MÁCIO AUGUSTO DE ALBUQUERQUE

ESTABILIDADE EM ANÁLISE DE AGRUPAMENTO

(CLUSTER ANALYSIS)

Dissertação apresentada a Universidade Federal Rural de Pernambuco, para obtenção do titulo de Mestre em Biometria, Área de concentração: Modelagem e planejamento de experimentação.

Orientador: Prof°. Rinaldo Luiz Caraciolo Ferreira, Dr.
Co-orientadores: Prof°. José Antônio Aleixo da Silva, PhD.
Profº. Borko D. Stosic, PhD.

RECIFE
Estado de Pernambuco – Brasil
Fevereiro – 2005

Ficha catalográfica

Setor de Processos Técnicos da Biblioteca Central – UFRPE

A345e Albuquerque, Mácio Augusto de
Estabilidade em análise de agrupamento (cluster analysis) / Mácio Augusto de Albuquerque. – 2005.
62 f. : il., tabs.

Orientador: Rinaldo Luiz Caraciolo Ferreira
Dissertação (Mestrado em Biometria) – Universidade Federal Rural de Pernambuco. Departamento de Física e Matemática.
Refêrencias.

CDD 574.018 2

1. Tabela de contingência
2. Análise de agrupamento
3. Algoritmo de agrupamento
4. Mahalanobis
5. Técnica de hierarquização aglomerativa

I. Ferreira, Rinaldo Luiz Caraciolo
II. Título

Universidade Federal Rural de Pernambuco
Departamento Física e Matemática
Mestrado em Biometria

ESTABILIDADE EM ANÁLISE DE AGRUPAMENTO

(CLUSTER ANALYSIS)

MÁCIO AUGUSTO DE ALBUQUERQUE

Dissertação foi julgada adequada para obtenção do titulo de mestre em Biometria, defendida e aprovada por unanimidade em 23/02/2005 pela Banca Examinadora:

Orientador:

_____

Prof. Dr. Rinaldo Luiz Caraciolo Ferreira – UFRPE

Examinadores:

_____

Prof. PhD. José António Aleixo da Silva – UFRPE

_____

Prof. Dr. Eufrázio de Souza Santos - UFRPE

_____

Prof. PhD. Mário de Andrade Lira Júnior - UFRPE

RECIFE – PE

Fevereiro/2005

A minha família, em especial à minha esposa Edna e aos meus filhos Tarsyla e Tárcio e a minha mãe Luzia, 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 coordenador do curso de Biometria professor doutor Eufrázio de Souza Santos, pela orientação, pela dedicação e esforço pelo curso. Meu respeito e gratidão.

Especialmente ao professor Borko Stosic, pelas sugestões na elaboração da dissertação.

Aos professores, Gauss Moutinho Cordeiro, Paulo de Paula Mendes e Maria Adélia Oliveira Monteiro da Cruz, pela dedicação, apoio e pela transmissão do conhecimento no decorrer do curso.

Ao meu amigo e cunhado José Jerônimo de Araújo, pelo incentivo e ajuda nas pesquisas pela internet, dedicando todo o seu carinho e atenção e principalmente por poder ter contado sempre com seu conselho amigo.

E a Cilene Augusta da Nóbrega Veloso, pelo incentivo e ajuda no português, dedico todo carinho e atenção.

Ao colega amigo Antonio Lopes Pessoa, pelas caronas, por todo apoio nas horas difíceis e também pelos ótimos momentos vivenciados juntos.

Aos amigos Paulo Duarte, Adalberto Gomes, Marcela Verônica, Luiz de França, Cleto Bezerra, Nedson Pereira, Arundo Nunes, Ady Marinho, Heliovânio Torres, Ilzes Celi, Carlos André, Sérgio de Sá, Dâmocles Aurélio, Antônio José de Oliveira, pela ótima convivência durante o curso.

Ao amigo Bruno Cunha Coutinho, aluno do curso de Computação da UFPB, que contribuiu com os seus conhecimentos em programação C.

As secretárias, Josemar, Mary e ao secretário Marcos pelo carinho, respeito e amizade.

A Universidade Federal Rural de Pernambuco, pela oportunidade de realização do meu mestrado.

Á coordenadoria de Aperfeiçoamento de Pessoal de Nível Superior – CAPES, pela bolsa concedida.

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

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

Estabilidade em Análise de Agrupamento

Autor: Ms. Albuquerque, Macio Augusto
Orientador: Dr. Rinaldo Luiz Caraciolo Ferreira

RESUMO

Objetivou-se propor uma sistemática para o estudo e a interpretação da estabilidade dos métodos em análise de agrupamento, através de vários algoritmos de agrupamento em dados de vegetação. Utilizou-se dados provenientes de um levantamento na Mata da Silvicultura, da Universidade Federal de Viçosa-MG. Para análise de agrupamento foram estimadas as matrizes de distância de Mahalanobis com base nos dados originais e via reamostragem "bootstrap" e aplicados os métodos da ligação simples, ligação completa, médias das distâncias, do centróide, da mediana e do Ward. Para a detecção de associação entre os métodos foi aplicado o teste qui-quadrado. Para os diversos métodos de agrupamento foi obtida a correlação cofenética. Os resultados de associação dos métodos foram semelhantes, indicando em princípio que qualquer algoritmo de agrupamento estudado está estabilizado e existem, de fato, grupos entre os indivíduos observados. No entanto, observou-se que os métodos são coincidentes, exceto os métodos do centróide e Ward e os métodos do centróide e mediana quando comparados com o de Ward, respectivamente, com base nas matrizes de Mahalanobis a partir dos dados originais e "bootstrap". A sistemática proposta é promissora para o estudo e a interpretação da estabilidade dos métodos de análise de agrupamento em dados de vegetação.

STABILITY IN CLUSTER ANALYSIS

Autohr: Ms. Albuquerque, Mácio Augusto
Advisor: Dr. Rionaldo Luiz Caraciolo Ferreira

ABSTRACT

The main objective of this research was to propose a systematic to the study and interpretation of the stability of methods in cluster analysis through many cluster algorithms in vegetation data. The data set used came from a survey in the Silviculture Forest at Federal University of Viçosa – MG. To perform the cluster analysis the matrices of Mahalanobis distance were estimated based on the original data and by “bootstrap” resampling. Also the methods of single linkageage, complete linkageage, the average of the distances, the centroid, the medium and the Ward were used. For the detection of the association among the methods it was applied the chi-square test. For the various methods of clustering it was obtained a cofenetical correlation. The results of the associations of methods were very similar, indicating, in principle, that any algorithm of cluster studied is stabilized and exist, in fact, groups among the individuals analyzed. However, it was concluded that the methods coincide with themselves, except the methods of centroid and Ward. Also the centroid methods and average when compared to the Ward, respectively, based on the matrices of Mahalanobis starting from the original data set and “bootstrap”. The methodology proposed is promising to the study and interpretation of the stability of methods concerning the cluster analysis in vegetation data.

1- INTRODUÇÃO

As técnicas de análise multivariada possibilitam avaliar um conjunto de características, levando em consideração as correlações existentes, que permitem que inferências sobre o conjunto de variáveis sejam feitas em um nível de significância conhecido.

Nas diversas áreas do conhecimento uma das técnicas multivariadas mais utilizadas é a análise de agrupamento. O seu emprego em áreas tais como engenharia florestal, experimentos agronômicos, medicina, sociologia, administração, entre outras, vem aumentando muito nos últimos anos.

A análise de agrupamento tem por finalidade reunir, por algum critério de classificação as unidades amostrais em grupos, de tal forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos (Johnson & Wichern, 1992; Cruz & Regazzi, 1994).

O processo de agrupamento envolve basicamente duas etapas. A primeira se refere à estimação de uma medida de dissimilaridade entre os indivíduos e a segunda, refere-se à adoção de uma técnica de formação de grupos.

Um grande número de medidas de similaridade ou de dissimilaridade tem sido proposto e utilizado em análise de agrupamento, sendo a escolha entre elas baseada na preferência e/ou na conveniência do pesquisador (Bussab et al., 1990).

Com a definição da medida de dissimilaridade a ser utilizada, a etapa seguinte é a adoção de uma técnica de agrupamento para formação dos grupos. Para realização desta tarefa, existe um grande número de métodos disponíveis, dos quais o pesquisador tem de decidir qual o mais adequado ao seu propósito, uma vez que as diferentes técnicas podem levar a diferentes soluções (Souza et al., 1997).

As técnicas de análise de agrupamento exigem de seus usuários a tomada de uma série de decisões independentes, que requerem o conhecimento das propriedades dos diversos algoritmos à disposição e que podem representar diferentes agrupamentos. Além disso, o resultado dos agrupamentos pode ser influenciado pela escolha da medida de dissimilaridade, bem como pela definição do número de grupos (Gower & Legendre, 1986; Jackson et al., 1989; Duarte et al., 1999).

O incremento na capacidade computacional promoveu grandes avanços na análise multivariada, reduzindo simultaneamente o esforço e os custos. Uma dessas

possibilidades de otimização é o procedimento de reamostragem “bootstrap". Esses procedimentos têm sido utilizados para avaliar a estabilidade dos agrupamentos obtidos a partir de matrizes de dissimilaridade (Weir, 1990; Meyer, 1995; Hillis et al., 1996; Manly, 1997). Logo, aplicação do procedimento de reamostragem “bootstrap” pode fornecer um ponto de equilíbrio que permite uma estimativa precisa dos grupos.

Assim, objetivou-se propor uma sistemática para o estudo e a interpretação da estabilidade dos métodos em análise de agrupamento, através de vários algoritmos de agrupamento em dados de vegetação.

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.

2.1 A Análise de Agrupamento

A análise de agrupamento é uma técnica multivariada que tem por objetivo proporcionar uma ou várias partições na massa de dados, em grupos, por algum critério de classificação, de tal forma que exista homogeneidade dentro e heterogeneidade entre grupos (Sneath & Sokal, 1973; Mardia et al., 1997).

Essa técnica sumariza dados para interpretação e utiliza métodos que procuram grupos excludentes, ascendentes, reduzindo as informações de um conjunto de n indivíduos para informações de um novo conjunto de g grupos, onde g é significativamente menor que n, resultando um dendrograma de exclusão (Mardia et al., 1997).

Conforme Reis (1997), de modo sintético, a técnica pode ser descrita como se segue: dado um conjunto de n indivíduos para os quais existe informação sobre a forma p variáveis, o método de análise de agrupamento procede ao agrupamento dos indivíduos em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhante aos elementos do mesmo grupo do que aos elementos dos grupos restantes. Essa técnica é também chamada de técnica de partição, classificação ou taxonomia, embora o termo partição seja mais utilizado para uma das técnicas

especificas da análise: aquela em que os indivíduos são divididos por um número preestabelecido de grupos.

Segundo Aaker et al. (2001), a premissa mais importante da análise de agrupamento é a de que a medida de similaridade ou dissimilaridade na qual o processo de agrupamento se baseia é uma medida válida de similaridade ou dissimilaridade entre os indivíduos. A segunda premissa mais importante é a de que existe uma justificativa teórica para estruturar os indivíduos em grupos. Como em outras técnicas multivariadas, também há teoria e lógica guiando e dando base à análise de agrupamento.

Geralmente, 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 medida, legitimidade do resultado, aparência de uma hierarquia natural (quando for empregado um método não hierárquico) e confiabilidade de testes de divisão de amostra, oferecem informações úteis (Bussab et al., 1990). Entretanto, é difícil saber, exatamente, quais os grupos são muito parecidos e quais objetos são difíceis de serem inseridos. Geralmente, não é fácil selecionar um critério e programa de agrupamento por meio de outra referência que não a disponibilidade.

Na análise de agrupamento, é fundamental ter particular cuidado na seleção das variáveis de partida que vão caracterizar cada indivíduo, e determinar, em última instância, qual o grupo em que deve ser inscrito. Nesta análise não existe qualquer tipo de dependência entre as variáveis, isto é, os grupos se configuram por si mesmo sem necessidade de ser definida uma relação causal entre as variáveis utilizadas. Essa análise não faz uso de modelos aleatórios, mas é útil por fornecer um sumário bem justificado de um conjunto de dados. As técnicas são exploratórias e a idéia é, sobretudo gerar hipóteses, mais do que testá-las, sendo necessária a validação posterior dos resultados encontrados através da aplicação de outros métodos estatísticos (Reis, 1997).

Genericamente, a análise de agrupamento compreende cinco etapas (Aaker et al., 2001):

1. A seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
2. A definição de um conjunto de variáveis a partir das quais serão obtidas informações necessárias ao agrupamento dos indivíduos;

3. A definição de uma medida de semelhança ou distância entre os indivíduos;
4. A escolha de um algoritmo de partição/classificação;
5. Por último, a validação dos resultados encontrados.

2.2 As técnicas de análise de agrupamentos

A análise de agrupamento envolve algumas decisões subjetivas, como qual a técnica a mais conveniente, conforme as circunstâncias.

Vários são os tipos de técnicas de agrupamento encontradas na literatura (Johnson & Wichern, 1992; Cruz & Regazzi, 1994; Mardia el al., 1997; Aaker et al., 2001; Barroso & Artes, 2003), 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.

De maneira geral, ao empregar quaisquer procedimentos de análise de agrupamento, o pesquisador deve tomar cuidado com os seguintes aspectos (Aldenderfer & Brashield, 1984):

- A maioria dos métodos de análise de agrupamento é procedimento relativamente simples que, geralmente, não tem um embasamento teórico estatístico abrangente.

- Os métodos de análise de agrupamento foram desenvolvidos com base em diversas disciplinas, e os vieses herdados de cada uma delas podem diferir muito entre si.

- Métodos de agrupamentos diferentes geram soluções diferentes para o mesmo conjunto de dados.

- A estratégia da análise de agrupamentos busca uma estrutura, enquanto sua operação necessita de uma estrutura preestabelecida.

As próprias técnicas de agrupar podem ser “classificadas” em grupos, e diferentes autores produzem diferentes classificações. Cormack (1971), propõe a seguinte:

1) A técnica hierárquica de agrupamento consiste em uma série de sucessivos agrupamentos ou sucessivas divisões de elementos, em que os elementos são agregados ou desagregados. As técnicas hierárquicas são subdivididas em aglomerativas e divisivas.

Os grupos, na técnica hierárquica, são geralmente representados por um diagrama bi-dimensional chamado de dendrograma ou diagrama de árvore. Nesse diagrama, cada ramo representa um elemento, enquanto a raiz representa o agrupamento de todos os elementos.

2) As técnicas não-hierárquicas, ou por particionamento, foram desenvolvidos para agrupar elementos em K grupos, em que K é a quantidade de grupos definida previamente.

Nem todos valores de K apresentam grupos satisfatórios, sendo assim, aplica-se o método várias vezes para diferentes valores de K, escolhendo os resultados que apresentem melhor interpretação dos grupos ou uma melhor representação gráfica (Bussab et al., 1990).

A idéia central da maioria dos métodos por particionamento é escolher uma partição inicial dos elementos e, em seguida, alterar os membros dos grupos para obter-se a melhor partição (Anderberg, 1973).

Quando comparado com a técnica hierárquica, a técnica não-hierárquica ou por particionamento é mais rápido porque não é necessário calcular e armazenar, durante o processamento, a matriz de similaridade ou dissimilaridade (Johnson & Wichern, 1992).

Em geral, os métodos por particionamento diferem entre si pela maneira que constituem a melhor partição. Como qualquer classificação, existirão tipos que serão difíceis de classificar, ou que poderão caber em mais de um grupo.

2.2.1 Técnicas hierarquização

Segundo Reis (1997), as técnicas de hierarquização conduzem a uma hierarquia de partições P_1 P_2,...,P_n do conjunto total dos n objetos em 1, 2....n grupos. A denominação de hierárquicos advém do fato de, que, para cada par de partições P_i e P_{i+1}, cada grupo da partição P_{i+1} está sempre incluído num grupo da partição P_i.

Esse tipo de técnica se baseia na construção de uma matriz de dissimilaridade ou distâncias em que cada elemento da matriz descreve o grau de diferença entre cada dois casos com base nas variáveis escolhidas. Segundo Souza et al. (1997), os métodos hierárquicos se dividem em aglomerativos e divisivos. Entre os métodos aglomerativos, citam-se o do “vizinho mais próximo”, do “vizinho mais distante”, da “mediana”, do “centróide”, da “média das distância” e o proposto

por Ward (1963). Entre os métodos divisivos, o mais conhecido é o de “Edwards e Cavalli-Sforza (1965)”. Nos primeiros, parte-se de n grupos de apenas um indivíduo, que vão sendo agrupados, sucessivamente, até que se encontre apenas um grupo que incluirá a totalidade dos n indivíduos. O processo inverso é utilizado pelos métodos divisivos: parte-se de um grupo que inclui todos os indivíduos em estudo e por um processo sistemático de divisões sucessivas. Os métodos de análise de agrupamentos mais utilizados pelos pesquisadores são os hierárquicos aglomerativos.

O ponto de partida comum a todos os métodos hierárquicos é a construção de uma matriz de similaridade ou de distância, sendo este o terceiro problema a resolver em qualquer análise de agrupamento.

2.2.2 Definição do número de grupos

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

Para Barroso & Artes (2003), o número de grupos pode ser definido a priori, através de algum conhecimento que se tenha sobre os dados, pela conveniência do pesquisador, por simplicidade, ou ainda pode ser definido a posteriori com base nos resultados da análise.

De acordo com Aaker et al., (2001), 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 com a média de similaridade interna do agrupamento, é possível estabelecer certo nível que ditaria o número de agrupamentos. Uma terceira abordagem é determinar o número de agrupamentos com base no padrão gerado pelo programa. 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 quarta 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, seria a indicação do número adequado de agrupamentos. Aumentar esse número além desse ponto seria inútil, e diminuí-lo seria correr o risco de misturar objetos diferentes.

Qualquer que seja a abordagem empregada, geralmente é aconselhável observar o padrão total de agrupamentos. Isto pode proporcionar uma medida da qualidade do processo de agrupamento e do número de agrupamentos que emerge nos vários níveis do critério de agrupamento. Geralmente mais de um nível de agrupamento é relevante (Aaker et al., 2001).

No caso de não existir o conhecimento do número de grupos em que a população em estudo deverá ser dividida, um dos métodos usados quando se usam técnicas hierárquicas, consiste na comparação gráfica do número de agrupamento com o respectivo coeficiente de fusão, isto é, o valor numérico (distância ou semelhança) para o qual vários casos se unem para formar um grupo (Reis, 1997). Assim, quando a divisão de um novo grupo não introduz alterações significativas no coeficiente de fusão, poderá se tomar essa partição como sendo ótima. Outro procedimento utilizado é o da comparação dos resultados obtidos por vários métodos diferentes de agrupamento. Poder-se-á aferir o grau de convergência entre os vários métodos de agrupamento através de uma tabela de contingência, indicando o número de observações que se agrupam no mesmo agrupamento, para o mesmo número de agrupamento. Desta forma é possível verificar a maior ou menor estabilidade das soluções encontradas, de maneira a concluir acerca da qualidade do agrupamento efetuado.

2.2.3 Dendrograma

Dendrograma é uma representação matemática e ilustrativa de todo o procedimento de agrupamento através de uma estrutura de árvore (Everitt et al. 2001).

Os nós do dendrograma representam agrupamentos, E nós são compostos pelos grupos e ou objetos (grupos formados apenas por ele mesmo) ligados a ele (nó). Se cortarmos o dendrograma em um nível de distância desejado, obteremos uma classificação dos números de grupos existentes nesse nível e dos indivíduos que os formam. O número de grupo dos indivíduos é obtido pelo corte do dendrograma em um nível desejado e então cada componente conectado forma um grupo.

2.3 Medidas de distância

Para agrupar indivíduos, é necessário a definição de uma medida de similaridade ou dissimilaridade. Com base nessa medida os indivíduos similares são agrupados e os demais são colocados em grupos separados (Aaker et al., 2001):

As medidas de dissimilaridade têm papel central nos algoritmos de agrupamentos. Através delas são definidos critérios para avaliar se dois pontos estão próximos, e, portanto, podem fazer parte de um mesmo grupo, ou não.

Segundo Barroso & Artes (2003), há dois tipos de medidas de parecença: medidas de similaridade (quanto maior o valor, maior a semelhança entre os objetos) e medidas de dissimilaridade (quanto maior o valor, menor a semelhança entre os objetos).

De um modo geral, as medidas de similaridade e de dissimilaridade são interrelacionadas e, facilmente, transformáveis entre si (Bussab et al., 1990). Há um grande número de coeficientes de similaridade e/ou de dissimilaridade para caracteres binários disponíveis na literatura. Segundo Clifford & Stephenson (1975), tais coeficientes podem ser, facilmente, convertidos para coeficientes de dissimilaridade: se a similaridade for denominada **s**, a medida de dissimilaridade será o seu complementar (**1** – **s**).

A maioria dos métodos de análise de agrupamento requer uma medida de similaridade ou dissimilaridade entre os elementos a serem agrupados, normalmente expressa como uma função distância ou métrica (Doni, 2004).

Seja M um conjunto, uma métrica em M é uma função d: MxM $\rightarrow \Re$, tal que para quaisquer i, j, z $\in$ M, tenhamos:

1. d (i, j) = d (j, i) (simétrica);
2. d (i, j) > 0, se i≠ j;
3. d (i, j) = 0, se e somente se, i = j; e
4. d (i, j) ≤ d (i, z) + d (z, j) (desigualdade triangular).

Além disso, é esperado que d (i, j) aumente quando a dissimilaridade entre i e j aumentar.

Existem várias medidas que podem ser utilizadas como medidas de distâncias ou dissimilaridade entre elementos de uma matriz de dados, (Cormack, 1971) descreve uma série de medidas possíveis: distâncias euclidiana, euclidiana quadrada e euclidiana padronizada, distância corda, distância de Nei, distância

absoluta ou City – Block Metric, distância de Minkowski, distância Mahalanobis, distância de Chebychev.

Segundo Cormack (1971) as distâncias mais utilizadas em análise de agrupamento são:

1. Distância Euclidiana: a distância entre dois casos (i e j) é a raiz quadrada do somatório dos quadrados das diferenças entre valores de i e j para todas as variáveis (v = 1, 2, ,,,,, p).

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

em que

X_{iv} representa a característica do indivíduo i,

X_{jv} representa a característica do individuo j,

p é o número de parcelas na amostra,

v é o número indivíduo na amostra.

2. Distância Euclidiana quadrada: a distância entre dois casos (i e j) é definida como o somatório dos quadrados das diferenças entre os valores de i e j para todas as variáveis (v = 1, 2.....p).

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

em que:

X_{iv} representa a característica do indivíduo i,

X_{jv} representa a característica do individuo j,

p é o número de parcelas na amostra,

v é o número indivíduo na amostra.

3. Distância euclidiana ponderada.

Para Bussab et al., (1990), na realidade, este coeficiente de similaridade e dissimilaridade está associado a uma questão freqüente em análise de agrupamento, que é o da ponderação das variáveis, ou seja, o de dar mais peso para variáveis que o pesquisador julgar mais importante para definir semelhança.

$$d = \sqrt{(X_i - X_j)' S (X_i - X_j)}$$

S é uma matriz diagonal, tendo i-ésimo componente a variância s^2_i , isto é, S = diag. $(s_1^2, s_2^2, ..., s_p^2)$ então:

$(X_i - X_j)'$ é matriz transposta;

$(X_i - X_j)$ é uma matriz ;

X_i é o vetor de médias do i'ésimo indivíduo;

X_j é o vetor de médias do j'ésimo indivíduo.

Os casos particulares mais importantes, sendo a distância ponderada por S, são:

i) S = 1, a ponderação é a matriz identidade, tem-se então a distância euclidiana usual.

ii) S = [diag. $(s_1^2, s_2^2, ..., s_p^2)]^{-1}$, e tem-se a distância das variáveis padronizadas.

iii) S = V^{-1}, onde V é matriz de covariâncias, tem-se então a "distância de Mahalanobis".

Esta última distância, além de ponderar pela variabilidade de cada uma das componentes, leva em conta também o grau de correlação entre elas. Este fato torna muito difícil a interpretação de resultados baseados neste coeficiente de similaridade dissimilaridade.

4. Distância de Mahalanobis também chamada distância generalizada, Esta medida, ao contrário das apresentadas anteriormente, considera a matriz de covariância $\sum$ para o cálculo das distâncias: Esta é a distância escolhida para o nosso trabalho pois ela leva em consideração a estrutura de correlação existente nos dados.

Para Reis (1997), a distância generalizada D^2 de Mahalanobis também pode ser usada como técnica de comparação quando na separação entre diversos grupos, permitindo avaliar a extensão e a direção dos afastamentos entre os valores médios das variáveis usadas na discriminação. As diferenças entre cada par de grupos que estão sendo comparados são, assim, examinados simultaneamente por meio das diversas variáveis, que podem ser correlacionadas, de modo que, a

informação fornecida por uma delas pode não ser independente da fornecida pelas demais.

O valor numérico da maior separação possível entre dois grupos quaisquer é chamado distância generalizada entre os grupos e mede, em escala independente da originalmente utilizada para as várias variáveis, a clareza das disjunções entre elas.

Assim, o valor da distância generalizada D^2, ligando dois grupos, é um número puro, com propriedades da distância comum, e mede a extensão com que diferem entre si em tamanho e forma.

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

$$D^2_{ij} = (\underset{\sim}{X}_i - \underset{\sim}{X}_j)'. \Sigma^{-1} . (\underset{\sim}{X}_i - \underset{\sim}{X}_j)$$

em que:

$\underset{\sim}{X}_i$ é o vetor de médias do i'ésimo grupo;

$\underset{\sim}{X}_j$ é o vetor de médias do j'ésimo grupo;

Σ é a estimativa combinada da matriz da Covariância/variância dentro dos grupos.

Para o cálculo de D^2, supõe-se a existência de distribuição multinormal **p**-dimensional e a homogeneidade da matriz de covariância residual das unidades amostrais, restringindo-se, portanto, o seu uso. Entretanto, considerável robustez para violação dessas hipóteses já foi demonstradas, que faz da distância de Mahalanobis uma opção de grande utilidade, principalmente pelo fato de D^2 ter analogia com outras técnicas multivariadas (Cruz e Regazzi, 1994). Além disso, a distância de Mahalanobis considera a variabilidade dentro de cada unidade amostral, e não somente a medida de tendência central, sendo, portanto, uma medida mais aceitável, quando as unidades amostrais constituem um conjunto de indivíduos e, principalmente, quando as variáveis são correlacionadas (Riboldi, 1986).

Este método de representação de diferenças entre grupos leva em consideração qualquer correlação que exista as variáveis utilizadas, e é também independente das unidades de medida com que as variáveis estão expressas.

2.4 Algoritmos de agrupamento

Nos algoritmos de agrupamentos hierárquicos, conhecidos como SAHN ("Seqüencial, Agglomerative, Hierarquic, Nonoverlapping Clustering Methods"), em cada passo do agrupamento há a necessidade de recalcular o coeficiente de dissimilaridade entre os grupos estabelecidos e os possíveis candidatos a futuras admissões de novos membros nos grupos já estabelecidos (Sneath & Sokal, 1973).

Os vários métodos de agregação das espécies diferem no modo como estimam distância entre grupo já formado e outros grupos ou indivíduos por agrupar. O processo de agrupamento de indivíduos já agrupados depende da similaridade e dissimilaridade entre os grupos. Portanto, diferentes definições destas 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, mas alguns são mais indicados para determinadas situações do que outros (Kaufmann & Rosseeuw, 1990). É prática comum utilizar vários critérios e fazer a comparação dos resultados, se tais resultados forem semelhantes, é possível concluir que eles possuem um elevado grau de estabilidade e, portanto, são confiáveis.

Os métodos mais comuns de agrupamento para determinar a distância entre agrupamentos são: ligação simples, ligação completa, dos centróides, da mediana, das médias das distâncias e da soma de erros quadráticos ou variância (método Ward) (Anderberg, 1973).

2.4.1 Método da Ligação Simples

Este algoritmo, também denominado de método do elemento mais próximo (*Neighbourhoods*) é um dos mais simples, sendo de uso geral e de rápida aplicação.

O método da ligação simples, segundo Orlóci (1978); Gama (1980); 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, temos:

Seja $E = \{E_1, E_2, ,,, , E_p\}$ um conjunto de elementos 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 de vegetação, cada dimensão do espaço corresponde a uma espécie diferente, então, qualquer medida de distância estatística ou de similaridade pode ser empregada neste algoritmo.

Suponha 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).

No método da ligação simples, os agrupamentos entre objetos e grupos ou entre grupos são feitos por ligações simples entre pares de objetos, ou seja, a distância entre os grupos é definida como sendo aquela entre os objetos mais parecidos entre esses grupos. Este método leva a grupos longos se comparados aos grupos formados por outros métodos de agrupamentos SAHN (Meyer, 2002). Os dendrogramas, resultantes deste procedimento, são, geralmente, pouco informativos, devido à informação dos indivíduos intermediários que não são evidentes (Carlini-Garcia, 1998). De acordo com Sneath & Sokal (1973), agrupamentos pelo método de ligação simples podem ser obtidos tanto pelo procedimento aglomerativo quanto divisivo.

Anderberg (1973) cita as seguintes características desse método:

- Em geral, grupos muito próximos podem não ser identificados;
- Permite detectar grupos de formas não-elipticas;
- Apresenta pouca tolerância a ruído, pois tem tendência a incorporar os ruídos em um grupo já existente;
- Apresenta bons resultados tanto para distância Mahalanobis quanto para outras distâncias;
- Tendência a formar longas cadeias (encadeamento).

Encadeamento é um termo que descreve a situação em que há um primeiro grupo de um ou mais elementos que passa a incorporar, a cada interação, um grupo de apenas um elemento. Assim, é formada uma longa cadeia, onde se torna difícil definir um nível de corte para classificar os elementos em grupos (Romesburg, 1984).

2.4.2 Método da Ligação Completa

Este método é também denominado de método do elemento mais distante, sendo uma das técnicas de hierarquização aglomerativa de maior aplicação na

análise de agrupamento (Gama, 1980). Como no método de 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, também denominado vizinho mais distante, a dissimilaridade entre dois grupos é definida como sendo aquela apresentada pelos indivíduos de cada grupo que mais se parecem, ou seja, formam-se todos os pares com um membro de cada grupo, e a dissimilaridade entre os grupos é definida pelo par que mais se parece. Este método, geralmente, leva a grupos compactos e discretos, tendo os seus valores de dissimilaridade relativamente grande.

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

- Apresenta bons resultados tanto para a distâncias Mahalanobis quanto para outras distancias;
- Tendência a formar grupos compactos;
- Os ruídos demoram para serem incorporados ao grupo.

2.4.3 Método do Centróide

O método do centróide foi proposto por Sokal & Michener (1958) e teve como origem, a caracterização da matriz de dados como pontos do espaço Mahalanobis (I_p). Cada agrupamento é considerado um simples ponto, representado pelo seu centro de massa, chamado centróide. O presente método utiliza uma função de agrupamento para medir a distância entre os centros de massa dos dados. Esta técnica é de hierarquização aglomerativa.

Este algoritmo se caracteriza pela redefinição, a cada passo, da matriz de dados, em que cada agrupamento é representado pelo vetor médio das p variáveis envolvidas. Na realidade, uma nova matriz de distâncias é determinada a cada interação.

No método do centróide, a distância entre dois grupos é definida como a distância entre os seus centróides, pontos definidos pelas médias das variáveis caracterizadoras dos indivíduos de cada grupo, isto é, o método do centróide calcula a distância entre dois grupos como a diferença entre as suas médias, para todas as variáveis. Uma desvantagem desse método é que se os dois grupos forem muito diferentes em termos de dimensão, o centróide do novo agrupamento estará, mas

próximo daquele que for maior e as características do grupo menor tenderão a se perde. De fato, com esse método, o centróide do novo grupo é uma combinação ponderada dos centróides dos dois grupos separados, sendo as ponderações proporcionais ao tamanho destes grupos (Reis, 1997).

Uma característica importante desse algoritmo é a de que a distância entre agrupamentos é determinada pela distância entre os pontos representativos dos seus respectivos centros de massa (centróide).

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

- Robustez à presença de ruídos;
- Fenômeno da reversão.

O fenômeno da reversão ocorre quando a distância entre centróides é menor que à distância entre grupos já formados, isso fará com que os novos grupos sejam formados em um nível inferior aos grupos já existentes, tornando o dendrogama confuso (Romesburg, 1984).

2.4.4 Método da Mediana

Com base na metodologia proposta por Orlóci (1978) e Gama (1980), este algoritmo é um caso particular do método do centróide. A determinação da distância entre dois agrupamentos por meio do cálculo do centro de massa não considera o número de elementos em cada um dos agrupamentos. Assim, o vetor médio que representa o novo agrupamento, pode eventualmente, ficar situado entre os elementos do agrupamento com maior número de elementos. Para contornar este problema, Gower (1967) desenvolveu um procedimento de cálculo que pondera a medida de distância pelo número de elementos de cada agrupamento.

A rigor, os dois métodos são um só, não havendo razões para classificá-los como métodos distintos.

Para Barroso & Artes (2003), o método da mediana é uma modificação do método do centróide para a independência da distância do tamanho dos grupos. Se agregarem os grupos, com centróides a e b, para formar um novo grupo, a distância desse novo grupo a outro grupo, de centróide c, é a mediana.

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

- Apresenta resultado satisfatório quando os grupos possuem tamanhos diferentes;

- Pode apresentar resultado diferente quando permutados os elementos na matriz de dissimilaridade;
- Robustez à presença de outliers;
- Fenômeno da reversão.

2.4.5 Método das Médias das Distâncias (da Média de agrupamento)

Este método define a distância entre dois grupos como sendo a média das distâncias entre todos os pares de elementos, 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 (Everitt, 1974).

Os grupos são reunidos em um novo grupo quando a média das distâncias entre seus elementos é mínima.

No método das médias das distâncias se define a distância entre dois grupos, i e j, como sendo a média das distâncias entre todos os pares de objetos constituídos por elementos dos dois grupos. A estratégia e o valor médio tem 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 indivíduos no qual cada um tem mais semelhanças, em média, com todos os membros do mesmo grupo do que com todos os elementos de qualquer outro grupo (Reis, 1997).

Kaufmann & Rosseeuw (1990) destaca as seguintes características desse método:

- Menor sensibilidade a ruídos que os métodos de ligação simples e completa;
- Apresenta bons resultados tanto para a distância Mahalanobis quanto para outras distâncias;
- Tendência a formar grupos com número de elementos similares.

2.4.6 Método de Ward

Ward (1963) propõe um processo geral de classificação em que n elementos são progressivamente reunidos dentro de grupos através da minimização de uma função objetiva para cada (n -2) passos de fusão.

Inicialmente, este algoritmo admite que cada um dos elementos se constituía em um único agrupamento. Considerando a primeira reunião de elementos em um

novo agrupamento, a soma dos desvios dos pontos representativos de seus elementos, 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 os elementos de E em um agrupamento (Gama, 1980).

Conforme a proposta de Bouroche & Saporta (1972), quando os indivíduos são pontos de um espaço Mahalanobis (I_p), a qualidade de uma partição é definida por sua inércia intraclasse ou por sua inércia interclasse. Quando se parte de K+1 classes para K classes, ou seja, agrupando duas classes numa só, a inércia interclasse só pode diminuir. A inércia interclasse é a média da soma dos quadrados das distâncias entre os centros de gravidade de cada classe e o centro de gravidade total.

Gama propôs (1980), que a reunião de elementos em grupos é feita pela análise dos valores da função de agrupamento, reunindo-se os elementos mais próximos, isto é, aqueles que apresentassem Min (d_{ij}).

Conforme Reis (1997), o método de Ward se baseia na perda de informação resultante do agrupamento das espécies e medida através da soma dos quadrados dos desvios das observações individuais relativamente às médias dos grupos em que são classificadas.

Cada grupo se caracteriza por uma soma dos quadrados dos desvios de cada observações do centróide do mesmo (é uma soma dos numeradores dos estimadores das variâncias de cada variável dentro do grupo, é também a soma de distância Mahalanobis do quadrado de cada observação do centróide). A distância entre dois grupos se define como o aumento que se pronunciaria nesta soma de quadrados, se ambos os grupos se agregassem para a formação de um único grupo. O método de Ward é atraente por se basear numa 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 desse método:

- Apresenta bons resultados tanto para distâncias Mahalanobis quanto para outras distâncias;
- Pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual;
- Tem tendência a combinar grupos com poucos elementos;
- Sensível à presença de outliers.

2.5 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 então examinar os perfis médios de valores de cada agrupamento mediante subamostras. Outra alternativa é deletar diversas colunas (variáveis) 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 (variáveis). 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, mas não contenham nenhum agrupamento. Em seguida, aplicam-se os métodos de agrupamento nos dados reais e nos artificiais, e comparam-se as soluções resultantes (Aaker et al., 2001).

2.6 Bootstrap

O “bootstrap” é uma técnica estatística computacionalmente intensiva que permite a avaliação da variabilidade de estatística, com base nos dados de uma única amostra existente. Essa técnica foi introduzida por Efron (1979), e, desde então, tem merecido profundo estudo por parte dos estatísticos que trabalha com análise multivariada, não só na parte teórica, como também na aplicada.

Na estatística, as situações difíceis podem ser vistas como os problemas de soluções analíticas complexas, e as variadas soluções possíveis seriam a utilização

de uma metodologia com grande quantidade de cálculos, para analisar um pequeno conjunto de dados. A solução para esses casos, com o uso de métodos computacionalmente intensivos, é obtida substituindo-se o poder analítico das expressões teóricas pelo poder de processamento dos computadores.

A idéia chave do método é a amostra “bootstrap”, que é retirada da amostra original com reposição. Dessa forma, todo resultado “bootstrap” depende diretamente da amostra original observada, isto é, os resultados “bootstrap” são robustos para a amostra original. Algumas considerações de regularidade sob as quais esse método é consistente foram discutidas por Bickel & Freedman (1981). Os conceitos básicos, propriedades teóricas e aplicações podem ser encontrados em Efron & Tibishirani (1993).

O “bootstrap” pode ser implementado tanto na estatística não paramétrica quanto na paramétrica, dependendo apenas do conhecimento do problema. No caso não-paramétrico, o método “bootstrap” reamostra os dados com reposição, de acordo com uma distribuição subjacente aos dados. No caso paramétrico, quando se tem informação suficiente sobre a forma da distribuição dos dados, a amostra “bootstrap” é formada realizando-se a amostragem diretamente nessa distribuição com os parâmetros desconhecidos substituídos por estimativas paramétricas. A distribuição da estatística de interesse aplicada aos valores da amostra “bootstrap”, condicional aos dados observados, é definida como a distribuição “bootstrap” dessa estatística.

Operacionalmente, o procedimento “bootstrap” consiste na reamostragem de mesmo tamanho e com reposição dos dados da amostra original (Figura 1), e cálculo da estatística de interesse para cada reamostra “bootstrap” (pseudo-dados).

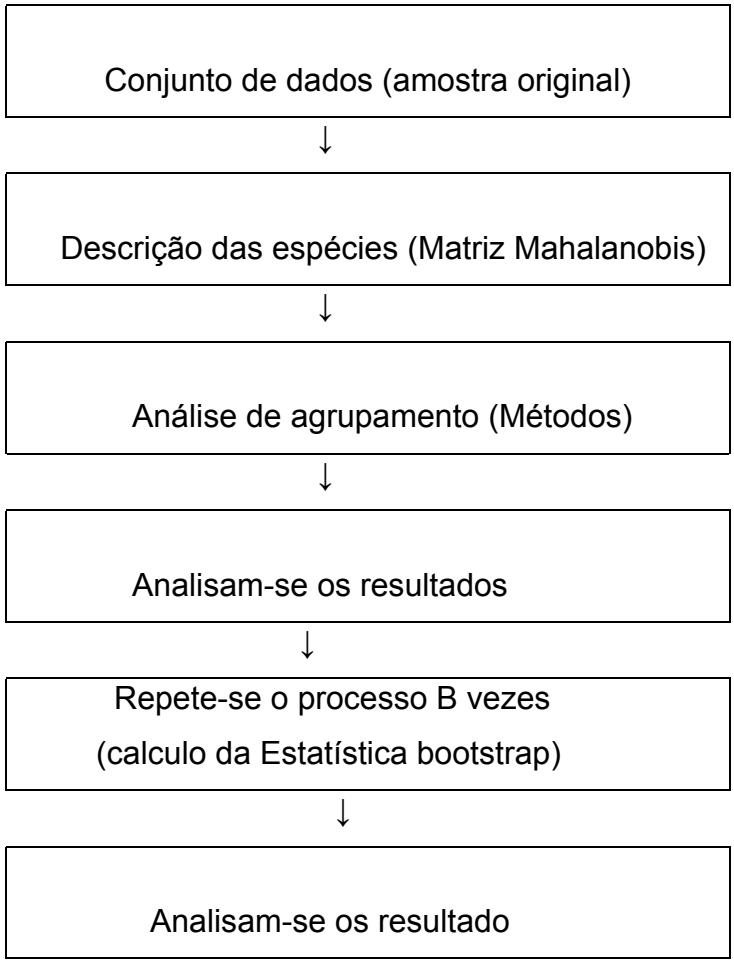

Figura 1 - Diagrama esquemático ilustrando a construção da distribuição “bootstrap”.

Na prática constrói-se a distribuição “bootstap” por Monte-Carlo com um número de repetições B, suficientemente grande. Um indicador do tamanho adequado de B, independente do custo computacional, é a qualidade da convergência da estimativa “bootstrap” do parâmetro para estimativa natural do parâmetro (Lavaranti, 2003).

O método mais simples para estimar intervalos de confiança por “bootstrap” é o método percentil. Esse método consiste em encontrar a distribuição F ”bootstrap” e calcular os percentis da distribuição que correspondem aos limites inferiores e superiores, respectivamente, do intervalo de confiança (Efron & Tibshirani, 1993). Uma versão melhorada desse método é chamada Bca, que é uma abreviação de (“bias – corrected and accelerated”). O intervalo Bca é um intervalo rigorosamente exato naquelas situações em que a estatística teórica tem uma resposta exata adaptada de Araújo (2003), dando uma precisão na estimação dos intervalos

confiança em todas as situações para dados obtidos da distribuição “bootstrap” (Diciccio & Efron, 1996).

Finalmente, a grande vantagem dos procedimentos estatísticos “bootstrap” é a estimação da precisão de qualquer estatística (multivariada na análise de agrupamento) com a comparação dos métodos. Esses métodos começaram a se tornar ferramentas bastante úteis e poderosas na construção de procedimentos estatísticos, evitando obtenção de fórmulas via argumentos analíticos.

2. 7 Correlação Cofenética

A correlação cofenética é uma medida de validação utilizada, principalmente, nos métodos de agrupamento hierárquicos. A idéia básica é realizar uma comparação entre as distâncias efetivamente observadas entre os objetos e distâncias previstas a partir do processo de agrupamento (Barroso & Artes, 2003).

A correlação cofenética mede o grau de ajuste entre a matriz de dissimilaridade original (matriz D) e a matriz resultante da simplificação proporcionada pelo método de agrupamento (matriz C). No caso, C é aquela obtida após a construção do dendrograma. Tal correlação foi calculada conforme Bussab et al., (1990):

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

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

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

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

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

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

Para Bussab et al., (1990), problema é responder se o valor observado é alto ou baixo?. Responder a isto é tão difícil como responder, na maioria das situações, o que é um alto coeficiente de correlação entre duas variáveis. Depende da área de estudo e de padrões que vão se desenvolvendo com a prática. Pode-se adiantar que em análise de agrupamento, algo em torno de 0,8 já pode ser considerado bom ajuste.

3. MATERIAL E MÉTODOS

3.1 Dados

Neste trabalho foram utilizados dados de um levantamento da vegetação da mata da Silvicultura (Tabela 1), da Universidade Federal de Viçosa-MG, retirado de Souza et al. (1997).

Tabela 1 – Densidade de 17 espécies da mata da silvicultura, em parcelas de 20 X 50 m, Universidade Federal de Viçosa – MG

Espécies	Parcelas										
	1	2	3	4	5	6	7	8	9	10	11
Casearia decandra Jacq.	8	1	27	0	1	9	2	3	22	15	7
Anadenanthera peregrina Spreg.	0	0	0	0	0	0	12	1	17	1	9
Apuleia leiocarpa (Vog.) Macbr.	3	9	4	6	22	9	5	2	7	4	4
Mabea fistulifera Mart.	6	3	3	4	29	12	0	4	4	4	4
Anadenanthera macrocarpa(Benth.) Brenan.	0	12	0	1	0	0	1	0	2	0	0
Platypodium elegans Vog.	0	0	1	1	9	1	0	0	5	11	1
Machaerium floridum (Benth.) Ducke	0	0	10	1	9	2	1	0	0	11	5
Copaifera langsdorffii Desf.	1	1	0	2	1	13	0	0	0	3	1
Ocotea pretiosa Mez.	2	0	2	2	2	6	0	5	0	2	2
Cabralea cangerana Saldanha	1	0	0	2	0	0	1	6	2	3	1
Piptadenia gonoacantha Macbr.	0	0	0	0	0	0	6	0	1	0	5
Dalbergia nigra Allem. Ex Benth.	5	0	7	0	5	0	0	0	0	1	0
Luehea divaricata Mart.	7	0	0	0	0	1	0	1	0	0	0
Melanoxylon brauna Schott.	0	0	0	0	0	0	0	0	0	2	1
Cedrela fissilis Vell.	0	0	0	0	0	0	1	0	0	0	0
Croton floribundus Spreng.	0	0	1	0	0	0	0	0	0	0	0

Fonte: Souza et al. (1997)

Considerando que a densidade da espécie é uma variável quantitativa discreta, os dados originais da matriz foram transformados por meio de raiz quadrada, para tornar a sua distribuição mais apropriada a uma análise de agrupamento.

3.2 Métodos Estatísticos

Com proposta de metodologia estatística para análise de agrupamento se considerou a seguinte ordem: processo reamostragem “bootstrap” na matriz de dados originais, cálculo da matriz de distância, considerando-se a distância de Mahalanobis, utilização dos algoritmos ligação simples, ligação completa, do centróides, da mediana, das distâncias médias e de Ward, dendrogramas, tabelas de contingências, e teste χ^2 para comparação dos métodos.

3.2.1 Estabilidade via método “bootstrap”

Para se estabilizar os métodos em análise de agrupamentos via bootstrap, foram seguidos os seguintes passos:

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

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

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

2. Com a matriz primária, encontrou-se a matriz de distância Mahalanobis, para aplicação dos algoritmos de agrupamento.

3. De posse da matriz de Mahalanobis, aplicou-se “bootstrap” e calculou-se uma nova matriz de distância Mahalanobis para aplicação dos algoritmos de agrupamento e comparação com a aplicação do item 2.

4. Construção de Tabelas de contingência 2x2 para comparação entre algoritmos de agrupamentos (número de observações que se agrupam no mesmo grupo para o mesmo número de grupo).

$$\text{Porcentagem} = \frac{\text{Número de observações}}{\text{Número de grupos}} \cdot 100\%$$

5. Conforme Jairo & Gilberto (1996) foi utilizado o indicador do grau de associação entre dois métodos analisandos dados por:

$$C = \sqrt{\frac{\chi^2 \text{calculado}}{\chi^2 \text{calculado} + N}}$$

Esses coeficiente podem variar entre [0 , 1], estando mais associados os métodos quanto maior é o valor de C.

3.2.2 Medida de distância

Como medida de dissimilaridade foi utilizada a distância de Mahalanobis (D^2) calculada conforme a seguinte expressão:

$$D^2 = (\underset{\sim}{X}_i - \underset{\sim}{X}_j)' \cdot \Sigma^{-1} \cdot (\underset{\sim}{X}_i - \underset{\sim}{X}_j),$$

em que $D^2 = d(\underset{\sim}{X}_i - \underset{\sim}{X}_j)$; Σ^{-1} é a inversa da matriz de covariância residual de X, e D^2 tem a característica de ser invariante para qualquer transformação linear não-singular. Ao contrário da distância Euclidiana, D^2 pode ser utilizada quando existe correlação entre as variáveis. Sendo os coeficientes de correlação nulos, o valor de D^2 equivale à distância Euclidiana para variáveis padronizadas.

3.2.3 Algoritmos de agrupamento

Foram utilizados os seguintes algoritmos de agrupamento, por serem os mais usados na prática pela facilidade de serem encontrados nos mais diversos programas computacionais, tais como:

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 seqüência de cálculos:

1. Com base na matriz de Mahalanobis original ou estabilizada via bootstrap foram determinados os valores da função de agrupamento d_{ij}, que foram representados na forma matricial (D_1);

2. Localizou-se o valor mínimo de $d_{ij} > 0$. Os elementos 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 os demais elementos, por meio da relação:

$d_{(i,j)\,l} = \min (d_{i1}, d_{i2})$, 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 os elementos 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 agrupamento formado na etapa anterior e os demais, considerando-se um elemento isolado de E como um agrupamento. Retornou-se a seguir a etapa 4.

Os processos foram repetidos até que todos os 11 elementos de E fossem alocados a um só agrupamento.

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

Dados n elementos e se admitindo 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 Gama (1980) e Mardia et al, (1997), nas seguintes etapas:

1.Determinou-se, com base na matriz de Mahalanobis, 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, os elementos E_i e E_j reunidos num primeiro grupo. Então se pressupõe que os (n-2) elementos restantes constituíssem, cada um, um agrupamento distinto;

3. Com base na matriz D_1, determina-se a distância entre cada um dos (n-2) elementos e o novo agrupamento formado pelos elementos (E_i, E_j). Esta distância é calculada pela relação:

$d_{(i, j) l}$ = máx (d_{i1}, 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 os elementos correspondentes, dando 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 do Centróide*.

Seja X (n x p) a matriz de dados básicos em que cada linha representa o conjunto de valores observados para cada espécie, e cada coluna representa uma parcela de área fixa ou variável.

Este algoritmo pode ser desenvolvido nas seguintes etapas:

1. Com base na matriz de dados e em uma função de agrupamentos, calcula-se as distâncias (d_{ij}) entre as parcelas (i, j), i = 1, 2...., n. Esses dados são reunidos numa matriz de distâncias (D).

2. A primeira operação em D, consiste em procurar pelo menor valor de d_{ij}, excluindo os valores de d_{ij} em que i = j, ou seja, excluir-se os elementos da diagonal principal. Os elementos X_i e X_j cuja similaridade é maior, são reunidos num mesmo agrupamento.

3. Em seguida, calcula-se uma nova matriz de distância e identificou-se os elementos do agrupamento mais próximo, e se constrói um novo agrupamento, Segundo Lance e Williams (1967) as distâncias podem ser obtidas através da seguinte expressão:

$$d_{k(ij)} = \left(\frac{n_i}{n_i + n_j} \right) \cdot d_{ki} + \left(\frac{n_j}{n_i + n_j} \right) \cdot d_{kj} - \left[\frac{n_i . n_j}{(n_i + n_j)^2} \right] \cdot d_{ij}$$

em que:

$d_{k(ij)}$, d_{ki}, d_{kj} e d_{ij} = distâncias Mahalanobis entre os elementos k e agrupamento ij, k e i, k e j, e i e j, respectivamente.

n_i, n_j e n_k = número de elementos nos grupamentos i, j e k, respectivamente.

4. Verifica-se se o número de agrupamentos determinados é igual ao valor fixado, $g \leq n$, se fosse verdade, termina-se o processo, caso contrário, retorna-se ao item 2.

d) *Método da Mediana*

Esse algoritmo foi desenvolvido nas seguintes etapas:

1. Com base na matriz de dados e em uma função de agrupamentos, calcula-se as distâncias (d_{ij}) entre as parcelas (i, j), i = 1, 2,...., n, Estes dados são reunidos numa matriz de distâncias (D).

2. A primeira operação em D, consistiu em procurar menor valor de d_{ij}, excluindo os valores de d_{ij} onde i = j, ou seja, excluir os elementos da diagonal principal, Os elementos (parcelas) X_i e X_j cuja dissimilaridade é menor são reunidos num mesmo agrupamento.

3. Em seguida é calculada uma nova matriz de distância e identificou-se os elementos do agrupamento mais próximo, constroem-se um novo agrupamento. Segundo Lance e Williams (1967), as distâncias podem ser obtidas através da seguinte expressão:

$$d_{k(ij)} = \left(\frac{1}{2} \right) \cdot d_{ki} + \left(\frac{1}{2} \right) \cdot d_{kj} - \left[\frac{1}{4} \right] \cdot d_{ij}$$

em que:

$d_{k\ (ij)}$, d_{ki}, d_{kj} e d_{ij} = distâncias Mahalanobis entre os elementos k e agrupamento ij, k e i, k e j, e i e j, respectivamente.

n_i, n_j e n_k = número de elementos nos grupamentos i, j e k, respectivamente.

4 Verifica -se se o número de agrupamentos determinados era igual ao valor fixado, g ≤ n, Se for verdade, termina-se o processo, caso contrário, retorna-se ao item 2.

e) *Método das Médias das distâncias (da Média de agrupamento)*

O método pode ser resumido nos seguintes passos:

1. Determina-se a matriz de distâncias inicial.
2. Localiza-se os dois elementos 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 seus elementos, sendo um elemento de cada um dos grupos.
4. Os dois grupos que apresentam menor distâncias são reunidos em um único grupo.
5. Se o número de grupos obtidos é igual a um número g ≤ n, o processo termina caso contrario, retorna-se ao passo 3.

Esta fórmula fornece das 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 os elementos k e agrupamento ij, k e i, k e j, e i e j, respectivamente.

n_i e n_j = número de elementos nos grupamentos i e j, respectivamente.

f) *Método de Ward*

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

1. Determina-se a matriz de distâncias e localiza-se os dois agrupamentos para os quais d_{ij} é mínimo;
2. Reúne-se estes agrupamentos, formando um novo agrupamento, e se verifica, se o número de agrupamentos (g) 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. Determina-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 Mahalanobis 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 dos elementos 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 elementos, 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:

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

3.2.4 Dendrogramas

A seqüência de fusão dos agrupamentos é representada graficamente nos dendrogramas, que foram divididas com a estatística descritiva usando o percentil, com um corte de 50% da distância Mahalanobis, para determinar o número de grupos. Os dendrogramas são construídos usando o programa computacional Minitab. Os diferentes dendrogramas obtidos são então comparados para possibilitar a análise da associação entre métodos.

3.2.5 Correlação Cofenética

Para os diversos métodos de agrupamento utilizados foram obtidas as respectivas matrizes cofenéticas resultantes da simplificação proporcionada pelo método. A matriz cofenética foi obtida após a construção do dendrograma. Com base nas matrizes de dissimilaridade original e cofenética, foi obtida a correlação cofenética conforme a expressão (Bussab et al., 1990):

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

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

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

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

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

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

3.2.6 Distorção entre a matriz de dissimilaridade e matriz cofenética

O grau da distorção (1 – α) foi calculado conforme Kruskal (1964):

$$\alpha = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}c_{ij}}{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}d_{ij}},$$

em que:

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

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

Esse parâmetro mede a distorção entre a matriz original e bootstrap e aquela obtida após a construção do dendrograma.

3.2.7 Tabelas de contingência

Foram construídas tabelas de contingência bi-dimensional (Tabela 2), na qual uma amostra de N observações foi classificada com relação os dois métodos de agrupamentos aplicados. Desta forma pôde-se cruzar as diversas características relevantes aos métodos pesquisados com diversas variáveis, tomadas duas a duas. Cada uma das células (n_{11}, n_{12}, n_{ij}, etc.) representou a associação ou a contagem de grupos em cada um dos métodos aplicado.

Tabela 2. Forma geral de uma tabela de contingência de duas dimensões

Método I	Método J					Total
	1	2	j	...	J	
1	n_{11}	n_{12}	n_{1j}	...	n_{1J}	n_1
2	n_{21}	n_{22}	n_{2j}	...	n_{2J}	n_2
...	...	...	...	...	...	...
i	n_{i1}	n_{i2}	n_{ij}	...	n_{iJ}	n_i
...	...	...	...	...	...	...
I	n_{I1}	n_{I2}	n_{Ij}	...	n_{IJ}	n_I
Total	$n_{.1}$	$n_{.2}$	$n_{.j}$	...	$n_{.J}$	N

Fonte: modificado de Everitt (1992)

3.2.7.1 Independência e associação entre métodos

Para a detecção de associação entre os métodos, ou seja, saber se as diferenças observadas entre métodos são significativas o suficiente para serem atribuídas a outros fatores que não aleatórios, foi aplicado o teste qui-quadrado (χ^2) por meio da expressão:

$$\chi 2 = \sum \frac{(\text{frequência observada} - \text{frequência esperada})}{\text{freqüência esperada}}$$

onde:

$$\text{freqüência esperada} = \frac{\sum L \sum C}{N}$$

em que:

L = número de categorias da variável disposta na linha da tabela de contingência;
C = número de categorias da variável disposta na coluna da tabela de contingência.

O método usado para decidir se o teste χ^2 é independente ou se não estão associado (ou seja se é possível ou não rejeitar a hipótese de nulidade, H_0), foi baseado na distribuição de probabilidade para χ^2 sob a pressuposição de que a hipótese nula é verdadeira. Dessa forma, quando a estatística χ^2 calculada foi maior do que o valor tabulado para um determinado nível de significância, H_0 foi rejeitada. Caso contrário, a hipótese nula não foi rejeitada.

O número de graus de liberdade (GL), para decidir quando o valor de χ^2 obtido de alguma tabela de contingência leva a uma rejeição ou não da hipótese nula, foi assim definido:

$$GL = (L - 1)(C - 1)$$

Todas os gráficos e as análises, ao longo deste trabalho foram implementados através dos programas computacionais EXCEL, STATISTICA, MINITAB e a construção de um programa na linguagem C.

4. RESULTADOS E DISCUSSÃO

4.1 Análise de agrupamento a partir da matriz de Mahalanobis original

Com base na matriz de dissimilaridade de Mahalanobis obtida a partir dos dados originais (Tabela 3) foram aplicados os métodos da ligação simples, da ligação completa, do centróide, da mediana, da média das distâncias e de Ward e obtidos os respectivos dendrogramas (Figuras de 2 a 7).

Tabela 3. Matriz de distância Mahalanobis dos dados originais para as 11 parcelas da Mata da Silvicultura, da Universidade Federal de Viçosa - MG

Parcela	1	2	3	4	5	6	7	8	9	10	11
1	00,00	16,81	20,01	15,93	20,35	14,40	19,50	13,37	20,11	18,53	17,66
2		00,00	18,50	7,721	18,51	13,81	18,80	18,59	15,59	20,89	19,03
3			00,00	17,68	20,80	18,21	22,27	20,08	18,48	20,55	19,10
4				00,00	12,57	10,32	16,73	7,092	14,54	15,04	13,84
5					00,00	17,65	22,86	20,99	19,20	20,37	17,55
6						00,00	19,90	16,21	16,89	18,27	13,12
7							00,00	20,02	17,36	22,69	16,47
8								00,00	15,11	18,29	14,05
9									00,00	14,87	13,20
10										00,00	11,90
11											00,00

Embora a estrutura geral dos agrupamentos seja bastante similar, pode-se observar que há pequenas alterações nos níveis em que os indivíduos são agrupados, ou seja, os indivíduos que estão dentro de um mesmo grupo podem ser agrupados em outra ordem, quando se mudam os métodos.

Pode-se observar que há divergências entre os métodos, corroborando com a afirmativa de Johnson & Wichern (1992), de que dificilmente os dendrogramas obtidos por métodos de agrupamentos diferentes sejam semelhantes. 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.

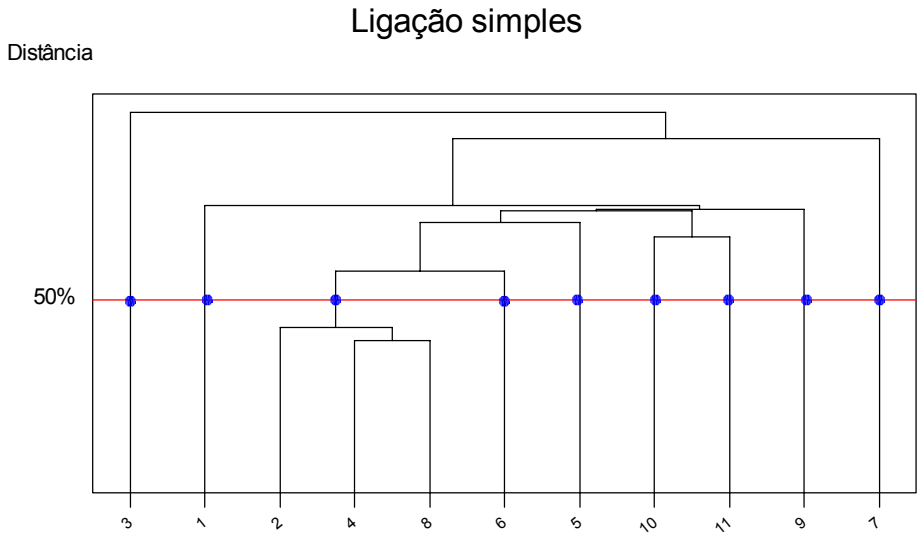

Figura 2 – Dendrograma representando as seqüências das fusões das parcelas, obtidas pelo emprego do método da ligação simples, com base na distância Mahanalobis dos dados originais.

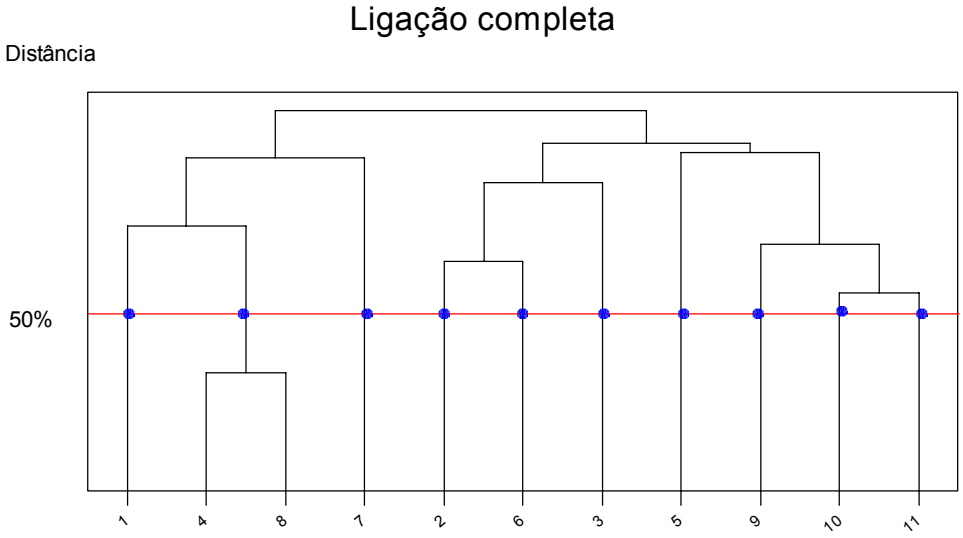

Figura 3 – Dendrograma representando as seqüências das fusões das parcelas, obtida pelo emprego do método ligação completa, com base na distância Mahalanobis dos dados originais.

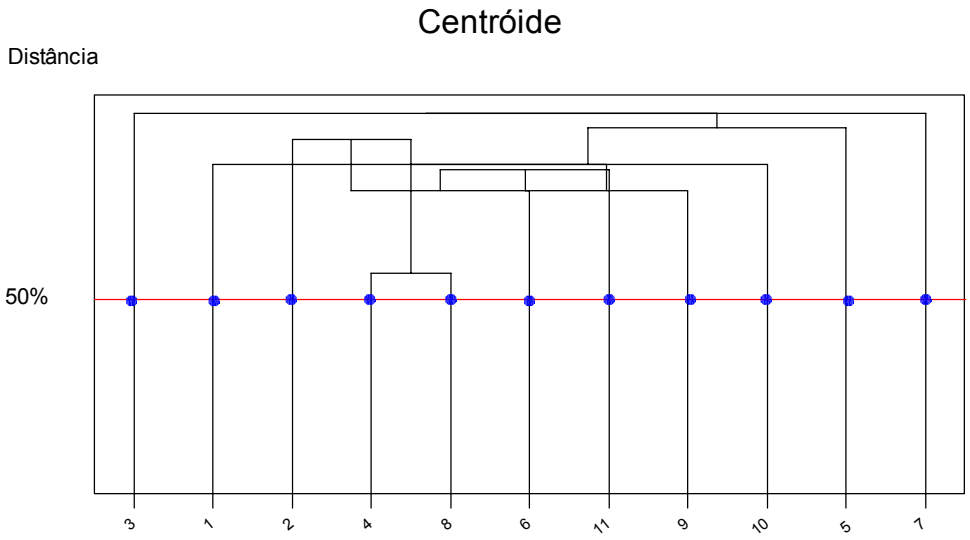

Figura 4 – Dendrograma representando as seqüências das fusões das parcelas, obtida pelo emprego do método da centróide, com base na distância Mahalanobis dados originais.

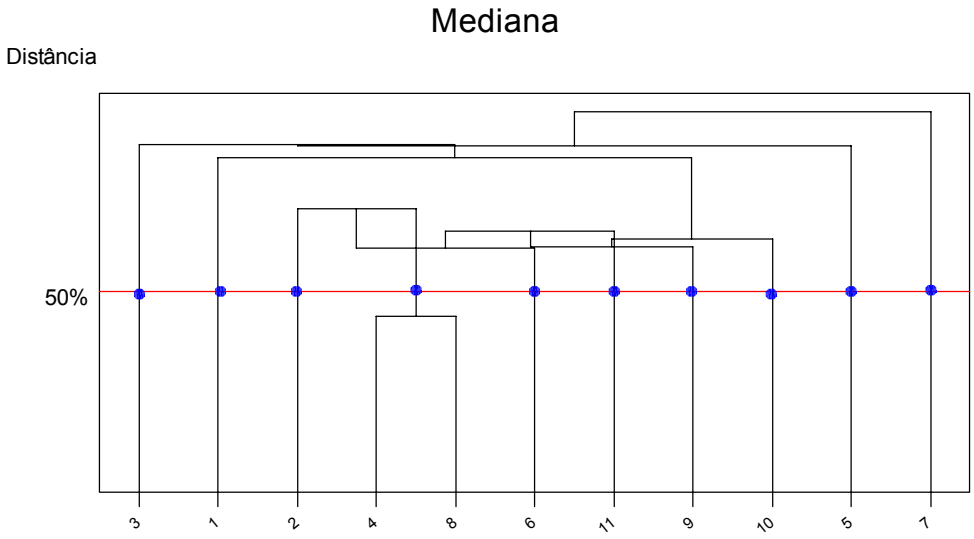

Figura 5 – Dendrograma representando as seqüências das fusões das parcelas, obtido pelo emprego do método mediana, com base na distância Mahalanobis dos dados originais.

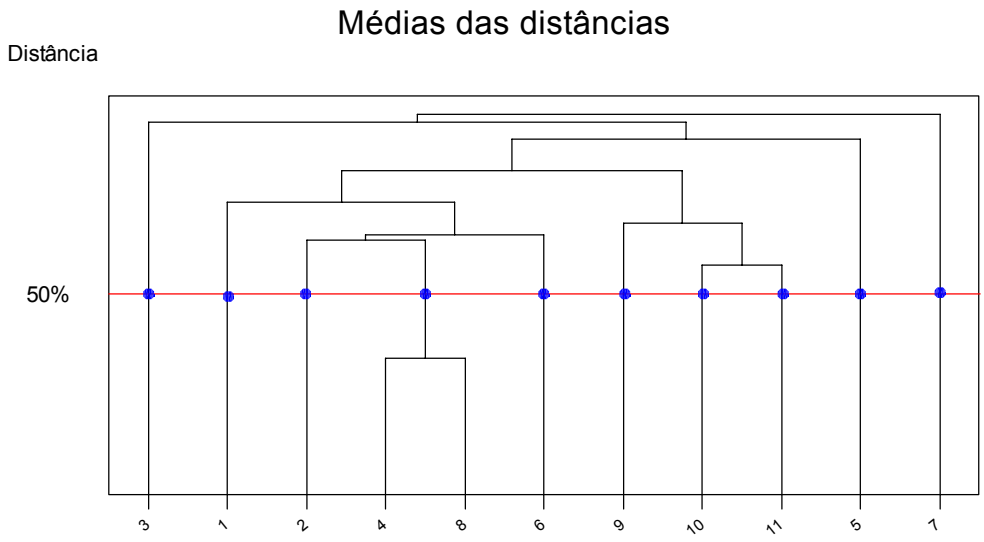

Figura 6 – Dendrograma representando as seqüências das fusões das parcelas, obtida pelo emprego do método das médias da distância, com base na distância Mahalanobis dos dados originais.

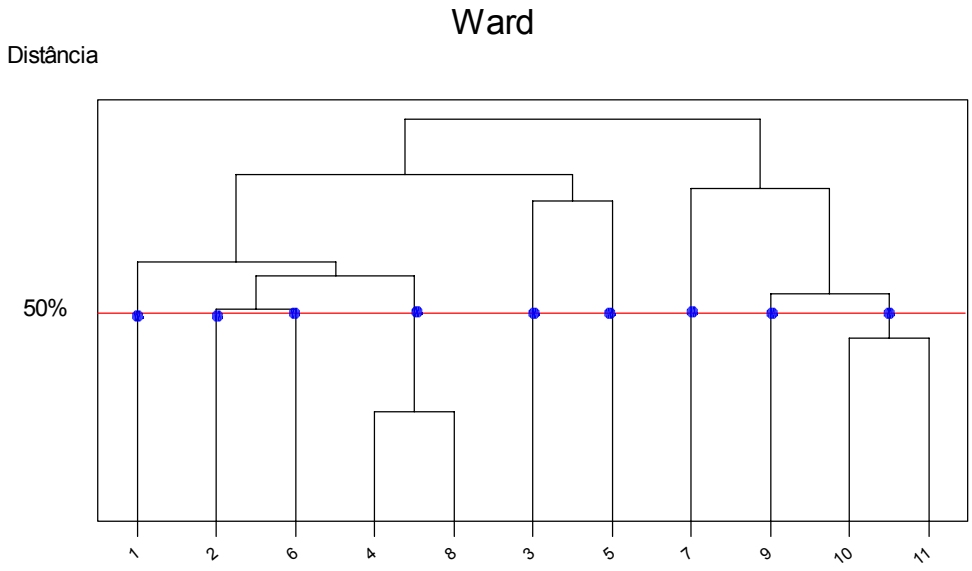

Figura 7 – Dendrograma representando as seqüências das fusões das parcelas, obtidas pelo emprego do método do Ward, com base na distância Mahalanobis dos dados originais.

De forma geral os dendrograma apresentaram estruturas de agrupamentos de objetos homogêneos, embora não exista um critério objetivo para determinar um ponto de corte no dendrograma, ou seja, para determinar quais os grupos foram formados.

Para os dendrogramas obtidos a partir dos métodos ligação simples (Figura 2), ligação completa (Figura 3), centróide (Figura 4), mediana (Figura 5), médias das distâncias ((Figura 6) e Ward ((Figura 7), observa-se que eles foram cortados utilizando-se a medida do percentil, com um corte de 50% da distância Mahalanobis, observando-se que nos métodos ligação simples e Ward, ligação completa, mediana e médias das distâncias, e centróide foram obtidos nove, dez e onze grupos, respectivamente.

Para dados originais no grupo I, verificou-se que os métodos (médias das distâncias e mediana, médias das distâncias e ligação completa, mediana e ligação completa), formaram dez agrupamentos, com semelhança de 100%, e os métodos do grupo II (Medias das distâncias e centróide, centróide e mediana, centróide e ligação completa, médias das distâncias e ligação simples, médias das distâncias e Ward, mediana e ligação simples, mediana e Ward, ligação simples e ligação completa, ligação completa e Ward), com 86% e 84% de semelhança e o III grupo com os métodos (centróide e ligação simples, centróide e Ward, ligação simples e Ward) com 80%, 70% e 67% de semelhança respectivamente (Tabela 4).

Tabela 4.- Porcentagem de grupos coincidentes entre métodos de agrupamento, com base na matriz de Mahalanobis dos dados originais, (nível de significância do teste de independência do χ^2), a partir da tabela de contingência e do grau de associação

Métodos	Médias das distâncias	Centróide	Mediana	Ligação simples	Ligação completa
Centróide	86% (0,58)				
Mediana	100% (0,71)	86% (0,58)			
Ligação simples	84% (0,57)	80% (0,52)	84% (0,59)		
Ligação Completa	100% (0,71)	86% (0,58)	100% (0,71)	84% (0,57)	
Ward	84% (0,57)	70% (0,38)	84% (0,57)	67% (0,61)	84% (0,57)

Entre parentes - nível de significância do teste de independência do χ^2.

Observa-se ainda que os resultados de associação dos métodos foram semelhantes e o nível de significância relativamente alto, sendo possível concluir que, em principio qualquer algoritmo de agrupamento estudado está estabilizado e existem, de fato, grupos entre os indivíduos observados e que existe estabilidade entre os métodos.

Quanto aos resultados da qui-quadrado (Tabela 5) para um nível de significância de 1% e 5% e de um grau de liberdade é igual 3,84 e 6,64 respectivamente, que não deixa dúvida que se pode rejeitar H_0, isto é, concluindo-se com risco de 1% e 5% que os métodos são dependentes, ou estão associados, excluindo-se os métodos do centróide e Ward.

Tabela 5 - Resultados dos dados originais com associação dos métodos obtidos a partir da qui-quadrado

Método	Médias das distâncias	Centróide	Mediana	Ligação simples	Ligação completa
Centróide	10,84				
Mediana	20,00	10,82			
Ligação simples	8,99	7,60	10,42		
Ligação Completa	20,00	10,82	20,00	9,00	
Ward	9,00	3,43	9,00	11,00	9,00

4.2. Análise de agrupamento a partir da matriz de Mahalanobis via “bootstrap”

Com base na matriz de dissimilaridade de Mahalanobis obtida via bootstrap (Tabela 6) foram aplicados os métodos da ligação simples, da ligação completa, do centróide, da mediana, da média das distâncias e de Ward e obtidos os respectivos dendrogramas (Figuras de 8 a 13).

Com base na análise dos dendrogramas formado pelos métodos, verificou-se que, com um corte de 50% nas matrizes de distância, foram formados três grupos tantos para os dados originais como para os dados da reamostragem “bootstrap”.

Tabela 6. Matriz de Mahalanobis obtida via reamostragem “bootstrap” após com 10000 iterações

Parcelas	1	2	3	4	5	6	7	8	9	10	11
1	00,00	20,01	14,40	20,11	16,81	17,66	20,35	17,66	19,50	16,81	20,11
2		00,00	18,51	18,50	18,80	15,59	18,59	15,59	19,03	20,89	18,51
3			00,00	20,80	18,21	20,80	19,10	19,10	20,08	22,27	19,10
4				00,00	12,57	14,54	10,32	16,73	16,73	13,84	7,092
5					00,00	17,55	20,99	19,20	19,20	20,99	17,65
6						00,00	16,21	19,90	19,90	16,21	16,21
7							00,00	22,69	22,69	16,47	16,47
8								00,00	14,05	14,05	18,29
9									00,00	13,20	14,87
10										00,00	11,90
11											00,00

Embora não exista um critério objetivo para determinar um ponto de corte no dendrograma, ou seja, para determinar quais os grupos foram formados, os dendrogramas obtidos a partir dos métodos com a reamostragem bootstrap com 10000 interações, foram cortados utilizando-se a medida percentil, com um corte de 50% na matriz de distância, observando-se que as Figuras 10 e 11, obtiveram onze grupos, e as Figuras 8, 9 e 12, obtiveram dez grupos e a Figura 13, obteve a formação de nove grupos.

Para os dados de reamostragem “bootstrap”, no grupo I, verificou-se que os métodos médias das distâncias e ligação simples, médias das distâncias e ligação completa, com dez grupos, centróide e mediana com onze grupos, com 100% semelhança, grupo II com os métodos médias das distâncias e centróides, médias das distâncias e mediana, centróide e ligação simples, centróide e ligação completa, mediana e ligação simples, mediana e ligação completa com 86% de semelhança e os métodos médias das distâncias e Ward, ligação simples e Ward, ligação completa e Ward, com 84% de semelhança e o III grupo com os métodos ligação simples e ligação completa com 80% de semelhança e centróide e Ward, mediana e Ward com 70% de semelhança (Tabela 7). Estes resultados demonstraram que existe boa possibilidade de estabilidades entre os métodos

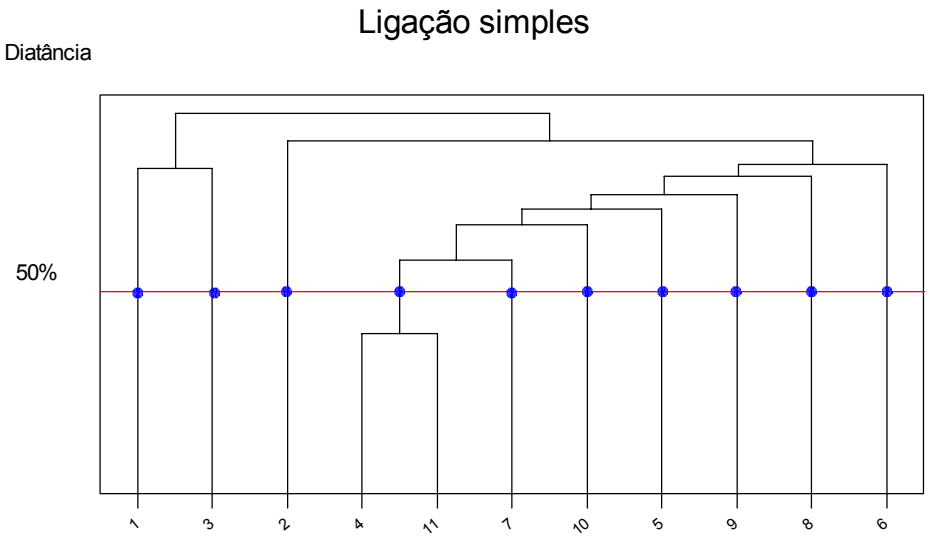

Figura 8 – Dendrograma representando as seqüências das fusões das parcelas, obtida pelo emprego do método ligação simples, com base na matriz de distância de Mahalanobis via bootstrap.

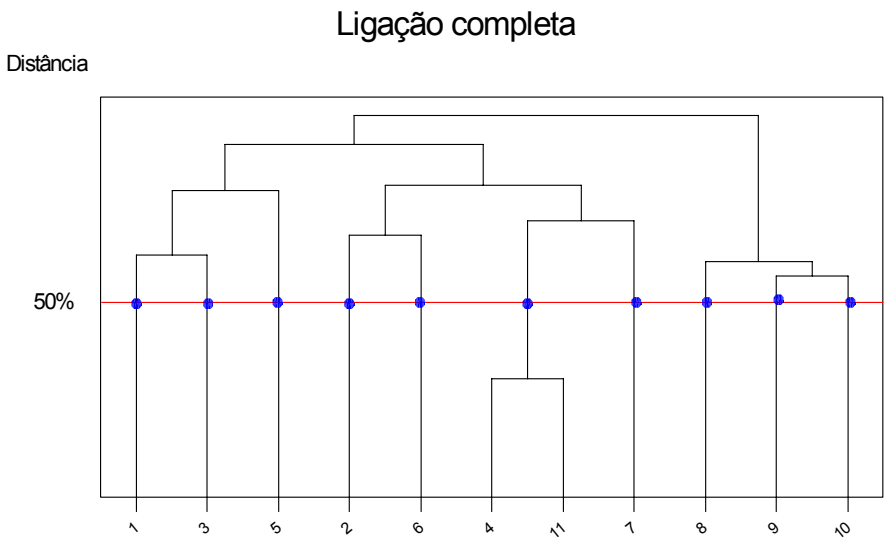

Figura 9 – Dendrograma representando as seqüências de fusão das parcelas, obtida pelo emprego do método ligação completa, com base na matriz de distância de Mahalanobis via bootstrap.

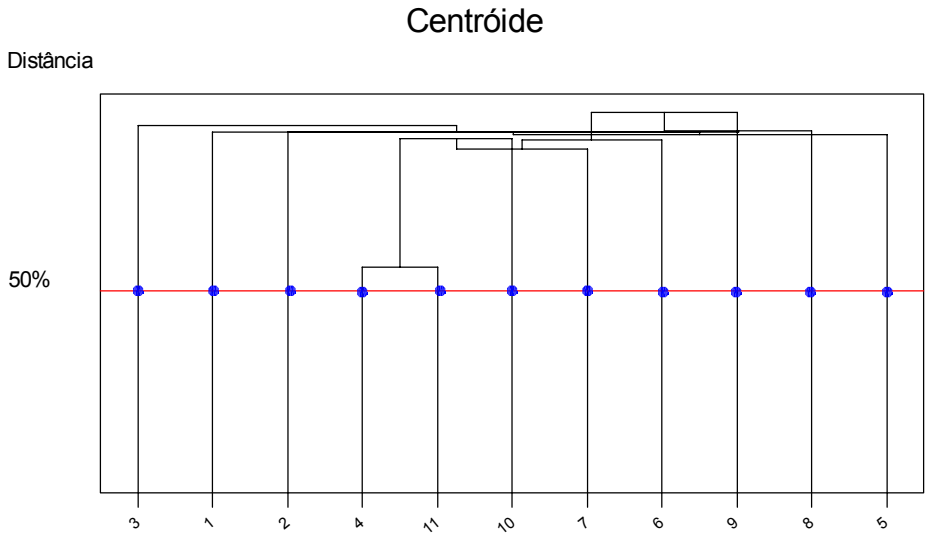

Figura 10 – Dendrograma representando as seqüências de fusão das parcelas, obtido pelo emprego do método do centróide, com base na matriz de distância de Mahalanobis via bootstrap.

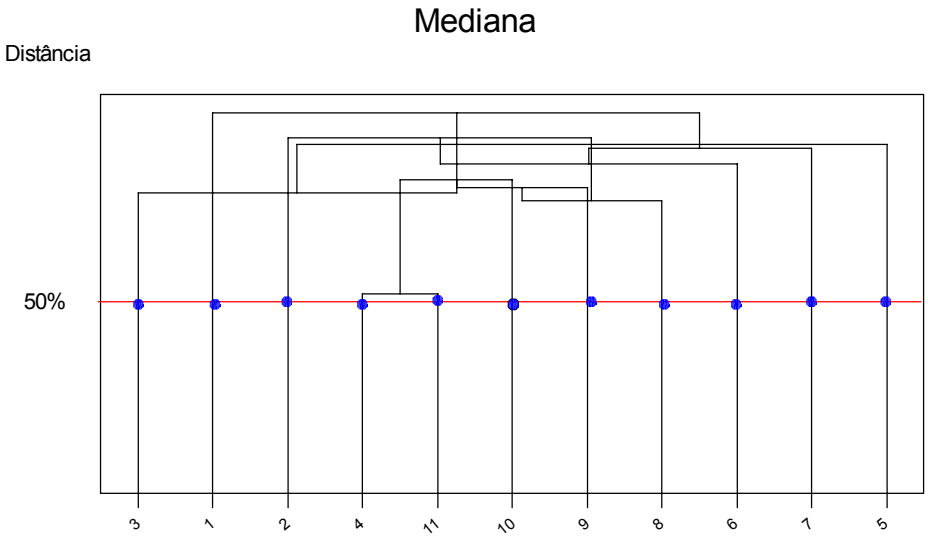

Figura 11 – Dendrograma representando as seqüências de fusões das parcelas, obtido pelo emprego do método da mediana, com base na matriz de distância de Mahalanobis via bootstrap.

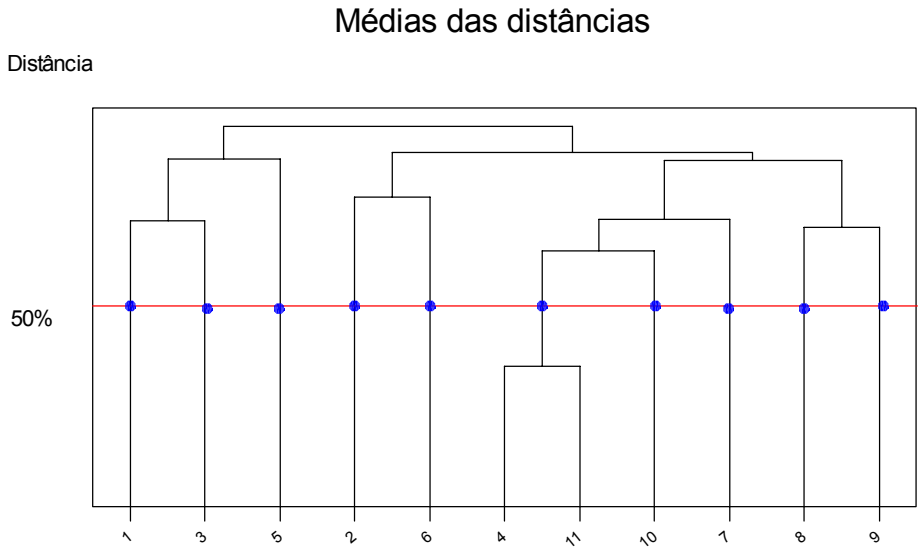

Figura 12 – Dendrograma representando as seqüências das parcelas, obtido pelo emprego do método das médias das distâncias, com base na matriz de distância de Mahalanobis via bootstrap.

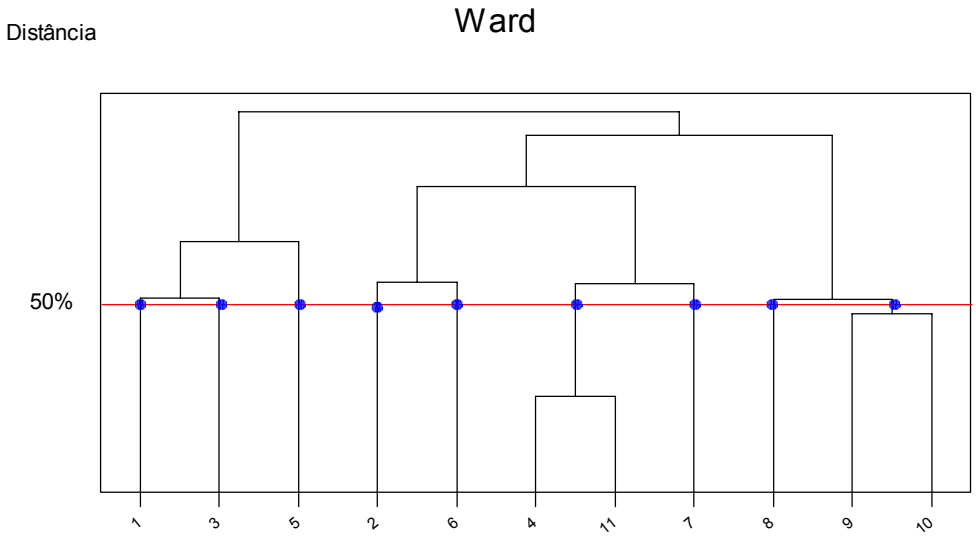

Figura 13 – Dendrograma representando as seqüências das parcelas, obtido pelo emprego do método do Ward, com base na matriz de distância de Mahalanobis via bootstrap.

Tabela 7 - Porcentagem de grupos coincidentes entre métodos de agrupamento, com base na matriz de Mahalanobis via bootstrap, (nível de significância do teste de independência do χ^2), a partir da tabela de contingência e do grau de associação

Métodos	Médias das distâncias	Centróide	Mediana	Ligação simples	Ligação completa
Centróide	86% (0,58)				
Mediana	86% (0,58)	100% (0,71)			
Ligação simples	100% (0,71)	86% (0,58)	86% (0,58)		
Ligação Completa	100% (0,71)	86% (0,58)	86% (0,58)	80% (0,51)	
Ward	84% (0,57)	70% (0,38)	70% (0,38)	67% (0,57)	84% (0,57)

Entre parentes - nível de significância do teste de independência do χ^2.

Verificando que os resultados da qui-quadrado (Tabela 8) para um nível de significância de 1% e 5% e de um grau de liberdade é igual 3,84 e 6,64 respectivamente, que não deixa dúvida que se pode rejeita-se H_0, isto é, concluindo-se com risco de 1% e 5% que os métodos são dependentes, ou estão associados, excluindo-se os métodos do (centróide e Ward) e (mediana Ward).

Tabela 8 - Resultados dos dados de reamostragem bootstrap com 10000 interações com associação dos métodos obtidos a partir da qui-quadrado

Métodos	Médias das distâncias	Centróide	Mediana	Ligação simples	Ligação completa
Centróide	10,82				
Mediana	10,82	22,00			
Ligação simples	20,00	10,82	10,82		
Ligação Completa	20,00	10,82	10,82	7,20	
Ward	9,00	3,48	3,46	9,00	9,00

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

4.3 Correlação cofenética

Os valores das correlações cofenética (Tabela 9) foram todas de magnitude elevada, para os dados originais e bootstrap. Isso mostra que há boa representação das matrizes de dissimilaridade na forma de dendrogramas e que isso independe do método usado.

Tabela 9 - Correlações cofenética entre as matrizes cofenética e a de dissimilaridade obtidas conforme método de agrupamento utilizado

Métodos de Agrupamento	Matriz	
	Original	Bootstrap
Ligação simples	0,99	0,99
Ligação completa	0,98	0,99
Centróide	0,99	0,99
Mediana	0,99	0,99
Média das distâncias	0,99	0,99
Ward	0,99	0,99

4.4 Distorção entre a matriz de dissimilaridade e a matriz cofenética

Tanto para os dados originais como os obtidos via “bootstrap” (Tabela 10), o métodos média das distâncias apresentou distorção nula e Ward “bootstrap” apresentou a maior distorção, corroborando com o observado na análise da correlação cofenética, ou seja, de que há uma boa representação das matrizes de dissimilaridade na forma de dendrograma e que isso independe do método usado e dos dados.

Tabela 10 - Grau de distorção (%) entre as distâncias original e bootstrap e a obtida por meio dos dendrogramas obtidos conforme método de agrupamento utilizado

Métodos de Agrupamento	Matriz	
	Original	Bootstrap
Ligação simples	0,18	0,16
Ligação completa	0,21	0,16
Centróide	0,35	0,36
Mediana	0,26	0,33
Média das distâncias	0,00	0,00
Ward	0,31	0,40

Apesar do presente trabalho não ter como objetivo comparar os métodos de análise, algumas considerações podem ser feitas. Com base em tudo que foi

apresentado, que, de forma geral não se deve utilizar vários métodos de agrupamento e a comparação posterior dos resultados obtidos, pois este procedimento é muito vulgarizado (Reis, 1997).

É interessante notar que os métodos dentro de cada categoria possuem princípios comuns e podem apresentar resultados muito parecidos.

Como já foi dito, existem diferentes distâncias, técnicas e métodos para agrupar indivíduos. O importante é conhecer suas propriedades, qualidade e deficiências, pois irá ajudar à escolha daquele que melhor responde ao objetivo do trabalho.

A principal dificuldade para interpretar os resultados da análise de agrupamento com construção de dendrogramas se deve ao fato de não haver um critério objetivo para identificar os grupos formados. Em diversos trabalhos os pesquisadores têm os seus próprios critérios.

Algoritmos que produzem árvores (dendrograma) são difíceis de analisar na presença de muitos objetos (Bussab et al., 1990), pois os mesmo dificultam sua visualização.

Apesar da versatilidade do modelo “bootstrap”, mais pesquisas devem ser conduzidas visando o pleno entendimento desse fenômeno, tão intrigante nos programas de análise de agrupamento, que é a interação dos dados. Um dos pontos que merecem estudos mais detalhados é a definição dos níveis de estabilidade e dos intervalos de confiança.

Finalmente, é preciso mencionar que a técnica “bootstrap” proposta, exige esforço computacional. Entretanto, pode ser vantajosa sua utilização quando se deseja alta qualidade na informação sobre a estabilidade dos dados em estudo.

5. CONCLUSÕES

A sistemática proposta é promissora para o estudo e a interpretação da estabilidade dos métodos em análise de agrupamento, através de vários algoritmos de agrupamento em dados de vegetação.

Houve correlação entre os métodos de estimação da distância Mahalanobis baseado na associação da tabela de contingência. Pórem, independente do método utilizado, mostrou que há significativa estabilidade entre os métodos.

Consideremos que esses resultados preliminares podem orientar pesquisas futuras no sentido de investigar correlações que podem justificar ou explicar os diferentes agrupamentos encontrados. Podem, ainda, subsidiar estudos posteriores sobre fatores críticos na análise de agrupamento.

6. REFERÊNCIAS BIBLIOGRÁFICAS

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

ALDENDERFER, M. S.; BLASHFELD, R.K. **Cluster analysis.** Beverly Hills; Sage, 1984, 547p.

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

ANDERBERG, M. R. **Cluster analysis for applications.** New York: Acafenic press, 1973, 359p.

ARAÚJO, R. C. C. **Aplicação das técnicas de DEA e Bootstrap para avaliar a eficiência do metrô do Recife**. Recife: UFRPE, 2003. 56f. Dissertação (Mestrado em Biometria) – Universidade Federal Rural de Pernambuco, 2003.

BARROSO, L. P., ARTES, R. **Análise de Multivariada**. Lavras: UFLA, 2003. 157p.

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

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

BUSSAB, W. DE O; MIAZAKI, E. S; ANDRADE, D. **Introdução à análise de agrupamentos**. São Paulo: Associação Brasileira de Estatística, 1990. 105p.

CARLINI-GARCIA, L. A. **Estudo da estrutura genética populacional através de marcadores moleculares**. Piracicaba: ESALQ, 1998. 118f. Monografia (Pós-graduação) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo.

CLIFFORD, H. T.; STEPHENSON, W. **An introduction to numerical taxonomy.** London: Academic Press, 1975. 229p.

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

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. 1994, cap. 6, p. 287-323.

DICICCIO, T. J.; EFRON, B. Bootstrap confidence interval. **Statistical Science**, v, 11, n. 11, p.189-228, 1996.

DONI, M. V. **Análise de Cluster**: métodos hierárquicos e de partição, São Paulo: Mackenzie: 2004. 93f. Monografia (Pós-graduação) – Universidade Presbiteriana Mackenzie, 2004.

DUARTE, M. C.; SANTOS, J. B.; MELO, L. C. Comparison of similarity coefficients based on RAPD markers in the common bean. **Genetics and 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.

EFRON, B. Bootstrap methods: another look at jackknife. **Annals of Statistics**, v. 7, n.1, p.1-26, 1979,

EFRON, B.; TIBSHIRANI, R. J. **An introduction to the bootstrap.** London: Champman & Hall, 1993. 579p.

EVERITT, B. **Cluster analysis**, London: Heinemann Educational Books, 1974. 136p.

EVERITT, B.S. **The analysis of contingency tables***.* 2. ed. London: Chapman & Hall, 1992. 164p.

EVERITT, B. S, LANDAU, S., LEESE, M. **Cluster analysis**. 4º ed. London: Arnold. 2001. 207p.

GAMA M. de P. **Bases da análise de agrupamentos ("Cluster Analysis")**. Brasília: UnB, 1980. 229f. Dissertação (Mestrado em Estatística e Métodos Quantitativos) - Universidade de Brasília, 1980,

GOWER, J. C.; LEGENDRE, P. Metric and 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.

HILLIS. D. M.; MORITZ. C.; MABLE. B. K. **Molecular systematics**. Massachusetts: Sinauer Associates, 1996. 655p.

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

JAIRO, S. F.; GILBERTO A. M. **Curso de Estatística.** São Paulo: Atlas. 1996, 320p.

JOHNSON, R. A.; WICHERN, D. W. **Applied multivariate statistical analysis.** 3 ed. New Jersey: Prantice Hall, 1992. 642p.

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

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

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

LAVARANTI, O. J. **Estabilidade e adatabilidade fenotipica da reamostragem "bootstrap" no modelo AMMI**. Piracicaba: ESALQ, 2003. 166f. Tese (Doutorado em Agronomia) - Escola Superior de Agricultura Luiz de Queiroz, 2003.

MANLY, B. F. J.; **Randomization and Monte Carlo methods in biology**. Cambridge*:* Cambridge University Press, 1997. 215p.

MARDIA, A. K. V.; KENT. J. T.; BIBBY, J.M. **Multivariate analysis.** London: Academic Press, 1997, 518p.

MEYER, A. S.; **Comparação de coeficientes de similaridade usados em análise de agrupamento com dados de marcadores moleculares dominantes.** Piracicaba*:* ESALQ, 2002. 106f. Dissertação (Mestrado em Agronomia) - Escola Superior de Agricultura Luiz de Queiroz, 2002.

MEYER, D. **Árvores evolutivas humanas: uma discussão sobre inferência filogenética**. Ribeirão Preto: Sociedade Brasileira de Genética, 1995. 136p. (Série Monografias, 3)

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

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

REIS, E.; **Estatística multivariada aplicada**. Lisboa: Edições Silabo, 1997. 342p.

RIBOLDI, J. **Análise de agrupamento** "Cluster Analysis", Piracicaba: ESALQ/USP, 1986.49p. (Monografia).

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

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

SOKAL, R. R.; MICHENER, C.D. A statistical method for evaluating systematic relationships. **Bulletin of the Society University of Kansas**, n.38, p.109-1438, 1958.

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

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

WEIR, B. W.; **Genetic data analysis**: Methods for discrete population genetic data. Sunderland: Sinauer, 1990. 445p.