Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

ANÁLISE DE CLUSTER PARA AVALIAÇÃO DA DIVERGÊNCIA GENÉTICA EM VARIEDADES DE MANGABA

Luana Baia Sousa

Uberlândia-MG
2018

Luana Baia Sousa

ANÁLISE DE CLUSTER PARA AVALIAÇÃO DA DIVERGÊNCIA GENÉTICA EM VARIEDADES DE MANGABA

Projeto de trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Lúcio Borges de Araújo

Uberlândia-MG

2018

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

_____

Prof. Dr. Lúcio Borges de Araújo

_____

Prof. Dr. Janser Moura Pereira

_____

Profª. Dr. Patrícia Ferreira Paranaíba

Uberlândia-MG

2018

Agradecimentos

Primeiramente, agradeço a Deus por ter me dado paciência o suficiente para conseguir vencer todas as batalhas durante o tempo de curso. Agradeço por ter me dado força, foco e fé para superar todas as pedras pelo caminho.

Agradeço a toda minha família, especialmente, a meu paizinho amado Olair Reis e minha amada mãezinha Maria Luisa por todos esses anos de incentivo, ensinamento, compreensão e luta, por nunca desistirem de lutar junto comigo e vencermos todas as dificuldades encontradas ao longo desses anos. As minhas irmãs Lorena Baía e Loriana Baía pelos seus sábios conselhos, por terem paciência comigo sempre que eu precisava e por ter me ajudado a superar cada empecilho juntas. Ao meu amado irmão Luís Hernandes, que foi um presente de Deus e que me ensinou muito, mesmo com sua pouca idade. Quero agradecer também a todo o restante da minha família que me deram apoio e entendem o quanto foram importantes nessa batalha, especialmente ao meu tio Nazareno Baía, um dos meus maiores incentivadores. Ao meu avô querido Jaime Maués que, infelizmente, não está mais entre nós, mas que eu tenho certeza que sempre desejou o meu melhor.

Ao meu orientador, Lúcio Borges de Araújo, é um prazer concluir o curso sob sua orientação. Obrigada por ter tido paciência durante esse ano e compartilhando seu conhecimento, sempre com carisma e humildade.

Aos técnicos que convivi esses anos, principalmente aos servidores da DIRPS que convivemos durante dois anos de estágio, especialmente ao Alécio Dantas e Odilon Tudini que me ensinaram muito sobre amizade, ética e transparência, que pra mim foram mais que supervisores, ao Alexandre Soares por ter sido meu amigo e ótimo conselheiro. Aos meus professores a quem tive a honra de ser aluna no decorrer da graduação, que foram muito importantes no meu desenvolvimento e ensinamento. Em especial, ao Edson Agustini, Janser Moura Pereira, José Waldemar da Silva, Lúcio Borges de Araújo, Marcelo Tavares, Patrícia Ferreira Paranaíba, Priscila Neves Faria e Raquel Romes Linhares.

Aos meus queridos amigos que me apoiam e me encheram de alegria quando eu estava triste, em especial aos que fiz ao decorrer da graduação, a Isabella Silva que já conhecia antes da UFU, mas que não imaginava que nos tornaríamos grandes amigas, a Marcella Spini por ter me aberto portas ao mercado, aos meus amigos de estágio da DIRPS Luana de Fátima, Pedro Vitorino, Samanta Carvalho e Thiago Amorim. Aos meus amigos e colegas da Estatística que me ajudaram muito sempre que precisei, principalmente a Andriely Antunes, Gabriela Bolaina, Jéssica Mendes, João Pedro Peres, Paloma Larissa, Patrícia Lorena, Vivian Barreto e Weila Freitas. Aos meus colegas de estágio Lidiana e Arthur e a Aleixa Reis que se tornou minha amiga tão facilmente. A Cristiane Silva uma amiga que sempre me escutou e deu conselhos quando precisava. E por fim, a Universidade Federal de Uberlândia por ter me acolhido tão bem.

Resumo

A mangaba é uma fruta típica da Caatinga, e pode também ser encontrada no Cerrado. Ela é pequena tem um formato parecido com o de uma pera, polpa branca, cremosa e suculenta, um pouco ácida e leitosa. É uma fruta comum no litoral do Nordeste, produzida principalmente em Sergipe e possui diversas variedades. Assim, este estudo teve por objetivo utilizar a Análise de *Cluster* para avaliar a divergência genética entre as variedades do fruto. O presente estudo permitiu concluir que a Análise de Cluster mostrou eficaz na identificação de 4 grupos das trinta e uma variedades da mangaba, em que foram agrupados pelo método de *K-means*, onde a maior divergência genética aconteceu entre os grupos 1 e 2.

Palavras-chave: Distância, Variedades, Métodos Hierárquicos.

Sumário

Lista de Figuras I

Lista de Tabelas III

1 Introdução 1

2 Material e Métodos 3

2.1 Análise de Agrupamentos . . . 3

2.2 Medidas de Similaridade ou Distância . . . 4

2.2.1 Medidas de Distância . . . 4

2.2.2 Medidas Correlacionais . . . 5

2.3 Métodos Hierárquicos . . . 6

2.4 Métodos Não Hierárquicos . . . 8

2.5 Material . . . 9

3 Resultados 11

4 Conclusões 19

Referências Bibliográficas 21

Lista de Figuras

2.1 Dendrograma de agrupamento de 5 indivíduos 7

3.1 Dendrograma: Distância Euclidiana e Método da Ligação Individual 12
3.2 Dendrograma: Distância Euclidiana e Método da Ligação Completa 13
3.3 Dendrograma: Distância Euclidiana e Método da Ligação Média 13
3.4 Dendrograma: Distância Euclidiana e Método de Ward 14
3.5 Dendrograma: Distância Euclidiana e Método do Centroide 14

Lista de Tabelas

3.1 Análise Descritiva das Variáveis Características da Mangaba 11
3.2 Coeficiente de Correlação Cofenético Por Método de Agrupamento 12
3.3 Grupos obtidos pela Análise de *Cluster* pelo método *K-Means* 15
3.4 Média Para Cada Grupo 17

1. Introdução

A mangaba, também chamada de mangaba-ovo, é o fruto da mangabeira (*Hancornia speciosa*), árvore rústica que pode chegar a dez metros de altura e é típica do bioma Caatinga, e pode também ser encontrada em fisionomias do Cerrado. É resistente a seca e se desenvolve bem em solos ácidos e pobres em nutrientes. Parecida com as plantas da Caatinga, a mangabeira possui tronco tortuoso com casca rugosa e áspera. A fruta é pequena tem um formato parecido com de uma pera, polpa branca, cremosa e suculenta, um pouco ácida e leitosa, daí surgiu o nome, de origem tupi-guarani, que significa "coisa boa de comer". As sementes são achatadas e arredondadas e ficam no interior da polpa. É uma fruta comum no litoral do Nordeste, mas vem sofrendo com o desmatamento da vegetação nativa para dar lugar a plantações de cana-de-açúcar e a empreendimentos imobiliários, correndo risco de sumir dessa região[1].

A mangaba é conhecida pelo seu ótimo aroma e sabor sendo consumida naturalmente ou processada na forma de doces, compotas, licor, vinagre e, principalmente, suco e sorvete, os quais são bem aceitos pelas agroindústrias e consumidores. O fruto apresenta um alto rendimento de polpa, em torno de 94%[13].

No litoral do Nordeste, a mangabeira, apresenta duas florações e frutificações ao longo do ano. Geralmente, a produção de frutos acontece entre dezembro e abril, na safra de verão, e de junho a julho, na safra de inverno. No verão, a produção de frutos é maior e os frutos tendem a ter uma aparência melhor, e no inverno a produção é menor e os frutos apresentam manchas escuras que mudam a sua aparência[5].

Sergipe é o maior estado produtor brasileiro, sendo quase toda sua produção da vegetação nativa, embora as primeiras áreas cultivadas estejam entrando em produção. Neste estado, a mangaba é a fruta mais consumida na forma de sorvete e polpa concentrada[2]. A mangaba também é utilizada na indústria alimentícia e medicinal, e a mangabeira já foi muito utilizada para a extração do seu látex para produção de borracha, e embora ela seja uma planta produtora de látex, o seu fruto, de sabor e aroma bastante apreciados, é o principal produto explorado, sobretudo pelas indústrias de polpas, sucos e sorvetes. Algumas partes da planta têm aplicação na medicina popular, como a casca, a folha e as raízes[11].

As variedades botânicas de mangabeira se diferem por algumas de suas características morfológicas, principalmente relacionadas à folha e à flor. Geralmente, são aceitas seis variedades botânicas de mangabeira as quais são: *Hancornia speciosa var. speciosa* (variedade típica), *H. speciosa var. maximiliani*, *H. speciosa var. cuyabensis*, *H. speciosa var. lundi*, *H. speciosa var. gardneri* e *H. speciosa var. pubescens*. Nos estados do Norte e Nordeste, a variedade bo-

tânica *H. speciosa var. speciosa* tem maior ocorrência, e as demais concentram-se nas regiões Centro-Oeste e Sudeste[5].

Assim, o estudo da divergência genética é uma técnica que auxilia na identificação de indivíduos geneticamente divergentes que quando combinados, possam aumentar o efeito heterótico na progênie. Um método estatístico conveniente para ser aplicado nesse tipo de estudo é a análise de agrupamento (Análise de *cluster*). Porém, antes desse método ser aplicado, deve ser obtida uma matriz de similaridade (ou distância) entre os genótipos. Essas distâncias podem ser calculadas de diversas maneiras, como por exemplo, distância euclidiana ou distância de Mahalanobis[8].

Diante disto, decidiu-se analisar a similaridade entre as diferentes variedades botânicas de mangabeira, com o intuito de verificar quais variedades possuem maior características em comum. Portanto, utilizou-se a Análise de *Cluster* para avaliar as variedades botânicas que possuem o maior grau de similaridades entre as diferentes variedades.

2. Material e Métodos

2.1 Análise de Agrupamentos

Análise de agrupamentos também conhecida como análise de conglomerados (*cluster analysis*) é uma técnica estatística de interdependência que segundo [9] permite agrupar casos ou variáveis em grupos homogêneos em função de grau de similaridade entre os indivíduos, a partir de variáveis predeterminadas.

Com a análise de agrupamentos, é possível classificar objetos de modo que cada objeto seja semelhante aos outros no agrupamento de acordo com o conjunto de características escolhidas. Assim, a homogeneidade dos objetos dentro dos grupos e a heterogeneidade entre os demais grupos são maximizadas[10].

Segundo [9], a análise de agrupamentos pode ser utilizada para vários tipos de pesquisa, como por exemplo, identificar grupos de investimentos de acordo com os perfis de risco, e identificar grupos de alunos mais propensos à evasão. Em análise de agrupamentos, o grupo de variáveis se assemelha a análise fatorial, visto que ambas as técnicas tem como objetivo identificar grupos de variáveis relacionadas. Entretanto, a análise fatorial se mostra mais robusta para o agrupamento de variáveis em detrimento do agrupamento de observações, foco da Análise de *cluster*.

Para [15], temos cinco etapas para a aplicação de uma análise de agrupamentos:

1. Escolha do critério de parecença (proximidades): consiste em definir se as variáveis devem ou não ser padronizadas se o critério que será utilizado na determinação dos grupos.

2. Definição do número de grupos: o número de grupos pode ser definido a priori, através de algum conhecimento que se tenha sobre os dados, conveniência de análise ou definido a posteriori com base nos resultados da análise.

3. Formação dos grupos: nesta etapa é definido o algoritmo que será utilizado na identificação dos grupos. Método hierárquico ou não hierárquico.

4. Validação do agrupamento: deve-se garantir que de fato as variáveis têm comportamento diferenciado nos diversos grupos. Nesta etapa, é comum supor que cada grupo seja da amostra aleatória de alguma sub população e aplicar técnicas inferenciais para compará-las.

5. Interpretação dos grupos: ao final do processo de formação dos grupos é importante

caracterizar os grupos formados. O uso de estatísticas descritivas é recomendado para esta fase da análise.

As medidas de proximidades têm um papel central nos algoritmos de agrupamentos. Através delas são definidos critérios para avaliar se dois pontos estão próximos, e, portanto podem fazer parte de um mesmo grupo ou não. Essas medidas são divididas em dois grupos: medida de similaridade e medida de dissimilaridade[7].

2.2 MEDIDAS DE SIMILARIDADE OU DISTÂNCIA

De acordo com [10], a similaridade entre os objetos é uma medida empírica de correspondência, ou semelhança, entre objetos a serem agrupados. Quanto maior o valor, maior é a semelhança entre os objetos. De modo geral, as medidas de similaridade ou distância (dissimilaridade) podem ser classificadas em três tipos: medidas de distância, medidas correlacionais e medidas de associação.

Para [9] a escolha das medidas de similaridade implica o conhecimento da natureza das variáveis (discreta, contínua, binária) e da escala de medida (nominal, ordinal, intervalar ou razão). Além disso, tanto as medidas de distância quanto as medidas correlacionais requerem dados métricos, enquanto as medidas de associação são destinadas ao tratamento de dados não métricos (nominal ou ordinal).

2.2.1 MEDIDAS DE DISTÂNCIA

As medidas de distância são consideradas medidas de dissimilaridade, pois, quanto maiores os valores, menor é a semelhança entre os objetos. As principais medidas de dissimilaridade segundo [9] são:

a. Distância Euclidiana: a distância entre duas observações (i e j) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j) para as p variáveis.

$$d_{ij} = \sqrt{\sum_{k=1}^{p}(x_{ik} - x_{jk})^2} \tag{2.1}$$

$$d_{ij} = \sqrt{(x_{i1} - x_{j1})^2 + (x_{i2} - x_{j2})^2 + ... + (x_{ip} - x_{jp})^2} \tag{2.2}$$

Em que x_{ik} é o valor da variável k referente à observação i e x_{jk} representa a variável k para a observação j. Nesta abordagem, quanto menor a distância, mais similares serão as observações.

b. Distância Quadrática Euclidiana: a distância entre duas observações (i e j) corresponde à soma dos quadrados das diferenças entre i e j para as p variáveis.

$$d_{ij}^2 = \sum_{k=1}^{p}(x_{ik} - x_{jk})^2 \tag{2.3}$$

A distância quadrática euclidiana é recomendada para métodos de agrupamento Centroide e Ward.

c. Mahalanobis: a distância estatística entre dois indivíduos i e j, considerando a matriz de covariância para o cálculo das distâncias.

$$d_{ij} = \sqrt{(\boldsymbol{x_i} - \boldsymbol{x_j})'\boldsymbol{S^{-1}}(\boldsymbol{x_i} - \boldsymbol{x_j})} \tag{2.4}$$

2.2.2 MEDIDAS CORRELACIONAIS

Segundo [10], "as medidas correlacionadas representam similaridade pela correspondência de padrões ao longo das características (variáveis)".

A correlação de Pearson, dada pela Fórmula (2.5), dentre as medidas correlacionais, é mais utilizadas nas ciências sociais [19].

$$r_{ij} = \frac{\sum_{k=1}^{p}(x_{1k} - \bar{x_k})(x_{1j} - \bar{x_j})}{\sqrt{\sum_{k=1}^{p}(x_{1k} - \bar{x_k})^2}\sqrt{\sum_{k=1}^{p}(x_{1j} - \bar{x_j})^2}} \tag{2.5}$$

Em que,

x_{ik} é o valor da variável k para a observação i;

x_{ij} é o valor da variável j para a observação i;

$\bar{x_k}$ representa a média da variável k;

$\bar{x_j}$ representa a média da variável j e

p representa o número de variáveis.

O valor do coeficiente varia entre -1 e 1, em que o zero significa que não há associações. Assim, quanto maiores as correlações, maior é associação entre as variáveis. Lembrando que a medida de similaridade mais utilizada em Análise de cluster são as medidas de distância, visto que as medidas correlacionadas não focam a magnitude dos objetos, e sim a correlação entre perfis [9].

O coeficiente de correlação cofenético, dado pela Fórmula mede o grau de ajuste entre a matriz de dissimilaridade (matriz fenética F) e a matriz resultante da simplificação devido ao método de agrupamento (matriz cofenética C). Quando o coeficiente de correlação cofenético for maior do que 0,7, concluí-se que o método de agrupamento é adequado.

$$ccc = \frac{\hat{Cov}(F, C)}{\sqrt{\hat{V}(F) * \hat{V}(C)}} \tag{2.6}$$

2.3 MÉTODOS HIERÁRQUICOS

Os métodos hierárquicos são técnicas simples onde os dados são particionados sucessivamente, produzindo uma representação hierárquica dos agrupamentos[3].

Nas técnicas hierárquicas, existem dois tipos de procedimentos de agrupamento: os métodos aglomerativos e os divisivos. No método aglomerativo, cada sujeito começa com seu próprio agrupamento e, a partir deste ponto, novos agrupamentos são realizados por similaridade, ou seja, no início cada indivíduo representa um grupo. Na etapa seguinte, os dois indivíduos mais similares são agrupados primeiramente e, nas etapas subsequentes, vão se fundindo com os demais grupos de acordo com a proximidade. Assim, em cada etapa reduz-se o número de agrupamentos em uma unidade. Já no método divisivo todas as observações começam em um grande agregado, sendo separadas, primeiramente, as observações mais distantes, até cada observação se tornar um grupo isolado[9].

Segundo [10], uma característica importante dos procedimentos hierárquicos é que os resultados de um estágio anterior são sempre aninhados com os resultados de um estágio posterior, semelhante a estrutura de uma árvore.

Para [16], pode-se construir um gráfico chamado dendrograma que tem como objetivo apresentar o arranjo entre os objetos em uma escala de distância. Esse gráfico tem a forma de árvore no qual a escala vertical indica o nível de similaridade (ou dissimilaridade). No eixo horizontal, são marcados os elementos amostrais numa ordem conveniente relacionada à história de agrupamento. As linhas verticais, partindo dos elementos amostrais agrupados, têm altura correspondente ao nível em que os elementos foram considerados semelhantes, isto é, a distância do agrupamento ou o nível de similaridade.

Na Figura 2.1 pode-se ver um exemplo de um dendrograma em que foi usada a distância do coeficiente de correlação e o método da ligação média.

Segundo [9], o algoritmo do método aglomerativo se desenvolve nos seguintes passos:

1. "Começar com N *clusters* (um para cada sujeito ou variável) e calcular a matriz de distância (ou matriz de proximidade) D_{NxN}.

2. Procurar na matriz D os pares de sujeitos (ou variáveis) i e j mais semelhantes (com menor d_{i-j}). Caso existam vários grupos com d_{i-j} iguais, usar como primeiro agrupamento o que possuir o sujeito de menor valor numérico.

3. Combinar os *clusters* i e j (os dois com menores d_{i-j}) para formar o *cluster ij*. Atualizar a matriz D, eliminando a linha e a coluna correspondentes ao *cluster* j e adicionando uma nova linha e coluna com as distâncias entre o novo *cluster ij* e os restantes *clusters* originais.

4. Repetir os passos 2 e 3 N-1 vezes, tomando nota dos *clusters* criados em cada um dos passos e das distâncias entre estes. Na última iteração do algoritmo, todos os sujeitos (ou variáveis) são agrupados em um único *cluster*".

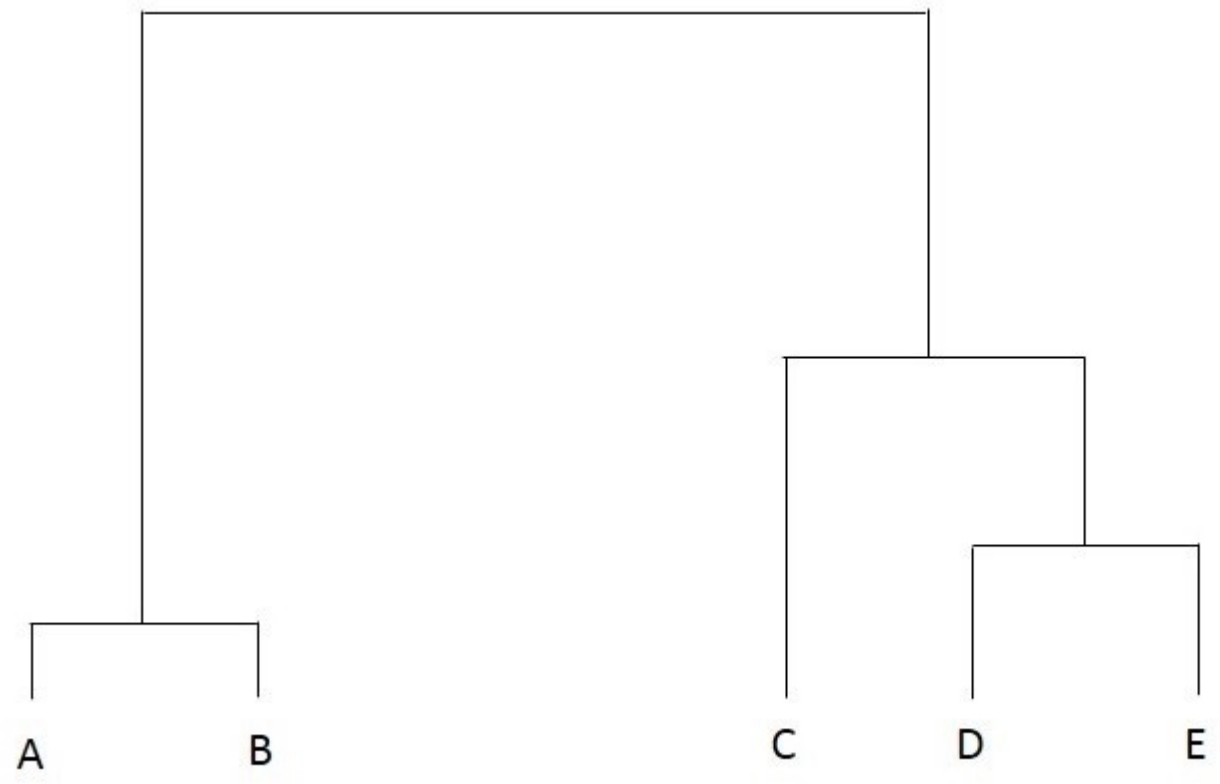

Figura 2.1: Dendrograma de agrupamento de 5 indivíduos

Depois da formação do primeiro *cluster*, é preciso definir como a distância entre dois *clusters* será computada. Assim, há diferentes métodos para a formação dos agrupamentos, sendo diferenciados, principalmente, pela maneira como as distâncias são calculadas entre os grupos já formados e os que faltam ser agrupados [9].

De acordo com [17] os métodos mais comuns são: *Single Linkage* (conhecido como método do vizinho mais próximo), *Complete Linkage* (conhecido como método do vizinho mais longe), *Avarage Linkage* (método da distância média), *Ward's Method* (método de Ward) e *Centroid Method* (método do centroide). Para [9] esses métodos são definidos da seguinte maneira:

1. **Método da Ligação Individual ou Menor Distância (*Single Linkage*)**: "baseia-se na distância mínima entre dois grupos de elementos, buscando agrupar inicialmente os objetos separados pela menor distância. Neste método, o primeiro grupo é formado pelos dois elementos que possuírem a menor distância entre eles, ou seja, será formado pelo vizinho mais próximo. Na próxima etapa, será agregado a este grupo o elemento que tiver menor distância em relação a eles, sucessivamente, até que se chega a um único grupo formado por todos os elementos."

   Dados dois grupos (i e k) e (k), a distância entre eles é representada pela distância mínima

de qualquer ponto de uma grupo até qualquer ponto do outro:

$$d_{(ij)k} = min\{d_{ik}, d_{jk}\} \tag{2.7}$$

2. **Método da Ligação Completa ou Maior Distância (*Complete Linkage*)**: "baseia-se na distância máxima, ao contrário do método da ligação individual. Neste método, a distância entre os dois grupos é definida como a distância máxima entre todos os pares de possibilidades de observações nos dois grupos. O método busca agrupar elementos cuja distância entre os mais afastados seja a menor."

   Dados dois grupos (i e k) e (k), a distância entre eles é representada pela distância máxima de qualquer ponto de um grupo até qualquer ponto do outro:

$$d_{(ij)k} = max\{d_{ik}, d_{jk}\} \tag{2.8}$$

3. **Método da Ligação Média ou Distância Média *Avarage Linkage***: "trata a distância entre dois grupos como sendo a distância média entre todos os pares de indivíduos dos dois grupos, buscando agrupar os agregados cuja distância média é a maior."

   Dados dois grupos (i e k) e (k), a distância entre eles é representada da seguinte maneira:

$$d_{(ij)k} = media\{d_{ik}, d_{jk}\} \tag{2.9}$$

4. **Método do Centroide *Centroid Method***: "baseia-se na distância (geralmente euclidiana ou quadrática euclidiana) entre os centroides, priorizando a menor distância entre eles. Este método identifica os dois grupos separados pela menor distância entre os pontos mais próximos e os coloca no mesmo agrupamento."

5. **Método do Ward *Ward's Method***: "busca agrupar os agregados que apresentam menor soma dos quadrados entre os dois agrupamentos, calculada sobre todas as variáveis."Trata-se de um método que tende a proporcionar agregados com aproximadamente o mesmo número de observações.[9]

Para [19], o método da ligação completa tende a formar grupos compactos e com indivíduos muito semelhantes entre si. Já o método de Ward é resumido em etapas, primeiro calcula a média das variáveis para cada grupo, depois é calculado o quadrado da distância euclidiana entre estas médias e os valores das variáveis para cada indivíduo, em seguida é somado a distância para todos os indivíduos e por último pretende-se minimizar a variância dentro dos grupos.

2.4 MÉTODOS NÃO HIERÁRQUICOS

No método não hierárquico o algoritmo não estabelece uma relação de hierarquia entre os sujeitos e os grupos, pois quando especificado o número de *clusters*, o processo é dinâmico e

interativo, tendo como objetivo identificar a melhor solução. Esse método é utilizado para agrupar indivíduos, e não variáveis, onde o número inicial de agrupamentos é definido pelo pesquisador[9].

O método *k-means*, um dos mais populares, é um algoritmo iterativo com a função de minimizar a soma das distâncias de cada padrão ao centroide de cada *clusters*, sobre todos os *clusters*. Assim, cada *cluster* é representado pelo centro do grupo e cada padrão é disposto ao *cluster* que está mais próximo[3].

De acordo com [14], no método *k-means*, o processo é composto de três passos:

1. "Partição inicial dos indivíduos em K *clusters* definidos pelo analista.

2. Cálculo dos centroides para cada um dos K *clusters* [...] e cálculo da distância euclidiana dos centroides a cada sujeito na base de dados.

3. Agrupar os sujeitos aos *clusters* cujos centroides se encontram mais próximos, e voltar ao passo 2 até que não ocorra variação significativa na distância mínima de cada sujeito da base de dados a cada um dos centroides dos K *clusters* (ou até que o número máximo de interações ou o critério de convergência, definido pelo analista, seja alcançado)".

2.5 Material

A partir dos dados de 2012 de variedades de mangaba fornecidos pela Embrapa (Empresa Brasileira de Pesquisa Agropecuária) Meio- Norte será aplicado a análise de agrupamento. Todos os acessos avaliados foram introduções da Empresa Estadual de Pesquisa Agropecuária da Paraíba (Emepa). Estes acessos pertencem a Embrapa Meio-Norte e tem duplicatas na Emepa. Utilizou-se trinta e uma variedades de mangaba, com dez repetições cada. As informações coletadas foram as características físicas e características químicas de cada variedade.

Para as características físicas foram coletadas as seguintes informações: massa do fruto em gramas, comprimento em milímetros, diâmetro em milímetros, massa da casca em milímetros, número de sementes do fruto, massa da semente, polpa em gramas e rendimento da polpa em percentual. Já as características químicas foram: grau Brix (Bx), alíquota de Acidez Total Titulável (ATT), volume de NaOH (Hidróxido de Sódio), pH, Acidez Total Titulável e a razão de Brix por ATT. Assim, realizou-se as análises dos dados no software R[18].

3. Resultados

Analisando os resultados obtidos, pode-se observar na Tabela 3.1 que a variável 8 (Percentual de Rendimento da polpa) apresentou os maiores valores de desvio padrão, o que significa que houve uma dispersão maior dos dados em relação à média, observa-se essa discrepância no valor mínimo (54,109) e máximo (74,106) da variável 8. Já a variável 10 (Alíquota de Acidez Total Titulável), obteve o menor desvio padrão, o que indica uma dispersão menor entre os dados em relação média.

Tabela 3.1: Análise Descritiva das Variáveis Características da Mangaba

Nome da Variável	Variável	Mínimo	Máximo	Média	Desvio Padrão
Massa do Fruto	V1	11,041	29,817	18,898	3,880
Comprimento	V2	26,699	38,518	32,826	2,696
Diâmetro	V3	25,589	37,172	31,380	2,502
Massa da Casca	V4	1,612	5,008	2,845	0,782
Número de Sementes por Fruto	V5	8,600	22,800	14,737	3,511
Massa da Semente	V6	2,198	6,919	4,197	1,180
Polpa	V7	6,668	17,890	11,860	2,450
Percentual de Rendimento da Polpa	V8	54,109	74,106	62,343	4,589
^{o}BRIX	V9	13,500	21,667	17,622	2,280
Alíquota de Acidez Total Titulável	V10	1,002	1,014	1,007	0,003
Volume de NaOH	V11	1,460	2,727	2,031	0,365
pH	V12	2,490	4,270	3,633	0,417
Acidez Total Titulável	V13	0,907	1,688	1,258	0,226
BRIX por Acidez Total Titulável	V14	9,418	22,313	15,164	3,372

Os resultados dos agrupamentos por cada método estão resumidos em dendrogramas nas figuras 3.1, 3.2, 3.3, 3.4 e 3.5. O método hierárquico de agrupamento escolhido foi o Método do Centroide obtido pela matriz de Distância Euclidiana (Figura 3.5), visto que o coeficiente de correlação cofenético mostra uma melhor distinção dos grupos, pois quando esse coeficiente for maior do que 0,7 pode-se dizer que o método de agrupamento foi adequado, observe na Tabela 3.2 os coeficientes de correlação de cofenéticos encontrados em cada método hierárquico. Assim, pelo dendrograma do Método do Centroide escolheu-se quatro grupos, sendo que os grupos retratados relacionaram as mangabas com com maior semelhança genética.

TABELA 3.2: Coeficiente de Correlação Cofenético Por Método de Agrupamento

Método de Agrupamento	Coeficiente de Correlação Confenético
Método da Ligação Individual	0,713
Método da Ligação Completa	0,489
Método da Ligação Média	0,708
Método de Ward	0,713
Método do Centroide	0,713

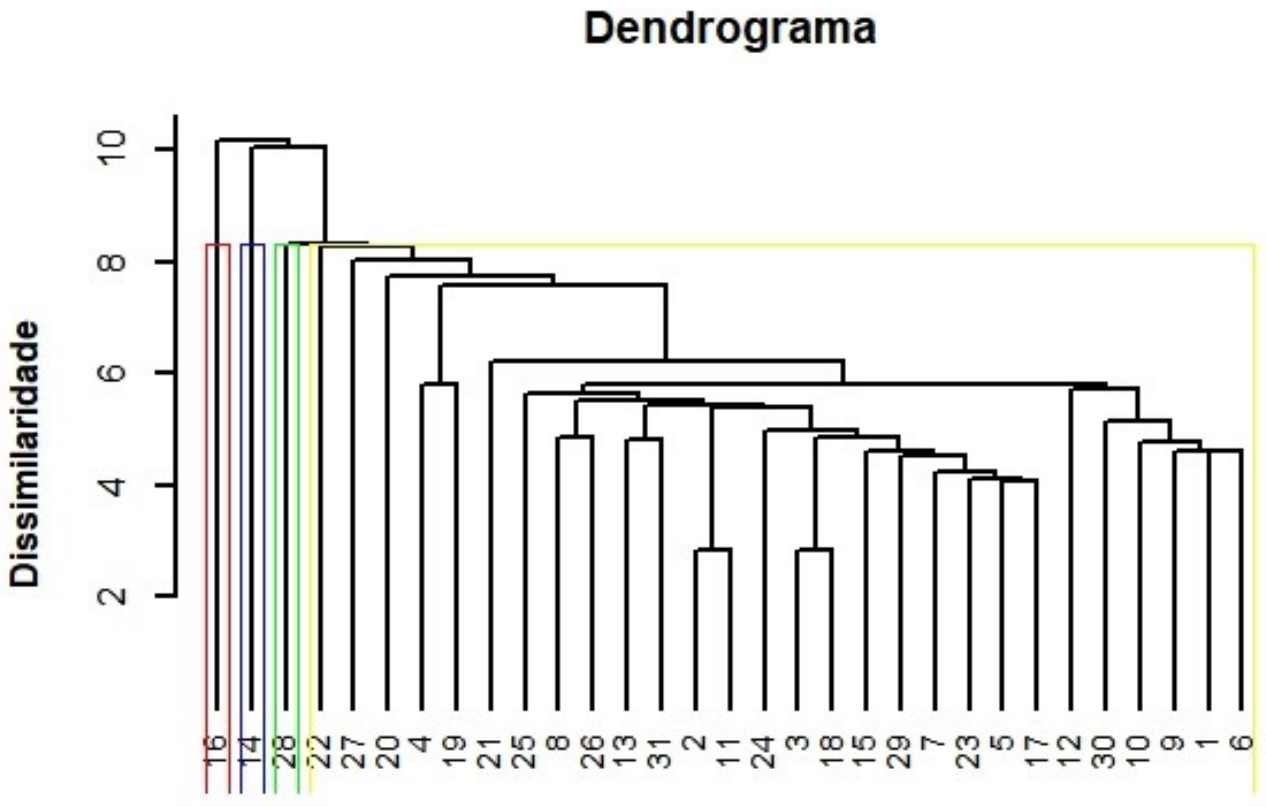

FIGURA 3.1: Dendrograma: Distância Euclidiana e Método da Ligação Individual

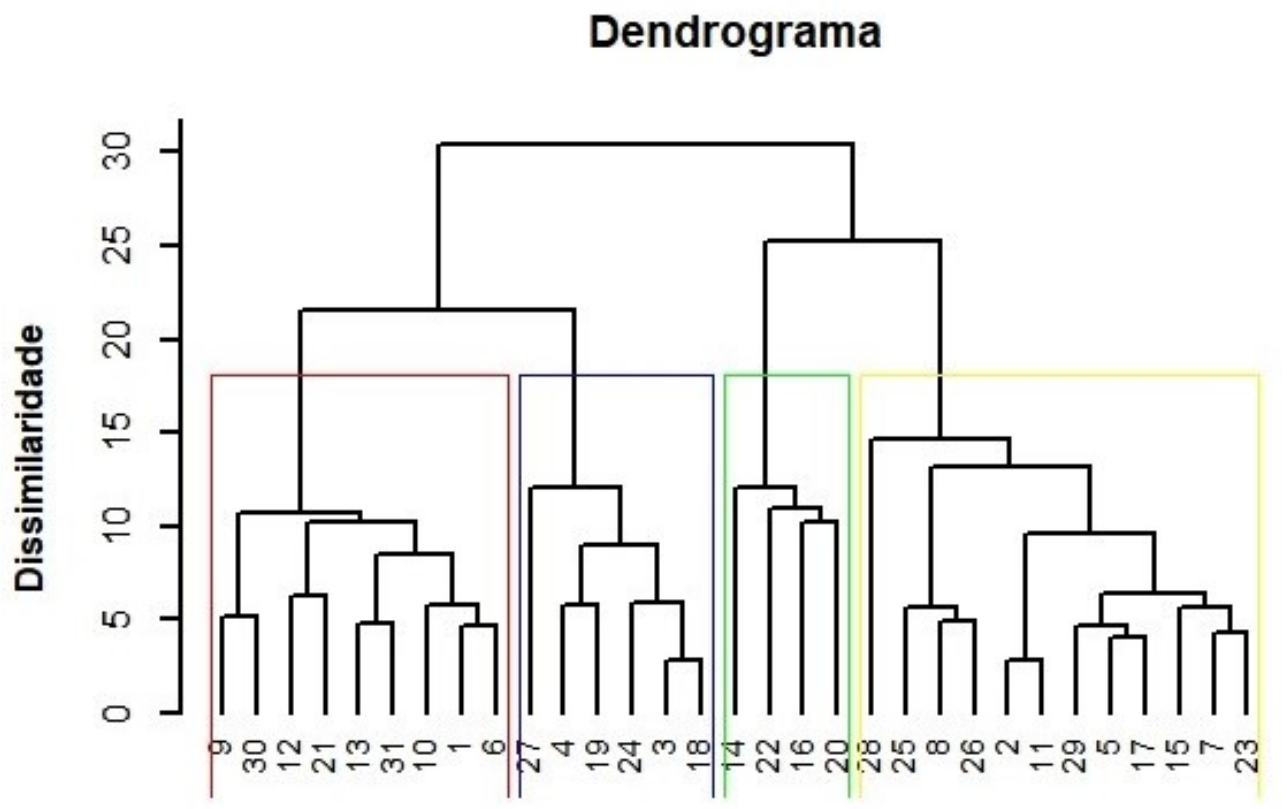

Figura 3.2: Dendrograma: Distância Euclidiana e Método da Ligação Completa

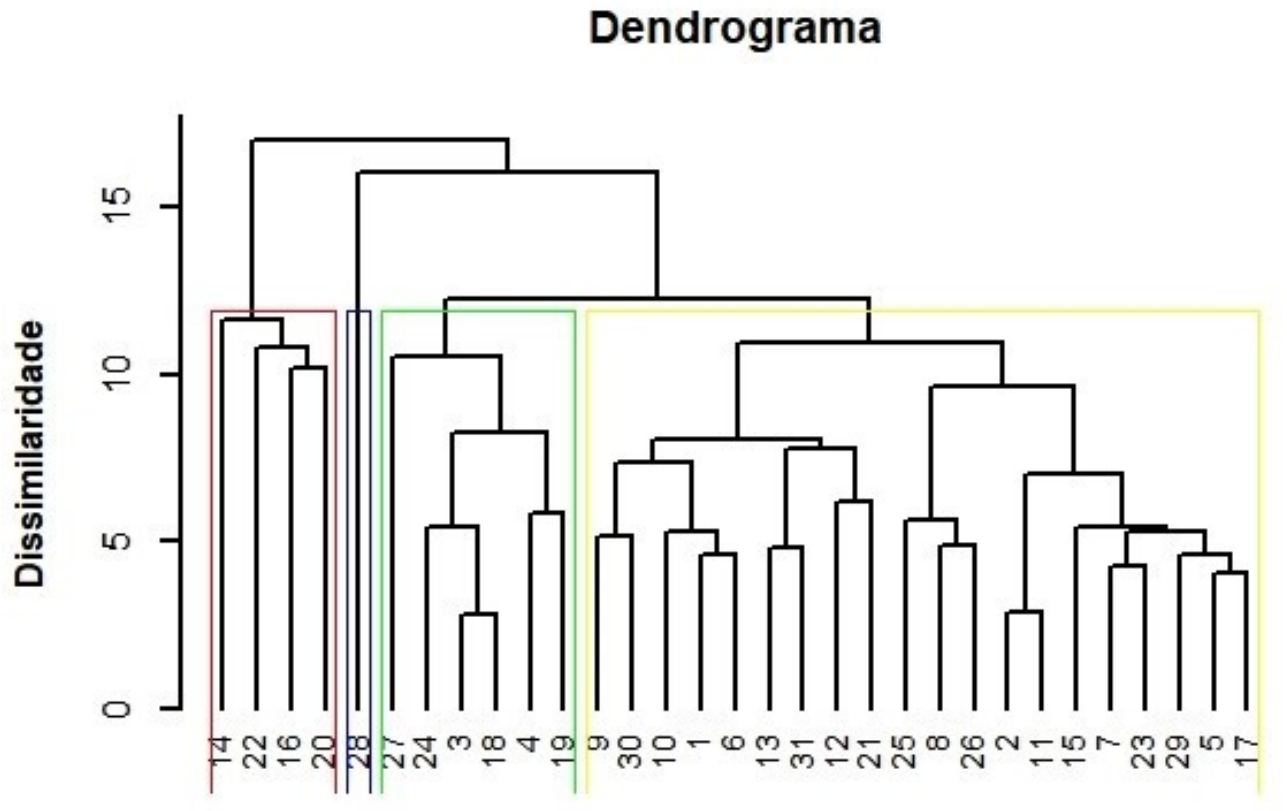

Figura 3.3: Dendrograma: Distância Euclidiana e Método da Ligação Média

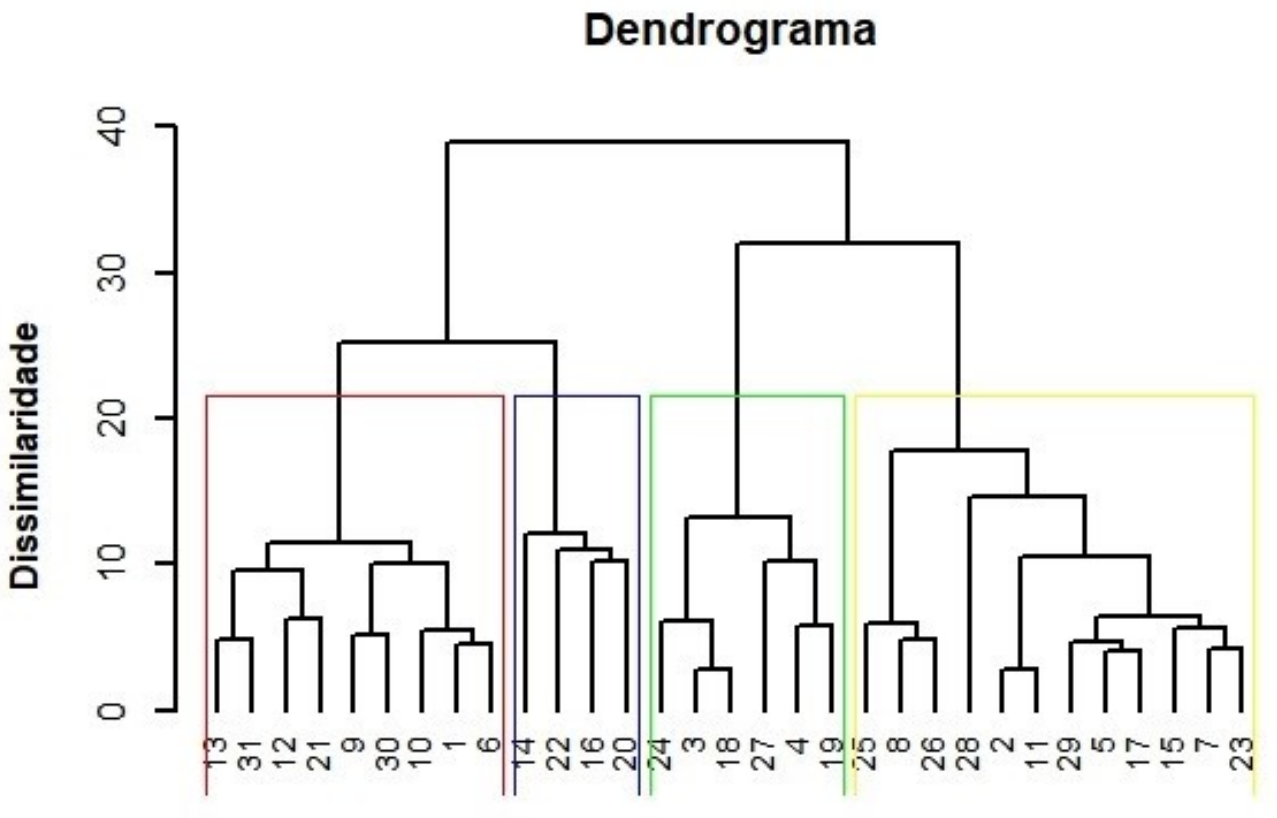

FIGURA 3.4: Dendrograma: Distância Euclidiana e Método de Ward

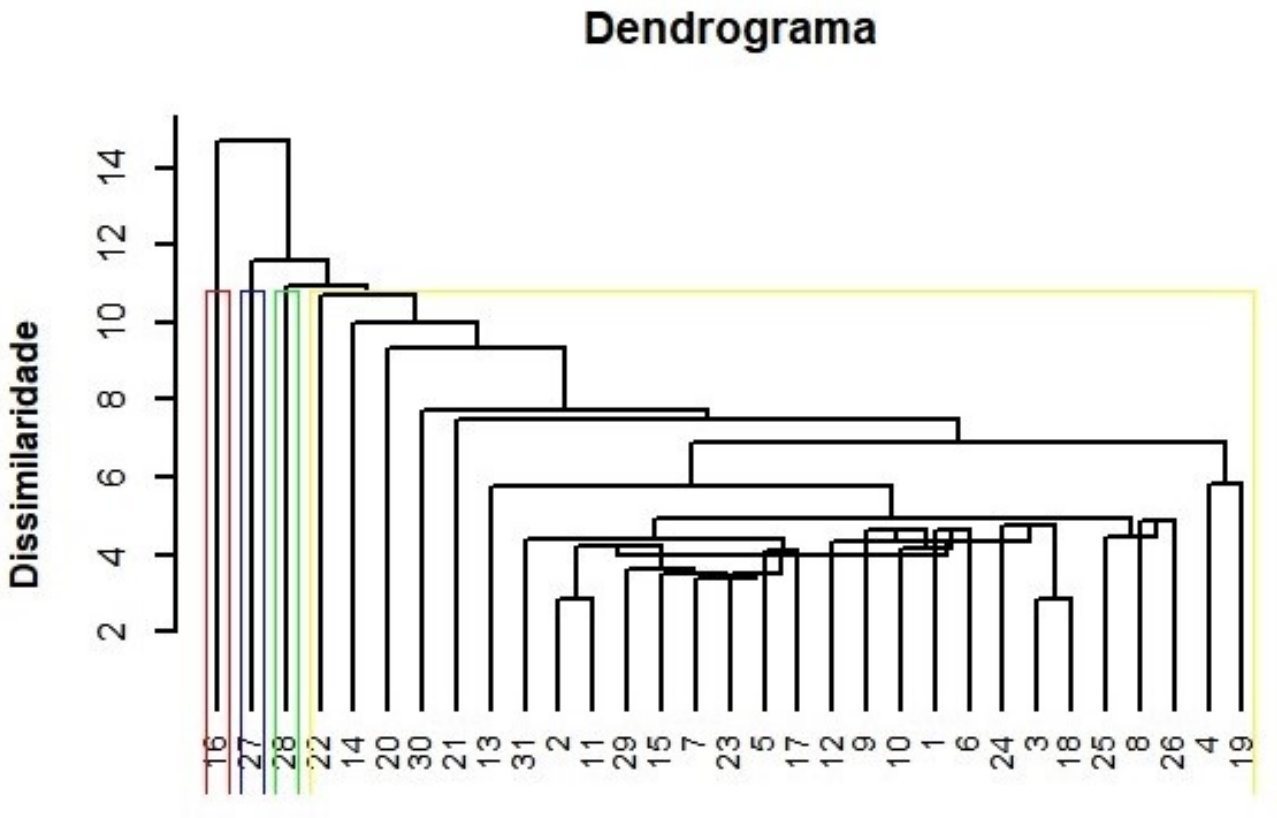

FIGURA 3.5: Dendrograma: Distância Euclidiana e Método do Centroide

Como o método de *K-means* é um processo dinâmico e interativo, que tem como objetivo identificar a melhor solução de agrupamento, usou-se esse método para definir os 4 grupos, esses grupos foram retratados na Tabela 3.3. O grupo 1 foi formado com 5 variedades genéticas, o grupo 2 com 6 variedades e o grupo 3 e 4 com 10 variedades genéticas cada um. Pode-se notar que as variedades 12 e 13, que são da variedade F14 pertencem ao mesmo grupo, o mesmo ocorre com as variáveis 20 e 21 que pertencem ao grupo F21 e com as variáveis 30 e 31 do grupo F30.

Tabela 3.3: Grupos obtidos pela Análise de *Cluster* pelo método *K-Means*

Grupo 1	**Grupo 2**	**Grupo 3**	**Grupo 4**
		F1	F1
F1			
F2			F4
		F4	F4
	F6	F7	
		F8	F11
		F14	
		F14	
	F15		F15
	F17	F17	
F19			
F20	F21		
	F21		
	F24		
			F25
		F26	F27
			F27
F28			F28
			F28
		F30	
		F31	

Na Tabela 3.4, é possível ver a média dentro de cada grupo. Pode-se observar que o grupo 2 é o grupo com a maior média de massa do fruto, comprimento, diâmetro, massa da casca, número de sementes por fruto, massa da semente, polpa, ºBRIX e BRIX por ATT, ou seja o grupo com maior média nas características físicas, exceto para o rendimento da polpa que teve maior média no grupo 4. Assim, pode-se supor que as variedades do grupo 2 são as maiores em tamanho, peso e polpa. Grau Brix é uma escala numérica que mede a quantidade de sólidos solúveis em uma solução de sacarose. Essa escala é geralmente utilizada na indústria alimentícia para medir a quantidade aproximada de açúcares em sucos de fruta e vinhos[6].

O grupo 2 apresentou uma massa média do fruto de 24,418 gramas e uma polpa média de 14,795 gramas, o que corresponde com o percentual médio de rendimento da polpa encontrado. E para a média do Grau Brix desse grupo, pode-se dizer que as variedades apresentaram as maiores quantidades de compostos solúveis numa solução de sacarose em relação aos outros grupos.

Já as características químicas ficaram divididas entre o grupo 1, 2 e 3. O grupo 1 apresentou maior média para as variáveis volume de Hidróxido de Sódio (NaOH) e ATT, o grupo 2 como já mencionado, para as variáveis ºBRIX e BRIX por ATT, já o grupo 3 alcançou a maior média na variável pH e o grupo 4 no percentual de rendimento da polpa. A variável alíquota de acidez total titulável teve média aproximada para os quatro grupos.

TABELA 3.4: Média Para Cada Grupo

Variável	$\bar{X}_{Grupo_1}$	$\bar{X}_{Grupo_2}$	$\bar{X}_{Grupo_3}$	$\bar{X}_{Grupo_4}$
Massa do Fruto	13,388	24,418	18,730	18,510
Comprimento	28,403	36,137	33,065	32,811
Diâmetro	27,880	34,695	31,372	31,150
Massa da Casca	2,262	3,837	3,014	2,372
Número de Sementes por Fruto	11,320	19,467	15,538	12,807
Massa da Semente	2,943	5,787	4,494	3,468
Polpa	8,183	14,795	11,222	12,576
Percentual de Rendimento da Polpa	60,269	60,641	59,419	67,324
ºBRIX	18,047	19,079	18,066	16,092
Alíquota de Acidez Total Titulável	1,007	1,008	1,008	1,007
Volume de NaOH	2,081	1,896	2,071	2,047
pH	3,491	3,747	3,779	3,489
Acidez Total Titulável	1,290	1,174	1,282	1,269
BRIX por Acidez Total Titulável	14,982	18,931	14,592	13,566

O pH tem uma importância relevante para os alimentos, visto que quando o pH for dito ácido pode haver uma boa digestão alimentícia e bom aproveitamento dos nutrientes e vitaminas. O pH é uma escala numérica medida 0 a 14, quando abaixo de 7 é considerado ácido, se for 7 é neutro e maior do que 7 é básico[4]. Assim, todos os grupos tiveram uma média menor do que 7, logo podem ser considerados frutos ácidos.

Pode-se observar também na Tabela 3.4 que o grupo 1 obteve as menores médias em massa do fruto, comprimento, diâmetro, massa da casca, número de sementes por fruto, massa da semente e polpa. Assim, pode-se dizer que o grupo 1 foi o grupo com menor tamanho, peso e polpa.

O grupo 2 teve a menor média para volume de Hidróxido de Sódio (NaOH) e ATT, já o grupo 3 teve a menor média somente no percentual de rendimento da polpa e o grupo 4 para as variáveis ºBRIX, pH e BRIX por ATT.

4. Conclusões

O presente estudo permitiu concluir que a Análise de Clusters mostrou possível agrupar as trinta e uma variedades de mangaba em 4 grupos. Esses grupos foram agrupados pelo método de *K-means*. O grupo 1 e 2 são os grupos com maior divergência genética.

Foi possível verificar que dentro desses 4 grupos, o grupo 1 foi o grupo em que se encontram as variedades com menores tamanho, peso e polpa, isto é, menores frutos, já o grupo 2 foi o grupo com as maiores variedades em tamanho, peso e polpa, o grupo 3 são as variedades com maiores grandezas de pH, ou seja, os frutos mais ácidos e, o grupo 4 é o grupo com maior rendimento de polpa. Podemos dizer também que as variedades de mangaba são ácidas, visto que o pH médio de todos os grupos não ultrapassou 7.

Referências Bibliográficas

[1] *Mangaba*. http://www.cerratinga.org.br/mangaba/, acesso em 10-05-2018.

[2] *Mangaba*. http://www.wikiwand.com/pt/Mangaba, acesso em 10-05-2018.

[3] *Métodos de Agrupamentos de Dados*. https://www.maxwell.vrac.puc-rio.br/7975/7975_4.PDF, acesso em 31/05/2018.

[4] *O pH dos Alimentos*. https://www.portaleducacao.com.br/conteudo/artigos/nutricao/o-ph-dos-alimentos/52114, acesso em 02-12-2018.

[5] *Sistema de Produção de Mangaba para a Região Nordeste do Brasil*. https://www.spo.cnptia.embrapa.br/conteudo?p_p_id=conteudoportlet_WAR_sistemasdeproducaolf6_1ga1ceportlet&p_p_lifecycle=0&p_p_state=normal&p_p_mode=view&p_p_col_id=column-1&p_p_col_count=1&p_r_p_-76293187_sistemaProducaoId=7719&p_r_p_-996514994_topicoId=10322, acesso em 15/11/2018.

[6] ANDRADE, C. e MENDES, L.: *Procedimento de Análises Laboratoriais - Grau Brix*. http://cienciadeagricultor.blogspot.com/2013/07/grau-brix.html, acesso em 26-11-2018.

[7] BARROSO, L. e ARTES, R.: *Análise Multivariada*. Lavras: UFLA, 2003.

[8] BARROSO, N. C.: *Categorização de dados quantitavos para estudos de diversidade genética*. Dissestação de Mestrado, 2010. http://www.locus.ufv.br/bitstream/handle/123456789/4037/texto%20completo.pdf?sequence=1&isAllowed=y.

[9] FÁVERO, L. P., BELFIORE, P., SILVA, F. L. e CHAN, B. L.: *Análise de dados: modelagem multivariada para tomada de decisões*. Elsevier, 8ª ed., 2009.

[10] HAIR, J. F., ANDERSON, R. E., TATHAM, R. L. e BLACK, W. C.: *Análise multivariada de dados*. Porto Alegre: Bookman, 5ª ed., 2005.

[11] JÚNIOR, J. F. S. e LÉDO, A. S.: *Mangaba*. http://www.agencia.cnptia.embrapa.br/gestor/territorio_mata_sul_pernambucana/arvore/CONT000fdkckctq02wx5eo0a2ndxy7t9pn7e.html, acesso em 10-05-2018.

[12] JOHNSON, R. A. e WICHERN, D. W.: *Applied Multivariate Statisitical Analysis*. NJ: Pearson/Prentice Hall, 6ª ed., 2007.

[13] LEDÓ, A. S.: *A cultura da mangaba*. Embrapa, 1ª ed., 2015.

[14] MARÔCO, J.: *Análise Estatística Com o SPSS Statistics*. Lisboa: Edições Sílabo, 3ª ed., 2007.

[15] MENDES, J. V. M.: *Avaliação das regionais de uma empresa de telecomunicações, através de análise de Cluster*. Trabalho de Conclusão de Curso, 2017. https://repositorio.ufu.br/handle/123456789/19266.

[16] MINGOTI, S. A.: *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Belho Horizonte: Editora UFMG, 2005.

[17] MORAES, M. B. C.: *Análise Multivariada Aplicada à Contabilidade*, 2016. https://edisciplinas.usp.br/pluginfile.php/2232110/mod_resource/content/1/An%C3%A1liseMultivariada-Aula12.pdf, acesso em 31/05/2018.

[18] R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. https://www.R-project.org/.

[19] REIS, E.: *Estatística multivariada*. Lisboa: Sílabo, 2ª ed., 2001.