ANÁLISE DE AGRUPAMENTOS NO ESTUDO DOS INDICADORES DE MORTALIDADE NO BRASIL[1]

Priscila Neves Faria[2]
Anísio Pereira dos Santos Júnior[3]
Mirian Fernandes Carvalho Araújo[4]

Resumo. O Sistema de Informações sobre Nascidos Vivos e o Sistema de Informações sobre Mortalidade são as fontes de informação do Ministério da Saúde que possibilitam o monitoramento dos eventos vitais no Brasil e permitem a construção de indicadores de saúde de forma contínua para todo o país. Assim, objetivou-se com o presente trabalho analisar as informações de óbitos e aplicar metodologias de Estatística Multivariada na análise dos dados do Ministério da Saúde e do Sistema de Informações sobre Mortalidade a fim de obter respostas quanto à semelhança entre Estados da Federação no que diz respeito à mortalidade no ano de 2011. Para isso, foi utilizada a distância de Mahalanobis para a obtenção da matriz de distâncias e, para a formação dos grupos, o método aglomerativo do Centroid. O método de otimização utilizado para determinar o número ótimo de grupos no dendrograma foi o de Tocher original, que resultou na escolha de três grupos, onde dois grupos apresentaram estados que ficaram isolados, formando um grupo: São Paulo formou o primeiro grupo e o estado do Rio de Janeiro outro, possivelmente devido aos elevados números de óbitos em ambos os estados. Cerca de 32% dos óbitos nos dois estados estão relacionadas a causa doenças do aparelho circulatório. Tais motivos dessa elevada taxa estão associados principalmente aos fatores de risco como a obesidade, o fumo, hipertensão, diabetes, hipercolesterolemia, sedentarismo e estresse. Outro fator dessa porcentagem elevada pode estar relacionado à qualidade da assistência médica disponível.

Palavras-chave: Óbitos; Mahalanobis; Centroid; Tocher.

Resumen. Análisis de agrupamentos en el estudio de los indicadores de mortalidad en Brasil. El Sistema de Información sobre Nacidos vivos y el Sistema de Información sobre Mortalidad son las fuentes de información del Ministerio de Salud que posibilitan el monitoreo de los eventos vitales en Brasil y permiten la construcción de indicadores de salud de forma continua para todo el país. Así, se objetivó con el presente trabajo analizar las informaciones de óbitos y aplicar metodologías de Estadística Multivaria en el análisis de los datos del Ministerio de Salud y del Sistema de Información sobre Mortalidad a fin de obtener respuestas en cuanto a la semejanza entre Estados de la Federación en lo que se refiere a la mortalidad en el año 2011. Para ello se utilizó la distancia de Mahalanobis para la obtención de la matriz de distancias y para la formación de los grupos el método aglomerativo del Centroid. El método de optimización utilizado para determinar el número óptimo de grupos en el dendrograma fue el de Tocher original, que resultó en la elección de tres grupos, donde dos grupos presentaron estados que quedaron aislados, formando un grupo: São Paulo formó el primer grupo y el estado de Río de Janeiro otro, posiblemente debido a los elevados números de muertes en ambos estados. Cerca del 32% de las muertes en los dos estados están relacionadas con la causa de las enfermedades del aparato circulatorio. Tales motivos de esta elevada tasa se asocian principalmente a los factores de riesgo como la obesidad, el humo, la hipertensión, la diabetes, la hipercolesterolemia, el sedentarismo y el estrés. Otro factor de este porcentaje elevado puede estar relacionado con la calidad de la asistencia médica disponible.

Palabras clave: Las muertes; Mahalanobis; Centroid; Tocher.

[1] Agradecimento ao CNPq pelo auxílio financeiro durante a IC.

[2] Doutora em Estatística e Experimentação Agronômica pela ESALQ/USP. Docente da Faculdade de Matemática da Universidade Federal de Uberlândia. E-mail: priscilaneves@ufu.br.

[3] Discente do curso de Bacharelado em Estatística da Universidade Federal de Uberlândia. E-mail: juninho4322@gmail.com.

[4] Doutora em Estatística e Experimentação Agronômica pela Escola Superior de Agricultura Luiz de Queiroz. Docente da Faculdade de Matemática da Universidade Federal de Uberlândia. E-mail: mirian@ufu.br.

Abstract. Analysis of groups in the study of mortality indicators in Brazil. The Information System on Live Births and the Mortality Information System are the information sources of the Ministry of Health that allow the monitoring of vital events in Brazil and allow the construction of health indicators on an ongoing basis all country. Thus, the objective of this study was to analyze the information on deaths and to apply Multivariate Statistics methodologies in the analysis of data from the Ministry of Health and Mortality Information System in order to obtain answers regarding the similarity between States of the Federation with regard to mortality in the Year of 2011. For this, the distance of Mahalanobis was used to obtain the distance matrix and, for the formation of the groups, the Centroid agglomerative method. The optimization method used to determine the optimum number of groups in the dendrogram was that of the original Tocher, which resulted in the choice of three groups, where two groups presented states that were isolated, forming a group: São Paulo formed the first group and the state Of Rio de Janeiro, possibly due to the high numbers of deaths in both states. About 32% of deaths in both states are related to diseases of the circulatory system. Such reasons for this high rate are mainly associated with risk factors such as obesity, smoking, hypertension, diabetes, hypercholesterolemia, sedentary lifestyle and stress. Another factor of this high percentage may be related to the quality of available medical care.

Keywords: Deaths; Mahalanobis; Centroid; Tocher

1 Introdução

Nas últimas décadas, o Ministério da Saúde (MS) desenvolveu sistemas nacionais de informação sobre nascimentos, óbitos, doenças de notificação, atenção hospitalar, ambulatorial e básica, orçamento público em saúde e outros. Há ampla disponibilidade eletrônica desses dados, cada vez mais utilizados no ensino de saúde pública. O MS também promove investigações sobre temas específicos, ainda que de forma assistemática. Outras fontes relevantes para a saúde são os censos e pesquisas de base populacional do Instituto Brasileiro de Geografia e Estatística (IBGE), que cobrem aspectos demográficos e socioeconômicos. O mesmo se aplica aos estudos e análises do Instituto de Pesquisa Econômica Aplicada (IPEA), referentes a políticas públicas.

O Sistema de Informações sobre Nascidos Vivos (SINASC) e o Sistema de Informações sobre Mortalidade (SIM) são as fontes de informação do MS que possibilitam o monitoramento dos eventos vitais no Brasil e permitem a construção de indicadores de saúde de forma contínua para todo o país. De acordo com Mello-Jorge et al. (2002), o reconhecimento da importância de monitoramento das informações sobre óbitos e nascimentos junto à facilidade de acesso aos dados têm resultado no aumento substancial na cobertura e na qualidade das informações de ambos os sistemas. O estudo do número de óbitos por grupo de causas das doenças tem sua importância também devido aos problemas de saúde pública que assolam o país: em 2013, doenças cardiovasculares, neoplasias, doenças respiratórias crônicas e diabetes responderam por 79,4% dos óbitos registrados no Brasil (MS, 2014).

Neste sentido, a Análise Multivariada pode ser de grande importância na análise de dados deste tipo, uma vez que corresponde a um conjunto de métodos e técnicas que analisam

simultaneamente todas as variáveis na interpretação teórica do conjunto de dados. Dentro desta análise, tem-se a técnica de Análise de Agrupamentos, que é classificada como técnica de interdependência por se tratar de uma análise simultânea de todas as variáveis em estudo, na tentativa de se encontrar uma estrutura subjacente ao conjunto inteiro de variáveis, objetivo do presente estudo.

A Análise de Agrupamentos é uma metodologia da Estatística Multivariada que possibilita a criação de agrupamentos de itens diversos, de acordo com as semelhanças apresentadas por esses itens em relação a algum critério de seleção, determinado previamente pelo analista/pesquisador. O objetivo da ferramenta é o de classificar um pequeno número de grupos que tenham a característica de ser homogêneos internamente, heterogêneos entre si e mutuamente excludentes (HAIR JR. et al., 2005).

Tendo em vista o exposto acima, este trabalho tem como objetivo analisar as informações de óbitos e aplicar metodologias de Estatística Multivariada na análise dos dados do MS e do SIM, a fim de obter respostas quanto à semelhança entre Estados da Federação no que diz respeito à mortalidade.

2 Metodologia

2.1 Dados analisados

Os dados do presente estudo são referentes às taxas de óbitos por estado de causas determinadas no ano de 2011, extraídos da página do DATASUS (2012) que registra a participação relativa dos grupos de causas de mortalidade, em relação ao total de óbitos informados entre os que tiveram a causa. O método do cálculo é dado pelo número de óbitos de residentes, por causa ou grupo de causas determinadas dividido pelo número total de óbitos de residentes, por causas determinadas. O resultado é multiplicado por 100. Os grupos de causas analisados foram os dos capítulos da 10ª Revisão da Classificação Internacional de Doenças (CID-10).

A motivação para seleção do ano de 2011, como pode ser conferido no próprio site, é que esta foi a última atualização deste banco de dados no qual foi disponibilizado os indicadores de mortalidade por grupos de causas considerando cada unidade federativa do país.

Foi realizada a análise descritiva dos dados como média, desvio padrão, mediana, mínimo, máximo e coeficiente de variação para cada estado brasileiro. Feita a análise descritiva, foi realizada a Análise de Agrupamentos com o objetivo de analisar a semelhança entre os

estados com relação aos grupos de causas de mortalidade com o auxílio do software estatístico R (R DEVELOPMENT CORE TEAM, 2013).

A Análise de Agrupamentos partiu de uma matriz de dissimilaridade, obtida através de da medida de distância de Mahalanobis. De acordo com Quintal (2006), a distância de Mahalanobis 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 $D_{ij}^2 = \left(\overline{X}_i - \overline{X}_j\right)' \cdot S^{-1} \cdot \left(\overline{X}_i - \overline{X}_j\right)$ 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 S é a estimativa combinada da matriz da covariância/variância dentro dos grupos.

Após a construção da matriz de dissimilaridade foram aplicados vários métodos hierárquicos para a formação dos agrupamentos: método da Ligação Simples, Ligação Completa, Ligação Média, Método do Centroide e o Método de Ward. Cada método de ligação utilizado foi representado graficamente pelo dendrograma (diagrama bidimensional em forma de árvore), que auxiliou na identificação da formação dos grupos dos estados brasileiros. Para a escolha do método de ligação mais adequado, foi realizado o diagnóstico do Coeficiente de Correlação Cofenético (CCC) entre as matrizes e os agrupamentos (SOKAL; ROHLF, 1962), onde o método de ligação mais adequado para a análise é o que possuir maior CCC. Para determinar o número ótimo de grupos no dendrograma, foi utilizado o método de Tocher.

2.2 Mortalidade no Brasil

Desde fins da década de 1940, reiniciando-se nos anos 1970, alguns autores chamam a atenção para o fato de que, no Brasil, particularmente em áreas menos desenvolvidas, a situação do sub-registro de óbitos era alarmante. Nesse contexto, a mortalidade infantil, definida, conceitualmente, como o número de mortes para cada mil nascidos vivos, ganha um destaque especial por expressar as condições de vida e de saúde, o acesso aos serviços de saúde, e o desempenho dos programas dirigidos à sua redução (FRIAS; SZWARCWALD; LIRA, 2011). O MS dispõe de dois sistemas de informações para o cálculo do Coeficiente de Mortalidade Infantil (CMI): o SIM, que foi implantado em 1975 e o SINASC, a partir de 1990.

Os dados do SINASC (Sistema de Informações sobre Nascimentos) medem a participação relativa dos grupos de causas de mortalidade, em relação ao total de óbitos informados entre os que tiveram a causa determinada. Além disso, informam, por exemplo, que proporções elevadas de óbitos como as de doenças infecciosas e parasitárias, estão em geral associadas a precárias condições socioeconômicas da população. O SIM capta informações sobre as características sociais, demográficas e epidemiológicas dos óbitos, possibilitando o

monitoramento, e um maior detalhamento da mortalidade e seus determinantes para diversos níveis de agregação geográfica.

A Declaração de Óbito (DO) é o instrumento oficial de coleta de dados do SIM e deve ser preenchida pelo médico. O fluxo de encaminhamento dessas declarações, as normas quanto ao seu preenchimento e o processamento das informações são definidos pelo MS e estão detalhados nos manuais de procedimentos e de preenchimento da DO. O MS detém a gestão nacional do sistema e é responsável pela consolidação e divulgação dos dados. Estes são disponibilizados por município brasileiro, desde 1979, por meio de consulta ao site do DATASUS (http://tabnet.datasus.gov.br), de onde serão retirados os dados referentes ao ano de 2011 para o presente estudo.

2.3 Análise Estatística Multivariada

A técnica multivariada de Análise de Agrupamentos (também conhecida como Análise de Clusters - AC) é uma técnica que tem como objetivo básico descobrir os agrupamentos naturais das variáveis, onde estes são feitos com base nas similaridades ou dissimilaridades (caracterizadas por diversas formas de cálculo de "distâncias"). Esse método 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). É uma maneira de se obter grupos homogêneos após a aplicação de alguma medida de distância, por um esquema que possibilite reunir os dados em questão em um determinado número de grupos, de modo que exista grande homogeneidade dentro de cada grupo e heterogeneidade entre eles (JONHSON; WICHERN, 1992; CRUZ; CARNEIRO, 2003). Além de possibilitar a construção de grupos de acordo com as similaridades dos indivíduos, a análise de agrupamento possibilita também representá-los de maneira bidimensional, por meio de um dendrograma (diagrama em foram de árvore) (MOITA NETO; MOITA, 1998).

As medidas de distâncias (ou de dissimilaridade) são utilizadas para a representação dos pontos na estrutura de similaridade. Uma das mais conhecidas é a Distância Euclidiana, que representa o menor espaço entre dois pontos, sendo uma extensão do teorema de Pitágoras para o caso multidimensional. O termo dissimilaridade surgiu em função da relação da distância entre dois pontos P e Q, definida como d (P,Q), pois, à medida que ela cresce, diz-se que a divergência entre os pontos (unidades amostrais) P e Q aumenta, ou seja, tornam-se cada vez mais dissimilares. Os valores de distâncias são geralmente obtidos a partir de informações de "n" unidades amostrais, mensurados em relação a "p" caracteres (variáveis). A Análise de

Agrupamentos é iniciada ao organizar a matriz de dados (**Tabela 1**), onde as linhas são os indivíduos e as colunas são as variáveis, sendo que o número de indivíduos deverá ser maior do que o número de variáveis (n > p).

Tabela 1 - Matriz de dados de n indivíduos e p variáveis

Indivíduo	**Variável**					
	X_1	X_2	...	X_j	...	X_p
1	X_{11}	X_{12}	...	X_{1j}	...	X_{1p}
2	X_{21}	X_{22}	...	X_{2j}	...	X_{2p}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
i	X_{i1}	X_{i2}	...	X_{ij}	...	X_{ip}
⋮	⋮	⋮	⋯	⋮	⋯	⋮
n	X_{n1}	X_{n2}	...	X_{nj}	...	X_{np}

Fonte: Elaborado pelos autores.

De acordo com Cruz, Ferreira e Pessoni (2011), é necessário especificar um coeficiente de semelhança que indique a proximidade entre os indivíduos sendo importante considerar, em todos os casos semelhantes a este, a natureza da variável (discreta, contínua, binária) e a escala de medida (nominal, ordinal, real ou razão).

2.4 Medidas de Dissimilaridade

A "distância euclidiana", como uma das principais medidas de dissimilaridade, preconiza que a distância entre duas observações (*i* e *j*) corresponda à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações (i e j), para todas as p variáveis (FÁVERO et al., 2009), conforme a fórmula $d_{ij} = \sqrt{\sum_j \left(Y_{ij} - Y_{i'j}\right)^2}$, considerando Y_{ij} a observação no *i*-ésimo indivíduo para a *j*-ésima variável.

Na definição das variáveis, espera-se que elas apresentem contribuição equivalente na análise de agrupamento, de modo que a distância entre as amostras não seja alterada com a adoção de características com unidades de medidas distintas, num mesmo conjunto de dados, de forma que as variáveis admitidas apresentem poder discriminatório semelhante, não baseado na amplitude de seus valores. Caso uma determinada característica apresente uma maior amplitude em seus valores, em comparação às demais, ela terá um maior peso na análise.

Conforme descrito por Johnson e Wichern (1998), a Distância Euclidiana é insatisfatória para muitas situações estatísticas. Isso ocorre devido à contribuição de cada coordenada ter o mesmo peso para o cálculo da distância. Segundo Fávero et al. (2009, p. 198), "A maior parte

das medidas de distância sofre influência das diferentes escalas ou magnitudes das variáveis de similaridade".

De acordo com Carvalho (2007), para sanar essa limitação da distância faz-se necessária a normalização dos dados, que consiste em fazer com que os dados tenham a mesma ordem de grandeza. Por questão de simplicidade, por meio dessa normalização os dados ficam todos no intervalo [0;1]. Assim, se todos os dados estiverem num mesmo padrão de medida, as variabilidades de cada característica serão homogêneas ou quase homogêneas, podendo neste caso, utilizar os dados originais, isto é, sem realizar a normalização. A normalização transforma os dados das variáveis originais em escores padrão (também denominados de escores Z), de maneira a apresentar média 0 (zero) e desvio padrão 1 (um), conforme a fórmula $Z_j = \frac{Y_j - \bar{Y}}{\sigma_j}$, em que σ_j é a estimativa do desvio-padrão e $\bar{Y}$ é a estimativa da média, ambos associados à *j*-ésima variável.

A distância Euclidiana cresce de acordo com o número de variáveis, e quanto maior for esse número, maior será o valor da distância calculada. Para resolver isso, o valor da distância Euclidiana foi dividido pela raiz quadrada do total de variáveis. A este método dá-se o nome de distância Euclidiana Média, que é calculada por $\Delta_{ik} = \frac{1}{\sqrt{p}} d_{ik}$.

A chamada Distância Euclidiana Quadrada consiste exatamente na Distância Euclidiana, porém sem extração da raiz quadrada: "A Distância Euclidiana Quadrada tem a vantagem de que não é necessário calcular a raiz quadrada, o que acelera o tempo de computação, e é a distância recomendada para o método de agrupamento Ward" (HAIR JR. et al., 2005, p. 394).

Uma das principais medidas de distância é a de Mahalanobis, pois ela é muito rica em informações por trabalhar com a correlação entre as variáveis. A vantagem dessa distância é que ela evita problemas de escala das variáveis. Para saber se existe ou não correlação entre as variáveis, aplica-se o teste de Bartlett, usado para testar a hipótese nula de que a matriz de correlações é igual à matriz de identidade (BARTLETT, 1937). Se o resultado do teste for significativo, as variáveis são correlacionadas. Sua distância é calculada por $d_{ik} = \sqrt{(X_i - X_k)' S^{-1} (X_i - X_k)}$ em que S é a matriz de covariância das variáveis, X_i é a diferença da variável *i* dos indivíduos *i* e *j*, e X_k é a diferença da variável k dos indivíduos *i* e *j*. Caso os dados estiverem padronizados, a matriz S de covariância é igual à matriz de correlação de Pearson dos dados não padronizados.

Após ser calculada todas as possíveis distâncias entre os indivíduos, forma-se uma matriz denominada matriz de distâncias, onde as linhas e as colunas são enumeradas do indivíduo i_1 até o indivíduo i_n, de modo que sua diagonal principal apresente todos os valores iguais à zero, pois a distância das variáveis de um indivíduo com ele mesmo é zero. Além disso, sabe-se que a distância do indivíduo 2 com o indivíduo 1 é a mesma do indivíduo 1 com o indivíduo 2, ou seja, forma-se uma matriz simétrica.

2.5 Métodos Hierárquicos de agrupamento

Basicamente, os Métodos Hierárquicos dividem os indivíduos em grupos sendo este processo repetido até a formação do gráfico conhecido como Dendrograma. Existem vários métodos de agrupamentos hierárquicos, onde cada um formará um tipo diferente de agrupamento. Os mais comuns e disponíveis na maioria dos softwares estatísticos são Ligação Simples, Ligação Completa, Ligação Média, Centroides e Ward.

No método da Ligação Simples, a similaridade entre dois conglomerados é definida pelos dois elementos mais parecidos entre si (MINGOTI, 2005). Já 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).

A Ligação Média entre grupos, também conhecida por *Unweighted Pair-Group Method using the Average* (UPGMA), 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, 2003). 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.

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. De acordo com Mingoti (2005), 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.

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. Cada elemento é considerado como um único conglomerado, e em cada passo do algoritmo de agrupamento é calculada a soma de quadrados dentro de cada conglomerado.

2.6 Coeficiente de Correlação Cofenético (CCC)

É muito difícil saber qual dos Métodos Hierárquicos é mais adequado para usar no conjunto de dados, pois cada um gera um tipo de agrupamento diferente uma vez que após a formação do gráfico dendrograma podem ocorrer distorções entre os padrões de dissimilaridade dos indivíduos estudados, além de uma elevada simplificação das informações originais (EVERITT, 1993). Desta forma, é importante que seja efetuada alguma medida de ajuste entre a matriz de distâncias (matriz fenética) com a matriz resultante do processo de agrupamento (matriz cofenética). De acordo com Sokal e Rohlf (1962), o CCC mede esse ajuste entre a matriz de dissimilaridade e a matriz resultante da simplificação proporcionada pelo método de agrupamento. Assim, o CCC avalia a consistência do agrupamento por meio da obtenção do gráfico dendrograma. Para a comparação e escolha do método de agrupamento mais adequado aos dados, foi aplicado o cálculo do CCC para cada método de agrupamento realizado.

Sokal e Rohlf (1962) criaram o CCC usando a ideia da correlação de Pearson. O CCC efetua medidas do grau de ajuste entre a Matriz Fenética (matriz de distâncias) com a Matriz Cofenética (matriz obtida por meio do Dendrograma). Esse grau de ajuste é dado em porcentagem, e quanto maior for essa porcentagem, maior será a consistência dos dados. Vários autores recomendam que o CCC deve ser maior ou igual a 70%. Caso isso não ocorra, o ideal é partir para outro Método Hierárquico. O CCC é dado pela seguinte expressão:

$$r_{mn} = \frac{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(C_{jj'} - \bar{C}\right)\left(f_{jj'} - \bar{f}\right)}{\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(C_{jj} - \bar{C}\right)^2}\sqrt{\sum_{j=1}^{n-1}\sum_{j'=j+1}^{n}\left(f_{jj} - \bar{f}\right)^2}}$$

em que $\bar{C}$ e $\bar{f}$ são as médias aritméticas, definidas por $\bar{C} = \frac{\sum_{j=1}^{n} C_i}{n}$ e $\bar{f} = \frac{\sum_{i=1}^{n} f_i}{n}$.

2.7 Métodos de Otimização

Nos métodos de Otimização os grupos serão formados pela adequação de algum critério de agrupamento. Um método bastante utilizado é o método de otimização de Tocher. A partir da matriz de distância, identifica-se o par mais próximo, formando-se o primeiro grupo. Simultaneamente, obtém-se θ, que é o maior valor dentre os grupos formados a partir do método de agrupamento escolhido, isto é, θ é o valor máximo da medida de dissimilaridade encontrado no conjunto das menores distâncias envolvendo cada indivíduo. A inclusão, ou não, do indivíduo *k* no grupo é, então, feita considerando:

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

Se $\frac{d_{(ij)k}}{n} > \theta$, não se inclui o indivíduo K no grupo (*ij*);

em que n é o número de indivíduos que constitui o grupo original. Além disso, a distância entre o indivíduo k e o grupo formado pelos indivíduos ij é dada por $d_{(ij)k} = d_{ik} + d_{jk}$.

Esses indivíduos formarão o primeiro grupo e a partir dele é avaliada a possibilidade de inclusão de novos indivíduos no grupo, adotando o critério de que a distância média intragrupo deve ser menor que a distância média intergrupo (CRUZ; CARNEIRO, 2003). Assim, o método de otimização de Tocher se baseia na identificação do par mais similar dentro da matriz de dissimilaridade, isto é, aquele com menor estimativa de distância.

2.8 Dendrograma

Utilizando todas as variáveis disponíveis e, depois de aplicado algum método hierárquico à matriz de distâncias, os agrupamentos são representados de maneira bidimensional por meio de um dendrograma. Nele estão dispostas linhas ligadas segundo os níveis de similaridade, que agrupará pares de indivíduos ou de variáveis segundo Everitt (1993).

O dendrograma ilustra as fusões ou partições efetuadas em cada nível sucessivo do processo de agrupamento, no qual o eixo das abscissas representa os indivíduos e o eixo das ordenadas as distâncias obtidas após a utilização de uma metodologia de agrupamento. Os ramos da árvore fornecem a ordem das (n-1) ligações, em que o primeiro nível representa a primeira ligação, o segundo a segunda ligação, e assim sucessivamente, até que todos se juntem.

3 Resultados e Discussões

Inicialmente foi realizada a análise descritiva dos dados (**Tabela 2**) em relação ao total de óbitos por grupos de causas determinadas no Brasil, a saber: (C.4.a) Doenças infecciosas e parasitárias, (C.4.b) Neoplasias, (C.4.c) Doenças do aparelho circulatório, (C.4.d) Doenças do aparelho respiratório, (C.4.e) Afecções originadas no período perinatal, (C.4.f) Causas externas e (C.4.g) Demais causas definidas. De acordo com os resultados obtidos, tem-se que o número médio de óbitos por doenças do aparelho respiratório apresentou o maior valor em 2011. Já os óbitos por afecções originadas no período perinatal foi a causa que apresentou o menor número médio de óbitos. Em geral, os estados brasileiros apresentam alta variação em relação ao número de óbitos nesse período, o que pode ser observado pelos resultados dos coeficientes de

variação de obtidos. A causa doença do aparelho respiratório foi o que apresentou maior variação, enquanto a causa externa apresentou menor variação.

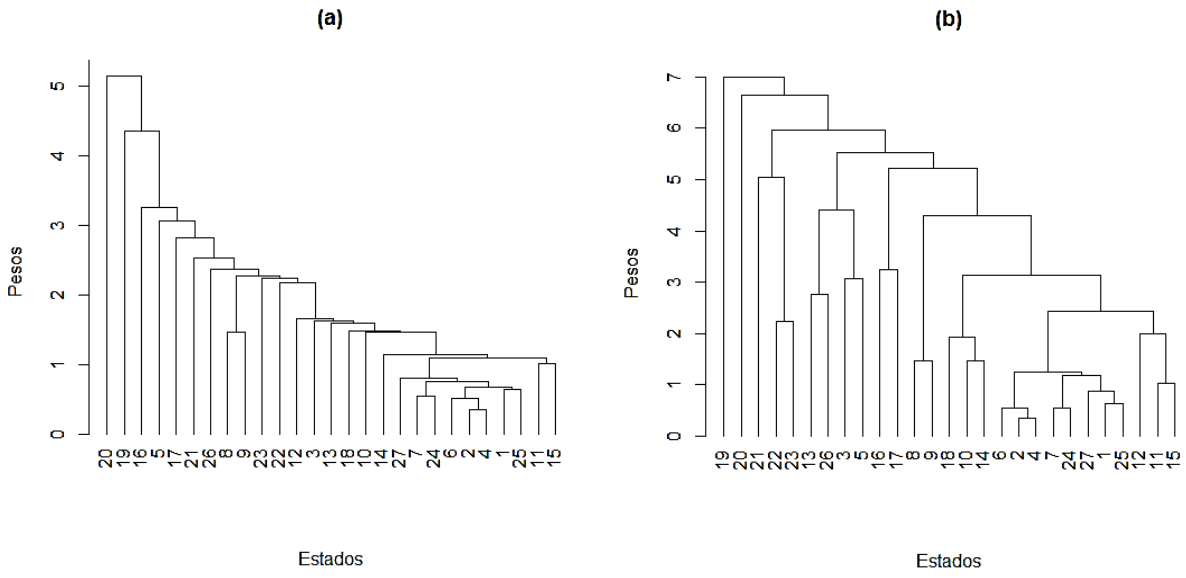
Tabela 2 - Estatística descritiva dos dados em relação aos tipos de causas determinadas de óbito

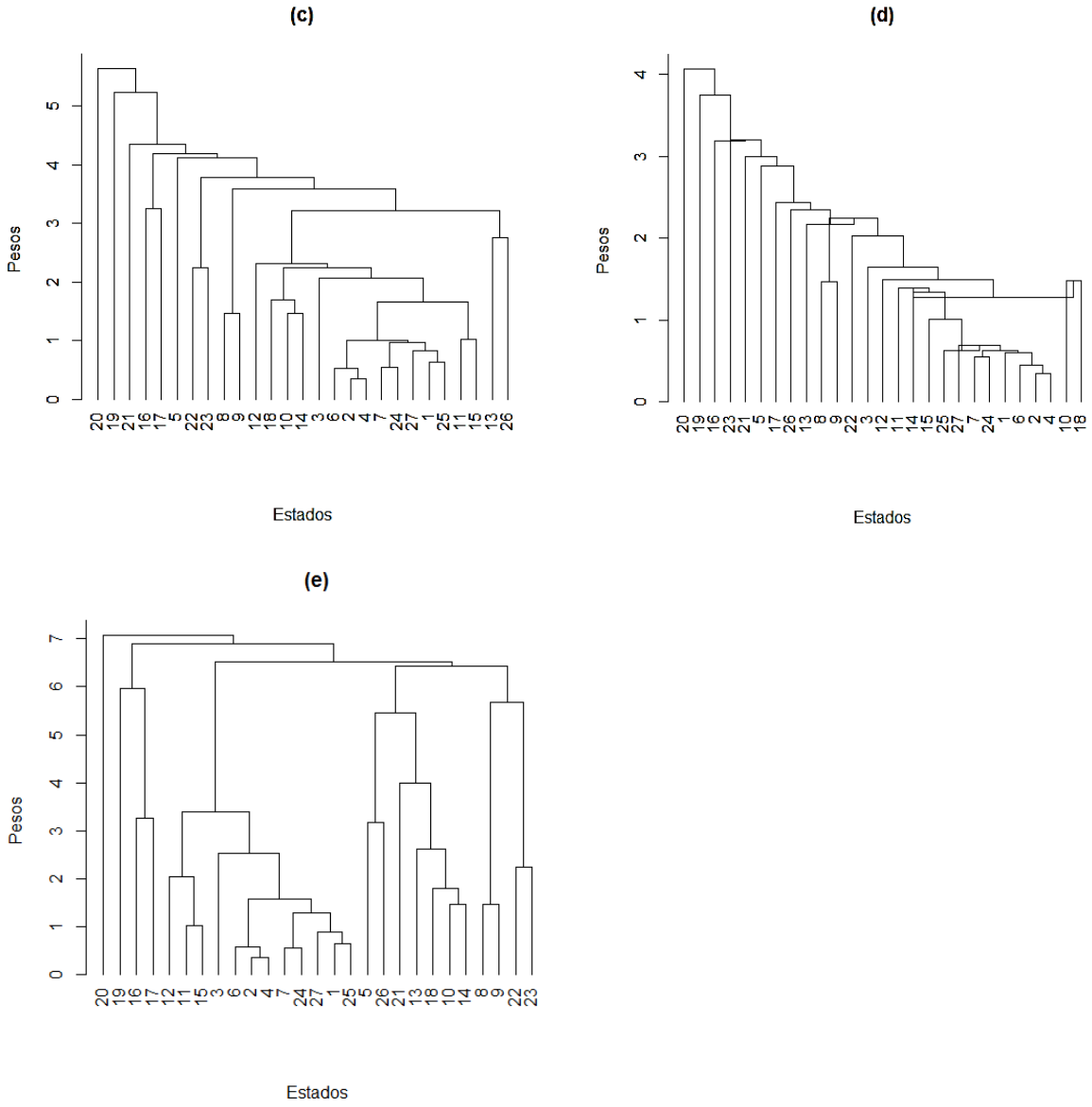
	Mínimo	Média	Mediana	Máximo	Desvio Padrão	Coeficiente de Variação (%)
C.4.a	88	1821,3	810	10409	2322,7	127,5
C.4.b	218	6829	3203	47276	9827,7	143,9
C.4.c	380	12415,3	6461	81182	16826,2	135,5
C.4.d	120	4692,3	1864	34679	7130,9	152
C.4.e	68	873,3	551	4164	890,5	102
C.4.f	350	5401,5	3572	24276	5436,2	100,6
C.4.g	321	8416,63	4317	52282	10989,5	130,6

Fonte: Elaborado pelos autores.

Em seguida, a análise de agrupamento foi aplicada com o intuito de agrupar os estados brasileiros de maior similaridade em relação ao número de óbitos por grupos de causas determinadas. A medida de distância que se mostrou mais adequada à análise foi a distância de Mahalanobis (verificação feita pelo teste de Bartlett entre as variáveis, que apresentou um p-valor significativo p< 0,0001). Foram aplicados os métodos Ligação Simples, Ligação Completa, Ligação Média, Método do Centroide e Método de Ward, sendo que os dendrogramas obtidos para cada método de agrupamento estão representados na **Figura 1**.

Figura 1 - Dendrogramas obtidos pelos métodos de agrupamento: (a) Ligação Simples; (b) Ligação Completa; (c) Ligação Média (UPGMA); (d) Método do Centroid (UPGMC); e (e) Critério de Ward





Fonte: Elaborado pelos autores.

Cada dendrograma obtido possui suas particularidades devido a cada método de ligação aplicado. O primeiro dendrograma formou longas cadeias (encadeamento), característica observada quando se aplica o Método da ligação Simples. O segundo dendrograma formou grupos pequenos que depois foram aglutinados para formar grupos maiores, tendência em geral observada quando é utilizado o método da ligação completa.

De acordo com Mingoti (2005), os diferentes métodos de ligação e combinações entre as medidas de similaridade ou dissimilaridade levam a padrões de agrupamento distintos. Deste modo, o ideal é utilizar vários métodos e comparar os resultados para que a análise dos dados seja realizada pela técnica mais adequada. Conforme Valentin (2000), um método é mais adequado que outro quando o dendrograma provê uma imagem menos distorcida da realidade. É possível avaliar o grau de deformação provocado pela construção do dendrograma obtendo-

se o CCC, segundo Sokal e Rohlf (1962). O menor grau de distorção será refletido pelo maior coeficiente cofenético, sendo que valores próximos à unidade indicam melhor representação (CRUZ e CARNEIRO, 2003). Como mostra a **Tabela 3**, o resultado indicou que o método que representou graficamente a matriz original com a maior consistência foi o do Centroid.

Tabela 3 - Coeficiente de Correlação Cofenético para cada Método de agrupamento

Método de Ligação	**CCC**
Ligação Simples	0.8567
Ligação Completa	0.8318
Ligação Média	0.9016
Método do Centroid	0.9187
Método de Ward	0.7228

Fonte: Elaborado pelos autores.

A análise de agrupamentos pelo Método de Otimização de Tocher possibilitou na formação de 3 grupos. Como mostra a **Tabela 4**, o grupo I englobou o maior número de estados, totalizando 24 estados e o Distrito Federal. O grupo II englobou apenas o Rio de Janeiro e o grupo III somente São Paulo.

Tabela 4 - Grupos obtidos por meio da análise de agrupamento

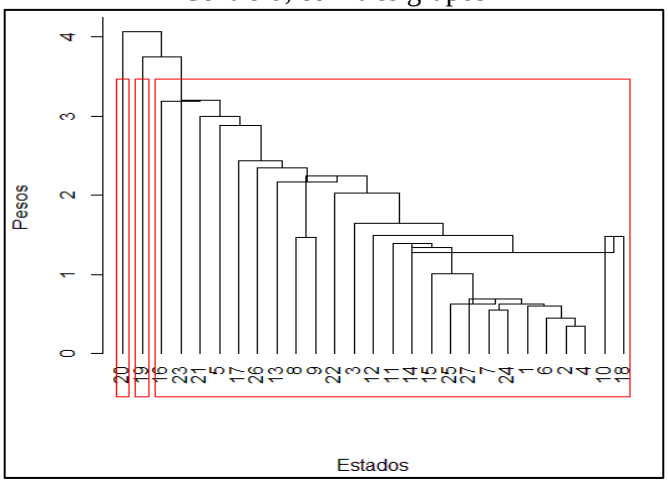
Grupo I		**Grupo II**	**Grupo III**
1 – Rondônia	14 – Alagoas	19 – Rio de Janeiro	20 – São Paulo
2 – Acre	15 – Sergipe		
3 – Amazonas	16 – Bahia		
4 – Roraima	17 – Minas Gerais		
5 – Pará	18 – Espírito Santo		
6 – Amapá	21 – Paraná		
7 – Tocantins	22 – Santa Catarina		
8 – Maranhão	23 – Rio Grande do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

Fonte: Elaborado pelos autores.

Logo, a Análise de Agrupamentos, utilizando o método hierárquico do Centroid, resultou na formação de três grupos, onde os estados brasileiros que constituem o mesmo grupo apresentam comportamento semelhante em relação ao total de óbitos de cada causa determinada, e se diferem dos estados que compõe os demais grupos. O grupo III é composto por São Paulo, o grupo II é composto pelo Acre e o grupo I composto pelos demais estados brasileiros, conforme mostram a **Figura 2** e a **Tabela 4**. Um dos possíveis motivos para o Acre ter formado um grupo único é que, comparado com os demais estados, ao coletar os 20%

maiores registros de óbitos pelas seis causas aqui avaliadas, este estado foi o único que atingiu este patamar em quatro causas - C.4.a, C.4.d, C.4.e e C.4.g.

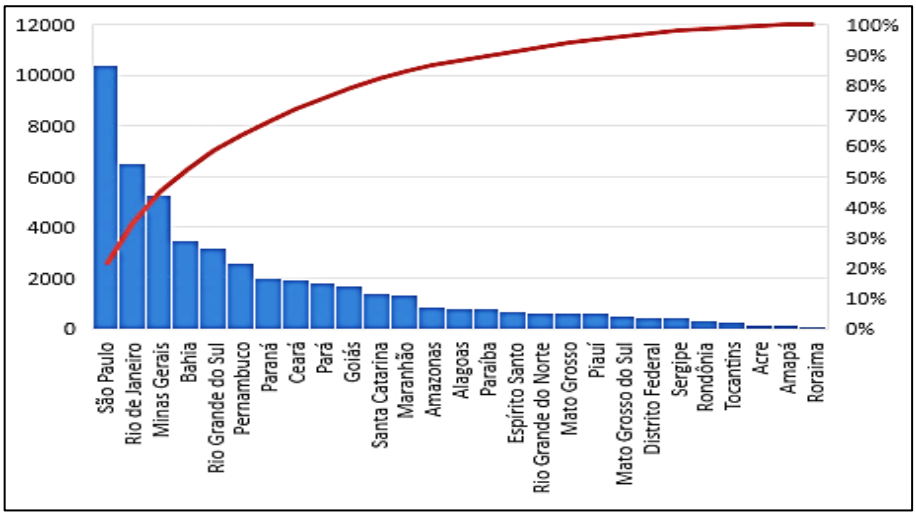
Figura 2 - Dendrograma formado a partir a Distância de Mahalanobis e o método Hierárquico do Centroid, com três grupos



Fonte: Elaborado pelos autores.

Em relação ao segundo e ao terceiro grupo formados, os estados de São Paulo e do Rio de Janeiro se destacam dos demais em relação ao grande número de óbitos. São Paulo foi o estado que apresentou os maiores índices de óbitos em relação a todas as causas determinadas. A **Figura 3** representa os estados brasileiros em relação ao total de óbitos por meio do diagrama de Pareto, um gráfico de barras que ordena as frequências das ocorrências, da maior para a menor, permitindo a priorização de problemas, mostrando ainda a curva de percentagens acumuladas (BUSSAB; MORETTIN, 2010). É possível observar que São Paulo e o Rio de Janeiro possuem o maior número de óbitos e juntos com Minas Gerais correspondem a quase 50% do número total de óbitos no país.

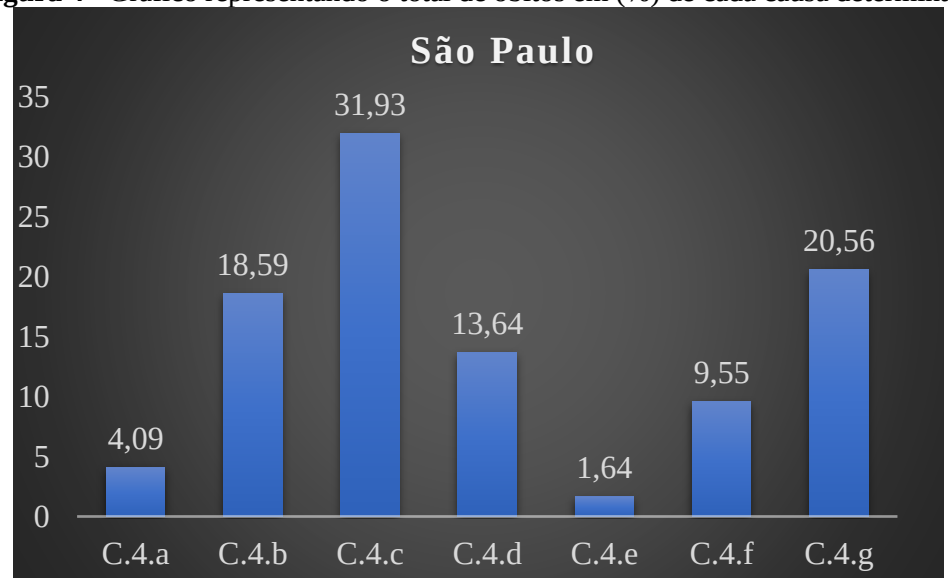
Figura 3 - Gráfico de Pareto representando os estados brasileiros em relação ao total de óbitos



Fonte: Elaborado pelos autores.

Como pode ser observado na **Figura 4**, aproximadamente 32% do total de óbitos do estado de São Paulo está relacionado a causa doença do aparelho circulatório (C.4.c). De acordo com os dados da Rede Integrada de Informações para a Saúde (RIPSA, 2000), cerca de 60% das mortes nesse grupo correspondem conjuntamente com as doenças isquêmicas do coração e as doenças cerebrovasculares. Tais motivos dessa elevada taxa de mortalidade do C.4.c então associadas principalmente aos fatores de risco como a obesidade, o fumo, hipertensão, diabetes, hipercolesterolemia, sedentarismo e estresse. Outro fator dessa porcentagem elevada pode estar relacionado à qualidade da assistência médica disponível.

Figura 4 - Gráfico representando o total de óbitos em (%) de cada causa determinada



Fonte: Elaborado pelos autores.

Segundo o MS (2019), cerca de 37% das mortes relacionadas a causa doença do aparelho circulatório são idosos (pessoas com 60 anos ou mais). As mais comuns são infarto, derrame e hipertensão. De acordo com o IBGE (2019), os idosos representam cerca de 13% da população total brasileira. Em relação aos óbitos por neoplasias (C.4.b), o número alto de registros se concentra principalmente na região sul do país, onde os três estados que formam esta região estão entre os cinco que obtiveram os índices mais altos por esta causa de mortalidade no conjunto de dados avaliado.

A desigualdade de acesso e qualidade dos serviços de saúde, bem como a qualidade de vida das pessoas, é outro fator que aumenta a taxa de mortalidade por doenças do aparelho circulatório. De acordo com a Prefeitura de São Paulo (1992), regiões com piores condições socioeconômicas, como Itaim Paulista e Brasilândia, apresentam risco bem maiores (chegando a mais de 50%) do que as regiões onde a média dos moradores apresentam melhores condições de vida como Vila Mariana e Jardim Paulista.

4 Considerações finais

As desigualdades existentes entre os estados brasileiros que se referem ao atendimento a saúde da população permearam o presente estudo no sentido de averiguar quais morbidades acarretam o maior número de óbitos e em quais estados isso ocorre. Dentre os principais resultados obtidos, a análise de agrupamento aplicado nos índices de mortalidade nos estados brasileiros resultou em 3 grupos, onde um deles resultou somente o estado de São Paulo, o outro o Rio de Janeiro e o terceiro os demais estados. Tanto o São Paulo como o Rio de Janeiro apresentaram comportamento distinto devido ao alto índice de óbitos.

Dentre as causas de mortes estudadas nos estados, destaca-se as relacionadas às doenças do aparelho circulatório que constituem um importante problema de saúde pública. Alguns fatores que estão relacionados com essa doença são a obesidade, o fumo, a hipertensão, o diabetes, a hipercolesterolemia, o sedentarismo e o estresse. Cerca de 37% dos óbitos são idosos, e regiões com condições socioeconômicas mais baixas apresentam uma taxa maior de óbitos em comparação com regiões com melhores condições de vida. Logo cabe ao Estado do Rio de Janeiro e São Paulo melhorar a qualidade de vida e dos serviços de saúde e melhorar o atendimento para o idoso.

É necessário aprofundar os estudos existentes no sentido de conhecer melhor as determinações da morbimortalidade pelas referidas causas, assim como capacitar os serviços de saúde das unidades federativas com maior prevalência para o diagnóstico e tratamento das mesmas. A causa de mortalidade pelo aparelho respiratório ganhou ainda mais atenção devido a pandemia instalada no mundo em 2020, causando preocupação dos órgãos estaduais de todo o país devido ao grande registro de óbitos pelo novo coronavírus. Diante da atual realidade, tais órgãos deveriam atualizar o banco de dados no SIM, a fim de motivar mais estudos como este, proporcionando mais informações a respeito sobre esta causa de mortalidade, quais morbidades levam ao maior número de registros de óbitos e quais as possíveis soluções que podem surgir para um melhor enfrentamento desta importante causa de mortalidade no país e no mundo.

Referências

BARTLETT, M.S. Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London*, serie A, London, 160:268-282, 1937.

BUSSAB, W. de O.; MORETTIN, P.A. *Estatística básica*. 6.ed. São Paulo: Saraiva, 2010.

CARVALHO, T. B. A. *Um estudo sobre funções de distância aplicadas a algoritmos de aprendizagem de máquina.* Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade Federal de Pernambuco. 2007.

CRUZ, C.D.; CARNEIRO, P.C.S. *Modelos biométricos aplicados ao melhoramento genético.* 2.ed. Viçosa: UFV, 2003. 585p.

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

DA SILVA, A.R.; MALAFAIA, G.; MENEZES, I.P.P. (2017) biotools: an R function to predict spatial gene diversity via an individual-based approach. *Genetics and Molecular Research*, 16: gmr16029655.

DATASUS. IDB 2012 - Indicadores e Dados Básicos no Brasil – 2012. Ministério da Saúde. *Indicadores de Mortalidade. Mortalidade proporcional por grupos de causas.* Disponível em: < http://tabnet.datasus.gov.br/cgi/tabcgi.exe?idb2012/c04.def> Acesso: 13 mai. 2019.

EVERITT, B.S. *Cluster analysis*. 3rd ed. London: Heinemann Educational Books, 1993, 122p.

FÁVERO, L. P. et al. *Análise de dados: modelagem multivariada para tomada de decisões.* Rio de Janeiro: Campus Elsevier, 2009.

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

FRIAS, P. G. de; SZWARCWALD, C. L.; LIRA, P. I. C. de. Estimação da mortalidade infantil no contexto de descentralização do sistema único de saúde (SUS). *Rev. Bras. Saude Mater. Infant.* 2011, 11: 463-470.

HAIR JR., J. F. et al. *Análise multivariada de dados.* 5. ed. Tradução Adonai Schlup Sant'Anna e Anselmo Chaves Neto. Porto Alegre: Bookman, 2005.

IBGE – Instituto Brasileiro de Geografia e Estatística. 2019. Disponível em: < https://censo2020.ibge.gov.br/2012-agencia-de-noticias/noticias/24036-idosos-indicam-caminhos-para-uma-melhor-idade.html >. Acesso: 01 junho 2020.

JOHNSON, R.A. and WICHERN, D.W. *Applied Multivariate Statistical Analysis.* New Jersey-USA: Englewood Cliffs, 642p. 1998.

MELLO-JORGE, M.H.P., GOTLIEB, S.L.D., LAURENTI, R. O sistema de informações sobre mortalidade: problemas e propostas para seu enfrentamento. II - Mortes por causas externas. *Rev Bras Epidemiolol,* 2002; 5:212-23.

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

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

MS - Ministério da Saúde. *Hipertensão é diagnosticada em 24,7% da população, segundo a pesquisa Vigitel. 17 maio 2019.* Disponível em: < http://www.saude.gov.br/noticias/agencia-saude/45446-no-brasil-388-pessoas-morrem-por-dia-por-hipertensao >. Acesso: 01 jun. 2020.

_____. *Situação Epidemiológica – Dados*, 27 mar. 2014. Disponível em: < https://www.saude.gov.br/index.php?option=com_content&view=article&id=11232&catid=671&Itemid=250 >. Acesso: 01 jun. 2020.

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

PREFEITURA DE SÃO PAULO. *Boletim PRO-AIM nº 10 / 4º trimestre 1992*. Disponível em: <https://www.prefeitura.sp.gov.br/cidade/secretarias/subprefeituras/noticias/?p=8387>. Acesso: 01 jun. 2020.

QUINTAL, G. *Análise de clusters aplicada ao Sucesso/Insucesso em Matemática.* Dissertação de mestrado em Matemática, Universidade da Madeira, Funchal. 2006.

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

R DEVELOPMENT CORE TEAM. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2013. URL http://www.R-project.org/.

RIPSA. *Indicadores de mortalidade*. Disponível em: <http://tabnet.datasus.gov.br/cgi/idb200/fqc11.htm>. Acesso: 01 junho 2020.

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

VALENTIN, J. L. *Ecologia numérica: uma introdução à análise multivariada de dados ecológicos*. Rio de Janeiro: Interciência, 2000.

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