UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE CIÊNCIAS

DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA

CURSO DE GRADUAÇÃO EM ESTATÍSTICA

NATÁLIA LIMA DE OLIVEIRA

ANÁLISE DE AGRUPAMENTO HIERÁRQUICOS

FORTALEZA

2022

NATÁLIA LIMA DE OLIVEIRA

ANÁLISE DE AGRUPAMENTO HIERÁRQUICOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Estatística.

Orientadora: Prof. Dr. Silvia Maria de Freitas

FORTALEZA

2022

NATÁLIA LIMA DE OLIVEIRA

ANÁLISE DE AGRUPAMENTO HIERÁRQUICOS

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística do Centro de Ciências da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Estatística.

Aprovada em:

BANCA EXAMINADORA

Prof. Dr. Silvia Maria de Freitas (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Xxxxxxx Xxxxxx Xxxxxxx
Universidade do Membro da Banca Dois (SIGLA)

Prof. Dr. Xxxxxxx Xxxxxx Xxxxxxx
Universidade do Membro da Banca Três (SIGLA)

Prof. Dr. Xxxxxxx Xxxxxx Xxxxxxx
Universidade do Membro da Banca Quatro (SIGLA)

À minha família, por sua capacidade de acreditar em mim e investir em mim. Mãe, seu cuidado e dedicação foi que deram, em alguns momentos, a esperança para seguir.

AGRADECIMENTOS

Em primeiro lugar, Deus, por minha vida, por me auxiliar em todos os momentos, durante meus anos de estudos.

A minha família, que me incentivaram nos momentos difíceis e compreenderam a minha ausência enquanto eu me dedicava, agradeço todo o apoio.

A minha mãe, que esteve ao meu lado em todos os momentos, meu exemplo de pessoa e minha motivação para realizar meus sonhos.

Aos meus colegas de curso Luan, Nayara, Letícia, Marcelo, Elias, Rickson, Daniele, cm quem convivi intensamente durante os últimos anos, pelo companheirismo e pela troca de experiências que me permitiram crescer não só como pessoa, mas também como profissional.

As minhas amigas, Juliene e Alice que estiveram sempre ao meu lado, pela amizade incondicional e pelo apoio demonstrado ao longo de todos esses anos.

A professora Silvia, por ser minha orientadora e ter desempenhado tal função com dedicação e amizade.

Ao professor Gualberto, por acreditar no meu potencial, por todos os conselhos, pela ajuda e pela paciência na qual guiaram o meu aprendizado.

“Pés, para que os quero, se tenho asas para voar”

(Frida Kahlo)

RESUMO

A análise de agrupamento hierárquico é uma forma eficaz de agrupar elementos, com o objetivo de facilitar a análise exploratória de dados. Os grupos são selecionados de forma que dentro de cada grupo os elementos sejam homogênios e os grupos sejam heterogênios entre si. Sua importância se dá na necessidade de estabelecer um perfil e dessa forma entender o comportamento de cada grupo. Primeiramente foi realizada a apresentação dos coeficientes de similaridade e associação, utilizados para cálculos das distâncias com diferentes tipo de variáveis. Diante disso, foram diferenciados os métodos hieráquicos que são: método da ligação simples, ligação completa, centróide, média das distâncias e ward, com suas particularidades. Também foram apresentados os algoritmos Agnes e Diana. Em seguida foram apresentados métodos de ajute e escolha do número ideal de grupos. Por último, foi utilizado o software R para a demonstração dos métodos e a comparação das ditâncias euclidiana e manhattan.

Palavras-chave: agrupamento; dendograma; similaridade

ABSTRACT

Hierarchical cluster analysis is an effective way to group elements together to facilitate exploratory data analysis. The groups are selected so that within each group the elements are homogeneous and the groups are heterogeneous to each other. Its importance lies in the need to establish a profile and thus understand the behavior of each group. First, the similarity and association coefficients were presented, used to calculate distances with different types of variables. In view of this, the hierarchical methods were differentiated, which are: single link method, complete link, centroid, average of distances and ward, with their particularities. The Agnes and Diana algorithms were also presented. Then, methods of adjustment and choice of the ideal number of groups were presented. Finally, the R software was used to demonstrate the methods and compare the euclidean and manhattan distances.

Keywords: grouping; dendrogram; similarity

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A análise de agrupamento, também chamada de análise de cluster, se trata de uma técnica de estatística multivariada que tem como principal objetivo subdividir os dados em grupos, esses grupos possuem características semelhantes internas, entretanto, quando comparados os grupos entre si, possuem distinção externa. Dessa forma existe homogeneidade dentro de cada grupo e uma heterogeneidade entre os grupos formados.

Essa técnica pode ser aplicada em várias áreas do conhecimento, como na recomendação de produtos em sites de vendas, nos anúncios que são enviados; no ramo agrícola é possível separar grupos de cultivo; na ecologia com a segmentação de espécies; na formação de perfis para seleção de empresa ou até mesmo estudos psicológicos e sociais.

A análise de agrupamentos é um tipo de classificação, pois, quando os dados são divididos, de certa forma está classificando os mesmos em grupos. Diante disso é importante saber que existem dois tipos de classificações supervisionadas e não-supervisionadas. A supervisionada se dá quando tem classes predefinidas e é trabalhado a classificação, porém, nas não-supervisionadas criamos os grupos sem classes preestabelecidas, portanto, poderia ser concluído que a técnica estudada se trata de uma classificação não-supervisionada.

Para aplicar essa técnica é preciso entender que existem dois tipos de métodos para sua aplicação, que consistem em métodos hierárquicos e não hierárquicos. A grande diferença entre eles é a necessidade de uma partição inicial, os hierárquicos formam todas as possibilidades de partições que podem ser adotadas, ademais, com os não hierárquicos é preciso definir essa partição inicial que pode se originar de um método hierárquico. Neste presente estudo está focado no processo de construção e validação dos métodos hierárquicos.

2 ANÁLISE DE AGRUPAMENTOS

No dia-a-dia é muito comum se deparar com bancos de dados com muitas informações coletadas, com um grande volume de observações para serem analisadas e a partir delas a tomada decisão. Diante disso, uma solução viável ao pesquisador é utilizar os métodos de análise de agrupamento, com o intuito de facilitar a visão do comportamento de seus dados.

Por exemplo, em uma situação cujo interesse é descobrir o gênero de filme favorito de usuários de uma plataforma de *streaming* de acordo com o histórico desses usuários. Sabe-se que nos dias atuais existem vastas opções de gêneros e subgêneros de filmes, ou seja, o banco de dados terá um volume grande de observações e analisar cada usuário seria inviável, pois demandaria um alto custo, tempo e força de trabalho. A solução para isso seria classificar os usuários em grupos com gostos semelhantes, utilizando os métodos de análise de agrupamento.

2.1 Introdução aos métodos hirárquicos

A característica mais marcante dos métodos hierárquicos na análise de agrupamento é ser capaz de fornecer a jornada que cada elemento faz até chegar no seu grupo final, ou seja, geram várias possibilidades de partições dos dados.

O método hierárquico não necessita que já possua o número de grupos inicialmente, entretanto, não são vistos como flexíveis, por conta que se um determinado elemento foi alocado em um grupo não é possível realocar em outro, ele é necessariamente dependente do seu grupo anterior. Diante disso, é importante saber que eles podem ser classificados como aglomerativos e divisivos.

No método aglomerativo inicialmente, cada elemento é considerado ser um grupo e ao longo das etapas os elementos vão se agrupando até que no fim exista somente um grupo com todos os elementos. E o método mais utilizado entre os hierárquicos, pois exige uma capacidade computacional menor que os divisivos.

O método divisivo consiste em considerar todos os elementos inicialmente em um único grupo, ao longo das etapas os grupos vão se dividindo entre si, até que na última etapa cada grupo terá um único elemento.

2.2 Medidas de distâncias

Sabendo que os dados constituem de n objetos mensurados p variáveis organizados em uma matriz de dados x_{n*p} com $i = 1...n$ e $j = 1...p$ representada abaixo:

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_i^T \\ \vdots \\ \mathbf{x}_n^T \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{(1)} & \dots & \mathbf{x}_{(j)} & \dots & \mathbf{x}_{(p)} \end{bmatrix}$$

Considere que $\mathbf{X}_n^T$ um vetor linha p-dimensional de observações do i-ésimo objeto e $x_{(j)}$ trata-se do vetor coluna n-dimensional de observações da j-ésima variável. Nas fórmulas abaixo calcula-se a distância entre os objetos x_r e x_s, $r \neq s = 1, 2 \dots p$.

2.2.1 Distância Euclideana

A distância Euclidiana é uma das medidas de distância mais utilizadas da literatura, e sua fórmula mede a distância entre dois pontos, e pode ser provada pelo conhecido teorema de Pitágoras (Gower 1982). Dessa forma, a distância euclidiana entre os objetos x_r e x_s é definida por:

$$D_{r,s} = \sqrt{\sum_{j=1}^{j} (x_{rj} - x_{sj})^2} = \sqrt{(x_r - x_s)^T (x_r - x_s)} \tag{2.1}$$

A seguir a matriz M representando dois indivíduos para ser utilizada em cada exemplo.

$$M = \begin{bmatrix} Indivduo\ 1 \\ Indivduo\ 2 \end{bmatrix} = \begin{bmatrix} 23 & 163 & 65 \\ 39 & 171 & 94 \end{bmatrix}$$

Exemplo: Temos a matriz M com dois indivíduos e suas respectivas características de interesse, calcule a distância Euclidiana entre eles.

$$D_{(Indivduo1, Indivduo2)} = \sqrt{(23-39)^2 + (1,63-1,71)^2 + (65-94)^2} = 34,0734501$$

2.2.2 *Distância de Manhattan*

É uma distância também chamada de *city-block*, ou máxima, e tem como objetivo maximizar a distância entre dois elementos. Essa distância tem forte relação com a distância Euclidiana, sendo defina como:

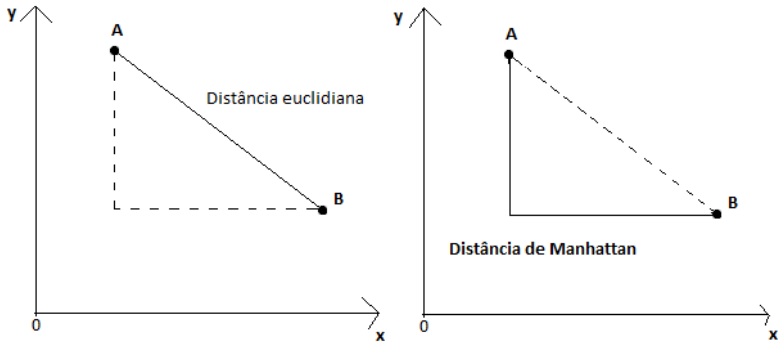
$$M_{r,s} = \sum_{j=1}^{p} |x_{rj} - x_{sj}| \tag{2.2}$$

Exemplo: Utilizando a matriz M, calcule a distância de Manhattan.

$$M_{(Indivduo1,Indivduo2)} = |23-39| + |163-171| + |65-94| = 53$$

Na figura 1 está apresentada, de forma gráfica a relação entre a distância Euclidiana e Manhanttan.

Figura 1 – Relação entre a distância Euclidiana e Manhattan



Fonte: Elaborado pela autora (2022).

2.2.3 *Distância de Minkowsky*

Observa-se que essa distância é a generalização da distância Euclidiana e Manhattan, respectivamente, quando $p = 2$ e $p = 1$,é dada por:

$$D_{(r,s)} = (\sum_{j=1}^{p} |x_{rj} - x_{sj}|^{p})^{\frac{1}{p}} \tag{2.3}$$

2.2.4 Distância de Camberra

A distância de Camberra apresentada por Lance y Williams (1967), é um caso particular da distância de Manhattan quando $|x_{ri}|$ e $|x_{si}|$ são iguais. Definida como:

$$D_{Can(r,s)} = \sum_{i=1}^{n} \frac{|x_{ri} - x_{si}|}{|x_{ri}| - |x_{si}|} \tag{2.4}$$

Exemplo: Utilizando a matriz M do exemplo anterior calcule a distância Camberra, é dada por:

$$D_{Can(Indivduo1,Indivduo2)} = \frac{16}{23-39} + \frac{8}{163-171} + \frac{29}{65-94} = -3$$

2.2.5 Distância de Mahalanobis

Essa distância foi apresentada por Mahalanobis em 1936 sendo uma distância Euclidiana que tem como pesos as covariâncias das variáveis. Considerando Σ a matriz das covariâncias das variáveis, a distância entre os indivíduos x_r ma x_s é definido por:

$$D_{Maha(r,s)} = (x_r - x_s)^T \Sigma^{-1} (x_r - x_s). \tag{2.5}$$

2.3 Coeficientes de concordância

Os coeficientes de concordância são quando os dados se apresentam na forma de classificação, em tais situações as medidas de distâncias já apresentadas não suprem essa necessidade para variáveis dessa natureza. Considere a seguir a Tabela 1 que apresenta os resultados associados às respostas de duas variáveis com resultados classificados como "presença= 1"e "ausência = 0".

Tabela 1 – Geral dos coeficientes de associação

	1	0
1	(1 , 1) a	(1 , 0) b
0	(0 , 1) c	(0 , 0) d

Na tabela é considerado o valor 1 como a presença e o valor 0 como ausência de determinada característica do objeto de estudo. A tabela nos apresenta os quatro casos de combinações de ausências e presenças dos possíveis resultados.

Nos exemplos dos coeficientes de concordância é utilizado os dados nos quais foram detectados a incidência de mamíferos de uma localidade montanhosa na região de Great Basin (EUA), conforme Grayson e livingston(1993), de acordo com o listado na Tabela 2.

Tabela 2 – Incidência de mamíferos

	Variáveis																		
Indivíduos	**y1**	**y2**	**y3**	**y4**	**y5**	**y6**	**y7**	**y8**	**y9**	**y10**	**y11**	**y12**	**y13**	**y14**	**y15**	**y16**	**y17**	**y18**	**y19**
x1	1	1	1	0	1	1	1	0	0	1	0	1	1	1	1	1	1	1	1
x2	1	1	0	1	1	0	1	0	1	0	0	1	1	0	1	1	1	1	1
x3	1	0	0	1	1	0	1	0	1	0	0	1	1	1	1	1	1	1	1
x4	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
x5	1	0	0	1	1	0	1	0	1	0	0	1	1	0	0	0	1	0	0
x6	1	1	1	0	1	0	1	1	0	0	0	1	1	1	1	1	1	1	1
x7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
x8	1	0	1	1	1	0	0	0	1	0	0	1	1	0	0	1	1	0	1
x9	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0
x10	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	1	0	0
x11	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1
x12	1	0	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1
x13	1	1	1	1	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1
x14	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

Fonte: Grayson, D.K. and S.D. Livingston. 1993.

Foram considerados dois indivíduos, conforme apresentado na Tabela 3, para exemplificar cada coeficiente apresentado nas próximas sessões.

Tabela 3 – Indivíduos x7 e x11

	y1	**y2**	**y3**	**y4**	**y5**	**y6**	**y7**	**y8**	**y9**	**y10**	**y11**	**y12**	**y13**	**y14**	**y15**	**y16**	**y17**	**y18**	**y19**
x7	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
x11	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	0	1	1

Utilizando a Tabela 1 geral dos coeficientes de associação, é possível contabilizar cada situação da Tabela 4 a seguir.

Tabela 4 – Coeficientes de associação entre os indivíduos x7 e x11

a	b	c	d
1	1	16	1

2.3.1 Correspondência múltipla (M)

Esse coeficiente chamado de correspondência múltipla ou de simple matching foi apresentado por Sokal e Michene no ano de 1953. Analisando a fórmula proposta é fácil ver que os autores consideram duas presenças (a) e duas ausências (d).

$$M = \frac{a+d}{a+b+c+d} \tag{2.6}$$

Exemplo: Considerando os indivíduos x_7 e x_{11} o coeficiente de correspondência múltipla é dado por:

$$M = \frac{1+1}{1+1+16+1} = \frac{2}{19} = 0,105263158$$

2.3.2 Rogers e Tanimoto(RT)

Em 1960, Roger e Tanimoto propuseram uma modificação no coeficiente de correspondência múltipla, em que adicionaram pesos aos casos com uma ausência e uma presença (b, c). Definido como:

$$RT = \frac{a+d}{a+2b+2c+d} \tag{2.7}$$

Exemplo: Utilizando os indivíduos x_7 e x_{11} para o cálculo do coeficiente de Roger e Tanimoto tem-se.

$$RT = \frac{1+1}{1+2.1+2.16+1} = \frac{2}{36} = 0,0555$$

2.3.3 Russell e Rao (RR)

Nesse caso os autores consideram apenas as duplas presenças, dado por:

$$RR = \frac{a}{a+b+c+d} \tag{2.8}$$

Exemplo: Utilizando x_7 e x_{11} para o cálculo do coeficiente de Russell e Rao, tem-se.

$$RR = \frac{1}{1+1+16+1} = \frac{1}{19} = 0,0526315789$$

2.3.4 Kulzynski(K)

Na distância de Kulzynski são considerados as duas ausências(d), ou seja, somente os casos em que se tem pelo menos uma presença (a, b, c).

$$K = \frac{a}{b+c} \tag{2.9}$$

Exemplo: Utilizando x_7 e x_{11} para o cálculo do coeficiente de Kulzynski.

$$K = \frac{1}{1+16} = \frac{1}{17} = 0,0588235294$$

2.3.5 Sokal e Sneath(SS)

Sokal e Sneath(1963) adicionaram pesos aos casos de dupla presença e dupla ausência e definiram da forma a seguir:

$$SS = \frac{2a+2d}{2a+b+c+2d} \tag{2.10}$$

Exemplo: Utilizando x_7 e x_{11} para o cálculo do coeficiente de Sokal e Sneath. Tem-se o resultado abaixo:

$$SS = \frac{2x1+2x1}{2x1+1+16+2x1} = \frac{4}{21} = 0,01904$$

2.3.6 Hamann (H)

Esse coeficiente tem as diferenças entre as somas das duplas presença e ausência(a, d) com as combinações (b, c). Diferente dos outros coeficientes apresentados esse está contido entre -1 e 1.

$$H = \frac{(a+d)-(b+c)}{a+b+c+d} \tag{2.11}$$

Exemplo: Utilizando x_7 e x_{11} para o cálculo do coeficiente de Hamann.

$$H = \frac{(1+1)-(1+16)}{1+1+16+1} = \frac{2-17}{19} = \frac{-15}{19} = -0,7894737$$

2.3.7 Jaccard(J)

Equivalente a Kulzynski, não é considerado a dupla ausência, porém, no denominador é adicionado a dupla presença (a).

$$J = \frac{a}{a+b+c} \tag{2.12}$$

Exemplo: Utilizando x_7 e x_{11} para o cálculo do coeficiente de Jaccard.

$$J = \frac{1}{1+1+16} = \frac{1}{18} = 0,0555555556$$

2.3.8 Dice- Sorensen (DS)

Esse coeficiente coloca pesos na dupla presença (a) e não considera as ausências (d) no cálculo conforme a definição apresentada a seguir.

$$DS = \frac{2a}{2a+b+c} \tag{2.13}$$

Exemplo: Utilizando x_7 e x_{11} para o cálculo do coeficiente de Dice- Sorensen.

$$DS = \frac{2x1}{2x1+1+16} = \frac{2}{19} = 0,105263158$$

2.3.9 Simpsom (S)

O coeficiente de Simpsom não utiliza a dupla ausência em sua fórmula sendo notório a priorização das duplas, o coeficiente é dado por:

$$S = \frac{a}{a+min(b,c)} \tag{2.14}$$

Exemplo: Utilizando x_7 e x_{11} para o cálculo do coeficiente de Simpsom.

$$S = \frac{1}{1+1} = \frac{1}{2} = 0,5$$

Na Tabela 5 a seguir estão apresentados os resultados dos coeficientes calculados para os indivíduos x_7 e x_{11} da Tabela 3.

Tabela 5 – Comparação entre os resultados dos coeficientes de associação

Resultados dos coeficientes	
Correspondência simples	0,1052
Rogers e Tanimoto	0,5555
Russell e Rao	0,05263
Kulzynski	0,05882
Sokal e Sneath	0,01904
Hamann	- 0,7894
Jaccard	0,05555
Dice- Sorensen	0,10526
Simpsom	0,5000

2.4 Coeficientes para dados categóricos e quantitativos

2.4.1 Gower(G)

Gower(1971) apresentou um coeficiente que pode ser utilizado para dados qualitativos e quantitativos que herdou o seu nome, possibilitando uma maior flexibilidade para o uso dos dados de natureza qualitativa/quantitativos, entretanto, para cada situação o cálculo é realizado de uma forma diferente. O caso geral, é definido por:

$$G = \frac{1}{p}\sum_{j=1}^{p} S_{ABj} \tag{2.15}$$

Sendo S_{ABj} a medida de similaridade entre os indivíduos A e B na variável J. Diante disso, existem quatro situações em que é possível aplicar Gower.

- Dados qualitativos nominais: Os dados qualitativos nominais podem ser representados por 1 = presença e 0 = ausência. Ou seja, serão quatro situações possíveis de resultados: duas concordâncias sendo a = (1,1) com duas presenças, e d = (0,0) com duas ausências e duas discordâncias b = (1,0) e c = (0,1), sendo uma presença e uma ausência cada. Dessa forma considera-se similaridade igual a um (S = 1) para as concordâncias e similaridade igual a zero (S = 0) para as discordâncias.

- Dados quantitativos: Para dados quantitativos contínuos ou discretos, calcula-se.

$$S_{ABj} = 1 - \frac{|X_{Aj} - X_{Bj}|}{max(X_j) - min(X_j)} \tag{2.16}$$

– Dados qualitativos ordinais: Para dados qualitativos ordinais é necessário se atentar nas diferenças entre valor em cada categoria, pois elas devem ter a mesma distância enter si. Se essas diferenças não for em consideráveis valores devem ser transformados. Podani(1999) apresentou um método para quantificar a similaridade neste caso, em que os valores são transformados em ranks (representado por r na fórmula):

$$S_{ABj} = 1 - \frac{|r_{Aj} - r_{Bj}| - \frac{T_{Aj}-1}{2} - \frac{T_{Bj}-1}{2}}{max(r_j) - min(r_j)| - \frac{max(T_j)-1}{2} - \frac{min(T_j)-1}{2}} \tag{2.17}$$

T_{Aj} e T_{Bj} :Valores totais de cada grupo, no caso, grupo A e B, respectivamente.

$max(T_j)$ e $min(T_j)$: são os números totais dos grupos que têm o alcance máximo e mínimo, respectivamente.

O coeficiente de Gower não utiliza dados ausentes em seu método. Na fórmula a seguir será considerado w_j representante existe ausência compartilhada $w_j = 0$, quando nenhuma informação é perdida $w_j = 1$. Temos assim a forma final do coeficiente de Gower:

$$G = \frac{\sum_{j=1}^{p} w_j S_{ABj}}{\sum_{j=1}^{p} w_j} \tag{2.18}$$

Exemplo: Considere um conjunto de dados fictícios, descrito na Tabela 6. O cálculo do coeficiente de Gower em cada caso da Tabela a seguir, será apresentado para os indivíduos x1 e x2.

Tabela 6 – Tabela com dados fictícios

	y1	**y2**	**y3**	**y4**	**y5**
x1	1	NA	0	1	5,6
x2	0	0	0	2	2,5
x3	1	1	1	3	9,0

A variável $y1$ é categórica, então, $w_j = 1$, pois se tem pelo menos uma presença em $y1$. E a similaridade $S_{AB1} = 0$, pois existe pelo menos uma ausência.

A variável $y2$ possui um dado faltante no indivíduo $x1$, por conta disso $w_2 = 0$.

A variável $y3$ é categórica como podemos notar observando a tabela, por conta disso podemos afirmar que $w_3 = 0$ por causa da dupla ausência e sua similaridade é $s_{AB3} = 0$.

A variável $y4$ apresenta dados qualitativos ordinais, no exemplo será considerado que esses já são seus ranks. Existe a presença de ambos $w_4 = 1$, A seguir o cálculo de sua

similaridade:

$$S_{AB4} = 1 - \frac{1 - 2 - \frac{1-1}{2} - \frac{2-1}{2}}{3 - 1 - \frac{3-1}{2} - \frac{1-1}{2}} = 0,5$$

Na variável $y5$ existe informações para ambos, logo $w_5 = 1$, entretanto a similaridade tem que ser calculada, Assim tem-se:

$$S_{AB5} = 1 - \frac{5,6 - 2,5}{9 - 2,5} = 1 - \frac{3,1}{6,5} = 1 - 0,476923077 = 0,523076923.$$

Agora podemos finalizar o cálculo do coeficiente de Gower:

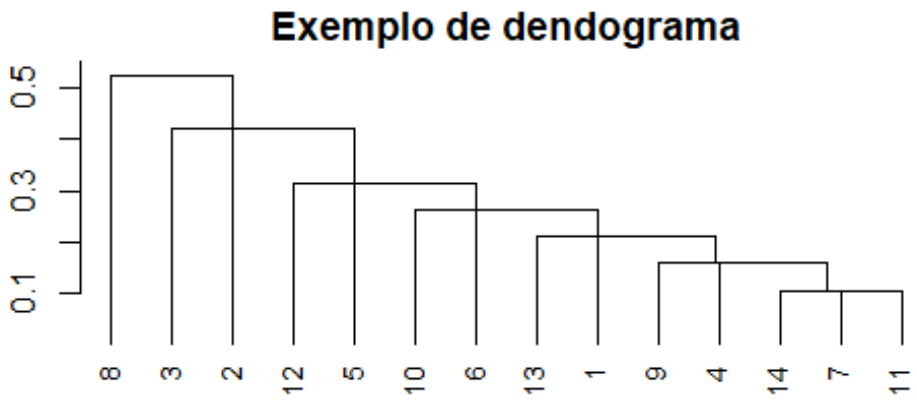
$$G = \frac{w_1.S_{AB1} + w_2.S_{AB2} + w_3.S_{AB3} + w_4.S_{AB4} + w_5.S_{AB5}}{w_1 + w_2 + w_3 + w_4 + w_5}$$

$$G = \frac{1.0 + 0 + 0 + 1.0,5 + 1.0,523076923}{1 + 0 + 0 + 1 + 1} = \frac{1,02307692}{3} = 0,34102564$$

3 MÉTODOS HIERÁRQUICOS

A definição de hierarquia consiste em uma ordem de prioridade entre os elementos de um conjunto, diante disso, os métodos hierárquicos são uma técnica que consiste em particionar os dados e é geralmente representada pelo gráfico dendrograma, em que é fácil notar a hierarquia entre os grupos. Observe no dendrograma a seguir.

Figura 2 – Dendograma



Fonte: Elaborado pela autora (2022).

3.1 Método de Ligação Simples

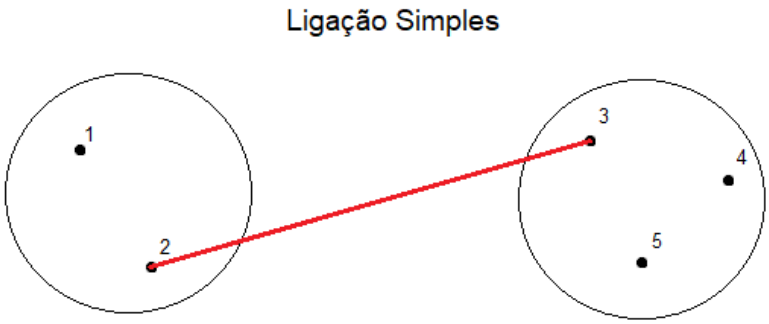
Este método também é conhecido como agrupamento dos vizinhos mais próximos, ou seja, é utilizada a menor distância entre os elementos dos grupos.

$$D(A,B) = min\{d(y_i,y_j)\ para\ y_i\ \varepsilon\ A\ ,y_j\ \varepsilon\ B\} \tag{3.1}$$

Em que $d(y_i,y_j)$ é a distância calculada, geralmente a distância euclidiana ou umas das apresentadas no capítulo anterior, entre ou vetores y_i,y_j . Em cada etapa a distância encontrada para o par é calculado com o grupo de menor distância, E assim, esse procedimento é repetido até que reste somente um grupo.

Observa-se na imagem dois grupos que no método de ligação simples busca-se encontrar o vizinho mais próximos, no caso da imagem o elemento 2 e 3. A distância representada pela linha tracejada em vermelho.

Figura 3 – Método de Ligação Simples



Fonte: Elaborado pela autora (2022).

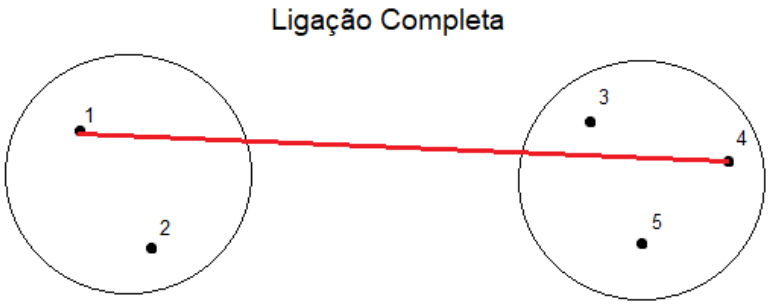
3.2 Método da ligação completa

Esse método diferente da ligação simples é utilizado a distância máxima entre os grupos e é conhecido por método dos vizinhos mais distantes.

$$D(A,B) = max\{d(y_i,y_j)\ para\ y_i\ \varepsilon\ A\ ,y_j\ \varepsilon\ B\} \tag{3.2}$$

Em que $d(y_i,y_j)$ é distância calculada, geralmente a distância euclidiana ou umas das apresentadas no capítulo anterior, entre ou vetores y_i,y_j . Em cada etapa a distância encontrada para par é calculado com o grupo de maior distância. E assim, esse procedimento é repetido até que reste somente um grupo.

Figura 4 – Método da ligação completa



Fonte: Elaborado pela autora (2022).

Observa-se na imagem dois grupos que no método de ligação completa busca-se encontrar o vizinho mais distante, no caso da imagem o elemento 1 e 4. A distância representada pela linha tracejada em vermelho.

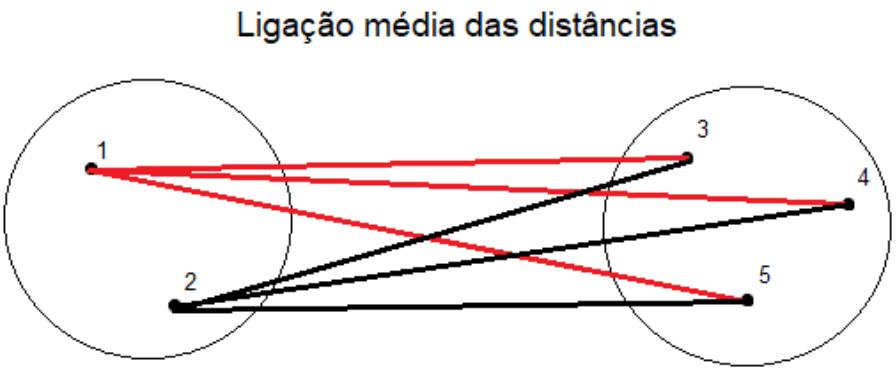
3.3 Método média das distâncias

Esse método como o próprio nome já sugere, é calculado a média das distâncias de todos os pares de elementos em que podem ser formados.

$$D(A,B) = \sum_{i \epsilon A}^{n_1} \sum_{j \epsilon B}^{n_2} \frac{1}{n_1 n_2} d(y_i, y_j) \tag{3.3}$$

Observa-se na imagem a linha tracejada em vermelho representa a interação do elemento 1 com os elementos do outro grupo no cálculo do método da ligação da média das distâncias.

Figura 5 – Método média das distânciaa



Fonte: Elaborado pela autora (2022).

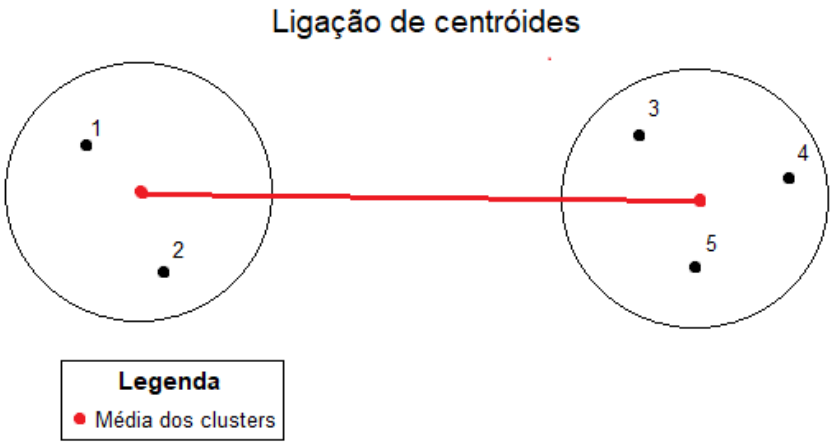
3.4 Método do centróide

O método do centroide consiste no cálculo da média em cada grupo e somente depois desse processo, é calculado a distância entre as médias, utilizando uma das distâncias apresentadas anteriormente.

$$D(A,B) = d(\bar{y}_i, \bar{y}_j) \tag{3.4}$$

É notório que a linha tracejada em vermelha representa a ligação de centroides, que está posicionada no centro dos grupos representado a média de cada grupo.

Figura 6 – Método do centróide



Fonte: Elaborado pela autora (2022).

3.5 Método de Ward

Esse método consiste no cálculo da distância entre dois grupos com a soma das distâncias ao quadrado.

$x_{l,i,k}$; Valor para a variável p na observação j pertencente ao grupo l

SS_l : soma dos erros quadrados no grupo L.

$SST_{l,i}$: soma total dos erros quadrados agrupamentos dos grupos l e i.

$$SS_l = \sum_{k=1}^{n_l} \sum_{j=1}^{p} x_{l,k,j} - \bar{x_{l,j}} \tag{3.5}$$

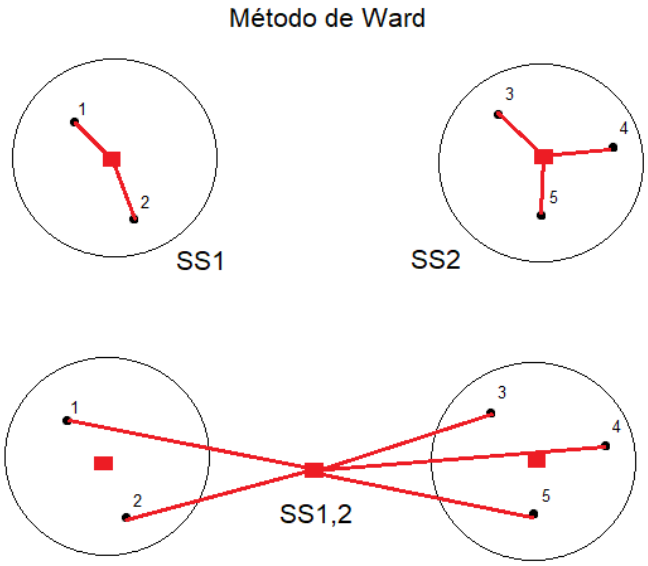
$$\bar{x_j} = n\frac{1}{n_l} \sum_{k=1}^{n_l} x_{l,k,j} \tag{3.6}$$

$$SS_{l,i} = \sum_{k=1}^{n_l} \sum_{j=1}^{p} (x_{l,k,j} - \bar{x_j})^2 + \sum_{k=1}^{n_l} \sum_{j=1}^{p} (x_{i,k,j} - \bar{x_i})^2 \tag{3.7}$$

$$\bar{x_j} = \frac{1}{n_l + n_i} (\sum_{j=1}^{p} x_{l,k,j} + \sum k = 1^{n_i} x_{i,k,l}) \tag{3.8}$$

$$D(C_l, C_i) = SS_{li} - (SS_l + SS_i) = \frac{n_l n_i}{n_l + n_i} \sum_{j=1}^{p} (x_{l,i} - \bar{x}_{i,j})^2 \tag{3.9}$$

Figura 7 – Método de ward



Fonte: Elaborado pela autora (2022).

É constatado na imagem que existem dois momentos representados, na primeira etapa é calculado SS_1 e SS_2 soma dos erros quadrados dentro de cada grupo. Na segunda etapa é calculado a soma total dos erros quadrados agrupamentos dos grupos 1 e 2.

3.6 Comparação entre os métodos hierárquicos

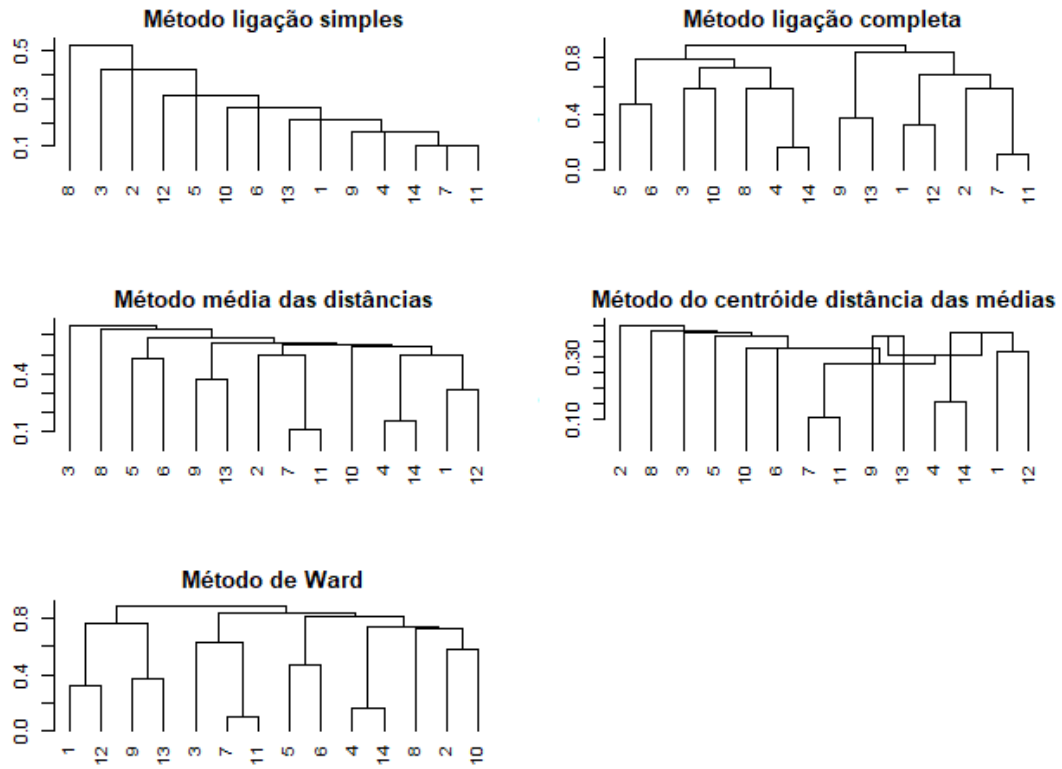
Segundo Sueli Mingoti (2005), a grande parte dos métodos apresentados geram grupos esféricos ou elipsoides, exceto o método da ligação simples, que pode produzir formas geométricas diferentes, porém, obtêm grupos mais próximos. Já o método da ligação completa gera grupos com diâmetros aproximados e tende a isolar valores considerados outiliers.

O método da média das distâncias, na maioria das vezes gera grupos com a variância interna parecida. Diante disso, se obtêm melhores divisões em relação aos métodos de ligação simples e completa. O método Ward tende a gerar grupos com o mesmo número de elementos e sua base teórica são os princípios de análise de variância.

Os métodos de ligação simples, completa e da média são utilizados com variáveis qualitativas e quantitativas, diferente dos métodos do centroide e ward indicados para variáveis quantitativas, pois utilizam média em seus métodos.

Na imagem a seguir Observa-se o comportamento do mesmo banco de dados utilizando o exemplo de coeficiente de associação em que é utilizado a técnica de correspondência simples pára o cálculo das distâncias. Observa-se que cada técnica possui resultados diferentes nos gráficos denominados dendogramas.

Figura 8 – Comparação entre os métodos com dados categóricos



Fonte: Elaborado pela autora (2022).

4 ALGORITMOS DE MÉTODOS HIERARQUICOS

4.1 Agglomerative Nesting (AGNES)

O algoritmo agglomerative nesting (agnes) utiliza o método aglomerativo. Dessa forma, os agrupamentos desse método usa a dissimilaridade entre os grupos. No início do método cada elemento do banco de dados é considerado um grupo, e assim os dois grupos mais semelhantes se tornam um e assim suscetivamente, recalculando a cada etapa a distância. Observe a fórmula a seguir:

$$D(M,N) = \frac{1}{n_M n_N} \sum_{y_i \in M, y_j \in N} d(y_i.y_j) \tag{4.1}$$

$$= \frac{1}{n_M n_N} \sum_{y_i \in A, y_j \in N} d(y_i.y_j) + \frac{1}{n_M n_N} \sum_{y_i \in B, y_j \in N} d(y_i.y_j) \tag{4.2}$$

$$= \frac{A}{M} \frac{1}{n_A n_N} \sum_{y_i \in A, y_j \in N} d(y_i.y_j) + \frac{B}{M} \frac{1}{n_B n_N} \sum_{y_i \in B, y_j \in N} d(y_i.y_j) \tag{4.3}$$

$$d(M,N) = \frac{A}{M} d(A.N) + \frac{B}{M} d(B.N) \tag{4.4}$$

A e B: Agrupamentos.

M; agrupamentos de A e B.

n_M, n_N, n_A, n_B : Quantidade de elementos em cada grupo.

4.2 Divisive Analysis (DIANA)

O algoritmo Divisive Analysis é chamado de Diana e foi baseado na proposta de Macnaughton-Smith (1964). Diana foi descrito por Kaufman Rousseeuw (1990) como algoritmo hierárquico divisivo. O algoritmo Diana consiste em selecionar o elemento com o maior diâmetro, considerando n como número de elementos no banco de dados, dado essa informação sabemos que teremos $n-1$ divisões no dendograma. Considere que selecionamos o agrupamento D, definido como:

$$Dimetro(D) = max(dist(i,j)), i, j \in A \tag{4.5}$$

Em que $d(i,j)$ é a dissimilaridade entre os elementos i e j pertencem ao agrupamento D. Segundo Struyf em 1996, se diâmetro(D)0, o agrupamento N pode ser dividido em dois agrupamentos A e B:

- Primeiramente considera-se que o agrupamento A = D e B=∅.
- Para transferir um elemento de A para B Para cada elemento $i \in A$, é calculada a dissimilaridade média do objeto i para todos os objetos de A denotado pela(i). O objeto d de A para o qual a(d) é o maior, é movido para B.
- Mover outros objetos de A para B segundo a regra abaixo: Se total de objetos de A = 1: Pare Senão para $i \in A$ calcule a dissimilaridade média de i para todos os objetos pertencentes ao agrupamento A denotado pela(i) e a dissimilaridade média de i para todos os objetos pertencentes a B denotado por d(i, B). Selecione o objeto $h \in A$ para o qual:

  $$a(h) - d(h,B) = max(a(i) - d(i,B)), iinA \tag{4.6}$$

  considere que d(h, B) é a dissimilaridade média entre h e B.
- Se $a(h)\breve{}d(h,B) > 0$ então mover h de A para B e repetir o item dois desse passo a passo.
- Se $a(h)\breve{}d(h,B) \leq 0$ então Manter A e B e parar o processo.

5 MÉTODOS DE ESCOLHA DO NÚMERO DE GRUPOS E AJUSTE

5.1 Coeficiente de correlação cofenética (CCC)

O Coeficiente de correlação cofenética é utilizado para medir o ajuste entre a matriz de dissimilidade (matriz fenética) e a matriz originada depois das aplicações dos métodos que é denominada como matriz cofenética, ou seja, o dendograma. Temos a seguir a fómula que se calcula o coeficiente de correlação cofenética.

$$CCC = \frac{Cov\hat{(}F,C)}{\sqrt{Var\hat{(}F)Var\hat{(}C)}} \tag{5.1}$$

Na fórmula acima o F representa a matriz fenética e o C retrata a matriz cofenética. A interpretação do coeficiente é que quando o CCC é maior que 0,7 conclui-se que o método de agrupamento foi adequado.

5.2 Método Elbow

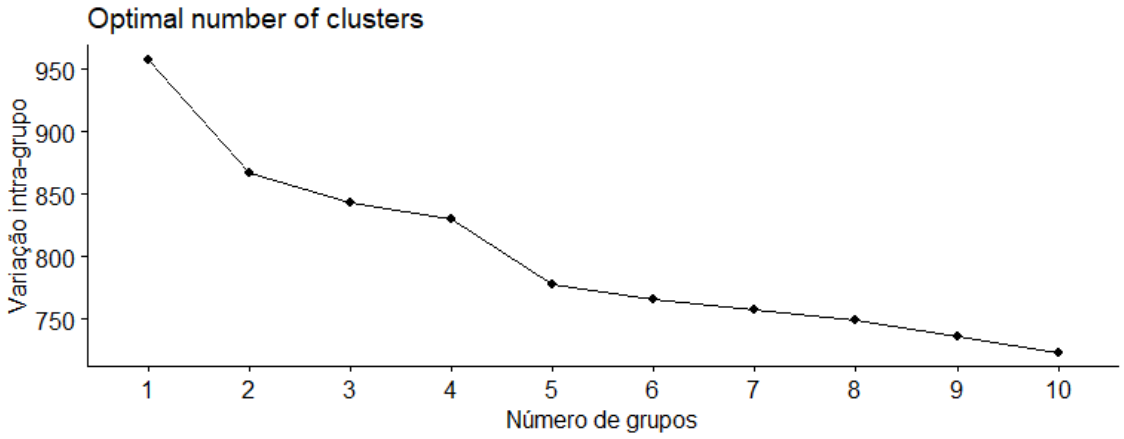
O método de Elbow também é chamado de método do cotovelo e seu principal objetivo é identificar o melhor número de grupos que pode ser obtido com a utilização dos métodos apresentados. Dessa forma, o método testa as variâncias dos dados em relação à quantidade de grupos, para o cálculo é utilizado a fórmula do erro quadrático.

$$SSE = \sum_{k=1}^{k} \sum_{x_i \in S_k} |X_i - C_k|^2 \tag{5.2}$$

Observa-se que os primeiros agrupamentos vão conter muitas informações. Entretanto, em um determinado momento obtêm-se a forma de um “cotovelo”, e quando se identifica esse ângulo se formar, é fácil notar a condição ideal da quantidade de grupos, por conta disso origina-se o nome do método.

No exemplo a seguir pode-se ver o "cotovelo"se formar ja no segundo grupo. .

Figura 9 – Exemplo de gráfico método de Elbow



Fonte: Elaborado pela autora (2022).

5.3 Análise do comportamento do nível de similaridade

Nesse método observa-se o nível de similaridade em cada etapa da análise de agrupamento. É identificado a quantidade de grupos a ser selecionado quando o nível de similaridade cai drasticamente. A fórmula utilizada no tópico a seguir:

$$S_{AB} = (1 - \frac{d_{AB}}{max(d_{r,s})}).100 \quad r,s = 1,...,n \tag{5.3}$$

Em que $max(d_{r,s})$ é a maior distâncias entre os n elementos inicias da primeira etapa da análise de agrupamentos.

6 BIBLIOTECAS

Este capítulo consiste em apresentar bibliotecas utilizadas no software R para aplicar os métodos apresentados e suas funções.

6.1 Stats

Essa biblioteca tem várias funções utilizadas na estatística, diante disso será focado as funções que auxiliam a aplicar as técnicas apresentadas.

6.1.1 Função dist

Esta função calcula e retorna a matriz de distância calculada usando a medida de distância especificada para calcular as distâncias entre as linhas de uma matriz de dados.

Figura 10 – Função dist

```
dist(x, method = "euclidean", diag = FALSE, upper = FALSE, p = 2)
```

Fonte: R Documentation

x: uma matriz numérica, quadro de dados ou objeto "dist".

method: a medida de distância a ser usada. Deve ser "euclidiano", "máximo", "manhattan", "canberra", "binário"ou "minkowski". Qualquer substring inequívoca pode ser fornecida.

diag: valor lógico que indica se a diagonal da matriz de distância deve ser impressa por print.dist.

upper: valor lógico indicando se o triângulo superior da matriz de distância deve ser impresso por print.dist.

p: O poder da distância de Minkowski.

6.1.2 Função hclust

Essa biblioteca do software R é utilizada para gerar agrupamentos hierarquicos. Em que funciona da seguinte forma:

d: É a distância calculada pela função por dist.

método: o método de aglomeração a ser utilizado. Deve ser (uma abreviação inequí-

Figura 11 – Função hclust

```
hclust(d, method = "complete", members = NULL)
```

Fonte: The hclust function is based on Fortran code contributed to STATLIB by F. Murtagh.

voca de) um de "ward.D", "ward.D2", "single", "complete", "average"(= UPGMA), "mcquitty"(= WPGMA), "median"(= WPGMC) ou "centróide"(= UPGMC).

membros: NULL ou um vetor com tamanho de comprimento d. Veja a seção 'Detalhes'.

6.2 Proxy

Nessa biblioteca existem várias funções, uma delas é a dist semelhante a do pacote stats calculando as medidas de distância, mas o diferencial desse pacote é função simil que calcula as similaridares. Observe:

6.2.1 simil

Função utilizada para calcular as similaridades, observe a sua estrutura:

Figura 12 – Função simil

```
simil(x, y = NULL, method = NULL, ..., diag = FALSE, upper = FALSE,
      pairwise = FALSE, by_rows = TRUE, convert_distances = TRUE,
      auto_convert_data_frames = TRUE)
```

Fonte: David Meyer [aut, cre], Christian Buchta [aut]

x: Para dist e simil, um objeto de matriz numérica, um quadro de dados ou uma lista.

y: NULL, ou um objeto semelhante ao x

method: método utilizado para calcular a similaridade. O padrão para dist é "Euclidean", e para simil "correlation". Para simil os métodos disponíveis são "simple matching", "Tanimoto", "Russel", "Kulczynski1", "Hamman", "Jaccard","Sorensen"e "Simpson".

diag: valor lógico que indica se a diagonal da matriz distância/semelhança deve ser impressa por print.dist/ print.simil. Observe que os valores diagonais nunca são armazenados em distobjetos.

upper valor lógico que indica se o triângulo superior da matriz de distância/semelhança deve ser impresso por print.dist/print.simil

pairwise valor lógico que indica se as distâncias devem ser calculadas para os pares de x e y apenas.

by rows: lógico indicando se as proximidades entre linhas ou colunas devem ser calculadas.

convert similarities, convert distances: lógico indicando se as distâncias devem ser automaticamente convertidas em similaridades (e vice-versa) se necessário.

auto convert data frames: lógico indicando se os quadros de dados devem ser convertidos em matrizes se todas as variáveis forem numéricas, ou todas forem lógicas, ou todas forem complexas.

6.3 Vegan

O pacote Vegan tem como principal objetivo ser usado para ecologia descritiva. Entretanto, podemos utilizar as funções desse pacote para o cálculo das dissimilaridades.

6.3.1 Vegdist

Essa função é uma alternativa para abranger mais distâncias, é notório na descrição dos métodos.

Figura 13 – Função vegdist

```
vegdist(x, method="bray", diag=FALSE, upper=FALSE,
        na.rm = FALSO, ...)
```

Fonte: Jari Oksanen, com contribuições de Tyler Smith (índice Gower), Michael Bedward (índice Raup-Crick) e Leo Lahti (Aitchison e distância robusta de Aitchison).

x: Matriz de dados da comunidade.

method: Índice de dissimilaridade, correspondência parcial para "manhattan", "euclidean", "canberra", "clark", "bray", "kulczynski", "jaccard", "gower", "altGower", "morisita", "horn", "mountford", "raup", "binomial", "chao", "cao", "mahalanobis", "chisq", "chord", "hellinger", "aitchison", ou "robust.aitchison".

diag: Calcular as diagonais.

upper: Retorne apenas a diagonal superior.

na.rm: Exclusão de pares de observações ausentes ao calcular dissimilaridades.

7 APLICAÇÃO

O banco de dados utilizado foi uma base de dados com as 91 principais faixas no Spotify.Os dados contêm cerca de 12 colunas e focado no gênero rock dos anos de 2009-2019. Com as seguintes variáveis:

song: Nome da Faixa.

duration ms: Duração da trilha em milissegundos.

year: Ano de lançamento da faixa.

popularity : quanto maior o valor, mais popular é a música.

danceability: Dançabilidade descreve o quão adequada uma faixa é para dançar com base em uma combinação de elementos musicais. Um valor de 0,0 é o menos dançável e 1,0 é o mais dançável.

energy: A energia é uma medida de 0,0 a 1,0 e representa uma medida perceptiva de intensidade e atividade.

key: A tonalidade em que a faixa está. Os inteiros são mapeados para pitchs usa-se a notação padrão de Pitch Class. Por exemplo, $0 = C$, 1 = C/D, 2 = D, e assim por diante. Se nenhuma chave foi detectada, o valor é -1.

Loudness: O volume geral de uma faixa em decibéis (dB). Os valores de volume são calculados em média em toda a faixa e são úteis para comparar o volume relativo das faixas. A sonoridade é a qualidade de um som que é o principal correlato psicológico da força física (amplitude). Os valores geralmente variam entre -60 e 0 db.

Speechiness: Speechiness detecta a presença de palavras faladas em uma faixa. Quanto mais exclusivamente falada a gravação (por exemplo, talk show, audiolivro, poesia), mais próximo de 1,0 o valor do atributo. Valores acima de 0,66 descrevem faixas que são provavelmente feitas inteiramente de palavras faladas. Valores entre 0,33 e 0,66 descrevem faixas que podem conter música e fala, seja em seções ou em camadas, incluindo casos como música rap. Os valores abaixo de 0,33 provavelmente representam músicas e outras faixas que não são de fala.

acousticness: Uma medida de confiança de 0,0 a 1,0 se a faixa é acústica. 1.0 representa alta confiança de que a faixa é acústica.

instrumentalness: prevê se uma faixa não contém vocais. Os sons "Ooh"e "aah"são tratados como instrumentais neste contexto. Faixas de rap ou de palavras faladas são claramente "vocais". Quanto mais próximo o valor de instrumentalidade estiver de 1,0, maior a probabilidade

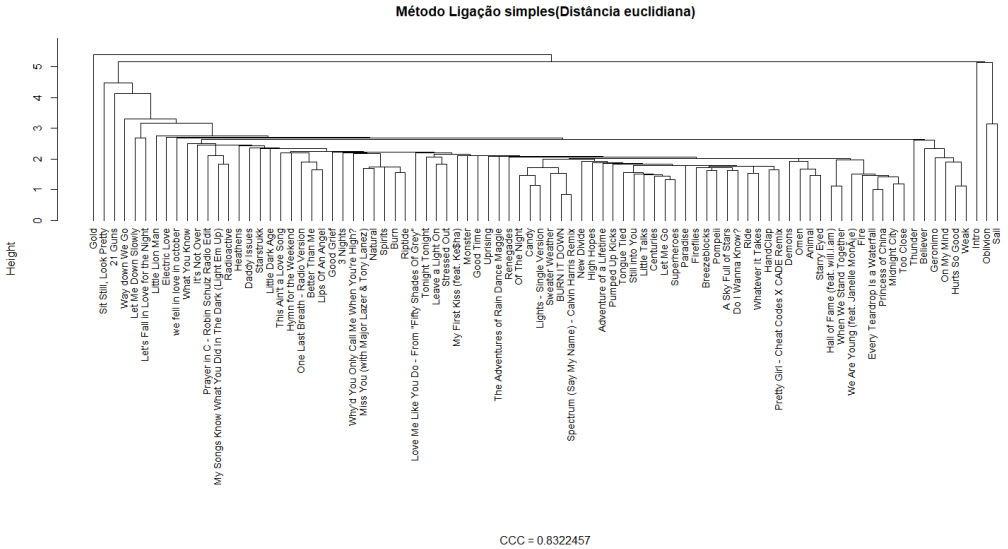
de a faixa não conter conteúdo vocal. Valores acima de 0,5 destinam-se a representar faixas instrumentais, mas a confiança é maior à medida que o valor se aproxima de 1,0.

liveness: Detecta a presença de uma audiência na gravação. Valores mais altos de vivacidade representam uma probabilidade maior de que a faixa tenha sido executada ao vivo. Um valor acima de 0,8 fornece uma forte probabilidade de que a faixa esteja ativa. Foram aplicados os métodos apresentados neste presente trabalho comparando duas ditâncias que são respectivamente euclidiana e manhattan.

7.1 Método de ligação Simples

Pode-se observar a disposição dos grupos Utiliza-se o método de ligação simples e a distância euclidiana. Também consta na imagem o coeficiente de correlação cofenética (CCC) que nesse caso é maior que 0,7, ou seja, pode-se considerar que o método de agrupamento foi adequado.

Figura 14 – Método ligação simples com distância euclidiana

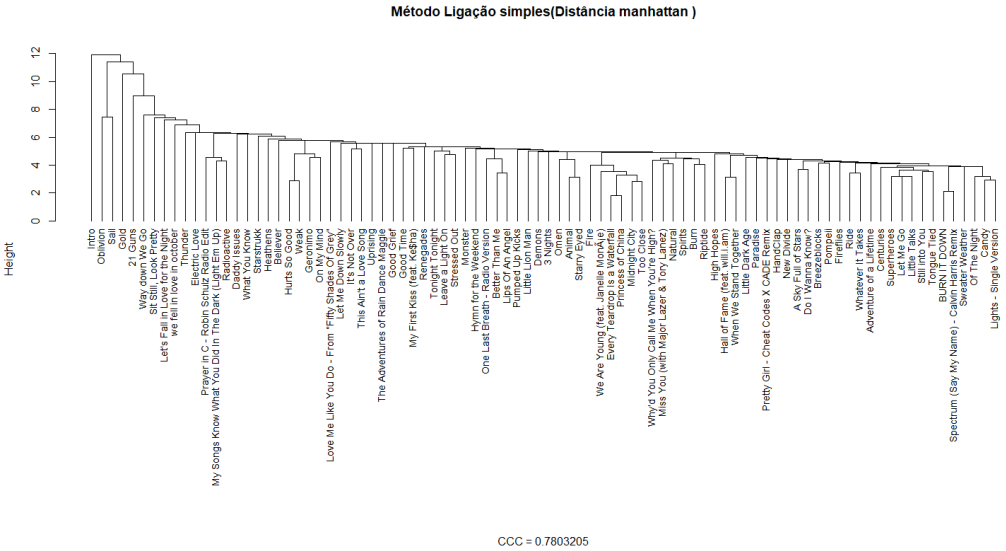


Fonte: elaborado pela autora (2022).

Utiliza-se o método de ligação simple e a distância de manhattan tem-se o seguinte dendograma. Conclui-se que o método é adequado, pois o Coeficiente de correlação cofenética (CCC) é maior que 0,7.

Observa-se a diferença que a utilização das distâncias pode influenciar na construção do dendograma e nas divisões formadas.

Figura 15 – Método ligação simples com distância de manhattan

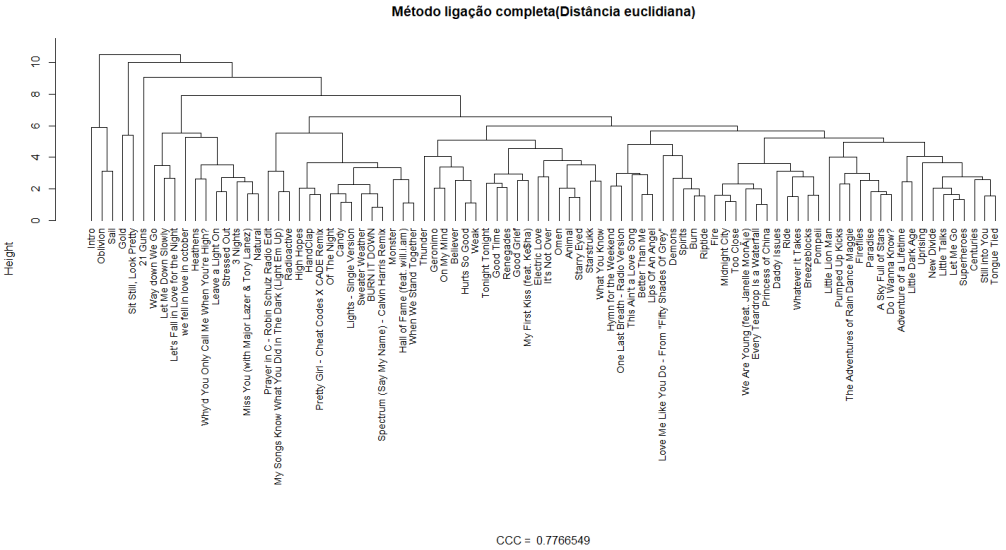


Fonte: elaborado pela autora (2022).

7.2 Método de ligação da Completa

Utilizando o método ligação completa e a distância euclidiana foi obtido o dendograma a seguir. É notório que o método é adequado pelo valor de seu coeficiente.

Figura 16 – Método ligação completa com distância euclidiana



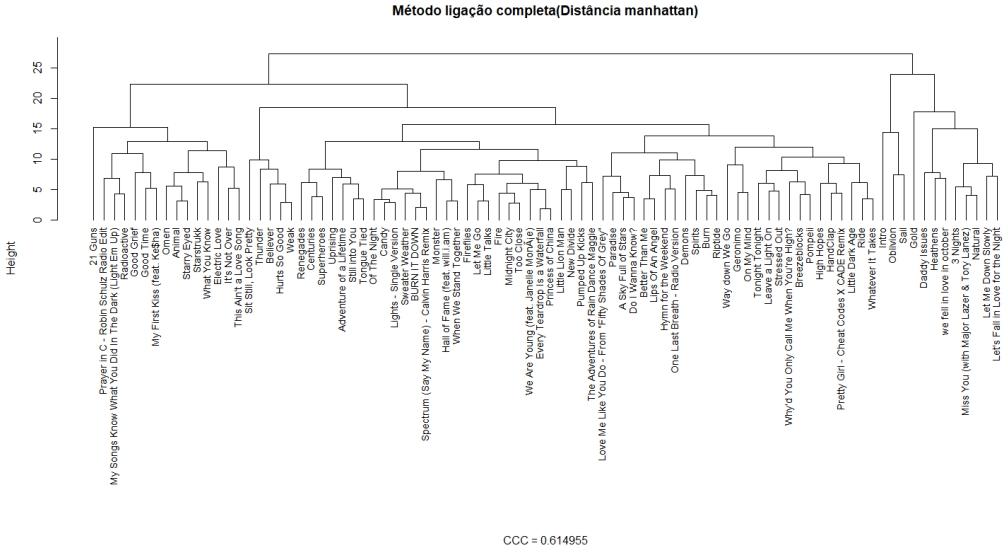
Fonte: elaborado pela autora (2022).

Com a distância de manhattan diferente do que vimos na euclidiana o método não é adequado para uso por conta que o Coeficiente de correlação cofenética (CCC) está abaixo de 0,7.

Pode-se notar que a escolha da distância pode influenciar além das disposições do

grupo, também interfere se o método é adequado para uso.

Figura 17 – Método ligação completa com distância de manhattan

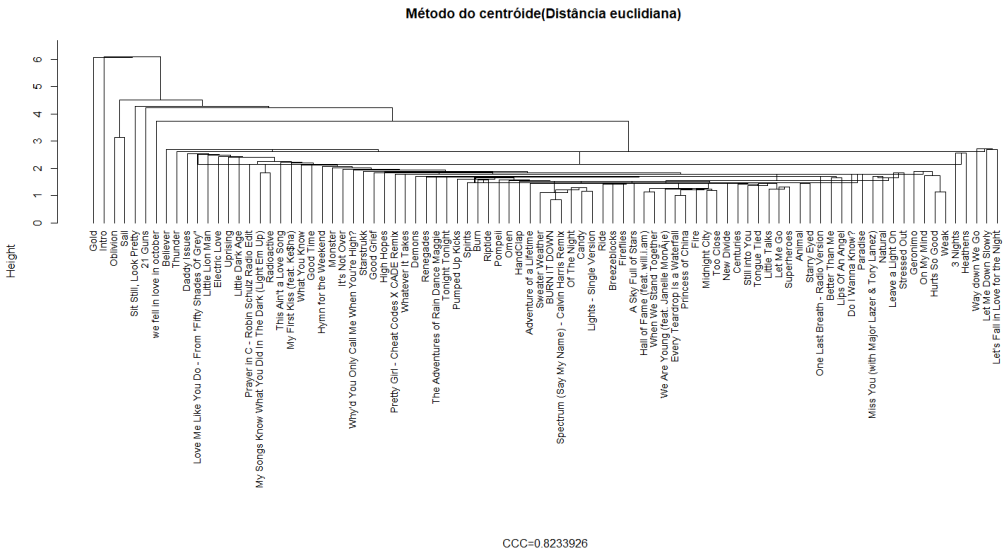


Fonte: elaborado pela autora (2022).

7.3 Método do centróide

Neste momento foi utilizado o método do centroide com a distância euclidiana para a construção do dendograma. É possível notar que o método é adequado.

Figura 18 – Método centróide com distância euclidiana

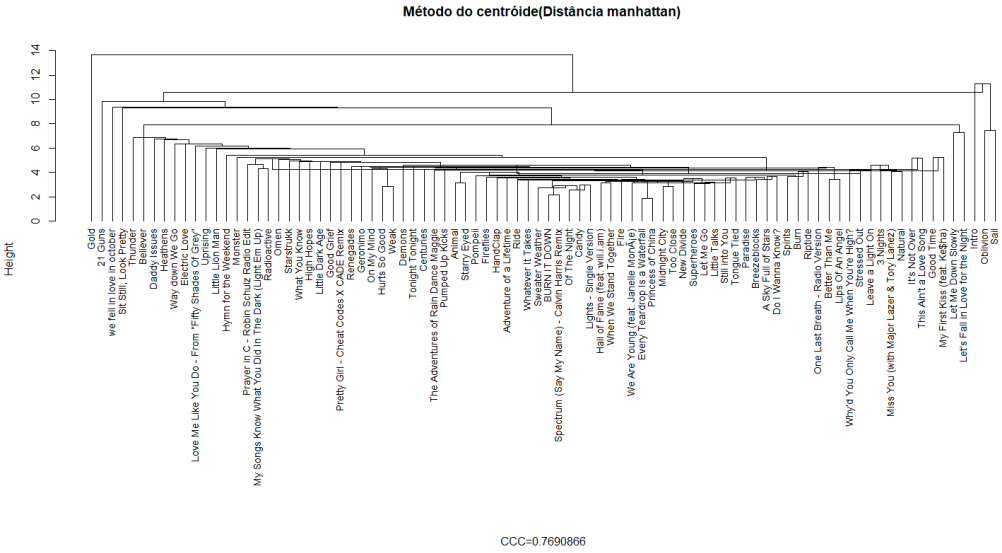


Fonte: elaborado pela autora (2022).

É fácil ver as diferenças entre os dendogramas em que Utilizam-se as distâncias distintas. Entretanto, os dois coeficientes de correlação cofenética (CCC) nos revela que os dois

estão adequados

Figura 19 – Método centróide com distância de manhattan

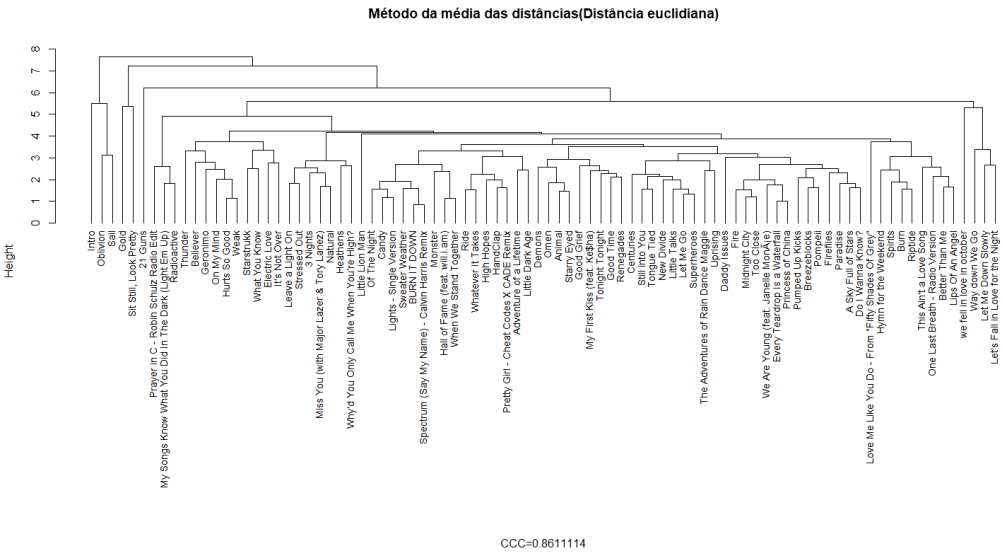


Fonte: elaborado pela autora (2022).

7.4 Método da média das distâncias

Aplica-se o método da média das distâncias com a distância euclidiana e têm-se o dendograma abaixo.

Figura 20 – Método da média das distâncias com distância euclidiana

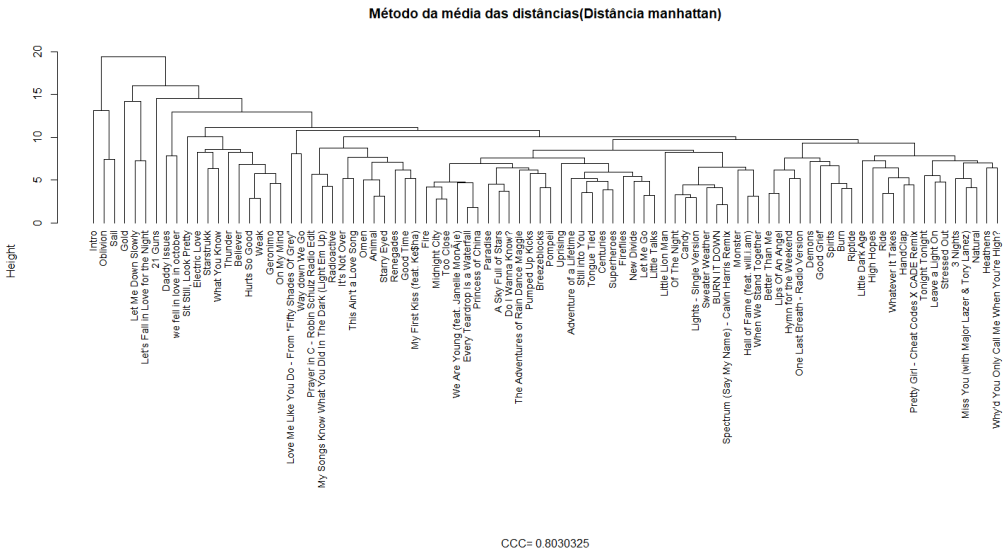


Fonte: elaborado pela autora (2022).

Aplica-se o método da média das distâncias com a distância de manhattan têm-se o dendograma abaixo. Entretanto, por mais diferenças nos dois gráficos apresentados é fácil ver

que os dois métodos estão adequados.

Figura 21 – Método da média das distâncias com distância de manhattan

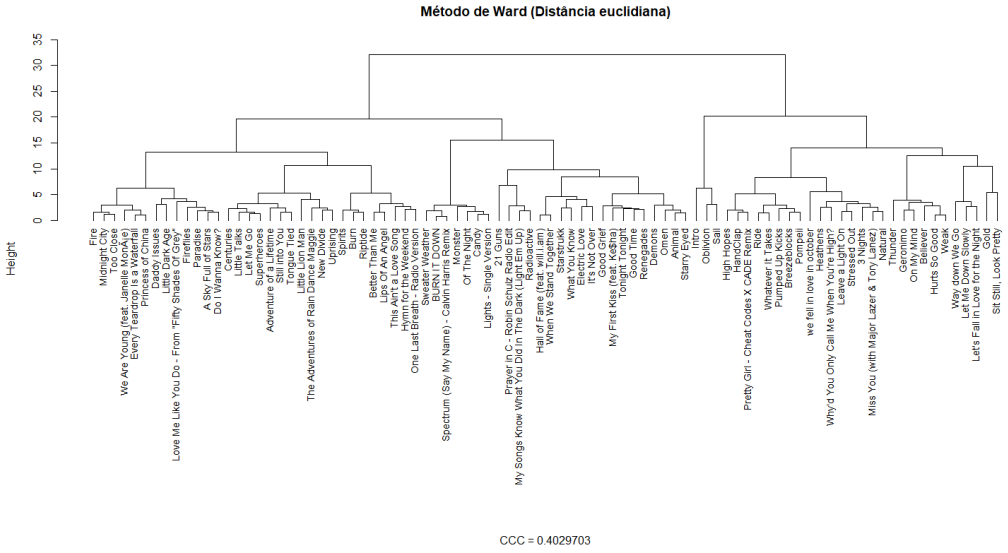


Fonte: elaborado pela autora (2022).

7.5 Método Ward

O dendograma a seguir foi utilizado o método de Ward e distância euclidiana, observe:

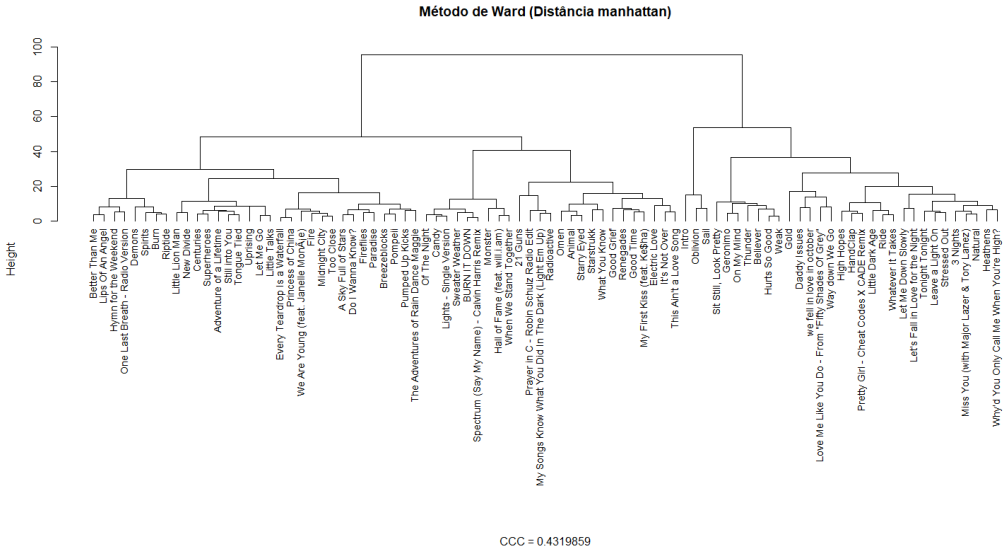
Figura 22 – Método ward com distância euclidiana



Fonte: elaborado pela autora (2022).

Pode-se ver que o método de ward em nenhum dos casos das distâncias euclidiana e manhattan, o CCC é maior que 0,7, ou seja, não é adequado o método para este caso.

Figura 23 – Método ward com distância manhattan

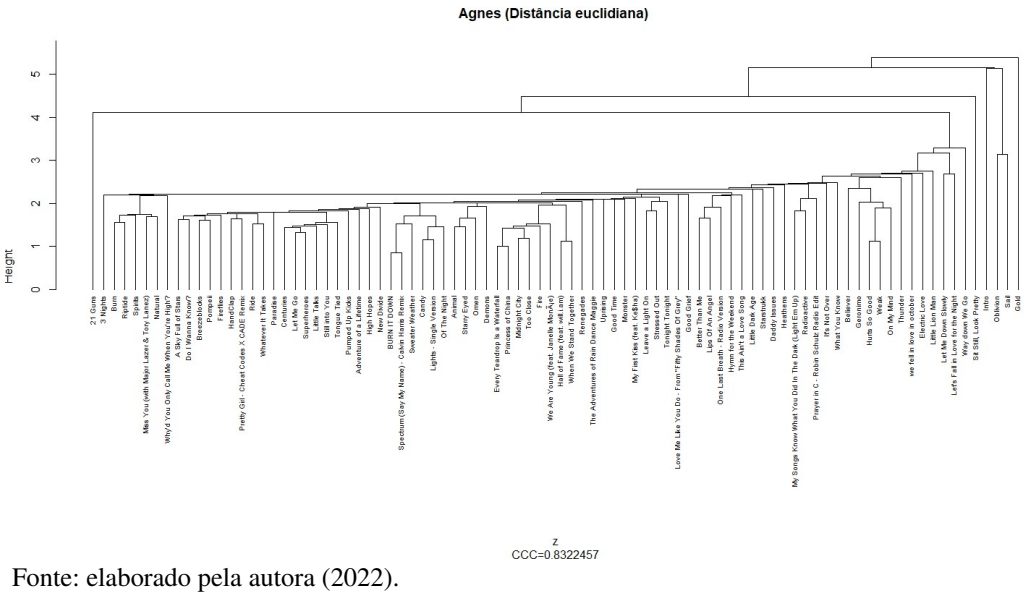


Fonte: elaborado pela autora (2022).

7.6 Agnes

Aplica-se o algoritmo agnes para a construção do dendograma têm-se os dois exemplo a seguir com as distâncias euclidiana e manhattan. São notórias as diferenças entre

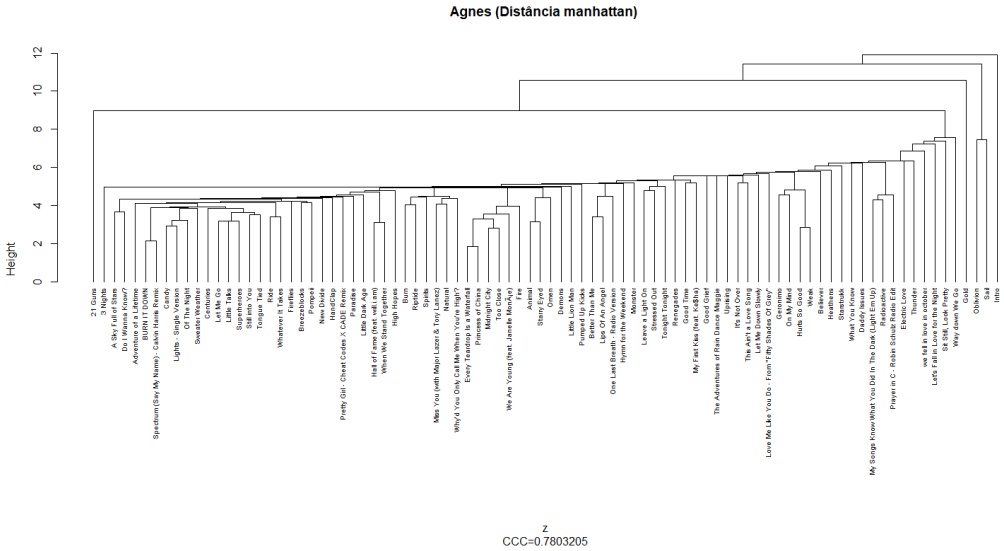
Figura 24 – Agnes com distância eucliana



Fonte: elaborado pela autora (2022).

as disposições dos grupos com as distâncias diferentes. Diante disso, pode-se ver que às duas distâncias aplicada a agnes estão adequadas de acordo com CCC.

Figura 25 – Agnes com distância manhattan

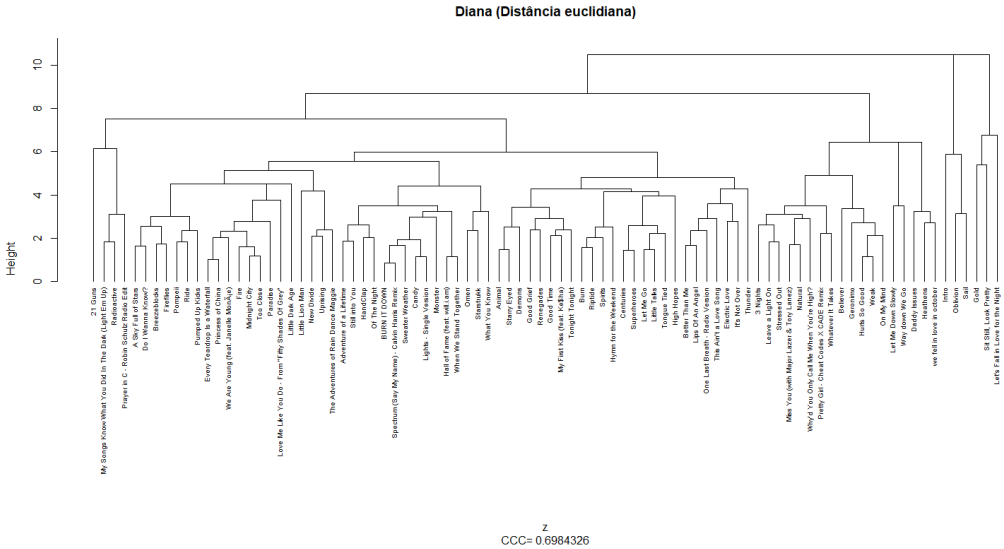


Fonte: elaborado pela autora (2022).

7.7 Diana

O algoritmo Diana foi aplicado as duas distâncias eucliana e mahattan, pode-se observar nos gráficos a seguir.

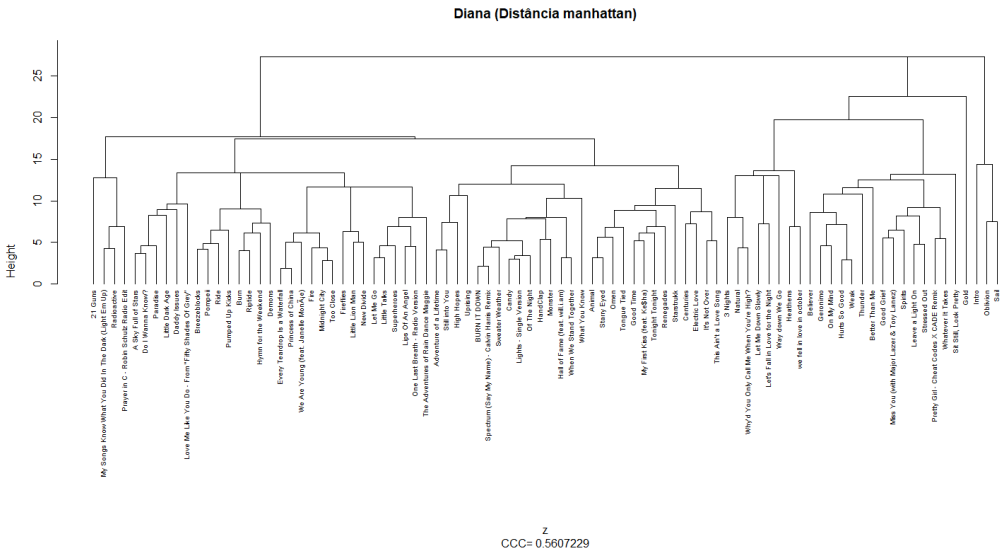
Figura 26 – Diana com distância euclidiana



Fonte: elaborado pela autora (2022).

O coeficiente de correlação cofenética (CCC) nos dois casos aplicados a Diana não foram considerados adequados, pois estão abaixo de 0,7.

Figura 27 – Diana com distância manhattan



Fonte: elaborado pela autora (2022).

8 CONCLUSÃO

A presente monografia exibiu coeficientes de similaridade e de associação, salientamos os tipos de variáveis para cada coeficiente, pois na rotina do profissional que lida diretamente com dados, existem uma variedade de maneiras em que as informações podem chegar até ele e visando isso, foram apresentadas variadas formas de medir a distância entre os elementos para que assim possa ser introduzido os métodos de agrupamento.

Conforme os textos apresentados anteriormente os métodos de agrupamento hierárquicos são eficazes para separação de grupos, com intuito de formação de perfil, estudos de comportamento, pesquisas climáticas, entre outros.

E para a escolha de tal método é necessário a visualização e validação, de como cada método se comporta com seus dados em específico. A seleção do método de agrupamento está diretamente ligada as variáveis, por conta disso é de suma importância saber como seus dados se comportam em cada método e se adéquam. Diante disso, é interessante a visualização com o auxílio de dendogramas e o cálculo do coeficiente de correlação cofenética.

Na aplicação foi possível notar como os mesmos dados podem se comportar de maneiras distintas mudando os métodos e as distâncias. Existem casos que o mesmo método com distâncias diferentes se comportam de forma diferente, por exemplo, um fica adequado e o outro não. Da mesma maneira, em que comparamos a mesma distância com métodos diferentes pode acontecer o mesmo. Além na diferença na disposição na separação dos grupos.

Portanto, deseja-se que este trabalho contribua de alguma forma nos estudos sobre análise agrupamento hierárquico e facilite o entendimento de suas etapas.

REFERÊNCIAS

Antoine Lucas e Sylvain Jasson, **Using amap and ctc Packages for Huge Clustering** , R News, 2006, vol 6, edição 5 páginas 58-60.

Bem, J. S. ; Giacomin, N. M. R. ; Waismann, M. **Utilização da técnica da análise de clusters ao emprego da indústria criativa entre 2000 e 2010: estudo da Região do Consinos, RS.**Campo Grande, v. 16, n. 1, p. 27-41, 2015

Cao Y, WP Williams, AW Bark. 1997. **Similarity measure bias in river benthic Aufwuchs community analysis.** Water Environment Research, 69(1): 95-106.

CRISPIM, D. L.; Fernandes, L. L.; Albuquerque, R. L. **Aplicação de técnica estatística multivariada em indicadores de sustentabilidade nos municípios do Marajó-PA.** Revista principia n. 46 João Pessoa, 2019. n. 4 (2009) pp. 18-36

Dellegosti Nell, M.;Henning , F. A. ;Mertz , L. M. ; Kopp, M. M. ; CRESTANI, M. ; Ivan Schuster2 ; ZIMMER, P. D. **Dissimilaridade ganética em população segregante de soja com variabilidade para caracteres morfológicos de semente.** Revista Brasileira de Sementes, vol. 33, nº 4 p. 689 - 698, 2011

FERREIRA, D. F **Estatística Multivariada.** Lavras : Ed. UFLA, 2008. 662 p. : il.

Ferreira, R. R. M. ;Paim, F. A. de P. ;Rodrigues, V. G. S. ;Castro, G. S. A. **Aná-lise de cluster não supervisionado em R: agrupamento hierárquico .** Campinas: Embrapa Territorial, 2020. 43 p.: il. ; (Documentos / Embrapa Territorial, ISSN 0103-7811; 133).

Grayson, D.K. and S.D. Livingston. 1993. **Missing mammals on Great Basin mountains:Holocene extinctions and inadequate knowledge.** Conservation Biology 7:527-532.

GN Lance e WT Williams, **Uma Teoria Geral de Estratégias de Classificação Classificatória 1. Sistemas Hierárquicos** The Computer Journal, vol. 9, nº 4, 1967, pp. 373-380.

Gower JC. 1966. **Some distance properties of latent root and vector methods used in multivariate analysis.** Biometrika, 53 (3/4): 325-338.

Gower JC. 1971. **A general coefficient of similarity and some of its properties.** Biometrics, 857-871.

Gower JC. 1982. **Euclidean distance geometry.** Mathematical Scientist, 7(1): 1-14.

Hamann U. 1961. **Merkmalsbestand und verwandtschaftsbeziehungen der farinosae: ein beitrag zum system der monokotyledonen.** Willdenowia, 639-768.

Jaccard P. 1900. **Contribution au problème de l'immigration post-glaciare de la flore alpine.** Bulletin de la Societe Vaudoise des Sciences Naturelles, 36: 87-130

Jari Oksanen [aut, cre], Gavin L. Simpson , F. Guillaume Blanchet , Roeland Kindt , Pierre Legendre , Peter R. Minchin , RB O'Hara [aut], Peter Solymos , M. Henry H. Stevens , Eduard Szoecs , Helene Wagner , Matt Barbour , Michael Bedward , Ben Bolker , Daniel Borcard , Gustavo Carvalho , Michael Chirico , Miquel De Caceres , Sebastien Durand , Heloisa Beatriz Antoniazi Evangelista , Rich FitzJohn , Michael Friendly , Brendan Furneaux , Geoffrey Hannigan , Mark O. Hill , Leo Lahti , Dan McGlinn , Marie-Helene Ouellette , Eduardo Ribeiro Cunha , Tyler Smith [aut], Adrian Stier , Cajo JF Ter Braak , James Weedon .2022 **Vegan: Pacote de Ecologia Comunitária** https://CRAN.Rproject.org/package=Rraven

Kaufman L, PJ Rousseeuw. 2009. **Finding groups in data: an introduction to cluster analysis** John Wiley and Sons. Hoboken.

LINDEN, R. **Técnicas de Agrupamento.** Revista de Sistemas de Informação da

FSMA n. 4, 2009 pp. 18-36

Nietto, P. R. ; Sampaio, H V. **O Uso do Algoritmo de Agrupamento Hierárquico Divisivo DIANA em uma Rede de Sensores Sem Fio Aplicada à Agricultura.** Faculdade Campo Limpo Paulista – FACCAMP Campo Limpo Paulista – SP, 2016

Macnaughton-Smith, P. (1965): **Some statistical and other numerical techniques for classifying individuals**, Home Office Res. Rpt. No. 6 (H.M.S.O., London)

Mahalanobis PC. 1936. **On the generalised distance in statistics.** Proceedings of the National Institute of Science of India, 12(1936): 49-55.

METZ, J. **Interpretação de clusters gerados por algoritmos de clustering hiérarquicos.** São Carlos : Ed. USP - São Carlos, 2006.

Meyer D, C Buchta. 2019. **proxy: distance and similarity measures. Paquete de R versión 0.4-23.** https:// CRAN.R-project.org/package=proxy

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

PALACIO,A. X.; APODACA,M.J.;CRISC. J. V. **Análisis multivarido para datos biológicos: Teoría y su aplicación utilizando el lenguaje R.** ad Autónoma de Buenos Aires : Fundación de Historia Natural Félix de Azara, 2020.

R Core Team and contributors worldwide **stats : The R Stats Package, Version: 4.3.0** 2022

RODRIGUES, F. S.; **Métodos de agrupamento na análise de dados de expressão genética.** São Carlos - UFSCar, 2009 93 f.

Rogers DJ, TT Tanimoto. 1960. **A computer program for classifying plants.**

Science, 132(3434): 1115- 1118

Sokal RR, FJ Rohlf. 1962. **The comparison of dendrograms by objective methods.** Taxon, 11: 33-40.

Sokal RR, PH Sneath. 1963. **Principles of numerical taxonomy.** WH Freeman Company. San Francisco.

Sokal RR. 1961. **Distance as a measure of taxonomic similarity.** Systematic Zoology, 10(2): 70-79.

Sokal RR, Michener D. **A statistical method for evaluation systematic relationships.** University of Kansas Scientific Bulletin, v. 38, n. 22, p. 1409-1438, 1958.

VALE, M. N.; **Agrupamentos de dados : avaliação de métodos e desenvolvimento de aplicativo para análise de grupos.** Rio de Janeiro : PUC, Departamento de Engenharia Elétrica, 2005.

Ward Jr JH. 1963. **Hierarchical grouping to optimize an objective function.** Journal of the American Statistical Association, 58(301): 236-244