto CRB6 2523

THAYNARA APARECIDA DE SOUZA NETO

ESTUDO DE DIVERGÊNCIA GENÉTICA EM ACESSOS DE *Capsicum annuum* L. UTILIZANDO MÉTODOS DE AGRUPAMENTO

Dissertação apresentada à Universidade Federal de Viçosa, como parte das exigências do Programa de Pós-Graduação em Estatística aplicada e Biometria, para a obtenção do título de *Magister Scientiae*.

APROVADA: 21 de fevereiro de 2022.

Assentimento:

Thaynara Aparecida de Souza Neto
Thaynara Aparecida de Souza Neto
Autor

Paulo Roberto Cecon
Orientador

AGRADECIMENTOS

Agradeço primeiramente a Deus e Maria Santíssima, por me fortalecerem mediante às situações de dificuldade e provações, por me guiarem e fortalecerem na caminhada, tanto nos momentos bons quantos nos momentos ruins.

Aos meus pais, Vanir e Conceição, às minhas irmãs, Thalissa, Thawanne e Manuela e minha vó, Maria por acreditarem em mim a todo momento, por me darem todo apoio, força e amor sempre. Vocês trazem alegria para minha vida e serei imensamente grata por todo sempre.

A todos os meus amigos, especialmente Wadlam, Vanusa, Talita, Alice e Iago por estarem sempre à disposição e por me alegrarem, mesmo que à distância. Às minhas parceiras de apartamento, Cíntia e Fernanda, por todos os momentos compartilhados, inclusive as dificuldades nesta etapa. Vocês foram e são parte essencial da minha caminhada.

Ao EJC (Encontro de Jovens com Cristo) e ao JSC (Jovens Seguidores de Cristo), por fortalecerem a minha fé, me trazer amizades muito especiais e por todos os momentos de descontração e oração. Isso, certamente, me ajudou muito.

Ao DMAFE (Departamento de Matemática, Física e Estatística), meu departamento no IF, por todo ensinamento e momentos compartilhados. De maneira especial, à professora Cristina, por sempre me apoiar, ensinar e incentivar.

Ao meu orientador, professor Paulo Roberto Cecon por todos os valiosos ensinamentos compartilhados, apoio, disponibilidade e por sempre me incentivar em todos os momentos. Ao meu coorientador, professor Sebastião Martins Filho, pelo apoio e excelentes sugestões para o desenvolvimento deste trabalho.

Aos membros da banca examinadora pela disponibilidade e enriquecedoras sugestões para este trabalho.

Aos colegas de turma por compartilharem momentos, tanto das conquistas quanto das dificuldades. Acredito que esse apoio mútuo foi muito importante para cada um de nós.

Ao GESTBIO (Grupo de Estudos em Estatística aplicada e Biometria), pela oportunidade de ser integrante do grupo, por terem o empenho, a disponibilidade e paciência em me inserir no programa e por compartilhar vários momentos durante as nossas reuniões.

Aos professores e funcionários do Programa de Pós-Graduação em Estatística aplicada e Biometria (PPESTBIO) por todo aprendizado, disponibilidade e incentivo durante este tempo.

À Universidade Federal de Viçosa pelo acolhimento e oportunidade.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), pela concessão da bolsa de estudo.

Foram tempos difíceis, enfrentamos grandes perdas e passamos por grandes dificuldades. Um tempo atípico, do qual muitos tiveram vontade de desistir, mas o fortalecimento da nossa fé, nos fez ter a esperança de tempos melhores e não nos tirou a alegria de viver. Logo, a todos que, direta ou indiretamente, contribuíram para a minha caminhada e consequente, titulação o meu muito obrigada!

“Nada é pequeno se feito com amor.”

- Santa Teresinha do Menino Jesus

RESUMO

SOUZA NETO, Thaynara Aparecida de, M.Sc., Universidade Federal de Viçosa, fevereiro de 2022. **Estudo de divergência genética em acessos de *Capsicum annuum* L., utilizando métodos de agrupamento.** Orientador: Paulo Roberto Cecon. Coorientador: Sebastião Martins Filho.

A *Capsicum annuum* L. é a espécie mais cultivada e economicamente importante do gênero e, por tal motivo, estudos em divergência genética são importantes, pois permitem conhecer o grau de seleção da variabilidade genética das populações vegetais e identificar combinações híbridas. Sendo assim, o objetivo foi avaliar a diversidade genética de genótipos de *C. annuum* L., por meio de técnicas multivariadas de agrupamentos, utilizando os métodos hierárquicos, Vizinho mais próximo e Ward e não-hierárquicos, k-médias e Tocher. Os dados foram provenientes de um experimento conduzido na área experimental na casa de vegetação do Departamento de Agronomia da Universidade Federal de Viçosa, sob delineamento inteiramente casualizado, com quatro repetições e uma planta como unidade experimental. Vinte e nove genótipos de *C. annuum* registrados no Banco de Germoplasma de Hortaliças da Universidade Federal de Viçosa (BGH/UFV) foram avaliados com base em 6 características, dentre elas, Comprimento e largura do fruto maduro, Peso total dos frutos por planta, Espessura da polpa, Teor de vitamina C e Número de sementes por fruto. Os nove genótipos foram divididos em 2, 5 e 2 grupos pelos métodos hierárquicos, de Tocher e $k-$médias, respectivamente. Para os cruzamentos, o método do Vizinho mais próximo reuniu 95% dos genótipos no grupo II e para o método de Ward, os grupos possuíram uma porcentagem bem próxima, no qual o grupo IV reuniu o maior número de genótipos (45%). Os métodos de Tocher e $k-$médias foram divididos em 5 e 3 grupos, respectivamente. Para todos os dados, o método do Vizinho mais próximo reuniu $86{,}20\%$ dos genótipos no grupo V e o de Ward, $44{,}83\%$ no grupo V, contendo este o maior número de genótipos. No método de Tocher, o grupo I reuniu $68{,}96\%$ dos genótipos e no de $k-$médias, $44{,}83\%$ no grupo III. Todos os métodos hierárquicos foram validados pelo coeficiente de correlação cofenética, no qual o de maior coeficiente ocorreu no método do Vizinho mais próximo para todos os dados, indicando uma melhor adequação deste método. Além disso, os genótipos 5 e 27 mostraram-se de extrema importância para o estudo da divergência genética. Assim, independentemente do procedimento, foi possível identificar os genótipos mais dissimilares, podendo contribuir para outras pesquisas.

Palavras-chave: Critério de agrupamento. Diversidade genética. Melhoramento de *Capsicum*. Método das k −médias. Método de Tocher.

ABSTRACT

SOUZA NETO, Thaynara Aparecida de, M.Sc., Universidade Federal de Viçosa, February 2022. **Study of genetic divergence in accessions of *Capsicum annuum* L., using clustering methods.** Advisor: Paulo Roberto Cecon. Co-Advisor: Sebastião Martins Filho.

Capsicum annuum L. is the most cultivated and economically important species of the genus and, for this reason, studies in genetic divergence are important because they allow to know the degree of selection of genetic variability of plant populations and to identify hybrid combinations. Therefore, the objective was to evaluate the genetic diversity of C. annuum L. genotypes, through multivariate clustering techniques, using the hierarchical methods, Closest Neighbor and Ward, and non-hierarchical, k-means and Tocher. The data came from an experiment conducted in the experimental area in the greenhouse of the Agronomy Department of the Federal University of Viçosa, under entirely randomized design, with four repetitions and one plant as an experimental unit. Twenty-nine C. annuum genotypes registered in the Germplasm Bank of the Federal University of Viçosa (BGH/UFV) were evaluated based on 6 characteristics, including ripe fruit length and width, total fruit weight per plant, pulp thickness, vitamin C content and number of seeds per fruit. The nine genotypes were divided into 2, 5 and 2 groups by hierarchical, Tocher's and k-means methods, respectively. For the crossings, the closest neighbor method gathered 95% of the genotypes in group II and for Ward's method, the groups had a very close percentage, in which group IV gathered the largest number of genotypes (45%). The Tocher and k-means methods were divided into 5 and 3 groups, respectively. For all data, the nearest neighbor method gathered 86.20% of the genotypes in group V and Ward's method, 44.83% in group V, which contained the largest number of genotypes. In genotypes the Tocher method, group I contained 68.96% of the genotypes and in the k-means method, 44.83% in group III. All hierarchical methods were validated by the cohenetic correlation coefficient, in which the highest coefficient occurred in the Nearest Neighbor method for all data, indicating a better adequacy of this method. Furthermore, genotypes 5 and 27 proved to be extremely important for the study of genetic divergence. Thus, regardless of the procedure, it was possible to identify the most dissimilar genotypes and may contribute to other research.

Keywords: Clustering criteria. Genetic diversity. *Capsicum* improvement. k −means method, Tocher's method.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS E ABREVIATURAS

BGH Banco de Germoplasma de Hortaliças

CAPES Coordenação de Aperfeiçoamento de Pessoal de Nível Superior

CCC Coeficiente de correlação cofenética

COM Comprimento do fruto maduro

DMAFE Departamento de Matemática, Física e Estatística

EJC Encontro de Jovens com Cristo

ESP Espessura da polpa

GESTBIO Grupos de estudos em Estatística aplicada e Biometria

JSC Jovens Seguidores de Cristo

LAR Largura do fruto maduro

NS Número de sementes por frutos

PPESTBIO Programa de Pós-Graduação em Estatística aplicada e Biometria

PT Peso total de frutos por planta

RMSSTD *Root Mean Square Standard Deviation*

UFV Universidade Federal de Viçosa

VIT Teor de vitamina C

SUMÁRIO

1 INTRODUÇÃO

O gênero *Capsicum* possui uma grande diversidade de pimentas e pimentões e o Brasil é um grande centro de diversidade genética (MIRANDA, 2014). De acordo com a Revista Campo e Negócios, em 2018, o Brasil apresentou uma área de 3,8 milhões de hectares destinados para cultivo de *Capsicum*, com produção mundial de 40,9 milhões de toneladas. Algumas coleções desse gênero estão conservadas em forma de sementes nos bancos de germoplasma a exemplo da Embrapa Hortaliças, com quase dois mil acessos, entre cultivares de polinização aberta, híbridos, populações de materiais cultivados, linhagens e materiais silvestres, provenientes de vários países e regiões brasileiras (Carvalho e Bianchetti, 2008).

Estudos sobre a diversidade genética do gênero *Capsicum* fornecem parâmetros para identificação de genitores que, quando cruzados, possibilitam o aparecimento de cultivares superiores facilitando o conhecimento da base genética da população que podem ser utilizados em programas de melhoramento genético (CARGNIN; SOUZA, 2007). Segundo Falconer (1981), a variabilidade genética de uma população segregante depende da diversidade genética entre os pais envolvidos nos cruzamentos.

Assim, estudos a respeito da diversidade genética apresentam grande relevância no melhoramento de plantas, por fornecerem parâmetros para identificação de genitores que, quando cruzados, possibilitam o aparecimento de cultivares superiores, além de facilitarem o conhecimento da base genética da população. (CARGNIN; SOUZA, 2007). Segundo Falconer (1981), a variabilidade genética de uma população segregante depende da diversidade genética entre os pais envolvidos nos cruzamentos.

Dada a sua importância, vários estudos com ênfase na predição da divergência genética são realizados. Dentre eles estão o uso de técnicas de análise multivariada, como a análise de agrupamento que, conforme Cruz *et al.* (2020), implica em avaliar a capacidade de alocação ou de discriminação de indivíduos, nos seus respectivos centros de referências, com base nas variáveis avaliadas, bem como formular e testar hipóteses sobre as causas dessa aglomeração ou dispersão. Segundo Machado (2011), a tarefa de agrupamento visa construir grupos conforme a similaridade dos elementos. O processo de agrupamento possui cinco etapas principais: preparação de padrões, a escolha de uma medida de similaridade, o agrupamento dos dados, a validação dos grupos formados e a interpretação dos mesmos.

Esses métodos são divididos em métodos hierárquicos e não-hierárquicos, no qual cada um deles podem levar a diferentes padrões de agrupamentos. Portanto, o pesquisador é quem define o mais adequado ao seu trabalho, de acordo com os seus objetivos.

Logo, este trabalho tem o objetivo de avaliar a diversidade genética dos genótipos de *Capsicum annuum L.*, comparando a qualidade dos agrupamentos obtidos por meio dos métodos hierárquicos (Vizinho mais próximo e Ward) e os métodos não-hierárquicos (k −médias e Tocher).

2 REVISÃO DE LITERATURA

2.1 Aspectos gerais da *Capsicum*

As pimentas e os pimentões do gênero *Capsicum* recebem a seguinte classificação botânica: Família: Solanaceae; Tribo: Solaneae; Subtribo: Solaninae; Divisão: Spermatophyta; Filo: Angiospermae; Classe: Dicotiledônea; Ramo: Malvales - Tubiflorae; Ordem: Solanales (Personatae). No geral, todas as espécies são originárias do hemisfério ocidental e nativas das regiões tropicais das Américas. Essa hortaliça está difundida em todas as regiões do Brasil, sendo que as principais áreas de cultivo estão localizadas nas regiões Sudeste e Centro-Oeste (FONSECA, 2016; WAGNER, 2003).

Segundo Ribeiro e Reifschneider (2008), as principais espécies cultivadas desse gênero são: *Capsicum annuum, C. baccatum, C. chinense*, *C frutenscens* e *C. pubescens* e as suas espécies, conforme Carvalho (2003), são classificadas de acordo com o nível de domesticação (domesticadas, semi-domesticadas e silvestres). As domesticadas são plantas nas quais o homem selecionou determinadas alterações genéticas, de tal modo que não são capazes de sobreviver em condições naturais. Já as semi-domesticadas são plantas selecionadas, cultivadas, mas ainda não completamente domesticadas. E por último, as silvestres são plantas que são encontradas e exploradas pelo homem, porém não são cultivadas e nem ocorrem em ambientes antrópicos (CARVALHO; BIANCHETTI, 2008). Carvalho *et al.* (2006), ressalta também que as espécies domesticadas de *Capsicum* em geral são autógamas, ou seja, são autopolinizadas (o pólen de uma flor fecunda o estigma da mesma flor).

Outra característica, a principal, que destaca-se exclusivamente nesse gênero é a pungência, conferida por alcaloides denominados capsaicinóides, análogos à capsaicina. Essa palavra é usada para designar sabor picante, quente, ardente ou condimentado (LUTZ; FREITAS, 2008).

Dentre as espécies citadas, a *C. annuum* é a mais cultivada e economicamente a mais importante, podendo apresentar frutos doces e pungentes (WANG; BOSLAND, 2006 *apud* PESSOA *et al.*, 2017). Ribeiro e Reifschneider (2008) enfatizam que as variedades mais comuns apresentam-se como cultivares de polinização aberta, híbridos de pimentões doces, pimentas doces para pápricas, pimenta–americana e as pimentas picantes jalapeño e cayenne e, ainda, poucas cultivares ornamentais.

Wagner (2003) ressalta também que os frutos da *Capsicum* são consumidos na forma *in natura* ou processados como condimentos, conservas, corantes, na composição de remédios e indústria bélica. Além disso, o agronegócio de pimenta e pimentão tornou-se um dos mais relevantes segmentos de produção no país, ocorrendo na maioria das regiões agrícolas nacionais, sendo um dos melhores exemplos de agricultura familiar e integração dos pequenos produtores com a agroindústria e com a produção nacional (RIBEIRO; FREITAS; CARVALHO, 2006 *apud* MENDES, 2018).

Logo, a grande diversidade observada concede à espécie grande potencial para o melhoramento genético, no qual o Brasil possui uma grande variabilidade genética entre espécies domesticadas, semi-domesticadas e silvestres sendo utilizada em programas de melhoramento genético, desenvolvidos por instituições públicas e privadas (MIRANDA, 2014; WAGNER, 2003).

2.2 Divergência genética

A divergência ou diversidade genética entre e dentro de populações encontradas em suas condições naturais, em bancos de germoplasma ou desenvolvidas nos programas de melhoramento genético pode ser predita pelas diferenças entre os valores fenotípicos mensurados em suas unidades (indivíduos, famílias etc.) (DINIZ FILHO, 2000 *apud* CRUZ *et al*., 2020). Assim, Carvalho (2003) frisa que a caracterização de germoplasma é importante, pois auxilia no conhecimento e no uso da variabilidade genética, permitindo aos melhoristas selecionar acessos para obtenção de populações e linhagens que atendam às necessidades específicas de um programa de melhoramento.

Segundo Amaral Jr e Thiébaut (1999) citado por Bento (2007), o estudo de divergência permite conhecer o grau de seleção da variabilidade genética das populações vegetais, e conforme CRUZ *et al*. (2012) identificar as combinações híbridas de maior efeito heterótico e maior heterozigose, de tal forma que, em suas gerações segregantes, haja maior possibilidade de recuperação de genótipos superiores.

Portanto, a divergência genética tem sido avaliada por meio de técnicas biométricas, baseadas na quantificação da heterose, ou por processos preditivos. Entre os métodos fundamentados em modelos biométricos, que se destinam à avaliação da divergência dos progenitores, citam-se as análises dialélicas, que avaliam tanto a capacidade específica quanto

a heterose manifestada nos híbridos. Nos dialelos, é necessária a avaliação de p progenitores e de todas as suas combinações híbridas $\left(\frac{p(p-1)}{2}\right)$. Assim, quando o valor de p é elevado, a obtenção do material experimental pode ser impraticável, e o estudo inviabilizado. Dessa forma, por dispensarem a obtenção prévia das combinações híbridas, os métodos preditivos têm merecido considerável ênfase (CRUZ *et al*., 2012).

Logo, entre os métodos preditivos da heterose, podemos citar aqueles que tomam por base as diferenças morfológicas, fisiológicas, ou moleculares, quantificando-as em alguma medida de dissimilaridade que expressa o grau de divergência genética entre os genitores. Para essa predição, vários métodos multivariados podem ser aplicados, dentre os quais a análise por componentes principais e por variáveis canônicas e os métodos aglomerativos (Cruz *et al.,* 2014; Cruz *et al.,* 2012).

Portanto, o estudo de divergência genética em espécies de *Capsicum spp.* tem resultado em vários trabalhos, como AQUINO, 2016; NASCIMENTO, 2018; SILVA *et al*., 2020; COSTA *et al*., 2021; JUNIOR *et al.*, 2021, entre tantos outros. Cruz *et al.* (2012) frisa ainda que, para escolher o método mais adequado é necessário que a precisão desejada pelo pesquisador seja determinada, além da facilidade da análise e pela forma como os dados foram obtidos.

2.3 Medidas de dissimilaridade

Ferreira (2008) relata que, em geral, deseja-se agrupar n objetos (indivíduos, itens, cidades, genótipos, etc.) em um número de grupos desconhecidos k ou agrupar as p variáveis, ao invés de objetos. Se houver apenas $p = 2$ variáveis, pode-se agrupar os objetos em um gráfico bidimensional. Porém, se $p > 2$, as análises vão além de agrupamentos gráficos e visam identificar padrões de agregação de acordo com suas similaridades.

Os dados observados em n indivíduos e p variáveis são representados da seguinte forma:

$$Y_{nxp} = \begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1p} \\ y_{21} & y_{22} & \cdots & y_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & y_{np} \end{bmatrix}$$

em que cada unidade amostral representada por uma linha da matriz de dados Y, sendo um vetor com p elementos (variáveis), e cada variável é representada por uma coluna de Y, sendo um vetor com n elementos, as observações.

Porém, conforme Ferreira (2011) *apud* Souza (2017) relata, a obtenção da matriz de dados a partir de uma amostra aleatória, como expressa na forma acima pode não ser muito informativa, principalmente se o tamanho amostral n for grande e houver um número excessivo de variáveis p. Desse modo, utiliza-se a matriz de proximidades por quantificarem e informarem sobre o grau de semelhança ou de diferença apresentado entre quaisquer genótipos (CRUZ et al., 2014).

A maior parte dos métodos de agrupamento requer que a matriz de proximidades entre objetos seja previamente obtida. A proximidade é o termo utilizado para indicar ou similaridade ou dissimilaridade, que é medida pelas distâncias. Na primeira quanto maior o valor observado, mais parecidos são os objetos. Já para a segunda quanto maior for o valor observado, menos parecidos (mais dissimilares ou divergentes) serão os objetos. (FERREIRA, 2008; REGAZZI; CRUZ, 2020).

Segundo Cruz *et al*. (2020), essas medidas são de grande importância em estudos de diversidade genética em que se procura identificar genitores a serem utilizados em programas de hibridação. Salienta-se também de que, se houver genitores de bom desempenho e com certo grau de diversidade, podem apresentar constituição genética complementar que proporcionariam, na F_1, maior heterose e, nas gerações segregantes, indivíduos transgressivos.

Portanto, Luz (2014) traz que, dentre as medidas de dissimilaridade comumente utilizada para características quantitativas nos estudos genéticos, cita-se a distância euclidiana, o quadrado da distância euclidiana, a distância euclidiana média e a distância generalizada de Mahalanobis.

Desse modo, Cruz *et al.* (2020) ressaltam que, para que uma medida de distância seja considerada métrica, é necessário satisfazer as propriedades abaixo. Considerando os genótipos i e j e um coeficiente d_{ij} capaz de medir a distância entre eles, tem-se:

$i)\ d_{ij} \geq 0$;

$ii)\ d_{ij} = d_{ji}$(simetria);

$iii)\ d_{ij} = 0, se\ e\ somente\ se\ i = j$;

$iv)\ d_{ij} \leq d_{il} + d_{jl}$ (desigualdade triangular).

O termo dissimilaridade apareceu em função de que, à medida que d_{ij} cresce, diz-se que a divergência entre i e j aumenta, ou seja, tornam-se cada vez mais dissimilares.

2.4 Análise de agrupamento

A análise de agrupamentos, também conhecida como análise de *cluster* ou de conglomerados é um procedimento de estatística multivariada que engloba técnicas que objetivam organizar objetos em grupos de acordo com a proximidade existente entre eles. Os objetos de um mesmo grupo são tão similares quanto possível (coesão interna) e ao mesmo tempo tão dissimilares quanto possível dos objetos dos demais grupos (isolamento externo). Mais especificamente, as técnicas de análise de agrupamento objetivam dividir um conjunto de observação em um número de grupos homogêneos, segundo algum critério conveniente de homogeneidade. (MATOS, 2007; REGAZZI; CRUZ, 2020).

Os métodos de agrupamento são divididos, de uma maneira geral, em métodos hierárquicos e não-hierárquicos. Nos hierárquicos, os indivíduos são reunidos em grupos e o processo repete-se em diferentes níveis até formar uma árvore. Nos não-hierárquicos, temos de definir o número k de grupos inicialmente e, então, alocar os n objetos aos k grupos de maneira otimizada (FERREIRA, 2008; REGAZZI; CRUZ, 2020).

Aaker *et al.* (2001) citado por Albuquerque (2005) expõe que, a análise de agrupamento compreende cinco etapas, sendo elas:

1. A seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;

2. A definição de um conjunto de variáveis a partir das quais serão obtidas informações necessárias ao agrupamento dos indivíduos;

3. A definição de uma medida de semelhança ou distância entre os indivíduos;

4. A escolha de um algoritmo de partição/classificação;

5. Por último, a validação dos resultados encontrados.

Matos (2007) relata que o principal fator que corrobora positivamente para o interesse de muitos pesquisadores no tema é a vasta aplicação desses procedimentos. Desse modo,

ressalta-se que são diversas as áreas de conhecimento que a utiliza para compreender e explicar fenômenos.

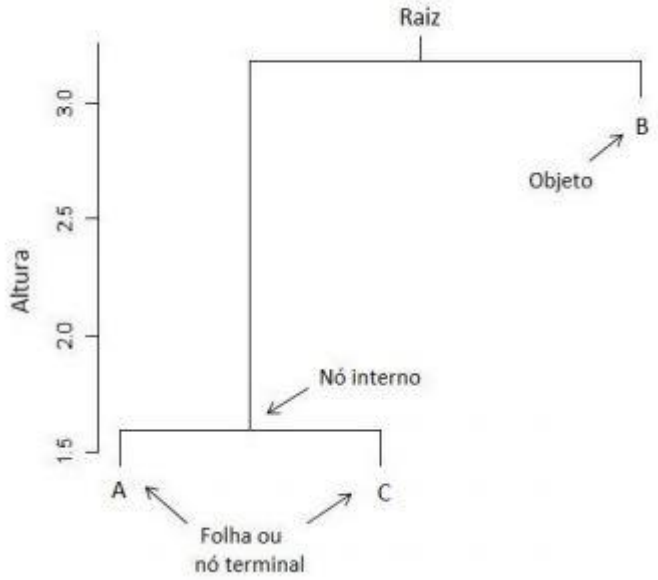
Portanto, a análise de agrupamento é uma técnica puramente exploratória, que visa a geração de hipóteses sobre o padrão de aglomeração estabelecido e pode ser suplementada ou complementada por outras técnicas de visualização (DIAS, 1998). Além disso, não é necessária qualquer hipótese acerca da distribuição de probabilidade dos dados. (CRUZ *et al.*, 2020)

2.4.1 Métodos Hierárquicos

Nos métodos hierárquicos, os genótipos são agrupados por um processo que se repete em vários níveis, até que seja estabelecido o dendrograma ou diagrama de árvore. Esse gráfico ilustra as fusões ou divisões realizadas a cada passo do processo. Ainda segundo os autores, o arranjo de nós e caules representam a topologia da árvore. O diagrama descreve o processo pelo qual foi obtida a hierarquia, assim há várias sub-árvores oriundas da raiz da árvore. O nó interno representa partições particulares, ou seja, os agrupamentos formados a partir dos nós terminais, que representam os objetos. A altura do nó interno corresponde ao ponto em que os objetos ou grupos foram unidos, ou seja, a proximidade entre eles. Dessa forma, a ordem de união dos grupos segue o princípio de ordem crescente da altura do nó (CRUZ *et al.*, 2014; EVERITT *et al.*, 2011 *apud* SOUZA, 2017).

A Figura 1, a seguir, representa essas terminologias descritivas dos dendrogramas.

Figura 1 - Terminologias descritivas dos dendrogramas



Fonte: De SOUZA, 2017, a partir de EVERITT; HOTHORN, 2011

Hair Júnior *et al.* (2009) ressalta que essas técnicas apresentam alguns pontos fortes, como alta difusão do método, simplicidade do processo, rapidez, habilidade para examinar uma gama de soluções e facilidade de comparações entre elas, com variação de medidas de dissimilaridade.

Bussab *et al.* (1990) enfatiza que esses métodos podem ainda ser subdivididos em dois tipos: aglomerativos, onde por meio de fusões sucessivas dos n objetos, vão sendo obtidos $n-1, n-2$, etc. grupos, até reunir todos os objetos num único grupo; divisivos, que partem de um único grupo, e por divisão sucessivas vão sendo obtidos $2,3,$ etc. grupos. O que caracteriza estes processos é que a reunião de dois agrupamentos numa certa etapa produz um dos agrupamentos da etapa superior, caracterizando o processo hierárquico.

O procedimento básico de todos os métodos hierárquicos aglomerativos de agrupamento é similar. Eles iniciam com o cálculo de uma matriz de proximidade entre as entidades e finalizam com um dendrograma, mostrando as fusões sucessivas dos indivíduos, o que culmina no estágio onde todos os indivíduos estão em um único grupo (REGAZZI; CRUZ, 2020).

Nascimento (2011) ressalta que dentre os métodos aglomerativos, citam-se o do vizinho mais próximo (*Single Linkage Method*); o do vizinho mais distante (*Complete*

Linkage Method); o da ligação média (*Average Linkage*), ponderado ou não; e o proposto por Ward (1963). Dentre os divisivos, o mais conhecido é o de Edwards e Cavalli-Sforza (1965).

O método do vizinho mais próximo define a semelhança entre agrupamentos como a menor distância de qualquer objeto de um agrupamento a qualquer objeto no outro, enquanto que o do vizinho mais distante baseia-se na distância máxima entre observações em cada agrupamento. Já o da ligação média considera-se que a similaridade de quaisquer dois agrupamentos é a similaridade média de todos os indivíduos em um agrupamento com todos os indivíduos em outro. Por outro lado, o método de Ward difere das técnicas anteriores no sentido de que a similaridade entre dois agrupamentos não é uma única medida de similaridade, mas a soma dos quadrados dentro dos agrupamentos feita sobre todas as variáveis (HAIR JÚNIOR et al., 2009).

Segundo Rosemburg (1984) *apud* Silva (2012), o método de Ward tem sido preferido, em alguns casos, devido ao efeito gráfico gerado pelo dendrograma, possibilitando a visualização de grupos bem definidos. Já Cruz *et al.* (2014) dá enfoque na utilização do método da ligação média, com frequência, em ecologia e sistemática e em taxionomia numérica.

Outro ponto importante refere-se à determinação do número de grupos para os métodos hierárquicos aglomerativos. Conforme Cruz *et al.* (2020) ressalta, pode ser definido pelo conhecimento que se tenha sobre os dados, pela conveniência do pesquisador ou por simplicidade, além da inspeção visual das ramificações do dendrograma ou recorrendo a algum procedimento estatístico, como o Método de Mojena (1977), do RMSSTD, do desvio padrão-médio, do coeficiente de determinação (R^2) ou do R^2 semiparcial.

Para avaliar a consistência do agrupamento, que é realizada após a obtenção do dendrograma, a literatura fornece algumas alternativas. a) validações externas – avaliam os agrupamentos com base em agrupamentos de diferença, b) validações internas – analisam as informações contidas nos grupos obtidos e c) validações relativas – compara o agrupamento com outros agrupamentos. Dentre as técnicas de validação interna citam-se: coesão, acoplamento, coeficiente de correlação cofenética e coeficiente de silhueta (OLIVEIRA, 2016). O coeficiente de correlação cofenética, segundo Sokal e Rohlf (1962), é um coeficiente de correlação momento-produto entre os elementos acima da diagonal da matriz de dissimilaridade e os valores da matriz cofenética. Quanto maior o seu valor, menor será a distorção provocada pelo agrupamento.

Logo, vale salientar que, não existe o que se possa chamar de melhor critério na análise de agrupamentos, mas alguns são mais indicados para determinadas situações do que outros (KAUFMANN; ROSSEEUW, 1990 *apud* ALBUQUERQUE, 2005). Portanto, é prática comum utilizar vários critérios e fazer a comparação dos resultados, se tais resultados forem semelhantes, é possível concluir que eles possuem um elevado grau de estabilidade e, portanto, são confiáveis (ALBUQUERQUE, 2005).

2.4.2 Métodos Não-hierárquicos

Nos métodos não-hierárquicos, é definido um número k de grupos inicialmente e, então, aloca-se os n objetos aos k grupos de maneira otimizada. Nestes casos, utiliza-se uma alocação arbitrária para iniciar o processo e iterativamente buscar a alocação ótima (FERREIRA, 2008).

Um fato interessante, ressaltado por Mingoti (2005), é que estes métodos não seguem a propriedade da hierarquia, ou seja, mesmo se dois objetos forem unidos em algum passo do processo, pode ser que eles não permaneçam no mesmo grupo na partição final. E, portanto, isso implica que não é possível construir dendrogramas para a representação dos agrupamentos formados passo a passo.

Ferreira (2008) enfatiza que existem muitos métodos não-hierárquicos baseados em misturas de distribuição, estimação de densidades e partição. Este último, também denominado de otimização – por ter como objetivo alcançar uma partição de indivíduos que otimize (maximize ou minimize) alguma medida predefinida (Cruz *et al.*, 2014) -, é o mais utilizado. Dentre eles, estão o método das k-médias (o mais popular), método de Tocher e Tocher Modificado.

O método das k-médias busca minimizar a distância dos elementos de um conjunto de dados com k centros de forma iterativa. No método de Tocher, adota-se o critério de que a média das medidas de dissimilaridade dentro de cada grupo deve ser menor que as distâncias médias entre quaisquer grupos. Por outro lado, no de Tocher modificado o processo de agrupamento é sequencial e não-simultâneo, não existindo influência dos indivíduos já agrupados, o que mostra-se mais eficaz (PALMA, 2018; CRUZ *et al.*, 2012; VASCONCELOS *et al.*, 2007 *apud* OLIVEIRA, 2016).

De acordo com o *site* Oper Data, o procedimento geral adotado para eles são:

- Escolher uma partição inicial (baseada em conhecimentos anteriores do problema);
- Realizar o deslocamento do objeto de seu grupo para outros grupos;
- Verificar o valor do critério utilizado, decidindo pela clusterização que apresentar melhoria.

Desse modo, as principais vantagens desses métodos são, que um item pode mudar de agrupamento com a evolução do algoritmo e, há possibilidade de se operar um maior conjunto de dados, pois os métodos particionais são rápidos de serem processados. A principal desvantagem destes métodos é que eles não trabalham bem em grupos de formas complexas e tamanhos diferentes (MAXIMILIANO; CORDEIRO, 2008; PEREIRA, 2007 *apud* MACHADO, 2011).

Portanto, Albuquerque (2005) enfatiza que os métodos não-hierárquicos diferem entre si pela maneira que constituem a melhor partição. Como qualquer classificação, existirão tipos que serão difíceis de classificar, ou que poderão caber em mais de um grupo.

3 MATERIAIS E MÉTODOS

3.1 Descrição do experimento

Os dados abaixo são provenientes de um experimento conduzido na casa de vegetação do Departamento de Agronomia da Universidade Federal de Viçosa (UFV), Viçosa, Minas Gerais, nas coordenadas geográficas: 20º 45' de latitude sul e 42º 51' de longitude oeste, com altitude média de 650 m.

O delineamento utilizado foi inteiramente casualizado, com quatro repetições e uma planta como unidade experimental, sendo 29 genótipos de *Capsicum annuum* L. registrados no Banco de Germoplasma de Hortaliças (BGH/UFV). Os 29 genótipos (tratamentos) foram constituídos por nove genótipos, divididos em dois grupos, e vinte combinações híbridas. O grupo 1 (G1) refere-se a cinco genótipos comerciais de *C. annuum* L. com pungência (Pimenta Vulcão, Pimenta Cayene, Pimenta Peter, Pimenta Picante para vaso e Pimenta Jamaica Yellow). O grupo 2 (G2) refere-se a quatro genótipos comerciais de *C. annuum* sem pungência (Pimenta Doce Italiana, Pimentão Quadrado, Pimentão Cascadura Ikeda e Pimentão Rubi Gigante).

As combinações são apresentadas na Tabela 1 a seguir.

Tabela 1 - Tabela com os nove genótipos e suas combinações híbridas resultantes.

Grupo *I*	Grupo *II*			
	Pimenta Doce Italiana (6)	Pimentão Quadrado (7)	Pimentão Cascadura Ikeda (8)	Pimentão Rubi Gigante (9)
Pimenta Vulcão (1)	$1x6$	$1x7$	$1x8$	$1x9$
Pimenta Cayene (2)	$2x6$	$2x7$	$2x8$	$2x9$
Pimenta Peter (3)	$3x6$	$3x7$	$3x8$	$3x9$
Pimenta picante para vaso (4)	$4x6$	$4x7$	$4x8$	$4x9$
Pimenta Jamaica Yellow (5)	$5x6$	$5x7$	$5x8$	$5x9$

Fonte: Oliveira, 2020.

As características avaliadas foram:

- Comprimento do fruto maduro (COM), expresso em milímetros;

- Largura do fruto maduro (LAR), expresso em milímetros, com médias de cinco frutos;
- Peso total de frutos por planta (PT), expresso em gramas;
- Espessura da polpa (ESP), porção mediana de cinco frutos por planta em milímetros;
- Teor de vitamina C (VIT);
- Número de sementes por frutos (NS), obtido da contagem das sementes de cinco frutos por planta.

Os dados obtidos com o experimento foram padronizados, justamente pelo fato de as variáveis terem sido quantificadas em diferentes medidas e consequentemente, isso evita que suas unidades afetem arbitrariamente a similaridade entre os indivíduos e faz com que as variáveis contribuam igualmente em sua avaliação.

3.2 Análise de agrupamento

Os dados foram submetidos às análises de medida de dissimilaridade, utilizando a Distância Generalizada de Mahalanobis; de agrupamento, dentre os quais, o método hierárquico Vizinho mais próximo e Ward e não-hierárquico, $k-$médias e Tocher. Para determinação do número de grupos, foram utilizados o Critério de Mojena e RMSSTD e para validação de agrupamento, o Coeficiente de correlação cofenética. Todas as análises foram utilizadas com o auxílio do *software* R (R Core Team, 2021).

3.2.1 Dissimilaridade entre acessos

A medida de dissimilaridade adotada foi a Distância Generalizada de Mahalanobis, pois considerou-se que há certo grau de correlação entre as características mensuradas e que são possíveis de serem quantificadas quando as avaliações são realizadas em genótipos avaliados em delineamentos experimentais.

Desse modo, é possível obter a matriz de dispersão residual (Ψ) e as médias das características (CRUZ et al., 2014). Assim, por meio dessas informações, obtêm-se as estimativas das distâncias de Mahalanobis por meio da seguinte expressão:

$$D_{ij}^2 = \delta' \Psi^{-1} \delta$$

em que:

D_{ij}^2 é a distância de Mahalanobis entre os genótipos i e j;

Ψ é a matriz de variâncias e covariâncias residuais;

$\delta' = [d_1 \quad d_2 \quad \ldots \quad d_v]$, sendo $d_p = Y_{ip} - Y_{jp}$;

Y_{ip} é a média do $i-$ésimo genótipo em relação à $p-$ésima variável.

Exemplo ilustrativo: Considerando um experimento instalado em um delineamento inteiramente casualizado, com seis genótipos, em relação a três variáveis (Y_1, Y_2, Y_3).

Tabela 2 - Exemplo ilustrativo para obtenção da matriz de dissimilaridade.

Genótipos	$\boldsymbol{Y_1}$	$\boldsymbol{Y_2}$	$\boldsymbol{Y_3}$
1	25,8	21,5	20,6
2	27,1	32,0	52,0
3	41,0	25,5	39,8
4	22,4	33,3	41,2
5	30,6	46,0	45,3
6	35,2	45,4	28,0

Etapa 1: Para obtenção da matriz Ψ, é necessário obter as variâncias e covariâncias residuais entre cada par de variáveis. Assim:

$$\Psi = \begin{bmatrix} 46{,}295 & 7{,}341 & -3{,}855 \\ 7{,}341 & 101{,}387 & 29{,}101 \\ -3{,}855 & 29{,}101 & 133{,}0657 \end{bmatrix} \therefore \quad \Psi^{-1} = \begin{bmatrix} 0{,}021 & -0{,}0018 & 0{,}001 \\ -0{,}0018 & 0{,}010 & -0{,}0023 \\ 0{,}001 & -0{,}0023 & 0{,}008 \end{bmatrix}$$

Etapa 2: Nesta etapa, calcula-se as distâncias entre os pares de genótipos em relação a cada variável. Vem,

$$\delta_{12} = \begin{bmatrix} 25{,}8 - 27{,}1 \\ 21{,}5 - 32{,}0 \\ 20{,}6 - 52{,}0 \end{bmatrix}, \delta_{13} = \begin{bmatrix} 25{,}8 - 41{,}0 \\ 21{,}5 - 25{,}5 \\ 20{,}6 - 39{,}8 \end{bmatrix}, \ldots, \delta_{56} = \begin{bmatrix} 30{,}6 - 35{,}2 \\ 46{,}0 - 45{,}4 \\ 45{,}3 - 28{,}0 \end{bmatrix}$$

Etapa 3: Agora, para o cálculo da distância entre os genótipos 1 e 2, procede-se da seguinte forma:

$$D_{12}^2 = [-1{,}3 \quad -10{,}5 \quad -31{,}4]\begin{bmatrix} 0{,}021 & -0{,}0018 & 0{,}001 \\ -0{,}0018 & 0{,}010 & -0{,}0023 \\ 0{,}001 & -0{,}0023 & 0{,}008 \end{bmatrix}\begin{bmatrix} -1{,}3 \\ -10{,}5 \\ -31{,}4 \end{bmatrix} = 7{,}627$$

Assim, segue sucessivamente, até o último par de genótipos (5 e 6). Portanto, a matriz de distâncias generalizadas de Mahalanobis é:

$$D^2 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}\begin{bmatrix} 0{,}000 & 7{,}627 & 8{,}241 & 4{,}008 & 8{,}752 & 6{,}938 \\ 7{,}627 & 0{,}000 & 5{,}507 & 1{,}641 & 2{,}940 & 8{,}727 \\ 8{,}241 & 5{,}507 & 0{,}000 & 8{,}715 & 7{,}261 & 7{,}799 \\ 4{,}008 & 1{,}641 & 8{,}715 & 0{,}000 & 2{,}764 & 6{,}395 \\ 8{,}752 & 2{,}940 & 7{,}261 & 2{,}764 & 0{,}000 & 2{,}677 \\ 6{,}938 & 8{,}727 & 7{,}799 & 6{,}395 & 2{,}677 & 0{,}000 \end{bmatrix}$$

3.2.2 Métodos

3.2.2.1 Agrupamento Hierárquico do Vizinho mais Próximo

No Agrupamento Hierárquico do Vizinho mais Próximo, os progenitores mais similares são identificados e reunidos, formando o grupo inicial por meio da matriz de dissimilaridade. As distâncias deste grupo em relação aos demais progenitores ou, nos estágios mais avançados, em relação a outros grupos são calculadas.

Assim, o processo de identificação das entidades mais similares se repete sobre a nova matriz de dissimilaridade, cuja dimensão é reduzida a cada passo, e finaliza quando todos os progenitores são reunidos em um único grupo (CRUZ *et al*., 2012).

Logo, o dendrograma é estabelecido pelos genótipos com maior similaridade, sendo a distância entre um indivíduo e um grupo, formado pelos indivíduos i e j, dada por:

$$d_{(ij)k} = min\{d_{ik}; d_{jk}\}$$

ou seja, $d_{(ij)l}$ é dada pelo menor elemento do conjunto das distâncias dos pares de indivíduos (i e l) e (j e l).

A distância entre dois grupos é dada por:

$$d_{(ij)(lm)} = min\{d_{il}; d_{im}; d_{jl}; d_{jm}\}$$

ou seja, a distância entre dois grupos formados, respectivamente, pelos indivíduos (i e j) e (l e m) é dada pelo menor elemento do conjunto, cujos elementos são as distâncias entre os pares de indivíduos (i e l), (i e m), (j e l) e (j e m).

Exemplo ilustrativo: Considerando a matriz de dissimilaridade obtida anteriormente (item 3.3.1), aplicar o método do Vizinho mais próximo, de acordo com as etapas abaixo:

Etapa 1: Considerando a matriz de distâncias calculada anteriormente, identificam-se os genótipos mais similares e, consequentemente, calcula-se as distâncias entre esse grupo e os outros genótipos.

$$D^2 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}\begin{bmatrix} 0{,}000 & 7{,}627 & 8{,}241 & 4{,}008 & 8{,}752 & 6{,}938 \\ 7{,}627 & 0{,}000 & 5{,}507 & \mathbf{1{,}641} & 2{,}940 & 8{,}727 \\ 8{,}241 & 5{,}507 & 0{,}000 & 8{,}715 & 7{,}261 & 7{,}799 \\ 4{,}008 & 1{,}641 & 8{,}715 & 0{,}000 & 2{,}764 & 6{,}395 \\ 8{,}752 & 2{,}940 & 7{,}261 & 2{,}764 & 0{,}000 & 2{,}677 \\ 6{,}938 & 8{,}727 & 7{,}799 & 6{,}395 & 2{,}677 & 0{,}000 \end{bmatrix}$$

Os genótipos mais similares são o 2 e 4, com $d_{24} = 1{,}641$.

As distâncias entre esse grupo e os demais genótipos, seguem:

$$d_{(24)1} = \min(d_{21}, d_{41}) = \min(7{,}627; 4{,}008) = 4{,}008$$

$$d_{(24)3} = \min(d_{23}, d_{43}) = \min(5{,}507; 8{,}715) = 5{,}507$$

$$d_{(24)5} = \min(d_{25}, d_{45}) = \min(2{,}940; 2{,}764) = 2{,}764$$

$$d_{(24)6} = \min(d_{26}, d_{46}) = \min(8{,}727; 6{,}395) = 6{,}395$$

A nova matriz é:

$$D_1^2 = \begin{matrix} 1 \\ (24) \\ 3 \\ 5 \\ 6 \end{matrix} \begin{bmatrix} 0{,}000 & 4{,}008 & 8{,}241 & 8{,}752 & 6{,}938 \\ 4{,}008 & 0{,}000 & 5{,}507 & 2{,}764 & 6{,}395 \\ 8{,}241 & 5{,}507 & 0{,}000 & 7{,}261 & 7{,}799 \\ 8{,}752 & 2{,}764 & 7{,}261 & 0{,}000 & \mathbf{2{,}677} \\ 6{,}938 & 6{,}395 & 7{,}799 & 2{,}677 & 0{,}000 \end{bmatrix}$$

Etapa 2: Verificar, novamente, os genótipos mais similares.

Os genótipos mais similares são o 5 e 6, com $d_{56} = 2{,}677$.

$$d_{(56)1} = \min(d_{51}, d_{61}) = \min(8{,}752; 6{,}938) = 6{,}938$$

$$d_{(56)3} = \min(d_{53}, d_{63}) = \min(7{,}261; 7{,}799) = 7{,}261$$

$$d_{(56)(24)} = \min\left(d_{(24)5}, d_{(24)6}\right) = \min(2{,}764; 6{,}395) = 2{,}764$$

A nova matriz é:

$$D_2^2 = \begin{matrix} 1 \\ (24) \\ 3 \\ (56) \end{matrix} \begin{bmatrix} 0{,}000 & 4{,}008 & 8{,}241 & 6{,}938 \\ 4{,}008 & 0{,}000 & 5{,}507 & \mathbf{2{,}764} \\ 8{,}241 & 5{,}507 & 0{,}000 & 7{,}261 \\ 6{,}938 & 2{,}764 & 7{,}261 & 0{,}000 \end{bmatrix}$$

Etapa 3: Idem etapa 3.

Tem-se aqui grupos mais similares, que são (24) e (56), com $d_{(24)(56)} = 2{,}764$.

$$d_{(2456)1} = \min(d_{21}, d_{41}, d_{51}, d_{61}) = \min(7{,}627; 4{,}008; 8{,}752; 6{,}938) = 4{,}008$$

$$d_{(2456)3} = \min(d_{23}, d_{43}, d_{53}, d_{63}) = \min(5{,}507; 8{,}715; 7{,}261; 7{,}799) = 5{,}507$$

A nova matriz é:

$$D_3^2 = \begin{matrix} 1 \\ (2456) \\ 3 \end{matrix} \begin{bmatrix} 0{,}000 & 4{,}008 & 8{,}241 \\ \mathbf{4{,}008} & 0{,}000 & 5{,}507 \\ 8{,}241 & 5{,}507 & 0{,}000 \end{bmatrix}$$

Etapa 4: Idem etapa 2.

O grupo (2456) e o genótipo 1 são os mais similares, sendo $d_{(2456)1} = 4{,}008$.

$$d_{(24561)3} = \min(d_{23}, d_{43}, d_{53}, d_{63}, d_{13}) = \min(5{,}507; 8{,}715; 7{,}261; 7{,}799; 8{,}241) = 5{,}507$$

Logo, a nova matriz considerada é:

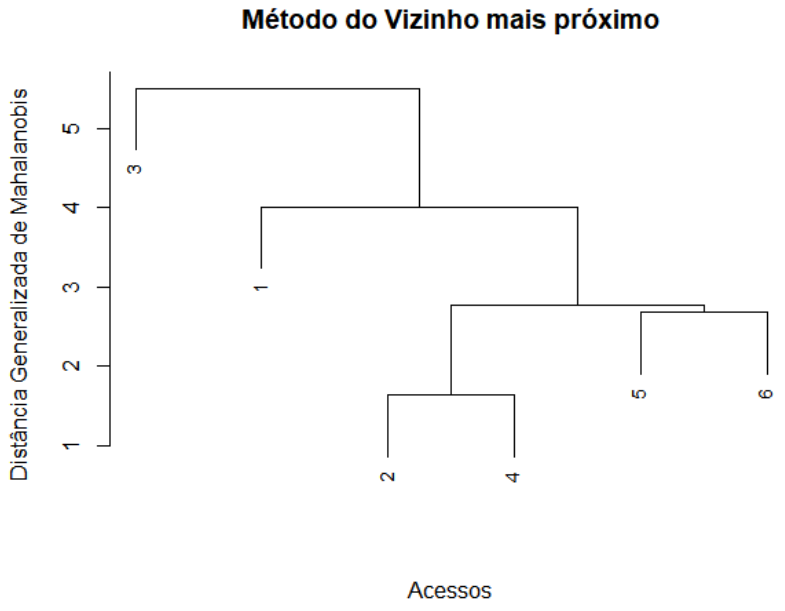
$$D_4^2 = \begin{matrix} \\ (24561) \end{matrix} \begin{matrix} 3 \\ \begin{bmatrix} 0{,}000 & 5{,}507 \\ 5{,}507 & 0{,}000 \end{bmatrix} \end{matrix}$$

Etapa 5: Por fim, o último grupo é composto por todos os genótipos. Assim, $d_{(24561)3} = 5{,}507$ e as fusões em cada fase segue abaixo.

Tabela 3 - Resultado do método do Vizinho mais próximo para o exemplo ilustrativo.

Etapa	Genótipo	Genótipo	Fusão
1	2	4	1,641
2	5	6	2,677
3	(24)	(56)	2,764
4	(2456)	1	4,008
5	(24561)	3	5,507

Figura 2 - Dendrograma obtido pelo método do Vizinho mais próximo, a partir da medida de dissimilaridade entre os genótipos do exemplo ilustrativo.



Fonte: Autor, 2021.

3.2.2.2 Agrupamento Hierárquico de Ward

O método de Ward, também denominado de método de variância mínima, foi proposto por Jr (1963). Diferentemente dos outros métodos hierárquicos aglomerativos, não segue o algoritmo básico de agrupamento apresentado. A razão dessa diferença é que ele não busca a menor distância entre dois grupos, mas sim a menor soma de quadrados dentro do grupo, ou seja, a menor variância interna. Num primeiro momento, ele permite a redução dos n conjuntos iniciais a $n - 1$ conjuntos mutuamente exclusivos considerando a fusão dos dois grupos, dentre todos os $\frac{n(n-1)}{2}$ pares possíveis, que resulte na menor soma de quadrados. Esse processo é repetido até haver um único grupo (SOUZA, 2017).

O agrupamento é feito a partir das somas dos quadrados dos desvios entre genótipos ou, alternativamente, a partir do quadrado da distância euclidiana, uma vez que se verifica a relação:

$$SQD_{ij} = \frac{1}{2} d_{ij}^2 = \sum_{j=1}^{n} SQD_{p(ij)}$$

sendo $SQD_{p(ij)}$ a soma de quadrados dos desvios, para a p −ésima variável, considerando os genótipos$i$ e j.

e

$$d_{ij}^2 = \sum_{p=1}^{n} (X_{ip} - X_{jp})^2$$

em que:

d_{ij}^2: quadrado da distância euclidiana entre os genótipos i e j;

n: número de características avaliadas;

X_{ij}: valor da característica p para o genótipo i.

A soma de quadrados dos desvios total é dada por:

$$SQDTotal = \frac{1}{g}\sum_{i<}^{g}\sum_{j}^{g} d_{ij}^2$$

sendo g o número de genótipos a serem agrupados.

Nesta análise de agrupamento, identifica-se, na matriz D (cujos elementos são os quadrados das distâncias euclidianas - d_{ij}^2) ou na matriz S (cujos elementos são as somas de quadrados dos desvios - SQD_{ij}), o par de genótipos que proporciona menor soma de quadrados dos desvios. Com esses genótipos agrupados, uma nova matriz de dissimilaridade, de menor dimensão, é recalculada, considerando que:

$SQD_{(ijk)} = \frac{1}{g} d_{(ijl)}^2$ (g é o número de genótipos no grupo, que neste caso é igual a 3).

$$d_{(ijl)}^2 = d_{(ij)}^2 + d_{(ij)l}^2 = d_{ij}^2 + d_{il}^2 + d_{jl}^2$$

e ainda que:

$SQD_{(ijlm)} = \frac{1}{g} d_{(ijlm)}^2$($g$ é numero de genótipos no grupo, que neste caso é igual a 4).

$$d_{(ijlm)}^2 = d_{ij}^2 + d_{il}^2 + d_{jl}^2 + d_{im}^2 + d_{jm}^2 + d_{lm}^2$$

e assim, sucessivamente.

No procedimento, realiza-se a análise, fornecendo os $g - 1$ (g é número de genótipos a serem agrupados) passos de agrupamento para que seja formado o dendrograma.

Lance e Williams (1967), porém, propuseram uma parametrização da fórmula de distâncias adaptável aos diversos métodos de agrupamento e assim, a implementação no método Ward possibilitou uma atualização da distância dos clusters nos agrupamentos, no qual otimizou a função.

O algoritmo é dado por:

$$d_{(ij)l} = \alpha_i d_{il} + \alpha_j d_{jl} + \beta d_{ij} + \gamma \left| d_{il} - d_{jl} \right|$$

$$\therefore\ d_{(ij)l} = \frac{n_i + n_l}{n_i + n_j + n_l} d_{il} + \frac{n_j + n_l}{n_i + n_j + n_l} d_{jl} + \frac{-n_k}{n_i + n_j + n_l} d_{ij}$$

em que:

$d_{(ij)l}$ é a distância entre os genótipos i e j e genótipo l;

n_i, n_j, n_l representam o número de genótipos envolvidos.

Assumindo $\alpha_i + \alpha_j + \beta = 1, \alpha_i = \alpha_j, \beta < 0$ e $\gamma = 0$.

Exemplo ilustrativo: Considerando a matriz de dissimilaridade obtida anteriormente (item 3.3.1), aplicar o método de Ward, de acordo com as etapas abaixo:

Etapa 1: Identifica-se, na matriz de dissimilaridade, os genótipos mais similares para compor o grupo I e, em seguida, calcular a distância em relação aos demais genótipos.

$$D^2 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}\begin{bmatrix} 0{,}000 & 7{,}627 & 8{,}241 & 4{,}008 & 8{,}752 & 6{,}938 \\ 7{,}627 & 0{,}000 & 5{,}507 & \mathbf{1{,}641} & 2{,}940 & 8{,}727 \\ 8{,}241 & 5{,}507 & 0{,}000 & 8{,}715 & 7{,}261 & 7{,}799 \\ 4{,}008 & 1{,}641 & 8{,}715 & 0{,}000 & 2{,}764 & 6{,}395 \\ 8{,}752 & 2{,}940 & 7{,}261 & 2{,}764 & 0{,}000 & 2{,}677 \\ 6{,}938 & 8{,}727 & 7{,}799 & 6{,}395 & 2{,}677 & 0{,}000 \end{bmatrix}$$

Grupo I:genótipos 2 e 4 $(d_{24} = 1{,}641)$

Calculando a distância com os demais genótipos, vem:

$$d_{(24)1} = \frac{n_2 + n_1}{n_2 + n_4 + n_1} d_{21} + \frac{n_4 + n_1}{n_2 + n_4 + n_1} d_{41} + \frac{-n_1}{n_2 + n_4 + n_1} d_{24}$$

$$= \frac{1+1}{1+1+1} 7{,}627 + \frac{1+1}{1+1+1} 4{,}008 + \frac{-1}{1+1+1} 1{,}641 = 7{,}209$$

$$d_{(24)3} = \frac{n_2 + n_3}{n_2 + n_4 + n_3} d_{23} + \frac{n_4 + n_3}{n_2 + n_4 + n_3} d_{43} + \frac{-n_3}{n_2 + n_4 + n_3} d_{24}$$

$$\frac{1+1}{1+1+1} 5{,}507 + \frac{1+1}{1+1+1} 8{,}715 + \frac{-1}{1+1+1} 1{,}641 = 8{,}934$$

$$d_{(24)5} = \frac{n_2 + n_5}{n_2 + n_4 + n_5} d_{25} + \frac{n_4 + n_5}{n_2 + n_4 + n_5} d_{45} + \frac{-n_5}{n_2 + n_4 + n_5} d_{24}$$

$$\frac{1+1}{1+1+1} 2{,}940 + \frac{1+1}{1+1+1} 2{,}764 + \frac{-1}{1+1+1} 1{,}641 = 3{,}255$$

$$d_{(24)6} = \frac{n_2 + n_6}{n_2 + n_4 + n_6} d_{26} + \frac{n_4 + n_6}{n_2 + n_4 + n_6} d_{46} + \frac{-n_6}{n_2 + n_4 + n_6} d_{24}$$

$$\frac{1+1}{1+1+1} 8{,}727 + \frac{1+1}{1+1+1} 6{,}395 + \frac{-1}{1+1+1} 1{,}641 = 9{,}534$$

A nova matriz de dissimilaridade(D_1^2) é:

$$D_1^2 = \begin{array}{c} \\ 1 \\ (2,4) \\ 3 \\ 5 \\ 6 \end{array} \begin{bmatrix} & & & & \\ 0{,}000 & 7{,}209 & 8{,}241 & 8{,}752 & 6{,}938 \\ 7{,}209 & 0{,}000 & 8{,}934 & 3{,}255 & 9{,}534 \\ 8{,}241 & 8{,}934 & 0{,}000 & 7{,}261 & 7{,}799 \\ 8{,}752 & 3{,}255 & 7{,}261 & 0{,}000 & \mathbf{2,677} \\ 6{,}938 & 9{,}534 & 7{,}799 & 2{,}677 & 0{,}000 \end{bmatrix}$$

Etapa 2: Identificar, nesta nova matriz (D_1^2), os genótipos mais similares.

Grupo *I1*: genótipos 5 e 6($d_{56} = 2{,}677$)

Distância entre os demais genótipos:

$$d_{(56)1} = \frac{n_5 + n_1}{n_5 + n_6 + n_1} d_{51} + \frac{n_6 + n_1}{n_5 + n_6 + n_1} d_{61} + \frac{-n_1}{n_5 + n_6 + n_1} d_{56}$$

$$= \frac{1+1}{1+1+1} 8{,}752 + \frac{1+1}{1+1+1} 6{,}938 + \frac{-1}{1+1+1} 2{,}677 = 9{,}567$$

$$d_{(56)3} = \frac{n_5 + n_3}{n_5 + n_6 + n_3} d_{53} + \frac{n_6 + n_3}{n_5 + n_6 + n_3} d_{63} + \frac{-n_3}{n_5 + n_6 + n_3} d_{56}$$

$$= \frac{1+1}{1+1+1} 3{,}255 + \frac{1+1}{1+1+1} 7{,}799 + \frac{-1}{1+1+1} 2{,}677 = 6{,}477$$

$$d_{(56)(24)} = \frac{n_5 + n_{24}}{n_5 + n_6 + n_{24}} d_{5(24)} + \frac{n_6 + n_{24}}{n_5 + n_6 + n_{24}} d_{6(24)} + \frac{-n_{24}}{n_5 + n_6 + n_{24}} d_{56}$$

$$= \frac{1+2}{1+1+2} 3{,}255 + \frac{1+1}{1+1+2} 9{,}534 + \frac{-1}{2+1+1} 2{,}677 = 6{,}539$$

A nova matriz de dissimilaridade (D_2^2) é:

$$D_2^2 = \begin{matrix} 1 \\ (24) \\ 3 \\ (56) \end{matrix} \begin{bmatrix} 0{,}000 & 7{,}209 & 8{,}241 & 9{,}567 \\ 7{,}209 & 0{,}000 & 8{,}934 & 6{,}539 \\ 8{,}241 & 8{,}934 & 0{,}000 & \mathbf{6{,}477} \\ 9{,}567 & 6{,}539 & 6{,}477 & 0{,}000 \end{bmatrix}$$

Etapa 3: Mesmo procedimento anterior.

Grupo *III*: genótipo 3 e grupo (56) $\left(d_{3(56)} = 6{,}477\right)$

$$d_{(563)1} = \frac{n_{56} + n_1}{n_{56} + n_3 + n_1} d_{(56)1} + \frac{n_3 + n_1}{n_{56} + n_3 + n_1} d_{31} + \frac{-n_1}{n_{56} + n_3 + n_1} d_{(56)3}$$

$$= \frac{2+1}{2+1+1} 9{,}567 + \frac{1+1}{2+1+1} 8{,}241 + \frac{-1}{2+1+1} 6{,}477 = 9{,}676$$

$$d_{(563)(24)} = \frac{n_{56} + n_{24}}{n_{56} + n_3 + n_{24}} d_{(56)(24)} + \frac{n_3 + n_{24}}{n_{56} + n_3 + n_{24}} d_{3(24)} + \frac{-n_{24}}{n_{56} + n_3 + n_{24}} d_{(56)3}$$

$$= \frac{2+2}{2+1+2} 6{,}539 + \frac{1+2}{2+1+2} 8{,}934 + \frac{-1}{2+1+1} 6{,}477 = 8{,}972$$

Logo, a nova matriz é:

$$D_3^2 = \begin{matrix} 1 \\ (2{,}4) \\ (563) \end{matrix} \begin{bmatrix} 0{,}000 & \mathbf{7{,}209} & 9{,}676 \\ 7{,}209 & 0{,}000 & 8{,}972 \\ 9{,}676 & 8{,}972 & 0{,}000 \end{bmatrix}$$

Etapa 4: Mesmo procedimento anterior.

Grupo *IV*: genótipo 1 e grupo (24) $\left(d_{1(24)} = 7{,}209\right)$

$$d_{(241)(563)} = \frac{n_{24} + n_{563}}{n_{24} + n_1 + n_{563}} d_{(24)(563)} + \frac{n_1 + n_{563}}{n_{24} + n_1 + n_{563}} d_{1(563)} + \frac{-n_{563}}{n_{24} + n_1 + n_{563}} d_{(24)1}$$

$$= \frac{2+3}{2+1+3} 8{,}972 + \frac{1+3}{2+1+3} 9{,}676 + \frac{-3}{2+1+3} 7{,}209 = 10{,}322$$

Logo, a nova matriz é:

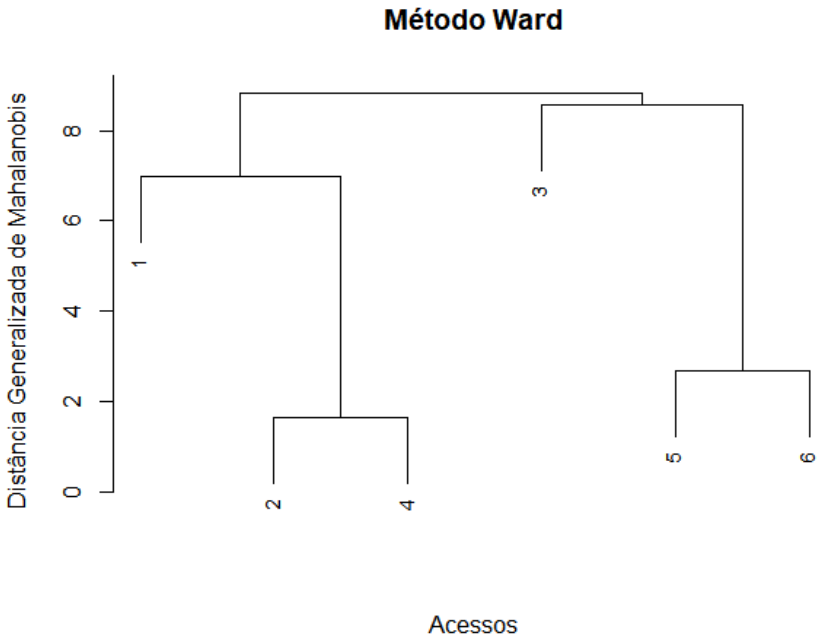
$$D_4^2 = \begin{matrix} (241) \\ (563) \end{matrix} \begin{bmatrix} 0{,}000 & 10{,}322 \\ 10{,}322 & 0{,}000 \end{bmatrix}$$

Etapa 5: O processo termina quando são formados somente um grupo. Assim, o nível de fusão é 10,322.

Tabela 4 - Resultado do método de Ward para o exemplo ilustrativo.

Etapas	**Fusão**		**Nível de fusão**	**Número de genótipos**
	Genótipo	Genótipo		
1	2	4	1,641	2
2	5	6	2,677	2
3	3	(56)	6,477	3
4	1	(24)	7,209	3
5	(241)	(563)	10,322	6

Figura 3 - Dendrograma obtido pelo método de Ward, a partir da medida de dissimilaridade entre os genótipos do exemplo ilustrativo.



Fonte: Autor, 2021.

3.2.2.3 Agrupamento de Otimização das k −médias

Um dos métodos não-hierárquicos mais conhecidos e utilizados atualmente é o k − médias, que é um algoritmo de agrupamento de dados não-hierárquico que utiliza uma técnica iterativa para particionar um conjunto de dados. Esse algoritmo busca minimizar a distância dos elementos de um conjunto de dados com k centros de forma iterativa (PALMA, 2018).

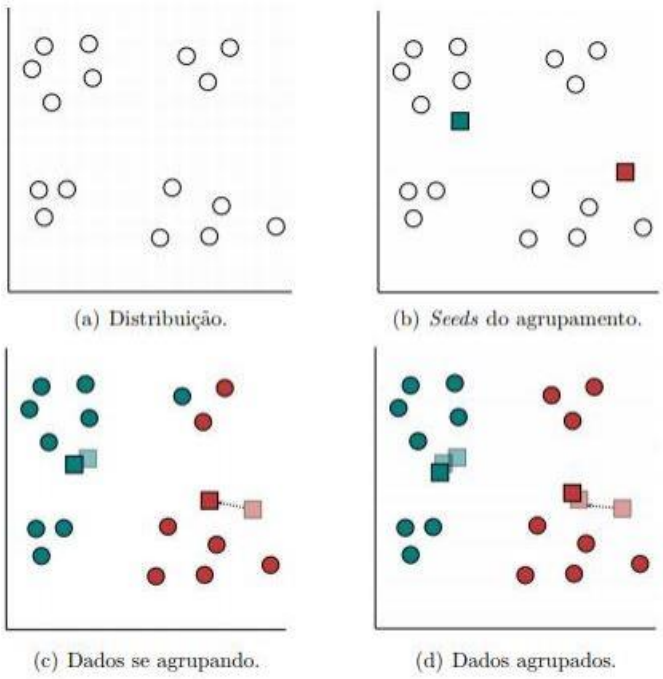
O processo iterativo do método pode ser descrito pelos seguintes passos:

1. Inicialmente são escolhidos k centroides, denominados de "sementes", calculados com base no número de grupos escolhido a *priori*;
2. Uma medida de distância é aplicada para comparar cada objeto a cada centroide inicial, sendo que o objeto é unido ao grupo de menor distância;
3. Os valores dos centroides são recalculados considerando cada grupo formado, então o passo 2 é repetido com os novos vetores de médias calculados para os novos grupos;
4. Os passos 2 e 3 são repetidos até que não haja mais realocação dos objetos entre os grupos.

O agrupamento final obtido por meio do método das k −médias depende diretamente da escolha das sementes (FERREIRA, 2011). Mingoti (2005) apresenta algumas propostas para a definição das sementes, sendo elas: aplicação de técnicas hierárquicas aglomerativas, escolha aleatória ou via observação dos valores discrepantes do conjunto de observações.

A Figura 4, a seguir, resume o método.

Figura 4 - Apresentação do passo a passo para a aplicação do método de Agrupamento de Otimização das k −médias.



Fonte: De PALMA, 2018, à partir de PRADO, 2008.

Exemplo ilustrativo: Considerando a matriz de dissimilaridade obtida anteriormente (item 3.3.1), aplicar o método de Agrupamento de Otimização das k −médias, de acordo com as etapas abaixo:

Etapa 1: Agrupa-se, inicialmente, os objetos de forma arbitrária em (12), (34), (56) e calcula-se os centroides dos grupos. Assim:

Tabela 5 - Objetos do exemplo ilustrativo reunidos de maneira arbitrária.

Grupos	$\overline{y_1}$	$\overline{y_2}$	$\overline{y_3}$
$g_1 = (12)$	26,45	26,75	36,3
$g_2 = (34)$	31,7	29,4	40,5
$g_3 = (56)$	32,9	45,7	36,65

Etapa 2: Calcula-se, a distância euclidiana de cada objeto para o centroide de seu grupo e para os demais grupos. Para o genótipo 1, tem-se:

$$d_{1,g_1} = \sqrt{(25{,}8 - 26{,}45)^2 + (21{,}5 - 26{,}75)^2 + (52{,}0 - 36{,}3)^2} = 16{,}567$$

$$d_{1,g_2} = \sqrt{(25{,}8 - 31{,}7)^2 + (21{,}5 - 29{,}4)^2 + (20{,}6 - 40{,}5)^2} = 22{,}208$$

$$d_{1,g_3} = \sqrt{(25{,}8 - 32{,}9)^2 + (21{,}5 - 45{,}7)^2 + (20{,}6 - 36{,}65)^2} = 29{,}894$$

A menor distância encontrada para o genótipo 1 foi, justamente, no grupo 1. Logo, não há realocação do genótipo e daí, passamos para o genótipo 2.

$$d_{2,g_1} = \sqrt{(27{,}1 - 26{,}45)^2 + (32{,}0 - 26{,}75)^2 + (52{,}0 - 36{,}3)^2} = 16{,}567$$

$$d_{2,g_2} = \sqrt{(27{,}1 - 31{,}7)^2 + (32{,}0 - 29{,}4)^2 + (52{,}0 - 40{,}5)^2} = 12{,}655$$

$$d_{2,g_3} = \sqrt{(27{,}1 - 32{,}9)^2 + (32{,}0 - 45{,}7)^2 + (52{,}0 - 36{,}65)^2} = 21{,}376$$

Tem-se que a menor distância para o genótipo 2 é com o grupo 2. Portanto, ele é realocado para o mesmo.

Etapa 3: Realoca-se o genótipo 2 para o grupo 2 e recalcula-se o centroide.

Tabela 6 - Realocação do genótipo 2 para o grupo 2.

Grupos	$\overline{y_1}$	$\overline{y_2}$	$\overline{y_3}$
$\boldsymbol{g_1 = 1}$	25,8	21,5	20,6
$\boldsymbol{g_2 = (234)}$	30,16	30,26	44,33
$\boldsymbol{g_3 = (56)}$	32,9	45,7	36,65

Para o genótipo 2, as distâncias com os centroides de cada grupo são:

$$d_{2,g_1} = \sqrt{(27{,}1 - 25{,}8)^2 + (32{,}0 - 21{,}5)^2 + (52{,}0 - 20{,}6)^2} = 33{,}134$$

$$d_{2,g_2} = \sqrt{(27{,}1 - 30{,}16)^2 + (32{,}0 - 30{,}26)^2 + (52{,}0 - 44{,}33)^2} = 8{,}439$$

$$d_{2,g_3} = \sqrt{(27{,}1 - 32{,}9)^2 + (32{,}0 - 45{,}7)^2 + (52{,}0 - 36{,}65)^2} = 21{,}376$$

Etapa 4: Como não houve realocação do genótipo 2, então passamos para o genótipo 3.

$$d_{3,g_1} = \sqrt{(41{,}0 - 25{,}8)^2 + (25{,}5 - 21{,}5)^2 + (39{,}8 - 20{,}6)^2} = 24{,}812$$

$$d_{3,g_2} = \sqrt{(41{,}0 - 30{,}16)^2 + (25{,}5 - 30{,}26)^2 + (39{,}8 - 44{,}33)^2} = 12{,}676$$

$$d_{3,g_3} = \sqrt{(41{,}0 - 32{,}9)^2 + (25{,}5 - 45{,}7)^2 + (39{,}8 - 36{,}65)^2} = 21{,}990$$

A menor distância encontrada para o genótipo 3 foi no grupo 2, no qual ele já pertence. Logo, não há realocação do genótipo e daí, passamos para o genótipo 4.

$$d_{4,g_1} = \sqrt{(22{,}4 - 25{,}8)^2 + (33{,}3 - 21{,}5)^2 + (41{,}2 - 20{,}6)^2} = 23{,}982$$

$$d_{4,g_2} = \sqrt{(22{,}4 - 30{,}16)^2 + (33{,}3 - 30{,}26)^2 + (41{,}2 - 44{,}33)^2} = 8{,}902$$

$$d_{4,g_3} = \sqrt{(22{,}4 - 32{,}9)^2 + (33{,}3 - 45{,}7)^2 + (41{,}2 - 36{,}65)^2} = 16{,}873$$

O mesmo acontece com o genótipo 4, logo, não é realocado.

Calcula-se, agora, a distância entre os grupos e o genótipo 5.

$$d_{5,g_1} = \sqrt{(30{,}6 - 25{,}8)^2 + (46{,}0 - 21{,}5)^2 + (45{,}3 - 20{,}6)^2} = 35{,}119$$

$$d_{5,g_2} = \sqrt{(30{,}6 - 30{,}16)^2 + (46{,}0 - 30{,}26)^2 + (45{,}3 - 44{,}33)^2} = 15{,}776$$

$$d_{5,g_3} = \sqrt{(30{,}6 - 32{,}9)^2 + (46{,}0 - 45{,}7)^2 + (45{,}3 - 36{,}65)^2} = 8{,}955$$

O genótipo 5 também não é realocado e assim, calcula-se as distâncias para o genótipo 6.

$$d_{6,g_1} = \sqrt{(35{,}2 - 25{,}8)^2 + (45{,}4 - 21{,}5)^2 + (28{,}0 - 20{,}6)^2} = 26{,}726$$

$$d_{6,g_2} = \sqrt{(35{,}2 - 30{,}16)^2 + (45{,}4 - 30{,}26)^2 + (28{,}0 - 44{,}33)^2} = 22{,}831$$

$$d_{6,g_3} = \sqrt{(35{,}2 - 32{,}9)^2 + (45{,}4 - 45{,}7)^2 + (28{,}0 - 36{,}65)^2} = 8{,}955$$

Etapa 5: Por fim, o genótipo 6 não é realocado e o agrupamento final é dado por:

Tabela 7 - Resultado obtido por meio da aplicação do método de Agrupamento de Otimização das k −médias para o exemplo ilustrativo.

Grupo	Genótipos
I	1
II	(234)
III	(56)

3.2.2.4 Agrupamento de Otimização de Tocher

No método de Tocher, adota-se o critério de que a média das medidas de dissimilaridade dentro de cada grupo deve ser menor que as distancias médias entre quaisquer grupos. O método requer a obtenção da matriz de dissimilaridade, sobre a qual é identificado o par de progenitores mais similar. Esses progenitores formarão o grupo inicial. A partir daí é avaliada a possibilidade de inclusão de novos progenitores, adotando-se o critério anteriormente citado (CRUZ *et al.*, 2012).

Assim, a entrada de um indivíduo em um grupo sempre aumenta o valor médio da distância dentro do grupo.

Portanto, a inclusão, ou não, do genótipo$l$ no grupo é feita considerando:

- Se $\frac{d_{(grupo)l}}{g} \leq \theta$, inclui-se o indivíduo l no grupo;

-Se $\frac{d_{(grupo)l}}{g} > \theta$, o indivíduo l não é incluído no grupo.

em que:

$d_{(grupo)l}$ é a distância média entre o genótipo l e um determinado grupo;

θ é o valor máximo da medida de dissimilaridade encontrado no conjunto das menores distâncias envolvendo cada genótipo;

g é o número de genótipos que constitui o grupo original.

Nesse caso, a distância entre o genótipo l e o grupo formado pelos indivíduos ij é dada por:

$$d_{(ij)l} = d_{il} + d_{jl}.$$

Exemplo ilustrativo: Considerando a matriz de dissimilaridade obtida anteriormente (item 3.3.1), aplicar o método de Tocher, de acordo com as etapas abaixo:

Etapa 1: É necessário identificar na matriz de dissimilaridade a menor distância em cada genótipo.

$$D^2 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{bmatrix} 0{,}000 & 7{,}627 & 8{,}241 & \mathbf{4{,}008} & 8{,}752 & 6{,}938 \\ 7{,}627 & 0{,}000 & 5{,}507 & \mathbf{1{,}641} & 2{,}940 & 8{,}727 \\ 8{,}241 & \mathbf{5{,}507} & 0{,}000 & 8{,}715 & 7{,}261 & 7{,}799 \\ 4{,}008 & \mathbf{1{,}641} & 8{,}715 & 0{,}000 & 2{,}764 & 6{,}395 \\ 8{,}752 & 2{,}940 & 7{,}261 & 2{,}764 & 0{,}000 & \mathbf{2{,}677} \\ 6{,}938 & 8{,}727 & 7{,}799 & 6{,}395 & \mathbf{2{,}677} & 0{,}000 \end{bmatrix}$$

Assim, as menores distâncias são 4,008 (A1), 1,641 (A2), 5,507 (A3), 1,641 (A4), 2,677 (A5), 2,677 (A6).

Desse modo, tem-se que o critério de agrupamento (o maior entre as menores distâncias) é $\theta = 5{,}507$.

Etapa 2: Nesta etapa, identifica-se na matriz de dissimilaridade os genótipos mais similares, de modo a formar o grupo I e, consequentemente, calcular a distância entre o grupo e os demais genótipos.

Grupo I: genótipos 2 e 4 $(d_{24} = 1{,}641)$

Calculando, agora, a distância entre o grupo e os demais genótipos:

$$d_{(24)1} = d_{21} + d_{41} = 7{,}627 + 4{,}008 = 11{,}635$$

$$d_{(24)3} = d_{23} + d_{43} = 5{,}507 + 8{,}715 = 14{,}222$$

$$d_{(24)5} = d_{25} + d_{45} = 2{,}940 + 2{,}764 = 5{,}704$$

$$d_{(24)6} = d_{26} + d_{46} = 8{,}727 + 6{,}395 = 15{,}122$$

De acordo com as distâncias calculadas, o genótipo 5 é o mais similar, pois possui a menor distância. Avaliando a possibilidade de inclusão no grupo *I*, tem-se:

$$\frac{d_{(24)5}}{2} \leq \theta \quad \therefore \quad \frac{5{,}704}{2} = 2{,}852 \leq \theta$$

Logo, o genótipo 5 é incluído no grupo *I*.

Etapa 3: Calcular a distância entre o novo grupo formado e os demais genótipos (1,3 e 6).

$$d_{(245)1} = d_{21} + d_{41} + d_{51} = 7{,}627 + 4{,}008 + 8{,}752 = 20{,}387$$

$$d_{(245)3} = d_{23} + d_{43} + d_{53} = 5{,}507 + 8{,}715 + 7{,}261 = 21{,}483$$

$$d_{(245)6} = d_{26} + d_{46} + d_{56} = 8{,}727 + 6{,}395 + 2{,}677 = 17{,}799$$

De acordo com as distâncias calculadas, o genótipo 6 é o mais similar. Avaliando a possibilidade de inclusão no grupo *I*, tem-se:

$$\frac{d_{(245)6}}{3} \leq \theta \quad \therefore \quad \frac{17{,}799}{3} = 5{,}933 > \theta$$

Logo, o genótipo 6 não pode ser incluído no grupo *I*.

Etapa 4: Para obtenção do grupo *II*, repete-se a etapa anterior com os genótipos não-agrupados. Considerando a matriz de dissimilaridade abaixo, temos:

$$D^2 = \begin{bmatrix} 0 & 8{,}241 & 6{,}938 \\ 8{,}241 & 0 & 7{,}799 \\ 6{,}938 & 7{,}799 & 0 \end{bmatrix}$$

Assim, $d_{16} = 6{,}938$, que é a menor dessa matriz, revela-se com um valor superior ao critério de agrupamento ($\theta = 5{,}507$). Do mesmo modo, o segundo e terceiro valores, $7{,}799$ e $8{,}241$, respectivamente, também são maiores que o critério.

Portanto, o agrupamento não é estabelecido, ficando cada genótipo em grupos separados.

O resultado do agrupamento pelo método de Tocher, é apresentado na Tabela 8 abaixo:

Tabela 8 - Resultado obtido por meio da aplicação do método de Agrupamento de Tocher para o exemplo ilustrativo.

Grupo	**Genótipos**	**Distância média intragrupo**
I	2,4,5	2,448
II	1	–
III	3	–
IV	6	–

As distâncias médias intergrupos são apresentadas na Tabela 9 abaixo:

Tabela 9 - Distâncias médias intergrupos obtidas por meio da aplicação do método de Agrupamento de Tocher.

Grupo	**II**	**III**	**IV**
I	6,795	7,161	5,933
II		8,241	6,938
III			7,799

3.3.3 Determinação do número de grupos

Uma dificuldade referente à organização dos dados é a determinação do número de grupos. Segundo Regazzi e Cruz (2020), uma sugestão para os métodos hierárquicos, é o

exame do dendrograma em busca de grandes alterações dos níveis de similaridade para as sucessivas fusões. Já para os métodos não-hierárquicos, com foco no k −médias, utiliza-se o Método de Cotovelo para encontrar o número ótimo de grupos. Os procedimentos para solucionar esse problema são explicitados abaixo.

3.3.3.1 Critério de Mojena

Mojena (1977) propôs um procedimento que baseia-se no tamanho relativo dos níveis de fusões (distâncias) no dendrograma. Sugere-se selecionar o número de grupos no passo j que, primeiramente, satisfizer a inequação abaixo:

$$\alpha_j > \theta_k$$

onde,

α_j é o valor da distância do nível de fusão correspondente ao passo j $(j = 1,2, \dots n-1)$;

θ_k é valor referencial de corte, dado por:

$$\theta_k = \bar{\alpha} + k\widehat{\sigma_\alpha}$$

em que:

$\bar{\alpha}$ e $\widehat{\sigma_\alpha}$ são a média e o desvio padrão dos $\alpha's$;

k é uma constante. Mojena sugere usar valores entre $2{,}75$ e $3{,}5$ e Milligan & Cooper (1985) sugerem utilizar $1{,}25$, baseando em resultados de simulação.

Assim,

$$\bar{\alpha} = \frac{1}{n-1}\sum_{j=1}^{n}\alpha_j$$

e

$$S_\alpha = \sqrt{\frac{\sum_{j=1}^{n}\alpha^2\frac{1}{n-1}(\sum_{j=1}^{n}\alpha_j)^2}{n-2}}.$$

Exemplo ilustrativo: Considerando a matriz de dissimilaridade obtida anteriormente (item 3.3.1) e os métodos do Vizinho mais próximo e Ward, itens 3.3.2.1 e 3.3.2.2, respectivamente, determinaremos aqui, o número de grupos pelo Critério de Mojena.

Método 1: Vizinho mais próximo

$$\theta_k = \bar{\alpha} + k\widehat{\sigma_\alpha}$$

$$\bar{\alpha} = \frac{1{,}641 + 2{,}677 + 2{,}764 + 4{,}008 + 5{,}507}{5} = 3{,}3194$$

$$\widehat{\sigma_\alpha} = \sqrt{\frac{1{,}641^2 + 2{,}677^2 + 2{,}764^2 + 4{,}008^2 + 5{,}507^2 - \frac{(16{,}597)^2}{5}}{4}} = 3{,}134$$

$$\therefore \theta_k = 3{,}3194 + 1{,}25 * 3{,}134 = 5{,}173$$

Na Tabela 10, a seguir, tem-se a determinação do número de grupos, por meio do Critério de Mojena para o método do Vizinho mais próximo.

Tabela 10 - Resultado da determinação do número de grupos para o método do Vizinho mais próximo.

Etapa	$\boldsymbol{\alpha}$	$\boldsymbol{\bar{\alpha}}$	$\boldsymbol{\widehat{\sigma_\alpha}}$	$\boldsymbol{\theta_k(k = 1{,}25)}$
1	1,641	−		
2	2,677	2,159	0,7325	3,074
3	2,764	2,3606	0,6247	3,141
4	4,008	2,7725	0,968	3,982
5	5,507	3,3194	1,483	5,173

Método 2: Ward

$$\theta_k = \bar{\alpha} + k\widehat{\sigma_\alpha}$$

$$\bar{\alpha} = \frac{1{,}641 + 2{,}677 + 6{,}477 + 7{,}209 + 10{,}322}{5} = 5{,}6652$$

$$\widehat{\sigma_\alpha} = \sqrt{\frac{1{,}641^2 + 2{,}677^2 + 6{,}477^2 + 7{,}209^2 + 10{,}322^2 - \frac{(28{,}326)^2}{5}}{4}} = 3{,}530$$

$$\therefore \theta_k = 5{,}6652 + 1{,}25 * 3{,}530 = 10{,}077$$

Na Tabela 11, a seguir, tem-se a determinação do número de grupos, por meio do Critério de Mojena para o método de Ward.

Tabela 11 - Resultado da determinação do número de grupos para o método de Ward.

Etapa	$\boldsymbol{\alpha}$	$\boldsymbol{\bar{\alpha}}$	$\boldsymbol{\widehat{\sigma_\alpha}}$	$\boldsymbol{\theta_k(k = 1,25)}$
1	1,641	–	–	
2	2,677	2,159	0,732	3,074
3	6,477	3,598	2,546	6,780
4	7,209	4,501	2,753	7,942
5	10,322	5,6652	3,530	10,077

3.3.3.2 RMSSTD

Sharma (1996) ressalta que o índice RMSSTD (*Root Mean Square Standard Deviation*), ou seja, raiz quadrada do desvio padrão médio, é usado para calcular a homogeneidade dos agrupamentos. Em outras palavras, quanto mais compactos forem os grupos formados, situação verificada na presença de um grande número de grupos, menores os valores para esta estatística (FARIA, 2009).

Desse modo, é possível verificar por um gráfico, que mostra o decréscimo do RMSSTD em função do aumento do número de clusters. Porém, é uma trajetória não linear e o seu ponto de máxima curvatura indica um limiar entre uma fase de decréscimo e uma fase de estabilização. Assim, depois deste ponto, que é denominado ótimo, mesmo que aumente o número de *clusters*, não verifica-se grandes declínios nos valores de RMSSTD.

Logo, o cálculo de RMSSTD, para cada novo grupo formado é realizado por meio da seguinte expressão:

$$RMSSTD = \sqrt{\frac{\sum_{\substack{k=1,\dots,nc \\ p=1,\dots,d}} \sum_{k=1}^{n_p} (x_i - \overline{x_p})}{\sum_{\substack{k=1,\dots,nc \\ p=1,\dots,d}} (n_{kp} - 1)}}$$

em que:

$k = 1, \dots nc$ e $p = 1, \dots d$;

nc é o número de grupos;

d é o número de variáveis;

$\overline{x_p}$é o valor esperado na p −ésima variável;

n_{kp} é o número de genótipos no k −ésimo grupo na p −ésima variável;

n_p é o número de genótipos na p −ésima variável em todo o conjunto de dados.

Exemplo ilustrativo: Considerando os métodos do Vizinho mais próximo e Ward, pode-se obter o RMSSTD para cada. Assim:

Método 1: Vizinho mais próximo

Grupos:

$$24 = (27{,}1 - 30{,}35) + (22{,}4 - 30{,}35) + (32 - 33{,}95) + (33{,}3 - 33{,}95) + (52 - 37{,}82) + (41{,}2 - 37{,}82) = 3{,}76$$

$$56 = (30{,}6 - 30{,}35) + (35{,}2 - 30{,}35) + (46 - 33{,}95) + (45{,}4 - 33{,}95) + (45{,}3 - 37{,}82) + (28 - 37{,}82) = 26{,}26$$

$$2456 = 3{,}76 + 26{,}26 = 30{,}02$$

$$24561 = (25{,}8 - 30{,}35) + (21{,}5 - 33{,}95) + (20.6 - 37.82) + 30{,}02 = -4{,}2$$

$$245613 = (41 - 30{,}35) + (25{,}5 - 33{,}95) + (39{,}8 - 37{,}82) + (-4{,}2) = -0{,}02$$

Logo,

$$RMSSTD = \sqrt{\frac{3{,}76 + 26{,}26 + 30{,}02 - 4{,}2 - 0{,}02}{1 + 1 + 3 + 4 + 5}}$$

$$RMSSTD = \sqrt{3{,}987143}$$

$$RMSSTD = 1{,}996783$$

Método 1: Ward

Grupos:

$$24 = (27{,}1 - 30{,}35) + (22{,}4 - 30{,}35) + (32 - 33{,}95) + (33{,}3 - 33{,}95) + (52 - 37{,}82) + (41{,}2 - 37{,}82) = 3{,}76$$

$$56 = (30{,}6 - 30{,}35) + (35{,}2 - 30{,}35) + (46 - 33{,}95) + (45{,}4 - 33{,}95) + (45{,}3 - 37{,}82) + (28 - 37{,}82) = 26{,}26$$

$$356 = (41 - 30{,}35) + (25{,}5 - 33{,}95) + 26{,}26 = 30{,}44$$

$$124 = (25{,}8 - 30{,}35) + (21{,}5 - 33{,}95) + (20.6 - 37.82) + 3{,}76 = -30{,}46$$

$$241563 = -30{,}46 + 30{,}44 = -0{,}02$$

Logo,

$$RMSSTD = \sqrt{\frac{3{,}76 + 26{,}26 + 30{,}44 - 30{,}46 - 0{,}02}{1 + 1 + 3 + 4 + 5}}$$

$$RMSSTD = \sqrt{2{,}998}$$

$$RMSSTD = 1{,}731473$$

3.3.3.3 Método do Cotovelo

A Curva de Cotovelo ou Método Elbow Curve é uma técnica usada para encontrar a quantidade ideal de clusters k. Este método testa a variância dos dados em relação ao número de clusters. O valor ideal de k é aquele que tem uma menor Soma de Quadrados interna e ao mesmo tempo o menor número de clusters. Chamamos de curva de cotovelo, porque a partir

do ponto que seria o "cotovelo" não existe uma discrepância tão significativa em termos de variância. Dessa forma, a melhor quantidade de clusters k seria exatamente onde o cotovelo estaria (*site* Kaggle, 2019).

3.3.4 Coeficiente de Correlação Cofenética

O coeficiente de correlação cofenética (CCC) é uma medida de validação utilizada, principalmente, nos métodos de agrupamento hierárquicos. Este 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 (ALBUQUERQUE, 2005).

O CCC é um coeficiente de correlação momento-produto calculado considerando a seguinte expressão:

$$r_{cof} = \frac{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}(c_{jj'} - \bar{c})(f_{jj'} - \bar{f})}{\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}(c_{jj'} - \bar{c})^2}\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}(f_{jj'} - \bar{f})^2}}$$

onde:

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

$$\bar{f} = \frac{2}{n(n-1)}\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n} f_{jj'}$$

$c_{jj'}$é o valor da distância entre indivíduos j e j' na matriz cofenética;

$f_{jj'}$ é o valor da distância entre os mesmos indivíduos na matriz original de distâncias;

n é a dimensão da matriz.

Logo, quanto maior for o CCC, menor será a distorção ocasionada pelo agrupamento dos genótipos frente à utilização de determinado método.

Exemplo ilustrativo: Considerando a matriz de dissimilaridade obtida anteriormente (item 3.3.1), pode-se obter o coeficiente de correlação cofenética para os métodos do Vizinho mais próximo e Ward.

Método 1: Vizinho mais próximo

- Matriz cofenética

$$C_1 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{bmatrix} 0{,}000 & 4{,}008 & 5{,}507 & 4{,}008 & 4{,}008 & 4{,}008 \\ 4{,}008 & 0{,}000 & 5{,}507 & 1{,}641 & 2{,}764 & 2{,}764 \\ 5{,}507 & 5{,}507 & 0{,}000 & 5{,}507 & 5{,}507 & 5{,}507 \\ 4{,}008 & 1{,}641 & 5{,}507 & 0{,}000 & 2{,}764 & 2{,}764 \\ 4{,}008 & 2{,}764 & 5{,}507 & 2{,}764 & 0{,}000 & 2{,}677 \\ 4{,}008 & 2{,}764 & 5{,}507 & 2{,}764 & 2{,}677 & 0{,}000 \end{bmatrix}$$

- Matriz de dissimilaridade

$$D^2 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix} \begin{bmatrix} 0{,}000 & 7{,}627 & 8{,}241 & 4{,}008 & 8{,}752 & 6{,}938 \\ 7{,}627 & 0{,}000 & 5{,}507 & 1{,}641 & 2{,}940 & 8{,}727 \\ 8{,}241 & 5{,}507 & 0{,}000 & 8{,}715 & 7{,}261 & 7{,}799 \\ 4{,}008 & 1{,}641 & 8{,}715 & 0{,}000 & 2{,}764 & 6{,}395 \\ 8{,}752 & 2{,}940 & 7{,}261 & 2{,}764 & 0{,}000 & 2{,}677 \\ 6{,}938 & 8{,}727 & 7{,}799 & 6{,}395 & 2{,}677 & 0{,}000 \end{bmatrix}$$

Definindo os cálculos para o coeficiente de correlação cofenética, tem-se:

$$\bar{c} = \frac{2}{6(6-1)}(4{,}008 + 5{,}507 + \cdots + 2{,}677) = 3{,}9294$$

e

$$\bar{d} = \frac{2}{6(6-1)}(7{,}627 + 8{,}241 + \cdots + 2{,}677) = 5{,}999$$

Logo, $r_{cof} = 0{,}635$.

Método 2: Ward

Sendo a matriz cofenética:

$$C_2 = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}\begin{bmatrix} 0{,}000 & 6{,}971 & 8{,}844 & 6{,}971 & 8{,}844 & 8{,}844 \\ 6{,}971 & 0{,}000 & 8{,}844 & 1{,}641 & 8{,}844 & 8{,}844 \\ 8{,}844 & 8{,}844 & 0{,}000 & 8{,}844 & 8{,}562 & 8{,}562 \\ 6{,}971 & 1{,}641 & 8{,}844 & 0{,}000 & 8{,}844 & 8{,}844 \\ 8{,}844 & 8{,}844 & 8{,}562 & 8{,}844 & 0{,}000 & 2{,}677 \\ 8{,}844 & 8{,}844 & 8{,}562 & 8{,}844 & 2{,}677 & 0{,}000 \end{bmatrix}$$

E a matriz de dissimilaridade usada desde o início, tem-se:

$$\bar{c} = \frac{2}{6(6-1)}(6{,}971 + 8{,}844 + \cdots + 2{,}677) = 7{,}665$$

e

$$\bar{d} = 5{,}999$$

Logo, $r_{cof} = 0{,}624$.

4 RESULTADOS E DISCUSSÃO

Os dados obtidos por meio do experimento e os mesmos padronizados estão apresentados na Tabela 12 a seguir.

Tabela 12 - Dados originais e dados padronizados obtidos por meio do experimento em suas respectivas características.

Trat	COM	LAR	PT	ESP	VIT	NS
1	27,55(-1,53)	8,89(-1,21)	1,51(-0,92)	0,99(-1,13)	114,67(0,75)	42,75(-1,21)
2	76,81(0,07)	8,03(-1,27)	3,48(-0,84)	0,91(-1,21)	109,84(0,50)	67,37(-0,49)
3	53,10(-0,69)	20,02(-0,43)	8,62(-0,63)	1,32(-0,76)	148,00(2,53)	32,12(-1,51)
4	17,87(-1,84)	6,93(-1,35)	0,67(-0,95)	0,72(-1,42)	128,50(1,49)	35,00(-1,43)
5	29,53(-1,46)	30,71(0,31)	9,17(-0,61)	1,26(-0,82)	98,95(-0,08)	52,25(-0,93)
6	121,57(1,53)	35,16(0,62)	52,96(1,12)	2,77(0,83)	91,28(-0,48)	151,37(1,93)
7	69,31(-0,17)	69,95(3,06)	109,92(3,39)	4,96(3,24)	76,26(-1,29)	110,00(0,73)
8	100,40(0,84)	44,32(1,26)	66,05(1,64)	3,48(1,61)	82,75(-0,94)	98,46(0,40)
9	76,68(0,07)	53,85(1,93)	79,29(2,17)	4,29(2,50)	77,37(-1,23)	143,25(1,69)
10	61,32(-0,43)	14,79(-0,80)	7,47(-0,68)	1,87(-0,16)	104,54(0,21)	57,62(-0,78)
11	62,32(-0,39)	18,21(-0,56)	10,56(-0,56)	1,90(-0,12)	112,66(0,65)	68,25(-0,47)
12	56,34(-0,59)	15,88(-0,72)	7,15(-0,69)	1,75(-0,28)	105,71(0,28)	15,31(-2,00)
13	58,94(-0,50)	18,87(-0,51)	9,66(-0,59)	1,86(-0,17)	104,08(0,19)	75,37(-0,26)
14	123,09(1,58)	16,62(-0,67)	15,13(-0,38)	1,38(-0,70)	102,81(0,12)	100,87(0,46)
15	116,40(1,36)	21,16(-0,35)	20,90(-0,15)	1,88(-0,15)	97,03(-0,18)	110,56(0,75)
16	127,43(1,72)	22,47(-0,26)	25,64(0,03)	1,89(-0,13)	102,07(0,08)	83,5(-0,03)
17	107,72(1,08)	22,02(-0,29)	19,64(-0,20)	1,74(-0,30)	94,40(-0,32)	87,25(0,07)
18	114,43(1,30)	26,36(0,01)	26,13(0,05)	1,75(-0,29)	129,44(1,54)	84,75(0,003)
19	94,01(0,63)	30,47(0,29)	29,78(0,20)	1,95(-0,06)	118,04(0,93)	107,87(0,67)
20	106,85(1,05)	28,42(0,15)	29,04(0,17)	1,78(-0,25)	103,55(0,16)	90,12(0,15)
21	93,21(0,60)	31,30(0,35)	30,96(0,24)	2,00(-0,02)	92,49(-0,42)	80,43(-0,12)
22	47,27(-0,89)	15,57(-0,74)	5,84(-0,75)	1,56(-0,49)	98,33(-0,11)	60,5(-0,69)
23	46,25(-0,92)	16,05(-0,71)	6,41(-0,72)	1,79(-0,24)	91,04(-0,50)	55,37(-0,84)
24	50,49(-0,78)	16,46(-0,68)	6,79(-0,71)	1,82(-0,21)	105,24(0,25)	80,5(-0,11)
25	47,17(-0,89)	16,70(-0,66)	7,26(-0,69)	1,79(-0,25)	101,16(0,03)	87,25(0,07)
26	79,42(0,15)	31,16(0,34)	23,63(-0,04)	2,10(0,09)	97,15(-0,17)	85,00(0,01)
27	61,58(-0,42)	38,33(0,84)	30,87(0,24)	2,15(0,14)	43,98(-3,01)	102,37(0,51)

28	69,71(-0,15)	39,94(0,96)	34,70(0,39)	2,26(0,26)	88,99(-0,61)	138,87(1,56)
29	64,63(-0,32)	41,73(1,08)	36,54(0,47)	2,53(0,56)	92,99(-0,39)	150,00(1,89)

4.1 Análise dos nove genótipos

Os nove genótipos envolvem do tratamento 1 ao 9 da Tabela 12 apresentada anteriormente. Por meio das estimativas das distâncias generalizadas de Mahalanobis (D_1^2), os genótipos que apresentaram menor dissimilaridade foram os genótipos 1 e 4, com $D_1^2 = 3{,}627$. Por outro lado, os que apresentaram a maior dissimilaridade foram os genótipos 3 e 5, com $D_1^2 = 15{,}788$.

4.1.1 Agrupamento com o Vizinho mais próximo e de Ward

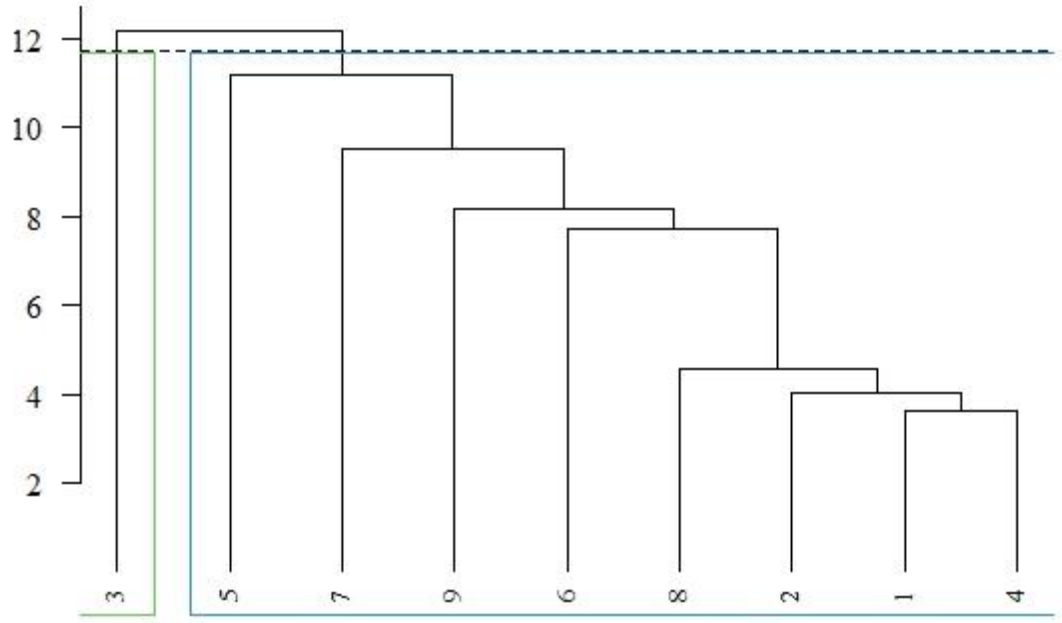
Aplicando o método de agrupamento do vizinho mais próximo entre grupos, baseado na distância citada (D_1^2), os nove genótipos de *C. annuum* foram divididos em dois grupos, onde o grupo I reuniu 1 genótipo $(11{,}11\%)$ e o grupo II reuniu 8 genótipos $(88{,}89\%)$, conforme a Tabela 13 abaixo:

Tabela 13 - Formação dos grupos de 9 genótipos de *Capsicum annuum L.,* por meio do método do Vizinho mais próximo, baseado na distância generalizada de Mahalanobis.

Grupos	**Genótipos**
I	3
II	1,2,4,5,6,7,8,9

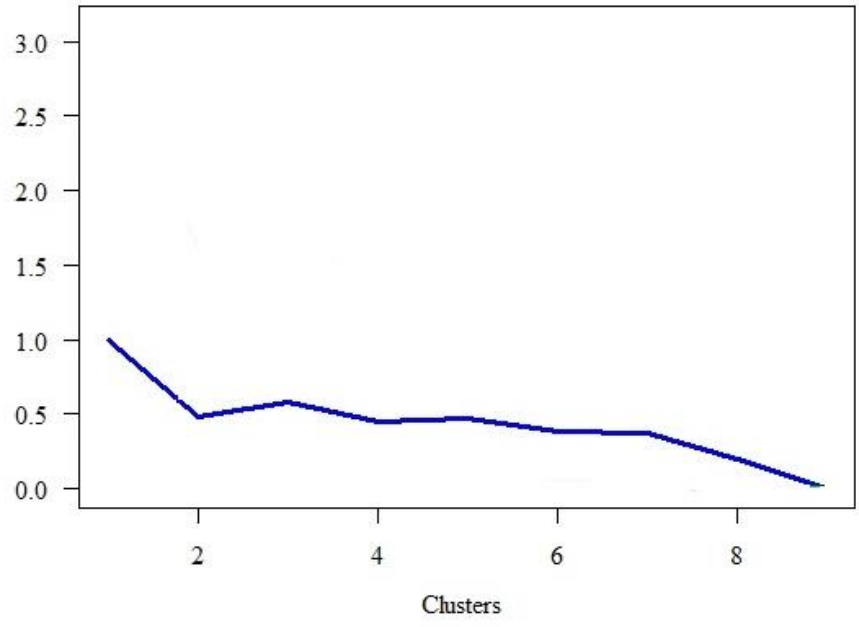
Após, foi feita a definição do número de grupos, baseada no critério de Mojena, que estabeleceu $\theta = 11{,}737$ como ponto de corte. Este valor corresponde a $96{,}23\%$ da distância máxima observada nos níveis de fusão. O RMSSTD, baseado na distância euclidiana, retornou um índice igual a $0{,}5$, que sugere dois como um bom número de grupos. Por meio do dendrograma e do gráfico a seguir, é possível visualizar esses pontos.

Figura 5 - Dendrograma obtido por meio do método do Vizinho mais próximo, com a separação dos grupos e delimitado pelo ponto de corte (θ)para os nove genótipos.



Fonte: Autor, 2021.

Figura 6 - Trajetória do índice RMSSTD para o método do vizinho mais próximo da análise dos genótipos.



Fonte: Autor, 2021.

O coeficiente de correlação cofenética deste método foi de 0,69, indicando um bom ajuste entre a matriz de dissimilaridade e a matriz cofenética.

Para o estudo de divergência genética, percebe-se que o genótipo 3, a Pimenta Peter, diferenciou-se dos demais e destacou-se como um potencial genótipo para realizar combinações híbridas de maior efeito heterótico e possuírem um desempenho superior à de seus genitores. Além disso, observa-se no grupo I, que a variável Vitamina apresentou-se com a maior média e Número de sementes como a menor. Já para o grupo II, constitui-se de genótipos que apresentaram-se com médias bem similares, referentes às características de porte, como Comprimento, Largura e Espessura da polpa.

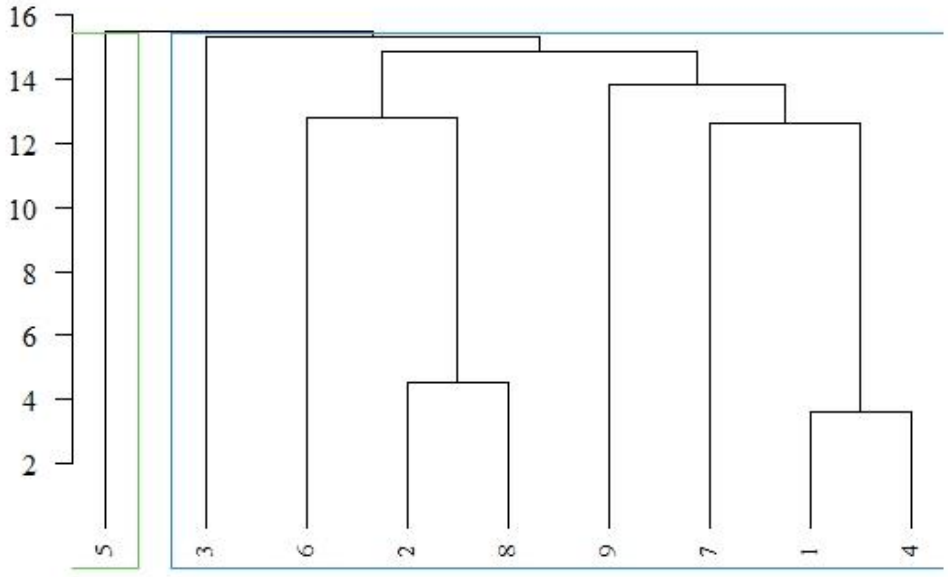
Aplicando o método de agrupamento de Ward entre grupos, baseado na distância citada (D_1^2), os genótipos de *C. annuum* foram divididos em dois grupos, dos quais o grupo I reuniu 1 genótipo $(11{,}11\%)$ e o grupo II reuniu 8 genótipos $(88{,}89\%)$, conforme a Tabela 14 abaixo:

Tabela 14 - Formação dos grupos de 9 genótipos de *Capsicum annuum L.,* por meio do método de Ward, baseado na distância generalizada de Mahalanobis.

Grupos	**Genótipos**
I	5
II	1,2,3,4,6,7,8,9

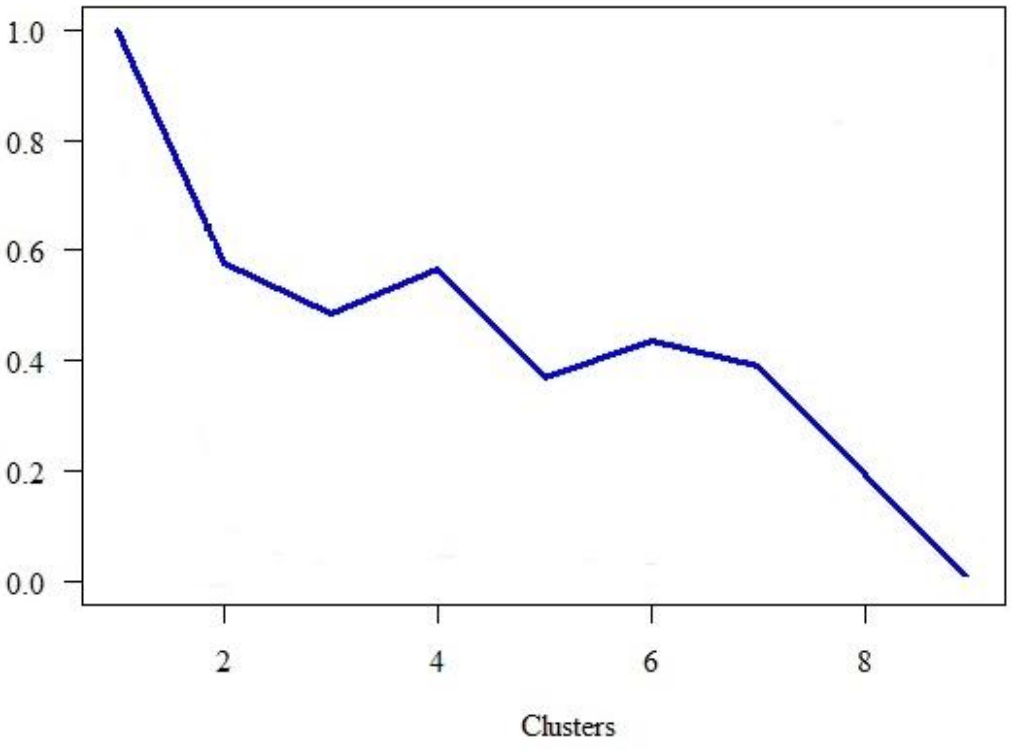
Após, foi feita a definição do número de grupos, segundo o critério de Mojena, que estabeleceu $\theta = 17{,}58$ como ponto de corte. Este valor supera a distância máxima observada nos níveis de fusão, o que se opõe ao critério global de Mojena e neste caso, este método mostrou-se inconsistente. O RMSSTD, baseado na distância euclidiana, retornou um índice entre $0{,}4$ e $0{,}6$, que sugere 2 como um número ótimo de grupos. Por meio do dendrograma e do gráfico a seguir, é possível visualizar esses pontos.

Figura 7 - Dendrograma obtido por meio do método de Ward, com a separação dos grupos para os nove genótipos.



Fonte: Autor, 2021.

Figura 8 - Trajetória do índice RMSSTD para o método de Ward da análise dos genótipos.



Fonte: Autor, 2021.

O coeficiente de correlação cofenética deste método foi de $0{,}636$, o que segundo Rohlf (1970), coeficientes menores que $0{,}7$ refletem a inadequação do método de agrupamento para resumir a informação do conjunto de dados. Porém, para Vaz Patto et al. (2004), o $r_{cof} \geq 0{,}56$ é considerado adequado, justamente por retratar menor distorção provocada pelo agrupamento. Desse modo, considera-se este coeficiente como relativamente bom.

É importante ressaltar que o grupo I foi formado por um único genótipo para ambos os métodos hierárquicos, o que provavelmente ocorreu pelo fato desses genótipos contribuírem positivamente para a diversidade genética. Neste caso, acredita-se que o genótipo 5, a Pimenta Jamaica Yellow, tenha apresentado uma variação interna muito alta e possivelmente, cruzamentos realizados com ele tendem a promover ganhos genéticos satisfatórios. Observa-se também que no grupo I, que a variável Largura apresentou-se com a maior média e Comprimento como a menor. Por outro lado, o grupo II apresentou Comprimento com a maior média e Largura com a menor.

Logo, comparando a solução destes agrupamentos, considera-se que o método do Vizinho mais próximo possuiu uma melhor consistência de agrupamento do que o de Ward, justamente por retornar um valor maior de coeficiente de correlação cofenética.

4.1.2 Agrupamento pelo método de Tocher e de k −médias

Por meio do método de Tocher (original) baseado na distância generalizada de Mahalanobis, os genótipos de *C. annuum* foram divididos em cinco grupos, onde o grupo I reuniu 5 genótipos $(55{,}56\%)$, o grupo II, III, IV e V reuniram 1 genótipo $(11{,}11\%)$ cada, conforme apresentado na Tabela 15 a seguir.

Tabela 15 - Formação dos grupos de 9 genótipos de *Capsicum annuum L.,* por meio do método de Tocher, baseado na distância generalizada de Mahalanobis.

Grupos	**Genótipos**
I	1,2,4,7,8
II	3
III	5
IV	6
V	9

Pode-se perceber que, por este método, os grupos II, III, IV e V reuniram exatamente um genótipo cada, o que mostra uma relevante variabilidade genética entre os genótipos de sua formação. Destaque, mais uma vez, para os genótipos 3 e 5 citados anteriormente. No grupo I, encontrou-se valores médios semelhantes nas variáveis Comprimento e Largura. Já para os grupos II, III, IV e V, que constituíram de grupos com um genótipo cada, apresentaram as maiores médias nas variáveis Vitamina, Largura, Comprimento e Número de sementes, respectivamente.

Resultados como estes, para o método de Tocher, foram encontrados por Vasconcelos *et al.* (2007) ao realizar simulações de coleções de acessos com diferentes características e em Puiatti *et al.* (2014) em estudo de divergência genética em acessos de alho.

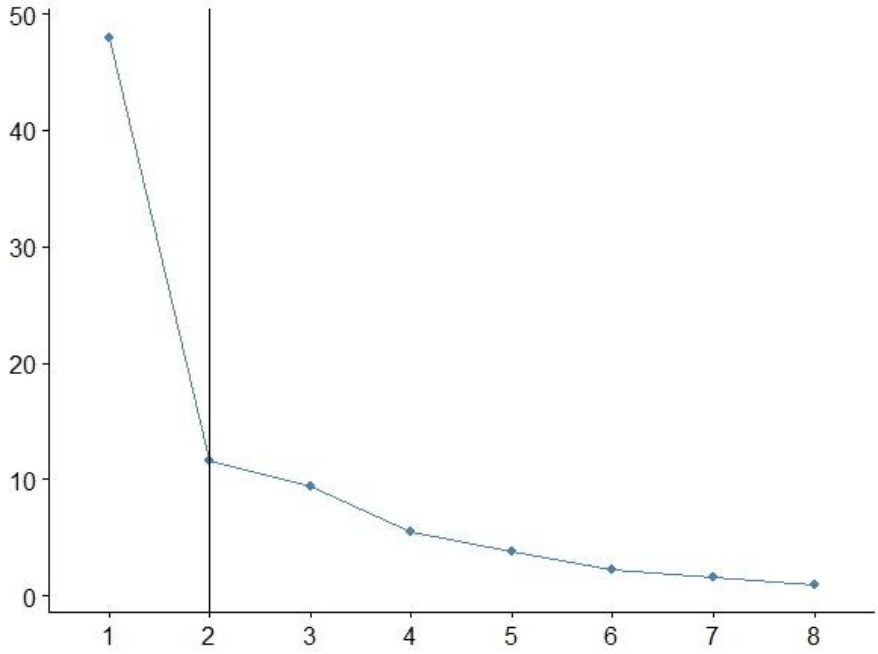
Para o método das $k-$médias, os genótipos foram divididos em dois grupos, onde o grupo I reuniu 5 genótipos $(55{,}56\%)$e o grupo II reuniu 4 genótipos $(44{,}44\%)$, conforme apresentado na Tabela 16 abaixo:

Tabela 16 - Formação dos grupos de 9 genótipos de *Capsicum annuum L.,* por meio do método de $k-$médias.

Grupos	**Genótipos**
I	1,2,3,4,5
II	6,7,8,9

O "método do cotovelo" foi utilizado e neste caso, o indicado seria dois grupos, como apresentado na Figura 9 abaixo. Um exemplo semelhante é apresentado em Matte (2020), quando apresenta um exemplo didático baseado em 40 instâncias, em que foram escolhidos aleatoriamente 3 grupos.

Figura 9 - Número ótimo de clusters para os 9 genótipos.



Fonte: Autor, 2021.

Pelo método de Tocher, os genótipos 3 e 5 ficaram em grupos separados e pelo das $k-$médias, ficaram juntos e com outros genótipos. Acredita-se que isso tenha ocorrido, pois, em alguma característica os genótipos tiveram um desempenho semelhante com os demais de seu grupo, levando a acreditar que esses genótipos tenham tido um desempenho superior ao do grupo II. Observa-se no grupo I, que a variável Vitamina apresentou-se com a maior média e Peso total dos frutos como a menor. Já o grupo II, apresentou-se com genótipos com dados médios semelhantes para o Peso total dos frutos, Espessura da polpa e Número de sementes.

4.2 Análise dos cruzamentos

Os cruzamentos envolvem do tratamento 10 ao 29 da Tabela 12 apresentada anteriormente. Por meio das estimativas das distâncias generalizadas de Mahalanobis (D_1^2), os genótipos que apresentaram menor dissimilaridade foram o 24 e 25, com $D_2^2 = 0{,}941$. Por

outro lado, os que apresentaram a maior dissimilaridade foram os genótipos 18 e 27, com $D_1^2 = 29{,}395$.

4.2.1 Agrupamento com o Vizinho mais próximo e de Ward

Aplicando o método de agrupamento do vizinho mais próximo entre grupos, baseado na distância citada (D_1^2), os cruzamentos de *C. annuum* foram divididos em dois grupos, onde o grupo I reuniu 1 genótipo (5%) e o grupo II reuniu 19 genótipos (95%), conforme a Tabela 17 abaixo:

Tabela 17 - Formação dos grupos de 20 genótipos de *Capsicum annuum L.,* por meio do método do vizinho mais próximo, baseado na distância generalizada de Mahalanobis.

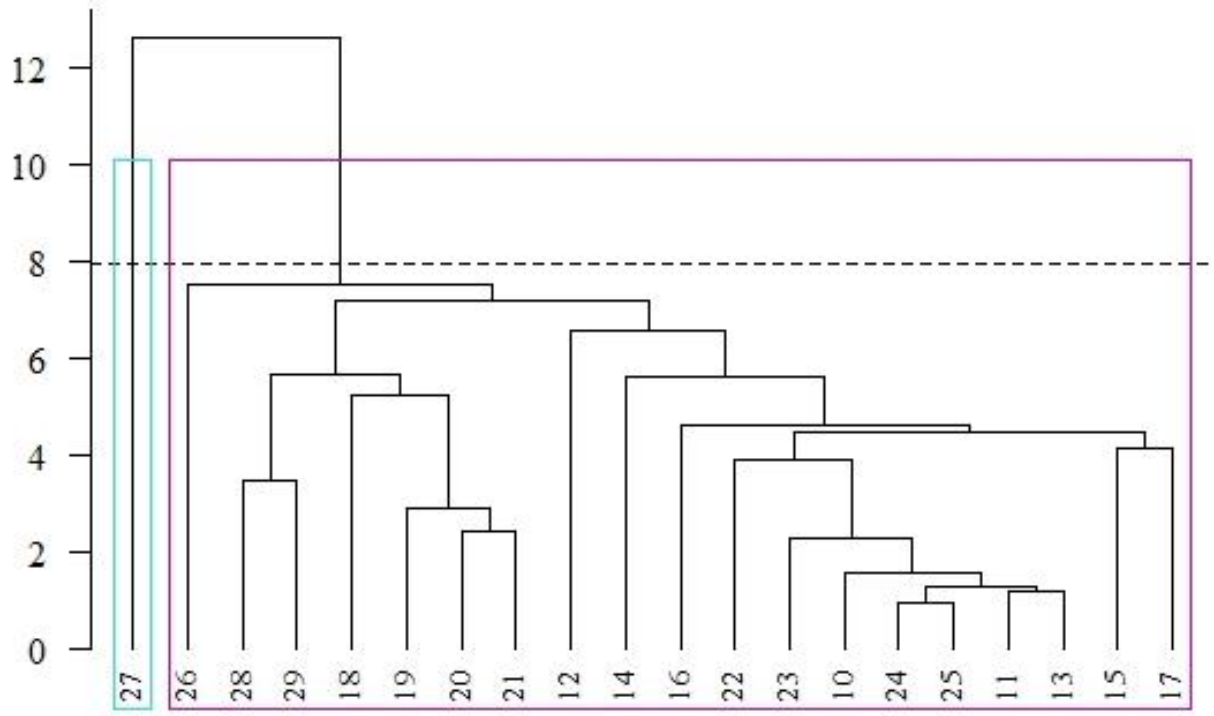
Grupos	**Genótipos**
I	27
II	10,11,12,13,14,15,16,17,18,19,20,21,22,23,24,25,26,28,29

Ressalta-se que, assim como a aplicação do método do Vizinho mais próximo dos genótipos, a aplicação do mesmo para os cruzamentos possuiu número igual de grupos. Importante visualizar também que o genótipo 27 do grupo I, cruzamento entre Pimenta Jamaica Yellow e Pimentão Quadrado, teve um desempenho superior aos demais. Interessante é que um de seus genitores é o genótipo 5, que destacou-se com um desempenho superior e que consequentemente, cruzamentos feitos à partir dele também o seriam. O grupo II reuniu a maior porcentagem de genótipos, sendo o grupo mais similar e que possuem características bem semelhantes.

Tem-se no grupo I, que a variável Largura apresentou-se com a maior média e Vitamina como a menor. Já o grupo II, apresentou genótipos com dados médios semelhantes para o Peso total dos frutos e Espessura da polpa.

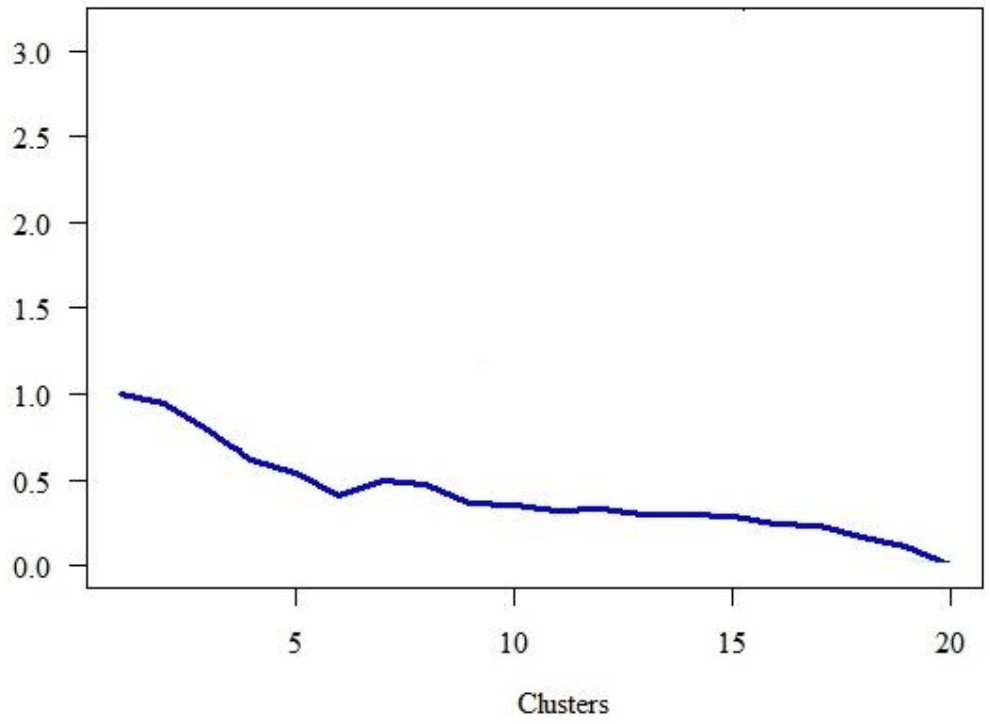
Baseado no critério de Mojena, foi feita a definição do número de grupos, que estabeleceu $\theta = 7{,}949$ como ponto de corte. Este valor corresponde a $62{,}92\%$ da distância máxima observada nos níveis de fusão. O RMSSTD, baseado na distância euclidiana, retornou um índice igual a $0{,}4$, que sugere entre cinco e seis como um bom número de grupos. Por meio do dendrograma e do gráfico a seguir, é possível visualizar esses pontos.

Figura 10 - Dendrograma obtido por meio do método do Vizinho mais próximo, com a separação dos grupos e delimitado pelo ponto de corte (θ) para os cruzamentos.



Fonte: Autor, 2021.

Figura 11 - Trajetória do índice RMSSTD para o método do Vizinho mais próximo da análise dos cruzamentos.



Fonte: Autor, 2021.

O coeficiente de correlação cofenética deste método foi de 0,65, valor inferior ao método do Vizinho mais próximo aplicado aos nove genótipos, mas que possui indicativo de uma boa consistência do agrupamento.

Por meio da aplicação do método de agrupamento de Ward entre grupos, os genótipos de *C. annuum* foram divididos em quatro grupos, dos quais o grupo *I* reuniu 4 genótipos(20%), o grupo *II* reuniu 6 genótipos(30%), o *III* reuniu 1 genótipo(5%) e o *IV* reuniu 9 genótipos (45%), conforme a Tabela 18 abaixo:

Tabela 18 - Formação dos grupos de 20 genótipos de *Capsicum annuum L.,* por meio do método de Ward, baseado na distância generalizada de Mahalanobis.

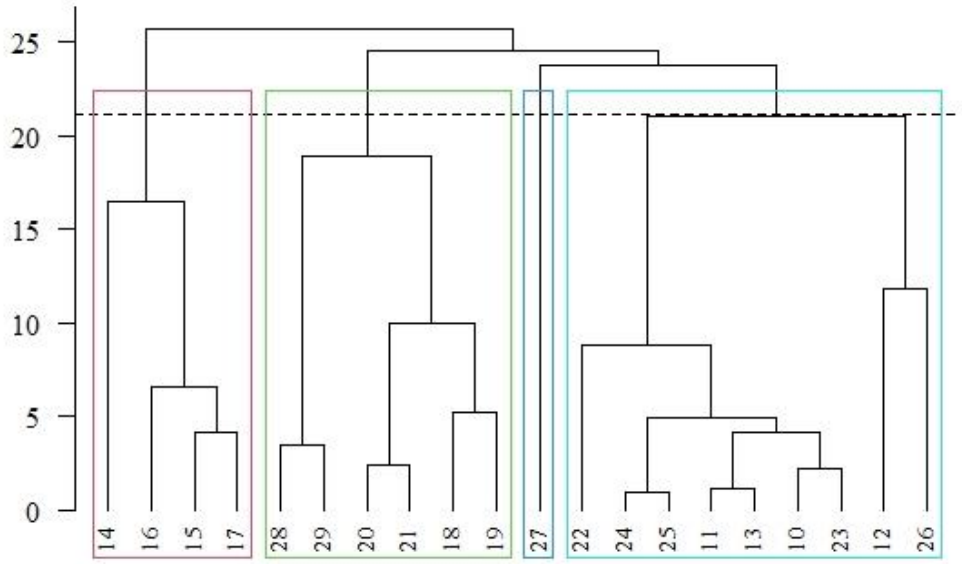
Grupos	**Genótipos**
I	14,15,16,17
II	18,19,20,21,28,29
III	27
IV	10,11,12,13,22,23,24,25,26

Para este método, os 20 genótipos foram divididos em grupos com poucos elementos, mostrando, assim, a tendência do método. Ressalta-se também o genótipo 27 que, assim como no método do Vizinho mais próximo, ficou isolado em um grupo, confirmando que possuiu uma variação interna alta e consequentemente, um desempenho superior. Um resultado semelhante foi visto em Silva *et al.* (2020), onde foram avaliados 11 genótipos de *Capsicum annuum* L. em estudo de divergência genética.

Observa-se no grupo I, que a variável Comprimento apresentou a maior média e Espessura da polpa como a menor e no grupo II foram encontrados valores médios similares para Largura e Peso total dos frutos. Já o grupo III apresentou valores bem próximos para Peso total dos frutos e Espessura da polpa. E por último, o grupo IV apresentou-se com valores médios semelhantes para quase todas as variáveis, exceto para Vitamina.

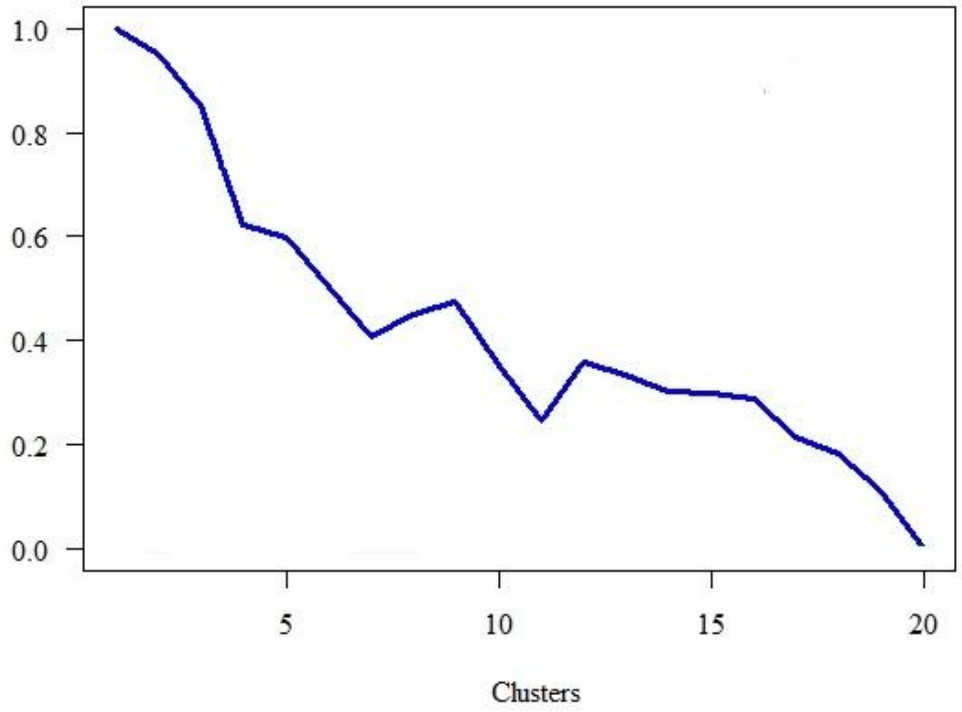
Para a definição do número de grupos, por meio do método de Mojena, obteve-se $\theta = 21{,}15$ como ponto de corte. Este valor corresponde a $81{,}10\%$ da distância máxima observada nos níveis de fusão. O RMSSTD, baseado na distância euclidiana, retornou um índice de aproximadamente, $0{,}61$, que sugere quatro como um bom número de grupos. Por meio do dendrograma e do gráfico a seguir, é possível visualizar esses pontos.

Figura 12 - Dendrograma obtido por meio do método de Ward, com a separação dos grupos e delimitado pelo ponto de corte (θ) para os cruzamentos.



Fonte: Autor, 2021

Figura 13 - Trajetória do índice RMSSTD para o método de Ward da análise dos cruzamentos.



Fonte: Autor, 2021.

Além disso, o que chama atenção, neste caso, é o valor do coeficiente de correlação cofenética que retornou um valor igual a 0,579. Mesmo refletindo uma adequação deste método de agrupamento, é um valor muito baixo e não há uma representação tão boa das distâncias do dendrograma.

4.2.2 Agrupamento pelo método de Tocher e de $k-$médias

Por meio do método de Tocher (original), os genótipos de *C. annuum* foram divididos em cinco grupos, onde o grupo I reuniu 16 genótipos (80%) e os grupos II, III, IV e V reuniram 1 genótipo (5%)cada, conforme apresentado na Tabela 19 a seguir.

Tabela 19 - Formação dos grupos de 20 genótipos de *Capsicum annuum L.,* por meio do método de Tocher, baseado na distância generalizada de Mahalanobis.

Grupos	Genótipos
I	10,11,12,13,15,16,17,18,19,20,21,22,23,24,25,28
II	14
III	26
IV	27
V	29

Pela tabela acima, tem-se que apenas o grupo I possuiu o maior número de genótipos e os grupos restantes com apenas um. Destaque para os grupos III, IV e V, que são cruzamentos em que um dos genitores é o genótipo 5, revelando, mais uma vez, a importância e o desempenho do mesmo. No grupo I, apresentou-se Vitamina com a maior média e Largura como a menor.Já para os grupos II, III, IV e V, grupos compostos por um genótipo cada, apresentaram as menores médias nas variáveis Espessura da polpa, Vitamina, Vitamina e Comprimento, respectivamente.

Trabalhos anteriores apresentaram resultados semelhantes a este, como em Silva (2012), onde foi avaliada a divergência genética de 89 acessos de alho e obteve-se quinze grupos, no qual do grupo IX ao XV possuíram o menor número de genótipos também.

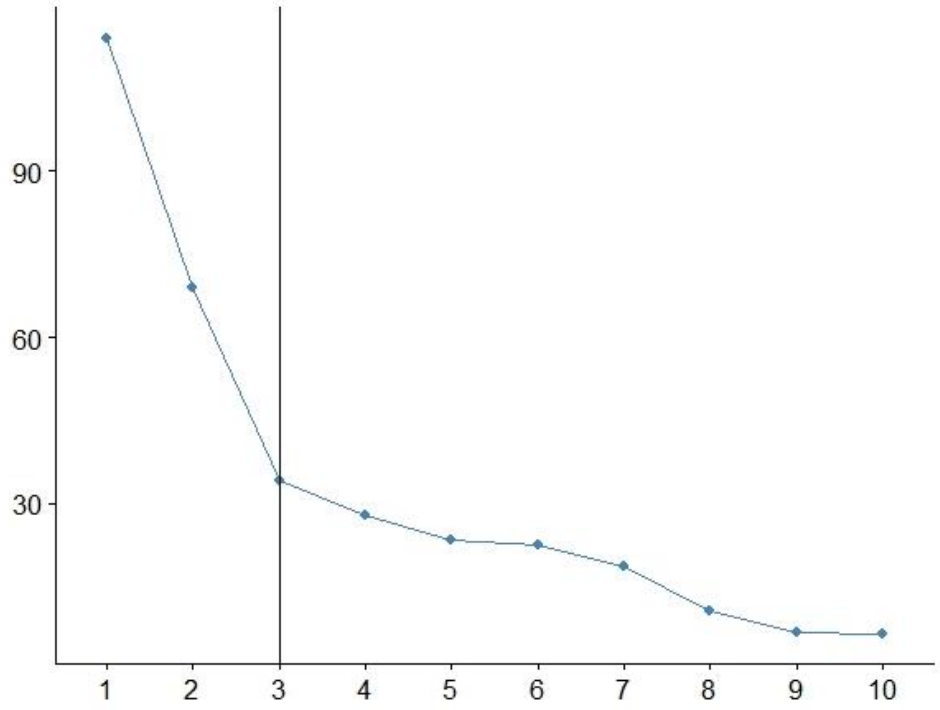
Pelo método das $k-$médias, os genótipos foram divididos em três grupos, onde o grupo I reuniu 9 genótipos(45%), o grupo II reuniu 3 genótipos (15%), e o grupo III reuniu 8 genótipos (40%), conforme apresentado na Tabela 20 abaixo:

Tabela 20 - Formação dos grupos de 20 genótipos de *Capsicum annuum L.*, por meio do método de $k-$médias.

Grupos	Genótipos
I	14,15,16,17,18,19,20,21,26
II	27,28,29
III	10,11,12,13,22,23,24,25

É importante ressaltar como os agrupamentos são bem definidos por este método e como não possui a tendência de agrupamentos com apenas 1 genótipo. O "método do cotovelo" foi utilizado e neste caso, o indicado seriam três grupos, como apresentado na Figura 14 abaixo.

Figura 14 - Número ótimo de clusters para os 20 genótipos.



Fonte: Autor, 2021.

Destaca-se o grupo II que tem como composição os genótipos 27,28 e 29 que, assim, como no método de Tocher, apresentou-se de extrema importância para o estudo de divergência genética pelo seu desempenho superior. O grupo I ressalta a variável Comprimento com a maior média e Espessura como a menor. No grupo II, possuem genótipos com características semelhantes em Largura e Espessura da polpa e o grupo III, genótipos semelhantes nas variáveis Comprimento e Largura.

4.3 Análise de todos os dados

Os dados desta análise contêm os genótipos e os cruzamentos juntos. Por meio das estimativas das distâncias generalizadas de Mahalanobis (D_1^2), os genótipos que apresentaram menor dissimilaridade foram o 24 e 25, com $D_2^2 = 0{,}296$, os mesmos genótipos da análise de cruzamentos. Por outro lado, os que apresentaram a maior dissimilaridade foram os genótipos 3 e 27, com $D_1^2 = 39{,}574$, os mesmos genótipos que estão entre a maior dissimilaridade da análise de genótipos e da análise de cruzamentos.

4.3.1 Agrupamento com o Vizinho mais próximo e de Ward

Aplicando o método de agrupamento do vizinho mais próximo entre grupos, os 29 genótipos de *C. annuum* foram divididos em cinco grupos, onde os grupos I, II, III e IV reuniram 1 genótipo $(3{,}45\%)$ cada e o grupo V reuniu 25 genótipos $(86{,}20\%)$, conforme a Tabela 21 abaixo:

Tabela 21 - Formação dos grupos de 29 genótipos de *Capsicum annuum L.,* por meio do método do Vizinho mais próximo, baseado na distância generalizada de Mahalanobis.

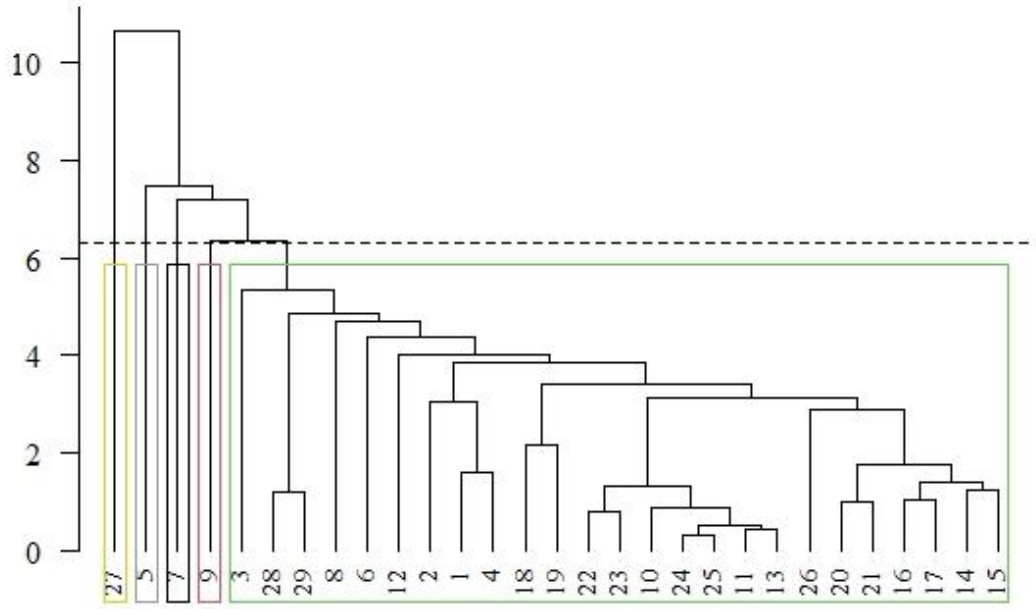
Grupos	**Genótipos**
I	27
II	5
III	7
IV	9
V	1, 2, 3, 4, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 28, 29

Destaca-se, novamente, o genótipo 27 que ficou isolado em um grupo único, tanto na aplicação do método do Vizinho mais próximo, Ward e Tocher para os cruzamentos, o que ocorreu por ter um desempenho superior e por ter como um dos genitores o genótipo 5, que também destacou-se por estar em um grupo isolado. Outro fator observado, é com relação a presença de um único grupo que possui a maior porcentagem de genótipos, o que pode ter ocorrido pela similaridade entre suas características.

Nos grupos I, II, III e IV apresentaram as maiores médias nas variáveis Largura, Largura, Peso total dos frutos e Espessura da polpa, respectivamente. Já para o grupo V, encontrou-se valores médios semelhantes nas variáveis Largura, Peso total dos frutos e Espessura da polpa.

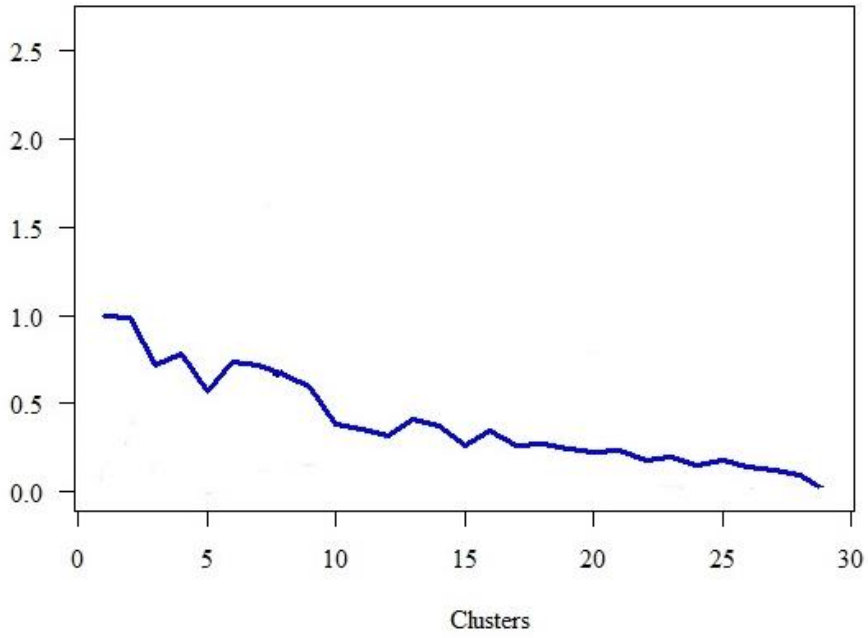
Pelo critério de Mojena, foi feita a definição do número de grupos, que estabeleceu $\theta = 6{,}298$ como ponto de corte. Este valor corresponde a $59{,}05\%$ da distância máxima observada nos níveis de fusão. O RMSSTD, baseado na distância euclidiana, retornou um índice de aproximadamente, $0{,}6$, que sugere entre três e cinco como um bom número de grupos. Por meio do dendrograma e do gráfico a seguir, é possível visualizar esses pontos.

Figura 15 - Dendrograma obtido por meio do método do Vizinho mais próximo, com a separação dos grupos e delimitado pelo ponto de corte (θ) todos os dados.



Fonte: Autor, 2021

Figura 16 - Trajetória do índice RMSSTD para o método do vizinho mais próximo da análise para todos os dados.



Fonte: Autor, 2021.

Pelo cálculo do coeficiente de correlação cofenética, obteve-se um valor igual a 0,70, que possui um indicativo de boa consistência do agrupamento.

Pelo método de agrupamento de Ward entre grupos, os genótipos de *C. annuum* foram divididos em cinco grupos, dos quais o grupo I e III reuniram 3 genótipos $(10{,}34\%)$ cada, o grupo II reuniu 2 genótipos$(6{,}9\%)$, o IV reuniu 8 genótipos $(27{,}59\%)$ e o V reuniu 13 genótipos $(44{,}83\%)$, conforme a Tabela 22 abaixo:

Tabela 22 - Formação dos grupos de 29 genótipos de *Capsicum annuum L.,* por meio do método de Ward, baseado na distância generalizada de Mahalanobis.

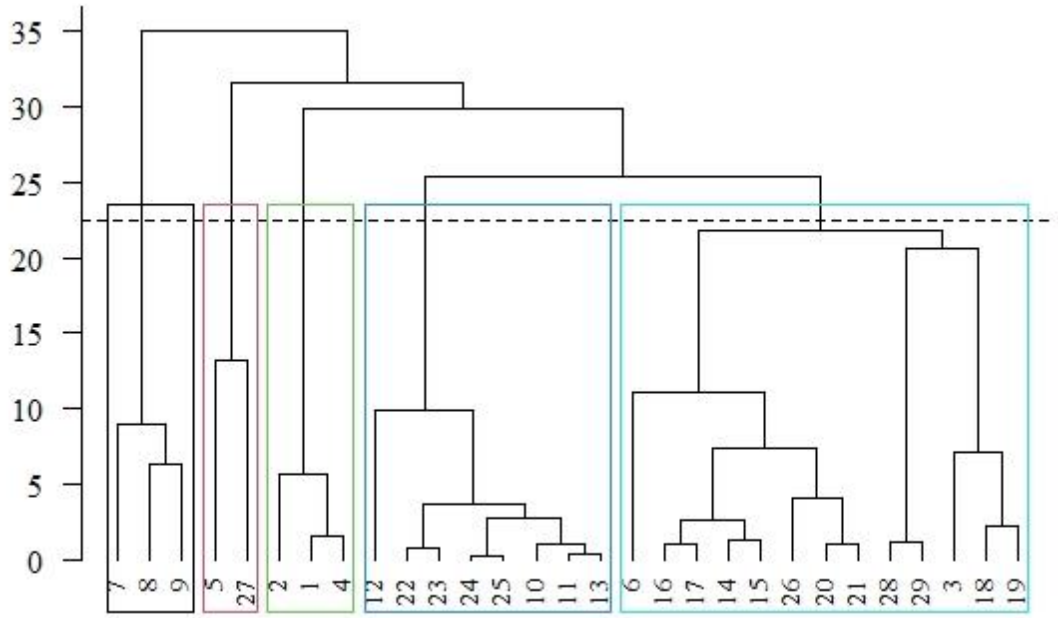
Grupos	**Genótipos**
I	7,8,9
II	5,27
III	1,2,4
IV	10,11,12,13,22,23,24,25
V	3,6,14,15,16,17,18,19,20,21,26,28,29

Na análise dos genótipos, o genótipo 5 apresenta-se isolado em único grupo e na análise dos cruzamentos, o genótipo 27 também. Um fato interessante é que, na análise de todos os dados, ambos aparecem juntos e acredita-se que deve-se ao fato de uma variação interna alta e semelhante, pois o genótipo 27 é cruzamento do genótipo 5, como dito anteriormente.

No grupo I, apresentaram médias similares as variáveis Peso total dos frutos e Espessura da polpa que, por sinal, são as maiores médias. No grupo II, essas mesmas variáveis destacam-se por estarem entre as menores médias, juntamente com a variável Comprimento. Já no grupo III, tem-se a variável Vitamina com a maior média e Largura com a menor. O grupo IV, destaca-se, novamente, nas características de porte, como Comprimento, Largura e Peso total dos frutos. E por fim, o grupo V, ressalta a característica Espessura com a menor média.

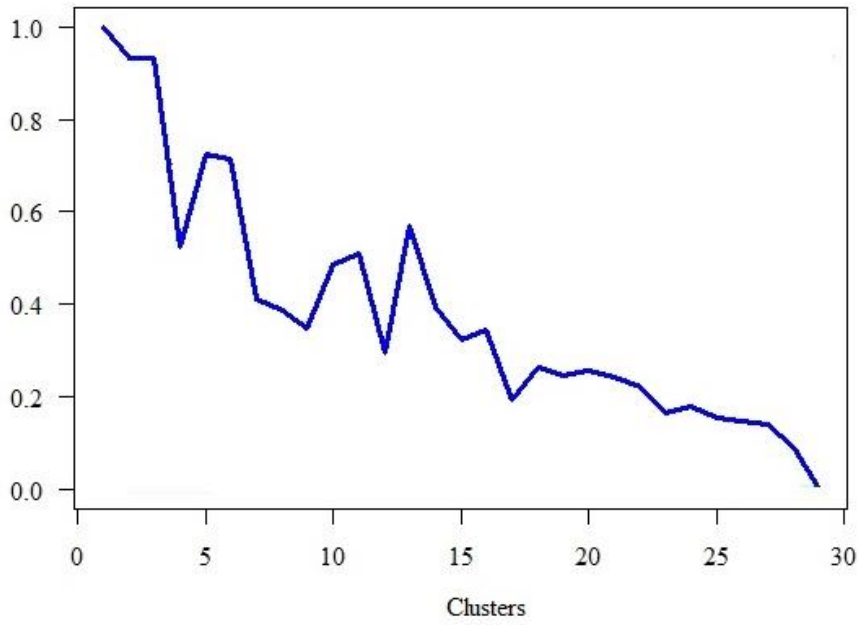
Pelo critério de Mojena, foi feita a definição do número de grupos, que estabeleceu $\theta = 22{,}40$ como ponto de corte. Este valor corresponde a $63{,}87\%$ da distância máxima observada nos níveis de fusão. O RMSSTD, baseado na distância euclidiana, retornou um índice entre $0{,}4$ e $0{,}6$ que sugere quatro como um bom número de grupos. Por meio do dendrograma e do gráfico a seguir, é possível visualizar esses pontos.

Figura 17 - Dendrograma obtido por meio do método de Ward, com a separação dos grupos e delimitado pelo ponto de corte (θ) todos os dados.



Fonte: Autor, 2021.

Figura 18 - Trajetória do índice RMSSTD para o método de Ward da análise para todos os dados.



Fonte: Autor, 2021.

Pelo cálculo do coeficiente de correlação cofenética, obteve-se um valor igual a 0,63, que possui um indicativo de boa representação das distâncias do dendrograma.

4.3.2 Agrupamento pelo método de Tocher e de k −médias

Por meio do método de Tocher (original), os genótipos de *Capsicum annuum L.* foram divididos em sete grupos, onde o grupo I reuniu 20 genótipos $(68{,}96\%)$, os grupos II, III e IV reuniram 2 genótipos $(6{,}89\%)$ cada, o V, VI e VII reuniram 1 genótipo $(3{,}45\%)$ cada, conforme apresentado na Tabela 23 a seguir.

Tabela 23 - Formação dos grupos de 29 genótipos de *Capsicum annuum L.,* por meio do método de Tocher, baseado na distância generalizada de Mahalanobis.

Grupos	**Genótipos**
I	2,6,8,10,11,12,13,14,15,16,17,18,19,20,21,11,23,24,25,26
II	28,29
III	1,4
IV	7,9
V	3
VI	5
VII	27

Destaca-se, nessa aplicação, o número de genótipos no grupo I, onde quase 70% dos genótipos estão reunidos nele. Resultados semelhantes são apresentados em Oliveira (2015), no qual avaliou a divergência genética de 25 genótipos de sorgo sacarino e por meio da aplicação do método de Tocher obteve oito grupos, sendo o grupo I com o maior número de genótipos. O mesmo ocorreu em Santos *et al.* (2012), no qual analisou a diversidade genética de 60 genótipos de babaçu, por meio da distância euclidiana padronizada e no qual o grupo I apresentou mais de 90% dos acessos reunidos.

No grupo I, apresentou-se Comprimento com a maior média e Largura como a menor. Já para os grupos II, III e IV destacaram-se variáveis Número de sementes, Vitamina e Espessura da polpa como as menores médias, respectivamente e os grupos V, VI e VII, as variáveis Espessura da polpa, Comprimento e Vitamina como as menores médias, respectivamente.

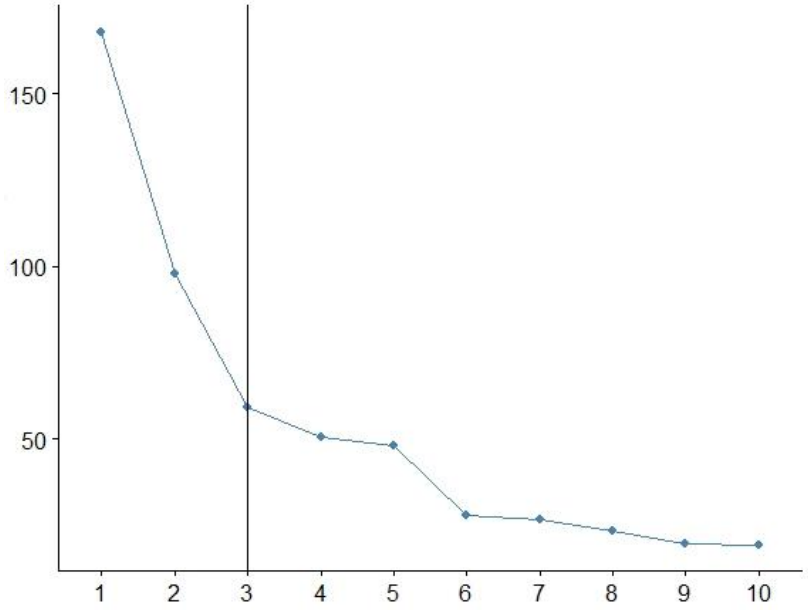
Por meio do método das $k-$médias, os genótipos foram divididos em três grupos, onde o grupo I reuniu 3 genótipos $(10{,}34\%)$, o grupo II e III reuniram 13 genótipos $(44{,}83\%)$ cada, conforme apresentado na Tabela 24 abaixo:

Tabela 24 - Formação dos grupos de 29 genótipos de *Capsicum annuum L.,* por meio do método de $k-$médias.

Grupos	**Genótipos**
I	7,8,9
II	6,14,15,16,17,18,19,20,21,26,27,28,29
III	1,2,3,4,5,10,11,12,13,22,23,24,25

Ressalta-se, nesta situação, que o grupo II possui os genótipos do grupo I e II da análise dos cruzamentos reunidos e o grupo III, a maior parte dos genótipos do grupo III dos cruzamentos. Então, percebe-se que, mesmo que possua algumas mudanças, a maioria dos genótipos se mantém unidos e desse modo, reafirmam a similaridade entre suas características. O "método do cotovelo" utilizado indicou que três grupos seria ideal, como apresentado na Figura 19 a seguir.

Figura 19 - Número ótimo de clusters para os 29 genótipos.



Fonte: Autor, 2021.

No grupo *I* foram encontrados valores médios similares para Peso total dos frutos e Espessura. Já o grupo *II* apresentou valores bem próximos para Largura e Peso total dos frutos. O grupo *III*, por sua vez, destacou valores médios similares para as variáveis Comprimento, Largura, Peso total dos frutos, Espessura e Número de sementes.

5 CONCLUSÕES

Para a análise dos nove genótipos, os métodos hierárquicos, Vizinho mais próximo e Ward, apresentaram resultados bem parecidos, no qual a diferença está no genótipo 3, que se encontra no grupo I do método do Vizinho mais próximo, enquanto que no método de Ward, o grupo I é formado apenas pelo genótipo 5; também obtiveram coeficientes de correlação cofenética (r_{cof}) distintos, onde o primeiro método possui uma adequação do método de agrupamento melhor do que o segundo. Os métodos não hierárquicos, Tocher e k-médias, formaram um maior número de grupos, 5 e 3, respectivamente. Adotando o método de Tocher, o grupo I reuniu o maior número de genótipos $(55{,}56\%)$, e o método $k-$médias reuniu os genótipos nos seguintes grupos: I $(1,2,3,4,5)$ e II $(6,7,8,9)$.

Para a análise dos cruzamentos (genótipos 10 a 29), os métodos hierárquicos apresentaram 2 e 4 grupos, respectivamente, e o método do Vizinho mais próximo apresentou um coeficiente de correlação cofenética maior que do método de Ward que, embora pequeno $(0{,}65)$, ainda representa uma boa adequação do método. Diferentemente, o método de Ward ressalta um $r_{cof} = 0{,}579$, que possui uma adequação do método, mesmo não sendo tão representativa. Por outro lado, o método de Tocher apresentou-se com a maior parte dos grupos com apenas 1 genótipo, onde reafirma-se, a tendência do método. O método das $k-$médias, por sua vez, possuiu um quantitativo de genótipos bem próximos uns dos outros nos grupos.

Por fim, para a análise de todos os dados, o grupo V do Vizinho mais próximo apresentou $86{,}20\%$ dos genótipos, destacando-se por reunir os genótipos mais similares. O método de Ward apresentou-se em grupos com muitos genótipos e em outros, com poucos. Em contrapartida, o método do Vizinho mais próximo apresentou um coeficiente de correlação cofenética superior ao de Ward, revelando, novamente, uma boa adequação do método de agrupamento. O método de Tocher, nesse caso, apresentou sete grupos, onde o grupo I reuniu 20 genótipos, o que pode ser explicado pelo critério de agrupamento do método. Já, pelo método das $k-$médias, foi mantido em mesmo grupo, a maior parte dos genótipos já agrupados na análise dos cruzamentos.

Desse modo, por meio das análises, ficou claro como a inclusão dos nove genótipos nos dados apresenta um coeficiente de correlação cofenética do método do Vizinho mais próximo melhor, o que pode ser explicado pela similaridade entre os genótipos.

Além disso, por meio do estudo de divergência genética ressalta-se que genótipos com um desempenho superior podem gerar cruzamentos com o mesmo ou superior desempenho. Um destaque em mais de um método de agrupamento, foram os genótipos 5 e 27, sejam por apresentarem dissimilaridade ou uma variação interna alta. Logo, acredita-se que cruzamentos com Pimenta Jamaica Yellow tendem a contribuir positivamente para o melhoramento genético, conservando essa espécie e possibilitando para um ganho significativo para o gênero *Capsicum annuum* L.

Portanto, as diferenças de tamanhos de grupos obtidos pelos agrupamentos hierárquicos e não-hierárquicos, pode ser associado à não-realocação de genótipos nos processos hierárquicos, gerando agrupamentos que explicam menos variações. Assim, independentemente do método adotado, foi possível identificar os genótipos mais dissimilares, possibilitar o estudo da divergência genética de *C.annuum* e contribuir para próximas pesquisas.

6 REFERÊNCIAS

AAKER, D. A.; KUMAR, V.; DAY, G. S. **Pesquisa de marketing**, São Paulo: Atlas, 2001. 745p.

ALBUQUERQUE, M.A.de. **Estabilidade em análise de agrupamento.** Dissertação (Mestrado em Biometria, Área de concentração: Modelagem e planejamento de experimentação.). Universidade Federal Rural de Pernambuco. Recife, p. 67. 2005.

AMARAL JÚNIOR, A.T. do; THIÉBAUT, J.T. de L. **Análise multivariada na avaliação da diversidade em recursos genéticos vegetais.** Campos dos Goytacazes: UENF, 1999. 55 p.

AQUINO, H.F.de. **Caracterização morfológica, agronômica e divergência genética de acessos de pimenta**. Dissertação (Mestrado em Agronomia, Área de Concentração: Melhoramento Genético de Plantas). Universidade Federal Rural de Pernambuco. Recife, p. 91. 2016.

BENTO, C. dos. S; SUDRE, C.P; RODRIGUES, R; RIVA, E.M; PEREIRA, M.G. Descritores qualitativos e multicategóricos na estimativa da variabilidade fenotípica entre acessos de pimenta. **Scientia Agraria**, Paraná, v. 8, n. 2, p. 149 – 157, 2007. Disponível em: https://www.redalyc.org/pdf/995/99516568006.pdf . Acesso em: 17 ago. 2021.

BUSSAB, W.de. O; MIAZAKI, E. S; ANDRADE, D.F. de. **Introdução à análise de agrupamentos.** São Paulo: ABE, 1990. 93p.

CARGNIN, A; SOUZA, M.A. 2007. Diversidade Genética em Cultivares de Arroz. 1-16. 1 ed. Coleção. Planaltina, **DF: Embrapa Cerrados**. – (Boletim de pesquisa e desenvolvimento/ Embrapa Cerrados, ISSN 1676 – 918x; 196).

CARVALHO. S.I.C.de; BIANCHETTI, L. de. B; BUSTAMANTE, P.G; SILVA, D.B. da. 2003. Catálogo de Germoplasma de Pimentas e Pimentões (*Capsicum* spp.) da Embrapa Hortaliças. 1-49. 1 ed. **Brasília: Embrapa Hortaliças**. Documentos, 49. ISSN: 1415-2312.

CARVALHO, S. I. C. de; BIANCHETTI, L.B de; RIBEIRO, C.S.C da; LOPES, C.A. Pimentas do gênero Capsicum no Brasil. - **Brasília: Embrapa Hortaliças**, 2006. 27p.

CARVALHO. S.I.C.de; BIANCHETTI, L. de.B. Botânica e recursos genéticos. *In*: RIBEIRO, C.S.da.C; LOPES, C.A; CARVALHO, S.I.C. de; HENZ, G.P; REIFSCHNEIDER, F.J.B. **Pimentas *Capsicum***. 1. ed. Brasília, DF: Embrapa Hortaliças, 2008. p. 1 – 202. Disponível em: https://www.infoteca.cnptia.embrapa.br. Acesso em: 03 set. 2021.

CASTANHA, R.C.G; GRÁCIO, M.C.C. Contribuição da análise multivariada para os indicadores de avaliação dos programas de pós-graduação: uma análise na área de Matemática (2007-2009). **Revista da Faculdade de Biblioteconomia e Comunicação da UFRGS**, v.21, n.1, p. 130 - 149, 2015.

CRUZ, C.D; REGAZZI, A.J; CARNEIRO, P.C.S. **Modelos biométricos aplicados ao melhoramento genético.** 4.ed. Viçosa: Editora UFV, 2012.

CRUZ, C.D; CARNEIRO, P.C.S; REGAZZI, A.J. **Modelos biométricos aplicados ao melhoramento genético.** 3.ed. Viçosa: Editora UFV, 2014.

CRUZ, C.D; FERREIRA, F.M; PESSONI, L.A. **Biometria aplicada ao estudo de diversidade genética.** 2.ed. Viçosa: UFV, 2020.

COSTA, M.do.P.S.D; RÊGO, E.R.do.; GUEDES, J.F.da.S; CARVALHO, M.G.de; SILVA, A.R.da; RÊGO, M.M.do. Selection of genotypes with ornamental potential in an F_4 population of ornamental peppers (*Capsicum annuum* L.) based on multivariate analysis. **Comunicata Scientiae Horticultural Journal**, v.12, Bom Jesus, Piauí, p. 1-8, 2021. Disponível em: https://www.comunicatascientiae.com.br/comunicata/article/view/3511. Acesso em: 21 ago. 2021.

DIAS, L. A. S. Análises multidimensionais. In: Alfenas, A.C. (ed). **Eletroforese de isoenzimas e proteínas afins.** Ed. UFV, Viçosa-MG, 1998, p. 405-475.

DINIZ FILHO, J. A. **Métodos filogenéticos comparativos.** Ribeirão Preto: Holos., 2000. 120 p.

EDWARDS, A.W.F.; CAVALLI-SFORZA, L.L. A method for cluster analysis. **Biometrics**, Washington, v.21, p. 362-375, 1965.

EVERITT, B. S. et al. **Cluster analysis**. 5. ed. Reino Unido: John Wiley & Sons, 2011.

FALCONER, D. S. **Introdução à genética quantitativa.** Viçosa, MG: UFV, 1981. 279p.

FARIA, P.N. **Avaliação de métodos para determinação do número ótimo de clusters em estudo de divergência genética entre acessos de pimenta.** Dissertação (Mestrado em Estatística Aplicada e Biometria). Universidade Federal de Viçosa. Viçosa, MG, p. 67. 2009.

FERREIRA, D.F. **Estatística multivariada**. 1 ed. Lavras: Editora UFLA, 2008. 662p.

FERREIRA, D. F. **Estatística multivariada.** 2. ed. Lavras: Editora UFLA, 2011. 675 p.

FONSECA, R.M. **Caracterização morfoagronomica de gerações de *Capsicum annuum* x *Capsicum chinense*.** Tese (Doutorado em Agronomia Tropical). Universidade Federal do Amazonas. Manaus, p. 142. 2016.

GUEDES, Clusterização (k −means). **Kaggle**, 2019. Disponível em: https://www.kaggle.com/code/eriveltonguedes/7-clusteriza-o-k-means-erivelton/notebook. Acesso em: 20 de mar. de 2022.

HAIR JR., J.F.; WILLIAM, B.C; BABIN, B.J; ANDERSON, R.E., TATHAM, R.L. **Análise multivariada de dados.** 6.ed. Porto Alegre: Bookman, 2009.

JUNIOR, J.C.C.; RODRIGUES, V.A.P; SOARES, I.F.G; ALMEIDA, R.de; MAURICIO, L.S; PAULA, F.C; NASCIMENTO, E.S.F; ALMEIDA, R.N. de; NASCIMENTO, L. de. C; ZAMPIERI, F.G; MENINI, L; MOULIN, M.M. Avaliação da diversidade genética de *Capsicum* spp. com base em descritores morfoagronômicos e bromatológicos. **Revista Ifes Ciência**, v.7, n.1, p. 1-11, 2021. Disponível em: https://ojs.ifes.edu.br/index.php/ric/article/view/1143. Acesso em: 21 ago. 2021.

KAUFMANN, L., ROUSSEEUW, P. J., **Finding groups in data:** an introduction to cluster analysis. New York: Jonh Wiley, 1990. 342p.

LANCE, G.N; WILLIAMS, W.T. A general theory of classificatory sorting strategies: 1. Hierarchical systems. **The Computer Journal**, v.9, n.4, p. 373 – 380, 1967.

LUTZ. D.L; FREITAS, S.C.de. Valor nutricional. *In*: RIBEIRO, C.S.da.C; LOPES, C.A; CARVALHO, S.I.C. de; HENZ, G.P; REIFSCHNEIDER, F.J.B. **Pimentas *Capsicum***. 1. ed. Brasília, DF: Embrapa Hortaliças, 2008. p. 1 – 202. Disponível em: https://www.infoteca.cnptia.embrapa.br. Acesso em: 03 set. 2021.

LUZ, O.D.S.L. **Dissimilaridade genética entre matrizes de Corymbia citriodora no município de Dueré-TO.** Dissertação (Mestrado em Produção Vegetal). Universidade Federal do Tocantins. Gurupi - TO, p. 60. 2014.

MACHADO, R.L. **Desenvolvimento de um algoritmo imunológico para agrupamento de dados.** Trabalho de Conclusão de Curso – Curso Ciência da Computação. Universidade de Caxias do Sul. Caxias do Sul, p. 122. 2011.

MAHALANOBIS, P.C. On the generalized distance in statistics. **Proceedings of The National Institute of Sciences of India**, v.12, p.49-55, 1936.

MATOS, R.A.de. **Comparação de metodologias de análise de agrupamentos na presença de variáveis categóricas e contínuas**. Dissertação (Mestrado em Estatística). Universidade Federal de Minas Gerais. Belo Horizonte, p. 156. 2007.

MATTE, M.K. **Impacto do Uso da Desigualdade Triangular para Acelerar o Algoritmo k-Means**. Dissertação (Mestrado em Ciência da Computação). Centro Universitário Campo Limpo Paulista. Campo Limpo Paulista, p. 168. 2020.

MAXIMILIANO, A. S. CORDEIRO, M. T. A. 2008. 45 fls. **Partição de grupos e validação da análise de agrupamento para equipamentos de fiscalização eletrônica de trânsito.** Monografia - Universidade Estadual do Paraná. Paraná. 2008.

MENDES, I. da.S. **Obtenção de híbridos de pimentas (*Capsicum* spp.) a partir de genótipos obtidos no estado do Maranhão.** Trabalho de Conclusão de Curso – Curso de Agronomia. Universidade Federal do Maranhão. Chapadinha, MA, p. 54. 2018.

MINGOTI, S. A. **Análise de dados por meio de métodos de estatística multivariada: uma abordagem aplicada.** Belo Horizonte: Editora UFMG, 2005.

MIRANDA, T.G; JÚNIOR, V.C. de A**. Características físico-química de genótipos de pimenta (*Capsicum chinense e Capsicum annuum)***. Dissertação (Mestrado em Produção Vegetal). Universidade Federal dos Vales do Jequitinhonha e Mucuri. Diamantina, p. 72. 2014.

MOJENA, R. Hierarchical grouping method and stopping rules: an evaluation. **Computer Journal.** v. 20, p. 359-363, 1977.

NASCIMENTO, M. **Análise de agrupamento para dados em painel.** Tese (Doutorado em Estatística e Experimentação Agropecuária). Universidade Federal de Lavras. Lavras, MG, p. 121. 2011.

NASCIMENTO, M.F. **Variabilidade genética, ação do Paclobutrazol e do etileno na arquitetura da planta e na longevidade de *Capsicum* spp. e *Solanum Pseuocapsicum*.** Tese (Doutorado em Genética e Melhoramento). Universidade Federal de Viçosa. Viçosa, MG. p. 71. 2018.

OLIVEIRA, T.C.de. **Divergência genética e correlação entre caracteres de genótipos de sorgo sacarino na região de Cáceres – MT**. Dissertação (Mestrado em Genética e Melhoramento de Plantas). Universidade do Estado de Mato Grosso. Cáceres, p. 90. 2015.

OLIVEIRA, A.C.R. **Análise Biométrica de acessos de *Capsicum chinense* Jacq. com ênfase na diversidade genética**. Dissertação (Mestrado em Estatística Aplicada e Biometria). Universidade Federal de Viçosa. Viçosa, MG, p. 52. 2016.

OLIVEIRA, A.C.R. **Análise Uni e multivariada para avaliação em cruzamentos dialélicos parciais**. Tese (Doutorado em Estatística Aplicada e Biometria). Universidade Federal de Viçosa. Viçosa, MG, p. 92. 2020.

OLIVEIRA, O que é análise de cluster? **Oper Data**, 2019. Disponível em: https://operdata.com.br/blog/analise-de-cluster/. Acesso em: 31 de ago. de 2021.

PALMA, L.F. **Agrupamento de dados: k-médias.** Trabalho de Conclusão de Curso – Curso Ciências Exatas e Tecnológicas. Universidade Federal do Recôncavo da Bahia, Cruz das Almas, p. 46. 2018.

PEREIRA, D. L. **Estudo do PSO na Clusterização de dados.** 2007. 52f. Monografia (Ciência da Computação). Universidade Federal de Lavras. Minas Gerais. 2007.

PESSOA, A.M.dos.S; RÊGO, E.R.do; RÊGO, M.M.do, 2017. **Divergência genética e análise dialélica em pimenteiras ornamentais (*Capsicum annuum* L.).** João Pessoa: Editora da UFPB. http://www.editora.ufpb.br/sistema/press5/index.php/UFPB/catalog/book/337.

PINTO, C.M.F; DONZELE, S.M.L. Diversidade das pimentas Capsicum. **Revista Campo e Negócios online**, 2021. Disponível em: https://revistacampoenegocios.com.br/pimentas-capsicum/. Acesso em: 23 de fev. de 2022.

PRADO, T. C. **Segmentação de Imagens Coloridas Utilizando Técnicas de Agrupamento de Dados.** Florianópolis, Santa Catarina, 2008.

PUIATTI, G.A; CECON, P.R; NASCIMENTO, M; NASCIMENTO, A.C.C; FINGER, F.L; PUIATTI, M; SILVA, F.F; SILVA, A.R.da. Comparação dos métodos de agrupamento de Tocher e UPGMA no estudo de divergência genética em acessos de alho. **Revista da Estatística da Universidade Federal de Ouro Preto**, Ouro Preto, v.3, n.3, p. 275 - 279, 2014.

R Core Team (2021). **R**: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

REGAZZI, A.J; CRUZ, C.D. **Análise Multivariada Aplicada.** Viçosa, MG, 2020. p. 1-401.

RIBEIRO, C.S. da C.; FREITAS, I.C.de.; CARVALHO, S.I.C. Produção de pimentas diversas na região de Bico de Papagaio – TO. **Horticultura Brasileira**, Brasília, v.24, n.2, p. 1218, 2006.

RIBEIRO, C.S.da.C; REIFSCHNEIDER, F.J.B. Genética e melhoramento. *In*: RIBEIRO, C.S.da.C; LOPES, C.A; CARVALHO, S.I.C. de; HENZ, G.P; REIFSCHNEIDER, F.J.B. **Pimentas *Capsicum***. 1. ed. Brasília, DF: Embrapa Hortaliças, 2008. p. 1 – 202. Disponível em: https://www.infoteca.cnptia.embrapa.br. Acesso em: 03 set. 2021.

ROSEMBURG, H. C. **Cluster analysis for researchers.** California, EUA: Lifetime Learning, 1984. 334p.

ROHLF, F.J. Adaptive Hierarchical clustering schemes. **Systematic Zoology**, v.18, p. 58 – 82, 1970.

SANTOS, M.F. dos; SOUSA, C.C.de.; CARVALHAES, M.A; SILVA, K.J.D; LIMA, P.S.da.C. Variação genética em populações naturais de babaçu (Orbignya Phalerata Mart.) por marcadores morfológicos. **II Congresso Brasileiro de Recursos Genéticos**, Belém, PA, p. 1 - 5, 2012.

SHARMA, S. **Applied multivariate techniquies.** New York: John Wiley, 1996. 493p.

SILVA, A.R. **Métodos de agrupamento: avaliação e aplicação ao estudo de divergência genética em acessos de alho**. Dissertação (Mestrado em Estatística Aplicada e Biometria). Universidade Federal de Viçosa. Viçosa, MG, p. 67. 2012.

SILVA, L.dos. S.N; MORAIS, G.C; COSTA, L.da.S; SANTOS, J.F.F.dos; FILHA, C.M.R.da.S; SILVA, R.N.O. Divergência genética em genótipos de Capsicum annuum L. (Solanaceae) promissores para uso ornamental. **Revista Brasileira de Gestão Ambiental e Sustentabilidade**, v.7, n.17, p. 1165-1174, 2020. Disponível em: http://revista.ecogestaobrasil.net/v7n17/v07n17a09a.html. Acesso em: 21 ago. 2021.

SOKAL, R.R; ROHLF, F. J. The comparison of dendrograms by objective methods. **Taxon**, v.11, p. 30 – 40, 1962.

SOUZA, L.G. **Agrupamento dos municípios do Sul/Sudoeste de Minas Gerais em relação ao envelhecimento populacional.** Dissertação (Mestrado em Estatística aplicada e Biometria). Universidade Federal de Alfenas. Alfenas, p. 83. 2017.

VASCONCELOS, E.S. de; CRUZ, C.D; BHERING, L.L; JÚNIOR, M.F.R.R. Método alternativo para análise de agrupamento. **Pesquisa agropecuária brasileira**, Brasília, v.42, n.10, p. 1421 - 1428, 2007.

VASCONCELOS, C.S; BARBIERI, R.L; NEITZKE, R.S; PRIORI, D.; FISCHER, S.Z; MISTURA, C.C. Determinação da dissimilaridade genética entre acessos de *Capsicum chinense* com base em caracteres de flores. **Revista Ceres**, Viçosa, v.59, n.4, p. 493 – 498, 2012.

VAZ PATTO, M.C; SATOVIC, Z; PÊGO, S; FEVEREIRO, P. Assessing the genetic diversity of Portuguese maize germoplasm using microsatellite markers. **Euphytica**, Wageningen, v. 137, n.1, p. 63 – 72, 2004.

WAGNER, C.M. **Variabilidade e base genética da pungência e de caracteres do fruto: implicações no melhoramento de uma população de *Capsicum annuum* L.** Tese

(Doutorado em Agronomia, Área de concentração: Genética e Melhoramento de Plantas). Universidade de São Paulo. Piracicaba, SP. p. 123. 2003.

WANG, D.; BOSLAND, P. W. The Genes of Capsicum. **HortScience**, v.41, n. 5, p. 1169 – 1187, 2006.

WARD, J. H. Hierarquical grouping to optimize an objective function. **Journal of the American Statistical Association**, v. 58, p. 236 – 244. Mar. 1963.

APÊNDICE A - Script das Análises no *Software* R

```
#************************************************************************#

1. Leitura dos dados

#*************************************************************************

dados=read.xlsx("Dissertação.xlsx")
dados
dados3=dados[ ,2:7]
dados3

#************************************************************************#

2. Padronização dos dados

#*************************************************************************

transformacao=preProcess(dados3, method = c("center", "scale"))
dados_transformado=predict(transformacao, dados3)
dados_transformado

#************************************************************************#

3. Distância generalizada de Mahalanobis

#*************************************************************************
S=cov(dados_transformado)

S
L=solve(S)
L

D2=D2.dist(data=dados_transformado,cov=S)
D2

#************************************************************************#

4. Agrupamentos hierárquicos

#*************************************************************************

 hc1=hclust(D2,method="single")
 hc1
 hc2=hclust(D2, method="ward.D2")
 hc2
#************************************************************************#

5. Pontos de corte dos dendrogramas(Método de Mojena, 1977)

#*************************************************************************

k=1.25 # Milligan e Cooper (1985)

PChc1=mean(hc1$height)+k*sd(hc1$height)
PChc1

PChc2=mean(hc2$height)+k*sd(hc2$height)
PChc2

par(family="serif",las=1)
plot(hc1, hang = -
1,cex=0.8,ylab="Dissimilaridade")abline(h=PChc1,v=NULL,col=3,lty=1)


G=rect.hclust(hc1,h=PChc1,which=c(1:2),border=15:20)#Identificar os grupos
```

```
 G

hc1$height
plot(hc2, hang = -
1,cex=0.8,ylab="Dissimilaridade")abline(h=PChc2,v=NULL,col=7,lty=1)

G=rect.hclust(hc2,h=PChc2,which=c(1:12),border=1:5)#Identificarosgrupos
  G
hc2$height

#*****************************************************************************#

6. Coeficiente de Correlação Cofenética (r)

#*****************************************************************************

D1=cophenetic(hc1)
D1
cor(D2,D1)

D3=cophenetic(hc2)
D3
cor(D2,D3)


#*****************************************************************************#7.
Método de Tocher

#*****************************************************************************


toc=tocher (D2, algorithm= "original")
toc
names (toc)
toc$clusters


#*****************************************************************************#8.
Método das k-médias

#*****************************************************************************

#Definindo o número ótimo de clusters
 fviz_nbclust (x=dados_transformado, FUNcluster=kmeans,
               method="wss")+
               geom_vline(xintercept=3)

#Fazendo análise de clusters
 kmeans=kmeans(x=X, centers=3)  #centers=n de clusters



#*****************************************************************************#9.
RMSSTD

#*****************************************************************************

hclus_eval <- function(dados_transformado, dist_m = 'euclidean', clus_m = 'single',
plot_op = T, dist_cust = NA){

  output_list <- list()

  ## Some Initial Calculations

  ### END INITIAL CALCULATIONS

  ### Create Distance matrix and Clustering ###
  Info <- paste('Creating Distance Matrix using', dist_m)
```

```
  print(Info)

  if (dist_m == 'custom' & !is.na(dist_cust)) {dist1 <- dist_cust}
  else dist1 <- dist(dados_transformado, method= dist_m)


  Info <- paste('Clustering using', clus_m)
  print(Info)

  clust1 <- hclust(dist1, method = clus_m)

  print('Clustering Complete. Access the Cluster object in first element of
output')

  output_list[[1]] <- clust1


  rs_out <- affected.rows(clust1$merge)

  me <- gen_cutmat(clust1, dim(dados_transformado)[1])
  rs <- rs_out[[2]]

  # Calculate Metrics #

  print('Calculating RMSSTD')
  output_list[[2]] <- RMSSTD_FUNC(dados_transformado, rs,
dim(dados_transformado)[1])
  print('RMSSTD Done. Access in Element 2')

  print('Calculating RSQ')
  output_list[[3]] <- RSQ_FUNC(dados_transformado, me, dim(dados_transformado)[1])
  print('RSQ Done. Access in Element 3')

  print('Calculating SPRSQ')
  output_list[[4]] <- SPRS_FUNC(dados_transformado, rs, dim(dados_transformado)[1],
me)
  print('SPRSQ Done. Access in Element 4')

  print('Calculating Cluster Dist. ')
  output_list[[5]] <- cluster.dis(dados_transformado, clust1, output_list[[4]])
  print('CD Done. Access in Element 5')

  ylim_n <- max(output_list[[2]], output_list[[3]], output_list[[4]],
output_list[[5]])

  if (plot_op == T){
    plot(output_list[[3]], type = 'l', col = 'pink', ylim = c(0, ylim_n), lwd=3)
    lines(output_list[[2]], col = 'blue', lwd=3)
    lines(output_list[[4]], col = 'green', lwd=3)
    lines(output_list[[5]], col = 'purple', lwd=3)
    legend('topright', c('RSQ', 'RMSSTD', 'SPRSQ', 'CD'), lty = c(1 ,1,1,1),
           col = c('pink', 'blue', 'green', 'purple'))
  }

  }

  return(output_list)

}


affected.rows <- function(object.merge){
  x <- object.merge
  list.affected.rows <- list()
  tot.cat <- nrow(x)+1
  cluster.mat <- matrix(0, nrow=nrow(x), ncol=tot.cat)
  cluster.mat[nrow(x), -x[1, ]] <- 1
```

```
  for (i in 2:nrow(x)){
    cluster.mat[tot.cat-i, ] <- cluster.mat[tot.cat+1-i, ]
    if (sum(sign(x[i, ])) == -2) cluster.mat[tot.cat-i, -x[i,]] <- i
    if (sum(sign(x[i, ])) == 0){
      cluster.mat[tot.cat-i, -x[i, 1]] <- x[i, 2]
      cluster.mat[tot.cat-i, cluster.mat[tot.cat-i, ] == x[i, 2]] <- i
    }
    if (sum(sign(x[i, ])) == +2){
      cluster.mat[tot.cat-i, cluster.mat[tot.cat-i, ] == x[i, 1]] <- i
      cluster.mat[tot.cat-i, cluster.mat[tot.cat-i, ] == x[i, 2]] <- i
    }
  }

  for (i in 1:nrow(cluster.mat)) list.affected.rows[[i]] <- which(cluster.mat[i, ]
== (tot.cat-i))
  return(list(cluster.mat, list.affected.rows))
}


gen_cutmat <- function(clus_out, mdim) {
  cut_mat <- matrix(nrow = mdim, ncol = mdim)
  for (i in mdim:1){
    cut_mat[, i] <- cutree(clus_out,k=i)
  }
  return(cut_mat)
}


RSQ_FUNC <- function(mod_dat, cut_mat, dim_obs) {

  # SS at levels
  SS_vect <- rep(NA, dim_obs)
  x_bar_all <- apply(mod_dat, 2, mean)
  tot_ss <- sum(apply(mod_dat, 1, function(x) (dist(rbind(x, x_bar_all))^2)))

  for (i in (dim_obs:1)) {
    # Get Clusters in H

    clus_g <- unique(cut_mat[, i])
    level_PG <- 0

    for (k in clus_g) {
      need_rows <- which(cut_mat[, i] %in% k)

      if (length(need_rows) > 1) {
        x_bar_k <- apply(mod_dat[need_rows, ], 2, mean)
        w_k <- sum(apply(mod_dat[need_rows, ], 1, function(x) (dist(rbind(x,
x_bar_k))^2)))
      }

      if (length(need_rows) == 1) {
        x_bar_k <- mod_dat[need_rows, ]
        w_k <- dist(rbind(mod_dat[need_rows, ], x_bar_k))^2
      }

      level_PG <- level_PG + w_k

    }

    SS_vect[i] <- 1 - (level_PG/tot_ss)
  }

  return(SS_vect)

}

SPRS_FUNC <- function(dados_transformado, rows, dim_m, cut_mat) {
```

```
  SPRS_vect <- rep(NA, dim_m)
  SPRS_vect[dim_m] <- 0
  x_bar_all <- apply(dados_transformado, 2, mean)
  tot_ss <- sum(apply(dados_transformado, 1, function(x) (dist(rbind(x,
x_bar_all))^2)))


  for (i in (dim_m - 1):1) {

    need_rows <- rows[[i]]

    if (length(need_rows) > 1) {
      x_bar_m <- apply(dados_transformado[need_rows, ], 2, mean)
      w_m <- sum(apply(dados_transformado[need_rows, ], 1,
                       function(x) (dist(rbind(x, x_bar_m))^2)))
    }

    if (length(need_rows) == 1) {
      x_bar_m <- dados_transformado[need_rows, ]
      w_m <- dist(rbind(dados_transformado[need_rows, ], x_bar_m))^2
    }

    # Get clusters of these rows from the previous step
    need_clus_prev <- unique(cut_mat[need_rows, i + 1])

    # Get Rows of cluster K
    need_rows_k <- which(cut_mat[, i + 1] %in% need_clus_prev[1])

    if (length(need_rows_k) > 1) {
      x_bar_k <- apply(dados_transformado[need_rows_k, ], 2, mean)
      w_k <- sum(apply(dados_transformado[need_rows_k, ], 1,
                       function(x) (dist(rbind(x, x_bar_k))^2)))
    }

    if (length(need_rows_k) == 1) {
      x_bar_k <- dados_transformado[need_rows_k, ]
      w_k <- dist(rbind(dados_transformado[need_rows_k, ], x_bar_k))^2
    }

    # Get Rows of cluster L
    need_rows_l <- which(cut_mat[, i + 1] %in% need_clus_prev[2])

    if (length(need_rows_l) > 1) {
      x_bar_l <- apply(dados_transformado[need_rows_l, ], 2, mean)
      w_l <- sum(apply(dados_transformado[need_rows_l, ], 1,
                       function(x) (dist(rbind(x, x_bar_l))^2)))
    }

    if (length(need_rows_l) == 1) {
      x_bar_l <- dados_transformado[need_rows_l, ]
      w_l <- dist(rbind(dados_transformado[need_rows_l, ], x_bar_l))^2
    }

    SPRS_vect[i] <- (w_m - w_k - w_l)/ tot_ss
  }

  return(SPRS_vect)
}

RMSSTD_FUNC <- function(mod_dat, row_v, dim_obs){
  RMSSTD_vec <- rep(NA, dim_obs)
  RMSSTD_vec[dim_obs] <- 0

  for (i in (dim_obs - 1):1) {
    x_bar_k <- apply(mod_dat[row_v[[i]], ], 2, mean)
    w_k <- sum(apply(mod_dat[row_v[[i]], ], 1, function(x) (dist(rbind(x,
x_bar_k))^2)))
```

```
    RMSSTD_vec[i] <- sqrt(w_k/(dim(mod_dat)[2]*(length(row_v[[i]]) - 1)))
  }

  return(RMSSTD_vec)

}

cluster.dis <- function(dados_transformado, hclust.obj, SPRS_V){
  meth <- hclust.obj$method
  cd_vec <- rep(0, nrow(dados_transformado)-1)
  cluster.mat <- affected.rows(hclust.obj$merge)[[1]]
  affected.rows <- affected.rows(hclust.obj$merge)[[2]]

  if (meth == "ward") cd_vec <- SPRS_V

  if (meth == "average"){
    for (i in 1:length(cd_vec)){
      merged.set <- hclust.obj$merge[nrow(dados_transformado)-i, ]
      if (sum(sign(merged.set)) == -2){
        temp.1 <- -merged.set[1]
        temp.2 <- -merged.set[2]
      }

      if (sum(sign(merged.set)) == 0){
        temp.1 <- -merged.set[sign(merged.set) == -1]
        temp.2 <- which(cluster.mat[nrow(dados_transformado)-
merged.set[sign(merged.set) == 1], ] == merged.set[sign(merged.set) == 1])
      }

      if (sum(sign(merged.set)) == 2){
        temp.1 <- which(cluster.mat[nrow(dados_transformado)-merged.set[1], ] ==
merged.set[1])
        temp.2 <- which(cluster.mat[nrow(dados_transformado)-merged.set[2], ] ==
merged.set[2])
      }

      each.dis <- rep(0, length(temp.1)*length(temp.2))
      temp.sum <- 0
      for (j in 1:length(temp.1)){
        for (k in 1:length(temp.2)){
          emp.sum <- temp.sum+dist(rbind(dados_transformado[temp.1[j], ],
dados_transformado[temp.2[k], ]))
        }
      }
      cd_vec[i] <- temp.sum/(length(temp.1)*length(temp.2))
    }
  }

  if (meth == "single"){
    for (i in 1:length(cd_vec)){
      merged.set <- hclust.obj$merge[nrow(dados_transformado)-i, ]
      if (sum(sign(merged.set)) == -2){
        temp.1 <- -merged.set[1]
        temp.2 <- -merged.set[2]
      }

      if (sum(sign(merged.set)) == 0){
        temp.1 <- -merged.set[sign(merged.set) == -1]
        temp.2 <- which(cluster.mat[nrow(dados_transformado)-
merged.set[sign(merged.set) == 1], ] == merged.set[sign(merged.set) == 1])
      }

      if (sum(sign(merged.set)) == 2){
        temp.1 <- which(cluster.mat[nrow(dados_transformado)-merged.set[1], ] ==
merged.set[1])
        temp.2 <- which(cluster.mat[nrow(dados_transformado)-merged.set[2], ] ==
merged.set[2])
```

```
      }

      each.dis <- matrix(0, nrow=length(temp.1), ncol=length(temp.2))
      for (j in 1:length(temp.1)){
        for (k in 1:length(temp.2)){
          each.dis[j, k] <- dist(rbind(dados_transformado[temp.1[j], ],
dados_transformado[temp.2[k], ]))
        }
      }
      cd_vec[i] <- min(each.dis)
    }
  }

  if (meth == "complete"){
    for (i in 1:length(cd_vec)){
      merged.set <- hclust.obj$merge[nrow(dados_transformado)-i, ]
      if (sum(sign(merged.set)) == -2){
        temp.1 <- -merged.set[1]
        temp.2 <- -merged.set[2]
      }

      if (sum(sign(merged.set)) == 0){
        temp.1 <- -merged.set[sign(merged.set) == -1]
        temp.2 <- which(cluster.mat[nrow(dados_transformado)-
merged.set[sign(merged.set) == 1], ] == merged.set[sign(merged.set) == 1])
      }

      if (sum(sign(merged.set)) == 2){
        temp.1 <- which(cluster.mat[nrow(dados_transformado)-merged.set[1], ] ==
merged.set[1])
        temp.2 <- which(cluster.mat[nrow(dados_transformado)-merged.set[2], ] ==
merged.set[2])
      }

      each.dis <- matrix(0, nrow=length(temp.1), ncol=length(temp.2))
      for (j in 1:length(temp.1)){
        for (k in 1:length(temp.2)){
          each.dis[j, k] <- dist(rbind(dados_transformado[temp.1[j], ],
dados_transformado[temp.2[k], ]))
        }
      }
      cd_vec[i] <- max(each.dis)
    }
  }

  if (meth == "mcquitty"){
    print("Mcquitty agglomeration for cluster distance metric not currently
supported")
  }

  if (meth == "median"){
    print("Median agglomeration for cluster distance metric not currently
supported")
  }

  if (meth == "centroid"){
    print("Centroid agglomeration for cluster distance metric not currently
supported")
  }

  return(cd_vec)
}


hclus = hclus_eval(dados_transformado, dist_m = 'euclidean', clus_m = 'single',
plot_op = T)
str(hclus)
```

```
RMSSTD <- hclus[[2]]


 #*****************************************************************************#
```