UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

REGINALDO FERREIRA NEVES

Análise de agrupamento aplicado nas distâncias das cidades do agreste paraibano

CAMPINA GRANDE – PB

OUTUBRO/2013

REGINALDO FERREIRA NEVES

Análise de agrupamento aplicado nas distâncias das cidades do agreste paraibano

Trabalho de conclusão de curso apresentado ao curso de Estatística da Universidade Estadual da Paraíba, em cumprimento à exigência para obtenção da conclusão da graduação em bacharel em Estatística.

Orientador: Prof. Dr. Edwirde Luiz Silva

CAMPINA GRANDE – PB

OUTUBRO/2013

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL – UEPB

N518a Neves, Reginaldo Ferreira

Análise de agrupamento aplicado nas distâncias das cidades do Agreste Paraibano [manuscrito] / Reginaldo Ferreira Neves. – 2013.

42 f. : il. color.

Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2013.

"Orientação: Prof. Dr. Edwirde Luiz Silva, Departamento de Estatística".

1. Análise de agrupamento. 2. Distância. 3. Estatística. I. Título.

21. ed. CDD 310

REGINALDO FERREIRA NEVES

Análise de agrupamento aplicado nas distâncias das cidades do agreste paraibano

Trabalho de Conclusão de Curso apresentado ao Curso de Bacharelado em Estatística, do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, em cumprimento às exigências legais para obtenção do título de Bacharel em Estatística.

Aprovado em: 01 / 11 / 2013

Banca Examinadora:

Prof. Dr. Edwirde Luiz Silva
Orientador

Prof. Dr. Tiago Almeida – DE/CCT/UEPB
Examinador (a)

Prof. Me. Silvio Fernando A. X. Jr. – DE/CCT/UEPB
Examinador (b)

DEDICATÓRIA

A minha família, em especial à minha esposa Glaucia e aos meus filhos Pâmela, Jonathan e Paola e a minha mãe Francisca, por sempre me incentivarem, apoiarem e darem força para seguir em busca dos meus ideais.

AGRADECIMENTOS

Agradeço ao Deus Altíssimo criador de céus e terra pelas oportunidades e graça que me permitiu concluir mas um desafio que a vida me proporcionou, a realização deste trabalho, agradeço as pessoas que colaboraram direta e indiretamente.

Por isso, tentarei agradecer a todos envolvidos na elaboração deste trabalho. A minha esposa Glaucia, aos meus filhos Jonathan, Pâmela e Paola a minha querida mãe Francisca e Maria Elita minha sogra, e a minha tia Maria que hoje descansa em Deus. Ao meu orientador Prof. Dr. Edwirde Luiz Silva, 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.

Aos coordenadores e Professores do curso de Estatística, professor Dr. Gustavo Henrique Esteves e sua esposa Dr[a]. Diana Esteves, Dr. Ricardo Alves de Olinda, Prof. Me. Juarez Fernandes de Oliveira, Me. Silvio Fernando, Prof. Dr. João Gil Luna, Dr. Tiago Almeida de Oliveira e sua esposa prof. Dr[a]. Ana Patrícia Bastos Peixoto, pela orientação, pela dedicação e esforço pelo curso. Meu respeito e gratidão.

Especialmente ao professor Prof. Dr. Edwirde Luiz Silva, pelas sugestões na elaboração do Trabalho de conclusão de curso.

Aos colegas e amigos Giovanni Barbosa, Genilson, Ricardo, Saulo, Eder Cabral, Priscila, Michele, Alessandra, por todo apoio nas horas difíceis e também pelos ótimos momentos vivenciados juntos.

A Universidade Estadual da Paraíba, pela oportunidade da realização da graduação.

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

RESUMO

Análise de agrupamento é uma técnica multivariada para formar grupos a partir de característica de similaridade ou dissimilaridade entre os objetos. Para melhor visualização do dendrograma aplicou-se a técnica hierárquica do vizinho mais próximo, em uma amostra aleatória de tamanho 12 do objeto distância entre municípios do agreste paraibano. O banco de dados entre as distância das cidades do agreste paraibano, foi criado através de pesquisados no Google maps, inseridos no software estatístico R 2.15.2. Conforme observado foi gerado uma matriz de similaridade ou distancia a partir da variável distância entre município, logo após uma nova matriz com valores padronizados; em seguida os agrupamentos formados entre as cidades, tomando como base a menor distância entre os municípios do agreste paraibano. A validação dos resultados da análise de agrupamentos da variável distância entre as cidades do agreste paraibano foi efetuado com o coeficiente de fusão e com o coeficiente de correlação cofenético.

Palavras-chaves: Distância, agrupamento e cidades.

ABSTRACT

Cluster analysis is a multivariate technique to form groups from feature similarity or dissimilarity between objects . For better visualization of the dendrograma, we applied the technique of hierarchical nearest neighbor in a random sample of size 12 from the object distance between cities arid Paraíba . The database distance between cities arid Paraíba , was created through researched on Google maps , inserted into the statistical software R 2.15.2. As noted was generated a matrix of similarity or distance from the variable distance between the municipality, after a new array with default values , then the groups formed between the cities , based on the shortest distance between municipalities arid Paraíba . The validation of the results of cluster analysis of the variable distance between cities arid Paraíba was performed with the fusion coefficient and the correlation coefficient cofenético

Keywords: Distance, grouping and cities.

SUMÁRIO

INTRODUÇÃO

A análise de agrupamento designa uma série de procedimentos estatísticos sofisticados que podem ser usados para classificar objetos e pessoas, sem preconceitos, isto é, observando apenas as semelhanças ou dessemelhanças entre elas, sem definir previamente critérios de inclusão em qualquer agrupamento. Mais concretamente, os métodos de análise de agrupamento são procedimento de estatística multivariada que tentam organizar um conjunto de indivíduos, para os quais é conhecida informação detalhada, em grupos relativamente homogêneos (clusters) (REIS, 2001).

A contribuição mais expressiva para a aplicação da análise de agrupamento foi dada por Sokal e Sneath em 1963 com o seu livro *Principies ot Numerical Taxanamy*. Sokal e Sneath mostraram que um método eficiente para se proceder à classificação biológica, seria juntar toda a informação existente sobre um conjunto de organismos, determinar a semelhança existente entre esses organismos e através de um método de análise de agrupamento colocarem organismos relativamente semelhantes num mesmo grupo. Uma vez agrupados os organismos, as características de cada grupo seriam analisadas de modo a determinar se tratava ou não de espécies diferentes. O número de publicações sobre o assunto multiplicou-se depois deste livro e podem apontar-se duas razões para isto ter acontecido; o desenvolvimento de computadores com elevado poder de cálculo; a importância da classificação como método científico (REIS, 2001).

A classificação de elementos é a base de compreensão de ciências como a química inorgânica e a teoria atômica da matéria e, ao mesmo tempo, a classificação de doenças fornece uma base estrutural em campos de estudo como a medicina. Nas ciências sociais, os métodos de análise de clusters foram utilizados pelos antropólogos para definirem áreas culturais homogêneas (DRIVER, 1965) e (JOHNSON, 1967), pelos psicólogos e pelos estudiosos da ciência política e da economia e ainda pelos geógrafos. Em marketing, a análise de clusters tem sido aplicada para proceder à segmentação de mercados a partir das características geográficas, demográficas e psicográficas dos consumidores, para identificar mercados potenciais para determinados produtos, determinar mercados idênticos em países diferentes ou encontrar grupos de consumidores que possam servir de referência na previsão de vendas.

Neste trabalho de conclusão de curso (TCC) aplicou-se conceitos e técnicas científica de análise de agrupamento da disciplina Multivariada, conhecimento

adquiridos no decorrer do curso de bacharel em Estatística na Universidade Estadual da Paraíba. Os procedimentos de análise de agrupamento com o método do vizinho mais próximo foram aplicados em uma amostra de 12 cidades do agreste paraibano, para tornar a análise gráfica e interpretação dos resultados mais simples para qualquer leitor, nas 66 cidades do que compõem o agreste paraibano também foram efetuado a análise de agrupamento, com a finalidade de formar grupos de cidades que possuam a mínima distância entre si, a análise de agrupamento utilizou-se em logística para criar grupos de cidades que apresente a menor distancias entre si, para distribuir recursos e matérias, com a finalidade de otimizar os atendimento técnico no agreste paraibano em diversos clientes.

A partir de levantamento no google maps, construiu-se um banco de dados das distâncias entre as 66 cidades do agreste paraibano. Desenvolveu-se algoritmos no software estatístico R, que realizou a análise de agrupamento com o método hierárquico do vizinho mais próximo entre as distancias das cidades, que compõe o agreste paraibano. O algoritmo gera a matrizes distância, e toma como base para medir a dessemelhanças entre as cidades a distância euclidiana, também forneceu informações dos agrupamentos através de um dendrograma, utilizou-se o coeficiente de correlação cofenético (ccc) para validar os resultados dos dados da amostra e população, o ponto ótimo dos agrupamentos da amostra foi obtido pela análise gráfica do coeficiente de fusão.

2 REVISÃO BIBLIOGRÁFICA

2.1 Análise de Agrupamento

De modo sintético, o método pode ser descrito como se segue: dado um conjunto de *n* indivíduos para os quais existe informação sobre a forma de *p* variáveis, o método de análise de clusters 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 semelhantes aos elementos do mesmo grupo do que a elementos dos restantes grupos. Este método é também chamado de método de partição, classificação ou taxonomia, embora o termo partição seja mais utilizado para uma das técnicas específicas da análise: aquela em que os indivíduos são divididos por um número preestabelecido de grupos (REIS, 2001).

Na análise de agrupamento, é fundamental ter particular cuidado na seleção das variáveis de partida que vão caracterizar cada indivíduo ou caso, 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 configuram-se por si mesmo sem necessidade de ser definida uma relação causal entre as variáveis utilizadas. A análise de clusters que aqui se apresenta não faz uso de modelos aleatórios, mas é útil por fornecer um sumário bem justificado de um conjunto de dados. Os métodos são exploratórios e a ideia é sobre tudo gerar hipóteses, mais do que testa-as, pelo que é necessária a validação posterior dos resultados encontrados através da aplicação de outros métodos estatísticos.

Uma dificuldade inicial é a de não existir uma única via de definição de grupos, isto é, um único critério de partição e/ou agrupamento dos indivíduos ou casos com base numa única medida de dessemelhanças. Em todos eles se pretende que os grupos sejam coerentes e que se distingam de maneira significativa uns dos outros genericamente, a análise de agrupamento compreende cinco etapas; a seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados; a definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos; a definição de uma medida de semelhança ou distância entre cada dois indivíduos; a escolha de um critério de agregação ou desagregação dos indivíduos, isto é, a definição

de um algoritmo de partição / classificação; por último, a validação dos resultados encontrados.

Discutem-se nos pontos seguintes os aspectos fundamentais de cada uma destas etapas (à exceção da primeira que, naturalmente, não diz respeito apenas ao método, dependendo de cada problema concreto de classificação) (REIS, 2001).

2.2 Métodos de análise de agrupamento

Análise de agrupamento é uma das técnicas de análise multivariada cujo propósito primário é reunir objetos, baseando-se nas características dos mesmos. Ela classifica objetos segundo aquilo que cada elemento tem de similar em relação a outro pertencentes a determinado grupo, considerando, é claro, um critério de seleção predeterminado. O grupo resultante dessa classificação deve então exibir um alto grau de homogeneidade interna (*within-cluster*) e alta heterogeneidade externa (*between-cluster*) (CORRAR, 2007).

Na aplicação do método, é necessário identificar a técnica de análise mais apropriada. É possível dividir as técnicas disponíveis em vários grupos (Hierárquica, Optimização, Densidade, outras), no entanto, fogem do objetivo deste trabalho as demais técnicas, logo, serão consideradas as técnicas hierárquicas.

Métodos hierárquicos

Estes métodos 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, para cada par de partições P_j e P_{j+1}, cada grupo da partição P_{j+1} estar sempre incluído num grupo da partição P_j.

Este tipo de técnica baseia-se na construção de uma matriz de semelhanças ou diferenças em que cada elemento da matriz descreve o grau de semelhança ou diferença entre cada dois casos com base nas variáveis escolhidas. Os métodos hierárquicos dividem-se em aglomerativos e divisivos**.** Nos primeiros, parte-se, de n grupos de apenas um indivíduo cada, que vão sendo agrupados sucessivamente até se encontrar

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 obtém-se n grupos de 1 elemento cada.

Os métodos de análise de clusters mais divulgados e mais utilizados são os hierárquicos aglomerativos, e isto porque os métodos divisivos, tal como os de optimização, são extremamente pesados em termos de capacidade informática.

O ponto de partida comum a todos os métodos hierárquicos é a construção de uma **matriz de semelhanças** ou de **distâncias**, sendo este o terceiro problema a resolver em qualquer análise de clusters (REIS, 2001).

2.3 Definição de medidas de semelhança/distância

Segundo Tversky (1977), a análise teórica das relações de semelhança tem sido dominada pelos modelos geométricos. Estes modelos representam os objetos como pontos num qualquer espaço de coordenadas de forma que as dissemelhanças observadas entre objetos correspondam a distâncias métricas entre os respectivos pontos. Os métodos de classificação exigem que os índices de semelhança respeitem as propriedades das métricas, dados por:

1. Simetria: dados dois objetos, x e y, a distância entre eles verifica a propriedade
$$d(x,y) = d(y,x) \geq 0$$
2. Desigualdade triangular: dados três objetos, x, y e z, as distâncias entre eles satisfazem a propriedade:
$$d(x,y) \leq d(x,z) + d(z,y)$$
3. Diferenciabilidade de não idênticos: dados dois objetos, x e y,
$$d(x, y) \neq 0 \Rightarrow x \neq y$$
4. Indiferenciabilidade de idênticos: dados dois objetos idênticos, x e y
$$d(x, y) = 0$$

Os índices de dessemelhanças mais comumente utilizados podem ser classificados em quatro categorias (Aldenderfer e Blashfield, 1985); coeficientes de correlação; medidas de distância; coeficientes de associação; medidas de semelhança probabilística.

Todas estas medidas tem vantagens e desvantagens, mas os mais utilizados nas ciências sociais e demais áreas são os dois primeiros tipos mencionados.

2.4 Coeficientes de Correlação

Estes coeficientes caracterizados por serem de fácil interpretação geométrica, são das medidas de semelhança mais utilizadas nas ciências sociais, em particular o coeficiente de correlação de Pearson, assim definido para dois indivíduos i e j, caracterizados por um conjunto de p atributos:

$$r_{ij} = \frac{\sum_{v=1}^{p}(x_{iv} - \bar{x}_i) + \sum_{v=1}^{p}(x_{Jv} - \bar{x}_J)}{\sqrt{\sum_{v=1}^{p}(x_{iv} - \bar{x}_i)^2 + \sum_{v=1}^{p}(x_{jv} - \bar{x}_j)^2}} \qquad (1)$$

Sendo

x_{iv} = valor da variável **v** para o indivíduo **i**, (v = 1, ..., p)

x_{jv} = valor da variável **v** para o indivíduo **j**

$\bar{x}_i$= média de todas as variáveis para o indivíduo i

$\bar{x}_j$ = média de todas as variáveis para o indivíduo j

p = número total de variáveis.

O valor do coeficiente varia entre - 1 e + 1, com o valor zero significando não existir tendência de correlação linear entre os indivíduos. Este coeficiente é particularmente insensível às diferenças de escala das variáveis, uma vez que o cálculo da média de todas as variáveis para cada indivíduo impõe a padronização prévia dessas variáveis. No entanto, é sensível às diferenças de forma de cada indivíduo e à dispersão dos valores das variáveis em torno das respectivas médias. Segundo (REIS, 2001), outra desvantagem do coeficiente de correlação reside no fato de uma média de valores de diferentes variáveis não ter um significado claro e daí, calcular correlações em algumas situações, pode não ter qualquer significado estatístico.

Além de tudo isto, este coeficiente não satisfaz a propriedade de desigualdade triangular das métricas. No entanto, a seguinte transformação do coeficiente pode dar lugar a uma métrica:

$$d_{ij} = [\ 0{,}5\ (1 - r_{ij})]^{1/2} \qquad (2)$$

Resultando $d_{ij} = 0$ para $r_{ij} = + 1$ e $d_{ij} = 1$ para $r_{ij} = - 1$.

Apesar destas desvantagens, o coeficiente de correlação tem sido utilizado com sucesso, precisamente quando se pretende que os resultados da classificação não sejam afetados por diferenças de dispersão e de escala das variáveis (HAMER e CUNNINGHAM, 1981).

2.5 Medidas de distância

Existem várias medidas que podem ser utilizadas como medidas de distância ou dessemelhança entre os elementos de uma matriz de dados. Cormack (1971) descreve uma série de medidas possíveis, de entre as quais, se podem destacar como mais utilizadas:

1. Distância Euclidiana: a distância entre dois individuos (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).

$$d_{ij} = \sqrt{\sum_{v=1}^{p} \left(x_{iv} - \overline{x}_{jv}\right)^2} \qquad (3)$$

2. Quadrado da Distância Euclidiana: a distância entre dois casos (i e j) para todas as variáveis (v = 1, 2, ..., p).

$$d_{ij}2 = \sum_{v=1}^{p} \left(x_{iv} - \overline{x}_{jv}\right)^2 \qquad (4)$$

3. Distância absoluta ou City - Block Metric: a distância entre dois elementos (i e j) é a soma dos valores absolutos das diferenças entre os valores das variáveis (v = 1, 2,., p) para aqueles dois casos:

$$d_{ij} = \sum_{v=1}^{p} \left|x_{iv} - x_{jv}\right| \qquad (5)$$

4. *Distância de Minkowskf:* Definida a partir da medida anterior, pode ser considerada como a generalização da distância Euclidiana (as duas coincidem quando r = 2):

$$d_{ij} = \left[\sum_{v=1}^{p} \left| x_{iv} - x_{jv} \right|^{r} \right]^{\frac{1}{r}} \qquad (6)$$

5. *Distância de Mahalanobis:* também chamada distância generalizada. Esta medida, ao contrário das apresentadas anteriormente, considera o inverso da matriz de covariância Σ^{-1}: para o cálculo das distâncias:

$$d_{ij} = \left(x_i - \overline{x}_j\right)' \ \Sigma^{-1} \left(x_i - \overline{x}_j \right) \qquad (7)$$

sendo x_i e x_j, respectivamente, os vetores de valores das das *variáveis* para os indivíduos *i* e *j*, $\boldsymbol{\Sigma^{-1}}$ é a inversa da matriz de covariância.

6. *Distância de Chebishev:* a distância entre dois indivíduos *i* e *j* é o *valor* máximo para todas as variáveis, das diferenças entre esses dois indivíduos.

$$d_{ij} = \max_{v} \left| x_{iv} - x_{jv} \right| \qquad (8)$$

A cada passo do processo *aglomerativo,* a matriz de semelhanças / distâncias é recalculada de modo, a saber-se qual a relação entre os grupos já formados e os elementos ainda não agrupados. De acordo com Johnson (1967), é nesta altura, quando se calcula a relação entre os grupos já formados e os casos restantes, que os métodos aglomerativo apresentam diferenças entre si. Mais precisamente, neste momento do processo, deverá ser satisfeita a seguinte fórmula de recorrência:

$$d_{k(i,j)=} \alpha_i . d_{ki} + \alpha_j . d_{kj} + \beta . d_{ij} + \gamma \left| d_{ki} - d_{kj} \right| \qquad (9)$$

que $d_k(_{i,\,j})$ é a distância entre o grupo k e o grupo $(_{i,\,j})$ formado pela fusão dos grupos (ou elementos) i e j. Embora a fórmula de recorrência seja sempre a mesma, os coeficientes α_i, α_j, β e diferem conforme o método aglomerativo escolhido.

Os valores dos parâmetros da fórmula de recorrência para os métodos de agregação são os seguintes para ligação simples (*Single linkage*), $\alpha_i = ½$; $\alpha_j = ½$; $\beta=0$; $y = -1/2$.

Apesar da sua importância, quer a distância Euclidiana, quer outras medidas de distância, tem vários problemas de utilização, sendo o mais importante o efeito que as diferenças de escala das variáveis provocam sobre o valor das distâncias. As variáveis que apresentam variações e unidades de medida elevadas, facilmente anularão o efeito das outras variáveis. Para resolver este problema, como já se referiu anteriormente é comum a prática de padronização das variáveis, de modo a tornar a sua média nula e o seu desvio-padrão unitário.

2.6 Medidas de semelhança probabilística.

A diferença entre este tipo de medidas de semelhança e todos os outros anteriormente apresentados reside no fato de não se calcular propriamente um valor para a semelhança entre os indivíduos. Para se formarem clusters avalia-se o ganho probabilístico da informação, a partir das variáveis iniciais, e agrupam-se os dois indivíduos que menos ganho de informação provoquem.

2.7 Critérios de agregação e desagregação dos casos

Escolhida uma medida de distância, surge o quarto problema a resolver em qualquer análise de clusters: a escolha do critério de (des)agregação dos indivíduos. Poder-se-á dizer que os vários métodos pretendem responder, de forma diferente, às seguintes questões; distância entre indivíduos do mesmo grupo e distância entre indivíduos de grupos diferentes; dispersão dos indivíduos dentro do grupo; densidade dos indivíduos dentro e fora dos grupos.

Os vários métodos de agregação dos indivíduos diferem no modo como estimam distâncias entre grupos já formados e outros grupos ou indivíduos por agrupar. O processo de agrupamento de indivíduos já agrupados depende da distância entre os grupos. Portanto, diferentes definições destas distâncias poderão resultar em diferentes soluções finais. (REIS, 2001)

Segundo (REIS, 2001), não existe aquilo a que se possa chamar o melhor critério de (des)agregação dos casos em análise de clusters. É prática comum utilizar vários critérios e fazer a comparação dos resultados. Se estes forem semelhantes, é possível concluir que se obtiveram resultados com elevado grau de estabilidade e, portanto, confiáveis. Os critérios de agregação mais utilizados são os seguintes:

1. Single linkage ou critério do vizinho mais próximo: Este método tem sido amplamente utilizado em diversas áreas. Aponta como desvantagens a incapacidade de não discernir grupos pobremente separados (Johnson e Wichern, 1988). Este critério define como semelhança entre dois grupos a semelhança máxima entre quaisquer dois casos pertencentes a esses grupos, ou dito de outro modo, dados dois grupos (i, j) e (k), a distância entre os dois é a menor das distâncias entre os elementos dos dois grupos:

$$d_{(i,j)k} = \min \{ d_{ik}; d_{jk} \} \qquad (10)$$

Ou seja, $d_{(i,j)k}$ é dada pelo menor elemento do conjunto das distâncias dos pares de indivíduos (i e k) e (j e k). A distância entre dois grupos é dada por:

$$d_{(ij)(kl)} = \min \{ d_{ik}, d_{il}, d_{jk}, d_{jl} \}$$

logo, a distância entre dois grupos formados, respectivamente, pelos indivíduos (i e j) e (k e l) é dada pelo menor elemento do conjunto, cujos elementos são as distâncias entre os pares de indivíduos (i e k), (i e l), (j e k) e (j e l). Na figura 1 ilustra o método do vizinho mais próximo.

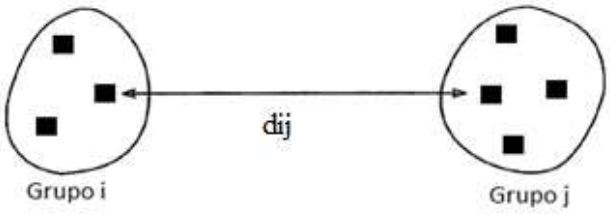

Figura 1 - Método do *single* linkage ou Vizinho mais próximo (REIS, 2001)

Este método torna-se, assim, um sistema contractor do espaço uma vez que cada caso terá mais tendência para se agrupar a um grupo já definido do que para formar o núcleo de um novo grupo (LANCE e WILLIAMS, 1967).

Esta característica torna-se numa desvantagem do método: a aptidão para que os agrupamentos finais se assemelhem a cadeias de elementos quando representados num espaço multidimensional, desvantagem esta que tem relegado para segundo plano a utilização do método de single linkage como método preferencial de agregação (CORMACK, 1971; LANCE e WILLIAMS, 1967; SNEAATH e SOKAL, 1973).
A maior vantagem deste método é ser insensível a transformações monótonas da matriz de distâncias e ainda por não ser afetado pela existência de relações nos dados iniciais.

2.8 Coeficiente de Correlação Cofenético (CCC)

Sokal e Rohlf (1962), definiram o coeficiente de correlação "cofenética" (CCC) que ainda hoje é a medida de validação mais utilizada pelos taxonomistas numéricos. Esta medida dá-nos a relação entre cada valor da matriz de semelhanças e um valor obtido a partir do dendrograma, significando, em última instância, à medida que o dendrograma resultante da aplicação de um método hierárquico, representa os valores da matriz de semelhanças / distâncias. Mais precisamente, a correlação cofenética é a correlação entre os elementos da matriz de distância (ou semelhanças) e os correspondentes coeficientes de fusão, ou seja, as distâncias (ou semelhanças) a que os indivíduos se juntam pela primeira vez para formar grupos. Embora este método de validação seja apropriado sobre tudo quando se utiliza um método hierárquico aglomerativo, foi criticado por Farris (1969) que referiu a sua sensibilidade ao tamanho dos grupos como razão suficiente para não ser aceite como justificação direta e final da técnica utilizada.

Com o objetivo de ilustrar esta técnica considere seis grandes empresas comerciais, para as quais foram medidas as seguintes dimensões estratégias (Reis, 2001):

X_1: número de lojas

X_2: dimensão média das lojas (m^2)

X_3: % da área alimentar no total das vendas

X_4: montante da área alimentar no total de caixa (10^3 contos).

Os dados obtidos foram os seguintes:

Quadro 1 – Caracterização de seis empresas comerciais

Empresa	Nº de lojas	Dimensão média lojas	% alimentar nas vendas	Área alimentar no total de caixa
1. Modelo	42	830	95	12.000
2. Pingo Doce	33	1215	92	11.500
3. Feira Nova	1	6000	70	18.500
4. Supa/Jumbo	6	5675	63	21.400
5. Minipreço	31	288	98	870
6. Continente	4	885	65	23.100

Para obtenção dos grupos estratégicos deverá ser utilizada a distância Euclidiana ao quadrado e o critério de agregação do vizinho mais próximo *(single linkage).*

Por estarmos em presença de variáveis com escalas de medida e dispersões muito diferentes, optou-se por uma padronização prévia das variáveis e, a partir daí, calcular a matriz de distâncias Euclidianas ao quadrado:

$$\mathbf{D} = \begin{matrix} 0,0 & & & & & \\ 0,3 & 0,0 & & & & \\ 12,2 & 9,2 & 0,0 & & & \\ 12,7 & 9,9 & 0,4 & 0,0 & & \\ 2,3 & 1,9 & 15,2 & 17,1 & 0,0 & \\ 9,8 & 7,4 & 4,3 & 3,4 & 13,8 & 0,0 \end{matrix}$$

O primeiro passo no processo de agrupamento consiste em encontrar a menor das distâncias entre cada par de empresas. Assim, o primeiro grupo será formado pelas empresas 1 e 2, ou seja, Modelo e Pingo Doce. Torna-se necessário recalcular as distâncias entre este grupo e as restantes empresas utilizando o critério do vizinho mais próximo que define a distância entre dois grupos como a menor das distâncias entre os seus elementos. Por exemplo, a distância entre o grupo (1, 2) e a empresa 3 será.

$$d_{(1,\,2)\,3} = \min \{d_{(1,3)} ; d_{(2,3)}\} = \min \{12{,}2 ; 9{,}2\} = 9{,}2$$

A nova matriz de distâncias passará então a ser.

$$\mathbf{D}_1 = \begin{array}{c} \begin{matrix} 1{,}2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{bmatrix} 0 & & & & \\ 9{,}2 & 0 & & & \\ 9{,}9 & 0{,}4 & 0 & & \\ 1{,}9 & 15{,}2 & 17{,}1 & 0 & \\ 7{,}4 & 4{,}3 & 3{,}4 & 13{,}8 & 0 \end{bmatrix} \end{array} \begin{matrix} 1{,}2 \\ 3 \\ 4 \\ 5 \\ 6 \end{matrix}$$

As empresas mais semelhantes são agora a 3 (Feira Nova) e a 4 (Supra/Jumbo) que passam a formar um grupo à distância de 0,4. Para o passo seguinte, a matriz de distância é:

$$\mathbf{D}_2 = \begin{array}{c} \begin{matrix} 1{,}2 & 3{,}4 & 5 & 6 \end{matrix} \\ \begin{bmatrix} 0 & & & \\ 9{,}2 & 0 & & \\ 1{,}9 & 15{,}2 & 0 & \\ 7{,}4 & 3{,}4 & 13{,}8 & 0 \end{bmatrix} \end{array} \begin{matrix} 1{,}2 \\ 3{,}4 \\ 5 \\ 6 \end{matrix}$$

À distância de 1,9 a empresa 5 (Minipreço) junta-se ao grupo já formado no primeiro passo e a matriz de distância, depois deste passo, passa a ser;

$$\mathbf{D}_3 = \begin{array}{c} \begin{matrix} 1,\,2,\,5 & 3{,}4 & 6 \end{matrix} \\ \begin{bmatrix} 0 & & \\ 9{,}2 & 0 & \\ 7{,}4 & 3{,}4 & 0 \end{bmatrix} \end{array} \begin{matrix} 1,\,2,\,5 \\ 3{,}4 \\ 6 \end{matrix}$$

A empresa 6 (Continente) se juntar ao grupo formado pela Feira Nova e Supra/Jumbo à distância 3,4. Por fim os dois grupos (1,2,5) e (3, 4, 6) vão juntar-se à distância de 7,4.

Todo este processo de agrupamento pode ser resumido num quadro do seguinte tipo:

$$\mathbf{D_4} = \begin{bmatrix} 0 & & 0 \\ 7{,}4 & & \end{bmatrix} \begin{matrix} 1, 2, 5 \\ 3, 4, 6 \end{matrix}$$

Quadro 2 - Processo de agrupamento das seis empresas segundo o critério do *single Iinkage*

PASSO	***DISTÂNCIAS***	***GRUPOS***
1	$d_{12} = 0{,}3$	(1, 2) (3) (4) (5) (6)
2	$d_{34} = 0{,}4$	(1, 2) (3, 4) (5) (6)
3	$d_{(1,2)5} = 1{,}9$	(1, 2, 5) (3, 4) (6)
4	$d_{(3,4)\,6} = 3{,}4$	(1, 2, 5) (3, 4, 6)
5	$d_{(1,2,5)\,(3,4,6)} = 7{,}4$	(1, 2, 3, 4, 5, 6)

Uma mais rápida e fácil visualização do processo de agrupamento é possível através de uma representação gráfica denominada de DENDROGRAMA:

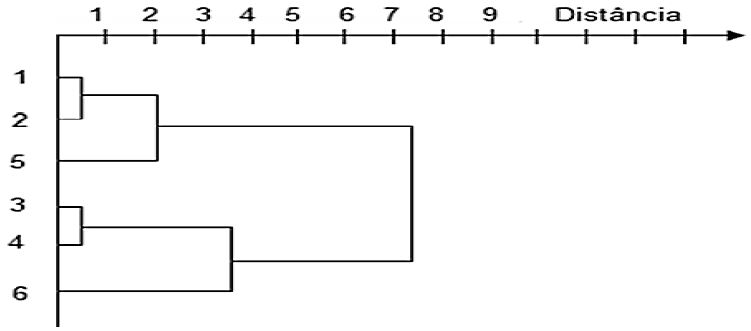

Figura 2 - Dendrograma obtido pelo método do vizinho mais próximo, a partir das medidas de dissimilaridade entre as 6 empresas.

A observação do dendrograma sugere a existência de dois grupos: um formado pelas empresas Modelo, Pingo Doce e Minipreço, e outro pelas empresas Feira Nova, Supa/Jumbo e Continente.

2. *Complete Iinkage* ou *critério do vizinho mais afastado:* utiliza o procedimento *inverso* ao anterior, uma *vez* que a distância entre dois grupos é agora definida como sendo a distância entre os seus elementos mais afastados ou menos semelhantes.

Dados dois grupos (I, J) e (k), a distância entre eles é a maior das distâncias entre os seus elementos:

$$d_{(I\,j)\,k} = \max \{ d_{lk}; d_{jk} \} \qquad (10)$$

De acordo com esta estratégia cada grupo passa a ser definido como um conjunto de elementos em que cada um é mais semelhante a todos os restantes elementos do grupo do que a qualquer dos elementos dos restantes grupos.

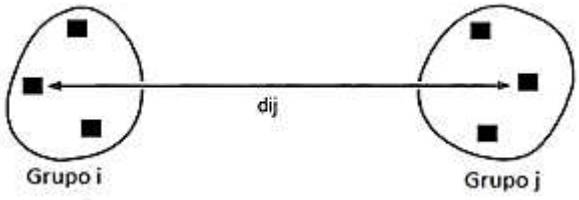

Figura 3 - Método do C*omplete linkage* ou *critério do vizinho mais afastado* (REIS, 2001)

Este método tem tendência para encontrar clusters compactos compostos de indivíduos muito semelhantes entre si. Embora os resultados da aplicação deste método deem uma visão nítida dos diferentes grupos encontrados, nem sempre apresentam um elevado grau de concordância com a estrutura inicial dos dados. Existem outros critério de agregação de casos como por ex: critério da média dos grupos, critério Ward.

2.9 Validação dos resultados obtidos

Uma vez que a análise de agrupamento tem como objetivo criar grupos homogêneos, surge um problema que é o da escolha do número adequado de agrupamentos ou grupos. A aplicação de métodos hierárquicos permite a apresentação dos resultados sob a forma de dendrograma ou de uma árvore de agrupamento. O dendrograma mostra todas as fases do processo de agrupamento desde a separação total dos indivíduos até à sua inclusão num grupo apenas.

O problema que põe é por onde cortar o dendrograma de modo a obter-se um número de grupos ótimo. Infelizmente, este passo fundamental da análise de clusters, não está ainda completamente resolvido, sendo motivo de estudos ainda.

Na figura 4, o corte do dendrograma a uma distância de aproximadamente 3 revela a existência de dois grupos: (2, 5, 3, 4) e (1, 6, 7)

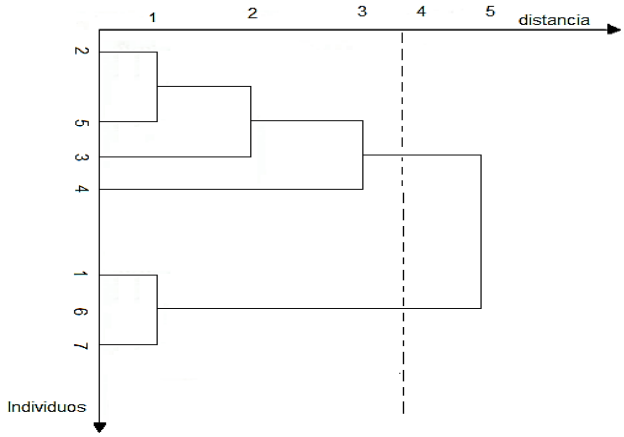

Figura 4 – Dendrograma com corte em aproximadamente na distância 3 formando dois grupos (2,5,3,4) e (1,6,7) (Fonte: Reis, 2001)

A árvore de agrupamento permite-nos também ter uma visualização, ao longo do processo de agrupamento, de quais os grupos que se vão subdividindo e do correspondente número de indivíduos. Na figura 5 apresenta-se uma árvore de agrupamento de objetos hipotéticos.

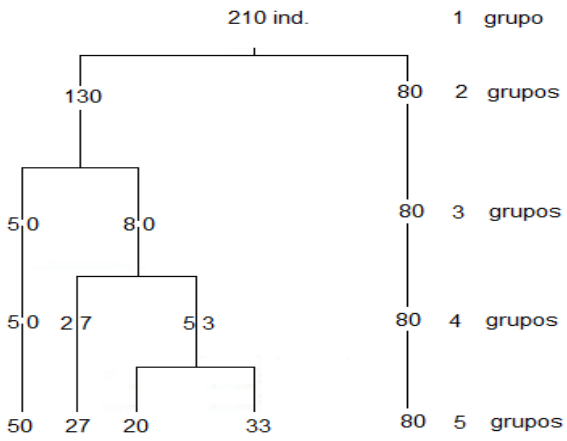

Figura 5 - Árvore de agrupamento de objetos hipotéticos) (Fonte: Reis, 2001)

Tal como acontece com os métodos de otimização em que é necessário definir previamente o número de grupos desejados, por vezes existe o conhecimento, por parte do investigador, do número aproximado de grupos em que a população em estudo se deverá dividir. Este método de escolha do número de grupos é muito subjetivo e não pode ser considerada satisfatória por se tornar enviesado pela necessidade de opiniões prévias quanto à correta estrutura dos dados.

Um método alternativo será a comparação gráfica do número de clusters com o coeficiente de fusão, isto é, o valor numérico (distância ou semelhança) para o qual vários indivíduos se une para formar um grupo.

Quando a divisão de um novo grupo não introduz alterações significativas no coeficiente de fusão poderá tornar-se essa partição como sendo ótima. Na figura 10, o exemplo indicado sugere que a partir de 3 grupos, a curva se torna quase paralela a um dos eixos, isto é, os "saltos" mais significativos no coeficiente de fusão dão-se quando se passa de 1 para 2 grupos, donde se poderá concluir que o agrupamento ótimo se verificará na formação de 2 grupos.

Um problema com a utilização deste método surge quando a representação gráfica mostra apenas pequenos saltos e não existe nenhuma maneira de avaliar, através da visualização gráfica, qual o melhor número de grupos. Pará resolver este problema, Mojena (1977) e Mojena e Wishart (1980) desenvolveram o método no sentido de encontrar uma partição ótima.

Outros métodos foram igualmente desenvolvidos, mas em qualquer deles, é necessário um cuidado redobrado na sua utilização que deverá ser sempre acompanhada por um processo de validação estatística dos resultados encontrados.

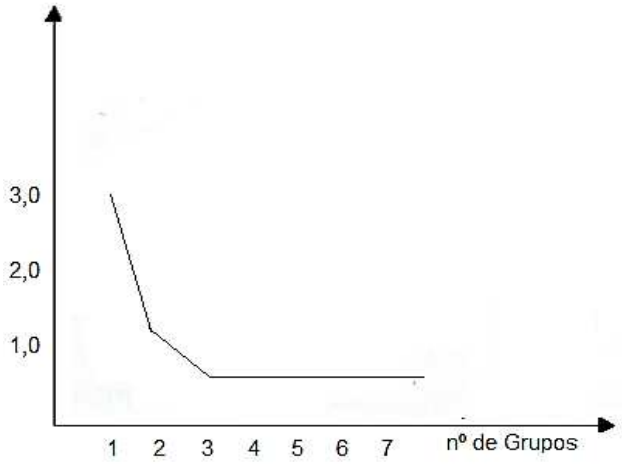

Figura 6 - Coeficientes de fusão

O gráfico do coeficiente de função e traçado a partir dos pontos entre o valor de distância que funde os grupos, cada valor de distância de agregação pode ser observado no dendrograma em cada grupo. Na figura 6, um único grupo atinge o valor máximo de distância, já formação de 2 grupos o salto é estatisticamente considerável, nas formações a partir de 3 grupos as diferenças de distância não apresentam valores significativos.

A escolha de diferentes técnicas de análise de clusters pode produzir resultados diferentes mesmo quando se utiliza uma mesma base de dados. Isto introduz outro problema, que é o da escolha do critério de agrupamento mais apropriado. Muitos estudos debruçaram-se sobre a escolha do melhor método de agrupamento e ao número de grupos e, não é de admirar, que os resultados a que chegaram sejam, por vezes, contraditórios.

Para ajudar à escolha do melhor método, Sokal e Rohlf (1962) definiram o coeficiente de correlação "cofenética" (r_c) que ainda hoje é a medida de validação mais utilizada pelos taxonomistas numéricos. Esta medida dá-nos a relação entre cada valor da matriz de semelhanças e um valor obtido a partir do dendrograma, significando, em última instância, à medida que o dendrograma resultante da aplicação de um método hierárquico, representa os valores da matriz de semelhanças / distâncias. Mais precisamente, a correlação cofenética é a correlação entre os elementos da matriz de distância (ou semelhanças) e os correspondentes coeficientes de fusão, ou seja, as distâncias (ou semelhanças) a que os indivíduos se juntam pela primeira vez para formar grupos. Embora este método de validação seja apropriado sobre tudo quando se utiliza um método hierárquico aglomerativo, foi criticado por Farris (1969) que referiu a sua

sensibilidade ao tamanho dos grupos como razão suficiente para não ser aceite como justificação direta e final da técnica utilizada.

3 MATERIAL E MÉTODOS.

Através da técnica multivariada análise de agrupamento, aplica-se o método hierárquico vizinho mais próximo para agrupar as cidades do agreste paraibano em agrupamento com a característica de menor distância entre elas. O banco de dados das distâncias das 66 cidades que compõe o agreste paraibano foi criado a partir de informações coletadas no google maps.

As informações foram colhidas no período de 20 de setembro de 2012 a 15 de outubro de 2012. O algoritmo de programação e a montagem do banco de dados foram concluídos em dezembro de 2012.

Os municípios cruzam-se entre si formando uma matriz de 66 linhas por 66 colunas, o que nos fornece 4.356 valores de distância entre os municípios.

Para melhor visualização do dendrograma obteve-se uma AAS (amostra aleatória simples) de tamanho 12, com a população em estudo é possível obter mais de 4 bilhões de amostra diferentes. Aplicou-se análise descritiva da amostra antes da análise de agrupamento, através de um algoritmo criado no software R 2.15.1, de forma randômica foi sorteado 12 cidades conforme a listagem. No Quadro 3 tem-se os municípios do agreste paraibano em estudo.

Quadro 3 - Cidades do Agreste Paraibano

1	Alagoa Grande	23	Cuité	45	Natuba
2	Alagoa Nova	24	Cuitegi	46	Nova Floresta
3	Alagoinha	25	Damião	47	Olivedos
4	Algodão de Jandaira	26	Dona Inês	48	Pilões
5	Araçagi	27	Esperança	49	Pilõezinhos
6	Arara	28	Fagundes	50	Pirpirituba
7	Araruna	29	Frei Martinho	51	Pocinhos
8	Areia	30	Gado Bravo	52	Puxinanã
9	Aroeiras	31	Guarabira	53	Queimadas
10	Areial	32	Gurinhém	54	Remigio
11	Bananeiras	33	Ingá	55	Riach. do Bacamart
12	Barra de Santa Rosa	34	Itabaiana	56	Salg. São Felix
13	Belém	35	Itatuba	57	Santa Cecilia
14	Boa Vista	36	Juárez Távora	58	São Sebast. L. Roça
15	Borborema	37	Lagoa de Dentro	59	Serra da Raiz
16	Cacimba de Dentro	38	Lagoa Seca	60	Serra Redonda
17	Caiçara	39	Logradouro	61	Serraria
18	Caldas Brandão	40	Massaranduba	62	Sertãozinho
19	Campo de Santana	41	Matinhas	63	Solânea
20	Campina Grande	42	Mogeiro	64	Sossêgo
21	Casserengue	43	Montadas	65	Tacima
22	Cubati	44	Mulungu	66	Umbuzeiro

(Fonte: http://pt.wikipedia.org/wiki/Mesorregi%C3%A3o_do_Agreste_Paraibano)

4 RESULTADOS E DISCUSSÕES

Neste momento, antes de processar a análise, convém estabelecer um número de cidades julgada adequada, tendo em vista os objetivos da pesquisa e o conhecimento que se tem do universo das distancias entre as cidades paraibanas. Para os fins deste trabalho, estabelece-se como solução final desejada um número de dois clusters: um estariam as cidades que apresentam pequenas distancias uma da outra, e no outro, estariam aqueles grupos de cidades mais distante uma da outra.

A matriz distância no Quadro 4 abaixo, exibe valores que correspondem as distâncias entre as primeiras cidades (Alagoa Nova e Campina Grande), por exemplo a distância entre elas é 27 km. A cidade de Alagoinha e Alagoa Nova a distância corresponde a 42 km e assim sucessivamente para todos os valores. Observou-se que a menor distância é entre a cidade de Belém e Bananeiras, apenas 13 km. Em contrapartida as mais distantes apresentaram 118 km de distância, é o caso das cidades de Natuba e Montadas.

Quadro 4 - Matriz distância da AAS de 12 cidades.

	C.Gra	Alnov	Alagoi	Araça	**Banan**	Belém	Cacim	Cassse	Ingá	Itabai	Mont
Alnov	**27**										
Alagoi	74	**42**									
Araça	92	67	25								
Banan	72	55	42	42							
Belem	86	69	30	28	**13**						
Cacim	84	64	64	63	22	35					
Casse	75	121	133	91	60	62	58				
Inga	37	56	48	74	84	77	93	96			
Itabai	77	89	59	81	99	85	120	126	35		
Mont	27	50	63	88	59	72	67	69	72	104	
Natub	96	103	109	131	148	135	158	160	66	49	118

C.Gra - Campina Grande, Alnov - Alagoa Nova, Alagoi-Alagoinha, Araça-Araçagi, Banan-Bananeiras, Belém, Cacim - Cacimba de Dentro, Casse-Casserengue, Itabai-Itabaiana, Mont-Montadas, Natub-Natuba

Pela distância euclidiana e o método do vizinho mais próximo. Obtém-se a formação do primeiro grupo que será entre a cidade de Belém e Bananeiras, pois apresenta a menor distância entre as cidades na matriz ($\mathbf{d}_{5,6}$ =13).

O Quadro 5, apresenta a formação de grupos com os valores de distância padronizada da matriz de similaridade, os grupos são formados a partir da menor distância entre as cidades, formando dois ou mais grupos, até que todas as cidades formem um único grupo.

Quadro 5 - Formação de grupos

Passo	Distância (Padronizada)	Grupos
1	**D_{56} = 48,97**	(1)(2) (3) (4) (5 6) (7) (8) (9) (10) (11) (12)
2	$D_{(56)\ 7}$= 57,9	(1) (2) (3) (4) (5 6 7) (8) (9) (10) (11) (12)
3	D_{34} = 67,95	(1)(2) (3 4) (5 6 7) (8) (9) (10) (11)(12)
4	D_{34} = 67,95	(1 11) (5 6 7) (3 4) (2) (8) (9) (10) (12)
5	$D_{1\ 11}$ = 72,47	(1 11) (2) (5 6 7) (3 4) (8) (9) (10) (12)
6	$D_{(1\ 11)\ 2}$ = 92,37	(1 2 11) (5 6 7) (3 4) (8) (9 10) (12)
7	$D_{9\ 10}$ = 92,4	(1 2 11) (5 6 7)(3 4) (8) (9 10) (12)
8	$D_{(3\ 4\ 5\ 6\ 7)}$ = 97 ,39	(1 2 11) (8) (9 10) (3 4 5 6 7) (12)
9	$D_{(1\ 2\ 9\ 10\ 11)}$ = 100,07	(1 2 11 8) (9 10) (3 4 5 6 7) (12)
10	$D_{(1\ 2\ 8\ 9\ 10\ 11\ )}$= 103,57	(1 2 8 9 10 11) (3 4 5 6 7) (12)
11	$D_{(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11)\ 12}$ = 136,53	(1 2 3 4 5 6 7 8 9 10 11) (12)
12	$D_{(1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9\ 10\ 11\ 12)}$ = 137,9	(1 2 3 4 5 6 7 8 9 10 11 12)

O dendrograma apresentado na Figura 7, foi construido pelo metodo hierarquico do vizinho mais proximo representando a aglomeração feita em uma escala de 40 a 140. A base do dendrograma representa as variáveis (cidades) que foram associadas, no entanto, este gráfico não imprime as distâncias que dariam uma informação sobre a homogeneidade dos cluster associados.

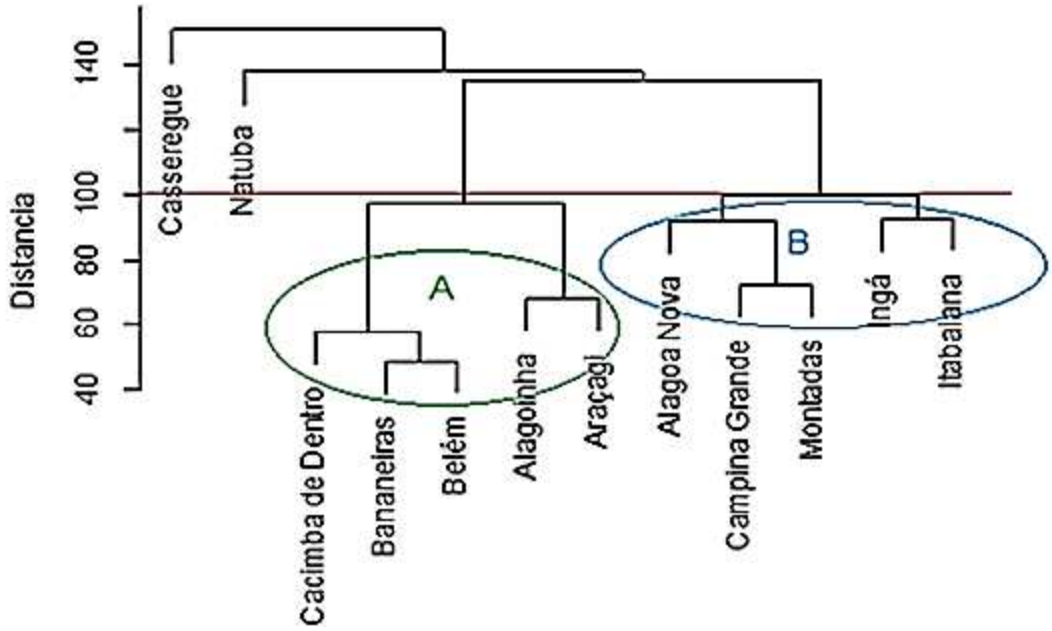

Figura 7. Dendrograma das 12 cidades da amostra do agreste paraibano.

O dendrograma é resultante do Quadro 6, ou seja, o dendrograma da AAS formada pela distância entre 12 cidades, a linha vermelha separa o agrupamento em dois grupos A e B, que possuem distâncias até 100km, o grupo A formado pelas cidades de Cacimba de Dentro, Bananeiras, Belem, Alagoinha e Araçagi, o grupo B formados pelas cidades de Alagoa Nova, Campina Grande, Montadas, Inga e Itabaiana.

A primeira formação de grupos por media ocorre a distância inferior a 50. É visto que as cidade formado pelo grupo A (Cacimba de Dentro, Bananeiras,Belem, Alagoinha e Araçagi) estão mais associadas. As cidades de Bananeiras e Belem apresenta uma grande homogeneidade. As cidades de Casserengue e Natuba apresenta uma certa distância das demais, ambas estão associadas a uma grande distância dos grupos A e B.

Os grupos A e B apresentam características internas homogêneas e entre si característica heterogêneas. O dendrograma é um gráfico em forma de uma árvore, onde podemos averiguar alterações dos níveis de similaridade, para as sucessivas etapas do agrupamento das cidades do agreste paraibano, no eixo vertical o nível de similaridade(distâncias) e no eixo horizontal as cidades, as linhas verticais partindo dos

indivíduos (cidades) agrupados tem altura correspondente ao nível que as cidades são considerados semelhantes. Um método alternativo será a comparação gráfica do número de cluster com o coeficiente de fusão, isto é, o valor numérico (distância ou semelhança) para o qual vários casos se unem para formar um grupo.

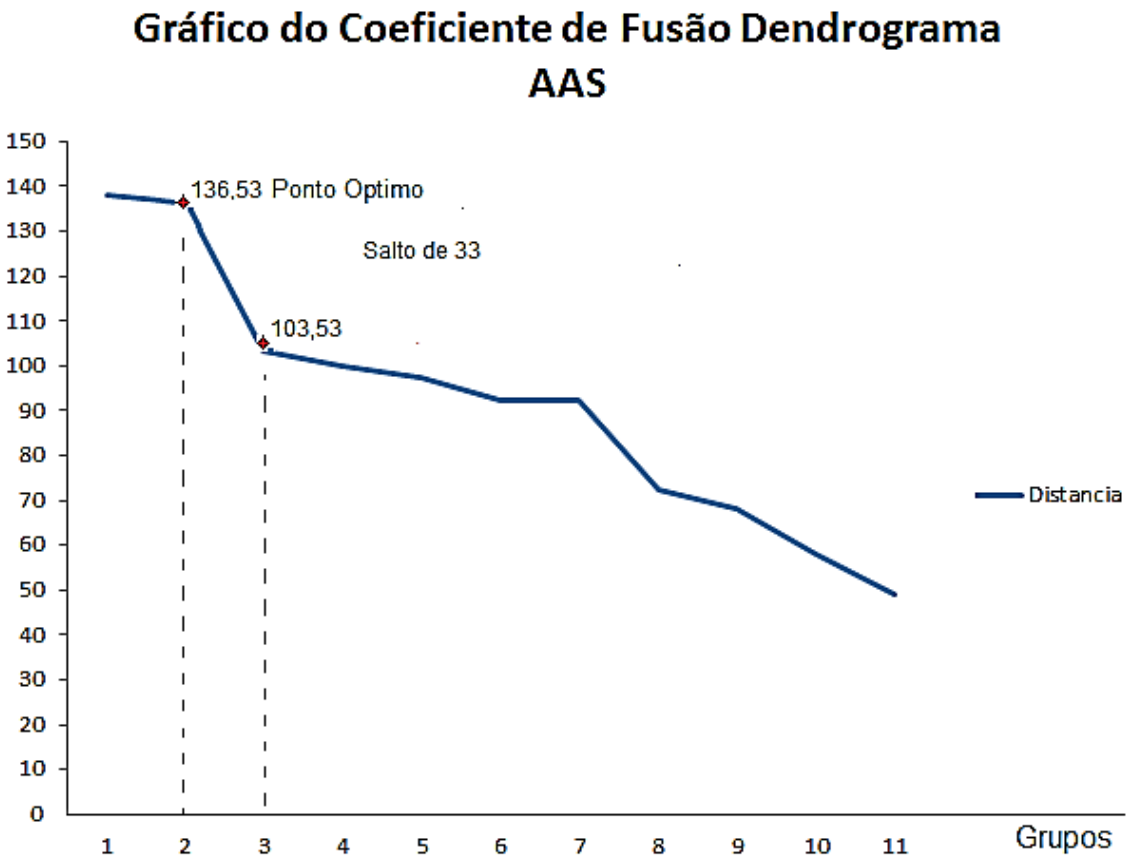

Figura 8. Gráfico do coeficiente de fusão

Observa-se na Fgura 8 que o ponto ótimo está no intervalo de distância de semelhança de 40 a 140, com valor numérico de 136,53 na escala de semelhança. O número de 2 grupos funde todos os objetos. O ponto ótimo obteve-se na formação de dois grupos, pois apresenta “salto” mais significativo com 33 unidades de semelhança, com este teste valida-se os resultados dos dados obtidos, com 2 grupos apresentam características homogêneas internamente e entre grupos características heterogêneas.

Para os agrupamentos hierárquico podemos utilizar uma medida bastante comum, que é a correlação cofenética (ALDENDERFER & BLASHFIEL, 1984; ROMESBURG, 2004) . O coeficiente de correlação cofenética mede o grau de

preservação das distancias emparelhadas pelo dendrograma resultante do agrupamento em relação às distancias originais (SNEATH & SOKAL, 1973).

A partir da inexistência de um método para selecionar a melhor técnica de agrupamento, é importante avaliar o grau de ajuste do agrupamento, coeficiente de correlação cofenética (CCC) menor que 0,7, indica inadequação do método de agrupamento, quanto maior o CCC melhor é o agrupamento. Nesse caso obteve-se um CCC de 0.7803, o que torna o agrupamento adequado para amostra aleatória simples (AAS) das 12 cidades.

A Figura 9 abaixo apresenta o dendrograma completo da população das 66 cidades do agreste paraibano.

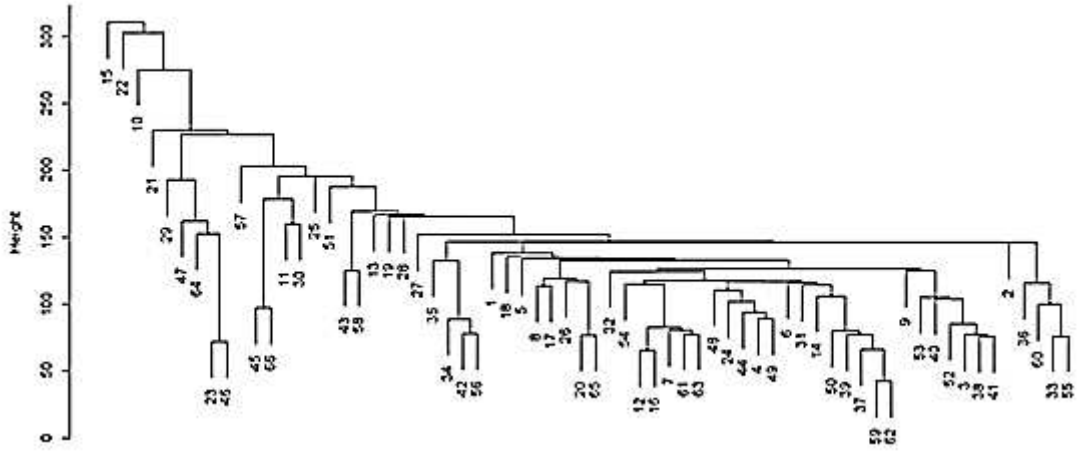

Figura 9. Dendrograma das cidades do agreste paraibano.

Percebe-se que na Figura 9, o lado esquerdo exibe uma régua de distância que vai de 0 a 300 unidades de semelhança, as cidades 59 e 62 agrupa-se a uma distância menor que 50 unidades de semelhança, a maioria das cidades se agrupam no intervalo de 50 a 200 unidades de semelhança, já as os agrupamentos formados pelas cidades 15, 22 e 10 possui afastamentos dos demais grupos.

Os resultados obtidos também foram validados pelo coeficiente de correlação cofenética e apresentou CCC = 0.70 através do algoritmo no software R.

5 CONCLUSÕES

Conclui-se que a análise de agrupamento é uma técnica multivariada que utiliza um conjunto de algoritmos com a finalidade formar grupos de objetos, através das similaridades e dissimilaridades, de forma que os grupos possuam internamente característica homogênea e entre grupos características heterogêneas.

A partir das cidades formou-se agrupamento pelo método hierárquico do vizinho mais próximo, ou seja, menor distância entre as cidades do agreste paraibano. Através do coeficiente de fusão conclui-se que as 12 cidades apresentaram-se discriminadas em dois grupos, ou seja, os dois grupos possuí características internas homogêneas e entre grupo característica heterogêneas das cidades constituintes.

Através do Coeficiente Correlação Cofenética (CCC) obteve-se a confirmação do método selecionado, o que valida à formação dos agrupamentos.

6 REFERÊNCIAS

ALDENDERFER, M. S. e R. K. BLASHFIEL (1984), Cluster Analysis, Sage University 88 p.

BAIN, L. J. ENGELHARDT, M. Introduction to probability and mathematical Statistics. Ed. Belmont-CA Duxbury Press, 1992. 644p

CORMACK, R. (1971), <<A review of classification>>, Journal of the Royal Statistical Society (Series A) 134: 321-367.

CASELLA; BERGER, R.L. Statitical Inference. 2. Ed. Pacific grove-CA: Duxbury Press 2002. 660p.

DRIVER, H. E. (1965), <<Survey of Numerical classification in anthropology>>, in D. Hymes (ed) The off computers in Anthoropology, pp 304-344.

FARRIS, J. S. (1969), <<On The cophenetic correlation coeffient>>, systematic Zoology 18:279-285.

JOHNSON, S. (1967), Hierarchical clustering schemes, Psychometrica, 38:241-254.

JOHNSON, R. A. e D. W. WICHERM (1988, 2º ed.), Applied Multivariate Statistical Analysis, Prentice Hall.

HAMER, R. E J. CUNNINGHAM (1981), <<Cluster analyzing profile data confounded with interrater differences: a comparison of profile measures>>, Applied Psychological Measurement, 5: 63-72.

MOJENA, R. (1977), <<Hierarchical grouping methods and stopping rules –an evaluation>>, computer journal, 20: 41-50.

MORETTIN. L. G. Estatistica basic V 1.7. Ed. Edifice São Paulo Makron Books, 1999. 210 p.

LANCE, G. E W. WILLIAMS (1967), << A general theory of classificatory sorting strategies>>, Computer Journal, 9: 373-380.

ELISABETH REIS (2001), Estatística Multivariada Aplicada, 2ª edição, 287-336

MOJENA, R. E D. WISHART (1980), <<Stopping rules for Ward´s clustering method>>, in proceedings of COMPSTAT 1980, PP. 426-423.

PESTANA, M. H. E GAGEIRO, J.N. (2000, 2ª ed.) pag.429, Analise de Dados para Ciências Sociais.

- A complementariedade do SPSS, Edições Silabo.

ROSS, S. M. Introduction to probability and statistics for engineers and scientists 2. Ed. San Diego-CA: Harcourt Academic Press, 2000. 578 p.

ROHLF F.J (1970)"Adaptive Hierarchical Clustering Schemes." Systematic Zoology 19:58.

SOKAL, R. E P. SNEADTH (1963), Principles of numerical taxonomy, Ed. W.H. Freeman.

SNEATH, P.H. A. E SOKAL, R. R. (1973), Numerical Taxonomy, W. H. Frenman.

SOKAL, R. E F. ROHLF (1962) , <<The Comparson of dendrograma by objective methods>>, Taxon 11:34-40

TVERSKY, A. (1977), << Features of similarity>>, Psychological, Review, 84 (4).

LUIZ, J. CORRAR, EDILSON P. FILHO, J.M (Coordenadores). Análise multivariada: para os cursos de administração, ciências contábeis e economia. São Paulo: Atlas, 2007.

7 APÊNDICE

Algoritmo da análise descritiva e da análise de cluster dos dados usando o software R 2.15.1.

rm(list=ls(all=TRUE)) # limpa memoria

bdados<-read.table("bdcidagpb.csv",header=T,sep=";") # importação dos dados do arquivo bdcidagpb.csv do excel

bdados

Distância em relação a Campina Grande

vetcamp<-bdados[,2]

relatorio<-bdados[,1]

#Uma Amostra Aleatória Simples tamanho de 12 municípios, para melhor visualização dos gráficos; Geração da Amostra aleatória Simples

rela<-c(1:66) ;rela

#vetor com sequência de 1 a 66 para listar as cidades do agrestre Paraibano para numerarmos os municípios, Mais de 490 bilhões de quantidade de amostras possíveis de tamanho 12 para 66 cidades, conforme a combinação de 66 tomada 12.

choose(66,12) ;#Combinação de 66 por 12: 66!/(12!(66-12)!)

amost<-sample(rela,12) # AAS de tamanho 12

amost # Resposta: 1, 17, 43, 3, 4, 14, 34, 21, 45, 6, 12, 33

sort(amost) # colocando a amostra em ordem crescente

1 3 4 6 12 14 17 21 33 34 43 45

Em ordem crescente (AAS) as cidade selecionadas são:

1 Campina Grande

3 Alagoa Nova

4 Alagoinha

6 Araçagi

12 Bananeiras

14 Belém

17 Cacimba de Dentro

21 Casserengue

33 Ingá

```
# 34 Itabaiana
# 43 Montadas
# 45 Natuba
# Vetor com valor das distância de campina as cidades conforme AAS
amostra<c(bdados[,1],bdados[,3],bdados[,4],bdados[,6],bdados[,12],bdados[,14],bdados
[,17],bdados[,21],bdados[,33],bdados[,35],bdados[,43],bdados[,45])
#todas as distâncias da amostras de tamanho 12.
```

Calculo da Moda

```
d<-amostra
 moda<-function(d)     {
  if ((is.vector(d) || is.matrix(d) || is.factor(d)==TRUE) &&
     (is.list(d)==FALSE))
     {
     dd<-table(d)
     valores<-which(dd==max(dd))
     vmodal<-0
     for(i in 1:(length(valores)))
        if (i==1) vmodal<-as.numeric(names(valores[i]))
           else

vmodal<-c(vmodal,as.numeric(names(valores[i])))
     if (length(vmodal)==length(dd))
        print("conjunto sem valor modal")
           else return(vmodal)
     }
     else print("o parâmetro deve ser um vetor ou uma matriz")
  }
```

Medidas de Posição Amostral

Moda(d) # a moda da amostra

mean(amostra) # Media da amostra Aleatória dos 12 municípios do agreste paraibano a Campina Grande

#[1] 70.72727

median(amostra) # Mediana da amostra das distâncias entre municípios do agreste paraibano e Campina Grande

#[1] 73

moda(d) # Moda da amostra das distâncias entre municípios do agreste paraibano e Campina Grande

Medidas de Dispersão Amostral

var(amostra) # Variância da amostra das distâncias entre municípios do agreste paraibano e Campina Grande

#[1] 1267.524

sd(amostra) # Desvio Padrão da amostra das distâncias entre municípios do agreste paraibano e Campina Grande

#[1] 35.60231

amplit<-max(amostra)-min(amostra) # Amplitude das distâncias entre municipios do agreste paraibano e Campina Grande

amplit #[1] 149

Medidas Diversas Amostral

sum(amostra) # Total do vetor

range(amostra) # mostra valor Máximo e mínimo das distâncias entre municípios do agreste paraibano e Campina Grande

#resumo da matrix dados

summary(bdados) # resumo descritivo - menor valor, 1º quantil, Mediana (2º quartil, centro dos dados), media, 3º quartil, Máximo Valor

summary(amostra) # resumo descritivo - menor valor, 1º quantil, Mediana (2º quartil, centro dos dados), media, 3º quartil, Máximo Valor

pnorm(20,70.73,35.60)

Probabilidade de chegar um número menor ou igual a 2**0.**

pnorm(70.73,70.73,35.60)

Probabilidade de ocorrer uma distância menor ou igual a medias das distância de campina aos municípios

Gráficos Boxplot

vetamostra<-amostra

vetcamp<-bdados[1:66,2]

vetalnova<-bdados[1:66,4]# gráfico boxplot das distância dos municípios a Alagoa Nova

vetalago<-bdados[1:66,5]# gráfico boxplot das distância dos municípios Alagoinha

vetaracagi<-bdados[1:66,6]# gráfico boxplot das distância dos municípios a Araçagi

vetbananeira<-bdados[1:66,13]# gráfico boxplot das distância dos municípios a Bananeiras

vetbelem<-bdados[1:66,15]# gráfico boxplot das distância dos municípios a Belém

vetcacimb<-bdados[1:66,18]# gráfico boxplot das distância dos municípios a Cacimba de Dentro

vetcasseren<-bdados[1:66,22]# gráfico boxplot das distância dos municípios a Casserengue

vetinga<-bdados[1:66,34]# gráfico boxplot das distância dos municípios a Ingá

vetitabaiana<-bdados[1:66,35]# gráfico boxplot das distância dos municípios a Itabaiana

vetmontada<-bdados[1:66,44]#grafico boxplot das distância dos municipios a montadas

vetnatuba<-bdados[1:66,46]# gráfico boxplot das distância dos municípios a natuba

Boxplot (amostra)

Boxplot(vetcampvetalnova,vetalago,vetaracagi,vetbananeira,vetbelem,vetcacimb,vetcasseren,vetinga,vetitabaiana,vetmontada,vetnatuba);title('Boxplot das cidades da Amostra Aleatória Simples do Agreste Paraibano') # boxplot de 12 amostras aleatórias simples

Transforma banco de dados da amostra em matriz distância

matdisaas<- as.dist(aas) #Matriz distância

matdisaas

#limite de casas decimais da matriz de semelhança

options(digits=4)

aasdist<-dist(aas,method='euclidean') # matriz de semelhança

aasdist # matrix semelhança

agrup<-hclust(aasdist, method='single')

agrup

Call:

hclust(d = aasdist, method = "single")

Cluster method : single

Distance : euclidean

Number of objects: 12

Gerando o dendrograma

Dendrograma da AAS de 12 municípios do Agreste paraibano

plot (hclust(dist(1-d), method="single"))

> #Calculo do coeficiente cofenético

> F <- dist (mdist)

> hc <- hclust (F, "single")

> C <- cophenetic (hc)