MÁCIO AUGUSTO DE ALBUQUERQUE

ANÁLISE DE AGRUPAMENTO HIERÁRQUICA E INCREMENTAL - ESTUDO DE CASO EM CIÊNCIAS FLORESTAIS

RECIFE-PE – FEV/2013.

UNIVERSIDADE FEDERAL RURAL DE PERNAMBUCO

PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO

PROGRAMA DE PÓS-GRADUAÇÃO EM BIOMETRIA E ESTATÍSTICA APLICADA

ANÁLISE DE AGRUPAMENTO HIERÁRQUICA E INCREMENTAL- ESTUDO DE CASO EM CIÊNCIAS FLORESTAIS

Tese apresentada ao Programa de Pós-Graduação em Biometria e Estatística Aplicada como exigência parcial à obtenção do título de Doutor em Biometria e Estatística Aplicada.

Área de Concentração: Estatística Aplicada

Orientador: Prof°. Dr. Rinaldo Luiz Caraciolo Ferreira

Coorientadores: Prof°. Dr. José Antônio Aleixo da Silva

RECIFE-PE – FEV/2013.

Ficha catalográfica

A345a Albuquerque, Mácio Augusto de
Análise de agrupamento hierárquica e incremental: estudo de caso em ciências florestais / Mácio Augusto de Albuquerque. – Recife, 2013.
160 f. : il.

Orientador: Rinaldo Luiz Caraciolo Ferreira.
Tese (Doutorado em Biometria e Estatística Aplicada) – Universidade Federal Rural de Pernambuco, Departamento de Estatística e Informática, Recife, 2013.
Referências.

1. Análise multivariada 2. Métodos de agrupamento 3. Método incremental 4. Métricas 5. Dendrograma 6. Índice de validação 7. Dados artificiais I. Ferreira, Rinaldo Luiz Caraciolo, orientador II. Título

CDD 574.018

A minha família, em especial à minha esposa Edna e aos meus filhos Tarsyla, Tárcio, a minha mãe Luzia (in menoriam) e meus irmãos, por sempre me incentivarem, apoiarem e darem força para seguir em busca dos meus ideais.

DEDICO

Agradecimentos

Agradecimento é o sentimento de principal importância dentro da realização deste trabalho. Acredito que seria impossível a evolução do ser sem que houvesse, direta e indiretamente a participação de outros. E que essa interação influenciou significativamente a minha vida, permitindo-me crescer no sentido mais amplo da palavra. Por isso, tentarei agradecer a todos envolvidos na elaboração deste trabalho.

A Deus pela força para realização desse trabalho.

Ao meu orientador professor doutor Rinaldo Luiz Caraciolo Ferreira, pela dedicação, praticidade, honestidade e orientação na execução deste trabalho; pela amizade e apoio durante todo o curso e principalmente pela confiança em mim depositada.

Ao professor Aleixo pela competência e atenção dispensada no desenvolvimento desta tese.

A coordenadora do curso de Biometria e Estatística Aplicada professora doutora Tatijana Stosic, pela orientação, pela dedicação e esforço pelo curso. Meu respeito e gratidão.

Aos professores do Programa de Doutorado em Biometria e Estatística Aplicada/UFRPE pelos conhecimentos transmitidos.

Aos meus colegas de turma pelas experiências trocadas e pelas lições apreendidas com cada um.

Aos colegas do mestrado e doutorado pelo bom convívio.

Ao secretário Marco Santos pelo carinho, respeito e amizade.

Aos colegas do Departamento de Estatística da Universidade Estadual da Paraíba pela compreensão, apoio e incentivo dado durante o doutorado.

Aos participantes da banca examinadora pelas sugestões.

À Universidade Federal Rural de Pernambuco, pela oportunidade de realização do meu doutorado.

A todos que de alguma forma contribuíram para o crescimento de cada momento para realização deste trabalho.

Resumo

A Estatística Multivariada, por avaliar múltiplas variáveis em uma observação de uma amostra, destaca-se por ter diversas aplicações, tanto no campo científico quanto em diversas áreas, como a Ciência Florestal, no entanto, mesmo em abordagens mais recentes apresentam dificuldades que limitam o seu uso por pesquisadores. Dentre a família de técnicas multivariadas, a análise de agrupamento é uma das mais utilizadas, dada sua utilidade pragmática. Neste trabalho, objetivou-se desenvolver uma nova abordagem para a análise de agrupamento, a partir da combinação de características das técnicas hierárquicas e não-hierárquicas. Assim, procurou-se fornecer uma análise exploratória mais completa dos dados, visando facilitar o trabalho dos pesquisadores quanto a *outliers*, a número de grupos, a técnicas de agrupamento, e de validação dos grupos, e aumentar o conhecimento que pode ser obtido com a aplicação de um conjunto de sentenças lógicas em análise de agrupamento. Utilizaram-se dados estruturais de um remanescente de Mata Atlântica, denominada por Mata das Caldeiras, localizado no município de Catende, PE. Os dados foram obtidos a partir de 40 parcelas de 250 m² (10 x 25 m, cada), alocadas sistematicamente ao longo de todo o remanescente, distando 25 m entre si. Em cada parcela, os indivíduos arbóreos vivos com CAP (circunferência à altura do peito – 1,30 m do solo) $\geq$ 15 cm receberam placas metálicas enumeradas e tiveram os seguintes dados anotados: o nome vulgar, a CAP e a altura. Posteriormente, foram calculados os diâmetros (DAP), as áreas basimétricas e a altura de Lorey. Realizaram-se três experimentos baseados na simulação de Monte Carlo, com dados artificiais que permitiu a avaliação dos comportamentos dos métodos por meio da normal bivariada. Foram aplicados, com base distância euclidiana, o método incremental, as técnicas hierárquicas de ligação simples, de ligação completa, de ligação média e de Ward. Para validação dos métodos foram utilizados os coeficientes cofenético, R^2, Pseudo F, Wilks e Rand ajustado. A metodologia apresentada foi aplicada a conjuntos de dados artificiais, comparando os resultados com os obtidos por outros métodos de agrupamentos. Observou-se que o uso da técnica incremental tem o potencial de melhorar significativamente a tomada de decisões sobre o número operacional grupos,

tornando-se, assim, uma técnica recomendável para buscar o número de grupos ideal de forma criteriosa. Como consideração final, sugere-se a utilização de outras medidas de dissimilaridade, com vistas a avaliar o desempenho da técnica em diversas circunstâncias.

Palavras-chave: Análise Multivariada, métodos de agrupamento, método incremental, métricas, algoritmo de agrupamento, dendrograma, índice de validação.

Abstract

The Multivariate Statistics, by evaluating multiple variables in a note of a sample, stands out for having several applications, both in science and in various areas, such as Forest Science, however, even in more recent approaches have difficulties that limit its use by researchers. Among the family of multivariate techniques, cluster analysis is one of the most used, given its pragmatic usefulness. This work aimed to develop a new approach to cluster analysis, based on a combination of characteristics of hierarchical and non-hierarchical technical. Thus, the objective was to provide a more complete exploratory analysis of data, to facilitate the work of researchers as outliers, the number of groups, clustering techniques, and validation groups, and increasing knowledge that can be obtained with applying a set of logical sentences in cluster analysis. Data were used from a structural remnant of Forest Atlântica, named Forest das Caldeiras, located in the municipality of Catende, PE. Data were obtained from plots of 40 m² 250 (10 x 25 m each), allocated systematically throughout the remainder, distant 25 m apart. In each plot, individual trees alive with CBH (circumference at breast height - 1.30m) ≥ 15 cm plates were enumerated and the following data were recorded: the common name, the CBH and height. Subsequently, we calculated the diameters (DBH), and basal areas of Lorey height. There were three experiments based on Monte Carlo simulation, with artificial data that permit evaluation of the condud of the methods through bivariate normal. Were applied, based on Euclidean distance, the incremental method, the techniques of hierarchical single linkage, complete linkage, average linkage and the Ward. For validation of methods were used cofenetic coefficients, R2, Pseudo F, Wilks and Rand adjusted. We have applied our methodology on simulated data sets, so as to evaluate its performance, comparing it with other clustering methods.It was observed that the use of the incremental technique has the potential to significantly improve the decisions about the number operating groups, becoming therefore recommended a technique to search for the optimal number of groups wisely. As a final consideration, we suggest the use of other measures of dissimilarity, in order to evaluate the performance of the technique in various circumstances.

Key words: Multivariate Analysis, clustering method., incremental method, metrics, clustering algorithm, indices of validation.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O levantamento dos dados de campo em uma floresta é uma atividade complexa, devido tanto às adversidades inerentes ao ambiente quanto à demanda pela qualidade dos dados a serem coletados. Além disso, a densidade e a diversidade de uma floresta também tornam complexa a coleta dos dados, tais como as alturas de fuste, comercial, total, do DAP e volume da árvore. Os estudos de crescimento e produção florestal necessitam desses dados, logo, torna-se imprescindível a identificação de metodologias que, apesar das dificuldades inerentes, possam gerar estimativas de qualidade (GONÇALVES et al., 2009).

Nas florestas tropicais, é característica a presença de um dossel, formado por espécies capazes de atingir elevadas alturas, e de um sub-bosque, formado por espécies tolerantes a sombra, de baixa estatura (GOURLET-FLEURY et al., 2005). Por isso, existe a procura por métodos que possa simplificar a sua estrutura e facilitar as interpretações quanto ao funcionamento desses ecossistemas. Dentre essas técnicas, podem destacar o agrupamento da floresta, quanto às espécies relacionadas com o tamanho das árvores.

Em uma floresta natural, é comum distribuir as suas plantas conforme uma arquitetura de agrupamento, ou seja, por meio de uma divisão em grupos, pois segundo Roberts e Gilliam (1995), esta arquitetura é um fator importante para manter maior diversidade de espécies, bem como subsidiar a gestão florestal. No entanto, o reconhecimento de grupos ainda é assunto muito controverso na literatura (VALE et al., 2009).

Em Ciências Florestais, nas suas diversas áreas, e conforme os objetivos do fenômeno que ocorrem nesta ciência sempre estão relacionados ao comportamento conjunto de várias variáveis, ou seja, as interpretações devem levar em conta as inter-relações que podem existir entre as variáveis estudadas.

Na Ciência Florestal, são utilizadas diversas técnicas multivariadas, destacando-se: análise de componentes principais (SCHEEREN et al., 2000; SANTOS et al., 2004; CLARKE et al., 2006; DOBBERTIN; NOBIS, 2010; ARAUJO et al., 2010); análise discriminante e classificação (SANTOS et al., 2004; HUANG

et al., 2007; ZHANG et al., 2009; LIMA JÚNIOR et al., 2009; BENITES et al., 2010; RODE et al., 2011; VALENTE et al., 2011; SEIDEL et al., 2012); análise fatorial (LIMA JÚNIOR et al., 2009; TOLEDO, 2009; VALENTE et al., 2011; SALOMÃO et al., 2012); e análise de agrupamento (SANTOS et al., 2004; BOSCARIOLI et al., 2006; FORTES et al., 2008; LIMA JÚNIOR et al., 2009; LUDEWIG et al., 2010; ARAUJO et al., 2010; BARALOTO et al., 2010;).

A utilização de uma dessas técnicas está atrelada ao objetivo do estudo, assim como ao tipo de espécies e suas inter-relações. Dentre as diversas técnicas multivariadas encontradas na literatura, sem dúvida, a mais utilizada é a análise de agrupamento, porque os pesquisadores sempre buscam o reconhecimento de um padrão, visando a simplificar e a explicar o comportamento de uma floresta, de um grupo de parcelas, ou de uma parcela, a partir da mensuração de várias espécies.

Vários são os tipos de técnicas de agrupamento encontradas na literatura MARDIA et al., 1979; ANDERSON, 1984; KAUFMANN; ROSSEEUW, 1990; REIS, 2001; EVERITT, 2005; JOHNSON; WICHERN, 2007; FERREIRA et al., 2008; HAIR et al., 2010), tendo o pesquisador que tomar a decisão de qual é a mais adequada ao seu propósito, uma vez que as diferentes técnicas também podem levar a diferentes soluções.

Na análise de agrupamento, o objetivo é reunir, por algum critério de classificação, os objetos ou parcelas de uma amostra da população estudada. Assim, a sua utilização exige a tomada de uma série de decisões independentes, que podem representar diferentes agrupamentos.

As técnicas de análise de agrupamento exigem, de seus usuários, a tomada de uma série de decisões independentes que, por sua vez requerem o conhecimento de suas propriedades, da escolha da similaridade ou dissimilaridade, dos diversos algoritmos e de um método de validade, que podem representar diferentes agrupamentos. Além disso, o resultado dos agrupamentos pode ser influenciado pela escolha da medida de similaridade ou dissimilaridade, bem como pela definição do número de grupos (GOWER; LEGENDRE, 1986; JACKSON et al., 1989; DUARTE et al., 1999; MEYER, et al., 2004; KUNZ et al., 2009).

O agrupamento é realizado de forma a minimizar as diferenças entre as parcelas em estudo dentro do agrupamento (*cluster*), e maximizar as diferenças entre as parcelas de agrupamentos diferentes.

A seleção de variáveis deve ser realizada com extremo cuidado, uma vez que os grupos a serem formados refletirão a estrutura inerente a elas, tendo-se em vista serão utilizadas para determinar a medida de similaridade ou dissimilaridade que corresponde ao critério de agrupamento (FORTES et al., 2008; BEZERRA NETO et al., 2010; BERTINI et al., 2010).

Vale observar que a técnica não distingue se as variáveis são ou não relevantes para o estudo, ficando essa tarefa a cargo do pesquisador.

Outro ponto importante a ressaltar é a inclusão de dados com comportamentos atípicos, isto é, com a presença de *outliers*, que podem ser definidos como observações que fogem do padrão esperado, em cada grupo, ou seja, referem-se a observações com características muito destoantes dos demais membros da população, podendo prejudicar a qualidade dos resultados (BEZERRA NETO et al., 2010; COIMBRA et al., 2010). Assim, antes de efetuar a análise de agrupamento, é recomendável verificar a existência de *outliers*, cabendo ao pesquisador decidir se devem continuar ou não na base de dados (LUDEWING et al., 2009; RODY et al., 2010).

Desta forma, neste trabalho se objetivou fornecer uma análise exploratória mais completa dos dados, visando facilitar o trabalho dos pesquisadores quanto a *outliers*, a número de grupos, a técnicas de agrupamento, e de validação dos grupos, e aumentar o conhecimento que pode ser obtido com a aplicação de um conjunto de sentenças lógicas em análise de agrupamento.

2 - REVISÃO DE LITERATURA

Em quase todas as áreas de pesquisa várias variáveis são mensuradas e, em geral, essas devem ser analisadas conjuntamente. A análise multivariada é a área da estatística que trata desse tipo de estudo e existem várias técnicas que podem ser aplicadas, sendo que, a utilização dessas depende do tipo de dado que se deseja analisar e dos objetivos do estudo.

Segundo Anderson (1984), existem, basicamente, duas formas de classificar as análises multivariadas: as que permitem extrair informações a respeito da independência entre as variáveis que caracterizam cada elemento, tais como análise fatorial, análise de agrupamento, análise canônica, análise de ordenamento multidimensional e análise de componentes principais; e as que permitem extrair informações a respeito da dependência entre uma ou mais variáveis ou uma com relação à outra, tais como análise de regressão multivariada, análise de contingência múltipla, análise discriminante e análise de variância multivariada.

A denominação “Análise Multivariada” corresponde a um conjunto de métodos e técnicas que analisam simultaneamente todas as variáveis na interpretação teórica do conjunto de dados. O primeiro passo para a utilização dessa análise é saber o que se pretende afirmar a respeito dos dados, visto que a técnica e o método estatístico ideal para a aplicação devem ser escolhidos de acordo com o objetivo da pesquisa.

Na Ciência Florestal, são utilizados vários métodos multivariados na resolução de problemas inerentes a ela, com destaque para as análises de componentes principais, discriminante, fatorial e de agrupamento (ALBUQUERQUE et al., 2006; OLIVEIRA et al., 2007; FERREIRA et al., 2008; MÁXIMO et al., 2009; LIMA JÚNIOR et al., 2009; BENITES et al., 2010; RODE et al., 2011; VASQUES et al., 2011). No entanto, a análise de agrupamento é a mais comumente utilizada para objetivos diversos, tais como, estratificação de áreas florestais e grupos de espécies funcionais.

O método da ligação completa e a distância euclidiana média padronizada foram utilizados por Fortes et al. (2008) objetivando agrupar por espécie as

matrizes de porta-sementes mais similares. Lobão et al. (2010) aplicaram a análise de componentes principais e de agrupamento visando agrupar espécies florestais pela similaridade das características físico-anatômicas e usos da madeira. Batista et al. (2011), visando conhecer e comparar a composição florística e a estrutura de duas áreas de florestas de várzea, utilizaram-se da distância euclidiana e do algoritmo de Ward. Assis et al. (2011) avaliaram semelhanças florísticas entre duas fisionomias de Floresta Atlântica por meio das distâncias de euclidiana simples e Bray-curtis, do coeficiente de Jaccard, do método de ligação média e, como medida de validação, o coeficiente cofenetico.

2.2 A Análise de Agrupamento

O grande sábio grego Aristóteles disse: “O homem vive classificando tudo o que vê”. Classificar significa agrupar, tendo por base aspectos de semelhança entre os elementos agrupados. Ao agrupar árvores, por exemplo, levam-se, em conta, critérios de semelhança como diâmetro, altura, área transversal, volume, idade, local, entre outros.

O termo Análise de Agrupamento (*Cluster Analysis*) foi introduzido por Tryon em 1939. No entanto, segundo Bergman e Feser (1998), ideias semelhantes já vinham ocorrendo desde o final do século XIX.

Os métodos de Análise de Agrupamento aparecem na literatura com diferentes denominações, como *Data Clustering*, Taxonomia Numérica, Análise de grupos ou Análise Aglomerados (ou Análise de *Clusters*), tipologia e reconhecimento de padrões não supervisionados. Provavelmente essa variedade de nomenclaturas deve-se à importância e a intensiva aplicação da Análise de Agrupamento, em diversas áreas de estudos. Análise de Agrupamento é o nome mais genérico, selecionado para se referir a tal processo, motivo pelo qual foi adotado nesta tese. É o nome dado às técnicas de análise que dividem os dados em grupos. Esses grupos podem ser constituídos por observações individuais multivariadas, ou por agrupamentos multivariados de variáveis.

A Análise de Agrupamento classifica objetos, parcelas (povoamento de espécies, número de árvores enumeradas), indivíduos, pessoas, animais, plantas ou itens, observando apenas as semelhanças ou as distâncias entre esses, sem

definir critérios de inclusão prévia em qualquer agrupamento. Os métodos de Análise de Agrupamento tentam organizar um conjunto de parcelas, para os quais é conhecida a informação, detalhada em grupos, relativamente homogêneos (agrupamento).

Essa técnica sumariza dados para interpretação, e utiliza métodos que procuram grupos excludentes, ascendentes, reduzindo as informações de um conjunto de n parcelas para informações de um novo conjunto de k grupos, onde k é significativamente menor que n, e aplicando a técnica hierárquica, resultando um dendrograma de exclusão (LUDEWING et al., 2009).

Segundo Scheeren et al. (2000), supondo-se que exista uma amostra de n parcelas, cada um dos quais tem um escore em p espécies, para planejar um esquema a fim de agrupar as parcelas em grupos, de modo que os similares estejam no mesmo grupo. Vale ressaltar que o método usado precisa ser complementado numérico, e o número de grupos não é conhecido usualmente.

Dentre a literatura, ainda não existe um procedimento padrão para resolver esta questão. Alguns pesquisadores das técnicas de agrupamento recomendam aplicar mais de um método sobre o mesmo conjunto de dados, e comparar os grupos formados para apresentar um melhor resultado.

Para Hair et al. (2010), o objetivo principal da Análise de Agrupamento é situar as observações homogêneas em grupos, a fim de definir uma estrutura para os dados. Para isso, são abordadas algumas questões básicas que devem ser consideradas durante a análise.

Quando o pesquisador optar pela técnica hierárquica, uma das primeiras decisões na análise se refere à medida de similaridade ou distância que deve ser estabelecida, ou seja, deve-se estabelecer a associação de dois objetos, baseada nas espéies da “espécies estatística de agrupamento”. Aleixo et al. (2008) define a “espécies estatística de agrupamento” como “o conjunto das espécies que representam as características usadas para comparar parcelas na análise de agrupamentos”.

De maneira geral, é difícil avaliar a qualidade do processo de agrupamento. Não existem testes estatísticos padrões para garantir que o resultado seja puramente aleatório. O valor do critério medido, da legitimidade do resultado, da aparência de uma hierarquia natural (quando for empregado um método não

hierárquico) e confiabilidade de testes de divisão de amostra oferecem informações úteis (OLIVEIRA et al., 2008). Entretanto, é difícil saber, exatamente, quais grupos são muito parecidos e quais parcelas são difíceis de serem inseridos. Em geral, não é fácil selecionar um critério de programa de agrupamento por meio de outra referência que não seja a disponibilidade.

Genericamente, a Análise de Agrupamento compreende cinco etapas:

1. Um padrão de representação que, de alguma forma, descreva cada parcela e que forneça o subsídio, para a diferenciação dos mesmos (que pode – ou não – incluir a ação de extração de características e/ou seleção das mesmas);
2. A definição de uma métrica apropriada para o padrão de representação das parcelas. Esta será usada para determinar um valor quantitativo de quão similares as parcelas do grupo são entre si;
3. A ação de agrupar os dados em seus determinados grupos;
4. A eventual necessidade de se abstrair os dados resultantes, para que possam ser interpretados para o fim que lhes cabe;
5. E por fim, a possível ação de avaliar o resultado do agrupamento em termos de um “bom resultado” ou um “resultado ruim”;

Os passos supracitados podem ser resolvidos de tantas formas diferentes que a ocorrência mais comum é de que se tenha uma técnica de agrupamento diferente para cada tarefa de agrupamento específica a ser implementada.

2.2.1 *Outliers* Multivariados

A inspeção inicial dos dados revela, com frequência, aspectos que podem surpreender o pesquisador, que pode se deparar com medidas incomuns ou informações inexistentes.

Como no caso univariado, antes de aplicar algum método multivariado, deve-se investigar a existência de valores discrepantes (*outliers* ou ruído), que podem afetar os resultados finais da análise estatística. Logo, é fundamental que

seja realizada uma análise exploratória das medidas, na tentativa de identificar pontos desse tipo (DUARTE, 2008; HAIR et al., 2010).

Em dados multivariados, uma medida é considerada *outliers,* caso esteja muito distante das restantes, no espaço p-dimensional definido pelas variáveis (PREARO et al., 2012), ou seja, deve ser uma medida não representativa da população, devendo, portanto, apresentar valores extremos em diversas variáveis e não apenas em outra. Vale salientar que é preciso ter muito cuidado com estes *outliers* multivariados, uma vez que é possível uma medida seja considerada um ponto discrepante, em termos multivariados e não, em termos univariados.

As medidas atípicas podem ser identificadas, sob uma perspectiva univariada, bivariada ou multivariada. A perspectiva univariada é aquela usual, para o caso de uma única variável (RENCHER; SCHAALJE, 2008). A bivariada refere-se aos casos dos gráficos de dispersão bidimensionais, aliados a elipses de confiança (JOHNSON; WICHERN, 2007). Já na perspectiva multivariada, gráficos de dispersão auxiliam na identificação de *outliers*, juntamente com a distância Mahalanobis, (D_i^2) e gráficos do tipo Q-Q plots (MINGOTI, 2007).

Segundo Cruz (1990) o uso da distância Mahalanobis, é sugerida por muitos textos como um método para detectar outliers em dados multivariados. Para indicar valores críticos *outliers*, sugere-se a estatística de teste $\left[\frac{p(n-1)}{(n-p)}\right] F_{(p,n-p,\alpha)}$, isto é, valores de D_i^2 maiores que o valor crítico dessa estatística são considerada *outliers*. Esta aproximação F é considerada mais adequada do que a distribuição $\chi^2_{(p,\alpha)}$, especialmente quando se lida com pequeno número de parcelas. No entanto, Penny (1996) afirmou que, na prática, a distribuição F é inapropriada para testar *outliers* para pequenas amostras.

No que tange a tais *outleirs,* uma *vez* identificadas como parcelas atípicas, por meio de um dos métodos, o pesquisador deve selecionar as que mostram verdadeira peculiaridade em comparação com o restante da população (MINGOTI, 2007).

Algumas técnicas multivariadas são também grandes aliadas na detecção de *outliers*, tais como:

1. Análise de Agrupamentos Hierárquicos: depois de realizado o agrupamento, pode-se identificar grupos, formados por apenas uma parcela. Cada um dessas parcelas pode ser classificada como possível *outliers*, visto que nenhuma outra parcela foi considerada similar para ser colocada no mesmo grupo dessas parcelas.

2. Análise de Componentes Principais: utilizam-se os escores das últimas componentes principais para a confecção de gráficos de dispersão, bi e tridimensional e Q-Q *plots*. Este método se justifica pelo fato de que a magnitude dos últimos componentes principais determina quão bem os primeiros se ajustam às parcelas. Na prática, as parcelas suspeitas serão aquelas que, no gráfico de dispersão dessas últimas componentes, se encontrarem distantes da nuvem de pontos (JOHNSON; WICHERN, 2007).

Convém observar que amostras grandes podem, eventualmente, exibir parcelas que aparentemente são atípicas, mas que não são essencialmente *outlies.* De fato, à medida que a amostra aumenta, é ampliada a chance de serem incluídos casos extremos que constituem parcelas legítimas da população, não sendo, dessa maneira, necessária, nem recomendada a sua remoção.

2.3 AS TÉCNICAS DE ANÁLISE DE AGRUPAMENTOS

Vários são os tipos de técnicas de agrupamento hierárquicas ou não hierárquicas encontradas na literatura (MARDIA et al., 1979; ANDERSON, 1984; KAUFMANN; ROSSEEUW, 1990; REIS, 2001; EVERITT, 2005; JOHNSON; WICHERN, 2007; FERREIRA et al., 2008; HAIR et al., 2010), tendo o pesquisador que tomar a decisão de qual é a mais adequada ao seu propósito, uma vez que as diferentes técnicas podem levar a diferentes soluções.

A escolha de uma técnica depende tanto dos tipos de dados disponíveis, quanto da aplicação desejada. Se a análise de agrupamento for usada como uma ferramenta para exploração dos dados, vários algoritmos podem ser executados sobre o mesmo conjunto de dados, a fim de avaliar os diferentes resultados de cada algoritmo e, dessa forma, comparar os resultados.

A classificação de técnicas de agrupamento não é uma tarefa direta ou canônica. A categorização, fornecida a seguir, baseia-se em categorização que vem sendo elaborada na literatura (JAIN et al., 1999; BENITEZ et al., 2010; MEIRELES et al., 2011). A divisão mais aceita é a classificação de técnicas hierárquicas ou não hierárquicas (partição) (MEIRELES et al., 2011).

2.3.1 Técnicas de hierarquização

Os métodos hierárquicos, como o próprio nome diz, envolvem a construção de uma hierarquia aglomerativa ou divisiva, em que as observações vão sendo combinadas passo a passo, e que não há um número predefinido de grupos que serão formados. Os resultados finais desses agrupamentos podem ser apresentados por uma árvore de classificação chamada de dendrograma, ilustrada na Figura 1. Esses agrupamentos podem ser utilizados tanto para agrupar parcelas, objetos, como para agrupar variáveis.

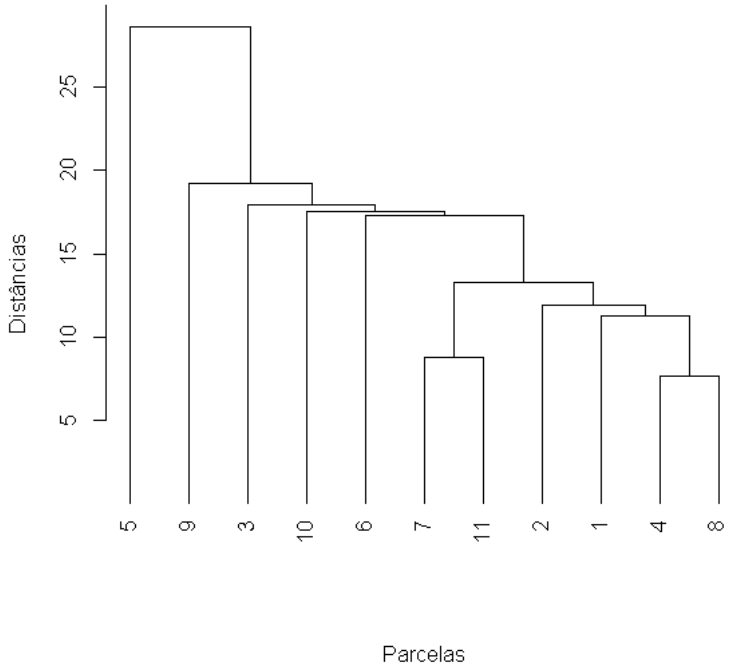

Figura 1. Dendrograma representando as sequências das fusões das parcelas, obtidas pelo emprego do método da ligação simples, com base na distância euclidiana.

Os métodos hierárquicos são numerosos e o pesquisador deverá decidir qual é o mais indicado ao seu trabalho, uma vez que as diversas técnicas podem levar a diferentes padrões de agrupamento.

Existem duas versões: a aglomerativa, que opera criando conjuntos a partir de parcelas isoladas; e a divisiva, que começa com um grande conjunto e vai quebrando-o em partes até chegar a parcelas isoladas.

Para Mingoti (2007), as técnicas hierárquicas aglomerativas partem do princípio de que, no início do processo de agrupamento, tem-se n grupos, ou seja, cada parcela do conjunto de dados observado é considerado como sendo um agrupamento isolado. Em cada passo do algoritmo, as parcelas amostrais vão sendo agrupados, formando novos agrupamentos até o momento no qual todas as parcelas consideradas inserem-se em único grupo. Portanto, no estágio inicial do processo de agrupamento, cada parcela amostral é considerada como um grupo de tamanho um e no último estágio de agrupamento, tem-se apenas um

único grupo constituído de todas as parcelas amostrais. Em termos de variabilidade, no estágio inicial, tem-se que a variância de cada agrupamento é igual a zero e, no estágio final, tem-se a maior dispersão interna possível. As principais etapas para a aplicação das técnicas hierárquicas aglomerativas podem ser resumidas da seguinte forma:

1. Cada parcela constitui um agrupamento de tamanho um. Portanto, tem-se n agrupamentos;
2. Os pares de agrupamentos mais "similares" são combinados e passam a constituir um único grupo. Apenas um novo grupo pode ser formado em cada passo. Dessa forma, em cada estágio do processo, o número de grupo vai sendo diminuído.
3. Em cada estágio da técnica, cada novo grupo formado é um agrupamento de grupo formado nos estágios anteriores. Se duas parcelas amostrais aparecem juntas em um mesmo grupo em algum estágio do processo de agrupamento, elas aparecerão juntas em todos os estágios subsequentes, ou seja, uma vez unidas, essas parcelas não poderão ser separadas.

Devido à propriedade de hierarquia, é possível construir um gráfico chamado de dendrograma ilustrado na Figura 1, que representa a "árvore" ou a história de agrupamento. A escolha do número final de grupos (k), em que o conjunto de parcelas deve ser repartido, é subjetiva. Existem alguns métodos que podem ser utilizados para auxiliar na determinação de k (índice Rand, índice Rand ajustado, silhueta, ÍIndice de Calinski e Harabasz, Índice de Hartigan, Índice de Duda e Hart, Índice C-Index, , Índice Gamma, Índice de Beale, Minimização do determinante de W, Maximização do traço de BW^{-1}, etc)

O propósito é encontrar o número k que esteja associado à "partição natural" das parcelas que estão sendo comparadas e agrupadas.

Na Figura 1, ilustra-se um exemplo de um dendrograma representando os grupos formados por meio das junções de 11 parcelas.

Os nós do dendrograma representam agrupamentos, e cada um deles é composto pelos grupos e/ou parcelas (grupos formados apenas por ele mesmo) ligados a eles (nós). Se cortar o dendrograma em um nível de distância desejado,

obtém-se um agrupamento dos números de grupos existentes nesse nível e das parcelas que os formam.

Os métodos hierárquicos aglomerativo ou divisivo não requerem o conhecimento, a priori, do número de grupos ou da partição inicial. Apresentam, no entanto, uma desvantagem, uma vez que uma parcela que foi designada a um grupo não pode ser realocada em um outro grupo (LATTIN et al., 2011).

Uma das vantagens dos algoritmos hierárquicos agromerativo ou divisivo é que esses permitem tratar facilmente quaisquer tipos de medidas de similaridade ou distâncias. Portanto, isso podem ser aplicáveis a quaisquer tipos de atributos.

Muitas vezes, são usados de forma exploratória, e a solução resultante é submetida a um método não hierárquico para refinar ainda mais a solução. Os dois métodos podem ser vistos como complementares ao invés de competidores.

A técnica hierárquica agromerativa ou divisiva se baseia na construção de uma matriz de (dis)similaridade ou distâncias em que cada elemento da matriz descreve o grau de diferença entre cada dois casos, com base nas variáveis escolhidas. Inicialmente, parte-se de n grupos de apenas uma parcela, que vai sendo agrupados sucessivamente, até que se encontre apenas um grupo que incluirá a totalidade dos n parcelas. O processo inverso é utilizado pelos métodos divisivos, isto é, o processo inicia com todas as parcelas formando um único grupo. Este grupo é particionado em dois bem distintos. Move-se para cada um dos grupos resultantes e reproduz-se o procedimento até cada parcela ficar isolada em um único agrupamento (BUSSAB et al., 1990). Observa-se que, na Figura 2, há distinção entre os métodos aglomerativo e o divisivo de Análise de Agrupamento, entretanto os pesquisados utilizam preferencialmente os hierárquicos aglomerativos.

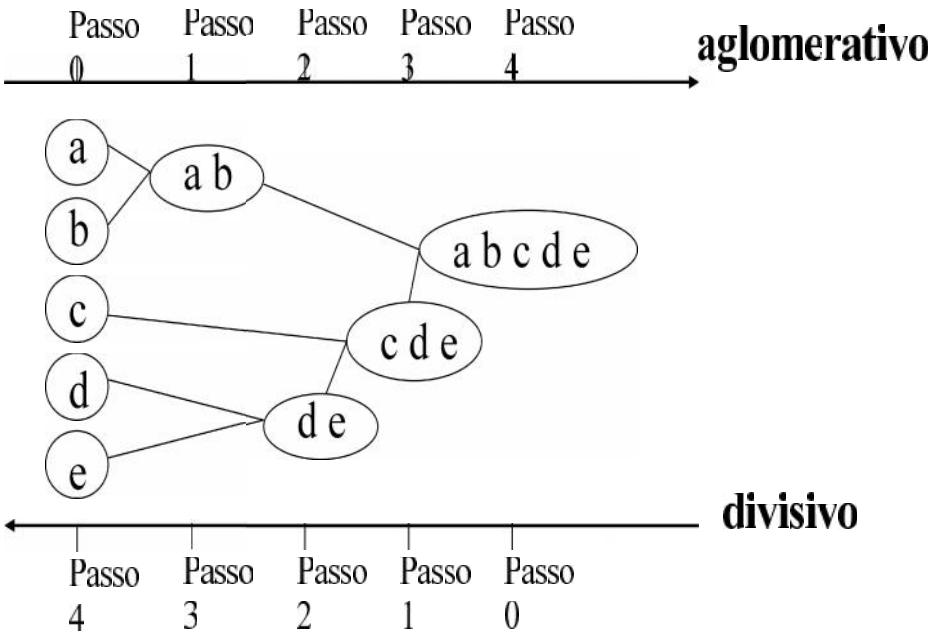

Figura 2. Distinção entre o método aglomerativo e o divisivo, Kaufman e Rousseuw (1990).

Métodos divisivos de agrupamento podem ser monotéticos ou politéticos. Os monotéticos usam um descritor (variáveis, espécies) único como base para a partição, visto que os modelos politéticos usam vários descritores que, na maioria dos casos, são combinados em uma associação de matrizes, antes do agrupamento. Os métodos divisivos monotéticos prosseguem escolhendo, para cada nível de partição, o descritor considerado melhor para esse nível, as parcelas são, em seguida, partidos, seguinte ao estado a que pertencem, em relação a esse descritor. Por exemplo, o descritor mais adequado a cada nível de partição poderá ser o que melhor representa a informação contida em todos os outros descritores, depois de medir a informação recíproca entre descritores. Quando uma única partição das parcelas é pedida, os métodos de produção do agrupamento monotético possuem uma única etapa (LEGENDRE; LEGENDRE, 1998).

Todavia, é possível construir métodos divisivos que não consideram todas as divisões.

Comparando os métodos aglomerativos com os divisivos, verifica-se que o método divisivo possui vantagem ao considerar, no primeiro estágio, muitas

divisões, diminuindo a probabilidade de uma decisão errada. Portanto, esse método torna-se mais seguro que o aglomerativo (KAUFMAN; ROUSSEEUW, 1990).

No entanto, apesar de os programas de computadores serem relativamente rápidos, os métodos hierárquicos podem não ser apropriados para analisar uma amostra muito grande, pois à medida que a amostra aumenta de tamanho, a necessidade de armazenamento de dados cresce dramaticamente. Por exemplo, uma amostra de 400 casos exige o armazenamento de 80.000 similaridades, e esse número cresce para 125.000, quando a amostra passa para 500 casos (CORRAR; PAULO; DIAS, 2007).

Os métodos hierárquicos aglomerativos (ligação simples, ligação completa, ligação média, da mediana, do centroide, do Ward, etc) podem ser confusos, porque combinações anteriores indesejáveis podem persistir no decorrer da análise, e levar a resultados artificiais.

O problema dos algoritmos hierárquicos é que, com eles, obtém-se apenas uma ordem de relacionamentos, e não grupos específicos. Para obter *k* grupos, basta apenas cortar as k - 1 arestas mais altas do dendrograma. Um exemplo disso pode ser visto na Figura 3.

Observa-se, na Figura 3, considerando corte 1, verifica-se a existência de nove grupos, sendo (5), (9), (3), (10), (7, 11), (6), (2), (1) e (4, 8). No corte 2, o número de grupos diminui para cinco, sendo (5), (9), (3, 10), (7, 11) e (6, 2, 1, 4, 8). Considerando o corte 3, o número de grupos diminui para três, sendo (1), (9, 3, 10) e (7, 11, 6, 2, 1, 4, 8). Dessa forma, o usuário deverá escolher o corte mais adequado às suas necessidades e à estrutura dos dados.

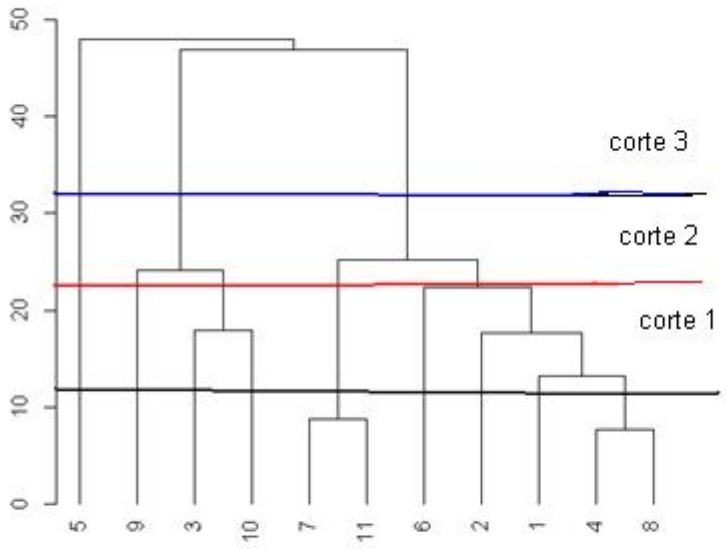

Figura 3. Exemplo no qual o dendrograma é cortado em três diferentes níveis.

2.3.2 Técnicas não hierárquicas ou por particionamento

As técnicas não hierárquicas têm alcançado crescente aceitabilidade e são aplicadas cada vez mais. Seu uso, entretanto, depende da habilidade do pesquisador, em selecionar os dados originais, de acordo com alguma base prática, objetiva ou teórica.

Os métodos não hierárquicos de agrupamento foram desenvolvidos para agrupar objetos, parcelas, ou itens, ao invés de variáveis, sobre os dados originais, em k grupos, que podem ser definidos antecipadamente, ou determinados durante a execução do procedimento (JOHNSON; WICHERN, 2007). Esses métodos exigem a pré-fixação de critérios que produzam medidas relativas à qualidade da partição produzida.

Para Lobão et al. (2010), os métodos não hierárquicos têm, como objetivo, encontrar diretamente uma partição de n elementos em k grupos (*clusters*), de modo que a partição satisfaça dois requisitos básicos: “coesão” interna (ou

“semelhança” interna) e isolamento (ou separação) dos grupos formados. Para se buscar a “melhor” partição de ordem k, algum critério de qualidade de partição deve ser empregado. É impossível, computacionalmente, criar todas as partições possíveis de ordem k e, a partir do conhecimento dessas partições, decidir qual será a mais adequada. Desse modo, são necessários processos que investiguem algumas das partições possíveis com o objetivo de encontrar a partição “quase ótima”.

Os métodos não hierárquicos diferem dos hierárquicos em vários aspectos. Primeiramente, requerem que o usuário especifique, a priori, o número de agrupamentos k desejado, onde o pesquisador pode optar por meio de algum conhecimento, pela conveniência, por simplicidade ou pelo método hierárquico, ao contrário das técnicas hierárquicas. Em cada estágio do agrupamento, os novos grupos podem ser formados por meio da divisão ou junção de grupos já combinados em passos anteriores. Nesse sentido, se em algum dos algoritmos, dois elementos tiverem sido colocados em um mesmo agrupamento, não necessariamente eles “estarão juntos” na partição final. Como consequência, não é mais possível a construção de dendrogramas. De forma geral, os algoritmos computacionais, utilizados nos métodos não hierárquicos, são do tipo iterativo e, em comparação com os métodos hierárquicos, têm uma maior capacidade de análise de conjunto de dados de maior porte, ou seja, com um grande número de observações.

Quando comparado com o método hierárquico, o não hierárquico é mais rápido, porque nele não é necessário calcular e armazenar, durante o processamento, a matriz de similaridade.

Segundo Hair et al. (2010), diferentemente dos métodos hierárquicos, os procedimentos não hierárquicos não envolvem o processo de construção em árvore. Em vez disso, designam objetos a agrupamentos, assim que os números de agregados a serem formados tenham sido especificados. O primeiro passo é selecionar uma semente de agrupamento como o centro inicial de um agregado; e todos os objetos (parcelas), selecionados dentro de uma distância de referência pré-especificada, são incluídos no agrupamento resultante. Em seguida, outra semente de agrupamento é escolhida, e o agrupamento continua até que todos os objetos tenham sido agrupados. Os objetos podem, assim, ser redesignados se

estiverem mais próximos de outro agregado do que daquele ao qual foram originalmente associados. Procedimentos de agrupamentos não hierárquicos frequentemente são chamados de agrupamentos de K médias, técnica de otimização, mistura de densidade, análise de agrupamento para dados estruturados, fuzzy, agrupamento com restrições, métodos em teria de grafos, método de partição para atributos mistos, etc.

De acordo com Lattin et al. (2011), procedimentos que produzem apenas uma solução de agrupamento, para um conjunto de pontos de agrupamentos. Ao invés de usar o processo de construção em forma de árvore, encontrado nos procedimentos hierárquicos, as sementes de agrupamentos são empregadas para reunir objetos, dentro de uma distância pré-especificada das sementes. Por exemplo, se quatro sementes de agrupamentos são especificadas, apenas quatro agrupamentos são formados. Os procedimentos não hierárquicos não produzem resultados para todos os possíveis números de agrupamentos, como ocorre com um procedimento hierárquico.

Os algoritmos não hierárquicos normalmente dependem de uma série de fatores que são determinados de forma arbitrária (subjetiva) pelo pesquisador, como número de grupos e as sementes de cada grupos. Isto pode causar impacto negativo na qualidade dos grupos gerados. Os algoritmos hierárquicos aglomerativos não são sujeitos a tais fatores, sendo totalmente determinísticos e independentes de parametrização (SCHULZ et al., 2003).

Um eventual problema é que essa condição enfatiza a questão da homogeneidade, e ignora a importante questão da boa separação dos agrupamentos. Isso pode causar uma má separação dos conjuntos, no caso de uma má inicialização dos centros, realizada de forma arbitrária (aleatória), no início do processo.

Outro ponto que pode afetar a qualidade dos resultados é a escolha do número de conjuntos por parte do usuário. Um número pequeno demais de conjuntos pode causar a junção de dois agrupamentos naturais, enquanto que um número grande demais pode fazer com que um agrupamento natural seja quebrado artificialmente em dois.

Uma das desvantagens dos métodos interativos é que não existe qualquer garantia da solução final ser ou não um ótimo agrupamento. Outra desvantagem é

requerer uma quantidade considerável de tempo de computação, uma vez que o modo mais lógico de fazer seria considerar todas as possíveis partições (k = 2, 3, 4, 5,...) e escolher a melhor de todas elas, o que torna necessária uma capacidade informática considerável, tendo em vista que é impossível tratar todas as partições possíveis do conjunto de dados k grupo pré-definidos (LATTIN, et al., 2011).

2.3.3 Método Incremental

Técnicas de análise de agrupamento, baseadas em distância, geralmente requerem que o número de grupos seja informado, a priori ou a posterior, e todas as parcelas sejam alocadas, ao menos, a um grupo. Algoritmos de Análise de Agrupamento e Incremental funciona, adicionando-se cada parcela apresentada ao grupo mais similar (KHALED; MOHAMED, 2003, tradução nossa). Quando a (dis)similaridade entre a parcela apresentada e as parcelas existentes não atende à área definida, um novo grupo é formado e a parcela é agrupado a esse grupo. O número de grupos não é um parâmetro predefinido, e sim o resultado da Análise de Agrupamento.

Os passos básicos utilizados pelo algoritmo incremental são:

1. Associar a primeira parcela a um grupo;
2. Pegar a próxima parcela e comparar com todos os grupos já existentes. A parcela deve ser associada a algum grupo, se atender o critério de associação, por exemplo, a distância média entre a parcela e o intervalo de ação do grupo. Se não atender o critério, ela deve ser associada a um novo grupo.
3. Repetir o passo 2 até que todas as parcelas tenham sido associadas.

Um dos motivos para que o método de agrupamento incremental fosse escolhido para esse trabalho é que ele revela como uma de suas características não ter como parâmetro de execução o número de grupos a ser gerado na análise de agrupamento, visto que o próprio algoritmo vai criando novos grupos, quando

necessário. Isso é muito útil quando não se conhece a amostra nem o grau de (dis)similaridade ou distância entre os objetos da mesma.

Essa técnica possui outras características e vantagens como: conhecer a distribuição e sua amostra, identificar o grau de (dis)similaridade ou distância dos objetos, identificar os agrupamentos, descobrir o número de grupos próximo do ideal, entre outros. Possui como desvantagem a dependência da ordem de entrada dos dados.

A maior vantagem do algoritmo de agrupamento método incremental é não precisar armazenar toda matriz de padrões na memória computacional (JAIN; MURTY; FLYNN, 1999). Sendo assim, o espaço necessário para a técnica ser executada é bem pequena. Geralmente, não é iterativo logo, o tempo para execução é menor do que nos métodos não hierárquico e hierárquico.

O método incremental pode ter diversas aplicações. Pode ser customizado para poder atender a algum problema específico. Por exemplo, para resolver o problema do agrupamento em conjuntos de objetos dinâmicos e manter os grupos de pequenos diâmetros quando novos objetos são inseridos. Para isso, desenvolveu-se algoritmo Incremental determinístico e randômico, baseado na análise dos requisitos da aplicação de recuperação da informação (MOSES et al., 2004).

Can (1993) apresentou um estudo sobre o método incremental de agrupamento de objetos textuais (documentos), onde os objetos são organizados automaticamente em grupos similares, facilitando sua localização, manipulação e análise.

Freitas e Prata (2007) utilizaram-se da distância euclidiana e dos algoritmos de agrupamentos, que combinam características dos métodos hierárquicos e não hierárquicos, com um conjunto de dados sobre a densidade de vegetação da Mata da Silvicultura, em parcelas. Castro et al. (2010) fizeram uma análise do método incremental e hierárquico.

3 DEFINIÇÃO DO NÚMERO DE GRUPOS

Determinação do número de grupos para uma base de dados é uma das tarefas mais difíceis no processamento de agrupamento.

Batista et al. (2011) afirmam que o número de grupos pode ser definido, a priori, por meio de algum conhecimento que se tenha sobre os dados, pela conveniência do pesquisador, por simplicidade, ou ainda pode ser definido, a posterior, com base nos resultados da análise, ou ainda pela experiência do pesquisador.

De acordo com Lattin et al. (2011), para determinar o número apropriado de grupos, existem diversas abordagens possíveis: em primeiro lugar, o pesquisador pode especificar antecipadamente o número de agrupamentos. Talvez, por motivos teóricos e lógicos, esse número seja conhecido. O pesquisador pode, também, ter razões práticas para estabelecer o número de agrupamentos, com base no uso que pretende fazer dele. Em segundo lugar, o pesquisador pode especificar o nível de agrupamento de acordo com um critério. Se o critério de agrupamento for de fácil interpretação, tal como a média de (dis)similaridade interna do agrupamento, é possível estabelecer certo nível que ditaria o número de agrupamentos. As distâncias entre os agrupamentos, em etapas sucessivas, podem servir de guia, e o pesquisador pode escolher interromper o processo, quando as distâncias excederem um valor estabelecido.

Uma terceira abordagem é representar, graficamente, a razão entre a variância total interna dos grupos e a variância entre os grupos em relação ao número de agrupamentos. O ponto em que surgir uma curva acentuada, um ponto de inflexão, poderá ser a indicação do número adequado de agrupamentos. Aumentar esse número, além desse ponto, seria inútil; e diminuí-lo, seria correr o risco de misturar parcelas diferentes.

Qualquer que seja a abordagem empregada, geralmente é aconselhável observar o padrão total de agrupamentos. Isso pode proporcionar uma medida da qualidade do processo de agrupamento e do número de agrupamentos que emergem nos diversos níveis do critério de agrupamento. De maneira geral, mais de um nível de agrupamento é relevante (BERTINI, 2010).

3.1 Validação dos agrupamentos

Para determinar se os grupos são significativos ou não, o resultado do agrupamento é validado para aferir a qualidade da solução encontrada. Vários trabalhos citam índices comumente empregados (MILIGAN; COOPER, 1985; YEUNG et al., 2001; JIANG; ZHANG, 2004; HAND et al., 2005; VANDRAMIN et al., 2008, 2009, 2010). Estes trabalham avaliam a qualidade dos grupos formados, com base na ideia de que, se eles refletirem a estrutura dos dados, então os índices de validação devem indicar um bom resultado.

Os trabalhos realizados, em validação, é geralmente desenvolvido no contexto de um dos três tipos diferentes de testes de validação: externos, internos e relativos:

- Os testes externos comparam um agrupamento ou parte dele com informação que não é usada para construir o agrupamento. Neste caso, o resultado do algoritmo é avaliado comparando-se com uma estrutura pré-definida que é imposta ao conjunto de dados, refletindo a estrutura real em grupos que se sabe ou que se pensa afetar os elementos do conjunto.
- Os testes internos comparam um agrupamento ou parte de um grupo com o conjunto de dados original, usando somente informação obtida a partir do processo do grupo, medindo-se essencialmente o desvio entre a estrutura gerada pelo algoritmo aplicado e os dados.
- Os testes relativos comparam várias estruturas de agrupamento do mesmo conjunto de dados, resultantes da aplicação do algoritmo com diferentes valores dos parâmetros de entrada ou ao conjunto de dados afetados de pequenas alterações mensuráveis.

Apresentaremos alguns índices de interesse dos pesquisadores. Esses índices podem ser utilizados tanto para medir a qualidade dos agrupamentos gerados quanto para compará-los.

3.1.1 Critérios de formação de grupos

Os critérios de formação de grupos que constituem um agrupamento que mais se destacam, são os critérios de formação de grupos usados na análise de uma matriz de dados contínuos, $X_{n\,x\,p}$, que usam a decomposição da matriz de dispersão T, dada por:

$$\mathrm{T} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (\mathrm{x_{ij}} - \bar{\mathrm{x}})(\mathrm{x_{ij}} - \bar{\mathrm{x}})^{\mathrm{T}}$$

em que $\mathrm{x_{ij}}$ é o vetor de dimensão p das observações do objeto i no grupo j e é o vetor de dimensão p das médias de cada variável.

$$\bar{x} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} x_{ij}$$

que é o vetor das médias das p variáveis nos n objetos

Esta matriz da variabilidade total pode ser decomposta em:

- matriz da dispersão dentro do grupo, W, definida por :

$$\mathrm{W} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (\mathrm{x_{ij}} - \bar{\mathrm{x}}_{\mathrm{j}})(\mathrm{x_{ij}} - \bar{\mathrm{x}}_{\mathrm{j}})^{\mathrm{T}}$$

em que $\bar{\mathrm{x}}_{\mathrm{j}}$ é o vetor de dimensão p das médias das variáveis dentro do grupo j .

- matriz da dispersão entre grupos, B, definida por :

$$\mathrm{B} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} \mathrm{n_j}(\bar{\mathrm{x}}_{\mathrm{j}} - \bar{\mathrm{x}})(\bar{\mathrm{x}}_{\mathrm{j}} - \bar{\mathrm{x}})^{\mathrm{T}}, \mathrm{com} \sum_{\mathrm{j}=1}^{\mathrm{k}} \mathrm{n_j} = \mathrm{n}$$

Então $T = B + W$

onde T, W e B são as matrizes associadas à variabilidade total dos dados, à variabilidade dentro dos grupos e à variabilidade entre os grupos, respectivamente.

Para dados univariados, p a equação $T = B + W$ representa a decomposição da soma total dos quadrados da variável em soma dos quadrados dentro dos grupos e a soma dos quadrados entre grupos, que é fundamental na análise de variância.

Como T é fixo, porque não depende do agrupamento que se realize, a melhor partição é aquela em que W é mínimo ou B máximo, isto é quanto maior a homogeneidade interna dos grupos maior é a separação entre os grupos.

3.1.2 Minimização do traço de W

No caso da análise multivariada, $p > 1$ generaliza-se o caso sugerido n análise univariada, embora o critério $T = B + W$, não seja tão claro como para $p = 1$.

Para determinar as três somas de quadrados acima referidas relativamente às p variáveis, necessitamos da soma dos elementos da diagonal principal destas matrizes. As três somas de quadrados são dadas por: $trT,\ trW,\ trB$.

A extensão óbvia no caso da análise multivariada, é a minimização da soma dos quadrados dentro dos grupos, que é equivalente a minimizar o traço da matriz W, trW, ou a maximizar o traço de B.

Minimizar o traço da matriz W é equivalente a minimizar a soma dos quadrados das distâncias euclidiana entre os objetos e as médias dos respectivos grupos,

$$\mathrm{E} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (\mathrm{x_{ij}} - \bar{\mathrm{x}}_{\mathrm{j}})(\mathrm{x_{ij}} - \bar{\mathrm{x}}_{\mathrm{j}})^{\mathrm{T}} = \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} \mathrm{d}_{\mathrm{ij,j}}^{2} = \sum_{\mathrm{l}=1}^{\mathrm{p}} \sum_{\mathrm{j}=1}^{\mathrm{k}} \sum_{\mathrm{i}=1}^{\mathrm{n_j}} (\mathrm{x_{ijl}} - \bar{\mathrm{x}}_{\mathrm{jl}})^{2}$$

em que

$d_{ij,j}$ é a distância euclidiana do objeto i do grupo j à média do grupo j. O critério pode também ser derivado do principio fundamental da matriz de distâncias:

$$E = \sum_{j=1}^{k} \frac{1}{2n} \sum_{i=1}^{n_j} \sum_{v=1, v \neq i}^{n_j} d_{ij,vj}^2$$

em que $d_{ij,vj}$ é a distância euclidiana entre o objeto i e o objeto v no grupo j . Assim, a minimização do traço de W é equivalente à minimização do critério de perda de homogeneidade para distâncias euclidiana usada por Ward no processo hierárquico para a formação de grupos**.**

3.1.3 Minimização do determinante de W

Na análise de variância múltipla, um teste para verificar se os vetores de médias são idênticos para os grupos considerados, é baseado na razão entre os determinantes da matriz da variabilidade total e da matriz da variabilidade dentro dos grupos $\frac{|T|}{|W|}$

Grandes valores de $\frac{\det(T)}{\det(W)}$ significa que os vetores de médias não são idênticos em todos os grupos.

Uma vez que para todas as partições de n objetos em K grupos, **T** permanece igual, a maximização de $\frac{\det(T)}{\det(W)}$, equivale a minimizar $\det(W)$.

Este critério foi estudado por Marriot, (1971).

3.1.4 Maximização do traço de BW^{-1}

Uma função usada, também, na análise de variância múltipla, é o $tr(BW^{-1})$ sendo B a matriz que representa a variabilidade entre os grupos e W, a

matriz que representa a variabilidade de dentro dos grupos. Grandes valores de $tr(BW^{-1})$ significa que os vetores de médias não são idênticos em todos os grupos. Baseando-nos neste critério, a melhor partição será a que corresponde ao máximo de $tr(BW^{-1})$. Quanto maior é o $tr(BW^{-1})$ e quanto menor$|W|$, maior é a diferença entre as médias dos grupos.

Propriedades do critério de agrupamento

Um dos critérios mais usados é o da minimização do traço da matriz W por ser simples e fácil de tratar computacionalmente. Consiste em minimizar a soma dos quadrados das distâncias euclidiana entre os objetos e os centroides dos respectivos grupos.

Apesar de ser o critério mais usado, este critério apresenta alguns inconvenientes, tais como:

- dependência da escala, ou seja, obtêm-se soluções diferentes com os mesmos dados estandardizados ou não estandardizados. É um grande inconveniente uma vez que o recurso à estandardização é muito frequente em análise de clusters;
- imposição de uma estrutura esférica, aos clusters observados, mesmo quando a estrutura “natural “ dos dados tem outra forma, porque nos cálculos só tem em conta os elementos da diagonal principal de W e B .

3.1.5 índice Friedman e Rubin

Friedman e Rubin, (1967), procuraram um critério alternativo à minimização do trW de tal forma que o resultado fosse independente da escala. Tal critério baseou-se na maximização de $\det(T) / \det(W)$ ou na maximização de $tr(BW^{-1})$ com a utilização dos valores próprios $\lambda_1, \lambda_2, \ldots, \lambda_p$, da matriz BW^{-1} :

$$traço(BW^{-1}) = \sum_{i=1}^{p} \lambda_l$$

$$\frac{\det(T)}{\det(W)} = \prod_{l=1}^{p}(1+\lambda_1)$$

Uma vez que os valores próprios da matriz BW^{-1} são os mesmos independentemente do fato desta matriz ser obtida da matriz original X ou a partir da matriz estandardizada, não são afetado pela escala.

O critério de minimização do tem sido o mais usado e não se restringe a grupos esféricos, ao contrário do critério do trW, que só identifica clusters esféricos. Estes critérios produzem grupos com o mesmo número de objetos, e o critério do determinante embora permita a formação de clusters elípticos, assume que todos os clusters têm a mesma forma; o que poderá, como é evidente causar alguns problemas. Por isso serão necessários outros critérios, abordados no ponto seguinte.

3.1.6 Coeficiente R^2

Quanto maior for o valor de R^2, maior será a soma de quadrados entre grupos SQE e menor será o valor da soma de quadrados residual SQD. Assim, o seguinte procedimento pode ser adotado como critério para a escolha do número k de agrupamento final da partição: faça um gráfico do tipo passo do agrupamento versus R^2. Neste caso, procure detectar se há algum "ponto de salto" relativamente grande em relação aos demais (REIS, 2001). Este ponto indica o "momento ideal de parada" do algoritmo de agrupamento. A função neste gráfico é sempre decrescente, pois, quanto maior for o valor de k, menor será a variabilidade interna dos grupos e, consequentemente, maior será o valor de R^2.

Root-mean square standard deviation de um novo grupo mede a homogeneidade dos grupos formados em cada passo, onde cada um deverá ser menor do que no passo anterior. R^2 mede a perda de homogeneidade, devido à reunião de dois grupos (caso o valor do índice seja zero, então o novo grupo é obtido pela reunião de dois grupos perfeitamente homogêneos). R^2 mede também o grau de diferença existente entre grupos, onde um valor zero indica que não

existe diferença entre grupos, enquanto que o valor 1 indica diferenças significativas entre eles, e a distância entre grupos mede a distância entre dois grupos que são reunidas num determinado passo (SILVA, 2005).

3.1.7 Estatística Pseudo F

Esta estatística é chamada de Pseudo F. Segundo Calinski e Harabasz (1974), se F é monotonicamente crescente com k, os dados sugerem que não existe qualquer estrutura “natural” de partição dos dados. Se, entretanto, isso não ocorrer e a função F apresentar um valor máximo, corresponderão à “partição ideal” dos dados. Pode-se mostrar que a estatística em F tem distribuição F com p($\bar{X}_{i1}$1) e p(n - k*) graus de liberdade, e que as n parcelas amostrais constituem-se uma amostra aleatória não ocorre, uma vez que a partição das parcelas amostrais é feita através de métodos de agrupamento com critério de similaridade previamente definidos por métricas matemáticas. Apesar desse fato, a ideia por detrás deste critério é bastante interessante. É como se em cada passo do algoritmo de agrupamento estivesse sendo feito um teste F de análise de variância, comparando-se, os vetores de médias dos grupos que foram formados no respectivo passo. Busca-se o maior valor de Pseud F, ou seja, aquele que estaria relacionado-se com menor probabilidade de significância do teste e, consequentemente, estaria rejeitando a igualdade de vetores de médias populacionais com maior significância, resultando, desse modo, na partição com maior heterogeneidade entre grupos (MINGOTI, 2007).

3.1.8 Lambda de Wilks

Lambda de Wilks é um teste estatístico usado em análise multivariada de variância (MANOVA), para testar se há diferenças entre as médias dos grupos identificados de matérias numa combinação de variáveis dependentes. Testa se o

escore médio de dois grupos, grupo I e grupo II, é o mesmo através de oito simultâneas, assim, eles estão considerando oito variáveis dependentes e comparando a média dessa combinação para dois grupos.

Lambda de Wilks faz, na configuração multivariada, com a combinação de variáveis dependentes, o mesmo papel que F-testa faz em análise de único modo de variância. Lambda de Wilks é uma medida direta da proporção da variância na combinação de variáveis dependentes que não é contada pela variável independente (o grupo ou fator). Se uma grande proporção da variância é contada pela variável independente, então ela sugere que há um efeito do agrupamento variável e que o grupo (nesse caso os grupos I e grupos II) tem diferentes valores médios (EVERITT, 2005).

Para a análise de variância simples, resulta do quociente entre os determinantes das matrizes de somas de quadrados e produtos cruzados dentro dos grupos e total.

O determinante de W é uma medida da variabilidade dentro dos grupos enquanto que o determinante de T nos dá uma medida da variabilidade total. Quanto maior for a semelhança entre os dois determinantes, menores serão as diferenças entre grupos B e mais o valor Lambda de Wilks se aproximará de 1. Pelo contrário, se as diferenças entre os grupos forem elevadas (heterogêneo), quando comparadas com a variabilidade dentro dos grupos, o valor de Lambda de Wilks tenderá a aproximar-se de 0. Assim, a estatística Λ de Wilks é uma medida inversa do grau de diferenciação entre os grupos: quanto menor o seu valor, maior esse grau de diferenciação (REIS, 2001).

3.1.9 Índice de Rand

O índice de Rand permite comparar duas partições com o número de grupos não necessariamente iguais. Basicamente, este índice baseia-se no número de pares de parcelas que foram atribuídos, da mesma maneira, em cada uma das partições, ou seja, baseia-se no número de pares de parcelas concordantes (tipo I e II). Assim, temos o índice de Rand, designado por IR é dado por (RAND, 1971):

$$IR = \frac{a+d}{a+b+c+d} = \frac{A}{\binom{n}{2}}$$

De um modo mais detalhada tem-se:

$$IR = \frac{\binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left[\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right]}{\binom{n}{2}}$$

Verifica-se $0 \leq IR \leq 1$, tomando o valor 0 quando as duas partições não têm qualquer semelhança (ou seja, quando uma partição é constituída por um só grupo com todos as parcelas, e a outra é constituída por n grupos com 1 parcela cada) e o valor 1 quando o acordo entre as duas partições é completo.

3.1.10 Índice de Rand Ajustado

Vejamos então como é feita a correção da estatística de Rand para o acaso, associando-lhe uma interpretação probabilística, ou seja, tomando em conta o contributo do acaso no valor do índice que terá de ser um valor constante (ex. zero). (HUBERT; ARABIE, 1985).

Sob o pressuposto hipergeométrico para as somas marginais da tabela de contingência, mostra-se que o valor esperado do índice de Rand é dado pela seguinte expressão:

$$E(IR) = 1 + 2\frac{\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\binom{n}{2}} - \frac{\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}}{\binom{n}{2}}$$

Usando a fórmula geral de um índice I corrigido para o acaso:

$$\frac{I - E(I)}{I_{\max} - E(I)}$$

que é limitado por 1 e toma o valor zero quando o índice iguala o seu valor esperado, o Índice de Rand Corrigido de Hubert e Arabie (IRC) é dado por (HUBERT; ARABIE, 1985).

$$IRC = \frac{\sum_{i=1}^{R}\sum_{j=1}^{C}\binom{n_{ij}}{2} - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{R}\binom{n_{i.}}{2} + \sum_{j=1}^{C}\binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1}\sum_{i=1}^{R}\binom{n_{i.}}{2}\sum_{j=1}^{C}\binom{n_{.j}}{2}}$$

Este índice toma valores em (−1, 1), onde o valor 1 indica um perfeito acordo entre as duas partições, enquanto que valores próximos de 0 correspondem a um acordo entre as partições devido ao acaso.

3.1.11 Estatística Silhouette de Kaufman & Rousseeuw

A estatística Silhouette de Kaufman e Rousseeuw proporciona um método gráfico de verificação da densidade e isolamento de grupos, diferenciando entre elementos do centro e da fronteira do grupo. A estatística é dada por (ROUSSEEUW, 1987; KAUFMAN; ROUSSEEUW, 1990).

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}$$

onde a(x) é a dessemelhança média do elemento x a todos os outros elementos pertencentes ao mesmo grupo, b(x) é a dessemelhança média do elemento x a todos os elementos pertencentes ao grupo que lhe está mais próxima, ou seja, o grupo que minimiza esta média. Verifica-se $-1 \leq s(x) \leq 1$, sendo os valores negativos indicativos de que os elementos são similares aos membros de outros grupos, e valores perto de 1 indicam que os elementos pertencem fortemente à o grupo na qual foram colocados, indicando que o elemento foi bem agrupado. O valor indicativo do número de grupos no conjunto de dados, é dado pelo valor que maximiza a média de s(x) para todos os elementos do conjunto de dados.

3.1.12 Gráfico da silhueta

O gráfico da silhueta Kaufman e Rousseeuw (1990) é um procedimento descritivo para verificar a qualidade dos agrupamentos formados. A ideia do método é verificar se uma parcela está mais próxima dos elementos do seu próprio grupo ou de elementos de grupos vizinhos. Ele baseia-se no cálculo de duas medidas: $a(i)$ a distância média entre a parcela i e os elementos do grupo e $b(i)$, a distância média entre a parcela i e os elementos do grupo mais próximo do de i, que não seja o seu próprio grupo (BARROSO; ARTES, 2003).

Seja G(i) o grupo que contém a parcela i, admita a existência de $n_{G(i)}$ observações neste grupo. Temos então que

$$a(i) = \frac{\sum_{JCG(i)} d_{ij}}{n_{G(i)} - 1}$$

Onde d_{ij} é a distância euclidiana entre as parcelas i e j.

Para cada grupo diferente de $G(i)$, determine a distância média entre seus elementos e i. Defina o grupo $H(i)$ como o de menor distância média entre seus elementos e a ponta i, admita que a cardinalidade de $H(i)$ seja $n_{H(i)}$. O grupo $H(i)$ é denominado vizinho de i. Assim, temos

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

Essa medida reflete quão adequada a alocação de i em seu grupo. Note que $s(i)$ é um número que varia entre -1 e 1. Valores próximos de 1 indicam boa alocação da parcela, uma vez que, nesse caso, $b(i) > a(i)$; por outro lado, valores negativos sugerem uma má alocação, uma vez que a parcela está em média mais próxima dos elementos do grupo vizinho do que de seu próprio grupo.

3.1.13 Índice Kappa de Cohen

Um dos primeiros índices que surgiram na literatura usando pares de elementos concordantes foi o índice Kappa de Cohen, aplicado no caso particular em que R = C (mesmo número de grupos nas duas partições), e é dado por (COHEN, 1968; HUBERT, 1977)

$$K = \frac{P_o - P_e}{1 - P_e}$$

onde

$$P_o = \sum_{i=1}^{R} \frac{n_{ii}}{n}$$

é a proporção observada de pares concordantes e

$$P_e = \sum_{i=1}^{R} \frac{n_{i.} n_{.i}}{n^2}$$

é a proporção esperada de pares concordantes devido ao acaso e sob a hipótese de independência entre partições.

3.1.14 Índice de Jaccard

O índice de Jaccard É dado por

$$j = \frac{a}{a + b + c}$$

Este índice ignora os casos de tipo II, ou seja, não considera os casos em que os pares de elementos são classificados em grupos diferentes em ambas as

partições. Em Milligan (1980), na realização pode algumas experiências, verifica-se que o índice de Rand e de Jaccard apresentaram uma elevada correlação (0.937), mostrando que de fato, os pares de tipo II podem não ter uma grande influência na comparação de partições.

3.1.15 Índice Dunn

Assim como o índice VRC e o da Silhueta, o índice de Dunn (DUNN, 1973) foi desenvolvido para privilegiar grupos compactos e bem separados. Para compreendê-lo, seja o diâmetro do grupo $\Delta(.)$ definido como

$$\Delta(C_i) = {}_{O_a} {}^{max}_{O_b \in C_i} \{d(O_a, O_b)\}$$

em que $d(.,.)$ é uma medida de distância (dissimilaridade) qualquer e C_i é um grupo qualquer.

Seja a distância entre grupos $\delta_{(C_i,C_j)} = {}_{O_a} {}^{max}_{O_b \in C_i} \{d(O_a, O_b)\}$ em que C_i, C_j são dois grupos quaisquer. O índice de Dunn é definido como

$$Índice\ de\ Dunn = \min_{1 \le i \le k} \left\{ \min_{1 \le i \le k, j \ne i} \left\{ \frac{\delta_{(C_i,C_j)}}{\max_{1 \le i \le k} \{\Delta(C_i)\}} \right\} \right\}$$

3.1.16 Índice de Validação SD

O Índice de validação SD é proposto em Halkidi et al. (2000) e a sua definição é baseada nos conceitos de dispersão média dos grupos e separação total entre grupos, definidas usando dessemelhanças intra-grupos e entre grupos. Basicamente, este índice seleciona a partição cuja densidade de pontos dentro dos grupos é muito maior do que a densidade de pontos entre grupos. Assim, o índice pretende identificar grupos compactas e isoladas. Consideremos = {x_1, . . . , x_n} o conjunto de n elementos descritos por t atributos numéricos, e C_1, . . . ,C_k uma partição em k grupos de n_1, . . . , n_k elementos e r_1, . . . , r_k representantes, respectivamente. Represente-se por $\varsigma(\Omega)$ a dispersão do conjunto de dados,

definida para a t^ésima coordenada, por $\varsigma_t(\Omega) = \frac{1}{n}\sum_{j=1}^{n}(x_{jt} - \bar{x}_t)^2$ onde $\bar{x}_t$ é a t^ésima coordenada de $\bar{X} = \frac{1}{n}\sum_{j=1}^{n} x_j,\ \forall_{x_j} \in \Omega$. Represente-se por $\varsigma(C_i)$ a dispersão do grupo C_i e defina-se, para a t^ésima coordenada, $\varsigma_t(C_i) = \frac{1}{n_i}\sum_{j \in A_i}(x_{jt} - \bar{x}_t)^2$ onde r_i é o representante da classe C_i e A_i é o conjunto dos índices dos elementos pertencentes ao grupo C_i.

O índice tem uma ordem de grandeza linear, e é definido por:

$$SD(k) = \gamma Scat(k) + Dis(k)$$

onde

$$Scat(k) = \frac{1}{k}\sum_{i=1}^{k}\frac{\|\varsigma(C_i)\|2}{\|\varsigma(C)\|2}(1)$$

e Dis(k) é a separação total entre grupos, dada por

$$Dis(k) = \frac{D_{max}}{D_{min}}\sum_{i=1}^{k}\left(\sum_{j=1}^{k}\|r_i - r_j\|2\right)^{-1}$$

onde

$$D_{max} = max\{\|r_i - r_j\|2\}\ \forall_{i,j} \in \{1, \dots, k\}\ e\ D_{min} = min\{\|r_i - r_j\|2\}\ \forall_{i,j} \in \{1, \dots, k\}$$

são respectivamente a distância máxima e mínima entre representantes dos grupos. O fator γ é um fator de balanço e é igual à $Dis_{(kmax)}$ quando k_{max} é o número máximo de grupos. A influência do número máximo de grupos relacionado com o fator de balanço é discutido em (HALKIDI, al et., 2001) sendo provado que o valor indicativo do número de grupos é obtido quase independentemente do valor máximo de grupos. O número de grupos presentes no conjunto de dados é aquele que minimiza o valor do índice.

3.1.17 Índice de Beale

Este índice aplica-se a métodos hierárquicos descendentes, e usa um F-ratio (razão) para testar a hipótese da existência de k_2 contra k_1 classes no conjunto de dados $k_2 > k_1$. O F-ratio avalia o aumento da média dos quadrados dos desvios dos valores dos elementos dos grupos em relação aos centroides, quando se vai de k_2 para k_1 grupos, contra a média dos quadrados dos respectivos desvios da partição com k_2 grupos. As tabelas F podem ser usadas com $t(k_2 - k_1)$ e $t(n - k_2)$ graus de liberdade, onde t é o número de atributos e n é o número de elementos no conjunto de dados. Em cada nível da hierarquia é feito um teste de duas contra um grupo, continuando o agrupamento até que a hipótese de um grupo seja rejeitada. A estatística de teste para decidir se um grupo deve ser dividido é dado por:

$$F = \frac{\frac{w_1 - w_2}{w_2}}{\frac{m-1}{m-2} 2^{\frac{2}{t}} - 1}$$

onde w_1, w_2, m e t são definidos como para o índice anterior) com uma distribuição $F_{t,(k-2)t}$, sob a hipótese nula, admitindo multinormalidade dos atributos. Rejeita-se a hipótese de uma classe única para valores significativamente grandes de F.

3.1.18 Índice de Fowlkes e Mallows

Segundo Fowlkes e Mallows (1983) é apresentada uma medida numérica do grau de semelhança entre agrupamento hierárquicas representada por B_k, sendo calculados a partir das partições obtidas pelo corte das árvores representativas dos agrupamentos hierárquicas, num nível tal que origine k grupos em cada um. Esta medida é aplicada a partições sem se ter em conta a estrutura

hierárquica em cada uma dos agrupamentos, por exemplo o nível onde se cortou cada uma das árvores é ignorado. Para este índice é calculada a média e a variância sob o pressuposto de que as somas marginais da matriz (n_{ij}) são constantes, ou seja, considera-se constante o número de elementos em cada uma dos grupos de cada uma dos agrupamentos. Através de algumas experiências baseadas no método Monte Carlo, compara-se B_k com o Índice de Baker [9] e com o Índice de Rand (RAND, 1971). Nesta secção vamos considerar o caso particular em que na tabela de contingência se tem R = S = k (apesar de esta restrição não ser necessária segundo Wallace (1983)). A medida proposta é dada por (FOWLKES; MALLOWS, 1983)

$$B_k = \frac{T_k}{\sqrt{P_k Q_k}}$$

$$T_k = \sum_{i=1}^{k}\sum_{j=1}^{k}\binom{n_{ij}}{2} = \sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}^2 - n$$

$$P_k = \sum_{i=1}^{k}\binom{n_{i.}}{2} = \sum_{i=1}^{k} n_{i.}^2 - n$$

$$Q_k = \sum_{i=1}^{k}\binom{n_{.j}}{2} = \sum_{i=1}^{k} n_{.j}^2 - n$$

$$n_{i.} = \sum_{j=1}^{k} n_{ij} \;;\; n_{.j} = \sum_{j=1}^{k} n_{ij} \;;\quad n.. = n = \sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij}$$

Verifica-se $0 \leq B_k \leq 1$, existindo uma correspondência direta entre valores altos do índice e um acordo entre as duas partições a serem comparadas. Sob o pressuposto hipergeométrico para as somas marginais da tabela de contingência, os autores deduzem as expressões para a média e a variância de B_k.

O índice de Fowlkes e Mallows também pode ser dado por

$$B_k = \frac{a}{\sqrt{(a+b)(a+c)}}$$

3.1.19 Goodman & Kruskal (G2)

Segundo Gordon (1999) este índice é muito utilizado em estudos de agrupamento. Após obtenção dos grupos são feitas comparações entre as dissemelhanças intragrupos e intergrupos. A comparação diz-se concordante (respectivamente discordante) se a dissemelhança dentro dos grupos é mais muito mais baixa (resp. elevada) que a dissemelhança intergrupos. O índice é então dado por:

$$G_2 = \frac{s(+) - s(-)}{s(+) + s(-)}$$

onde $s(+)$ e $s(-)$ representam o número de pares concordantes e discordantes respectivamente, envolvendo os valores das matrizes de dissemelhança e ultramétricas. O valor máximo desta medida indica o número de grupos a reter.

4 MEDIDAS DE DISSIMILARIDADE E SIMILARIDADE

Muitos problemas multivariados podem ser vistos em termos de distâncias entre observações individuais, entre amostras de observações ou entre populações de observações. Um grande número de medidas de similaridade e/ou dissimilaridade tem sido proposto e utilizado em Análise de Agrupamento (MEYER et al., 2004)

Com base na medida de similaridade ou dissimilaridade, as parcelas mais próximas são agrupadas e os demais são colocados em grupos separados (LINDEN, 2009).

Na matriz de distâncias, em cada célula, está presente o valor do coeficiente calculado para as parcelas posicionadas nas respectivas linha e coluna. Esse valor representa uma medida de distância entre essas duais parcelas e, dependendo da medida que foi escolhida, essa distância é considerada uma distância, como aquelas recomendadas por Johnson e Wichen (2007).

Estudos de (dis)similaridade atendem a determinados objetivos do pesquisador, por propiciarem informações acerca do grau de semelhança ou de diferença entre duas ou mais parcelas. Entretanto, o número de estimativas de (dis)similaridade obtido é relativamente elevado, quando se tem grande número de parcelas, o que torna, às vezes, impraticável o reconhecimento de grupos homogêneos por um simples exame visual. Portanto, o uso de métodos que agrupem as parcelas pode ser uma das melhores alternativas para análise e interpretação dos dados (BAFFETTA et al., 2011).

De modo geral, as medidas de similaridade e de dissimilaridade são interrelacionadas e facilmente transformáveis entre si (FERREIRA et al., 2008). Há um grande número de coeficientes de similaridade e/ou de dissimilaridade para caracteres binários disponíveis na literatura. Cargnelutti et al. (2009), afirmam tais coeficientes podem ser, convertidos, com facilidade em coeficientes de dissimilaridade. Se a similaridade for denominada s, a medida de dissimilaridade será o seu complementar, desde que sejam verificadas as propriedades das medidas de (dis)similaridade.

$$d_{ij} = 1 - s, \quad d_{ij} = \sqrt{1 - s} \ \ ou \ \ d_{ij} = 1 - s^2, \quad d_{ij} = \sqrt{1 - s^2}$$

A maioria dos métodos de análise de agrupamento requer uma medida de similaridade ou dissimilaridade entre as parcelas a serem agrupadas, normalmente expressos como uma função distância ou métrica (DALIRSEFAT; MEYER, 2009).

4.1 Estrutura de dados no agrupamento de dados

Para que os algoritmos de agrupamentos possam fazer a partição dos dados, é necessário que utilizem estruturas de dados, capazes de armazenar as parcelas a serem processadas ou as informações concernentes às relações entre elas (SIQUEIRA et al. 2010).

A maior parte dos algoritmos de agrupamento opera tipicamente com as duas seguintes estruturas de dados.

Matriz de dados (ou estrutura de parcelas por atributos): esta estrutura apresenta n parcelas, tais como árvores, com p atributos, e variáveis dendrométricas. A estrutura se apresenta na forma de uma tabela relacional, ou n x p matriz (n parcelas x p atributos).

$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ x_{n1} & \cdots & x_{nf} & \cdots & x_{np} \end{bmatrix}$$

Matriz de proximidade (também chamada de matriz de dissimilaridades, matriz de distâncias ou matriz de similaridades): Esta estrutura armazena uma coleção de proximidade que estão disponíveis para todos os pares de n parcelas. De maneira geral, é representada por uma matriz simétrica n x n:

$$\begin{bmatrix} 0 & & & & \\ d(2,1) & 0 & & & \\ d(3,1) & d(3,2) & 0 & & \\ : & : & : & & \\ d(n,1) & d(n,2) & \ldots & \ldots & 0 \end{bmatrix}$$

Em que: d_{ij} é a diferença ou dissimilaridade medida entre as parcelas i e j. Em geral, d_{ij} é um número não negativo que é perto de zero, quando as parcelas i e j são muito similares, e torna-se maior quanto maior for a diferença entre as indivíduos, itens, parcelas conforme pode ser visto na matriz $d_{ij} = d_{ij}$ e $d_{ii} = 0$.

A matriz de dados é geralmente chamada matriz de dois modos, enquanto a matriz de dissimilaridades é chamada de matriz de um modo, uma vez que as linhas e as colunas da matriz de dados representam diferentes entidades, enquanto que as da matriz de dissimilaridades representam a mesma entidade. Se os dados são apresentados no formato da matriz de dados, podem ser primeiro transformados em uma matriz de dissimilaridades, antes da aplicação dos algoritmos de agrupamento (JAIN et al., 1999; BERTINI et al., 2010).

4.2 Tipos de Dados

Na análise de agrupamento ter-se-á que lidar com atributos de diferentes tipos. Os dados podem ser descritos por atributos exclusivamente contínuos, discretos ou categóricos ou, eventualmente, por combinações desses dois tipos

de características (DUARTE, 2008). As características categóricas poderão ser binários, nominais ou ordinais. Tem-se uma descrição de cada um desses tipos de características:

Características contínuas: são medidas de uma escala contínua, a saber:

- O conjunto das alturas das árvores de uma floresta, já que no intervalo de 1,30m a 15m a variável pode assumir uma infinidade de valores, considerando-e que existem infinitas medidas entre 1,30 e 15m;
- O conjunto dos volumes individuais das árvores ocas de uma floresta.

Características discretas: admitem somente números inteiros, conforme se pode verificar:

1. Conjunto de árvores com qualidade de fuste 1;
2. Conjunto de árvores com sapopema;
3. Proporção de árvores sadias.

Características binários: estes atributos possuem apenas dois estados: 0 e 1, em que 0 significa que o atributo está ausente, e 1 significa que o atributo está presente; frequência de uma determinada espécie de árvore.

Atributos nominais: são generalizações das características binárias, visto que podem tomar mais do que dois estados. Considere-se *M* o número de estados de uma característica nominal. Os estados podem ser constituídos por letras, símbolos, ou por um conjunto de inteiros (1, 2,…, *M*). Tais inteiros são usados apenas para tratamento de dados e não representam qualquer ordenamento específico.

A qualidade fuste pode ser uma característica nominal que pode ter estados: reto, semirreto e curvo;

Sucupira amarela, Sucupira vermelha e/ou Sucupira preta;
Classificação de solos (argiloso e arenoso).

Características ordinais: podem ser discretos ou contínuos e são muito úteis para registrar avaliações subjetivas de qualidades que não podem ser medidas objetivamente.

1 = madeira leve;
2 = madeira moderada;
3 = madeira pesada.

Os métodos de agrupamento exigem que os coeficientes respeitem as propriedades métricas, como as que são apresentadas em Bertini et al. (2010), Seja M um conjunto, uma métrica em M é uma função d: M x M $\rightarrow \Re$, tal que para quaisquer i, j, z $\in$ M, tenha-se:

d (i, j) = d (j, i) (simétrica);
d (i, j) > 0, se i$\neq$ j;
d (i, j) = 0, se e somente se, i = j; e
d (i, j) $\leq$ d (i, z) + d (z, j) (desigualdade triangular).

Além disso, espera-se que d_{ij} aumente, quando a dissimilaridade entre i e j também aumente. Se, além de todas as propriedades citadas acima, a métrica também possui a propriedade d(ax, ay)= |a|d(x, y), esta última se torna uma norma.

5. MEDIDAS DE DISTÂNCIAS

Um conceito intrínseco da análise de agrupamento é a (dis)similaridade, ou seja, o conjunto de regras que servem como critério para agrupar ou separar parcelas. Um dos elementos para indicar dissimilaridade mais conhecido é a distância entre as parcelas.

Existem várias medidas que podem ser utilizadas como medidas de distâncias ou dissimilaridade entre parcelas de uma matriz de dados. Cormack, (1971) descreveu uma série de medidas possíveis: distância euclidiana, euclidiana quadrada, euclidiana média e euclidiana padronizada, distância euclidiana ponderada, distância Mahalanobis, distância corda, distância de Nei, distância absoluta ou *City – Block Metric*, distância de Minkowski, distância de Chebychev.

5.1 Distância euclidiana

De acordo com Hair et al. (2010), este é o coeficiente de dissimilaridade mais conhecido e utilizado para indicar a proximidade entre parcelas. É simplesmente a distância geométrica entre duas parcelas em um espaço multidimensional. A ideia básica é considerar cada observação como um ponto em um espaço euclidiano e, desse modo, calcular o coeficiente que representará a distância física entre os pontos. A distância euclidiana é frequentemente usada para avaliar a proximidade entre parcelas, em um espaço bidimensional ou tridimensional, e funciona bem para conjuntos de dados com grupos compactos e isolados.

Segundo Baraloto et al. (2010), a distância euclidiana e suas derivadas podem ser calculadas tanto para os dados puros ou crus, quanto para os dados padronizados. Essas distâncias possuem certas vantagens, a saber: apresentam simplicidade de cálculo; a distância entre quaisquer duas parcelas não é afetada pela inserção de novas parcelas na análise, mesmo que os novos valores sejam

classificados como valores atípicos; encontram-se associadas intuitivamente ao conceito usual de distância. Contudo, essas distâncias podem ser bastante afetadas pelas diferenças de escala associadas às dimensões, a partir das quais as distâncias são computadas, o que implicaria a necessidade de transformar os dados, antes da aplicação dessa medida.

Observa-se que, enquanto a distância euclidiana nos fornece o caminho mais curto entre dois pontos quaisquer do plano, medindo o segmento de reta que os une, a distância d_1 representa a soma da medida dos catetos do triângulo formado pelos pontos (x_1, y_1), (x_1, y_2) e (x_2, y_2), isto é, " contorna o quarteirão " como faria um motorista de táxi para ir do ponto *A* de coordenadas (x_1, y_1) até um ponto *B* de coordenadas (x_2, y_2), conforme observa-se na Figura 4.

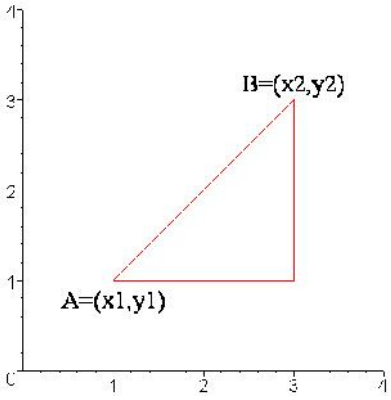

Figura 4. Distância euclidiana entre as árvores A e B no plano

Após escolher uma função para medir distâncias, pode-se definir a circunferência como o lugar geométrico dos pontos que equidistam de um ponto fixo *C*. O ponto fixo é chamado centro da circunferência e a distância de qualquer dos seus pontos ao centro é o raio dessa circunferência.

Mcroberts et al. (2007) afirmam que a distância entre duas observações (i e j) corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações de i e j para todas as variáveis (v = 1, 2,... p).

$$d_{ij} = \sqrt{\sum_{v=1}^{p} (X_{iv} - X_{jv})^2}$$

Quando existem apenas duas informações, essa expressão se torna a medida da hipotenusa de um triângulo retângulo.

5.2 Distância euclidiana quadrada

A distância entre duas parcelas (i e j) é definida como o somatório dos quadrados das diferenças entre os valores de i e j para todas as p variáveis.

$$d_{ij}^2 = \sum_{v=1}^{p} (X_{ij} - X_{jv})^2$$

Ao elevar a distância euclidiana ao quadrado, é gerada a distância euclidiana quadrática, sendo essas medidas bastante influenciadas por aquelas parcelas que se encontram mais distantes.

A distância euclidiana quadrada tem a vantagem de que não é necessário calcular a raiz quadrada, o que acelera sensivelmente o tempo de computação, além de ser a distância recomendada para os métodos de agrupamentos centroide e Ward (HAIR et al., 2010).

5.3 Distância euclidiana média (chamada distância de Penrose)

Para Messeti (2007), outra medida derivada da distância euclidiana muito utilizada é a distância euclidiana média, expressa pela raiz quadrada da divisão entre o somatório do quadrado das diferenças, pelo número de variáveis envolvidas.

Pode-se observar que o valor da distância euclidiana aumenta, quando novas variáveis são incorporadas às originais. Uma maneira possível de contornar esse problema é dividir esse valor pela raiz quadrada do número de caracteres, isto é:

$$\Delta_{ij} = \frac{1}{\sqrt{p}} d_{ij}$$

Essa distância é apenas um reescalonamento da distância anterior, possuindo as mesmas propriedades e, portanto, produzindo os mesmos resultados, caso seja submetida às técnicas de análise de agrupamentos. Esse coeficiente possui uma propriedade interessante, uma vez que garante essa a possibilidade de essa distância ser utilizada na ausência de dados para algumas coordenadas (“*missing values”).*

5.4 Distância euclidiana ponderada

Segundo Assis et al. (2011), deriva-se da distância euclidiana, onde está associada a uma questão frequente em análise de agrupamento, a ponderação das parcelas, ou seja, o ato de dar peso para as parcelas que o pesquisador julgar mais importantes. Assim, pode-se criar uma matriz diagonal S de ponderação para as parcelas x_1, x_2, ..., x_p com respectivos pesos w_1, w_2,....., w_v e a distância define-se:

$$d_{ij} = \sqrt{w_1|x_{i1} - x_{j1}|^2 + w_2|x_{i2} - x_{j2}|^2 + \cdots + w_v|x_{iv} - x_{jv}|^2}$$

onde $w_1, \ldots, w_v$ são os pesos de cada um dos atributos envolvidos na descrição das parcelas.

$$d_{ij} = \sqrt{w_i \sum_{v=1}^{p} (X_{iv} - X_{jv})^2}$$

5.5 Distância Mahalanobis

Conforme Mello et al. (2012), a distância de Mahalanobis, também chamada distância generalizada, foi desenvolvida em 1936, por Prasanta Chandra Mahalanobis, sendo baseada na correlação entre variáveis. Sua escala é invariante, isto é, não depende da escala de medida. Na estatística Multivariada,

esta distância é muito rica em informações. É usada em análise de agrupamento e outras técnicas de classificação, como também na distribuição de Hotteling's T^2, empregados em testes multivariados. Além disso, a distância Mahalanobis é utilizada para detectar agrupamento, em especial no desenvolvimento de modelos de regressão linear (*Chi-Squale plot ou Q-Q plot*).

A distância de Mahalanobis entre os grupos i e j é usualmente estimada segundo (RAO, 1952) por:

$$D_{ij}^2 = (X_i - X_j)'.\sum^{-1}(X_i - X_j)$$

$\sum$ é a estimativa combinada da matriz da covariância/variância dentro dos grupos.

5.5.1 Vantagens e desvantagens da utilização da distância Mahalanobis

Segundo Cormack (1971), existem duas razões para não se adotar uma distância para aplicação em analise de agrupamento, baseada em uma matriz de covariância em geral:

1. A maioria das inter-relações existentes é, provavelmente, causado pela existência dos grupos que estão sendo buscados;
2. A estrutura de inter-relação dentro dos grupos pode variar consideravelmente de um grupo para outro.

Todavia, utilizando-se a matriz de variâncias e covariâncias, na forma particionada, a adoção da extensão da distância de Mahalanobis para o contexto apresenta as seguintes vantagens:

1. Trata simultaneamente o vetor misto, sem atribuição arbitrária de pesos que combinem distâncias entre contínuas e entre categóricas, a sem a

subjetividade não só dos pesos como também das distâncias a serem adotadas para cada um dos grupos;

2. A distância entre parcelas pode ser decomposta em três parcelas que analisadas separadamente, podem esclarecer a forma como os dados afetam as distâncias por meio das contínuas, removido o efeito das categóricas e das contínuas e da presença conjunta de suas inter-relações.

3. As técnicas de agrupamento podem ser aplicadas a cada parcela separadamente, a combinação de parcelas e distância total. Diferenças nos agrupamentos produzidos podem ser, com isso, um auxilio a melhor caracterização dos grupos que estão sendo buscados.

5.6 Distância de Bray-Curtis

A distância de Bray-Curts (1957) é de uso frequente, por estar disponível na maioria dos pacotes estatísticos (MINITAB, SPSS, STATISTICA, S-PLUS, outros).

Em vez dos desvios quadráticos, é muito comum o uso do valor absoluto:

$$d_{ij} = \sum_{j=1}^{p} W_j |x_{ij} - x_{vj}|$$

onde os $w_j's$ representam as ponderações para as variáveis. Os valores mais usados são os $w_j = 1$ ou $w_j = \frac{1}{p}$

Essa medida é conhecida como métrica "*city-block*".:

Adaptando-se de Kaufman e Rousseeuw (1990), comentam sobre a origem do nome mencionado. Imagine uma floresta, em que se dividiam essas parcelas

de largura 1 (um), na Figura 5. Observa-se na Figura 5, se alguém quiser se mover entre as árvores A e B, percorrer-se-á, no mínimo, uma distância 3 (três), uma vez que não se pode cruzar uma parcela.

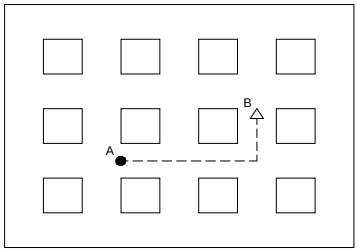

Figura 5. Distância parcela entre as árvores A e B.

A distância *city block* permite caminhar em quatro direções para ir de um ponto a outro. Com isso, a distância de *city block* nem sempre fornece a menor distância em linha reta entre dois pontos, assim faz como a distância euclidiana. No entanto, é bastante utilizada por seu cálculo ser mais fácil do que o da distância euclidiana (BARROSO; ARTES, 2003).

5.7 Distância Chebyshev (DCHBY)

Conforme Reis (2001), essa distância apresenta o valor absoluto da máxima diferença existente entre as variáveis multidimensionais de dois elementos. A distância de Chebychev é apropriada no caso em que se deseja definir dois elementos como diferentes, se apenas umas das dimensões é diferente.

A distância entre dois casos i e j é o valor máximo para todas as variáveis, das diferenças entre esses dois indivíduos.

$$d_{ij} = \max_{v} |X_{iv} - X_{jv}|$$

5.8 Distância de Minkowsky

Constitui uma generalização da distância euclidiana e é dada por:

$$D_{ij} = \left[\sum_{j=1}^{p} W_j |x_{ij} - x_{vj}| \right]^{k}{}^{\frac{1}{k}}$$

Em que os $w_j's$ representam as ponderações para as variáveis, e k é um inteiro com quaisquer finais (BUSSAB et al., 1990). O inconveniente do uso das métricas de Minkowski reside no fato destas métricas apresentarem uma tendência, para que os atributos de maior escala dominem os restantes. Soluções para esse problema incluem a normalização dos atributos contínuos (para uma escala ou variância comum) ou outro tipo de normalização ponderada.

Na Figura 6, evidencia-se a diferença de interpretação entre as várias versões da métrica de Minkovsky, quando se está em duas dimensões.

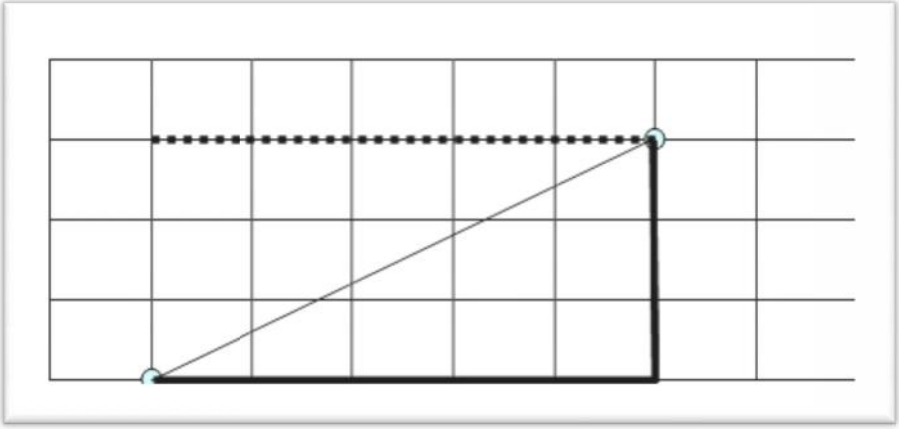

Figura 6: Exemplo de distância calculada pelas distintas métricas. A distância euclidiana (linha fina) é calculada por meio de uma linha reta entre os pontos. A distância *city block* é calculada um quarteirão (unidade) de cada vez. O fato de não se ir reto não muda a distância (comprove-se). A distância de Chebyschev (linha tracejada) é dada pela maior das duas dimensões da distância (LINDEN, 2009).

6. ALGORITMOS DE AGRUPAMENTO

Os algoritmos utilizados na formação dos grupos podem ser classificados em métodos hierárquicos e não hierárquicos.

Os algoritmos de agrupamentos hierárquicos, conhecidos como SAHN ("*Sequencial, Agglomerative, Hierarquic, Nonoverlapping Clustering Methods*"), são formados a partir de uma matriz de (dis)similaridade, na qual se identifica o par de parcelas que mais se parecem. Nesse instante, o par é agrupado, formando uma única parcela. Esse processo requer uma nova matriz de similaridade ou dissimilaridade. Em seguida, identifica-se o par mais semelhante que formará o novo grupo, e assim sucessivamente, até que todas as parcelas fiquem reunidos em um só grupo (SNEATH; SOKAL, 1973).

Os vários algoritmos de agrupamento das espécies diferem no modo como estimam distância entre grupo já formado, e outros grupos ou parcelas por agrupar. O processo de agrupamento de parcelas já agrupadas depende da similaridade e dissimilaridade entre os grupos. Portanto, diferentes definições dessas distâncias poderão resultar em diferentes soluções finais (BUSSAB et al., 1990).

A seguir, são apresentados diversos métodos de agrupamentos que fazem parte dos métodos SAHN. Vale salientar que não existe o que se possa chamar de melhor critério na análise de agrupamentos, embora alguns sejam mais indicados para determinadas situações do que outros (KAUFMANN; ROSSEEUW, 1990). É prática comum utilizar diversos critérios e comparar os resultados. Se tais resultados forem semelhantes, é possível concluir que eles possuem um elevado grau de estabilidade, sendo, pois, confiáveis.

Os métodos mais comuns de agrupamento para determinar a distância entre agrupamentos são: ligação simples, ligação completa, centroides, mediana, médias das distâncias e método Ward (ANDERBERG, 1973).

Um algoritmo que aplica a técnica de aglomeramento hierárquico de agrupamento, para N parcelas, e descrito em (JOHNSON; WICHERN, 2007), apresenta-se:

1. Começar com N grupo, cada um contendo uma entidade unitária e uma matriz simétrica e de distâncias (ou similaridades);
2. Procurar a distância da matriz para o par de (grupo) mais próximo;
3. Deixar a distância entre os agrupamentos “mais similares” I e J;
4. Fusionar os agrupamentos I e J, renomear o grupo recentemente formado (IJ). Atualizar a matriz de distâncias;
5. Eliminando as linhas e as colunas correspondentes aos agrupamentos I e J, e;
6. Adicionar uma linha e uma coluna, preenchendo as distâncias entre o grupo (IJ) e os agrupamentos restantes;
7. Repetir os passos 2 e 3 até N - 1 vezes. (Todos os elementos estarão em um grupo unitário (quando o algoritmo terminar));
8. Salvar a identidade dos agrupamentos que estão fusionados e os níveis (distância ou similaridade) nos quais a fusão efetivada.

Esse algoritmo é referência necessária para entender a forma da operação das diferentes técnicas de agrupamento que serão expostas a seguir:

6.1 Método da Ligação Simples (*Single Linkage*)

Este método também denominado “Método do Vizinho mais Próximo” (*Neighbourhoods*), foi proposto originalmente por Florek et al. (1951) e depois foi revisado por McQuitty (1960), sendo um dos algoritmos mais antigos, mais simples utilizados na literatura nele as conexões entre parcelas e grupos, ou entre grupos, são realizadas por ligações simples entre pares de parcelas. Observa-se na Figura 7, ou seja, a distância entre os grupos é definida como sendo aquela entre as parcelas mais parecidas entre esses grupos.

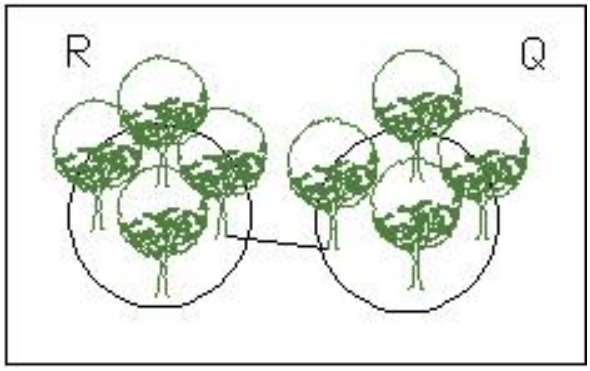

Figura 7. Distância entre agrupamento de ligação simples

O método da ligação simples, segundo Orlóci (1978) e Mardia et al. (1997), é uma técnica de hierarquização aglomerativa e tem, como uma de suas características, não exigir que o número de agrupamentos seja fixado a priori. Assim, tem-se:

Seja E = {E_1, E_2, ,,, , E_p} um conjunto de parcelas em que cada um é representado por um vetor $\underset{\sim}{X}_i$, para i = 1, 2, p pontos do espaço p-dimensional (I_p). No caso de análise da vegetação, cada dimensão do espaço corresponde a uma espécie diferente. Desse modo, qualquer medida de distância estatística ou de similaridade pode ser empregada em tal algoritmo.

Suponha-se que tenham sido determinados todos os n(n - 1)/2 diferentes valores de d_{ij} ou S_{ij} (i = j = 1, 2..... n), representados na forma de uma matriz de distância (D_1) ou de similaridade (S_1).

Este método leva a grupos longos, comparado-se aos grupos formados por outros métodos de agrupamentos SAHN (MEYER et al., 2004).

Os dendrogramas, resultantes desse procedimento, são geralmente pouco informativos, devido à informação das parcelas intermediários que não são evidentes (CARLINI-GARCIA et al., 2001). De acordo com Sneath e Sokal (1973), agrupamentos pelo método de ligação simples podem ser obtidos tanto pelo procedimento aglomerativo quanto pelo divisivo.

Anderberg (1973) cita as seguintes características do referido método:

1. Em geral, grupos muito próximos podem não ser identificados;
2. Permite detectar grupos de formas não-elípticas;

3. Apresenta pouca tolerância a *outliers*, por ter tendência a incorporar os *outliers* em um grupo já existente;
4. Apresenta bons resultados tanto para distância euclidiana quanto para outras distâncias;
5. Tendência a formar longas cadeias (encadeamento).

Encadeamento é um termo que descreve a situação em que há um primeiro grupo de um ou mais parcelas que passa a incorporar, a cada interação, um grupo de apenas uma parcela. Assim, é formada uma longa cadeia, em que se torna difícil definir um nível de corte para classificar as parcelas em grupos (ROMESBURG, 1984), como pode se observar na Figura 8.

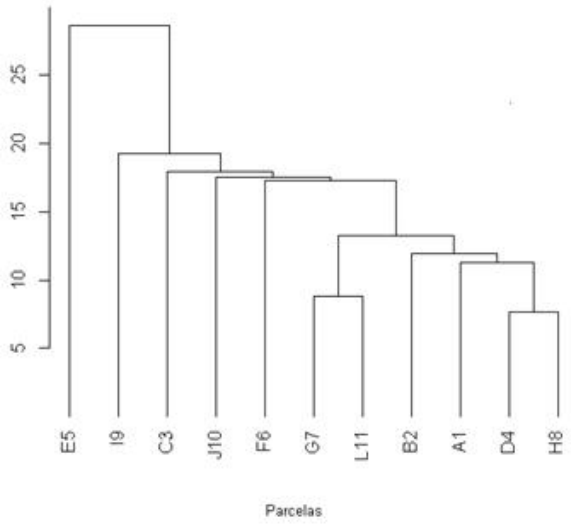

Figura 8. Fenômeno do encadeamento

6.2 Método da Ligação Completa ("*Complete Linkage*")

Este método foi introduzido em 1948, sendo oposto ao método de ligação simples. É também denominado método do elemento mais distante, sendo uma das técnicas de hierarquização aglomerativa de maior aplicação na Análise de Agrupamento (ALBUQUERQUE et al., 2006). Como no método da ligação

simples, aqui também não é exigida a fixação a priori do número de agrupamentos.

Conforme Bussab et al. (1990), no método da ligação completa observada Figura 9, também denominado vizinho mais distante, a dissimilaridade entre dois grupos é definida como sendo aquela apresentada pelas parcelas de cada grupo que mais se parecem, ou seja, formam-se todos os pares com um membro de cada grupo, e a (dis)similaridade entre os grupos é definida pelo par que mais se parece. Esse método em geral leva a grupos compactos e discretos, tendo seus valores de dissimilaridade relativamente grandes.

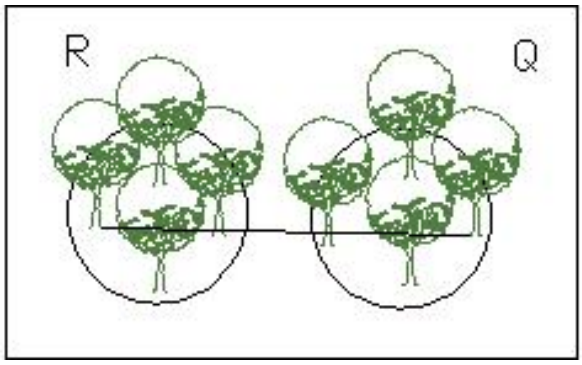

Figura 9. Distância entre agrupamento de ligação completa

Kaufmann e Rosseeuw (1990) cita as seguintes características desse método:

1. Apresenta bons resultados tanto para a distâncias euclidiana quanto para outras distancias;
2. Tendência a formar grupos compactos;
3. Os *outliers* demoram a ser incorporados ao grupo.

O método descrito tem a desvantagem de poder produzir agrupamento diferente, quando a dissimilaridade mínima ocorre para mais de um par de grupos e é necessário escolher um, para ser unido.

Os métodos de ligação simples e ligação completa trabalham em direções opostas. Se eles apresentam resultados semelhantes, significa que o grupo está bem definido no espaço, ou seja, o grupo é real. Todavia se ocorre o contrário, é provável que os grupos não existam (ROMESBURG, 1984).

6.3 Método das Médias das Distâncias (“*Averege Linkage*”)

Este método, também denominados método das médias das ligações e método da média de grupo observa-se na Figura 10, foram propostos originalmente por Sokal e Michener (1958) são uma ponderação entre os métodos de ligação simples e de ligação completa. Usa-se a (dis)similaridade média das parcelas ou do grupo que pretende unir a um grupo já existente. Há vários tipos de métodos, uma vez que há vários tipos de médias, sendo que quatro são mais comuns, provenientes da combinação de dois critérios alternativos: agrupamento em função da média aritmética versus agrupamento com base no centroide, podendo ser ou não ponderados em ambos os casos.

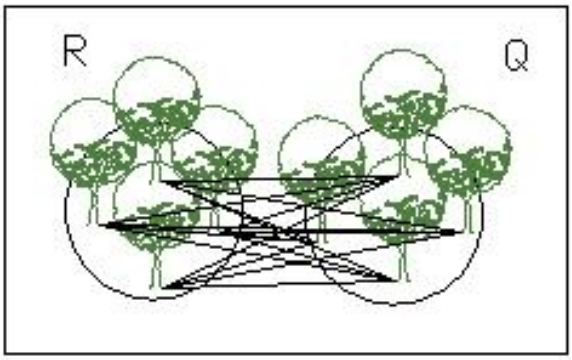

Figura 10. Distância entre agrupamento de ligação média

Nos métodos de agrupamento com base na média aritmética, os coeficientes de similaridade (ou dissimilaridade) médios entre o indivíduo que se pretende agrupar e as parcelas do grupo já existente são calculados. O método do centroide busca o centroide das parcelas para construir grupos, e medir a (dis)similaridade relativa a esse ponto, entre qualquer parcela ou grupo candidato. Os métodos nomalizados pretendem dar pesos iguais a todos os ramos do dendrograma, sendo que o número de parcelas que compõem cada ramo não é considerado (BUSSAB et al., 1990).

Sneath e Sokal (1973) descrevem as quatro combinações possíveis para esses critérios descritos:

Este método define a distância entre dois grupos como sendo a média das distâncias entre todos os pares de parcelas, sendo um em cada grupo. Este procedimento pode ser utilizado tanto para medidas de similaridade como de distância, contanto que o conceito de uma medida média seja aceitável; os grupos são reunidos em um novo grupo quando a média das distâncias entre suas parcelas é mínima;

No método das médias das distâncias, define-se a distância entre dois grupos, i e j, como sendo a média das distâncias entre todos os pares de parcelas constituídos por parcelas dos dois grupos. A estratégia e o valor médio têm a vantagem de evitar valores extremos e de tomar em consideração toda a informação dos grupos.

Um grupo passa a ser definido como um conjunto de parcelas, em que cada um tem mais semelhanças, em média, com todos os membros do mesmo grupo do que com todos os elementos de qualquer outro grupo.

Kaufmann e Rosseeuw (1990) destacam as seguintes características desse método descrito:

1. Apresenta menor sensibilidade a *outliers*, comparando-se com os métodos de ligação simples e completa;
2. Apresenta bons resultados tanto para a distância euclidiana quanto para outras distâncias;
3. Revela tendência a formar grupos com número de parcelas similares.

Segundo Hartigan (1981) esse método também é inconsistente na detecção de grupos de “alta densidade”. Todavia Milligan e Cooper (1985), afirmam que em um estudo comparativo envolvendo os métodos ligação simples, ligação completa e ligação médias, classificaram o método da ligação média como o melhor, visto que o último método tira proveito da homogeneidade do método de ligação completa e da estabilidade do método da ligação simples.

6.4 Método de Ward

Ward (1963) propôs um processo geral de classificação em que n parcelas são progressivamente reunidos dentro de grupos, por meio da minimização de uma função objetiva para cada (n - 2) passos de fusão.

Inicialmente, nesse algoritmo, admite-se que cada uma das parcelas se constituía em um único agrupamento. Considerando a primeira reunião de parcelas em um novo agrupamento, a soma dos desvios dos pontos representativos de suas parcelas, em relação à média do agrupamento, é calculada, e dá uma indicação de homogeneidade do agrupamento formado. Esta medida fornece a “perda de informação”, que se produz, ao reunir as parcelas em um agrupamento (LATTIN et al., 2011).

A proposta de Bouroche e Saporta (1972) demonstra quando as parcelas são pontos de um espaço euclidiano (I_p). A qualidade de uma partição é definida por sua inércia intragrupo ou por sua inércia intergrupo. Quando se parte de K+1 grupos para K grupos, ou seja, agrupando-se dois grupos em uma só, a inércia intergrupo só pode diminuir. A inércia intergrupo é a média da soma dos quadrados das distâncias entre os centros de gravidade de cada grupo e o centro de gravidade total.

Mingoti (2007) propôs que a reunião de parcelas em grupos fosse realizada pela análise dos valores da função de agrupamento, reunindo-se as parcelas mais próximas, isto é, aqueles que apresentassem Min (d_{ij}).

Conforme Reis (2001), o algoritmo de Ward se baseia na perda de informação resultante do agrupamento das espécies e medida por meio da soma dos quadrados dos desvios das parcelas individuais relativamente às médias dos grupos em que são classificadas.

Cada grupo se caracteriza por uma soma dos quadrados dos desvios de cada parcela do centroide do mesmo algoritmo (é uma soma dos numeradores dos estimadores das variâncias de cada variável dentro do grupo; é também a soma de distância euclidiana do quadrado de cada parcela do centroide). A distância entre dois grupos se define como o aumento que se pronunciaria nessa soma de quadrados, se ambos os grupos se agregassem para a formação de um

único grupo. O algoritmo de Ward é atraente por se basear em uma medida com forte apelo estatístico e por gerar grupos que, assim como os do método vizinho mais longe, possuem uma alta homogeneidade interna (BARROSO; ARTES, 2003).

Romesburg (1984) cita as seguintes características do método ora descrito:

1. Apresenta bons resultados tanto para distâncias euclidiana quanto para outras distâncias;
2. Pode apresentar resultados insatisfatórios quando o número de parcelas em cada grupo é praticamente igual;
3. Tem tendência a combinar grupos com poucas parcelas;
4. É sensível à presença de *outliers*.

Os algoritmos de ligação simples, completa e média podem ser utilizados tanto para variáveis quantitativas quanto qualitativas, ao contrário dos métodos de centroide e de Ward, que são apropriados apenas para variáveis quantitativas, já que têm como base a comparação de vetores de médias (BARROSO; ARTES, 2003).

7 INFERÊNCIA ESTATÍSTICA

Apesar das tentativas de construção de vários testes para a confiabilidade estatística dos agrupamentos, nenhum procedimento totalmente comprovado está ainda disponível. A ausência de testes adequados provém da dificuldade de especificação de hipóteses nulas realísticas.

No que se refere aos enormes problemas associados à inferência estatística na análise de agrupamentos, os pesquisadores podem lançar mão de alguns procedimentos práticos para conferir, de maneira superficial, os resultados dessas análises. Por exemplo, eles podem aplicar duas ou mais rotinas diferentes de agrupamento ao mesmo conjunto de dados ou realizar a análise de agrupamentos com os mesmos dados, empregando diferentes medidas de distância e comparando os resultados por meio de algoritmos e medidas de distância. Pode-se, também, repartir os dados aleatoriamente em duas metades, realizar agrupamentos diferentes, e, portanto, examinar os perfis médios de valores de cada agrupamento mediante subamostras. Outra alternativa é deletar diversas colunas (parcelas) nos dados originais de perfis, calcular as medidas de dissimilaridade entre as colunas remanescentes, e comparar esses resultados com os agrupamentos encontrados por meio do uso do conjunto total de colunas (parcelas). Outra abordagem de validação seria a utilização de procedimentos de simulação que empreguem geradores de números aleatórios para criar um conjunto de dados com propriedades que combinem com aquelas dos dados originais, não contendo, entretanto, segue um agrupamento. Em seguida, aplicam-se os métodos de agrupamento nos dados reais e nos artificiais, e comparam-se as soluções resultantes (GAULI et al., 2009).

7.1 Simulação de Monte Carlo

Trata-se de um método para estimar parâmetros e taxas de probabilidade por meio de amostragem aleatória por computador. É utilizada quando tais valores são muito difíceis ou impossíveis de serem calculados analiticamente.

Em validação de agrupamentos, uma das formas mais comuns de utilização de simulação de Monte Carlo é no estabelecimento da distribuição referência de um índice sob a hipótese nula. Inicialmente, é gerada uma grande quantidade de conjuntos de dados artificiais de acordo com a distribuição considerada na hipótese nula H_0. Cada um desses conjuntos é agrupado e o valor do índice é calculado em cada caso. Com esses valores do índice é traçado um gráfico de dispersão, que é uma aproximação da função de densidade de probabilidade do índice. Dado o valor do índice para o agrupamento que está sendo validado e a distribuição estimada, determina-se a possibilidade da hipótese H_0 ser aceita ou rejeitada.

A falta de conhecimento da estrutura dos dados, necessária à aplicação de critérios de validação, inviabiliza muitas vezes a sua aplicação. No entanto, usando-se o método de Monte Carlo, e, para alguns conjuntos de dados muito específicos (exibindo propriedades de coesão interna e isolamento externo), em (MILLIGAN, 1981), prova-se que existe uma alta correlação entre alguns índices de validação. Assim, estes índices poderão ser utilizados, para avaliar o grau de variabilidade de um dado algoritmo da estrutura em grupos efetivamente existente nos dados. No referido trabalho, estudaram-se 30 métodos internos de avaliação e dois índices externos de validação que efetivamente medem o grau de variabilidade da estrutura existente no conjunto de dados, o índice de Rand e o índice de silhueta. Concluiu-se que quatro métodos de agrupamento e que os índices mostraram ser fiáveis, apresentando uma grande correlação com o índice Rand, podendo ser assim considerados índices válidos na identificação do grau de validação da estrutura nos dados.

No contexto dos índices de validação de agrupamento, um dos métodos de validar o resultado consiste na avaliação da sua estabilidade (RAND, 1971; MILLIGAN, 1980; GORDON, 1999). O processo engloba a reanálise de uma versão modificada do conjunto de dados, avaliando a extensão das diferenças em relação o agrupamento original.

8 MATERIAL E MÉTODOS

8.1. Área de Estudo

Este estudo foi realizado a partir de um banco de dados obtidos por Costa Júnior (2006) em um remanescente de Mata Atlântica denominado Mata das Caldeiras, localizado no município de Catende-PE, que está situado na mesorregião da Mata Pernambucana .

Os dados foram obtidos a partir de 40 parcelas de 250 m² (10 x 25 m, cada) e 29 espécies, alocadas sistematicamente ao longo de todo o remanescente, distando 25 m entre si. Nas parcelas, foram amostrados apenas os indivíduos arbóreos vivos com CAP (circunferência à altura do peito – 1,30 m do solo) $\geq$ 15 cm, que receberam placas metálicas enumeradas e tiveram os seguintes dados anotados: o CAP, mensurada com fita métrica, e a altura, pela estimativa visual, utilizando-se como base as hastes do podão, as quais medem 2 m. Posteriormente, a partir da CAP foram calculados os diâmetros (DAP) e as áreas basimétricas (COSTA JÚNIOR, 2006).

No presente estudo, foram utilizados os dados de espécies com um mínimo de 10 indivíduos amostrados (Tabela 1).

Tabela 1. Listagem das espécies arbóreas com respectiva média e desvio padrão do diâmetro a 1,30 m do solo (DAP), da altura e da área basimétrica para amostra de 805 indivíduos arbóreos, remanescente de Floresta Atlântica, Mata das Caldeiras, município de Catende, PE.

Família/Espécie	DAP(cm)		Altura(m) (H)		Área (m^2)(G) basimétrica	
	Média	DP	Média	DP	Média	DP
Anacardiaceae						
Tapirira guianensis Aubl	18,7	9,8	17,6	8,4	0,0360	0,0355
Thyrsodium spruceanum Benth	9,57	6,0	12,1	6,2	0,0099	0,0148
Arliaceae						
Schefflera moratotoni (Aubl.) Maguire, Steyerm & Frodin	15,0	11,3	16,4	6,2	0,0260	0,0388
Bombaceae						
Eriotheca gracilipes (K. Schum.) A. Robyns	11,6	7,3	11,7	6,1	0,0144	0,0218
Burseraceae						
Protium heptaphyllum (Aubl.) Marchand	8,1	4,0	11,1	5,3	0,0063	0,0066
Caesalpiniaceae						
Copaifera langsdorffii Desf	15,1	10,3	14,0	7,0	0,0257	0,0335
Dialium guianense (Aubl.) samdwirh	10,1	8,8	15,1	7,9	0,0099	0,0281
Cecropiaceae						
Cecropia palmata Willd.	13,7	8,2	11,2	4,9	0,0199	0,0250
Cedrela sp. P. Browne	10,2	6,0	11,1	4,9	0,0106	0,0137
Chrysobalanaceae						
Licania rigida Benth.	16,9	8,1	17,6	11,8	0,0271	0,0214
Erythroxylaceae						
Erythroxylum squamatum Sw.	7,8	4,6	9,7	4,4	0,0064	0,0093
Euphorbiaceae						
Mabea occidentalis Benth	15,3	10,2	19,4	8,9	0,0257	0,0286
Fabaceae						
Pterocarpus violaceus Vogel	15,6	10,8	12,6	5,1	0,0260	0,0363
Lauraceae						
Nectandra cuspidata (Nees et. Mart.) Nees	12,2	9,8	13,4	8,2	0,0189	0,0322
Ocotea gardneri (Meisn.) Mez	12,1	8,6	13,3	6,3	0,0172	0,0253
Ocotea opifera Mart.	11,4	8,5	10,2	4,3	0,0155	0,0216
Lecythidaceae						
Eschweilera ovata (Cambess.) Miers	9,67	8,9	10,5	4,7	0,0134	0,0391
Melastomataceae						
Miconia albicans (Sw.) Triana	7,53	2,3	10,2	2,9	0,0048	0,0031
Meliaceae						
Pouteria grandiflora (A.Dc.) Baehni	14,1	5,1	10,7	4,4	0,0215	0,0108
Mimosaceae						
Inga thibaudiana DC.	10,2	6,1	10,6	5,0	0,0109	0,0136
Parkia pendula (Willd.) Benth. ex Walp.	32,6	20,9	15,9	7,4	0,1160	0,1398
Plathymenia foliolosa Benth.	24,8	21,6	18,7	8,2	0,0837	0,1592
Stryphnodendron pulcherrimum (Willd.) Hochr.	15,9	12,4	14,1	9,0	0,0307	0,0463
Moraceae						
Brosimum discolor Schott	13,8	9,4	14,2	8,7	0,0218	0,0297
Helicostylis tomentosa (Poepp. & Endl.) Rusby	10,1	7,4	12,4	5,9	0,0122	0,0322
Ochnaceae						
Ouratea hexasperma (A. St.-Hil.) Baill.	9,29	4,9	12,3	5,6	0,0085	0,0091
Sapindaceae						
Cupania racemosa (Vell.) Radlk.	8,46	4,5	9,9	6,4	0,0071	0,0017
Cupania revoluta Rolfe	6,96	1,4	10,5	3,6	0,0039	0,0080
Geral	13,1	8,5	13,2	6,4	0,0225	0,03151

Para o estudo da similaridade entre parcelas foi utilizada a altura de Lorey média obtida por:

$$\overline{H}_{Ljk} = \frac{\sum_{k=1}^{nk} h_{ijk} \cdot g_{ijk}}{\sum_{k=1}^{nk} g_{ijk}}$$

Em que:

$\overline{H}_{Ljk}$ = altura de Lorey média da j-ésima espécie, na k-ésima parcela;

h_{ijk} = altura do do i-ésimo indivíduo, da j-ésima espécie, na k-ésima parcela;

g_{ijk} = área basimétrica do i-ésimo indivíduo, da j-ésima espécie, na k-ésima parcela,j = 1, ..., 29; k = 1, ..., 40; i = 1, ..., n_i (n_i = número total de indivíduo da j-ésima espécie, na k-ésima parcela).

8.2 Métodos Estatísticos

Com a proposta de metodologia estatística para análise de agrupamento considerou-se o seguinte: cálculo da matriz de distância, considerando-se método incremental, a distância euclidiana, a normalização, a utilização dos algoritmos de ligação simples, de ligação completa, de ligação média e de Ward, do dendrogramas, do coeficiente R^2, do Pseudo F, de Wilks, da correlação de cofenética, de Rand ajustado e o processo de dados artificiais (empíricos) “Monte Carlo”. Os métodos, algoritmos e figuras foram implementados utilizando-se o ambiente R.

8.2.1 Medida de distância

A distância euclidiana foi utilizada como medida de distância para o método incremental hierárquico e do algoritmo de ligação simples, de ligação completa, de ligação média e de Ward, a qual foi obtida conforme a expressão:

$$d_{ij} = \sqrt{\sum_{v=1}^{p} (X_{iv} - X_{jv})^2}$$

Em que:

d_{ij} = distância euclidiana entre o i-ésimo e o j-ésimo parcela;

X_{IV} representa a característica da parcela i,

X_{IV} representa a característica da parcela j,

p é o número de parcelas na amostra,

v é o número de individuo na amostra.

8.2.2 Método Incremental

A proposta desse método tem, como interesse principal, agregar a filosofia dos agrupamentos dos métodos hierárquicos e não-hierárquicos, com o intuito de (CAN, 1993):

1. Evitar a subjetividade inerente ao pesquisador, comumente utilizada; e
2. Reduzir o número de interações utilizadas até a convergência, para o agrupamento desejado.

A seguir, é apresentado o algoritmo do método denotado método incremental.

I. Passo 1: Determinou-se, para cada parcela do conjunto de dados analisado, o somatório das distâncias a todas as demais parcelas do conjunto. Ordenaram-se as parcelas, em ordem crescente, e se associou a primeira parcela a um grupo;

II. Passo 2: Determinou-se o intervalo de abrangência de cada parcela. A abrangência é dada pelo somatório das distâncias de cada parcela aos demais, dividido pelo número de observações.

III. Passo 3: Avaliou-se, para a primeira parcela ordenada (primeiro nó semente), quais parcelas estavam contidas dentro de seu intervalo, compondo, então, um grupo.

IV. Passo 4: Repetiu-se o passo 3, para as parcelas subsequentes das parcelas ordenadas fora do intervalo de abrangência dos seus antecessores. Caso uma parcela que já compunha um grupo estivesse mais próxima de outra parcela do nó semente, ela saiu do agrupamento inicial e passou a compor um novo grupo com essa nova parcela.

8.2.3 Algoritmos de agrupamento

Utilizou-se os seguintes algoritmos de agrupamento, por serem os mais usados na prática pela facilidade de serem encontrados nos mais diversos programas computacionais:

a) Método da Ligação Simples ou do Vizinho mais Próximo (*Single Linkage*)

De posse da matriz primária de dados X (n x p), o método de ligação simples foi resolvido na seguinte sequência de cálculos:

1. Com base na matriz de distância euclidiana, determinou-se os valores da função de agrupamento d_{ij}, que foram representados na forma matricial (D_1);
2. Localizou-se o valor mínimo de $d_{ij} > 0$. As parcelas E_i e E_j, correspondentes a este valor, foram reunidos em um mesmo grupo, ficando (n - 1) agrupamentos remanescentes;
3. Com base na matriz de distância inicial (D_1), determinou-se a distância entre o novo agrupamento e as demais parcelas, por meio da relação:
   $d_{(i,j)\,l} = \min (d_{il}, d_{il})$, $l = 1, (n - 2)$
   $l \neq i \neq j$ e construiu-se nova matriz de distância (D_2);
4. Localizou-se em D_2 o menor valor de $d_{ij} > 0$ e, em seguida, agrupou-se as parcelas que deram origem a esta nova distância, formando-se novo agrupamento. Neste passo, têm-se (n - 2) agrupamentos;
5. Compôs-se nova matriz de distâncias, baseando-se na matriz de distância. Para isto, calculou-se a distância entre o agrupamento formado na etapa anterior e os demais, considerando-se uma parcela isolada de E como um agrupamento. Em seguida, retornou-se à etapa 4.

Os processos foram repetidos até que todas as parcelas de E fossem alocados a um só agrupamento.

b) Método da Ligação Completa ou do Vizinho mais Longe *(complete linkage)*

Dados n parcelas e admitindo-se conhecidos os n(n - 1)/2 valores de uma função de agrupamentos, d_{ij}, i = j = 1, 2.... n, apresentados na forma de uma D, este método pode ser sintetizado, segundo Mardia et al. (1997), nas seguintes etapas:

1. Determinou-se, com base na matriz de distância euclidiana, o conjunto de valores de uma função de agrupamento. Estes valores constituem medida de distância estatística (D) que formam a matriz D_1;
2. Decidiu-se o valor mínimo de d_{ij}, sendo as parcelas E_i e E_j reunidos num primeiro grupo. Então se pressupõe que os (n - 2) parcelas restantes constituam cada um, um agrupamento distinto;
3. Com base na matriz D_1, determina-se a distância entre cada um dos (n - 2) parcelas e o novo agrupamento formada pelas parcelas (E_i, E_j). Esta distância é calculada pela relação:

$$d_{(i,\,j)\,l} = \text{máx}\,(d_{il}, d_{jl}),\ l = 1, (n - 2).$$
$$l \neq i \neq j$$

   Sendo estas distâncias reunidas numa matriz D_2
4. Determina-se, com base na matriz D_2, o maior valor d_{ij}, agrupando-se as parcelas correspondentes, dando-se origem a um novo agrupamento, e, então, obtendo-se (n - 2) agrupamentos;
5. Construiu-se um novo conjunto de valores de distâncias com base na matriz D_2 (interação anterior), entre o novo agrupamento e os demais.

c) Método das Médias das distâncias

O método pode ser resumido nos seguintes passos:

1. Determina-se a matriz de distâncias inicial.
2. Localiza-se os duas parcelas que apresentam a menor distância, reunindo em um único grupo.
3. Calcula-se a distância entre os diversos pares de grupos como sendo a média das distâncias entre todos os pares de suas parcelas sendo uma parcela de cada um dos grupos.
4. Os dois grupos que apresentam menor distância são reunidos em um único grupo.

Se o número de grupos obtidos é igual a um número k ≤ n, o processo termina. Caso contrário, retorna-se ao passo 3.

Esta fórmula fornece as Médias das distâncias.

$$d^{*}k_{(i,j)} = \left[\frac{n_i}{n_i + n_j}\right] \cdot d_{ki} + \left[\frac{n_j}{n_i + n_j}\right] \cdot d_{kj}$$

em que:

$d^{*}_{k(ij)}$, d_{ki} e d_{kj} = distâncias entre as parcelas k e o agrupamento ij, k e i, k e j, e i e j, respectivamente.

n_i e n_j = número de parcelas nos agrupamentos i e j, respectivamente.

d) Método de Ward

Segundo Orlóci (1978), o algoritmo de Ward pode ser resumido nas seguintes etapas:

1. Determina-se a matriz de distâncias e localizam-se os dois agrupamentos para os quais d_{ij} é mínimo;
2. Reúnem-se estes agrupamentos; forma-se um novo agrupamento, e se verifica, se o número de agrupamentos (k) já foi alcançado, senão, segue-se à etapa 3, caso contrário, termina-se a análise;
3. Calcula-se o valor do aumento a ser obtido na soma dos quadrados pela reunião de qualquer dos agrupamentos: I = (1/2), d_{pq}.;
4. Determinam-se os dois agrupamentos que apresentam um menor incremento na matriz D, isto é, Min (I_{ij}) e volta-se à etapa 2.

Este método tem como função de agrupamentos a distância euclidiana, e o critério de agrupamento é dado pelo valor do incremento, que se obtém na soma de quadrados do erro.

Observação:

$d^2_{pk} = (X_p - X_k)^2$, é a distância entre as médias das parcelas de G_p e G_k sendo G_p e G_k, respectivamente, os grupos p e k,

$I_{pk} = \frac{N_p \cdot N_k}{N_p + N_k} \cdot d_{pk}$, em que as reuniões dos agrupamentos G_p e G_k será feita se I_{pk} = mínimo.

Admita-se que para o agrupamento $G_p \cup G_k = G_r$, o incremento na soma das médias do erro é dado por:

$$I_{tr} = \frac{n_t \cdot n_r}{n_t + n_r} \cdot d_{tr}, \quad \text{onde} \quad d^2_{tr} = (\overline{X}_t - \overline{X}_r)^2$$

Podendo ser escrita por:

$$d_{tr} = \frac{n_p}{n_r} \cdot d_{tp} + \frac{n_k}{n_r} \cdot d_{tk} - \frac{n_p \cdot n_k}{n_r^2} \cdot d_{pk},$$

Substituindo-se cada distância, em função do número de parcelas, do agrupamento, obteve-se:

$$I_{tr} = \frac{1}{n_t + n_r} \cdot \left[(n_t + n_p) \cdot I_{tp} + (n_t + n_k) \cdot I_{tk} - n_r \cdot I_{pk} \right]$$

Ou ainda, considerando $d_{tp} = 2.I_{tp}$, tem-se:

$$d_{tr} = \frac{1}{n_t + n_r} \left[(n_t + n_p).d_{tp} + (n_t + n_k).d_{tk} - n_t d_{pk} \right]$$

8.3 Comparação dos métodos

8.3.1 Correlação cofenética

A correlação cofenética mede o grau de ajuste entre a matriz de dissimilaridade original (matriz D) e a matriz resultante da simplificação proporcionada pelo algoritmo de agrupamento (matriz C). No caso, C é aquela obtida após a construção do dendrograma. Tal correlação foi calculada conforme Mayer et al. (2004):

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

em que;

$c_{i\ j}$ = valor de dissimilaridade entre as parcelas i e j, obtidos a partir da matriz cofenética;

$d_{i\ j}$ = valor de dissimilaridade entre as parcelas i e j, obtidos a partir da matriz de dissimilaridade;

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

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

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

8.4 Validação

8.4.1 Coeficiente R^2

Em cada passo do algoritmo de agrupamento, é possível calcular a soma de quadrados entre os grupos, e dentro dos grupos da partição correspondente. Os critérios de formação de grupos que constituem um agrupamento que mais se destacam, são os critérios de formação de grupos usados na análise de uma matriz de dados contínuos, X_{nxp}, que usam a decomposição da matriz de dispersão T, dada por:

$$T = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x})(x_{ij} - \bar{x})^T$$

e é o vetor de dimensão p das médias de cada variável.

$$\bar{x} = \sum_{j=1}^{k}\sum_{i=1}^{n_j} x_{ij}$$

Esta matriz da variabilidade total pode ser decomposta em**:**

- matriz da dispersão dentro do grupo, W, definida por :

$$W = \sum_{j=1}^{k}\sum_{i=1}^{n_j}(x_{ij} - \bar{x}_j)(x_{ij} - \bar{x}_j)^T$$

em que $\bar{x}_j$ é o vetor de dimensão p das médias das variáveis dentro do grupo j .

- matriz da dispersão entre grupos, B, definida por :

$$B = \sum_{j=1}^{k} \sum_{i=1}^{n_j} n_j (\bar{x}_j - \bar{x})(\bar{x}_j - \bar{x})^T, \text{com} \sum_{j=1}^{k} n_j = n$$

Então $T = B + W$

onde T, W e B são as matrizes associadas à variabilidade total dos dados, à variabilidade dentro dos grupos e à variabilidade entre os grupos, respectivamente.

Para dados univariados, p a equação $T = B + W$ representa a decomposição da soma total dos quadrados da variável em soma dos quadrados dentro dos grupos e a soma dos quadrados entre grupos, que é fundamental na análise de variância.

Como T é fixo, porque não depende do agrupamento que se realize, a melhor partição é aquela em que W é mínimo ou B máximo, isto é quanto maior a homogeneidade interna dos grupos maior é a separação entre os grupos.

Define-se o coeficiente R^2 da partição como:

$$R^2 = \frac{B}{T}$$

8.4.2 Estatística Pseudo F

Segundo Ferreira et al. (2008), este critério que pode-se utilizar tanto nos métodos hierárquicos como nos não hierárquicos e que baseia-se em uma aproximação F. para compararmos duas soluções de agrupamento.

Conforme Calinski e Harabasz (1974) sugerem, o cálculo da estatística F em cada passo do agrupamento, isto é,

$$F = \frac{B/(k^* - 1)}{W/(n - k)} = \left(\frac{n - k^*}{k^* - 1}\right)\left(\frac{R^2}{1 - R^2}\right)$$

Em que k é o número de grupos relacionado com a partição do respectivo estágio de agrupamento.

8.4.3 Teste de Wilks

A razão Λ (lambda de Wilks), que é a estatística do teste para a hipótese proposta e para a análise de variância simples, resulta do quociente entre os determinantes das matrizes de somas de quadrados e produtos cruzados dentro dos grupos e total.

$$\boldsymbol{\Lambda} = \frac{|\boldsymbol{И}|}{|\boldsymbol{И} + \boldsymbol{B}|}$$

Uma análise de variância permite que, comparando-se vários grupos a um só tempo, utilizam-se variáveis contínuas. O teste é paramétrico (a variável de interesse deve ter distribuição normal) e os grupos devem ser independentes.

8.4.4 Índice de Rand ajustado

O índice de validação externo Rand ajustado é muito utilizado na comparação de algoritmos de agrupamento, e oferece como vantagens a independência do número de grupos (JAIN; DUBES, 1988). O índice de Rand ajustado determina a semelhança entre duas parcelas P_1 e P_2 examinando a qual

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

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

$$Rand\ ajustado = \frac{\sum_{i=1}^{k_{P_1}} \sum_{j=1}^{k_{P_2}} \binom{n_{ij}}{2} - \binom{n}{2}^{-1} \sum_{i=1}^{k_1} \binom{n_{i.}}{2} \sum_{j=1}^{k_{P2_1}} \binom{n_{.j}}{2}}{\frac{1}{2}\left[\sum_{i=1}^{k_{P_1}} \binom{n_{i.}}{2} \sum_{j=1}^{k_{P_2}} \binom{n_{.j}}{2}\right] - \binom{n}{2}^{-1} \sum_{i=1}^{k_{P_1}} \binom{n_{i.}}{2} \sum_{j=1}^{k_{P_2}} \binom{n_{.j}}{2}}$$

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

8.4.5 Dados artificiais via método “Monte Carlo”

A simulação de dados artificiais, a partir dos dados originais, foi realizada conforme os estudos de Milligan (1981) e Milligan e Cooper (1985). O gerador de dados foi desenvolvido seguindo as descrições desses autores. Em nosso trabalho, utilizou-se 40 parcelas e 29 espécies cada, incorporado na distância euclidiano com três experimentos n = 60, 80 e 100. Cada conjunto de dados, aplicou-se “método incremental” e calculou-se o número de grupos ideal contém k = 6, 6 e 7 agrupamentos distintos para cada um agrupamento é permitido em todos. A disposição das parcelas com agrupamentos segue a distribuição normal bivariada, em cada experimento o resultado estruturado pode ser considerado

para consistir o agrupamento natural que exibe as propriedades de isolação externa e de coesão interna, comprovadas pelos métodos de validação.

Para validar os métodos em análise de agrupamentos via Monte Carlo, foram seguidos os seguintes passos:

1. Considerou-se a seguinte matriz X, denominada de matriz de dados ou matriz original (primaria).

$$\mathbf{X}_{(nxp)} = \begin{bmatrix} X_{11} & X_{12} & \cdots\cdots\cdots & X_{1n} \\ X_{21} & X_{22} & \cdots\cdots\cdots & X_{2n} \\ . & . & & . \\ . & . & & . \\ X_{p1} & X_{p2} & \cdots\cdots\cdots & X_{pn} \end{bmatrix}$$

Em que i = 1, 2 ,....., p espécie na amostra e j = 1, 2,....,n parcelas.

2. Com a matriz original, encontrou-se a matriz de distância euclidiana, para aplicação dos métodos de agrupamento.

3. De posse da matriz euclidiana, aplicou-se “método incremental” e calculou-se o número de grupos ideal.

4. De posse da matriz de euclidiana, aplicou-se “Monte Carlo” e calculou-se uma nova matriz de distância para aplicação dos métodos de agrupamento e comparação com a aplicação do item 2.

5. De posse dos métodos de agrupamentos, aplicaram-se os métodos de validação para a comparação dos mesmos.

9 RESULTADOS E DISCUSSÃO

9.1 Matriz de distância euclidiana

Com base na matriz de distância euclidiana obtida a partir dos dados originais e transformados na altura Lorey (Tabela 2) foram aplicados o método incremental, os métodos de ligação simples, da ligação completa, da média das distâncias e de Ward e obtidos os respectivos dendrogramas (Figuras de 14 a 17).

Na Tabela 2, observa-se, como base na distância euclidiana, que as parcelas mais similares são 2, 32, 6, 12 e 38 e as mais distantes 5, 24, 39, 3, 27 e 13.

9.2 Método Incremental

Nesse método, como observa-se na Tabela 3, foram ordenadas as somas das distâncias de cada parcela. Por meio das médias das somas das distâncias de cada parcela, obteve-se o intervalo inferior e o superior de cada parcela. Por fim, identifica-se qual (ou quais) parcela(s) pertencem ao um determinado grupo (Tabela 4). Assim, descobre-se o número de grupos ideal, conhece-se o número das parcelas em seus respectivos grupos, verificando o grau de variabilidade dos grupos e identificando-se as parcelas *outliers*.

Verifica-se que o método de agrupamento incremental não precisar armazenar toda matriz na memória computacional (JAIN et al., 1999). Logo, o espaço necessário para o método ser executado é bem pequeno. Geralmente não são iterativos e, assim, o tempo para execução é menor do que os métodos hierárquicos e não hierárquicos.

Tabela 2. Matriz de distância euclidiana entre as 40 parcelas obtida a partir da altura de Lorey

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
1	0.00	62.60	71.50	66.60	72.40	59.40	55.80	78.00	64.90	49.20	65.40	55.10	73.40	66.00	76.30	75.90	78.90	62.30	57.20	61.00	64.20	53.60	59.60	70.10	60.10	55.10	69.90	41.30	72.00	59.60	60.00	59.40	61.60	56.20	61.10	48.00	66.30	58.60	67.60	71.10
2	62.60	0.00	56.50	48.20	59.20	46.50	44.00	71.20	47.60	64.50	59.50	44.30	49.40	55.00	58.80	75.00	63.60	44.80	52.30	45.50	46.20	47.20	52.90	45.20	41.20	34.40	63.20	63.30	66.30	50.70	57.40	39.80	44.40	53.80	57.10	58.40	32.80	46.40	62.00	44.30
3	71.50	56.50	0.00	62.90	73.20	59.60	56.90	72.50	74.20	74.20	82.00	53.70	59.70	64.90	67.40	74.70	79.00	56.60	57.00	63.20	60.30	63.10	59.60	62.30	65.80	53.20	74.30	72.20	72.60	64.50	72.50	57.60	55.40	72.60	72.20	64.60	67.10	56.80	76.40	64.70
4	66.60	48.20	62.90	0.00	61.50	44.90	54.50	64.70	59.60	68.50	63.50	41.50	70.40	58.60	65.00	59.90	66.10	56.40	56.00	57.00	51.20	51.50	54.80	59.50	58.60	45.90	58.20	54.60	67.20	55.50	60.40	47.30	55.90	53.30	57.70	58.90	57.00	53.30	63.40	46.20
5	72.40	59.20	73.20	61.50	0.00	55.50	54.20	74.60	64.40	70.00	74.70	59.90	59.40	61.40	78.80	77.70	59.40	72.40	62.70	67.50	62.30	63.00	55.90	66.20	64.90	48.60	68.30	70.00	81.50	72.70	61.10	48.60	59.60	65.80	57.40	69.80	49.90	50.20	55.50	58.50
6	59.40	46.50	59.60	44.90	55.50	0.00	50.30	65.30	52.30	57.60	60.10	39.80	65.70	47.00	59.60	69.80	48.20	52.90	45.80	57.60	57.20	46.00	49.90	64.80	50.20	37.60	63.10	53.70	71.40	56.50	58.40	44.70	46.60	52.40	49.80	47.90	49.50	43.40	49.40	52.30
7	55.80	44.00	56.90	54.50	54.20	50.30	0.00	70.90	58.70	61.70	66.10	40.90	57.20	59.60	75.50	74.60	67.10	48.40	53.30	57.30	63.60	48.60	43.10	55.60	58.20	39.20	59.40	54.40	67.40	58.70	61.30	48.30	51.10	52.60	42.70	61.70	39.70	51.60	56.10	59.20
8	78.00	71.20	72.50	64.70	74.60	65.30	70.90	0.00	70.20	82.90	77.40	66.30	84.40	79.50	81.20	93.10	65.80	75.20	73.00	69.70	65.10	68.30	74.40	77.90	78.60	65.80	76.40	73.90	85.70	68.80	69.90	65.90	66.80	59.70	75.70	72.30	75.10	62.80	77.10	67.90
9	64.90	47.60	74.20	59.60	64.40	52.30	58.70	70.20	0.00	68.30	54.50	63.00	72.90	65.00	71.30	83.90	53.30	58.80	59.90	52.30	62.90	59.00	61.00	67.80	61.70	53.20	71.60	66.00	80.90	51.00	51.80	54.50	50.60	63.00	65.40	63.10	44.30	57.70	67.50	65.10
10	49.20	64.50	74.20	68.50	70.00	57.60	61.70	82.90	68.30	0.00	74.00	67.30	75.90	71.00	73.60	82.70	79.60	70.10	63.80	72.50	62.00	51.00	61.90	73.60	57.20	58.90	65.70	55.70	71.30	71.70	67.40	54.20	67.50	64.40	62.70	58.40	64.00	60.80	77.90	70.60
11	65.40	59.50	82.00	63.50	74.70	60.10	66.10	77.40	54.50	74.00	0.00	63.10	84.80	70.10	83.70	85.00	65.40	72.40	71.50	63.30	75.70	51.30	70.60	80.60	59.10	64.10	72.20	64.10	79.70	63.10	71.70	63.10	59.70	62.30	75.70	60.60	59.00	63.50	66.40	70.60
12	55.10	44.30	53.70	41.50	59.90	39.80	40.90	66.30	63.00	67.30	63.10	0.00	64.40	50.50	70.90	58.60	61.60	48.10	44.10	54.50	56.70	49.70	48.30	61.70	52.40	34.50	60.70	44.90	71.10	54.10	64.50	46.10	44.40	43.90	54.30	52.80	50.60	47.60	51.20	47.00
13	73.40	49.40	59.70	70.40	59.40	65.70	57.20	84.40	72.90	75.90	84.80	64.40	0.00	68.30	64.30	82.70	76.60	59.20	64.80	66.00	58.50	69.40	60.10	39.50	59.20	54.10	70.10	78.70	75.70	69.20	68.50	59.00	66.90	68.40	66.60	73.60	57.10	58.30	71.20	66.30
14	66.00	55.00	64.90	58.60	61.40	47.00	59.60	79.50	65.00	71.00	70.10	50.50	68.30	0.00	68.70	74.00	70.10	64.00	57.80	59.20	64.90	58.90	61.40	71.80	49.40	51.00	70.00	64.80	78.50	64.20	67.00	52.00	42.30	64.10	62.20	55.00	59.50	58.10	60.00	57.00
15	76.30	58.80	67.40	65.00	78.80	59.60	75.50	81.20	71.30	73.60	83.70	70.90	64.30	68.70	0.00	77.80	80.30	60.20	67.40	63.30	49.00	66.00	68.90	65.60	60.90	60.40	77.80	78.00	83.10	75.80	58.30	63.40	68.80	75.20	71.70	60.40	67.40	51.80	79.40	63.30
16	75.90	75.00	74.70	59.90	77.70	69.80	74.60	93.10	83.90	82.70	85.00	58.60	82.70	74.00	77.80	0.00	89.00	66.10	68.30	74.80	70.20	72.30	74.30	81.70	66.70	66.60	70.70	67.70	92.00	81.60	74.30	63.80	74.40	78.40	70.30	69.40	78.00	66.70	70.40	74.30
17	78.90	63.60	79.00	66.10	59.40	48.20	67.10	65.80	53.30	79.60	65.40	61.60	76.60	70.10	80.30	89.00	0.00	70.60	64.70	68.50	71.50	61.30	68.80	78.50	69.60	58.00	75.00	69.20	85.20	61.50	63.00	67.00	65.30	71.30	71.40	74.60	65.00	62.90	64.30	63.70
18	62.30	44.80	56.60	56.40	72.40	52.90	48.40	75.20	58.80	70.10	72.40	48.10	59.20	64.00	60.20	66.10	70.60	0.00	53.90	50.40	64.40	54.90	54.80	63.00	56.70	47.90	66.70	60.50	78.80	61.90	63.60	53.10	56.60	61.80	57.20	56.80	50.90	46.90	65.50	65.00
19	57.20	52.30	57.00	56.00	62.70	45.80	53.30	73.00	59.90	63.80	71.50	44.10	64.80	57.80	67.40	68.30	64.70	53.90	0.00	65.40	61.20	53.40	63.60	73.00	59.90	47.70	73.70	67.00	75.30	67.20	62.70	52.50	54.20	60.50	55.60	53.60	53.30	56.80	52.90	55.80
20	61.00	45.50	63.20	57.00	67.50	57.60	57.30	69.70	52.30	72.50	63.30	54.50	66.00	59.20	63.30	74.80	68.50	50.40	65.40	0.00	52.80	56.10	54.80	55.70	57.30	48.00	62.30	63.30	75.20	57.40	58.90	53.10	41.30	58.70	63.80	67.10	58.30	43.60	61.80	57.40
21	64.20	46.20	60.30	51.20	62.30	57.20	63.60	65.10	62.90	62.00	75.70	56.70	58.50	64.90	49.00	70.20	71.50	64.40	61.20	52.80	0.00	60.20	63.90	49.30	52.70	45.80	65.50	66.50	76.30	59.00	55.70	47.60	59.20	65.30	64.00	61.40	60.00	48.50	75.00	38.20
22	53.60	47.20	63.10	51.50	63.00	46.00	48.60	68.30	59.00	51.00	51.30	49.70	69.40	58.90	66.00	72.30	61.30	54.90	53.40	56.10	60.20	0.00	50.50	65.10	51.40	42.50	51.10	51.90	49.30	63.80	60.70	43.50	49.20	54.30	60.30	50.50	51.90	48.00	61.00	53.20
23	59.60	52.90	59.60	54.80	55.90	49.90	43.10	74.40	61.00	61.90	70.60	48.30	60.10	61.40	68.90	74.30	68.80	54.80	63.60	54.80	63.90	50.50	0.00	55.90	68.50	42.00	59.40	48.90	60.30	62.00	59.20	50.60	53.30	51.50	52.30	62.70	53.10	50.60	63.40	62.60
24	70.10	45.20	62.30	59.50	66.20	64.80	55.60	77.90	67.80	73.60	80.60	61.70	39.50	71.80	65.60	81.70	78.50	63.00	73.00	55.70	49.30	65.10	55.90	0.00	59.10	54.20	60.50	73.80	66.50	65.20	62.30	59.50	61.70	58.20	63.40	74.20	58.90	59.30	74.30	59.70
25	60.10	41.20	65.80	58.60	64.90	50.20	58.20	78.60	61.70	57.20	59.10	52.40	59.20	49.40	60.90	66.70	69.60	56.70	59.90	57.30	52.70	51.40	68.50	59.10	0.00	51.50	61.10	61.70	78.00	61.20	63.90	50.00	51.50	58.70	61.10	51.30	48.60	53.10	60.90	56.80
26	55.10	34.40	53.20	45.90	48.60	37.60	39.20	65.80	53.20	58.90	64.10	34.50	54.10	51.00	60.40	66.60	58.00	47.90	47.70	48.00	45.80	42.50	42.00	54.20	51.50	0.00	58.30	50.30	62.90	47.10	58.00	33.90	38.20	54.10	52.40	55.20	42.60	36.00	55.90	40.60
27	69.90	63.20	74.30	58.20	68.30	63.10	59.40	76.40	71.60	65.70	72.20	60.70	70.10	70.00	77.80	70.70	75.00	66.70	73.70	62.30	65.50	51.10	59.40	60.50	61.10	58.30	0.00	60.50	60.40	71.30	80.50	53.90	62.50	62.50	64.70	77.60	63.40	60.10	64.00	69.00
28	41.30	63.30	72.20	54.60	70.00	53.70	54.40	73.90	66.00	55.70	64.10	44.90	78.70	64.80	78.00	67.70	69.20	60.50	67.00	63.30	66.50	51.90	48.90	73.80	61.70	50.30	60.50	0.00	69.30	51.40	65.30	58.10	61.90	53.40	65.30	57.00	66.40	58.00	71.20	66.90
29	72.00	66.30	72.60	67.20	81.50	71.40	67.40	85.70	80.90	71.30	79.70	71.10	75.70	78.50	83.10	92.00	85.20	78.80	75.30	75.20	76.30	49.30	60.30	66.50	78.00	62.90	60.40	69.30	0.00	75.70	79.30	69.40	71.40	71.00	81.90	81.70	75.70	75.10	81.50	69.50
30	59.60	50.70	64.50	55.50	72.70	56.50	58.70	68.80	51.00	71.70	63.10	54.10	69.20	64.20	75.80	81.60	61.50	61.90	67.20	57.40	59.00	63.80	62.00	65.20	61.20	47.10	71.30	51.40	75.70	0.00	63.30	61.20	58.50	67.30	69.70	67.70	64.50	63.50	75.10	65.40
31	60.00	57.40	72.50	60.40	61.10	58.40	61.30	69.90	51.80	67.40	71.70	64.50	68.50	67.00	58.30	74.30	63.00	63.60	62.70	58.90	55.70	60.70	59.20	62.30	63.90	58.00	80.50	65.30	79.30	63.30	0.00	60.30	64.10	61.80	60.00	55.40	61.90	54.50	70.30	64.90
32	59.40	39.80	57.60	47.30	48.60	44.70	48.30	65.90	54.50	54.20	63.10	46.10	59.00	52.00	63.40	63.80	67.00	53.10	52.50	53.10	47.60	43.50	50.60	59.50	50.00	33.90	53.90	58.10	69.40	61.20	60.30	0.00	43.50	54.50	58.10	53.10	42.40	33.40	61.00	50.80
33	61.60	44.40	55.40	55.90	59.60	46.60	51.10	66.80	50.60	67.50	59.70	44.40	66.90	42.30	68.80	74.40	65.30	56.60	54.20	41.30	59.20	49.20	53.30	61.70	51.50	38.20	62.50	61.90	71.40	58.50	64.10	43.50	0.00	50.10	62.90	58.60	48.90	44.50	59.10	51.70
34	56.20	53.80	72.60	53.30	65.80	52.40	52.60	59.70	63.00	64.40	62.30	43.90	68.40	64.10	75.20	78.40	71.30	61.80	60.50	58.70	65.30	54.30	51.50	58.20	58.70	54.10	62.50	53.40	71.00	67.30	61.80	54.50	50.10	0.00	60.80	57.60	54.50	53.00	60.30	63.00
35	61.10	57.10	72.20	57.70	57.40	49.80	42.70	75.70	65.40	62.70	75.70	54.30	66.60	62.20	71.70	70.30	71.40	57.20	55.60	63.80	64.00	60.30	52.30	63.40	61.10	52.40	64.70	65.30	81.90	69.70	60.00	58.10	62.90	60.80	0.00	64.80	54.30	60.80	50.40	65.60
36	48.00	58.40	64.60	58.90	69.80	47.90	61.70	72.30	63.10	58.40	60.60	52.80	73.60	55.00	60.40	69.40	74.60	56.80	53.60	67.10	61.40	50.50	62.70	74.20	51.30	55.20	77.60	57.00	81.70	67.70	55.40	53.10	58.60	57.60	64.80	0.00	57.80	50.30	65.30	65.00
37	66.30	32.80	67.10	57.00	49.90	49.50	39.70	75.10	44.30	64.00	59.00	50.60	57.10	59.50	67.40	78.00	65.00	50.90	53.30	58.30	60.00	51.90	53.10	58.90	48.60	42.60	63.40	66.40	75.70	64.50	61.90	42.40	48.90	54.50	54.30	57.80	0.00	48.40	55.40	55.90
38	58.60	46.40	56.80	53.30	50.20	43.40	51.60	62.80	57.70	60.80	63.50	47.60	58.30	58.10	51.80	66.70	62.90	46.90	56.80	43.60	48.50	48.00	50.60	59.30	53.10	36.00	60.10	58.00	75.10	63.50	54.50	33.40	44.50	53.00	60.80	50.30	48.40	0.00	56.90	54.50
39	67.60	62.00	76.40	63.40	55.50	49.40	56.10	77.10	67.50	77.90	66.40	51.20	71.20	60.00	79.40	70.40	64.30	65.50	52.90	61.80	75.00	61.00	63.40	74.30	60.90	55.90	64.00	71.20	81.50	75.10	70.30	61.00	59.10	60.30	50.40	65.30	55.40	56.90	0.00	70.60
40	71.10	44.30	64.70	46.20	58.50	52.30	59.20	67.90	65.10	70.60	70.60	47.00	66.30	57.00	63.30	74.30	63.70	65.00	55.80	57.40	38.20	53.20	62.60	59.70	56.80	40.60	69.00	66.90	69.50	65.40	64.90	50.80	51.70	63.00	65.60	65.00	55.90	54.50	70.60	0.00

Tabela 3. Soma das distâncias euclidianas de uma parcela (p) em relação as demais, às médias da soma das distâncias e intervalo inferior e superior de seus grupos obtidos conforme o método incremental.

Soma das distâncias de uma parcela (p) em relação as demais	Média das soma das distâncias das parcelas	Intervalo inferior	Intervalo superior
1946 (26)	**49**	**1897**	**1995**
2056(2)	51	2005	2107
2078(32)	52	2026	2078
2083(6)	**52**	**2031**	**2135**
2084(12)	52	2032	2136
2106(38)	53	2053	2159
2163(22)	54	2110	2217
2180(7)	55	2125	2235
2184(33)	**55**	**2129**	**2239**
2209(37)	55	2154	2264
2236(4)	56	2180	2292
2269(23)	57	2212	2326
2289(25)	57	2232	2346
2316(20)	**58**	**2258**	**2374**
2330(18)	58	2272	2388
2334(21)	58	2276	2392
2339(19)	59	2280	2398
2344(40)	59	2285	2403
2350(34)	59	2291	2409
2394(36)	60	2334	2454
2413(9)	60	2353	2473
2421(28)	60	2361	2481
2423(14)	**60**	**2363**	**2483**
2423(35)	60	2363	2483
2467(1)	62	2405	2529
2474(31)	62	2412	2474
2479(30)	62	2417	2536
2489(5)	62	2427	2489
2495(24)	62	2433	2551
2528(39)	**63**	**2465**	**2591**
2568(3)	64	2504	2568
2576(27)	64	2512	2660
2590(13)	65	2525	2655
2604(10)	65	2539	2669
2670(11)	**67**	**2603**	**2737**
2675(17)	67	2608	2742
2689(15)	67	2622	2756
2844(8)	71	2773	2915
2886(29)	**72**	**2814**	**2958**
2907(16)	73	2834	2980

Pelo método incremental foram formados oito grupos (Tabela 4). Observa-se que o grupo I, formado apenas pela parcela 26, pode ser um grupo discrepante (*outliers*) que merece uma atenção especial do pesquisador. Entre esses grupos, quatro deles possuem o mesmo número de parcelas, grupo II e III com cinco parcelas e o grupo IV e V com oito parcelas, o grupo VI com seis parcelas, o grupo VII com quatro e o grupo VIII com três parcelas. Observa-se que os coeficientes de variação indica a consistência dos grupos, e que mostra-se os grupos III, IV, V e VIII com uma boa homogeneidade, e o grupo VI mostrou-se um pouco menos consistente com alta dispersão.

Tabela 4. Grupos de parcelas obtidos por meio do método incremental.

Grupos	Parcelas	Coeficiente de variação
Grupo I	26	-
Grupo II	2, 32, 6, 12 e 38	0,85
Grupo III	22, 7, 33, 37 e 4	1,29
Grupo IV	23, 25, 20, 18, 21 19, 40 e 34	1,22
Grupo V	36, 9, 28, 14, 35, 1, 31 e 30	1,30
Grupo VI	5, 24, 39, 3, 27 e 13	1,70
Grupo VII	10, 11, 17 e 15	1,42
Grupo VIII	8, 29 e 16	1,11

Observaram-se abaixo na Tabela (5) com a divisão de oito grupos (I, II, III, IV, V, VI, VII e VIII) e suas respectivas parcelas, e com a média, com o desvio padrão e com o coeficiente de variação das espécies arbóreas. A quantidade de grupos e os números de parcelas de cada grupo foram obtidos pelo método incremental.

Observou-se Grupo I, com uma parcela unitária e com a sua listagem das espécies arbóreas com respectiva média, desvio padrão e coeficiente de variação. Entre as espécies do Grupo I, de maior média, destacou-se *Parkia pendula* pelo seu alto valor e de menor média *Cupania racemosa*, e os restantes das espécies tiveram médias semelhantes.

Dentre as espécies arbóreas do Grupo II de maiores médias, relevou-se também *Nectandra cuspidata* e *Parkia pendula* pelos seus altos valores, *Helicostylis*

tomentosa, Inga thibaudiana e Miconia albicans também merecem destaque por ter sido as espécies com menores médias.

Nas espécies arbóreas do Grupo III, verificou-se que as maiores médias foram*, Brosimum discolor Schott, Licania rígida, Plathymenia foliolosa* e *Tapirira guianensis* pelos seus altos valores, *Ouratea hexasperma* também merecem atenção por terem sido a espécie com menor media.

Entre as espécies arbóreas no Grupo IV, de maiores médias, salientou-se também *Chrysophyllum splendens, Copaifera langsdorffii, Parkia pendula, Schefflera morototoni* e *Stryphnodendron pulcherrimum* pelos seus altos valores, *Miconia albicans e Cupania revoluta* também merecem atenção por terem sido as espécies com menores médias.

Dentre as espécies arbóreas no Grupo V, observou-se que as maiores médias foram *Copaifera langsdorffii, Ocotea gardneri, Ouratea hexasperma e Plathymenia foliolosa* pelos seus altos valores, *Cedrela, Nectandra cuspidata, Parkia pendula e Plathymenia foliolosa* também merecem destaque por terem sido as espécies com menores médias.

Nas espécies arbóreas do Grupo VI, de maiores médias, sobressaíram-se também *Dialium guianense, Mabea occidentalis e Tapirira guianensis* pelos seus altos valores, *Protium heptaphyllum e Cupania racemosa* também merecem atenção por terem sido as espécies com menores médias.

A partir das espécies arbóreas Grupo VII, de maiores médias, destacou-se também*, Cecropia palmata, Ocotea opifera, Parkia pendula, Schefflera morototoni e Stryphnodendron pulcherrimum* pelos seus altos valores, *Erythroxylum squamatum, Licania rígida e Miconia albicans* também merecem atenção por terem sido as espécies com menores médias.

Nas espécies arbóreas do Grupo VIII, de maiores médias, verificou-se que *Brosimum discolor, Ocotea gardneri e Tapirira guianensis* possui as maiores médias, *Cecropia palmata* também merecem destaque por terem sido a espécie com menor media.

O coeficiente de variação dos grupos (I, II, III, IV, V e VI) são semelhantes e com boa homogeneidade, o grupo VII com boa hogeneidade e o grupo VIII observou-se com o maior valor do coeficiente de variação.

Tabela 5. Distribuição das espécies arbóreas conforme Grupo e respectivas média, desvio padrão e coeficiente de variação para altura de Lorey.

Espécies	*Grupo (Parcelas)*							
	I (26)	II (2, 6, 12, 32 e 38)	III (4, 7, 22, 33 e 37)	IV (18, 19, 20, 21,23,25, 34 e 40)	V (1, 9,14, 28, 30, 31, 35 e 36)	VI (3,5,13,24,27 e 39)	VII (10, 11, 15 e 17)	VIII (8, 16 e 29)
Brosimum discolor	-	12,80	24,80	10,70	11,40	11,10	17,80	19,60
Cecropia palmata	-	8,60	-	12,50	10,25	14,50	20,30	4,50
Cedrela sp.	*5,00*	15,00	-	20,00	7,98	10,40	12,20	-
Chrysophyllum splendens	-	-	13,00	23,30	9,00	-	-	-
Copaifera langsdorffii	-	19,60	18,00	28,48	21,00	13,50	-	15,00
Cupania racemosa	*4,00*	10,00	22,90	12,70	9,92	5,00	-	9,50
Cupania revoluta	-	-	10,00	8,00	11,80	7,64	13,30	-
Dialium guianense	-	-	18,90	17,50	12,50	18,40	17,00	13,20
Eriotheca gracilipes	*6,00*	11,10	8,00	21,10	10,50	9,23	-	-
Erythroxylum squamatum	-	12,00	8,50	6,71	15,70	5,00	6,65	8,40
Eschweilera ovata	*5,00*	-	11,00	11,40	11,40	7,67	12,00	14,54
Helicostylis tomentosa	*7,00*	6,91	20,50	14,60	12,00	11,80	8,96	8,25
Inga thibaudiana	*8,00*	7,00	12,00	15,00	12,70	9,65	7,00	10,10
Licania rígida	-	-	25,30	15,85	19,34	-	-	-
Mabea occidentalis	-	8,00	12,00	-	13,00	18,00	-	-
Miconia albicans	-	7,00	11,41	7,87	14,00	-	7,00	10,00
Nectandra cuspidata	-	25,50	25,30	17,20	10,80	11,40	10,00	14,00
Ocotea gardneri	-	11,00	19,90	-	20,21	10,30	15,53	19,00
Ocotea opifera	-	-	9,00	14,50	8,17	7,36	20,00	8,00
Ouratea hexasperma	*4,00*	13,00	8,00	14,76	21,00	10,00	17,00	12,00
Parkia pendula	*14,41*	23,00	20,00	23,50	17,90	12,80	20,00	-
Plathymenia foliolosa	-	5,00	25,20	17,30	22,50	12,70	14,00	18,60
Pouteria grandiflora	-	12,00	12,08	10,70	11,00	14,80	10,00	11,00
Protium heptaphyllum	*5,66*	12,70	13,50	14,00	-	4,00	14,00	-
Pterocarpus violaceus	-	19,70	15,00	-	-	12,20	15,00	-
Schefflera morototoni	-	16,50	16,20	22,47	11,01	14,00	22,00	-
Stryphnodendron pulcherrimum	-	-	14,00	27,00	11,00	10,00	20,00	-
Tapirira guianensis	*6,16*	27,90	27,00	25,40	19,80	16,00	18,90	18,90
Thyrsodium spruceanum.	*4,67*		18,20	14,00	10,20	7,67	17,80	15,30
Média Geral (m)	*6,35*	13,54	16,29	20,11	13,56	10,97	14,63	12,77
Desvio padrão (m)	*2,94*	6,43	6,05	9,80	3,99	4,43	3,36	5,95
CV (%)	*46,30*	44,44	37,50	40,84	31,34	38,44	22,75	45,11

Verificando-se os resultados das análises da soma das distâncias de uma parcela em relação as demais, das parcelas de 1 até 40, utilizou-se os métodos gráficos, para verificar a normalidade das parcelas, observa-se (Figura 11), que a mesma tem distribuição normal sem a presença de *outliers*, e que, utilizando-se o método incremental (Tabela 3), com a soma das distâncias das parcelas ordenadas e seus respectivos intervalos, observamos na formação dos grupos pelo método incremental (Tabela 4), que houve um parcela (26) isolado ou seja um grupo outliers, que não foi confirmado pelo Boxplot e pelo Q-Q plot (Figura 11).

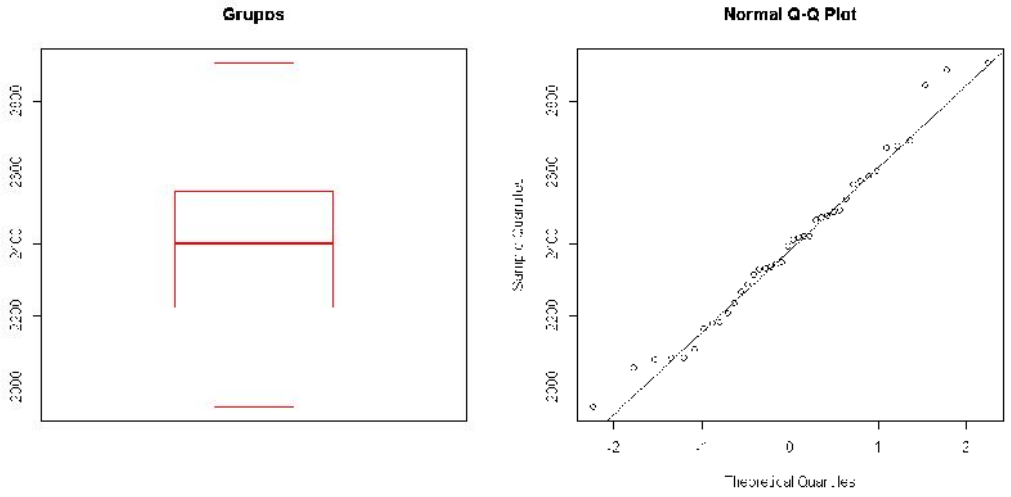

Figura 11 – Boxplots e Q-Q plot das somas das distâncias de uma parcela em relação as demais para identificar outliers.

9.3 Métodos Hierárquicos

9.3.1 Normalidade das distâncias euclidianas entre parcelas

Os *outliers* distorcem a verdadeira estrutura e tornam os grupos derivados não representativos da verdadeira estrutura da população (parcelas). Por essa razão, uma triagem preliminar, em busca de *outliers,* é necessária. Provavelmente, a maneira mais fácil de conduzir essa triagem é por meio de um *boxplot* ou um *Q-Q plot.*

Na Figura 12, pode-se observar a presença de 18 *outliers* (círculos) nas parcelas das distâncias euclidianas, parcelas (1, 2, 6, 7, 9, 12, 20, 21, 22, 23, 26, 29, 30, 31, 32, 34, 35 e 39).

Observa-se (Figura 13), Q-Q plot da matriz de distância euclidiana das 40 parcelas mostraram falta de normalidade, que as distâncias euclidianas entre as parcelas não têm uma distribuição normal. Porém, nesses casos o recomendado-se a padronização das parcelas, por conta das diferentes escalas de medida. No nosso caso não foi necessário, pois apenas se deseja obter novos grupos.

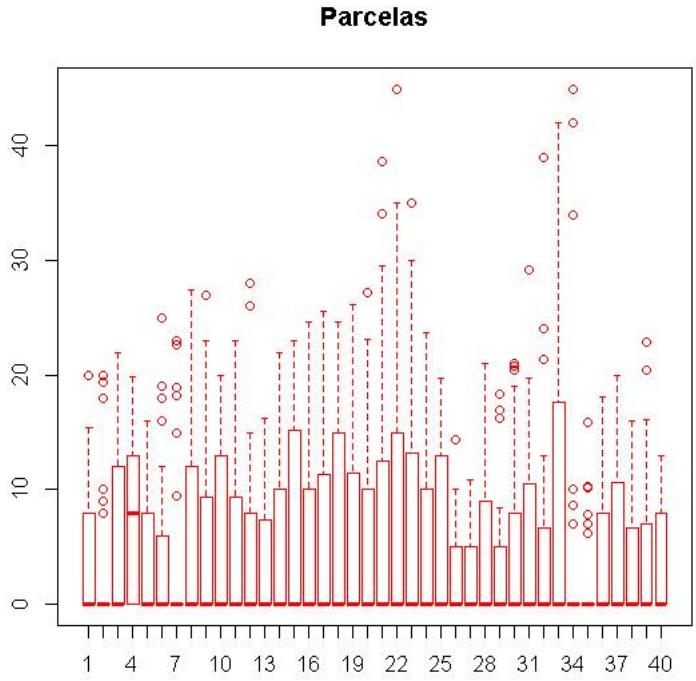

Figura 12 – Boxplots das distâncias para identificar *outliers*

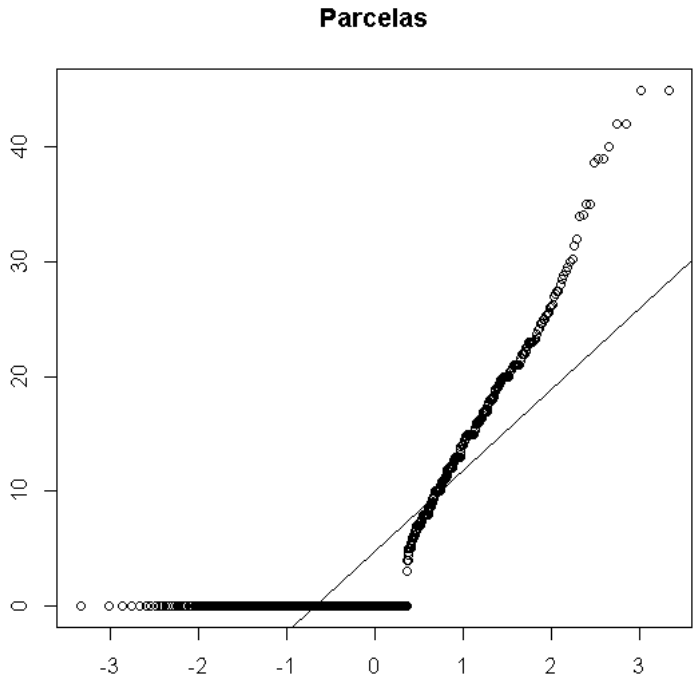

Figura 13 – Q-Q plot das distância para identificar *outliers*.

9.4 Dendrogramas

9.4.1 Inspeção visual

Uma observação visual dos dendrogramas pode ser feita com base nas (Figuras 14, 15, 16 e 17). Para os dendrogramas obtidos, observa-se a presença de oito grupos em cada figura, o algoritmo de ligação simples indica a presença de sete grupos unitários (I, II, III, IV, V, VI e VII) com as respectivas parcelas (33, 34, 21, 8, 22, 19, e 3), ligação completa indica a presença de cinco grupos unitários (I, IV, V, VI e VII) com as respectivas parcelas (34, 21, 23, 33 e 22), o algoritmo de ligação média indica a presença de cinco grupos unitários (I, II, III, IV e V) com as respectivas parcelas (33, 21, 22, 34, e 23), e Ward com dois grupos unitários (I e IV) com as respectivas parcelas (33 e 21). Se um destes algoritmos fosse escolhido, haveria a necessidade de considerar um grupo para o algoritmo de ligação simples, três grupos, para o algoritmo de ligação completa e ligação

média e seis grupos, para o algoritmo Ward, pois os mesmos têm dois grupos unitário. Os dendrogramas dos algoritmos apresentam aspectos diferentes, sendo poucas as parcelas em grupos idênticos, porém o número de parcelas incluídas em cada um dos grupos foi bem distinto. O método de Ward apresentou a solução mais apropriada ao problema, sugerindo-se a construção de oito grupos com diferentes números de parcelas.

É importante destacar-se que o fato desse tipo de análise não apresenta um critério objetivo para identificação dos grupos e dificulta muito a interpretação dos resultados.

Embora a estrutura geral dos agrupamentos seja bastante similar, pode-se observar que há pequenas alterações nos níveis em que as parcelas são agrupadas, ou seja, as parcelas que estão dentro de um mesmo grupo podem ser agrupadas em outra ordem, quando se mudam os algoritmos. Entretanto, isso causa poucos problemas práticos.

Pode-se observar que há divergências entre os algoritmos, corroborando com a afirmativa de Johnson e Wichern (2007). De forma geral, os dendrogramas apresentaram estruturas de agrupamentos de parcelas homogêneos, embora não exista critério objetivo para determinar um ponto de corte no dendrograma, ou seja, para determinar quais os grupos foram formados. No entanto, segundo Bussab et al. (1990), a grande vantagem do dendrograma é permitir observar graficamente o quanto é necessário “relaxar” o nível de dissimilaridade para considerar grupos próximos.

Observa-se que, como não existe critério, para determinar um ponto de corte no dendrograma, os pesquisadores utilizam-se de um percentual das respectivas distâncias para um ponto de corte no dendrograma, assim, dependendo do algoritmo utilizado, obtém-se um número de grupos diferente. Nesse trabalho, optou-se por determinar o número de grupos estabelecido no método incremental (Tabela 4), em vez de cortar o dendrograma aleatoriamente.

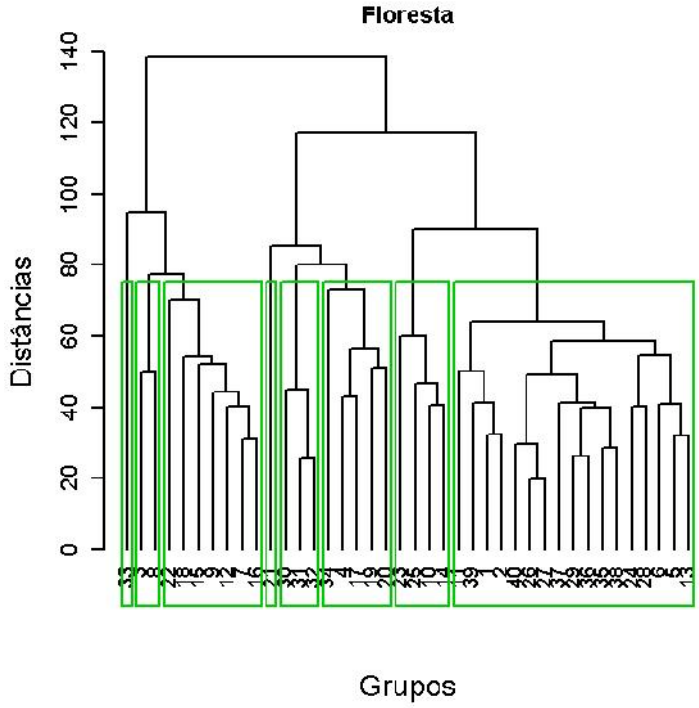

Figura 14. Dendrograma obtido por meio do algoritmo de Ward, baseando-se na distância euclidiana.

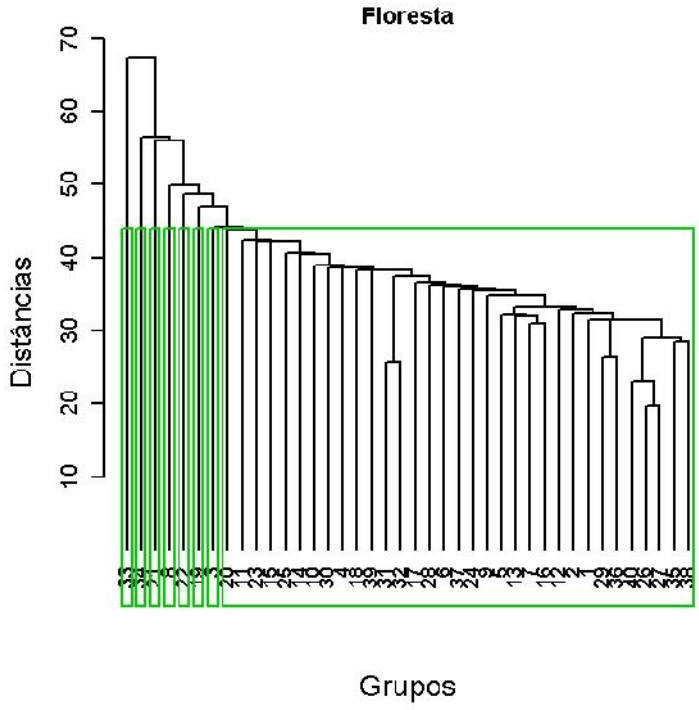

Figura 15. Dendrograma obtido por meio do algoritmo de ligação simples, baseando-se na distância euclidiana.

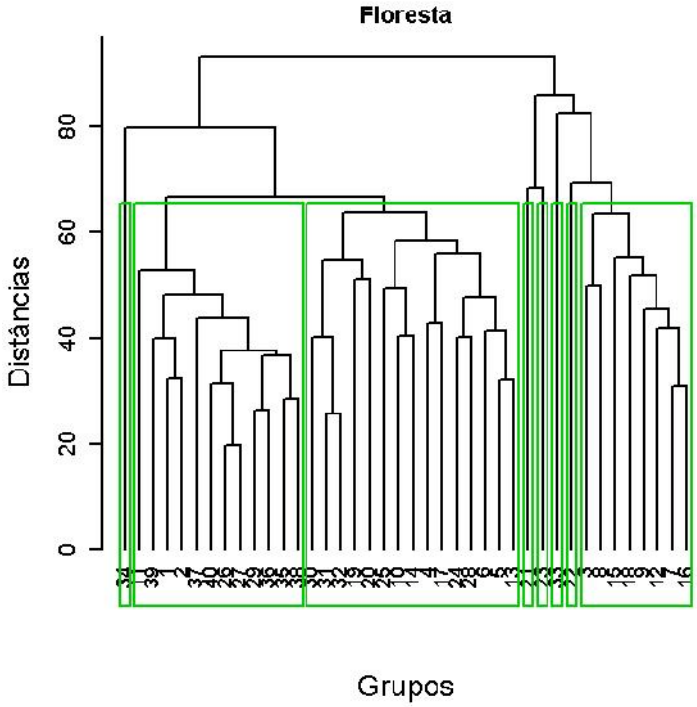

Figura 16. Dendrograma obtido por meio do algoritmo de ligação completa, baseando-se na distância euclidiana.

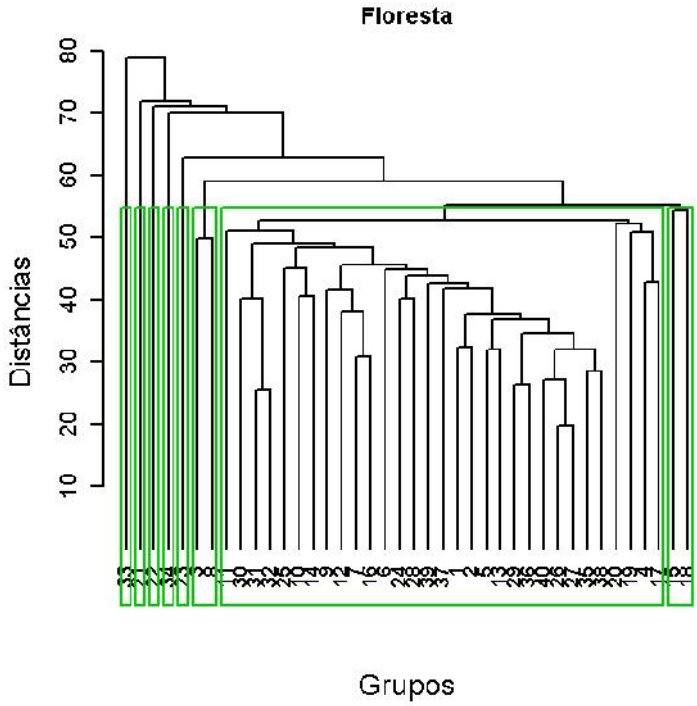

Figura 17. Dendrogramas obtido por meio do algoritmo de ligação média, baseando-se na distância euclidiana.

Máximo (2009) descreveu que, além de essa técnica determinar um número de grupo, dois requisitos básicos são levados em consideração: maior coesão interna e menor dispersão interna (variância) dos dados dentro dos grupos; maior isolamento possível entre os grupos gerados, isto é, cada grupo formado possui características distintas um do outro. Vale a pena repetir que o objetivo mais comum da análise de agrupamentos talvez seja tratar homogeneidade, dentro dos grupos; e da heterogeneidade, entre os grupos nos dados. O resultado esperado é um pequeno número (administrável) de grupos, cada um consistindo-se em um número de parcelas relativamente homogêneos, com uma variação dentro do grupo consideravelmente menor do que o total de variação no conjunto completo de dados (ou dados originais).

O uso da técnica de análise de agrupamento pode auxiliar bastante o pesquisador na construção de grupos, baseando-se em informações de mais de uma característica. Na decisão da melhor solução, recomenda-se que o pesquisador avalie a qualidade dos agrupamentos obtidos, compare as variâncias internas de cada grupo e a variância total da matriz de distância.

Vale ressaltar que, na matriz euclidiana, foram estimadas a média de 54,2, a mediana 52,7, o desvio padrão 12,7, o máximo e o mínimo das distâncias, respectivamente, de 93 e 19,8, e com valor do coeficiente de variação de 23,43%, observando-se uma boa homogeneidade.

Observou-se (Tabela 6), e considerando-se, os dados da matriz euclidiana e do algoritmo de Ward com oito grupos: os (grupo I e IV), com uma parcela; grupo II, com duas parcelas; com valor do coeficiente de variação de 8,89%, verificou-se uma ótima homogeneidade; grupo III, com sete parcelas também com valor do coeficiente de variação de 17,94%, observando-se uma boa homogeneidade; grupo V, com três parcelas; com valor do coeficiente de variação de 25%, obteve-se uma média homogeneidade; grupo VI, com cinco parcelas; com valor do coeficiente de variação de 17,74%, observando-se uma boa homogeneidade; grupo VII, com quatro parcelas; com valor do coeficiente de variação de 3,74%, observando-se uma excelente homogeneidade; grupo VIII, com dezessete parcelas; com valor do coeficiente de variação de 20,6%, observou-se uma boa homogeneidade, o método apresenta se com dois grupos

com uma parcela ou seja, *outliers*, precisando de uma atenção do pesquisador para esses grupos com *outliers*.

Tabela 6. Valor mínimo, médio, mediano, do desvio-padrão e máximo da distância euclidiana nos grupos obtidos pelo método de Ward.

Grupos	I	II	III	IV	V	VI	VII	VIII
Nº de parcelas	1	2	7	1	3	5	4	17
	2,5%	5,0%	17,5	2,5%	7,5%	12,5%	10,0%	42,5%
Mínimo	67,2	44,0	30,9	56,1	25,6	36,7	38,9	19,8
Médio	67,2	46,9	37	56,1	30,0	44,5	40,6	30,9
Mediano	67,2	46,9	34,8	56,1	32,1	43,7	40,6	32
Desvio. padrão	-	4,17	6,64	-	7,51	7,88	1,37	6,39
Máximo	67,2	49,9	56,6	56,1	38,6	56,6	42,3	42,4

Observam-se os valores descritivos da matriz euclidiana e de ligação simples (Tabela 6), com oito grupos: os (grupo I, II, III, IV, V, VI e VII) com uma parcela; grupo VIII, com trinta e três parcelas; com valor do coeficiente de variação de 19,64%, observando-se uma boa homogeneidade. Observam-se que (Tabela 7), e no dendrograma (Figura 15), que apresenta uma forma de encadeamento e que o método de ligação simples apresentam-se com sete dos grupos isolados, ou seja, *outliers,* desconsiderar os grupos para obter-se uma solução mais satisfatória, porque nenhuma outra parcela foi considerada similar, cabe ao pesquisador tão decisão, as parcelas e grupo VIII com variabilidade inferior ao da matriz euclidiana.

Tabela 7. Valor mínimo, médio, mediano, do desvio-padrão e máximo da distância euclidiana nos grupos obtidos pelo método de ligação simples.

Grupos	I	II	III	IV	V	VI	VII	VIII
Nº de parcelas	1	1	1	1	1	1	1	33
	2,5%	2,5%	2,5%	2,5%	2,5%	2,5%	2,5%	82,5
Mínimo	67,2	56,6	56,1	49,9	48,8	46,9	44	19,8
Médio	67,2	56,6	56,1	49,9	48,8	46,9	44	33,5
Mediana	67,2	56,6	56,1	49,9	48,8	46,9	44	34,8
Desvio padrão	-	-	-	-	-	-	-	6,58
Máximo	67,2	56,6	56,1	49,9	48,8	46,9	44	43,7

A estatística descritiva, (Tabela 8), da matriz euclidiana e do algarismo de ligação completa com oito grupos: os (grupo I, IV, V, VI, VII) com uma parcela; grupo II, com doze parcelas, com valor do coeficiente de variação de 23,94%, com uma média homogeneidade, grupo III, com quinze parcelas, com valor do coeficiente de variação de 16,13%, observou-se uma boa homogeneidade, grupo VIII, com oito parcelas, com valor do coeficiente de variação de 17,36%, com uma boa homogeneidade. Verificou-se (Tabela 7), e no dendrograma (Figura 19), que o método de ligação completa apresentou-se com a maioria dos grupos isolados, ou seja, *outliers.*

Tabela 8. Valor mínimo, médio, mediano, do desvio-padrão e máximo da distância euclidiana nos grupos obtidos pelo método de ligação completa.

Grupos	I	II	III	IV	V	VI	VII	VIII
Nº de	1	12	15	1	1	1	1 2,5%	8
parcelas	2,5%	30,0%	37,5%	2,5%	2,5%	2,5%		20,0
Mínimo	56,6	19,8	25,6	56,1	42,3	67,2	48,8	30,9
Médio	56,6	29,4	36,5	56,1	42,3	67,2	48,8	38
Mediano	56,6	28,5	36,7	56,1	42,3	67,2	48,8	36,5
D. padrão	-	7,04	5,89	-	-	-	-	6,9
Máximo	56,6	42,4	46,9	56,1	42,3	67,2	48,8	49,9

De acordo com a estatística descritiva (Tabela 9), da matriz euclidiana e de ligação média com oito grupos dois quais com cinco grupos unitarios: os (grupo I,

II, III, IV, V) com uma parcela, apresentou-se com um *outliers,* e que deve se ser observado pelo pesquisador, grupo VI, com duas parcelas, com valor do coeficiente de variação de 8,89%, verificou-se uma ótima homogeneidade; grupo VII, com trinta e um parcela; com valor do coeficiente de variação de 20,40% observou-se uma boa homogeneidade; grupo VIII, com duas parcelas; com valor do coeficiente de variação de 5,69%, obteve-se uma excelente homogeneidade, todos os grupos com mais de uma parcela, mostram-se com melhor variância do que a matriz de distância euclidiana.

Tabela 9. Valor mínimo, médio, mediano, do desvio-padrão e máximo da distância euclidiana nos grupos obtidos pelo método de ligação média.

Grupos	I	II	III	IV	V	VI	VII	VIII
Nº de	1	1	1	1	1	2	31	2
parcelas	2,5%	2,5%	2,5%	2,5%	2,5%	5,0%	77,5%	5,0%
Mínimo	67,2	56,1	48,8	56,6	42,3	44,0	19,8	38,2
Médio	67,2	56,1	48,8	56,6	42,3	46,9	33,2	40,2
Mediano	67,2	56,1	48,8	56,6	42,3	46,9	32,8	40,2
D. padrão	-	-	-	-	-	4,17	6,79	2,99
Máximo	67,2	56,1	48,8	56,6	42,3	49,9	46,9	42,3

Observando-se os desvios padrões dos grupos, apresentados nas (Tabelas 6, 7, 8 e 9), percebe-se que o algoritmo do Ward apresentou com dois grupos unitários (grupo I e IV) e que os grupos II, III, V, VI, VII, VIII com desvio padrões inferior ao da matriz distância euclidiana. O algoritmo de ligação simples apresentou-se com sete grupos unitários (grupo I, II, III, IV, V, VI e VII), o grupo VIII, com o desvio padrão inferior ao da matriz de distância euclidiana. O algoritmo de ligação completa apresentou com quatro grupos unitários (grupo I, IV, V e VI), os (grupos II, III, VI e VIII) com o desvio padrão inferior ao da matriz euclidiana e o algoritmo de ligação média apresentaram-se com cinco grupos unitários (grupo I, II, III, IV e V), e o (grupo VI, VII e VIII) com o desvio padrão inferiores ao da matriz de distância euclidiana.

9.5 Dados artificiais

9.5.1 Análise dos dados artificiais e eficiência dos métodos

Neste trabalho, foi realizada uma análise de agrupamento e uma validação do método incremental, de Ward, de ligação simples, de ligação completa e de ligação média. A primeira parte da análise tem o objetivo de verificar se o número de grupos determinado pelo método incremental e um algoritmo executado diversas vezes chega a resultados parecidos. Isso porque as abordagens heurísticas tratadas neste trabalho contêm componentes aleatórias em suas várias fases, e, se executados diversas vezes, podem chegar a soluções diferentes. A ideia é que, mesmo se executado diversas vezes, um método não apresente soluções tão diferentes. A segunda parte da análise aborda a eficiência dos algoritmos através da comparação do índice Rand ajustado encontrado por todos os métodos para algumas bases de dados.

9.5.2 Dados artificiais I

O conjunto de dados artificiais I, observado pelo método incremental, mostrou a formação de seis grupos (Figura 18 e Tabela 10). È possível visualizar, na Figura 18, o resultado da formação dos grupos através de uma dispersão, onde a cor indica o grupo à qual pertence e na Tabela 10, e mostra os grupos com suas respectivas parcelas, e ao qual serão aplicados os algoritmos Ward, de ligação simples, de ligação completa e de ligação média. O grupo VI, formado apenas pelas 27 e 44, pode ser um grupo discrepante que merece uma atenção especial do pesquisador.

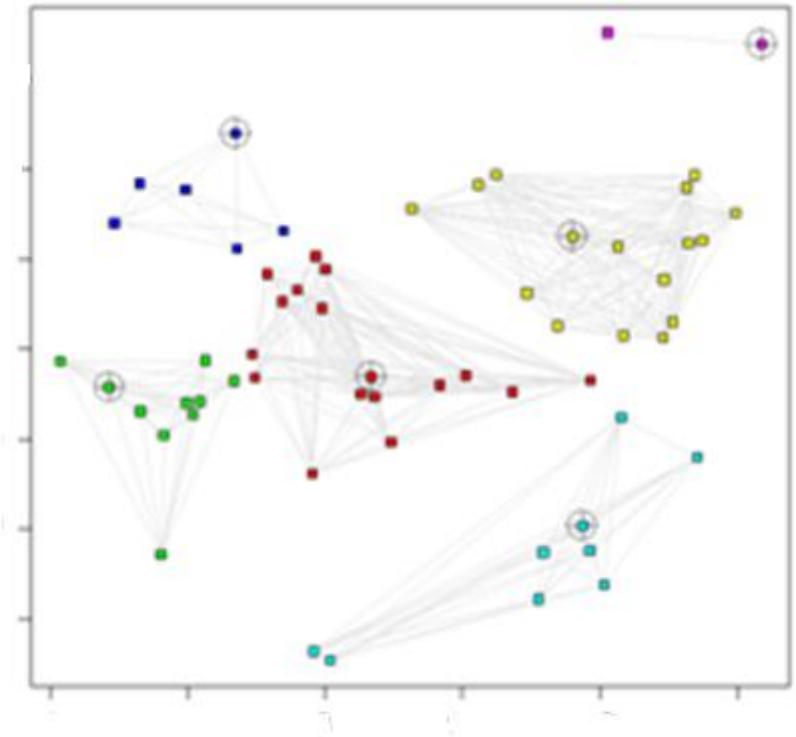

Figura 18. Dispersão obtida por meio do método incremental, baseando-se no dado artificial I.

Pelo método incremental baseando-se nos dados artificiais I foram formados seis grupos (Tabela 10). Observa-se que não foi apresentado nenhum *outliers.*

Tabela 10. Grupos de parcelas obtidos da distância euclidiana por meio do método incremental, baseando-se no dado artificial I.

Grupos	Parcelas
Grupo I	1, 9, 14, 15, 20, 23, 24, 25, 26, 29, 30, 32, 41, 46, 51, 53 e 57
Grupo II	3, 6, 7, 8, 10, 17, 28, 33, 42, 43, 47, 50, 55, 58, 59 e 60
Grupo III	5, 2, 4, 13, 16, 18, 31, 37, 39 e 45
Grupo IV	35, 11, 12, 19, 21, 34, 36, 48 e 54
Grupo V	40, 22, 38, 49, 52, 56,
Grupo VI	27 e 44

Notam-se, na Tabela 11, que os valores do coeficiente de variação são semelhantes para os grupos I, II, III e IV, e que mostraram que têm uma boa homogeneidade, os grupos V e VI, mostraram-se com valores semelhantes, mas com valores altos.

Tabela 11 Distribuição das espécies arbóreas conforme Grupo e respectivas média, desvio padrão e coeficiente de variação para altura de Lorey, baseando-se no dado artificial I.

Espécies	*Grupo (Parcelas)*					
	I (1, 9, 14, 15, 20, 23, 24, 25, 26, 29, 30, 32, 41, 46, 51, 53 e 57)	II (3, 6, 7, 8, 10, 17, 28, 33, 42, 43, 47, 50, 55, 58, 59 e 60)	III (2, 4, 5, 13, 16, 18, 31, 37, 39 e 45)	IV (11,12,19, 21,34, 35, 36, 48 e 54)	V (22, 38,40, 49, 52 e 56)	VI (27 e 44)
Brosimum discolor	12,40	11,37	17,42	20,60	9,11	14,41
Cecropia palmata	9,25	13,50	10,25	4,50	-	5,47
Cedrela sp.	8,98	20,22	7,98	8,45	-	3,50
Chrysophyllum splendens	9,00	23,30	9,00	-	-	-
Copaifera langsdorffii	19,00	18,48	21,00	15,00	3,00	12,00
Cupania racemosa	9,92	14,70	9,92	9,50	2,04	2,50
Cupania revoluta	13,80	8,00	11,80	-	3,33	5,00
Dialium guianense	12,50	17,50	12,50	13,20	13,76	15,74
Eriotheca gracilipes	11,50	21,10	10,50	-	1,00	4,50
Erythroxylum squamatum	14,70	6,71	15,70	8,40	-	-
Eschweilera ovata	13,40	11,40	11,40	14,54	3,25	2,50
Helicostylis tomentosa	12,80	14,60	12,00	8,25	3,33	11,93
Inga thibaudiana	13,70	15,00	12,70	10,10	10,51	9,65
Licania rígida	19,84	15,85	19,34	11,23	6,83	-
Mabea occidentalis	13,43	-	13,00	-	2,17	6,00
Miconia albicans	14,23	7,87	14,00	10,00	-	5,60
Nectandra cuspidata	11,80	17,20	10,80	14,00	3,18	-
Ocotea gardneri	20,21	18,14	20,21	19,00	3,06	-
Ocotea opifera	9,17	14,50	8,17	8,00	-	-
Ouratea hexasperma	21,67	14,76	21,00	12,00	2,29	-
Parkia pendula	18,90	23,50	17,90	-	3,96	5,25
Plathymenia foliolosa	22,50	17,30	22,50	18,60	7,54	22,50
Pouteria grandiflora	11,00	10,70	11,00	11,00	10,79	7,48
Protium heptaphyllum	7,54	14,00	14,34	-	7,08	-
Pterocarpus violaceus	11,24	11,24	9,17	7,34	6,50	-
Schefflera morototoni	11,01	22,47	11,01	-	2,84	-
Stryphnodendron pulcherrimum	11,00	27,00	11,00	-	-	-
Tapirira guianensis	19,80	25,40	19,80	18,90	14,27	19,75
Thyrsodium spruceanum.	10,20	14,00	10,20	15,30	10,23	12,08
Média Geral (m)	*13,60*	16,06	*13,64*	12,28	5,91	9,21
Desvio padrão (m)	*4,24*	5,28	4,35	4,43	3,99	5,94
CV (%)	*31,19*	32,87	31,91	36,11	67,50	64,45

Observaram abaixo, na Tabela 11, com a divisão de seis grupos (I, II, III, IV, V e VI) e suas respectivas parcelas, com a média, com o desvio padrão e com o coeficiente de variação das espécies arbóreas. A quantidade de grupos, e os números de parcelas de cada grupo foram obtidos pelo método incremental.

Uma observação visual dos dendrogramas pode ser feita com base nas (Figuras 14, 15, 16 e 17) dos dados originais e nas (Figuras 19, 20, 21,22, 23, 24, 25, 26, 27, 28, 29, 30, 31 e 32) dos dados artificiais. Verificando que as estrutura gerais dos dentrogramas e dos agrupamentos são bastante similares, pode-se observar que há pequenas alterações nos níveis em que as parcelas são agrupadas nos respectivos métodos, ou seja, as parcelas que estão dentro de um mesmo grupo podem ser agrupadas em outra ordem, quando se mudam os algoritmos. Entretanto, isso causa poucos problemas práticos.

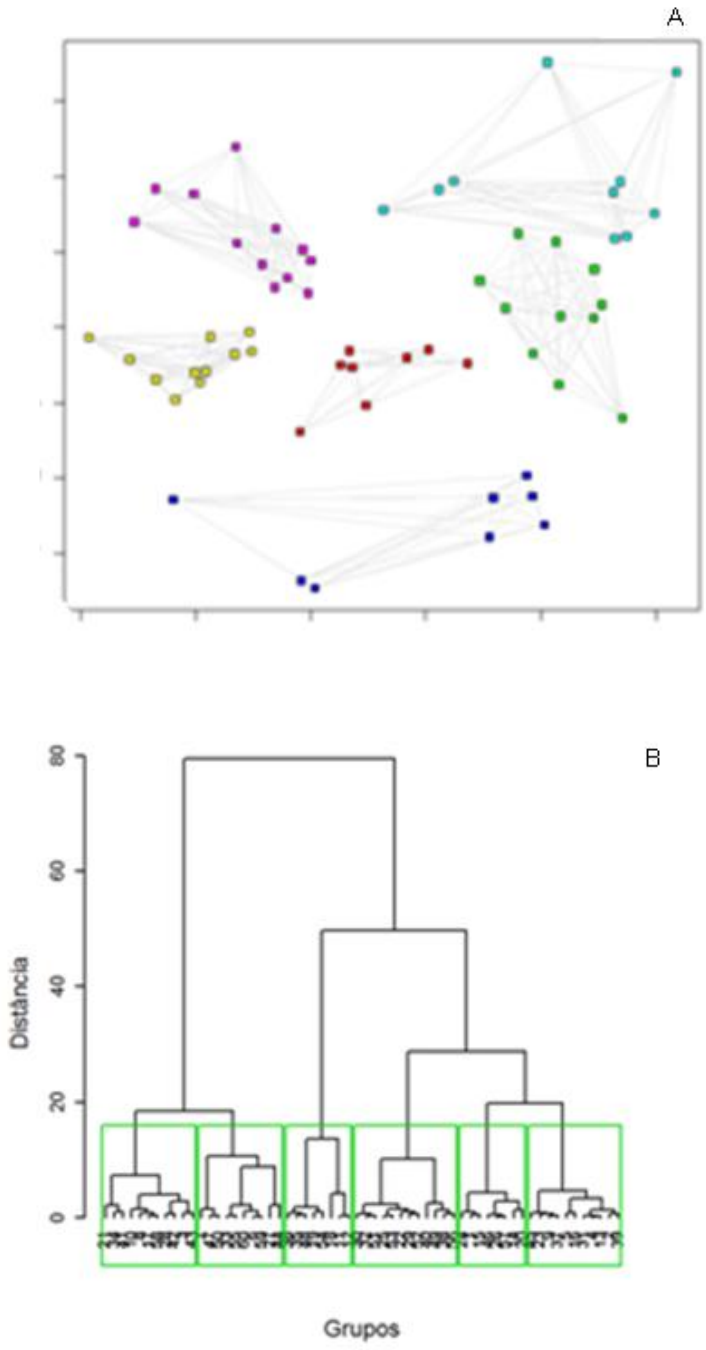

Figura 19. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de Ward, baseando-se no dado artificial I.

Observou-se, na (Figura 20), que mostra disprsão (A) e o dendrograma (B) típico de ligação simples. Há uma estrutura clara, ligando os grupos, que esta gradualmente se juntarem em um grande grupo, isolando cinco grupos até a etapa final.

Figura 20 (B) observa-se o encadeamento dos grupos, verificando-se uma das características do método ligação simples, e o isolamento de *outliers,* se dois grupos forem especificados (Se três grupos são especificados, o outro *outliers* é acantonada (recuperado), deixando ainda um grande agrupamento). Apesar da óbvia falta de sucesso na recuperação dos grupos *outliers*, um benefício potencial da aplicação de ligação simples, isto é que pode ser usado para identificar *outliers*, uma vez que esses são deixados como “singletons” se estiverem longe suficiente do seu vizinho mais próximo.

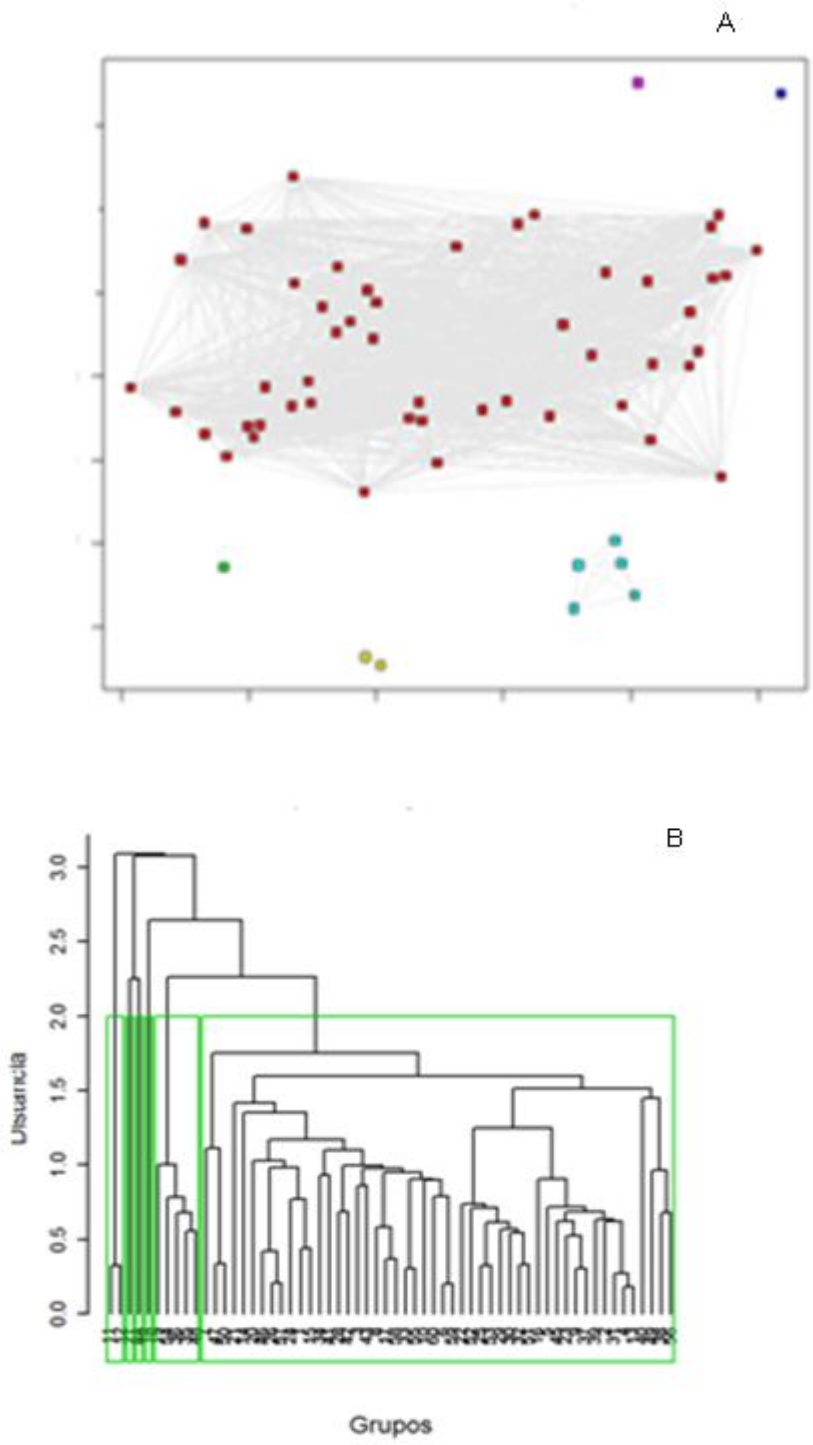

Figura 20. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de ligação simples, baseando-se no dado artificial I.

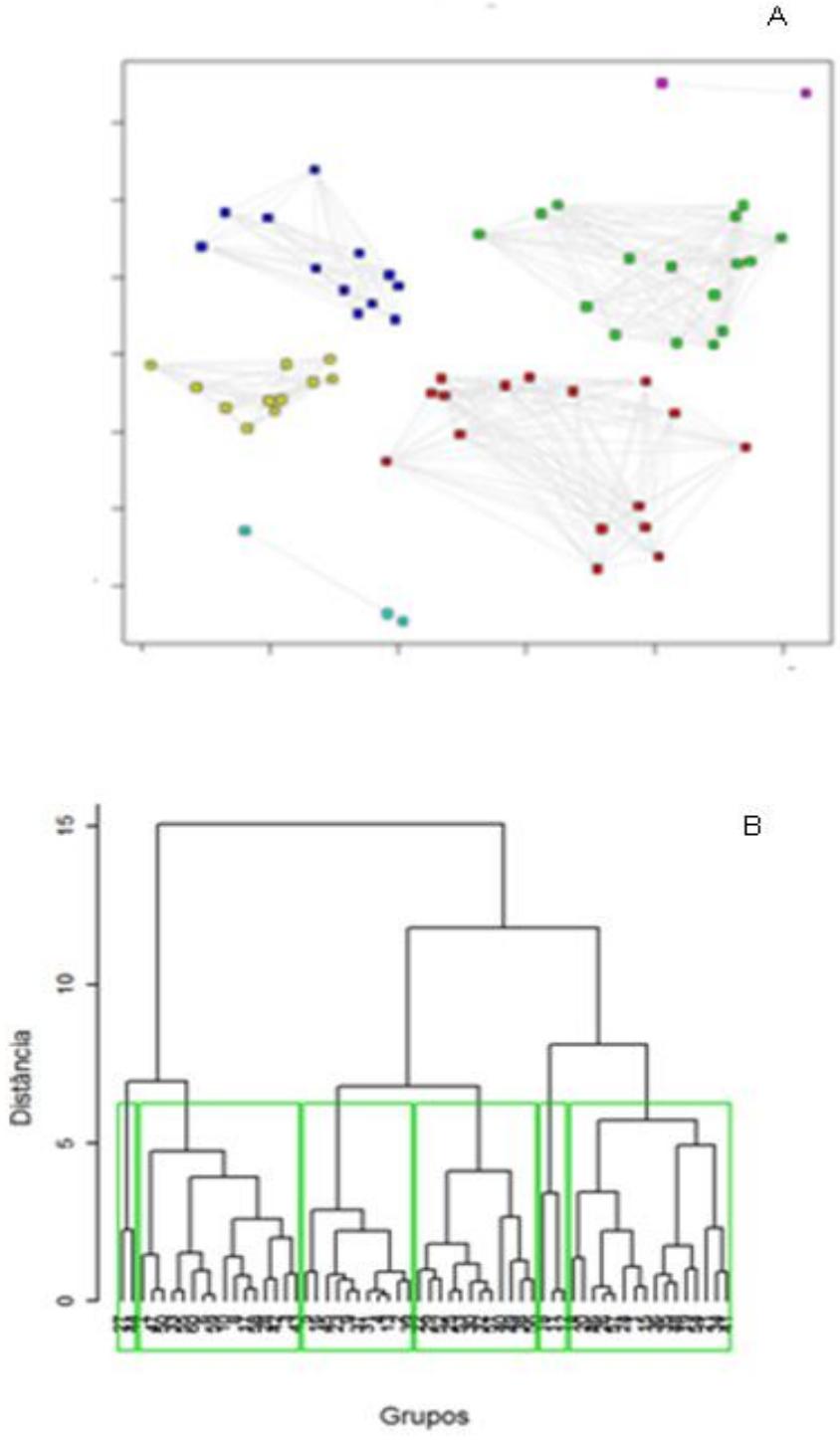

Figura 21. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de ligação completa, baseando-se no dado artificial I.

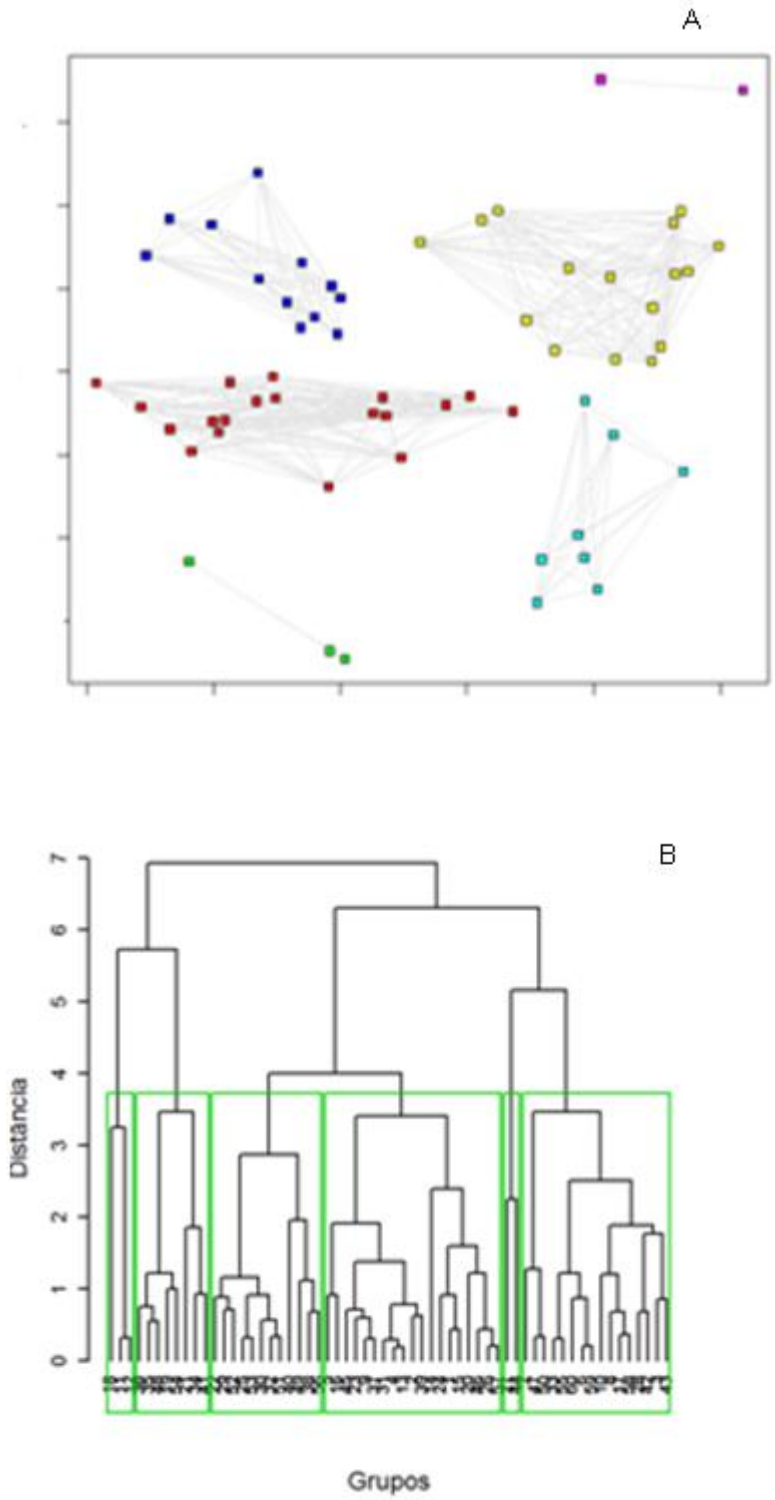

Figura 22. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de ligação média, baseando-se no dado artificial I.

9.5.3 Dados artificiais II

O conjunto de dados artificiais II, observado pelo método incremental, apresenta a formação de seis grupos (Figura 23 e Tabela 12). É possível visualizar o resultado da formação dos grupos através de uma dispersão, onde a cor indica o grupo à qual pertence e na Tabela 12, e mostra os grupos com suas respectivas parcelas, e aos quais serão aplicados os algoritmos Ward, de ligação simples, de ligação completa e de ligação média. O grupo VI, formado apenas pelas 27 e 44, pode ser um grupo discrepante que merece uma atenção especial do pesquisador.

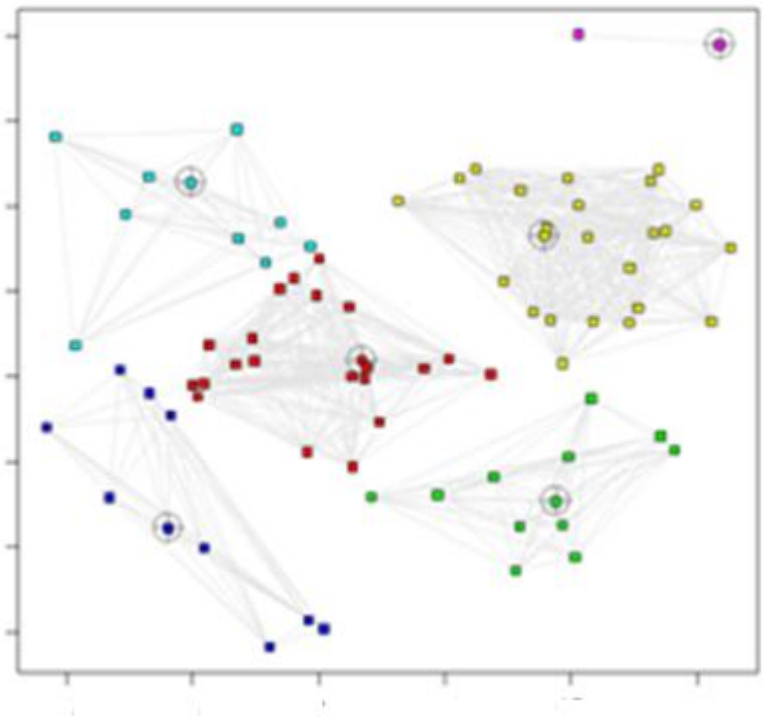

Figura 23. Dispersão obtida por meio do método incremental, baseando-se no dado artificial II.

Pelo método incremental baseando-se nos dados artificiais II foram formados seis grupos (Tabela 12). Observa-se que não foi apresentado nenhum *outliers.*

Tabela 12. Grupos de parcelas obtidos da distância euclidiana por meio do método incremental, baseando-se no dado artificial II.

Grupos	Parcelas
Grupo I	1, 4, 9, 13, 14, 15, 20, 23, 24, 26, 30, 31, 32, 37, 45, 46, 51, 53, 57, 61, 68 e 76
Grupo II	3, 6, 7, 8, 10, 17, 28, 33, 41, 42, 43, 47, 50, 55, 58, 59, 60, 66, 69, 70, 73, 74, 75 e 79
Grupo III	19, 21, 34, 35, 36, 48, 54, 62, 63, 64, 65 e 77
Grupo IV	16, 22, 25, 29, 38, 40, 49, 52, 56 e 72
Grupo V	2, 5, 11, 12, 18, 39, 67, 71, 78 e 80
Grupo VI	27 e 44

Observaram abaixo, na Tabela 13, com a divisão de seis grupos (I, II, III, IV, V e VI) e suas respectivas parcelas, com a média, com o desvio padrão e com o coeficiente de variação das espécies arbóreas. A quantidade de grupos, e os números de parcelas de cada grupo foram obtidos pelo método incremental.

Percebe-se, na Tabela 13, que os valores do coeficiente de variação são semelhantes para os grupos I, II, e III, IV e V, e mostrou-se que têm uma boa homogeneidade, o grupo VI, mostrou-se com valor alto.

Tabela 13. Distribuição das espécies arbóreas conforme Grupo e respectivas média, desvio padrão e coeficiente de variação para altura de Lorey, baseando-se no dado artificial II.

Espécies	*Grupo (Parcelas)*					
	I (1, 4, 9, 13, 14, 15, 20, 23, 24, 26, 30, 31, 32, 37, 45, 46, 51, 53, 57, 61, 68 e 76)	II (3, 6, 7, 8, 10, 17, 28, 33, 41, 42, 43, 47, 50, 55, 58, 59, 60, 66, 69, 70, 73, 74, 75 e 79)	III (19, 21, 34, 35, 36, 48, 54, 62, 63, 64, 65 e 77)	IV (16, 22, 25, 29, 38, 40, 49, 52, 56 e 72)	V (2, 5, 11, 12, 18, 39, 67, 71, 78 e 80)	VI (27 e 44)
Brosimum discolor	13,19	18,24	16,56	8,45	12,14	14,41
Cecropia palmata	7,38	14,66	16,86	20,47	11,09	5,47
Cedrela sp.	9,35	18,36	13,12	18,05	14,23	3,50
Chrysophyllum splendens	12,13	24,76	7,15	7,56	-	-
Copaifera langsdorffii	24,26	18,32	22,65	12,32	8,11	12,00
Cupania racemosa	7,14	12,11	7,87	19,18	12,04	2,50
Cupania revoluta	11,97	13,24	14,24	7,34	8,45	5,00
Dialium guianense	14,50	18.00	9,46	13,20	12,38	15,74
Eriotheca gracilipes	13,42	22,45	11,34	14,54	3,86	4,50
Erythroxylum squamatum	15,38	8,76	12,43	11,65	6,78	-
Eschweilera ovata	16,24	9,52	9,67	-	3,25	2,50
Helicostylis tomentosa	17,33	15,35	17,54	8,25	-	11,93
Inga thibaudiana	14,23	14,43	11,65	12,34	11,21	9,65
Licania rígida	18,78	16,21	18,65	19,31	7,43	-
Mabea occidentalis	17,47	13,76	14,98	12,43	12,17	6,00
Miconia albicans	15,48	17,54	8,23	-	11,67	5,60
Nectandra cuspidata	14,26	16,18	12,57	18,34	3,18	-
Ocotea gardneri	19,28	18,14	23,52	12,00	-	-
Ocotea opifera	8,57	21,23	16,23	14,23	-	-
Ouratea hexasperma	11,67	9,45	11,89	7,00	12,34	-
Parkia pendula	17,34	24,34	23,07	18,23	4,98	5,25
Plathymenia foliolosa	21,11	16,28	9,17	7,34	11,34	22,50
Pouteria grandiflora	12,26	12,28	17,43	9,00	9,44	7,48
Protium heptaphyllum	9,44	13,96	18,11	12,67	12,18	-
Pterocarpus violaceus	14,54	10,34	8,91	-	-	-
Schefflera morototoni	13,91	19,67	19,76	12,11	12,51	-
Stryphnodendron pulcherrimum	12,81	29,18	8,78	7,23	9,34	-
Tapirira guianensis	17,84	18,46	21,96	19,81	11,48	19,75
Thyrsodium spruceanum.	11,20	12,67	7,85	12,56	11,94	12,08
Média Geral (m)	14,22	16,48	*14,19*	12,91	9,73	9,21
Desvio padrão (m)	4,02	4,85	5,07	4,44	3,25	5,94
CV (%)	28,28	29,46	35,73	34,39	33,45	64,45

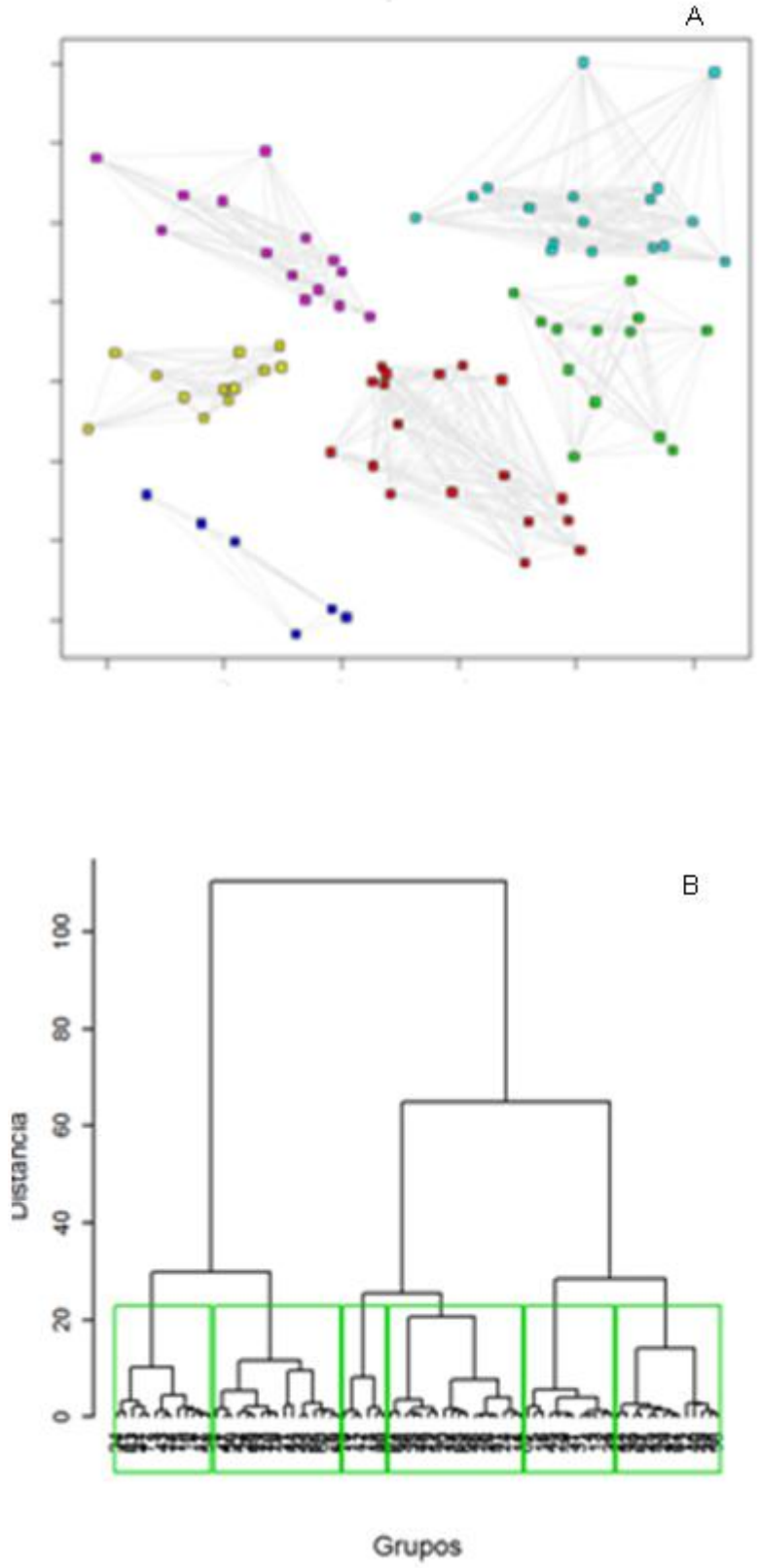

Figura 24. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de Ward, baseando-se no dado artificial II.

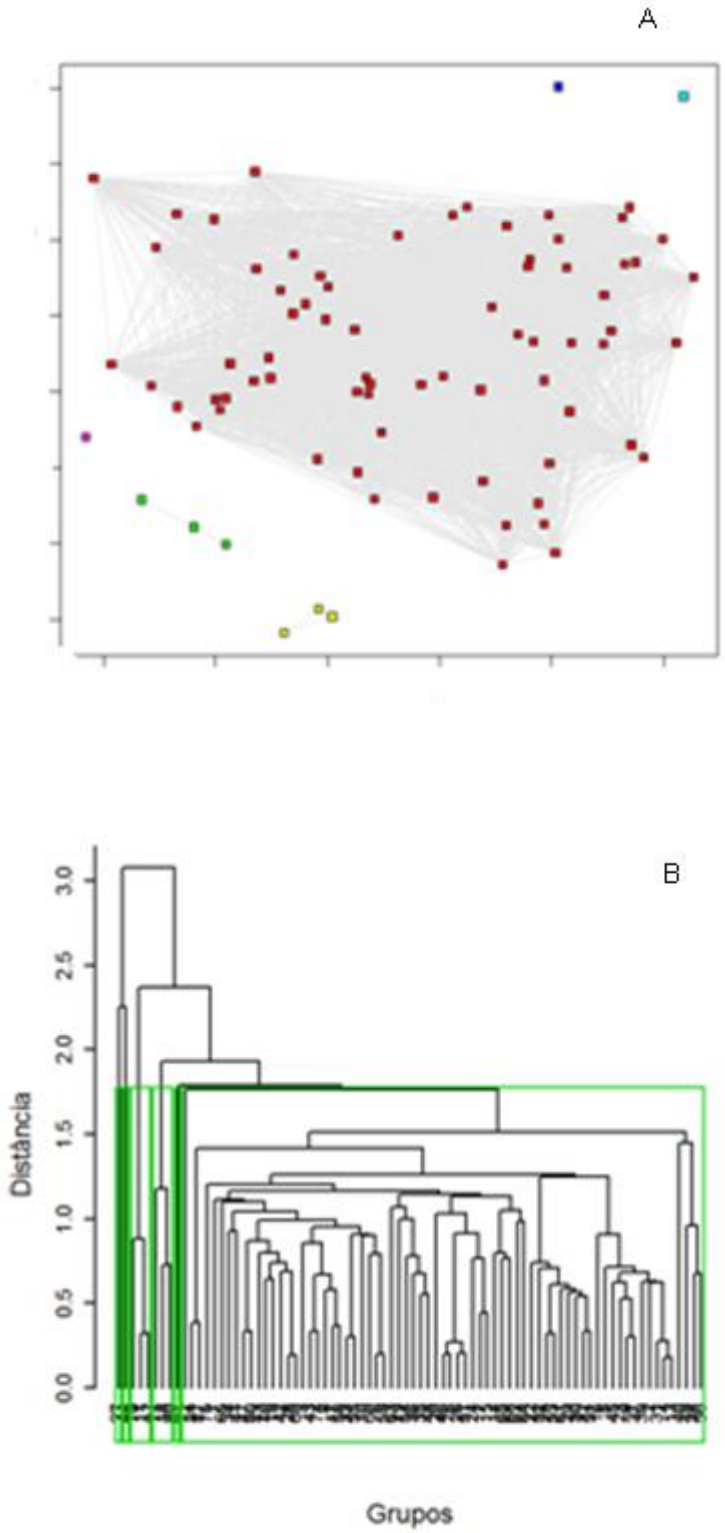

Figura 25. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de ligação simples, baseando-se no dado artificial II.

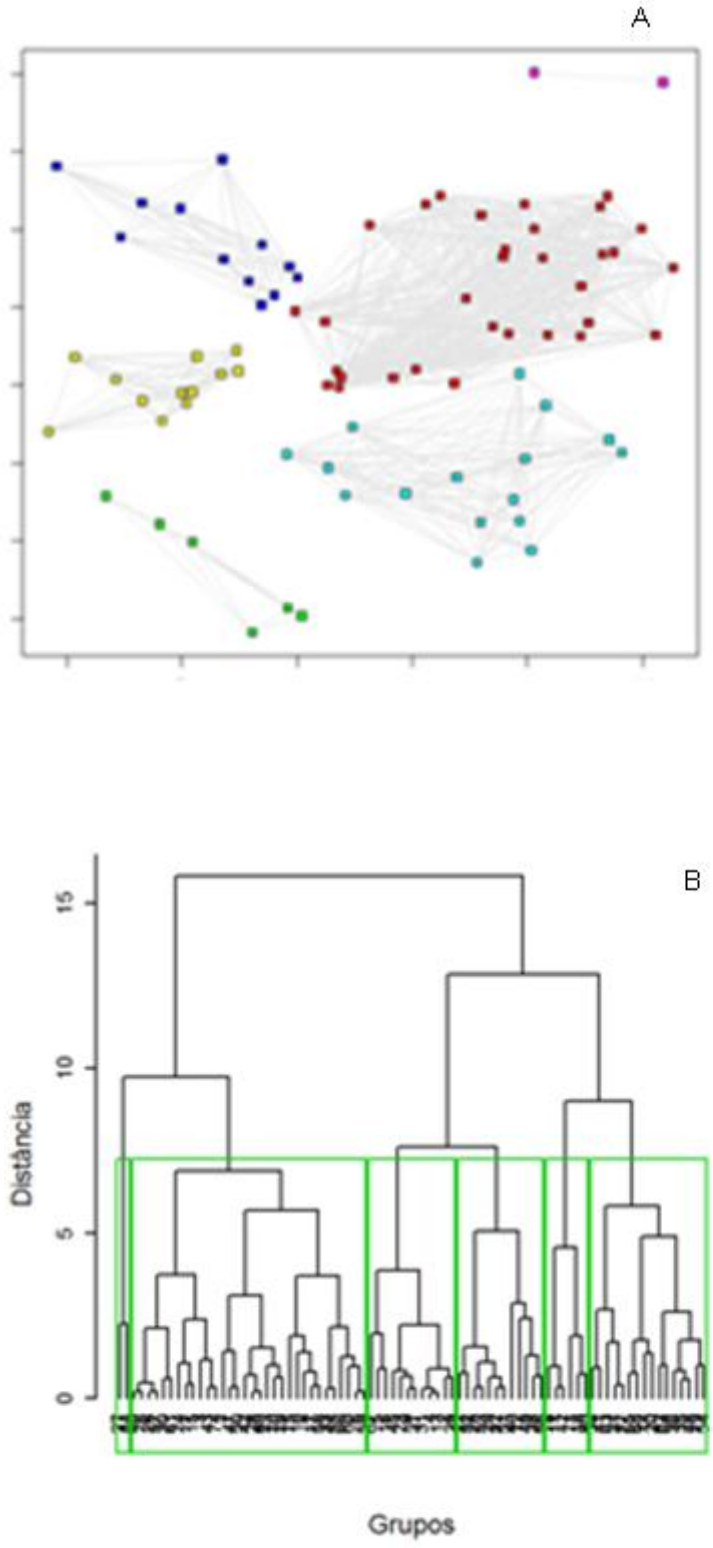

Figura 26. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de ligação completa, baseando-se no dado artificial II.

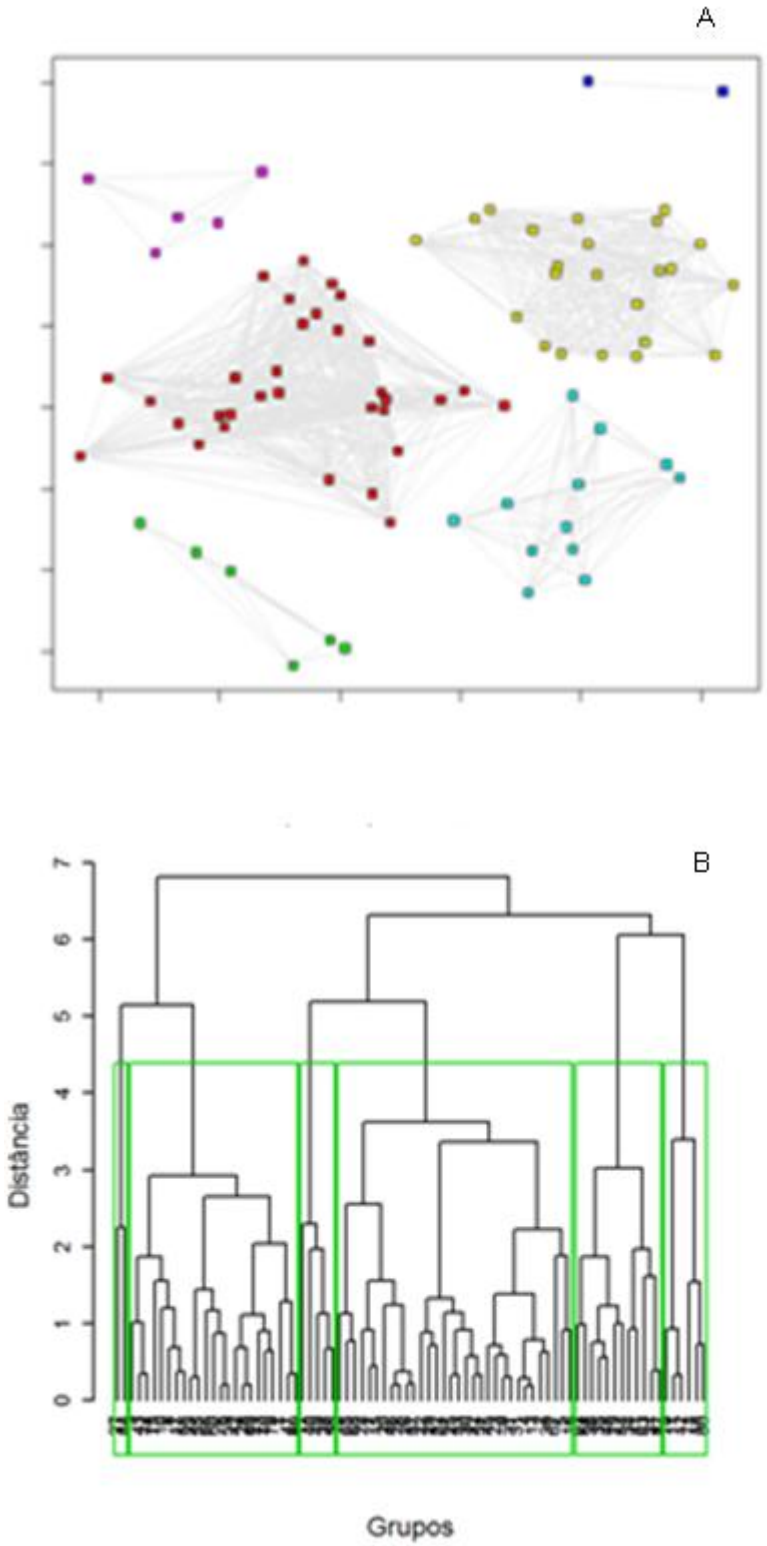

Figura 27. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de ligação média, baseando-se no dado artificial II.

Dados artificiais III

O conjunto de dados artificiais III, observado pelo método incremental, mostra a formação de sete grupos (Figura 28 e Tabela 14). É possível visualizar o resultado da formação dos grupos através de uma dispersão na Figura 28, onde a cor indica o grupo à qual pertence e na Tabela 14, e mostra os grupos com suas respectivas parcelas, que serão aplicados aos algoritmos Ward, de ligação simples, de ligação completa e de ligação média.

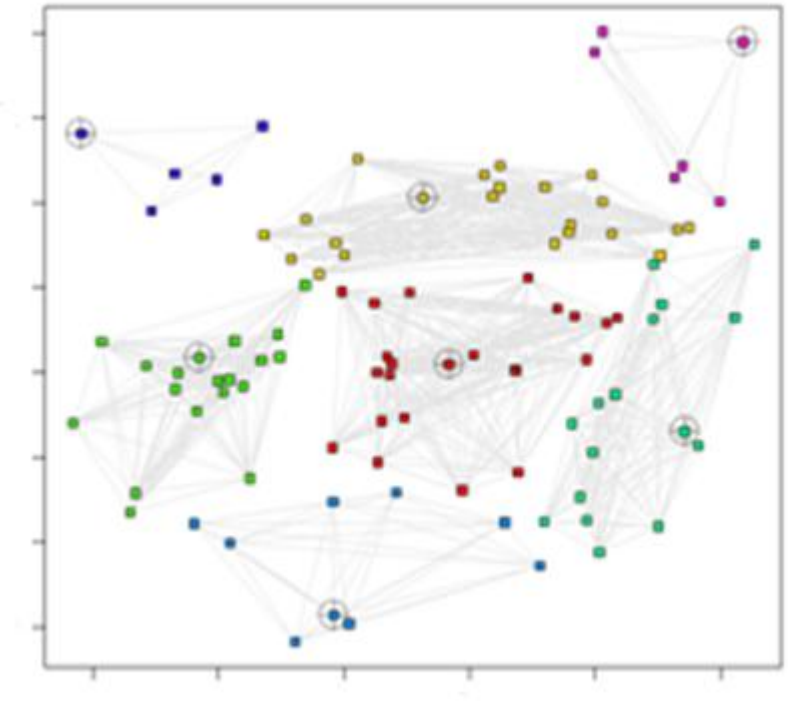

Figura 28. Dispersão obtida por meio do método incremental, baseando-se no dado artificial III.

Pelo método incremental baseando-se nos dados artificiais III foram formados sete grupos (Tabela 14). Observa-se que não foi apresentado nenhum *outliers.*

Tabela 14. Grupos de parcelas obtidos da distância euclidiana por meio do método incremental, baseando-se no dado artificial III.

Grupos	Parcelas
Grupo I	1, 3, 8, 14, 15, 20, 24, 26, 30, 41, 43, 46, 57, 61, 62, 64, 68, 74, 76, 83, 89, 94 e 95
Grupo II	6, 7, 22, 25, 28, 29, 42, 47, 50, 51, 52, 53, 59, 69, 70, 73, 79, 87, 88, 90, 91 e 96
Grupo III	2, 4, 5, 9, 13, 16, 23, 31, 32, 37, 39, 45, 67, 71, 86, 98, 92, 99 e 100
Grupo IV	10, 17, 21, 34, 35, 36, 48, 54, 58, 63, 66, 75, 77, 81, 82 e 84
Grupo V	11, 12, 18, 19, 65, 78, 80, 85 e 97
Grupo VI	38, 40, 49, 56 e 72
Grupo VII	27, 33, 44, 55, 60 e 93

Observaram abaixo, na Tabela 15, com a divisão de sete grupos (I, II, III, IV, V, VI e VII) e suas respectivas parcelas, com a média, com o desvio padrão e com o coeficiente de variação das espécies arbóreas. A quantidade de grupos, e os números de parcelas de cada grupo foram obtidos pelo método incremental.

Destaca-se, na Tabela 15, que os valores do coeficiente de variação são semelhantes para os grupos I, II, III, IV, V, VI e VII, e mostrou-se que têm uma boa homogeneidade.

Tabela 15. Distribuição das espécies arbóreas conforme Grupo e respectivas média, desvio padrão e coeficiente de variação para altura de Lorey, baseando-se no dado artificial III.

Espécies	*Grupo (Parcelas)*						
	I (1, 3, 8, 14, 15, 20, 24, 26, 30, 41, 43, 46, 57, 61, 62, 64, 68, 74, 76, 83, 89, 94 e 95	II (6, 7, 22, 25, 28, 29, 42, 47, 50, 51, 52, 53, 59, 69, 70, 73, 79, 87, 88, 90, 91 e 96	III (2, 4, 5, 9, 13, 16, 23, 31, 32, 37, 39, 45, 67, 71, 86, 98, 92, 99 e 100)	IV (10, 17, 21, 34, 35, 36, 48, 54, 58, 63, 66, 75, 77, 81, 82 e 84)	V (11, 12, 18, 19, 65, 78, 80, 85 e 97)	VI (38, 40, 49, 56 e 72)	VII (27,33, 44, 55, 60 e 93)
Brosimum discolor	17,43	18,26	14,45	12,40	11,14	7,23	18,01
Cecropia palmata	22,12	9,57	13,23	12,55	9,56	5,12	21,32
Cedrela sp.	15,23	15,34	10,47	10,34	11,34	9,76	9,34
Chrysophyllum splendens	14,23	8,65	11,53	23,44	-	8,11	14,11
Copaifera langsdorffii	12,34	19,60	17,50	27,45	22,45	3,50	12,43
Cupania racemosa	14,17	11,36	22,34	13,45	12,34	-	9,49
Cupania revoluta	23,45	9,65	12,33	7,98	11,35	8,34	-
Dialium guianense	18,13	-	17,91	17,67	13,25	8,11	17,00
Eriotheca gracilipes	7,12	15,23	9,33	22,76	9,98	8,56	-
Erythroxylum squamatum	11,32	14,23	8,50	8,00	16,71	4,98	6,65
Eschweilera ovata	8,34	12,34	11,22	13,21	11,38	7,77	12,15
Helicostylis tomentosa	11,23	15,23	21,54	14,60	13,12	2,45	8,01
Inga thibaudiana	18,13	17,22	11,98	15,00	18,81	1,23	-
Licania rígida	13,00	9,23	25,44	16,93	19,34	8,45	7,00
Mabea occidentalis	23,00	6,45	19,36	-	12,98	8,12	-
Miconia albicans	5,00	-	11,41	9,78	15,12	-	-
Nectandra cuspidata	12,00	25,12	25,12	17,20	9,67	10,98	10,18
Ocotea gardneri	12,70	14,11	14,11	-	20,11	-	4,23
Ocotea opifera	19,70	12,43	12,43	19,78	16,34	7,36	19,98
Ouratea hexasperma	16,21	9,49	7,98	24,45	21,09	8,34	17,01
Parkia pendula	11,13	7,45	19,78	11,28	13,90	2,80	20,32
Plathymenia foliolosa	7,32	-	24,45	15,46	18,51	2,56	13,24
Pouteria grandiflora	21,34	17,23	11,28	10,77	10,34	4,32	12,54
Protium heptaphyllum	7,35	12,70	15,57	9,47	8,45	-	14,11
Pterocarpus violaceus	12,43	21,16	12,00	11,32	-	21,34	3,12
Schefflera morototoni	9,45	18,50	16,38	22,47	19,43	4,18	22,17
Stryphnodendron pulcherrimum	8,45	-	12,70	18,96	11,54	1,13	19,01
Tapirira guianensis	15,32	26,98	21,16	25,12	17,33	4,01	8,67
Thyrsodium spruceanum.	14,77	15,24	18,57	15,81	10,18	2,11	14,14
Média Geral (m)	*13,87*	14,51	*15,52*	15,84	14,29	2,83	13,93
Desvio padrão (m)	*5,04*	5,23	5,13	5,59	4,10	4,04	4,61
CV (%)	*36,35*	36,08	33,07	35,28	28,70	31,46	33,09

Observou-se, na (Figura 30), que mostra disprsão (A) e o dendrograma (B) típico de ligação simples. Há uma estrutura clara, ligando os grupos, que esta gradualmente se juntarem em um grande grupo, isolando seis grupos até a etapa final.

Figura 30 (B) observa-se o encadeamento dos grupos principais juntos numa ligação simples, e o isolamento de *outliers,* se dois grupos forem especificados (Se três grupos são especificados, o outro *outliers* é acantonada (recuperado), deixando ainda um grande agrupamento.)

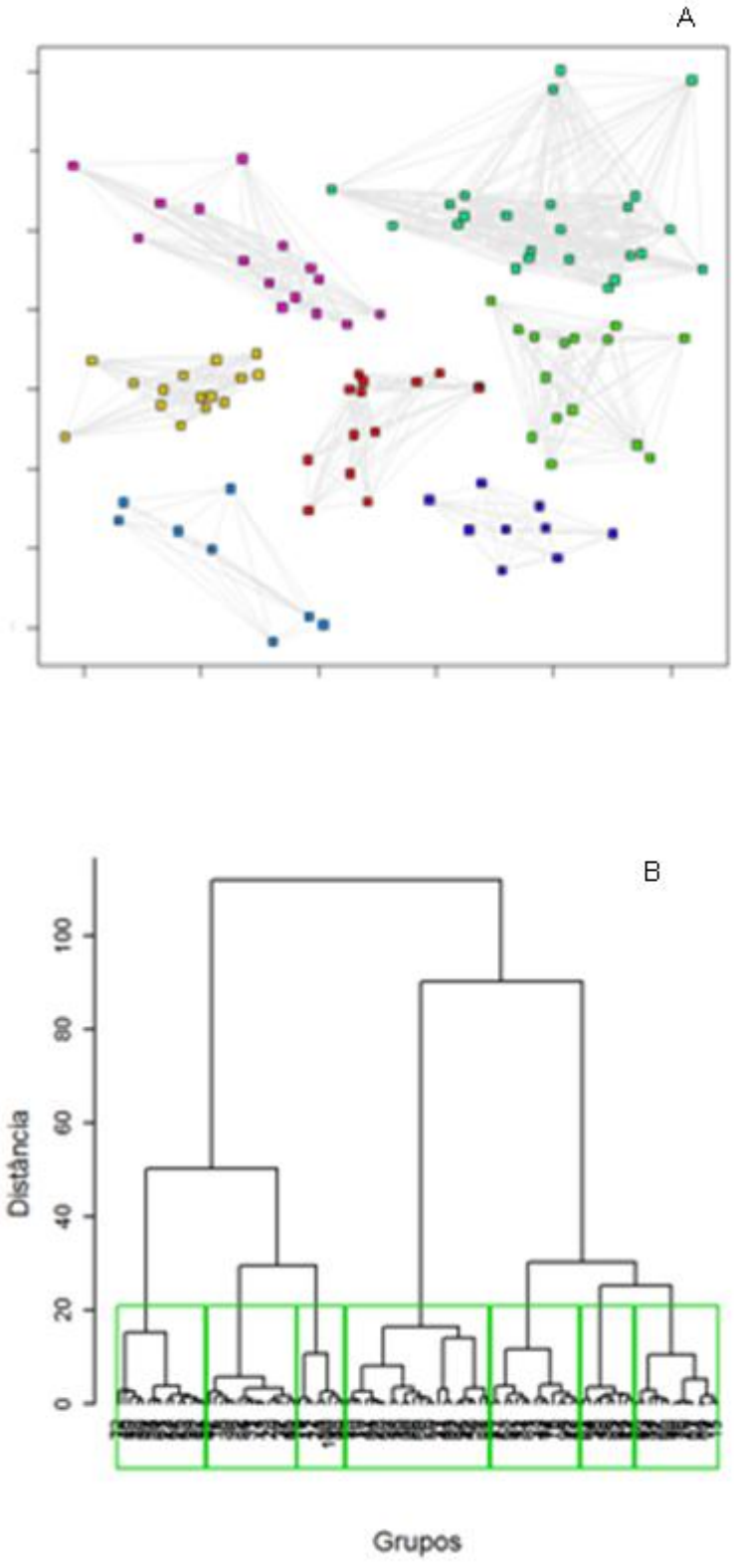

Figura 29. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de Ward, baseando-se no dado artificial III.

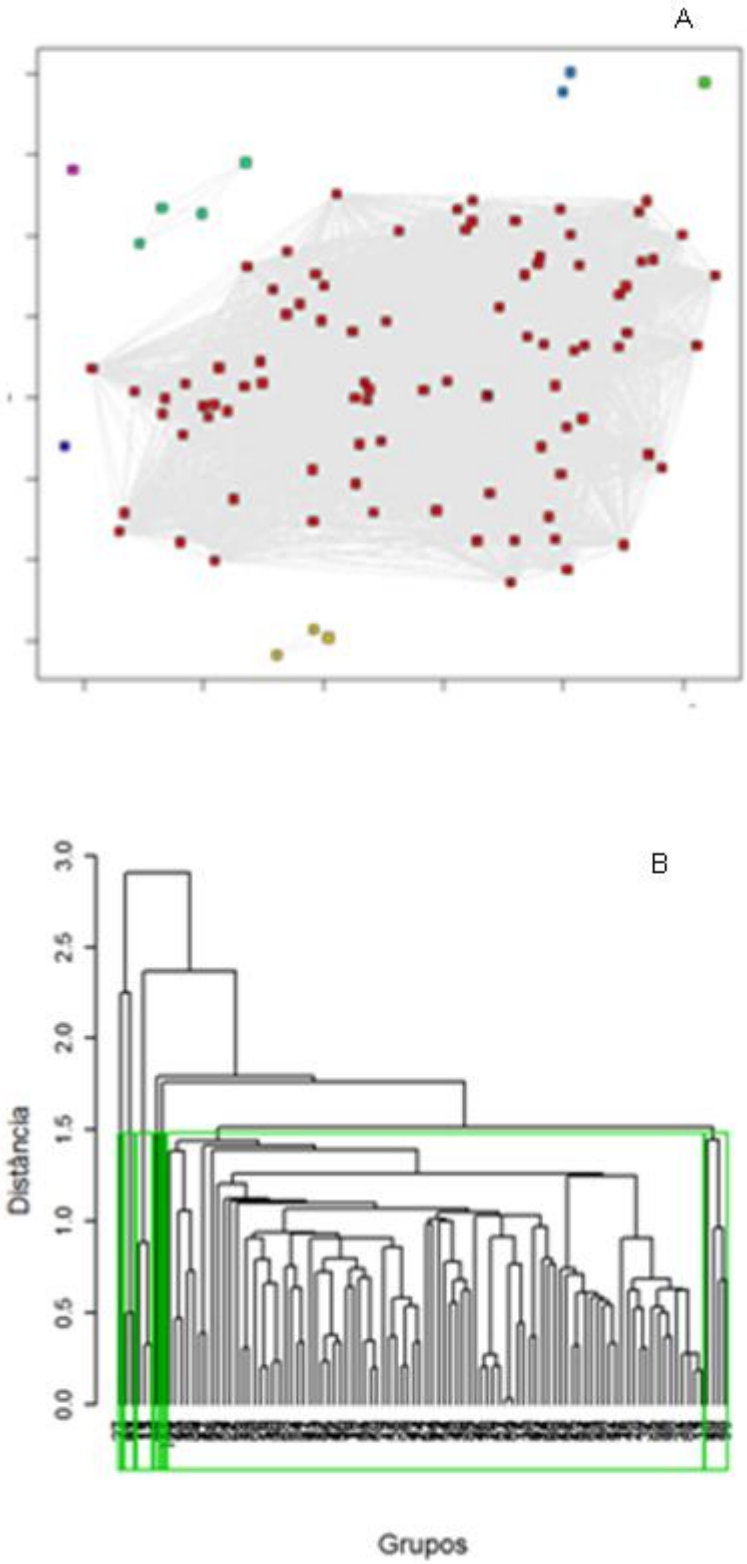

Figura 30. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de ligação simples, baseando-se no dado artificial III.

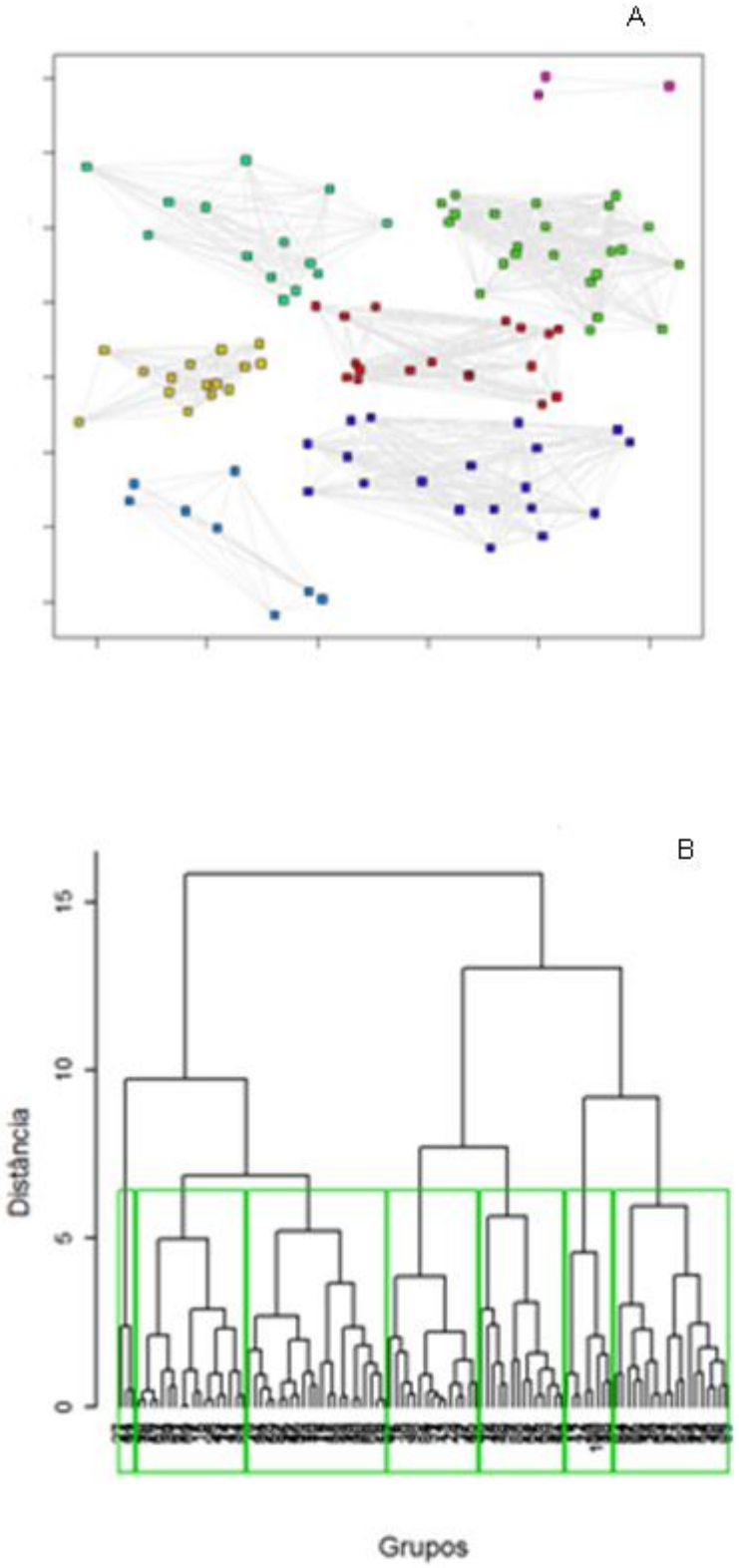

Figura 31 – Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de ligação completa, baseando-se no dado artificial III.

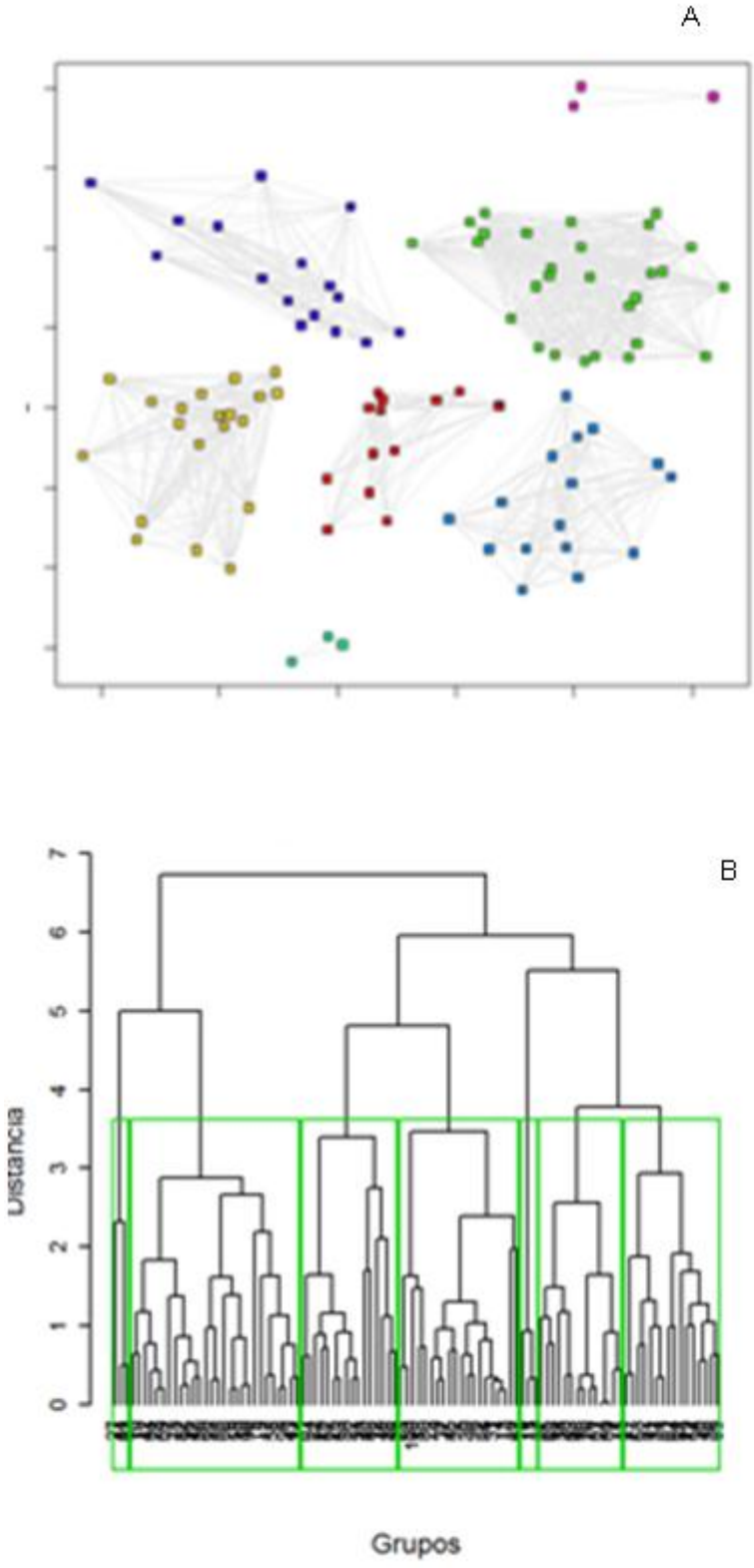

Figura 32. Dispersão (A) e dendrograma (B) obtido por meio do algoritmo de ligação média, baseando-se no dado artificial III.

Observou-se, na Tabela 16, os resultados obtidos pela aplicação dos métodos hierárquicos com os dados artificiais. Com os agrupamentos obtidos pelos algoritmos Ward, de ligação simples, de ligação completa, de ligação média e do método incremental foram calculados os índices de validação. Após uma análise em separado dos resultados de cada método, foi feita uma análise comparativa dos métodos, enfocando os cenários utilizados e o número de grupos.

Não existe um algoritmo específico que seja apropriado a todos os tipos de dados, e a adoção de um ou outro depende de uma análise aprofundada, como a aqui apresentada.

Os agrupamentos gerados pelo método incremental apresentaram melhor desempenho na validação estatística em relação aos algoritmos Ward, de ligação simples, de ligação completa e de ligação média. Tomando os valores obtidos pelos índices, nos algoritmos Ward, de ligação simples, de ligação completa e de ligação média respectivamente, tem-se os valores para o índice de Rand ajustado de magnitude elevada, e correlação cofenética, com uma baixa magnitude e as técnicas de variância multivariada mostrou que existe homogeneidade interna dos grupos, ou seja, a variação dentro dos grupos e heterogeneidade entre os grupos dos agrupamentos confirmando os resultados dos dados originais.

Como o interesse do nosso trabalho está em montar grupos, cuja variabilidade interna de cada grupo seja a menor possível, o algoritmo de Ward, excluindo os grupos unitários, forneceu a solução mais interessante ao problema, porque apresentou grupos com números de parcelas mais similares (homogênea).

Sempre que o pesquisador tiver interesse em criar grupos de tamanhos distintos e homogêneo, sugere-se que o mesmo inicie o processo de agrupamento com um número maior de parcelas ou de grupos e retire as parcelas de cada grupo ou o próprio grupo, somente após a aplicação da técnica (KEESE, 2012)

A motivação, para realizar-se a combinação de análise de agrupamento hierárquico e não hierárquico com o método incremental é combinar as vantagens das duas técnicas e criar um relacionamento entre os grupos hierárquico e não hierárquico de maneira incremental.

A análise de agrupamento hierárquico de dados organiza os grupos numa árvore hierárquica, facilitando-se a navegação entre os dados. Entendendo-se que um grupo pode representar uma parcela, é possível não só conhecer as parcelas tratadas pelos grupos, como também o relacionamento entre elas e estimar a hierarquia entre elas (MINGOTE, 2007).

A técnica hierárquica possui outras características e vantagens como: conhecer o relacionamento entre os dados, descobrir os grupos abordados pelas parcelas existentes e suas ligações, observa-se o número de grupos ideal para a técnica não hierárquica do conjunto, entre outros (SILVA, 2005).

A técnica hierárquica, quando aplicada para descobrir a relação de grupos e subgrupos, dentro de uma hierarquia do agrupamento, possui alguns problemas relacionados à alta dimensão, grande volume de dados, facilidade de navegação e parcelas significativos para os grupos (CAN; DROCHAK, 1990), devido à capacidade computacional necessária.

O resultado do agrupamento obtido com o método da técnica não hierárquica é melhor e mais eficaz do que a técnica hierárquica. Porém a classe de algoritmo que combina a técnica não hierárquica com a técnica hierárquica, obtém resultados ainda melhores do que a técnica de não hierárquica, além de permitir a redução dos erros no estágio inicial gerada pelas técnicas (JING et al., 2007). Baseando-se na premissa de que também construímos um método que combine uma técnica de não hierárquica com a técnica hierárquica, não é nova a ideia de combinar uma parcela do método incremental e um hierárquico.

Comparando-se a técnica hierárquica e o método incremental de agrupamento aplicado a ciência florestal, o algoritmo de Ward e de ligação completa foi o que obteve melhores resultados, tanto em relação ao número de parcelas agrupadas, quanto ao número de grupos formados, e quanto a parcela hierárquica obtida.

Comparando-se os resultados obtidos com o método incremental e os obtidos com os algoritmos hierárquica, utilizados nesse trabalho, nota-se que há diferenças nos grupos formados e na quantidade de dados agrupados.

Essas diferenças ocorrem devido à abordagem do algoritmo escolhido, pois a principal característica do método incremental é ele ser apto a escolher quando

um novo grupo deve ser criado durante sua execução, enquanto nos algoritmos escolhidos e comparados, o número de grupos é definido.

O método incremental proposto mostrou-se eficiente. Os grupos formados são coesos dentro dos intervalos indicados por meio do cálculo.

Utilizando-se do método incremental e hierárquico, é possível visualizar o comportamento dos grupos, descobrir o quão similar as parcelas são (CAN, 1993). Dessa forma, permite a um pesquisador conhecer sua base, saber qual a melhor maneira de organizá-la, e qual o intervalo de corte para o agrupamento.

9.6 Validação e interpretação dos agrupamentos

Baseando-se nas correlações cofenéticas dos métodos de agrupamentos hierárquicos apresentados (Tabela 16), pode-se avaliar que o método das médias das distâncias proporcionou um melhor agrupamento, apresentando-se a maior correlação cofenética e o método do Ward apresentou-se o pior resultado. Este resultado concorda com Barroso e Artes (2003) e Assis, et al. (2011) que afirmam, que o método das médias das distâncias produz melhores partições que os métodos de ligação simples e de ligação completa. Observou-se que o método de Ward proporcionou-se o pior agrupamento, apresentando-se a menor correlação cofenética. Comparando se os métodos hierárquicos e o método incremental (Tabela 16), verifica-se que o método incremental apresentou a maior correlação, mostrando que existe uma ótima correlação entre a soma das distâncias de uma parcela, em relação as demais, e as médias de seus respectivos grupos.

Não se deve utiliza-se da correlação cofenética para validar os agrupamentos, ou seja, número de grupos, e sim, para fazer uma comparação entre os métodos, pois, para qualquer número de grupos, a correlação cofenética tem a mesma mensuração (valor). Já se mudar o método, muda-se o valor da correlação cofenética, pois para cada método ele tem um valor diferente. Sokal e Rohlf (1962) definiram o coeficiente de correlação “cofenético”, como medida de validação do melhor método.

Meyer (2004) e Cargnelutti et al. (2008) referem-se à avaliação dos agrupamentos, o coeficiente de correlação cofenética, que relaciona a matriz de distâncias originais oriundas da classificação (matriz cofenética); algo em torno de 0,8 já seria bom.

Os valores das correlações cofenéticas, na Tabela 16, mostraram os algoritmos de ligação simples e de ligação média de magnitude elevada, para os respectivos algoritmos e o de ligação completa, e os de Ward mostraram-se de baixa magnitude e para os dados artificiais a correlação cofenética, verificou-se com uma baixa magnitude. Isso mostra que há uma boa representação das matrizes de dissimilaridade na forma de dendrogramas, para os algoritmos de ligação simples e de ligação média dos dados originais, e isso mostra que, através das correlações cofenéticas, devem-se ser escolhidos os algoritmos que sejam superiores ao valor de 0,8 (BUSSAB et al., 1990).

Comparando-se cada grupo como uma amostra de uma população, aplicou-se um teste F de comparação de média para cada parcela. Todos os níveis dos centros foram inferiores a 0,05, indica-se haver diferenças entre as médias dos grupos.

A técnica da análise de variância multivariada foi aplicada, no intuito de verificar a homogeneidade interna dos grupos, ou seja, a variação dentro dos grupos e heterogeneidade entre os grupos dos agrupamentos.

Por meio da análise de variância multivariada, aplicada para cada algoritmo, foram evidenciadas diferenças significativas entre os vetores de médias. Conforme observou-se, na Tabela 16, dados originais o F obteve 9,21, com a maior magnitude, algoritmo de Ward, e de menor magnitude 8,09 o algoritmo de ligação média, R^2 obteve o valor de 0,67, com maior magnitude, com o algoritmo de Ward. Esses valores são de boa magnitude, mas, mostrando-se a existência de heterogeneidade significativa entre grupos, e que os grupos são homogêneos internamente, e o valor do Pseudo-F com 1,99, mostrando-se existência de uma eficiência. Com os valores de Wilks mostrando-se uma elevada variabilidade, fica claro que são heterogêneos entre os grupos, e homogêneos intragrupos.

Observando-se o método incremental (Tabela 16), o F e o R^2, mostrando a existência de heterogeneidade significativa entre grupos, e que os grupos são

homogêneos internamente; e o valor do Pseudo-F mostrou a existência de uma boa eficiência. Com o valor de Wilks bem próximo de zero, mostra que os grupos são heterogêneos entre os grupos e homogêneos intragrupos. Comprando-se os métodos hierárquicos e método incremental (Tabela 16), observa-se que todos os valores são significativamente superiores ao método hierárquico. Como se pode ver pela (Tabela 16), os valores dos índices Rand ajustado mostrou-se com uma magnitude elevada para todos os métodos.

Tabela 16 Valores das estatísticas para testar H_0 (médias iguais para grupos) para os algoritmos de agrupamento, correlações cofenéticas e Rand ajustado entre as matrizes de dissimilaridade obtidas conforme algoritmos de agrupamento e o método incremental para dados originais e artificiais I, II e III obtidas conforme algoritmos de agrupamento e o método incremental.

Dados originais	Ward	L. simples	L. completa	Média	Incremental
$F_{calculado}$	9,21	8,32	8,50	8,09	285,71
R^2	0,67	0,65	0,65	0,64	0,98
Pseudo-F	1,99	1,82	1,90	1,77	62,5
Wilks	0,33	0,35	0,35	0,36	0,02
Correlação cofenética	0,49	0,84	0,58	0,89	0,99
Rand ajustado	0,90	0,87	0,91	0,87	0,97
Dados artificiais I Nº de grupos = 6	Ward	L. simples	L. completa	Média	Incremental
$F_{calculado}$	9,21	18,81	7,83	8,21	85,71
R^2	0,67	0,64	0,65	0,64	0,98
Pseudo-F	0,77	1,74	0,72	0,76	0,95
Wilks	0,33	0,36	0,35	0,36	0,02
Correlação cofenética	0,63	0,66	0,64	0,77	0,99
Rand ajustado	0,94	0,91	0,93	0,93	0,96
Dados artificiais II Nº de grupos = 6	Ward	L. simples	L. completa	Média	Incremental
$F_{calculado}$	9,21	10,65	7,83	13,04	87,02
R^2	0,98	0,46	0,62	0,53	0,99
Pseudo-F	1,77	1,74	0,72	0,76	8,50
Wilks	0,02	0,54	0,38	0,47	0,01
Correlação cofenética	0,59	0,66	0,59	0,77	0,99
Rand ajustado	0,95	0,93	0,95	0,95	0,96
Dados artificiais III Nº de grupos = 7	Ward	L. simples	L. completa	Média	Incremental
$F_{calculado}$	13,21	11,65	11,83	12,04	82,86
R^2	0,92	0,57	0,79	0,86	0,99
Pseudo-F	0,82	0,75	0,72	0,76	0,98
Wilks	0,08	0,43	0,21	0,14	0,01
Correlação cofenética	0,57	0,58	0,58	0,64	0,99
Rand ajustado	0,95	0,94	0,95	0,95	0,95

Em geral, quanto maior o pseudo-F, mais “eficiente” é a partição na redução da heterogeneidade, no interior do grupo (refletida em W). Embora B geralmente vá aumentando, o número de B não aumenta. Em geral, a medida do pseudo-F não aumenta monotonicamente, mas atinge um máximo para determinado valor especificado de K (dependendo, naturalmente, dos dados) (MELO; HEPP, 2008).

Usando-se os algoritmos de agrupamento de Ward, de ligação simples, de ligação completa e de ligação médias, obtendo-se os resultados observados na Tabela 16. O R^2, e os valores da Pseudo-F mostram que existem uma boa eficiência. Os algoritmos de Ward apresentou o melhor resultado referentes ao $F_{calculado}$, R^2, Pseudo-F e os valores da estatística lambda Wilks apresentou valores de magnitude semelhantes e mostrando que existe homogeneidade interna e heterogeneidade entre grupos.

Comparando-se os agrupamentos através do cálculo do valor de R^2 e da variabilidade residual média, como mostrado na Tabela 16. A variabilidade residual média é simplesmente a média dos valores das somas de quadrados dentro dos grupos, onde a média é calculada em relação ao número de grupos .

Num enfoque diferente, a análise de variância multivariada foi aplicada, de forma inédita, na área de ciência florestal, posteriormente a técnicas exploratórias multivariadas, no sentido confirmatório da divergência dos grupos selecionados, apoiando-se na estatística inferencial. Melo e Hepp (2008) e Ferreira (2008) aplicaram essa metodologia, com o objetivo de detectar-se a diferença entre os vetores de médias de todas as variedades e, assim, justificar a necessidade da busca da divergência ou da redução dimensional e dos descartes de variáveis. Ambos os objetivos são válidos como sugestão para um programa de melhoramento de árvores.

O determinante de W é uma medida da variabilidade dentro dos grupos, enquanto que o determinante de T nos dá uma medida da variabilidade total. Quanto maiores forem as semelhanças entre os dois determinantes, menores serão as diferenças entre grupos (B), mais o valor de Wilks se aproximarão de 1. Se as diferenças entre os grupos forem elevadas (heterogêneo), quando comparadas com a variabilidade dentro dos grupos, o valor de Wilks tenderá a aproximar-se de zero. Assim, a estatística Λ de Wilks é uma medida inversa do

grau de diferenciação entre os grupos: quanto menor o seu valor, maior esse grau de diferenciação (RODE et al., 2011; SANTOS; LONGHI, 2012).

Os resultados analisados que foram obtidos no contexto dos dados originais e dos artificiais I, II e III que, mostraram-se com uma magnitude elevada na qualidade dos agrupamentos gerados confirmando os resultados dos dados originais, demonstrando assim uma concordância entre as partições obtidas nos respectivos dados originais e artificiais.

10 CONCLUSÕES

Neste trabalho foi proposta a utilização do método Incremental que pode utiliza-se para a análise de agrupamento. Realizou-se uma aplicação com um conjunto de dados da área de Ciência Florestal e com dados artificiais, cujos resultados mostraram-se bastante satisfatórios. Constatou-se que o método exposto apresenta uma série de benefícios, dentre os quais podem ser destacados:

I. O método converge rapidamente, pois o número máximo de interações equivale, no pior caso, ao número de observações;

II. a determinação do número de grupos independe do pesquisador; portanto, o método não sofre influência de subjetividade;

III. o método fornece sempre as mesmas soluções, sendo, portanto, determinístico. Assim, para uma mesma massa de dados, a técnica só necessita ser executado uma vez;

IV. o método fornece apenas um resultado de agrupamento.

O método apresenta algumas limitações intrínsecas ou que decorreram de simplificações no corpo da pesquisa, dentre as quais pode se ressaltar que, por trabalhar com a distância euclidiana como critério para determinação dos intervalos de influência dos pontos sementes, a qualidade das soluções geradas é influenciada por *outliers*.

Como sugestões para o aprofundamento do tema pesquisado e aperfeiçoamento do algoritmo proposto, recomenda-se:

1. o método deve ser testada em outros conjuntos de dados, de modo a avaliar o desempenho da técnica em várias situações; tal análise permitirá uma definição precisa acerca das vantagens e desvantagens do método;

2. também devem ser empregadas outras medidas de (dis)similaridade, para avaliar o comportamento da técnica.

REFERÊNCIAS BIBLIOGRÁFICAS

ALBUQUERQUE, M. A. et al. Estabilidade em análise de agrupamento: estudo de caso em ciência florestal**. Revista Árvore**, Viçosa-MG, v. 30, n.2, p. 257-265, 2006.

ALDENDERFER, M. S.; BLASHFIELD, R. K. **Cluster Analysis.** Beverly Hills, CA: Sage Publications, 1984.

ANDERBERG, M. R. **Cluster analysis for applications.** London: Academic Press, 1973. 359 p.

ANDERSON, T. W. **Na introduction to multivariate statistical analysis.** New York; John Wiley & Sons, 1984. 675 p.

ARAUJO, et al. Análise de agrupamento em remanescente de Floresta Ombrófila Mista. **Ciência Florestal,** Santa Maria, v. 20, p. 1-18, 2010.

BAFFETTA, F; CORONA, P.; FATTORINI, L. Assessing the attributes of scattered trees outside the forest by a multi-phasesampling strategy. **Forestry An International Journal of Forest Research**, v. 84, n. 3, p. 315 – 325, 2011.

BAFFETTA, F.; FATTORINI, L.; CORONA, P. Estimation of small woodlot and tree row attributes in large-scale forest inventories . **Environ. Ecol. Stat.** v.18, p. 147 – 167, 2011.

BARROSO, L. P.; ARTES, R. **Análise Multivariada**. In: REUNIÃO ANUAL DA RBES E SEAGRO, 48., 100, Lavras. Curso. Lavras: Departamento de Ciências Exatas, 155p. 2003. 155 p.

BARALOTO, C. et al. Integrating functional diversity into tropical forest plantation designs to study ecosystem processes. **Annals of Forest Science**, Bethesda, v. 67, p. 303 - 313, 2010.

BATISTA, F. J. et al. M. Comparação florística e estrutural de duas florestas de várzea no estuário amazônico, Pará, Brasil. **Revista Árvore,** Viçosa-MG, v. 35, n. 2, p. 289 - 298, 2011.

BENIN, G. et al. Comparações entre medidas de dissimilaridade e estatísticas multivariadas como critérios no direcionamento de hibridações em aveia. **Ciência Rural,** Santa Maria, v.33, n.4, p.657-662, 2003.

BENITES, V. M et al. Análise discriminante de solos sob diferentes usos em área de mata atlântica a partir de atributos da matéria orgânica. **Revista Árvore,** Viçosa-MG, v. 34, n.4, p. 685 - 690, 2010.

BERTINI, C. H. M. et al. Análise multivariada e índice de seleção na identificação de genótipos superiores de feijão-caupi. **Acta Scientiarum Agronomy**, v. 32, n. 4, p. 613 - 619, 2010.

BERGMAN, E. M.; FESER, E. J. **Industrial and Regional Clusters:** Concepts and Comparative Applications. Virginia: West Virginia University – Regional Research Institute, 1998. [on-line] [citado em 17/08/2012] Disponível na World Wide Web: http://www.rri.wvu.edu/WebBook/Bergman-Feser/contents.htm

BEZERRA NETO, F. V. B. et al. Descritores quantitativos na estimativa da divergência genética entre genótipos de mamoneira utilizando análises multivariadas. **Revista Ciência Agronômica**, Fortaleza-CE, v. 41, n. 02, p. 294 - 299, abr-jun, 2010.

BICKEL, P.; FREEDMAN, D. Some asymptotic theory the bootstrap. **Annals of Statistics**. v. 1, n. 9, p.1196 - 1197, 1981.

BOSCARIOLI, C. **Análise de Agrupamentos baseada na topologia dos Dados e em Mapas Auto-organizáveis.** 2008. 118p. Escola Politécnica da Universidade de São Paulo. Tese de Doutorado, São Paulo, 2008.

BOSCARIOLI, C; SILVA, L. A.; HERNANDEZ, E. D. M. Análise de Agrupamentos UTILIZANDOS Mapas Auto-organizáveis EM AGRICULTURA DE PRECISÃO**. ANIS do Congresso Brasileiro de Agricultura de Precisão,** São Pedro, SP. Combap, 2006.

BOUROCHE, J. M. SAPORTA, G. **Análise de dados**, Rio de Janeiro: Zahar, 1972, 116p

BRAY, J. R.; CURTIS, J. T. An ordination of the upland forest communities of Southern Wisconsin. **Ecological Monographies**. v. 27, p. 325 - 349, 1957.

BRUN, M. et al. Model-based evaluation of clustering validation measures. Pattern Recognition. **Elsevier Science**, Oxford, v. 40, n. 2, 807 - 824, 2007.

BUSSAB, W. DE O; MIAZAKI, E. S; ANDRADE, D. **Introdução à análise de agrupamentos**. 9º Simpósio Nacional de Probabilidade e Estatística (Sinape). São Paulo. Associação Brasileira de Estatística, 1990. 105 p.

CALINSKI, T.; HARABASZ, J. A dendrite method for cluster analysis. **Communications in Statistics,** v. 3, p. 01 - 27, 1974.

CAMPOS, K. C.; CARVALHO, H. R. Análise estatística multivariada: uma aplicação na atividade agrícola irrigada do município de Guaiúba Ce. **Revista de Economia da UEG**. Anápolis (GO), vol. 3, nº 1, jan/jun-2007.

CAN, F. Incremental Clustering for Dynamic Information Processing. **ACM Transactions on Information Systems,** v. 11, n. 2, p. 143 – 164, april 1993.

CAN, F.; DROCHAK II N.D. Incremental Clustering for Dynamic Document Databases. **In Proceedings of the 1990 Symposium on Applied Computing**, p. 61-67, 1990.

CARLINI-GARCIA, L. A.; VENCOVSKY, R.; COELHO, A.S.G. Método bootstrap aplicados em níveis de reamostragem na estimação de parâmetros genéticos populacionais. **Scientia Agricola**, v.58, n.4, p.785-793, out./dez. 2001.

CARGNELUTTI, F. et al. Comparação de métodos de agrupamento para o estudo da divergência genética em cultivares de feijão. **Ciência Rural,** Santa Maria,. vol.38, n.8, p. 2138 – 2145, 2008.

CARRASQUINHO, et al. Selection of *Pinus pinea* L. plus tree candidates for cone production. **Ann. For. Sci**. v. 67, p. 814 - 821, 2010.

CASTRO, B. M. et al. Raio de Influência: um método de agrupamento alternativo para Análise de Cluster**.** 19º SINAPE. São Paulo. **Anais...**.ABE, 2010.

CHARIKAR, M. et al. **Algorithms for facility location problems with outliers**, in Proceedings of the 12th Annual ACM-SIAM Symposium on Discrete Algorithms, pp. 642‑651. 2001.

CLARKE, K. R.; SOMERFIELD, P. J.; CHAPMAN, M. G. On resemblance measures for ecological studies, including taxonomic dissimilarities and a zero-adjusted Bray-Curtis coefficient for denuded assemblages. **Journal of Experimental Marine Biology and Ecology.** v. 330, p. 55 - 80, 2006.

COIMBRA, R. R. et al. Caracterização e divergência genética de populações de milho resgatadas do Sudeste de Minas Gerais. **Revista Ciência Agronômica**, Fortaleza, CE, v. 41, n. 01, p. 159 – 166, 2010.

CORMACK, R. A review of classification. Journal of the Royal **Statistical Society** (Series A), v. 134, p. 321 - 367, 1971.

CORRAR, L. J.; PAULO, E.; DIAS, F. J. M. **Análise multivariada.** São Paulo: Atlas, 2007. 540.p.

COSTA JUNIOR, R. F. **Caracterização estrutural de um remanescente de Mata Atlântica do município de Catende-PE**. 2006. 52p. Dissertação (Mestrado em Ciências Florestais) – Universidade Federal Rural de Pernambuco, 2006.

CRUZ, C. D. A**plicação de algumas técnicas multivariadas no melhoramento de plantas.** 1990, 188p. Tese (Doutorado em Genética e Melhoramento de Plantas). Escola Superior de Agricultura “Luiz de Queiroz”. Universidade de São Paulo, Piracicaba, 1990.

CRUZ, T. L. **Utilizando uma Nova Abordagem para Análise de Cluster com Vistas a uma Gestão de Riscos mais Efetiva em Projetos de Software**. 2009. 101p. Tese (Doutorado em Informática) Universidade Federal do Rio de Janeiro. Programa de Pós-Graduação em Informática, Rio de Janeiro, 2009.

CRUZ, C. D.; REGAZZI, A. J. Divergência genética. In: CRUZ, C. D.; REGAZZI, A. J. **Métodos biométricos aplicados ao melhoramento genético**. Viçosa, UFV: Impressa Universitária. cap. 6, p. 287- 323, 1994.

DALIRSEFAT, S. B. ; MEYER, A. S. ; MIRHOSEINI, S. . Comparison of similarity coefficients used for cluster analysis with amplified fragment length polymorphism markers in the silkworm, Bombyx mori. **Journal of Insect Science** (Online), v. 09, p. 1- 8, 2009.

DOBBERTIN, K. M.; NOBIS, M. P. Exploring research issues in selected forest journals. **Annals of Forest Science**, Bethesda, vol. 66, n. 8 , p. 800 – 807, 2010.

DUARTE, F. J. F. **Otimização da Combinação de Agrupamentos baseada da Acumulação de Provas pesadas por Índices de validação e com uso de Amostragem.** Vila Real. 2008. 391p. Tese (Área de Engenharia Eletrotécnica e de Computadores). Universidade de Trás-os-Montes e Alto Douro, 2008.

DUARTE, M. C.; SANTOS, J. B.; MELO, L. C. Comparison of similarity coefficients based on RAPD markers in the common bean. **Genetics e Molecular Biology**, v. 22, n. 3, p. 427- 432, 1999.

EDWARDS, A. W. F; CAVALLI-SFORZA, L. L. A method for cluster analysis. **Biometrics**, v. 21, n. 2, p. 362 – 375, 1965.

EVERITT, B. S. **Companion to Multivariante Analysis**. 1. ed. London: Springer, 2005. 221 p.

FÀVERO, L. P. L. Análise de Dados: **Modelagem Multivariada para Tomada de Decisões**. 1. ed. Rio de Janeiro: Elsevier, 2009. 646 p.

FERREIRA, R. L. C. et al. Comparação de duas metodologias multivariadas no estudo de similaridade entre remanescentes de Floresta Atlântica. **Revista Árvore**, Viçosa-MG, v. 32, n. 3, p. 511- 521, 2008.

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

FLOREK, K. et al. "Sur la Liaison et la Division des Points d'un Ensemble Fini," **Colloquium Mathematicae**, v. 2, p.282 -285. 1951.

FORTES, F. O. et al. Agrupamento em amostras de sementes de espécies florestais nativas do Estado do Rio Grande do Sul - Brasil. **Ciência Rural**, Santa Maria, v. 38, n. 6, set, 2008.

FREITAS, S. M; PRATA, B. A. Uma nova abordagem para a análise de agrupamento com uma aplicação em agronomia. 12º Seagro. 2007.

GAULII, A.; GAILING, O.; STEFENON, V. M.; REINER, F. GENETIC. similarity of natural populations and plantations of *Pinus roxburghii* Sarg. in Nepal. **Annals of Forest Science.** v. 66, p. 693 – 903. 2009.

GONÇALVES, D. A.; ELDIK, T. V.; POKORNY, B. O uso do dendrômetro a laser em florestas tropicais: aplicações para o manejo florestal na Amazônia. **Floresta**, Curitiba, PR, v. 39, p. 175 - 187, 2009.

GOURLET-FLEURY, S. et al. Grouping species for predicting mixed tropical forest dynamics: looking for a strategy. **Annual Forest Science**, Bethesda, vol. 62, n. 8, p. 785 - 796, 2005.

GORDON, A. D. **Classification** - 2nd Edition. Chapman & Hall/CRC, second edition, 1999**.**

GOWER, J. C.; LEGENDRE, P. Metric e euclidean properties of dissimilarity coefficients, **Journal of Classification**, v. 3, p. 5 - 48, 1986.

GOWER, J. C. A comparison of some methods of cluster analysis. **Biometrics**, v. 23, p. 623 - 637, 1967.

HAIR, J. F. et al. **Multivariate Data Analysis**. 7. ed. Pearson Prentice Hall, 2010. 593 p.

HARTIGAN, J. A. "Consistency of Single Linkage for High-Density Clusters," **Journal of the American Statistical Association**, v. 76, p. 388 - 394, 1981.

HUANG, J. Y.; GUO X. P.; QIU Y. B.; and Chen Z. Y.; Cluster and discriminant analysis of electrochemical noise dta. **Electrochimica,** v. 53, p. 680 – 687, 2007.

HOLGERSSON, M. The limited value of cophenetic correlation as a clustering criterion. **Pattern Recognition**, v. 10, p. 287 – 295, 1978.

JACKSON, A. A.; SOMERS, K. M.; HARVERY, H. H. Similarity coefficients: measures for co-occurrence e association or simply measures of occurrence. **American Naturalist**, v.133, p. 436 - 453, 1989.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. Prentice Hall, Englewood Cliffs, NJ, 1988.

JAIN, A.; MURTY, M.; FLYNN, P. Data Clustering*: **A Review*.** **ACM Computing Surveys**, v. 31, n. 3, p. 264 - 323, 1999.

JING, L.; MICHEL, K. N.; JOSHUA, Z. H. “An Entropy Weighting k- Means Algorithm for Subspace Clustering of High-Dimensional Sparse Data”. **IEEE Transactions on Knowledge and Data Engineering**. v. 19, p. 1026-104, 2007.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical *Analysis**.6** ed. New Jersey: Upper Saddle River, 2007. 767 p.

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

KEESE A. M. **Àdaptação de viés indutivo de algoritmos de agrupamento de fluxos de dados .** 2012, 134 p. Tese (Doutorado em Ciências de Computação e Matemática Computacional). Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2012.

KHALED, M. H.; MOHAMED S. K. “**Incremental document clustering using cluster similarity histograms**”. IEEE/WIC International Conference on Web Intelligence (WI’03), p. 597, 2003.

KOOP, B. Hierarchical classification I: single method. **Biometrical Journal,** v**.** 20. p. 485 – 501, 1978.

KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nommetric hypothesis. **Psychometrika**, v. 29, p. 1 - 27, 1964.

KUNZ, S. H. et al. Análise da similaridade florística entre florestas do Alto Rio Xingu, da Bacia Amazônica e do Planalto Central. **Rev. bras. Bot**., São Paulo, v. 32, n. 4, p.725-736, Out/Dez, 2009.

LANCE, G. N., WILLIAMS, W. T. A general theory of classificatory sorting stategies, **Computer Journal**, v. 9, p. 373 - 380, 1967.

LATTIN, J. M.; DOUGLAS C.; PAUL E. G. **Análise de dados multivariados.** São Paulo: Cengage Learning, 2011. 455 p.

LEGENDRE, P.; LEGENDRE, L. **Numerical ecology**. Amsterdam, 2. ed English edition. Elsevier, 1998. 870 p.

LIMA JÚNIOR, L. M. et al. Utilização de técnicas multivariadas na classificação de fases de crescimentos de Leucaena leucocephala (Lam.) De Wit. **Floresta,** Curitiba, PR, v. 39, n. 4, p. 921-935, out./dez. 2009.

LINDEN, R. Técnicas de agrupamento, **Revista de Sistemas de Informação da FSMA**, Rio de Janeiro, vol. 4, p. 18 – 36, 2009.

LOBÃO, M. S. et al. Agrupamento de espécies florestais pela similaridade das características físico-anatômicas e usos da madeira. **Revista Cerne**, Lavras, v. 16, p. 97-105, 2010.

LUDEWIG, D. R. et al. O processo de gestão de custos e planejamento de resultados utilizando técnicas de análise estatística de agrupamentos**. Acta Scientiarum. Technology .** Maringá, v. 31, p. 215 – 220, 2009.

MAEDA, E. E.; FORMAGGIO, A. R.; SHIMABUKURO, Y. E. Análise histórica das transformações da Floresta Amazônica em áreas agrícolas na Bacia do Rio Suia-Miçu. **Sociedade & Natureza***,* Uberlândia, v. 20, n. 1, p. 5-24, 2008.

MARDIA, K. V.; KENT, J. T.; BIBBY, J. M. **Multivariate analysis**. London: Academic Press, 1979. 520.p.

MARTINS, et al. Conseqüências genéticas da regeneração natural de espécies arbóreas em área antrópica, AC, Brasil. **Acta Bot. Bras.**, São Paulo, v. 22, n. 3, set. 2008 .

MAXIMO, P. S. et al. Valoração de contingente pelas modelagens logit e análise multivariada: um estudo de caso da disposição a aceitar compensação dos cafeicultores vinculados ao PRO-CAFÉ de Viçosa - MG **Revista Árvore,** Viçosa-MG, v.33, n.6, pp. 1149 - 1157, 2009.

MCROBERTS, R. E. et al. Estimating areal means and variances of forest attributes using the k-nearest neighbours technique and satellite imagery . **Remote Sens. Environ**. v. 111, p. 466 – 480, 2007.

MCQUITTY, L. L. Hierarchical syndrome analysis for the isolation of types, Educational and Psychological. **Measurement**, v. 20, p. 55 – 67,1960.

MELO, A. S. ; HEPP, L. U. Ferramentas estatísticas para análise de dados provenientes de biomonitoramento. **Oecologia Brasiliensis**, v. 12, p. 463 - 486, 2008.

MEIRELES A. C. M; OLIVEIRA L. J. Sustentabilidade do modelo agrícola da bacia do riacho Faé **Revista Ciência Agronômica**, Fortaleza-CE, v. 42, n. 1, p. 84 - 91, jan-mar, 2011.

MESSETTI, A. V. L. **Estudo da semelhança de genótipos de girassol (Helianthus annuus L.) com o uso da distância generalizada de Mahalanobis na análise de agrupamento.** 2007. 87 p. Tese (Doutor em agronomia – área de concentração: energia na agricultura) – Universidade Estadual Paulista “Júlio de Mesquita Filho”. 2007.

MEYER, A. S. et al. Comparison of similarity coefficients used for cluster analysis with dominant markers in maize (Zea mays L.). **Genetics and Molecular Biology**, v. 27, n.1, p. 83 - 91, 2004.

MILLIGAN, G. N.; COOPER, M. C. An examination of procedures for determining the number of cluster in a data set. **Pfychometrika**, springer, Nova York, v. 50, n. 2, p. 159 – 179, 1985

MILLIGAN, G. W. An examination of the effect of six types of error perturbation on fifteen clustering algorithms. **Psychometrika**,v. 45, p. 325–342, 1980.

MILLIGAN, G. W. A Monte Carlo study of thirty internal criterion measures for cluster analysis. **Psychometrika**, v.46, n. 4, p.187–199, June 1981.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada**. Belo Horizonte: Editora UFMG, 1ª reimpressão. 2007. 297p.

MOSES, C. et al. “Incremental clustering and dynamic information retrieval”. **SIAM Journal on Computing**. v. 33, p. 1417 – 1440, 2004.

OLIVEIRA, S.N. et al. Delimitação automática de bacias de drenagens e análise multivariada de atributos morfométricos usando modelo digital de elevação hidrologicamente corrigido. **Revista Brasileira de Geomorfologia**, São Paulo, v. 8, n. 1, p. 3 – 21, 2007.

OKSANEN, J. **Cluster Analisys**: Tutorial with R, 2010. Disponível em HTTP://cc.oulu.fi/~jarioksa/opetus/metodi/session3.pdf. 2010.

ORLÓCI, L. **Multivariate analysis in vegetational research**. 2. ed. The Hague: Dr. W. Junk B. V. Publishers, 1978. 451 p.

PENNY. K. I. Appropriate Critical Values when Testing for a Single Multivariate Outlier by Using the Mahalanobis Distance. **Applied Statistics**, Londres, v. 45, n. 1, p. 73 – 81, 1996.

PREARO, L. C.; GOUVÊA, M. A.; ROMEIRO, M. C. Avaliação da adequação da aplicação de técnicas multivariadas de dependência em teses e dissertações de algumas instituições de ensino superior. **Ensaios FEE**, Porto Alegre, v. 33, n. 1, p. 261- 290, 2012.

RAO, C. R. **Advanced statistical methods in biometric research**. New York: John Wiley & Sons, 1952. 390 p.

RAND, W. M.. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical Association**, v.66, (336), p. 846–850, December 1971.

REIS, E. **Estatística multivariada aplicada**. Lisboa: Edições Silabo, 2001. 342 p.

RENCHER, A. C.; SCHAALJE, G. B. **Linear Models in Statistics**. 2. ed. New Jersey : John Wiley , 2008. 672 p.

ROBERTS, M. R.; GILLIAM, F. S. Patterns and mechanisms of plant diversity in forested ecosystems: implications for forest management. **Ecological Application**, v. 5, n. 4, p. 969 – 977, Nov. 1995.

RODE, R.; FIGUEIREDO F. A.; MACHADO, S. A.; GALVAO, F. Grupos florísticos e espécies discriminantes em povoamento de Araucaria angustifolia e uma floresta ombrófila mista. **Revista Árvore,** Viçosa-MG, v. 35, n. 2, p. 319 – 327, 2011.

RODY, Y. et al. Delimitação de sítios ambientais homogêneos no Estado do Espírito Santo, com base no relevo, solo e clima. **Ciência Rural,** Santa Maria, v.40, n.12, p. 2493 – 2498, 2010.

ROMESBURG, C. H. **Cluster analysis for researchers**. Belmont: Lifetime Learning Publications, 1984. 334 p.

SANTOS, N.; LONGHI, S. Percepção das paisagens da Floresta Nacional de Canela (RS) pelos turistas Landscapes perception of the of the National Forest of Canela (RS). **AMBIÊNCIA**, Guarapuava, v. 8, n. 1, p. 113 – 123, 2012.

SCHULZ H.; HÄRTLING S. Vitality analysis of Scots pines using a multivariate approach. **For. Ecol. Manage.** v. 186, p. 73 – 84, Dez, 2003.

SCHEEREN, L. W. et al. Agrupamento de unidades amostrais de Araucaria angustifolia (Bert.) O. Ktze. em função de variáveis do solo, da serapilheira e das acículas, na região de Canela, RS. **Ciência Florestal**, Santa Maria, v. 10, n. 2, p. 39 - 57, 2000.

SEIDEL, E.; OLIVEIRA, M.; TAVARES, B.; ANTONIALLI, L. Procedimento para Formação de Grupos de Empresas e para Construção de Índice de Avaliação dos Agrupamentos. **Revista Eletrônica Sistemas & Gestão**, América do Norte, v. 7, n. 1, 2012.

SCIPIONI, M. C. et al. Análise fitossociológica de um fragmento de floresta estacional em uma catena de solos no morro do cerrito, Santa Maria, RS. **Ciência Florestal,** Santa Maria, v. 22, n. 3, p. 457-466, jul.-set., 2012

SILVA, H. C. M. B. **Métodos de Partição e Validação em Análise Classificatória Baseados em Teoria de Grafos**. Porto. 2005. 275p. Tese (Doutorado em Matemática Aplicada). Faculdade de Ciências da Universidade do Porto, março de 2005.

SIQUEIRA, M. M. et AL. Análise de desempenho de vigas em madeira laminada colada de paricá (*Schizolobium amazonicum* Huber ex. Ducke). **Scientia Forestalis**, Piracicaba, v. 38, n. 87, p. 471- 480, 2010.

SNEATH, P. H. A; SOKAL, R. R. **Numeric taxonomy: the principles e practice of numerical classification**. San Francisco: W. H. Freeman, 1973. 573.p.

SOKAL, R. R.; MICHENER, C. D. A. statistical method for evaluating systematic relationships. **Bulletin of the Society** University of Kansas, n. 38, p. 109 - 143, 1958.

SOKAL, R.R.; ROHLF, F. J. The comparison of dendrograms by objective methods. **Taxon**, v. 11, p. 33 ‑ 40, 1962.

SOUZA M. D. et al. Análise de agrupamento e regressão não-linear aplicados ao crescimento in vitro de Leucoagaricus gongylophorus (Singer) Möller em meios de cultura acrescido com diferentes extratos vegetais. **Biotemas,** v. 24, n. 4, 85-93, Cuiabá – MT, dez. 2011.

SOUZA, A. L.; SOUZA, D. R. Análise multivariada para estratificação volumétrica de uma floresta ombrófila densa de terra firme, Amazônia oriental. **Rev**. **Árvore**, Viçosa-MG, v. 30, n. 1, p. 49 - 54, 2006.

SOUZA, A. L.; FERREIRA, R. L. C.; XAVIER, A. **Análise de Agrupamento aplicada à ciência florestal**. Viçosa: SIF, 1997, 109 f. (Documento SIF, 16).

SUSANTO, S.; KENNEDY, R. D.; PRICEE, J. H. A new fuzzy c-means and assignment technique based cell formation algorithm to perfor part-type cluster and machine-type cluster separately. **Production Planning and Control**, v. 10, n. 4, p. 375 - 388, 1999.

TOLEDO, L. O. et al. Análise multivariada de atributos pedológicos e fitossociológicos aplicada na caracterização de ambientes de cerrado no norte de Minas Gerais. **Revista Árvore**, Viçosa-MG**,** v. 33, n. 5, p. 957- 967, 2009.

VALENTE, M. D. R.; QUEIROZ, W. T.; PINHEIRO, J. G.; MONTEIRO, L. A. S. Modelo de predição para o volume total de Quaruba (*Vochysia inundata ducke*) via análise de fatores e regressão. **Revista Árvore,** Viçosa-MG**,** v. 35, n. 2, p. 307- 317, 2011.

VALE et al. Composição florística e estrutura do componente arbóreo em um remanescente primário de floresta estacional semidecidual em Araguari, Minas Gerais, Brasil. **Hoehnea**, v. 36, n. 3, p. 417- 429, 2009.

VASQUES, A., SILVA, J., ALMEIDA, A..A identificação da orientação estratégica da empresa florestal no Brasil – uma aplicação da teoria de porter. **Floresta**, Curitiba, PR, v. 41, n. 4, p. 695 - 706, 2011.

WARD, J. H. Hierarchical grouping to optimize an objective function. **Journal of. American Statistical Association**, v. 58, p. 236 - 244, 1963.

ZHANG J. et al. Density dependence on tree survival in an old-growth temperate Forest in northeastern China. **Annals of Forest Science**, Bethesda, v. 66, n. 7 , p. 204 – 210, 2009.