UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE MATEMÁTICA
CURSO DE ESTATÍSTICA

BRUNA QUEIROZ DE MELO PRADO

ANÁLISE DE AGRUPAMENTOS DAS TAXAS DE INCIDÊNCIA DE DENGUE NOS ESTADOS BRASILEIROS

Uberlândia – MG
Dezembro – 2015

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE MATEMÁTICA
CURSO DE ESTATÍSTICA

ANÁLISE DE AGRUPAMENTOS DAS TAXAS DE INCIDÊNCIA DE DENGUE NOS ESTADOS BRASILEIROS

Bruna Queiroz de Melo Prado

Priscila Neves Faria

Trabalho de Conclusão de Curso apresentado à Coordenação do Curso de Estatística, da Universidade Federal de Uberlândia, para a obtenção do grau de Bacharel em Estatística.

Uberlândia – MG
Dezembro – 2015

UNIVERSIDADE FEDERAL DE UBERLÂNDIA
FACULDADE DE MATEMÁTICA
CURSO DE ESTATÍSTICA

ANÁLISE DE AGRUPAMENTOS DAS TAXAS DE INCIDÊNCIA DE DENGUE NOS ESTADOS BRASILEIROS

Bruna Queiroz de Melo Prado

Priscila Neves Faria
Faculdade de Matemática

Homologado pela Coordenação do Curso de Estatística em ___ / ___ / ___.

Edmilson Rodrigues Pinto

Uberlândia – MG
Dezembro – 2015

RESUMO

A dengue é considerada um grave problema de saúde pública no Brasil, país que concentra quase 70% de notificações da doença nas américas desde 2000 e onde cerca de 4 milhões de casos foram registrados entre 2000 e 2009. Assim, objetiva-se com o presente trabalho analisar as taxas de incidência de dengue nos 27 estados brasileiros, no período de 2000 a 2014, por meio da Análise de Agrupamentos, com o intuito de identificar e descrever o perfil dos estados com comportamentos similares a fim de contribuir para o redirecionamento das estratégias dos programas de controle dessa endemia e fomentar ações de prevenção. Para isso, foi utilizada a distância Euclidiana Padronizada para a obtenção da matriz de distâncias, e o método hierárquico aglomerativo da Ligação Média para a formação dos agrupamentos. Quatro métodos foram utilizados para determinação do número ótimo de grupos no dendrograma, sendo eles o Método de Otimização de Tocher original e modificado, e os critérios Pseudo F e Pseudo T^2, que resultaram na escolha de dois grupos, onde houve a discriminação do estado do Acre em um único grupo devido aos altos índices de incidência da doença registrados. A introdução e recirculação de sorotipos da doença, além do tratamento inadequado dos criadouros do vetor, estão entre os possíveis motivos que ocasionaram nas epidemias de dengue verificadas no estado nesse período.

Palavras-chave: Epidemia; vetor; dissimilaridade; Ligação Média; dendrograma.

ABSTRACT

Dengue is considered a serious public health problem in Brazil, which accounts for almost 70% of notifications of the disease in the Americas since 2000 and where about 4 million cases were recorded between 2000 and 2009. Thus, this study aims to analyze the dengue incidence rates in 27 states, from 2000 to 2014, through cluster analysis, for the purpose of to identify and describe the profile of states with similar behaviors in order to contribute to redirect the strategies of control programs of this endemic disease and encourage preventive measures. For this, the Standardized Euclidean distance for obtaining the distance matrix was used, and the unweighted pair-group method using the average approach (UPGMA) for the formation of clusters. Four methods were used to determine the optimal number of groups in the dendrogram, being they Tocher Optimization Method original and amended, and the criteria Pseudo F and Pseudo T^2, which resulted in the choice of two groups, where there was the discrimination of the state of Acre on a single group because of the high incidence reported of the disease. The introduction and recirculation of serotypes of the disease and the inadequate treatment of vector breeding sites are among the possible reasons that caused the epidemics of dengue recorded in the state during this period.

Keywords: Epidemic; vector; dissimilarity; UPGMA; dendogram.

SUMÁRIO

1. INTRODUÇÃO

A dengue é uma doença que se tornou um grave problema de saúde pública no Brasil. Essa arbovirose é causada por quatro tipos de vírus do gênero *Flavivirus*: DENV-1, DENV-2, DENV-3 e DENV-4, e transmitida por mosquitos do gênero *Aedes*. A infecção pode tomar uma série de formas, desde formas de infecção leves e assintomáticas até as formas mais sérias e fatais (MARZOCHI, 1994). Segundo Ferreira et al. (2009), a importância dessa doença está relacionada à sua morbidade, mortalidade e necessidade de várias estratégias para o seu controle.

O aparecimento desta doença é apenas uma das consequências causadas pela urbanização desordenada que se verifica em países de economia emergente (CHIARAVALLOTI NETO et al., 2006). De acordo com o Boletim Epidemiológico de 1999 da Fundação Nacional de Saúde (FUNASA), são vários os fatores que contribuem para a proliferação do mosquito *Aedes aegypti*, seu principal vetor urbano, como o fluxo populacional, as condições ambientais precárias dos grandes centros urbanos, as condições climáticas favoráveis à reprodução dos vetores, além da falta de implementação de ações eficientes de combate e controle vetorial.

O *Aedes aegypti* possui uma grande capacidade de adaptação ao ambiente urbano, onde expõe mais de 2,5 bilhões de pessoas no mundo ao risco de contrair a doença. Esse vetor se reproduz principalmente em recipientes comumente encontrados nos domicílios, como vasos de flores, pneus velhos, baldes, e lixos em geral. Grandes reservatórios de água, como tambores e cisternas, localizados próximos à habitações humanas, são importantes produtores de grande quantidade de mosquitos na forma adulta (GLUBER, 1998).

No Brasil, a dengue possui um padrão sazonal, com maior taxa de incidência nos cinco primeiros meses do ano, período mais quente e úmido, característico dos climas tropicais (FUNASA, 1999). Segundo o Ministério da Saúde, o país concentra quase 70% dos casos notificados da doença nas américas desde 2000. Entre os anos 2000 e 2009, cerca de 4 milhões de casos de dengue foram notificados no Brasil, com ênfase para 2002 e 2008, anos em que foram registradas as maiores epidemias de dengue da década.

Diante do exposto, percebe-se a importância da produção científica em torno dessa temática, no intuito de descrever o perfil das taxas de incidência de dengue em território brasileiro, contribuindo, assim, para o redirecionamento das estratégias dos

programas de controle dessa endemia e fomentar ações de prevenção à doença e promoção à saúde, visto que ainda não há controle efetivo da dengue no país.

Apesar das técnicas de análise multivariada terem sido desenvolvidas para resolver problemas específicos, principalmente nas áreas de Biologia e Psicologia, essas podem ser também utilizadas para resolver outros tipos de problemas em diversas áreas do conhecimento (VICINI, 2005).

Um dos métodos mais utilizados para se classificar objetos em categorias de similaridade é a Análise de Agrupamentos (*Cluster Analysis*). Essa técnica considera um conjunto inicial de objetos, aos quais são associadas medidas de várias grandezas, denominadas variáveis classificatórias, utilizadas para se obter grupos de objetos assemelhados em relação aos valores assumidos por essas variáveis (EVERITT, 1993). Além de possibilitar a construção de grupos de acordo com as similaridades dos indivíduos, a Análise de Agrupamentos possibilita também representá-los de maneira bidimensional, através de um dendrograma (MOITA NETO e MOITA, 1997).

Um grande rol de aplicações utiliza a Análise de Agrupamentos para reunir regiões geográficas. Frei e Prado (1994) utilizam a Análise de Agrupamentos para reunir as 42 Regiões de Governo do Estado de São Paulo, por meio de variáveis de industrialização e urbanização, com o objetivo de analisar o comportamento dos grupos formados em relação às variáveis de mortalidade violenta. Pinto e Curi (1991) identificam grupos de estados brasileiros em função da mortalidade por tipo de neoplasia, onde estados como São Paulo, Rio de Janeiro e Rio Grande do Sul formaram um grupo homogêneo no que se refere à presença de alguns tipos específicos de neoplasias.

Na área de saúde pública, uma linha de aplicação da Análise de Agrupamentos é aquela que procura obter subgrupos de indivíduos na população com riscos em relação a determinados agravos para possíveis prevenções (FREI, 2006).

Assim, com base no exposto, pretende-se com este trabalho analisar a incidência de dengue nos 27 estados brasileiros, no período de 2000 a 2014, por meio da técnica multivariada de Análise de Agrupamentos. Pretende-se com esta análise identificar os estados com comportamentos similares em relação à incidência da dengue, além de descrever o perfil dos estados semelhantes, a fim de contribuir para o redirecionamento das estratégias dos programas de controle dessa endemia e fomentar ações de prevenção à doença.

2. REFERENCIAL TEÓRICO

2.1 Análise de Agrupamentos

São diversas as definições encontradas na literatura para a técnica de Análise de Agrupamentos (também conhecida como Análise de *Cluster*). De acordo com Everitt (1993) e Manly (1986), a Análise de Agrupamentos é uma técnica que objetiva agrupar os indivíduos (casos) que possuem características semelhantes em função de um conjunto de variáveis selecionadas. Para Khattree e Naik (2000), a Análise de Agrupamentos é uma técnica multivariada de grande aplicabilidade, principalmente o procedimento *Cluster*, cujo objetivo da classificação é repartir os indivíduos em grupos homogêneos, de modo que cada um seja bem diferenciado. Após a obtenção dos resultados, esses dados servirão para a definição do número de grupos distintos.

A Análise de Agrupamentos situa-se como uma técnica indivíduo-dependente, na qual valores de distâncias, sob a forma de matrizes, são arranjados entre os objetos. Neste caso a estimação de parâmetro não é requerida, o que lhe ratifica o caráter não-probabilístico (CHATFIEL & COLLINS, 1986). Segundo Mardia, Kent e Bibby (1995), a Análise de Agrupamentos apresenta a vantagem de reduzir o espaço multidimensional a uma medida de distância entre os objetos, representando esta em um espaço bidimensional, muito mais simplificado do que o espaço multidimensional.

A maioria dos métodos de Análise de Agrupamentos requer uma medida de similaridade ou dissimilaridade entre os elementos a serem agrupados, normalmente expressa como uma função distância para dados que têm propriedades métricas (DONI, 2004), ou então uma medida de "comparação" se considerarmos dados que têm componentes qualitativos.

O processo de agrupamento começa levando as medidas das p variáveis em cada um dos n objetos. Assim, a matriz $n \times p$ de dados é transformada em uma matriz $n \times n$ de semelhança ou, alternativamente, por medidas de distância onde são computadas as semelhanças ou distâncias entre pares de objetos pelas p variáveis. Logo, um algoritmo é selecionado com a finalidade de definir as regras que concernem ao agrupamento dos objetos em subgrupos com base nas semelhanças (VALLI, 2002).

Quando as variáveis são quantitativas, as medidas de proximidade entre os objetos são normalmente medidas de distância ou dissemelhança. Atualmente, diversas

medidas de dissimilaridade são propostos na literatura, principalmente devido ao grande desenvolvimento e utilização das técnicas multivariadas (KHATTREE e NAIK, 2000).

Segundo Khattree e Naik (2000) e Cruz e Carneiro (2006), dentre as medidas de dissimilaridades conhecidas, a distância Euclidiana e a distância de Mahalanobis estão entre as medidas que mais se destacam, devido a sua maior utilização.

A distância Euclidiana é, sem dúvida, a medida de distância mais utilizada para a Análise de Agrupamentos. Segundo Cormack (1971) a distância Euclidiana entre dois casos (i e j) é a raiz quadrada do somatório dos quadrados das diferenças entre os valores de i e j para todas as variáveis ($k = 1, 2, \dots, p$), isto é:

$$d_{ij} = \sqrt{\sum_{k=1}^{p} \left(X_{ik} - X_{jk}\right)^2}$$

em que X_{iv} representa a característica do indivíduo i; X_{jv} representa a característica do indivíduo j; e p representa o número de variáveis na amostra.

Quando se trabalha com a distância Euclidiana, geralmente são somadas grandezas não comparáveis, como por exemplo centímetros, quilos, anos, milhões, etc., muito embora, a mudança de uma das unidades possa alterar completamente o significado e o valor do coeficiente. Essa é uma das razões da padronização das variáveis dos elementos $x_1, x_2, \dots, x_p$ do vetor X, dada por:

$$z_i = \frac{x_i(.) - \overline{x_i}}{s_i}$$

onde $\overline{x_i}$ e s_i indicam, respectivamente, a média e o desvio padrão da *k-ésima* variável. Feita a transformação, a distância Euclidiana passa a ser (BUSSAB, MIAZAKI e ANDRADE, 1999):

$$d(A, B) = \left[\sum_{i=1}^{p} \left(\left(z_i(A) - z_i(B)\right)\right)^2\right]^{\frac{1}{2}}$$

que é a soma dos desvios padronizados. A expressão acima pode ser escrita da seguinte forma em notação vetorial:

$$d(A,B) = \left[\left(x(A) - x(B)\right)' D^{-1}\left(x(A) - x(B)\right)\right]^{\frac{1}{2}}$$

em que D é uma matriz diagonal, tendo como i-ésimo componente a variância s_i^2, isto é,

$$D = diag(s_1^2, s_2^2, \ldots, s_p^2)$$

Outra medida de distância bastante conhecida é a distância de Mahalanobis, que de acordo com Quintal (2006), além de reduzir a dependência das unidades de medição, reduz também a correlação entre variáveis. A distância de Mahalanobis entre os grupos i e j é usualmente estimada segundo Rao (1952) por:

$$D_{ij}^2 = \left(\overline{X}_i - \overline{X}_j\right)' \cdot \Sigma^{-1} \cdot \left(\overline{X}_i - \overline{X}_j\right)$$

em que $\overline{X}_i$ é o vetor de médias do *i-ésimo* grupo; $\overline{X}_j$ é o vetor de médias do *j-ésimo* grupo; e Σ é a estimativa da matriz da covariância/variância entre as variáveis.

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).

2.2 Técnicas Hierárquicas Aglomerativas

As técnicas de conglomerados ou agrupamentos são frequentemente classificadas em dois tipos: técnicas hierárquicas e não hierárquicas, sendo que as hierárquicas são classificadas em aglomerativas e divisivas. As técnicas hierárquicas, na maioria das vezes, são utilizadas em análises exploratórias dos dados com o intuito de identificar possíveis agrupamentos e o valor provável do número de grupos. Já para o uso de técnicas não hierárquicas, é necessário que o valor do número de grupos já esteja pré-estabelecido pelo pesquisador (MINGOTI, 2013).

Segundo Mingoti (2013), as técnicas hierárquicas aglomerativas partem do princípio de que no início do processo de agrupamento tem-se n conglomerados, ou seja, cada elemento do conjunto de dados observado é considerado como sendo um conglomerado isolado. Em cada passo do algoritmo, os elementos amostrais vão sendo agrupados, formando novos conglomerados até o momento no qual todos os elementos

considerados estão em um único grupo. Já nos métodos divisivos, todos os objetos pertencem inicialmente ao mesmo grupo, que vai sendo dividido, até que cada observação forme um grupo individualmente (JOHNSON e WICHERN, 1992).

Como resultado da Análise de Agrupamentos, tem-se o dendrograma, que apresenta o arranjo entre os objetos em uma escala de distância. O dendrograma é um gráfico em forma de árvore no qual a escala vertical indica o nível de similaridade (ou dissimilaridade). No eixo horizontal, são marcados os elementos amostrais numa ordem conveniente relacionada à história de agrupamento. As linhas verticais, partindo dos elementos amostrais agrupados, têm altura correspondente ao nível em que os elementos foram considerados semelhantes, isto é, a distância do agrupamento ou o nível de similaridade (MINGOTI, 2013). Na Figura 1, é ilustrado um exemplo de dendrograma, em que uma hierarquia significante está indicada pelas linhas tracejadas.

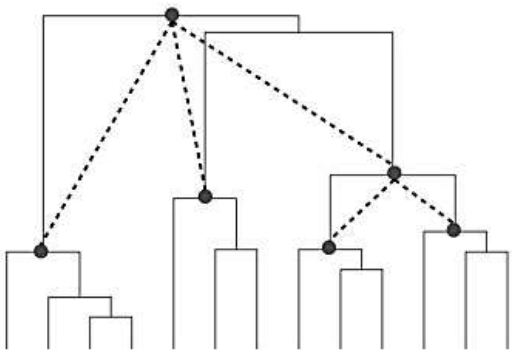

Figura 1: Exemplo de dendrograma.

Os métodos de agrupamentos mais utilizados são os hierárquicos aglomerativos. De acordo com Anderberg (1973), dentre os métodos hierárquicos aglomerativos mais utilizados, estão: Ligação simples (conhecido também como Single Linkage ou Critério do Vizinho mais Próximo); Ligação Completa (conhecido também como Complete Linkage ou Critério do Vizinho mais Distante); Ligação Média (ou UPGMA); Método do Centroide (ou UPGMC); Ligação Mediana (ou WPGMC) e Critério de Ward.

2.2.1 Método da Ligação Simples

No método da Ligação Simples, a similaridade entre dois conglomerados é definida pelos dois elementos mais parecidos entre si (SNEATH, 1957). Por meio desse

método, a distância entre uma observação k e um grupo formado pelas observações i e j é dada por:

$$D^2_{(ij)k} = \min(D^2_{ik}; D^2_{jk})$$

em que $D^2_{(ij)k}$ é a distância entre o grupo ij e a observação k; e $\min(D^2_{ik}; D^2_{jk})$ é a menor distância entre os grupos de observações ik e jk (AMARAL JÚNIOR e THIÉBAUT, 1999).

Segundo Anderberg (1973), algumas características desse método são: grupos muito próximos podem não ser identificados; permite detectar grupos de formas não-elípticas; apresenta pouca tolerância a ruído uma vez que tem tendência a incorporar os ruídos em um grupo já existente; apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias; e tem tendência a formar longas cadeias (encadeamento).

2.2.2 Método da Ligação completa

O agrupamento por Ligação Completa é exatamente o oposto do Método da Ligação Simples. Nesse caso, os elementos são agrupados considerando a distância máxima (ou similaridade mínima) (DILLON e GOLDSTEIN, 1984). O algoritmo é iniciado encontrando a menor distância $D = \{d_{ik}\}$ e agrupando os objetos correspondentes (U e V) para formar o grupo (UV). No passo 3 do algoritmo, as distâncias entre (UV) e qualquer outro grupo W são calculadas por (JOHNSON e WICHERN, 2002):

$$d_{(UV)W} = max\{d_{UW}, d_{VW}\}$$

Este método tem tendência para encontrar *clusters* compactos compostos de objetos muito semelhantes entre si. Quando um objeto é acrescentado a um grupo, a distância do novo grupo aos restantes aumenta ou então fica inalterada. O método de ligação completa tende a formar grupos pequenos que depois serão aglutinados para formar grupos maiores (QUINTAL, 2006).

2.2.3 Método da Ligação Média

A ligação média entre *clusters*, também conhecida por UPGMA (Unweighted Pair-Group Method using the Average approach), é um método não-ponderado de agrupamento aos pares, utilizando médias aritméticas das medidas de dissimilaridade, que evita caracterizar a dissimilaridade por valores extremos (máximo ou mínimo) (CRUZ & CARNEIRO, 2006). Este método trata a distância entre dois conglomerados como a média das distâncias entre todos os pares de elementos que podem ser formados com os elementos dos dois conglomerados que estão sendo comparados. Portanto, se o conglomerado C_1 tem n_1 elementos e o conglomerado C_2 tem n_2 elementos, a distância entre eles será definida por (MINGOTI, 2013):

$$d(C_1, C_2) = \sum_{l \in C_1} \sum_{k \in C_2} \left(\frac{1}{n_1 n_2}\right) d(X_1, X_k)$$

A grande vantagem deste critério é tornear as consequências da existência de valores extremos e considerar toda a informação dos grupos. No entanto, dependendo do tipo de *cluster* que se espera obter, esta propriedade também pode ser vista como uma desvantagem (QUINTAL, 2006).

2.2.4 Método do Centroide

No Método do Centroide (ou UPGMC), a distância entre dois grupos é definida como sendo a distância entre os vetores de médias, também chamados de centroides, dos grupos que estão sendo comparados. Logo, considerando dois conglomerados C_1 e C_2 e seus respectivos vetores de médias amostral $\bar{X}_1$ e $\bar{X}_2$, a distância entre C_1 e C_2 é definida por:

$$d(C_1, C_2) = (\bar{X}_1 - \bar{X}_2)'(\bar{X}_1 - \bar{X}_2)$$

que é a distância Euclidiana ao quadrado entre os vetores de médias amostral $\bar{X}_1$ e $\bar{X}_2$. O método do centroide também pode ser usado com a distância Euclidiana usual entre os vetores de médias. Em cada passo do algoritmo do agrupamento, os conglomerados que apresentam o menor valor de distância são agrupados (MINGOTI, 2013).

De acordo com Mingoti (2013), o Método do Centroide é direto e simples, porém, para fazer o agrupamento, é necessário em cada passo voltar-se aos dados originais para o cálculo da matriz de distâncias, exigindo um tempo computacional maior comparado com outros métodos. O método do centroide não pode ser usado em situações nas quais se dispõe apenas da matriz de distâncias entre os n elementos amostrais.

2.2.5 Ligação Mediana ou WPGMC

O Método da Distância Mediana, também conhecido por WPGMC (weighted pair-group method using the centroid approach), é semelhante ao Método do Centroide, com a diferença de que ao aglutinar os dois grupos A e B, os seus centroides, $\bar{X}_A$ e $\bar{X}_B$, recebem pesos iguais antes de produzirem o centroide do novo *cluster* resultante da aglutinação. O novo centroide $\bar{X}$ será definido como sendo a mediana dos centroides dos grupos aglutinados,

$$\bar{X} = \frac{(\bar{X}_A + \bar{X}_B)}{2}$$

com o objetivo de evitar que o grupo com maior número de objetos absorva o grupo com menor número de objetos. Repare que a mediana referida não corresponde à mediana estatística mas sim à mediana de um triângulo, isto é, um segmento de reta que une um vértice de um triângulo ao ponto médio do lado oposto (QUINTAL, 2006).

2.2.6 Critério de Ward

O método de agrupamento proposto por Ward (1963) é fundamentado na mudança de variação entre os grupos e dentro dos grupos que estão sendo formados em cada passo do agrupamento. Seu procedimento é também conhecido como Mínima Variância, e fundamenta-se nos princípios de que inicialmente, cada elemento é considerado como um único conglomerado, e em cada passo do algoritmo de agrupamento é calculado a soma de quadrados dentro de cada conglomerado. Esta soma é o quadrado da distância Euclidiana de cada elemento amostral pertencente ao conglomerado em relação ao correspondente vetor de médias do conglomerado, isto é,

$$SS_i = \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i.})' (X_{ij} - \bar{X}_{i.})$$

onde, n_i é o número de elementos no conglomerado C_i quando se está no passo k do processo de agrupamento, X_{ij} é o vetor de observações do *j-ésimo* elemento amostral que pertence ao *i-ésimo* conglomerado, $\bar{X}_{i.}$ é o centroide do conglomerado C_i, e SS_i representa a soma de quadrados correspondente ao conglomerado C_i. No passo k, a soma de quadrados total dentro dos grupos é definida como:

$$SSR = \sum_{i=1}^{g_k} SS_i$$

onde g_k é o número de grupos existentes quando se está no passo k.

A distância entre os conglomerados C_l e C_i é, então, definida como:

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i}\right] (\bar{X}_{l.} - \bar{X}_{i.})' (\bar{X}_{l.} - \bar{X}_{i.})$$

que é a soma de quadrados entre os *cluster*s C_l e C_i. Em cada passo do algoritmo de agrupamento, os dois conglomerados que minimizam a distância são combinados.

Tem-se que a medida de distância nada mais é do que a diferença entre o valor de SSR depois e antes de se combinar os conglomerados C_l e C_i em um único conglomerado. Portando, a cada passo do conglomerado, o método de Ward combina os dois conglomerados que resultam no menor valor de SSR (MINGOTI, 2013).

2.3 Método de Otimização de Tocher

Nos métodos de otimização os grupos são formados pela adequação de algum critério de agrupamento. Os métodos de otimização se diferem dos hierárquicos basicamente pelo fato dos grupos formados serem mutuamente exclusivos, ou seja, independentes.

Entre esses métodos, o de Tocher, citado por Rao (1952), utiliza um critério de agrupamento que possui a particularidade de apresentar a distância média intragrupo sempre menor que a distância média intergrupo.

O método de Tocher, como apresentado por Cruz, Ferreira e Pessoni (2011), requer a obtenção da matriz de distâncias, onde é identificado o par de indivíduos mais similares. Esses primeiros indivíduos localizados formarão o grupo inicial. A partir daí é avaliada a possibilidade de inclusão de novos indivíduos, adotando-se o critério de que a distância média intragrupo deve ser menor que a distância média intergrupo.

Neste método, a entrada de um indivíduo em um grupo sempre aumenta o valor médio da distância dentro do grupo. A inclusão, ou não, do indivíduo k no grupo, é feita considerando:

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

$$\text{Se } \frac{d_{(grupo)k}}{n} > \theta, \text{ o indivíduo } k \text{ não é incluído no grupo}$$

sendo n o número de indivíduos que constitui o grupo original; θ é o maior valor do conjunto de menores distâncias entre os indivíduos; e a distância entre o indivíduo k e o grupo formado pelos indivíduos ij dada por $d_{(ij)k} = d_{ik} + d_{jk}$.

Por meio desse agrupamento é possível compreender melhor o porquê do método ser mutuamente exclusivo; isto ocorre, pois os grupos formados são independentes, ou seja, não são relacionados. Daí preconizar-se que tal método é mais confiável do que os hierárquicos, pois nesses últimos os grupos formados podem ser identificados de forma subjetiva (AMARAL JÚNIOR e THIÉBAUT, 1999).

Encontra-se também na literatura o Método de Tocher Modificado (VASCONCELOS et al., 2007), que difere do original pelo fato de ser adotado critério de aglomeração inverso, de modo que o processo de agrupamento deixa de ser simultâneo para ser sequencial. Em relação ao método original, os autores afirmam a vantagem de, durante o processo de agrupamento, não mais haver influência dos indivíduos já agrupados.

Os procedimentos iniciais para formação de grupos propostos pelo método de Tocher modificado (sequencial) são os mesmos que o método de Tocher original. Para a formação dos demais grupos o procedimento é similar, porém, para Pereira et al. (2009), o melhor desempenho do método de Tocher modificado em relação ao método de Tocher original se deve ao fato de que no método sequencial indivíduos já agrupados não influenciam no agrupamento dos demais, ou seja, o melhor desempenho está no fato do método sequencial excluir as informações daqueles indivíduos anteriormente já

agrupados e assim sucessivamente. Com isso, à medida em que se realizam os agrupamentos, é reduzida a exigência no acréscimo médio da distância, proporcionando uma melhor formação dos grupos.

2.4 Coeficiente de Correlação Cofenético

O coeficiente de correlação linear de Pearson entre os elementos da matriz de dissimilaridade (matriz de distâncias entre os indivíduos, obtida a partir dos dados originais) e os elementos da matriz cofenética (matriz de distâncias entre os indivíduos, obtida a partir do dendrograma) é denominado Coeficiente de Correlação Cofenético (CCC). Esse coeficiente pode ser utilizado para avaliar a consistência do padrão de agrupamento de métodos de agrupamentos hierárquicos, sendo que valores próximos à unidade indicam melhor representação (BARROSO & ARTES, 2003; CRUZ & CARNEIRO, 2006).

O CCC avalia a consistência do agrupamento após a obtenção do dendrograma, pois após a formação do dendrograma pode ocorrer distorções entre os padrões de dissimilaridade dos indivíduos estudados, além de uma elevada simplificação das informações originais (EVERITT, 1993; CRUZ; CARNEIRO, 2006). Pode-se notar que o CCC equivale ao cálculo da correlação de Pearson entre a matriz de similaridade original e a matriz cofenética, obtida após a construção do dendrograma (MEYER, 2002).

O CCC é calculado por (BUSSAB et al., 1990):

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

em que c_{ij} é o valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz cofenética; s_{ij} é o valor de similaridade entre os indivíduos i e j, obtidos a partir da matriz de similaridade; $\bar{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} c_{ij}$; $\bar{s} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} s_{ij}$.

2.5 Critérios para seleção do número de agrupamentos

Uma questão de grande importância é de como se deve proceder na escolha do número final de grupos que define a partição do conjunto de dados analisado, ou em qual passo o algoritmo de agrupamento deve ser interrompido (MINGOTI, 2013).

A seleção do número ótimo de grupos na etapa final dos estudos que utilizam os métodos hierárquicos de agrupamento é uma tarefa difícil para os pesquisadores (DIAS, 2014). O critério mais simples utilizado para decidir qual o número de grupos a adotar é o corte do dendrograma pela análise subjetiva dos diferentes níveis do mesmo, o que torna esse procedimento naturalmente enviesado pelas necessidades e opiniões dos analistas e pesquisadores (MARTINS; PEDRO e ROSA, 2004).

De acordo com Milligan e Cooper (1985), ao aplicar um índice para determinar o número correto de grupos, podem ocorrer dois tipos de erros de decisão: o primeiro tipo de erro ocorre quando o índice indica a seleção de g grupos, sendo que na verdade há menos que g grupos no conjunto de dados. Já o segundo tipo de erro ocorre quando o índice indica menos grupos no conjunto de dados do que o real. Mesmo a severidade dos dois tipos de erros podendo mudar de acordo com o contexto do problema, tem-se que a ocorrência do segundo tipo de erro é considerado mais grave na maioria das análises, pois haverá perda de informação por fundir grupos distintos. Assim, os autores destacaram o fato de existirem poucas apresentações de métodos de comparação de desempenho de índices para determinar o número correto de grupos. Desse modo, Milligan e Cooper (1985) testaram um total de 30 medidas de validação de agrupamento visando determinar o número ideal de grupos para cada uma das medidas utilizadas no processo de agrupamento hierárquico, fazendo uso de dados artificiais com número de grupos conhecido (no intervalo de 2 a 5 grupos). Dentre os métodos avaliados, os que apresentaram melhor desempenho foram o índice de Calinski e Harabasz (CALINSKI e HARABASZ, 1974), e o índice de Duda e Hart (DUDA e HART, 1973), os quais serão descritos nas subseções 1.5.1 e 1.5.2.

Diversos autores na literatura destacam os resultados do estudo de Milligan e Cooper (1985). Segundo Everitt (2011), os dois melhores desempenhos no estudo de Milligan e Cooper (1985) foram as técnicas introduzidas por Calinski e Harabasz (1974) e Duda e Hart (1973) para o uso em dados contínuos. Timm (2007) também destaca o resultado obtido pelos autores de que o índice pseudo F é o mais útil para identificar o número de *cluster*s. Dias (2014) faz referência ao estudo dos autores em sua tese de

doutorado, onde utiliza o índice de Duda e Hart (1973) para obter o número ótimo de grupos para identificar modelos cujos avaliadores são semelhantes.

2.5.1 Estatística Pseudo *F*

Calinski e Harabasz (1974) sugerem para cada passo do agrupamento, o cálculo da estatística chamada Pseudo F, definida por:

$$F = \frac{SSB/(g^* - 1)}{SSR/(n - g^*)} = \left(\frac{n - g^*}{g^* - 1}\right)\left(\frac{R^2}{1 - R^2}\right)$$

em que SSB é a soma de quadrados total entre os g^* grupos da partição, dada por:

$$SSB = \sum_{i-1}^{g^*} n_i \, (\bar{X}_{i.} - \bar{X})'(\bar{X}_{i.} - \bar{X})$$

e SSR a soma de quadrados total dentro dos grupos da partição (Soma de Quadrados Residual), dada por:

$$SSR = \sum_{i=1}^{g^*} SS_i = \sum_{i=1}^{g^*} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_{i.}\right)' \left(X_{ij} - \bar{X}_{i.}\right)$$

R^2 é o coeficiente da partição, dado por $R^2 = \frac{SSB}{SST_c}$ em que SST_c é a soma de quadrados total corrigida para a média global em cada variável, dado por:

$$STT_c = \sum_{i=1}^{g^*} \sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}\right)' \left(X_{ij} - \bar{X}\right)$$

g^* é o número de grupos relacionado com a partição do respectivo estágio de agrupamento; n é o número de elementos amostrais; $X'_{ij} = \left(X_{i1j}\, X_{i2j}\, \ldots\, X_{ipj}\right)$ é o vetor de medidas observadas para o *j-ésimo* elemento amostral do *i-ésimo* grupo; $\bar{X}'_{i.} = \left(\bar{X}_{i1.}\, \bar{X}_{i2.}\, \ldots\, \bar{X}_{ip.}\right)$ é o vetor de médias do *i-ésimo* grupo; e $\bar{X}' = \left(\bar{X}_{.1}\, \bar{X}_{.2}\, \ldots\, \bar{X}_{.p}\right)$ é o vetor

de médias global, sem levar em conta qualquer posição, onde $\bar{X}_{.l} = \frac{1}{n}\sum_{i=1}^{g^*}\sum_{j=1}^{n_i} X_{ilj}\,, l = 1, 2 \ldots, p$.

Segundo os autores, se F é monotonicamente crescente com g^*, os dados sugerem que não existe qualquer estrutura natural de partição dos dados. Porém, se isso não ocorrer e a função F apresentar um valor de máximo, o número de conglomerados e a partição referente a esse valor máximo corresponderão à partição ideal dos dados.

2.5.2 Estatística Pseudo T^2

Foi proposta por Duda e Hart (1973) a estatística Pseudo T^2, sendo definida quando dois conglomerados se unem para formar um novo conglomerado. Se num determinado passo do agrupamento o conglomerado C_k é a união dos conglomerados C_i e C_l', então a estatística Pseudo T^2 é definida por:

$$Pseudo\ T^2 = \frac{B_{il}}{\left[\sum_{j\in C_i}\left\|X_{ij} - \bar{X}_{i.}\right\|^2 + \sum_{j\in C_l}\left\|X_{lj} - \bar{X}_{l.}\right\|^2\right](n_i + n_l - 2)^{-1}}$$

sendo $\left\|X_{kj} - \bar{X}_{k.}\right\| = \left[\left(X_{kj} - \bar{X}_{k.}\right)'\left(X_{kj} - \bar{X}_{k.}\right)\right]^{\frac{1}{2}}, k = i, l$

e B_{il} é definido como $B_{il} = \frac{n_i n_l}{n_i + n_l}(\bar{X}_{i.} - \bar{X}_{l.})'(\bar{X}_{i.} - \bar{X}_{l.})$, em que n_i é o número de elementos amostrais do conglomerado C_i e n_l é o número de elementos amostrais do conglomerado C_l.

A estatística Pseudo T^2 teria uma distribuição F com p e $(n_i + n_l - 2)$ graus de liberdade. Porém, na prática não se tem alocação aleatória devido aos critérios de agrupamento que são utilizados para a formação dos grupos. O valor da Pseudo T^2 é calculado em cada passo do algoritmo de agrupamento, e o valor de g correspondente ao ponto máximo, ou aquele imediatamente anterior, é escolhido como provável número de grupos da partição final (MINGOTI, 2013).

3. METODOLOGIA

Os dados em estudo são referentes as taxas de incidências anuais de casos de dengue nos 27 estados brasileiros, no período de 2000 a 2014, disponibilizadas no Sistema Nacional de Agravos de Notificação (SINAN), que registra os casos confirmados e suspeitos da doença. A incidência é dada pelo número de casos confirmados de dengue (clássica e febre hemorrágica), por 100 mil habitantes, sendo a ocorrência de casos relacionada à picada do mosquito *Aedes aegypti* infectado com o vírus do dengue, tipos 1, 2, 3 ou 4 (grupo dos flavivírus).

Foi realizada a análise descritiva dos dados utilizando estatísticas básicas, como: média, desvio padrão, mediana, coeficiente de variação e, por fim, obtido o gráfico boxplot para cada estado analisado. Feita a análise descritiva, partiu-se para a realização da Análise de Agrupamentos, que foi aplicada com o objetivo de analisar a similaridade entre os estados em relação à incidência dos casos de dengue, ou seja, avaliar quais estados possuem comportamento semelhante em relação à incidência da doença, além de descrever o perfil dos agrupamentos obtidos. Para isso as análises foram implementadas com o auxílio do software estatístico R (R Development Core Team, 2014).

A análise se inicia a partir de uma matriz de dissimilaridade, que é obtida por meio da medida de distância mais adequada, visto que os procedimentos em análise de agrupamento são influenciados pela natureza das variáveis ou dos atributos dos objetos. Deste modo, considerando a natureza dos dados em estudo, isto é, o fato dos dados apresentarem grandezas (escalas) distintas, foi utilizada no presente trabalho a distância Euclidiana Padronizada para a obtenção da matriz de distâncias.

Em seguida foram utilizados para a formação dos agrupamentos os métodos da Ligação Simples, Ligação completa, Ligação Média (UPGMA), Método do Centroide (UPGMC), Ligação mediana (WPGMC) e Critério de Ward. A sequência de fusão dos agrupamentos para cada método de ligação utilizado foi representada graficamente pelo dendrograma, que auxiliou na identificação dos agrupamentos dos estados brasileiros.

Para a comparação e escolha do método de ligação mais adequado aos dados, foi feito o diagnóstico por meio do CCC entre as matrizes e os agrupamentos (ROHLF E SOKAL, 1981), onde quanto maior a concordância entre os agrupamentos, maior deverá ser o respectivo CCC calculado (PINTO e CURI, 1991).

Quatro métodos foram utilizados para determinação do número ótimo de grupos no dendrograma, a saber: o método de Tocher original, o método de Tocher modificado, e os critérios Pseudo F e Pseudo T^2. Os critérios Pseudo F e Pseudo T^2 foram obtidos por meio do pacote NbClust (CHARRAD et al., 2015) do software estatístico R (R Development Core Team, 2014), que fornece os 30 índices presentes no estudo de simulação de Milligan e Cooper (1985) para determinar o número ótimo de grupos. A proposta do pacote é permitir ao pesquisador modificar simultaneamente o número de grupos, o método de agrupamento, a distância e os índices para decidir a melhor forma de agrupar as observações do seu conjunto de dados ou para comparar todos os índices ou métodos de agrupamento (DIAS, 2014).

Segundo Dias (2014), para decidir o número de grupos adequado para a situação em estudo, o pesquisador pode selecionar o número de grupos indicados pela maioria dos índices avaliados no estudo de Milligan e Cooper (1985) ou considerar apenas os índices de Calinski e Harabasz (CALINSKI e HARABASZ, 1974) e Duda e Hart (DUDA e HART, 1973), que foram os que apresentaram o melhor desempenho.

4. RESULTADOS E DISCUSSÕES

Inicialmente foi realizada a análise descritiva dos dados (Tabela 1) e, de acordo com os resultados obtidos, tem-se que o estado do Acre apresentou a maior incidência média de dengue no período em estudo. Já os estados que apresentaram as menores incidências médias da doença foram Santa Catarina e Rio Grande do Sul.

Em geral os estados brasileiros apresentaram alta variação em relação à incidência de dengue nesse período, o que pode ser observado pelos resultados dos coeficientes de variação obtidos. O estado do Rio Grande do Sul foi o que apresentou a maior variação, enquanto que os estados que apresentaram a menor variabilidade foram Pará e Piauí.

Por meio da Figura 2 é possível ver os gráficos boxplots dos 27 estados brasileiros em relação à incidência de dengue no período em estudo. Em grande parte dos estados foi verificado a presença de outliers, detectando o estado do Acre com a maior incidência do país.

Tabela 1: Estatísticas descritivas dos estados em relação à incidência de dengue.

Estado	Mínimo	Mediana	Média	Máximo	Desvio Padrão	Coeficiente de variação
Rondônia	109,4	263,4	395,1	1345,8	391,7	99%
Acre	39,0	354,3	1137,1	4793,3	1517,4	133%
Amazonas	19,3	143,9	286,9	1779,2	448,8	156%
Roraima	158,0	399,1	741,7	2248,8	645,7	87%
Pará	55,5	148,7	158,2	264,0	64,0	40%
Amapá	10,5	300,3	364,1	755,2	198,0	54%
Tocantins	122,8	417,3	471,6	957,0	285,4	61%
Maranhão	27,4	87,9	95,2	214,1	52,2	55%
Piauí	29,6	239,7	229,5	387,7	109,0	47%
Ceará	50,8	344,2	363,7	747,8	189,9	52%
Rio Grande do Norte	80,6	560,3	544,3	1331,5	356,8	66%
Paraíba	24,7	229,9	269,1	662,2	181,0	67%
Pernambuco	28,3	186,1	256,8	1235,7	295,4	115%
Alagoas	52,0	259,0	350,4	1517,6	385,2	110%
Sergipe	22,9	101,2	207,9	1065,6	269,7	130%
Bahia	34,4	238,2	256,1	683,4	194,7	76%
Minas Gerais	58,8	178,3	355,1	2021,3	523,4	147%
Espírito Santo	80,7	629,0	616,4	1771,0	451,4	73%
Rio de Janeiro	8,2	186,5	515,0	1691,9	588,6	114%
São Paulo	7,8	110,4	173,1	515,2	190,6	110%
Paraná	1,6	48,9	141,5	601,0	177,7	126%
Santa Catarina	0,3	1,5	2,0	5,4	1,6	80%
Rio Grande do Sul	0,2	1,2	3,8	34,1	8,5	223%
Mato Grosso do Sul	14,6	367,3	809,9	3051,8	1090,7	135%
Mato Grosso	88,9	348,8	543,5	1839,2	509,9	94%
Goiás	51,9	421,6	627,0	2165,9	632,9	101%
Distrito Federal	11,7	50,2	139,8	584,3	180,6	129%

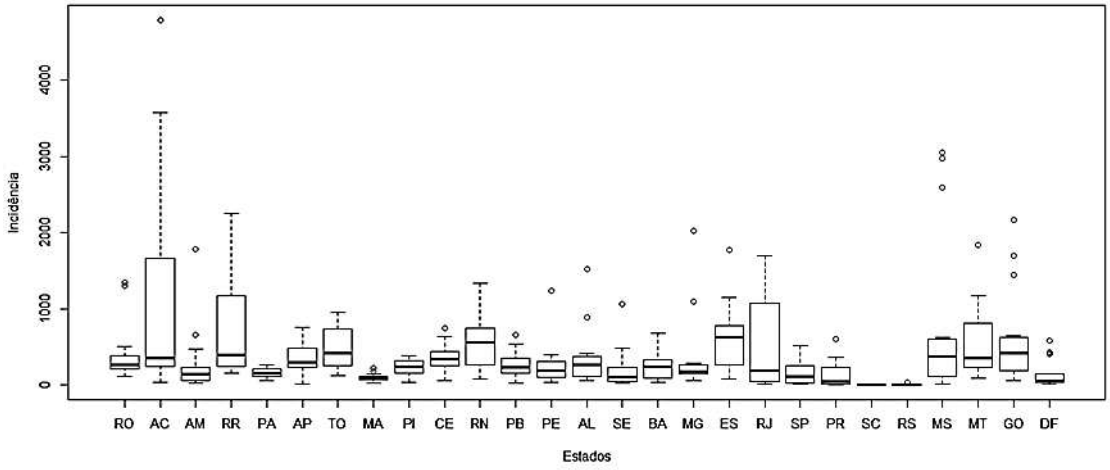

Figura 2: Gráficos boxplots dos estados brasileiros em relação à incidência de dengue.

Em seguida, a Análise de Agrupamentos foi aplicada com o intuito de agrupar os estados brasileiros de maior similaridade em relação à incidência de dengue no país. Considerando a natureza das variáveis agrupadoras, tem-se que a medida de distância que se mostrou mais adequada à análise foi a distância Euclidiana Padronizada, pois os dados apresentam a mesma escala, porém grandezas diferentes. Logo, devido a presença de grandezas diferentes, o uso da distância Euclidiana sem a padronização dos dados não apresenta bons resultados, o que justifica a não utilização dessa medida de distância.

Já a distância de Mahalanobis também não se aplica aos dados devido a matriz de correlação obtida entre as variáveis apresentar poucas correlações significativas, o que torna inviável a sua utilização, pois o uso dessa medida é adequado quando há um grau de correlação significativo entre os dados estudados (ALVES, 2012).

Assim, para a matriz de dissemelhança entre as variáveis foi utilizada a distância Euclidiana Padronizada, e para a formação dos agrupamentos foram aplicados os métodos Ligação simples, Ligação completa, Ligação Média, Método do Centroide, Ligação Mediana e Critério de Ward, cujos dendrogramas obtidos para cada método de agrupamento estão representados pela Figura 3.

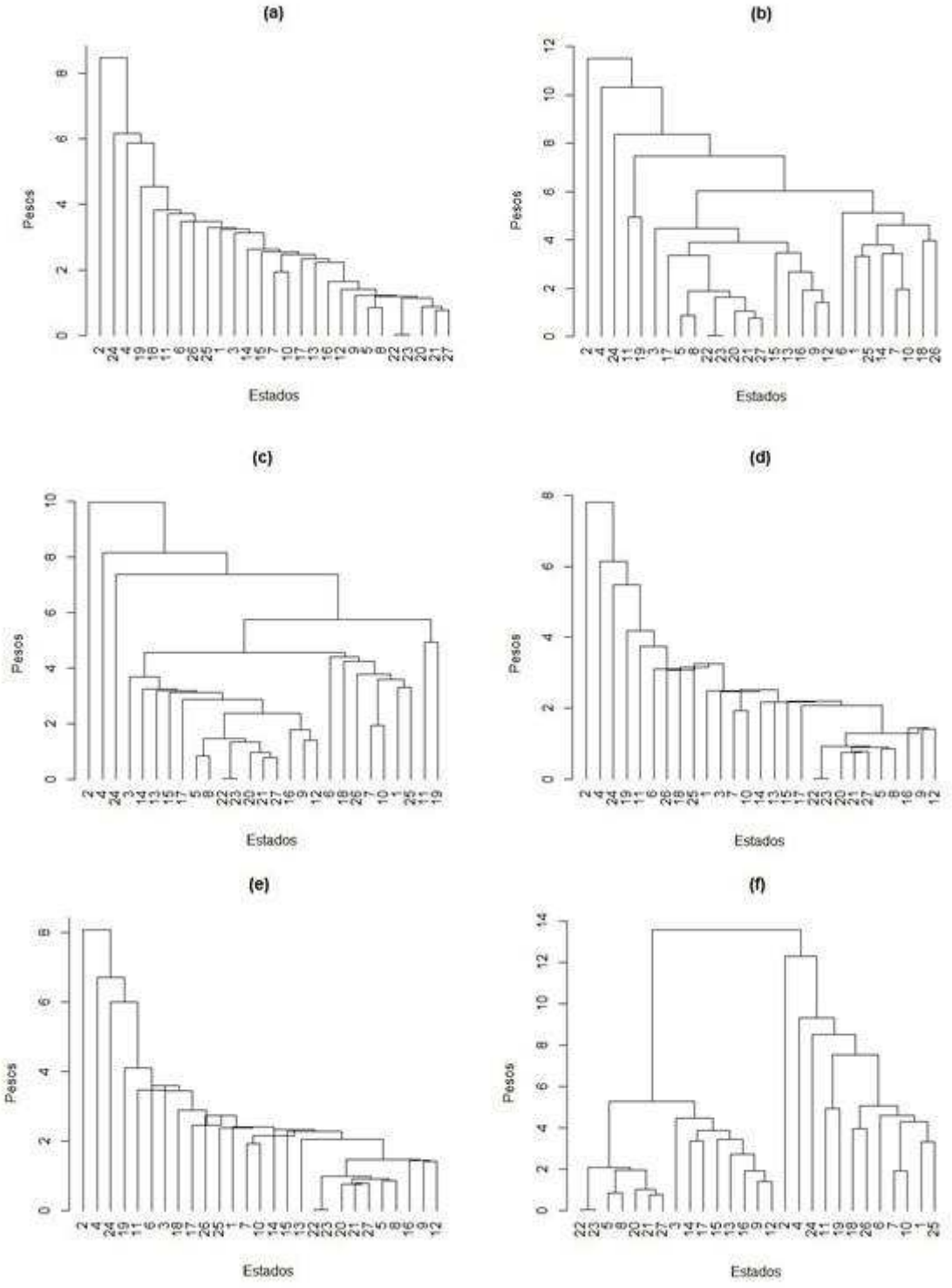

Figura 3: Dendrogramas obtidos pelos métodos de agrupamento: (a) Ligação simples; (b) Ligação completa; (c) Ligação Média (UPGMA); (d) Método do Centroide (UPGMC); (e) Ligação Mediana (WPGMC) e (f) Critério de Ward.

Para verificar o ajuste entre a matriz de dissimilaridade e os dendrogramas obtidos a partir de cada método de agrupamento, calculou-se o Coeficiente de Correlação Cofenético, segundo SOKAL e ROHLF (1962). O resultado indicou que o método que representou graficamente a matriz original com maior consistência foi o método da Ligação Média, como mostra a Tabela 2.

Tabela 2: Coeficiente de Correlação Cofenético para cada método de agrupamento.

Método de Ligação	**CCC**
Ligação Simples	0,9387
Ligação Completa	0,9356
Ligação Média	0,9556
Método do Centroide	0,9515
Ligação Mediana	0,9297
Critério de Ward	0,6199

A análise de agrupamentos tanto pelo Método de Otimização de Tocher original como também pelo Método de Otimização de Tocher Modificado (sequencial), possibilitou a formação de dois grupos (Tabela 3). O grupo I englobou o maior número de estados, totalizando 26 estados. O grupo II englobou apenas o estado do Acre.

De acordo os resultados obtidos por meio dos critérios Pseudo F e Pseudo T^2, tem-se que o critério Pseudo F indicou a formação de cinco grupos, enquanto que o critério Pseudo T^2 indicou a formação de dois grupos. Por meio do critério Pseudo T^2, houve novamente a discriminação apenas do estado do Acre no grupo II.

Tabela 3: Comparação dos critérios de determinação do número de grupos.

Critérios	**Número de grupos**
Tocher original (Rao, 1952)	2
Tocher modificado (Vasconcelos et al., 2007)	2
Pseudo F (Calinski e Harabasz, 1974)	5
Pseudo T^2 (Duda e Hart, 1973)	2

Para decidir o número de grupos adequado à situação em estudo, como houve divergência no número de grupos utilizando os dois índices que obtiveram melhor desempenho no estudo de simulação de Milligan e Cooper (1985), tem-se a escolha de dois grupos devido aos resultados obtidos pelo Método de Otimização de Tocher

original e Método de Otimização de Tocher Modificado (sequencial), que foram semelhantes ao resultado obtido pelo índice de Duda e Hart (1973).

Logo, a Análise de Agrupamentos, utilizando o método hierárquico aglomerativo da Ligação Média, resultou na formação de dois grupos, onde os estados brasileiros que constituem o mesmo grupo apresentam comportamento semelhante em relação à incidência de dengue no país, e se diferem dos estados que compõem os demais grupos. O grupo II é composto pelo estado do Acre, e o grupo I composto dos demais estados brasileiros, conforme mostram a Figura 4 e a Tabela 4.

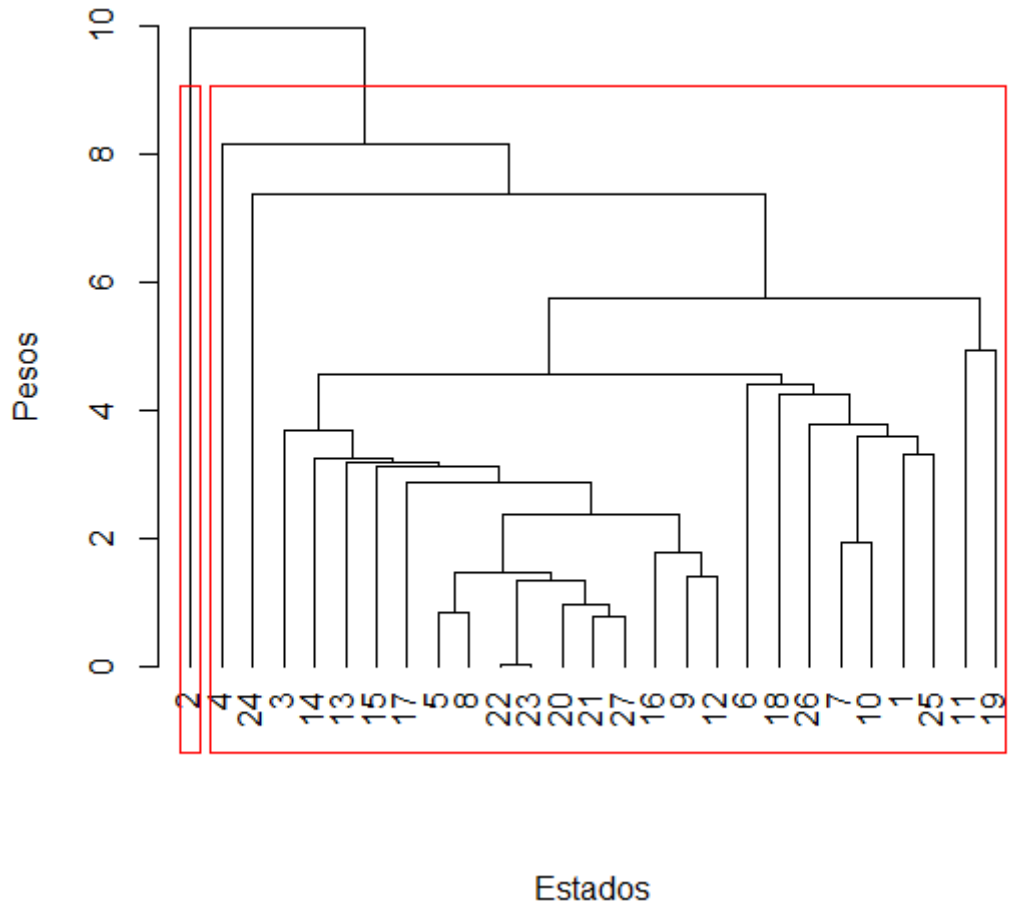

Figura 4: Dendrograma formado a partir da Distância Euclidiana Padronizada e Método Hierárquico da Ligação Média, com dois grupos.

O dendrograma retratou os dois grupos formados (Tabela 4), sendo o primeiro composto por 26 estados brasileiros com exceção do estado do Acre, que apresentou as maiores incidências médias de dengue em 2010 e 2013, enquanto que em 2004 foi verificada a menor incidência média da doença no período em estudo. Todos os anos apresentaram alta variabilidade em relação à incidência de dengue, sendo os anos 2000 e 2007 os que apresentaram as maiores variações, enquanto que em 2006 foi verificada a menor variação. As maiores incidências anuais foram verificadas em 2000, 2007, 2010 e 2013, como mostra a Figura 5.

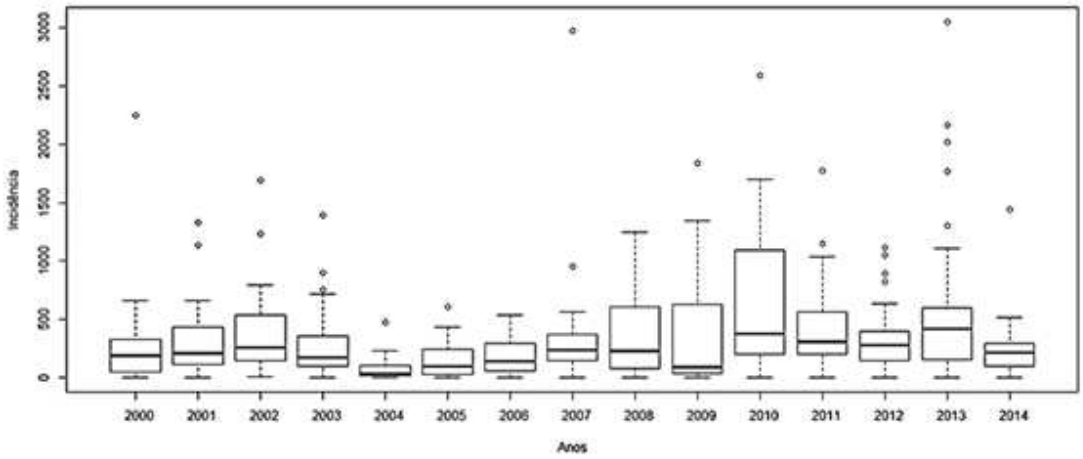

Figura 5: Gráficos boxplots referente à incidência de dengue nos anos em estudo representando os estados brasileiros pertencentes ao primeiro grupo.

Em relação ao segundo grupo formado, é possível perceber que o estado do Acre se destaca dos demais estados brasileiros em relação à incidência de dengue. O fato de o Acre ter se destacado em relação aos demais estados é devido aos altos índices de incidência que apresentou no período em estudo. O Acre foi o estado que obteve as maiores incidências do país em maior número de anos, apresentando a maior incidência nos anos de 2004, 2009, 2010, 2011 e 2014.

Tabela 4: Grupos obtidos por meio da Análise de agrupamento.

Grupo 1		Grupo 2
1 - Rondônia	15 - Sergipe	2 - Acre
3 - Amazonas	16 - Bahia	
4 - Roraima	17 - Minas Gerais	
5 - Pará	18 - Espírito Santo	
6 - Amapá	19 - Rio de Janeiro	
7 - Tocantins	20 - São Paulo	
8 - Maranhão	21 - Paraná	
9 - Piauí	22 - Santa Catarina	
10 - Ceará	23 - Rio Grande do Sul	
11 - Rio Grande do Norte	24 - Mato Grosso do Sul	
12 - Paraíba	25 - Mato Grosso	
13 - Pernambuco	26 - Goiás	
14 - Alagoas	27 - Distrito Federal	

Por meio da Figura 6, onde estão representados os estados brasileiros que apresentaram a maior incidência da doença em cada ano, percebe-se que o estado do Acre lidera em relação ao número de maiores incidências anuais registradas, sendo também o estado que apresentou a maior incidência registrada no país no período em estudo, com uma incidência de 4.793,3 casos por 100.000 habitantes em 2010.

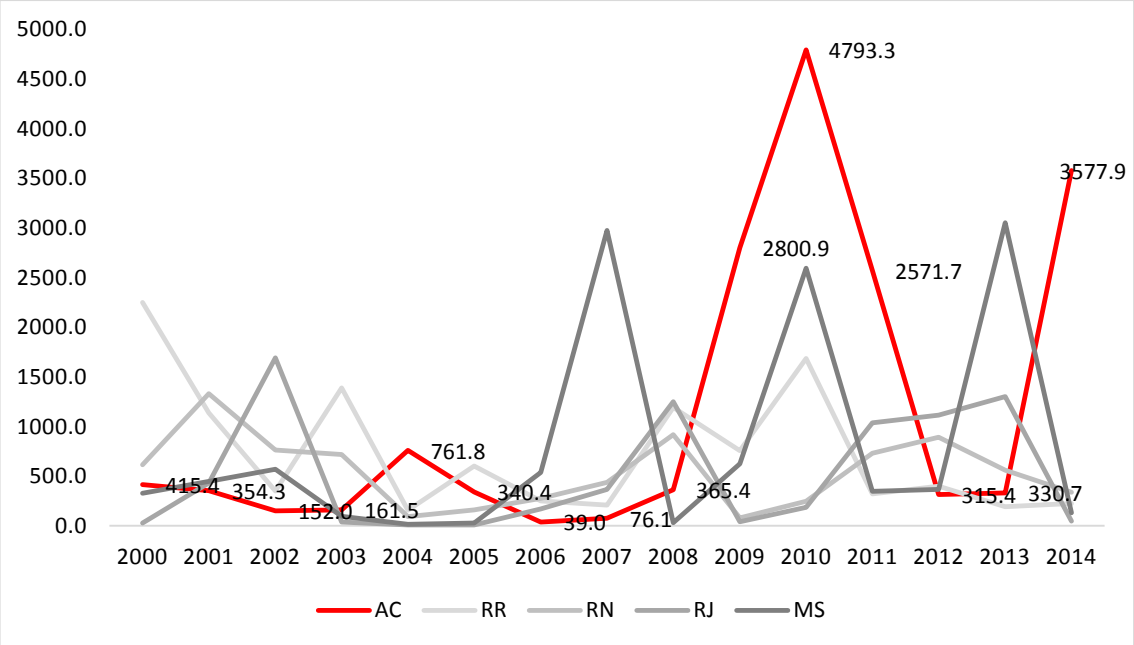

Figura 6: Gráfico representando as maiores incidências de dengue obtidas em cada ano.

Segundo Rocha (2010), a transmissão de dengue no estado do Acre vem ocorrendo desde 2000, sendo que é na capital Rio Branco, que concentra quase metade da população do estado, que é registrado atualmente a maioria dos casos da doença. Para o autor, a intensidade da epidemia verificada em 2004 se deu, possivelmente, pela introdução na capital do sorotipo DENV-3, detectada no período endêmico, sendo que nos anos anteriores foram detectados apenas os sorotipos DENV-1 e DENV-2 da doença. A entrada desse novo sorotipo ocasionou nos primeiros casos de febre hemorrágica e alguns óbitos.

De acordo com Rocha (2010), as atividades de pesquisas em criadouros realizada em Rio Branco mostraram que os depósitos mais infestados pelas larvas do vetor da dengue costumam ser os utilizados para armazenamento de água de uso doméstico, como tambores e caixas d'água, representando mais de 80% dos criadouros infectados. O hábito da população de estocar água nesses recipientes é comum devido às constantes intermitências e falta de abastecimento de água pelo setor público. Assim, a presença do vetor tem sido constante no município, apresentando índices elevados periodicamente.

Além disso, os programas nacional, estadual e municipal de controle da doença reconhecem haver oito municípios com transmissão autóctone de dengue, isto é, contraídos no próprio município, e 13 municípios infestados pelo Aedes aegypti (ROCHA, 2010).

As menores incidências registradas no Acre foram nos anos de 2006 e 2007, onde permaneceu entre os dez estados que registraram as menores incidências de dengue no país. Em uma área indene de transmissão da dengue, a entrada de um sorotipo resulta em epidemias aceleradas e explosivas onde, após o período de transmissão, demonstrado por um ciclo epidêmico bianual, ocorre o desaparecimento da doença e seu ciclo aparentemente se interrompe. Isso indica que a baixa incidência sucede devido à diminuição do número de pessoas que se tornaram susceptíveis à infecção pelos sorotipos causadores das epidemias anteriores, porém ficando susceptíveis às novas infecções causadas por outros sorotipos (GLUBER, 1997; HALSTED, 2006). Esse fato explica um dos possíveis motivos da baixa incidência de dengue verificada no estado do Acre nos anos de 2006 e 2007.

Por meio da Figura 3 é possível ver que após a epidemia da doença em 2004, uma nova epidemia é verificada no período de 2009 a 2011, onde o auge se deu em 2010. De acordo com o Informe Epidemiológico da Dengue disponibilizado pelo

Ministério da Saúde, dos casos notificados de dengue até a nona semana de 2010, o estado do Acre já liderava o ranking dos estados brasileiros com maior incidência da doença. Além disso, 35,4% do total de casos notificados no país estavam concentrados em seis municípios, dentre eles o município de Rio Branco.

O monitoramento de sorotipos circulantes ao longo de 2009 apontou para uma nova mudança no sorotipo predominante, com a recirculação do DENV-1. Segundo o Informe Epidemiológico, a recirculação do DENV-1 alerta para a possibilidade de grande circulação do vírus nos estados em que esse sorotipo se mostrou predominante, em virtude da população desses estados não estar em contato com o mesmo desde o início da década.

Segundo o Ministério da Saúde, a incidência de dengue do estado do Acre e do município de Rio Branco, no período de 2000 a 2010, seguiu o padrão observado na região Norte e no Brasil, com os ciclos de alta transmissão influenciados pela predominância de diferentes sorotipos no país: DENV-3 no período de 2001 a 2006, DENV-2 em 2007 a 2009, e a predominância de DENV-1 no ano de 2010. Logo, há indícios de que a recirculação e predominância do sorotipo DENV-1 na população do estado tenha sido um dos principais motivos para a ocorrência da epidemia da doença verificada no período de 2009 a 2011.

Após a epidemia de dengue no estado do Acre no período de 2009 a 2011, percebe-se pela Figura 4 que há uma redução na incidência da doença nos anos posteriores, com exceção de 2014, onde a incidência apresenta novamente um aumento significativo. De acordo com o Boletim Epidemiológico disponibilizado pelo Ministério da Saúde, tem-se que em 2014 o Acre liderou o ranking dos estados que apresentaram aumento no número absoluto de casos prováveis e incidência acima de 300 casos por 100 mil habitantes.

Segundo o Boletim Informativo Mensal de Dengue emitido pela Secretaria de Estado de Saúde do Acre, 77,8 % do total de casos notificados de dengue nas 12 primeiras semanas epidemiológicas de 2014 estão concentrados nos municípios Rio Branco e Cruzeiro do Sul, sendo confirmado no município de Cruzeiro do Sul os primeiros casos autóctones (contraídos do próprio município) de dengue.

Além disso, dentre os dez municípios que apresentaram maior registro de casos prováveis no país, somente Cruzeiro do Sul não apresentou redução dos casos a partir do mês de julho. Os resultados obtidos pelo LIRAa (Índice Rápido de Infestação por Aedes aegypti), em 2014, mostraram que Cruzeiro do Sul está entre os quatro

munícipios do estado do Acre que apresentaram situação de risco. Diante do exposto, tem-se que o surto de dengue e a consequente epidemia no município de Cruzeiro do Sul pode ter contribuído para o grande aumento da incidência de dengue no estado do Acre em 2014.

5. CONCLUSÃO

A Análise de Agrupamento aplicada às incidências de dengue nos estados brasileiros resultaram em dois grupos, discriminando o estado do Acre dos demais estados. O estado do Acre apresentou comportamento distinto em relação aos outros estados devido às altas incidências registradas da doença.

Dentre os possíveis motivos que ocasionaram nas epidemias verificadas no estado no período em estudo, estão a introdução e recirculação de sorotipos da doença, além do tratamento inadequado de recipientes que podem se tornar criadouros do vetor da doença. Logo, cabe ao estado do Acre melhorar o direcionamento das atividades de vigilância epidemiológica e de prevenção e controle da doença.

Para trabalhos futuros, está previsto a alteração no conjunto de dados, de modo que os novos dados correspondam a estudos dentro do estado de Minas Gerais e, posteriormente, outros estudos no âmbito do Triângulo Mineiro.

6. REFERÊNCIAS BIBLIOGRÁFICAS

ALVES, S. C. **Comparação de métodos para definição do número ótimo de grupos em análise de agrupamento.** 2012. Dissertação (Mestrado em Estatística Aplicada e Biometria) – Coordenação do Programa de Pós-Graduação de Estatística Aplicada e Biometria, Universidade Federal de Viçosa, Viçosa.

AMARAL JÚNIOR, A.T.; THIÉBAUT, J.T.L. **Análise multivariada na avaliação da diversidade em recursos genéticos vegetais.** Campos dos Goytacazes - Universidade Estadual do Norte Fluminense - UENF, CCTA, p. 55, 1999.

ANDERBERG, M. R. ***Cluster* analysis for applications**. New York: Academic Press, 1973.

BRASIL. **Saúde Brasil 2009: uma análise da situação de saúde e da agenda nacional e internacional de prioridades em saúde** / Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Análise de Situação de Saúde. Brasília: Ministério da Saúde, 2010.

BRASIL. **Informe Epidemiológico da Dengue: Análise de situação e tendências - 2010** / Ministério da Saúde, Secretaria de Vigilância em Saúde. Brasília: Ministério da Saúde, 2010.

BRASIL. **Sistema Nacional de Vigilância em Saúde: Relatório de situação: Acre** / Ministério da Saúde, Secretaria de Vigilância em Saúde. Brasília: Ministério da Saúde, 2011.

BRASIL. **Boletim Epidemiológico: Monitoramento dos casos de dengue e febre de chikungunya até a Semana Epidemiológica (SE) 47 de 2014** / Ministério da Saúde, Secretaria de Vigilância em Saúde. Brasília: Ministério da Saúde, 2014.

BRASIL. **Boletim Informativo Mensal de Dengue. Secretaria de Estado de Saúde do Acre/DVS/DVE.** Área técnica de vigilância epidemiológica da dengue/ATD. Ano III edição 02 de 2014.

BARROSO, L.P.; ARTES, R. **Análise multivariada**. Lavras: UFLA, 2003. 151 p.

BUSSAB, W. de O.; MIAZAKI, S. E.; ANDRADE, D. F. Introdução à análise de agrupamento. In: IX SIMPÓSIO BRASILEIRO DE PROBABILIDADE E ESTATÍSTICA, 1990, IME-USP São Paulo, 105 p.

CALINSKI, T.; HARABASZ, J. A Dendrite Method for *Cluster* Analysis. **Communications in Statistics**, v. 3, n. 1, p. 1-27, 1974.

CHATFIELD C.; COLLINS, A. J**. Introduction to multivariate analysis.** London: Chapman & Hall, 1986. 246 p.

CHARRAD, M., GHAZZALI, N., BOITEAU V.; NIKNAFS, A. **NbClust: An examination of indices for determining the number of *cluster*s.** Disponível em: <https://cran.r-project.org/web/packages/NbClust/NbClust.pdf>. Acesso em: 11 out. 2015.

CHIARAVALLOTI NETO, F.; BARBOSA, A. A. C.; CESARINO, M. B.; FAVARO, E. A.; MONDINI, A.; FERRAZ, A. A.; DIBO, M. R.; VICENTINI, M. E. Controle do dengue em uma área urbana do Brasil: avaliação do impacto do Programa Saúde da Família com relação ao programa tradicional de controle. **Cad. Saúde Pública**, Rio de Janeiro, v. 22, n. 5, p. 987-997, mai., 2006.

CORMACK, R.M. A review of classification. **Journal of the Royal Statistical Society**. Series A (General), v. 134, n. 3, p. 321-367, 1971.

CRUZ, C. D.; CARNEIRO, P. C. S. **Modelos Biométricos aplicados ao melhoramento genético.** Viçosa: UFV, 2006. 585 p.

CRUZ, C. D.; FERREIRA, F. M.; PESSONI, L. A. Biometria aplicada ao estudo da diversidade genética. Visconde do Rio Branco: **Suprema**, 2011. 620p.

DIAS, A. **Seleção multivariada e identidade de modelos não lineares para o crescimento e acúmulo de nutrientes em frutos de mangueira.** 2014. Tese (Doutorado em Estatística e Experimentação Agropecuária) – Coordenação do Programa de Pós-Graduação de Estatística e Experimentação Agropecuária, Universidade Federal de Lavras, Lavras.

DILLON; W. R.; GOLDSTEIN, M. **Multivariate Analysis – Methods and Applications**. New York: John Wiley & Sons, 1984.

DONI, M. V. **Análise de *cluster*: métodos hierárquicos e de particionamento.** São Paulo: Universidade Presbiteriana Mackenzie, 2004. Disponível em: < http://meusite.mackenzie.com.br/rogerio/tgi/2004Cluster.PDF>. Acesso em: 22 set. 2015.

DUDA, R. O.; HART, P. E. **Pattern classification and scene analysis**. New York: Wiley, 1973.

EVERITT, B.S. ***Cluster* analysis**. London: Heinemann Educational Books, 1993. 122 p.

EVERITT, B. S.; LANDAU, S., LEESE, M.; STAHL, D. ***Cluster* Analysis**. Chichester: Wiley, 2011. 330 p.

FERREIRA, B. J.; SOUZA, M. F. de M.; SOARES FILHO, A. M.; CARVALHO, A. A. Evolução histórica dos programas de prevenção e controle da dengue no Brasil. **Ciência & Saúde Coletiva**, vol. 14, n. 3, p. 961-972, 2009.

FREI, F. **Introdução à análise de agrupamentos: teoria e prática.** São Paulo: Editora Unesp, 2006. 112 p.

FREI, F., PRADO, B. B. A. Mortalidade devido a causas violentas no estado de São Paulo. 1994. Mimeografado. FRIEDMAN, H. P, RUBIN, J. "On some inavariant criteria for grouping data". **America Statistical Association Journal,** n. 30, p. 1 159-1 178, 1967.

FUNDAÇÃO NACIONAL DA SAÚDE. **Boletim Epidemiológico** / Ministério da Saúde. Brasília: Ministério da Saúde, 1999.

GLUBER, D. J**. Dengue and dengue hemorrhagic fever: its history and resurgense as a global public health problem.** In: Gluber D. J., Kuno G., eds. Dengue and dengue hemorrhagic fever. New York: CAB International, 1997.

GLUBER, D. J. **Dengue and Dengue Hemorrhagic Fever.** Cinical Microbiology Reviews, v. 11, n. 3, p. 480-496, July, 1998.

HALSTEAD, S. B. Dengue in the Americas and Southeast Asia: Do they differ? **Rev. Panam Salud Publica,** v. 20, n. 6, p. 407-415, 2006.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. N.J.: Prentice-Hall, 1992.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. New Jersey: Prentice Hall, 2002.

KHATTREE, R.; NAIK, D.N. **Multivariate data reduction and discrimination with SAS Software.** New York: BBU Press and John Wiley Sons Inc., 2000. 574p.

MANLY, B. J. F. Randomization and regression methods for testing for associations with geographical, environmental and biological distances between populations. **Research in Population Ecology**, v. 28, n. 2, p. 201–218, 1986.

MARDIA, K. V.; KENTK, J. T.; BIBBYB, J. M. **Multivariate analysis**. London: Academic Press, 1995. 518p.

MARTINS, M. do R. F. de O., SOFIA P., SOFIA, R. **Escolha do número de grupos e validação da solução em análise classificatória: da teoria à prática**. Disponível em: <http://run.unl.pt/handle/10362/7686>. Acesso em: 17 set. 2015.

MARZOCHI, K. B. F. **Dengue in Brazil – Situation, Transmission and Control – A Proposal for Ecological Control.** Mem. Inst. Oswaldo Cruz, Rio de Janeiro, v. 89, n. 2, p. 235-245, apr./jun. 1994.

MEYER, A. S. **Comparação de coeficientes de similaridade usados em análises de agrupamento com dados de marcadores moleculares dominantes.** 2002. 106 f. Dissertação (Mestrado em Agronomia) – Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba.

MILLIGAN, G. W.; COOPER, M. C. **An examination of procedures for determining the number of *cluster*s in a data set.** Psychometrika, v. 50, n. 2, p. 159-179, 1985.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada.** Belo Horizonte: Editora UFMG, 2013. 297p.

MOITA NETO J. M.; MOITA G. C. Uma introdução à análise exploratória de dados multivariados. **Química Nova**, v. 21, n. 4, p. 467-469, 1998.

PEREIRA. T. M.; ESPÓSITO, D. P.; SOUZA, A. O.; Couto, M. F.; CRUZ C. D. Estudo comparativo dos métodos de agrupamento otimizados de Tocher e Tocher modificado. In: XIII ENCONTRO LATINO AMERICANO DE INICIAÇÃO CIENTÍFICA E IX ENCONTRO LATINO AMERICANO DE PÓS-GRADUAÇÃO, 2009, Universidade do Vale do Paraíba. 2009, p. 1-3.

PINTO, F. G. & CURI, P. R. Mortalidade por neoplasias no Brasil (1980/1983/1985): agrupamento dos estados, comportamento e tendências. **Revista de Saúde Pública**, v. 25, n. 4, p. 276-281, 1991.

QUINTAL, G. **Análise de *cluster*s aplicada ao Sucesso/Insucesso em Matemática.** 2006. Dissertação (Mestrado em Matemática para Ensino) – Universidade da Madeira, Funchal.

R Development Core Team (2014). R: **A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. Disponível em: <http://www.rproject.org>. Acesso em: 20 out. 2014.

RAO, C. R. **An advanced statistical method in biometric research**. New York: Ed. John Wiley e Sons, 1952. 390 p.

RIBOLDI, J. **Análise de agrupamento "*Cluster* Analysis" e suas aplicações.** Piracicaba: ESALQ, 1986. 33 p.

ROCHA, R. da C. **Epidemiologia da dengue na cidade de Rio Branco-Acre, Brasil, no período de 2000 a 2007.** 2011. Tese (Doutorado em Ciências) – Coordenação do Programa de Pós-Graduação em Saúde Pública, Universidade de São Paulo, São Paulo.

SNEATH, P. H. A. Prejudice in bacterial classification. **Journal of general microbiology**, v. 17, 1957.

SOKAL, R. R.; ROHLF, F. J. The comparison of dendrograms by objective methods. **Taxon**, Berlin, v. 11, n. 1, p. 30-40, 1962.

SOKAL, R. T.; ROHLF, F. J. **Biometry - The principles and practice of statistics in biological research.** Ney York: W. H. Freeman & Company, 1981. 859 p.

TIMM, N. H. **Applied multivariate analysis**. Ney York: Springer, 2002. 695 p.

VALLI, M. Análise de *Cluster*. **Augusto Guzzo Revista Acadêmica**, São Paulo, v. 4, p. 77, 2002.

VASCONCELOS, E. S. de; CRUZ, C. D.; BHERING, L. L.; RESENDE JÚNIOR, M. F. R. Método Alternativo para Análise de Agrupamento. **Pesquisa Agropecuária Brasileira**, Brasília, v. 42, n. 10, p. 1421-1428, out. 2007.

VICINI, L.; SOUZA, A. M. **Análise multivariada da teoria à prática.** Santa Maria: Universidade Federal de Santa Maria, 2005. Disponível em: < http://w3.ufsm.br/adriano/livro/Caderno%20dedatico%20multivariada%20-%20LIVRO%20FINAL%201.pdf>. Acesso em: 23 set. 2015.

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