Revista Ibero- Americana de Humanidades, Ciências e Educação- REASE

doi.org/10.51891/rease.v9i9.11422

COMPARAÇÃO DAS TÉCNICAS DE AGRUPAMENTO: ESTUDO DE CASO EM DADOS DE VACINAÇÃO E MORTALIDADE INFANTIL ENTRE OS ANOS DE 2011 A 2021

COMPARISON OF CROUPING TECHNIQUES: A CASE STUDY ON VACCINATION AND CHILD MORTALITY DATA BETWEEN 2011 AND 2021

Mácio Augusto de Albuquerque[1]
Fabiano Florentino dos Santos[2]

RESUMO: O presente trabalho tem por objetivo mostrar como pode ser feita a análise de cluster, através da técnica hierárquica e não hierarquia. Foram utilizados dados de vacinação e mortalidade infantil de crianças na faixa etária de idade de 28 dias até um ano de vida em um recorte temporal que vai de 2011 a 2021, disponíveis no DATASUS dos estados brasileiros através dos números de infectados de cada estado para assim identificar a similaridades entre os estados através dos números de infectados, oferecendo um contraponto ao critério utilizado de análise do número de infectados dos estados, se baseando no tamanho da população e comparando com a sua população. Para a análise de agrupamento foi utilizado a matriz de Euclidiano com o método hierárquico, aplicou-se os métodos de ligação simples, completa, média, ligação de ward e um método não hierárquico através do método de K-means, também foram aplicados os métodos de determinação do número ideal de grupos como os métodos como cotovelo, coeficiente de silhueta média e Rand ajustado o coeficiente de correlação confénetica para medir o grau de ajuste entre as matrizes original e a matriz resultante da simplificação proporcionada dendrogramaa. No entanto foi verificado o método que melhor representa os dados é o de Ward. Ao agrupar os estados de ambos os dados levou em consideração a semelhança entre as variáveis dos dados e a correlação onde pode se observar que os dados são correlacionados.

Palavras-chave: Métodos. Vacinação. Cluster. Mortalidade.

ABSTRACT: The present work aims to show how cluster analysis can be carried out, using the hierarchical technique and not hierarchy. Data on vaccination and infant mortality of children aged between 28 days and one year of life were used in a time frame ranging from 2011 to 2021, available in DATASUS of the Brazilian states through the numbers of infected people in each state, in order to identify similarities between states through the numbers of infected people, offering a counterpoint to the criterion used to analyze the number of infected people in states, based on the size of the population and comparing it with their population. For cluster analysis, the Euclidean matrix was used with the hierarchical method, the simple, complete, average, ward linkage methods and a non-hierarchical method through the K-means method were applied, the methods of determining the ideal number of groups using methods such as elbow, average silhouette coefficient and Rand adjusted conphenetic correlation coefficient to measure the degree of adjustment between the original matrices and the matrix resulting from the simplification provided by the dendrogram. However, the method that best represents the data was found to be Ward's. When grouping the states of both data, the similarity between the data variables and the correlation were taken into account, where it can be seen that the data are correlated.

Keywords: Methods. Vaccination. Cluster. Mortality.

[1]Doutor em Biometria aplicada em Estatístic, Departamento de Estatística, Universidade Estadual da Paraíba, Campina Grande, Paraíba, Brasil, ORCID: https://orcid.org/0000-0002-0113-9130.
[2] Estudante de Graduação em Estatística, Departamento de Estatística, Universidade Estadual da Paraíba, Campina Grande, Paraíba, Brasil.

INTRODUÇÃO

Mediante a inúmeras pesquisas, estudos e questionamentos percebemos com um grande volume de informações mediante análise e identificação de padrões, sendo as estratégias de machine learning sendo dividido com os tipos incluindo aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. Deep learning com subárea de ML onde se faz uso de redes neurais artificiais com múltiplas camadas com o intuito de aprender cada vez mais dados cada vez mais complexos e contexto do trabalho em questão se faz uso de técnicas de aprendizado não supervisionado onde apenas especificamos o que se feito automaticamente independentemente do programa. Recordamos que a necessidade de agrupar é algo intuitivo e inerente ao ser humano se levarmos em consideração que para sobreviver e compreender os fenômenos precisamos ordenar segundo algum critério e posteriormente classificá-los. Podemos citar casos da vida cotidiana, por exemplo, quando uma criança na escola ao receber um conjunto de lápis para pintar um desenho, seleciona as principais cores de seu gosto, para daí começar a pintar. Outro exemplo é quando um assistente administrativo de uma empresa está organizando documentos em determinado setor, é normalmente utilizado sistemas computacionais onde determinadas pastas se localizam separadas e listadas pelos nomes para otimizar o tempo e organização, isso por si só se caracteriza uma segmentação, precisando tomar os devidos cuidados, pois cada documento tem uma característica semelhante ou não em relação às demais. Em áreas das ciências como a biologia, na antiguidade estudos sobre a taxonomia que dizem respeito à classificação dos seres vivos, é importante enfatizar que um dos primeiros a sugerir um modelo taxonômico foi o filósofo grego Aristóteles (384-322 a.C.). Em outras palavras podemos dizer que tanto na ciência como na vida cotidiana das pessoas a classificação é algo primordialmente necessário em vários ramos, por isso a necessidade de aplicação de técnicas para tanto dentro da estatística. Com isso em mente as técnicas multivariadas de associação, tem com base critérios de classificações, sendo essas de escolha por parte do pesquisador visando reunir parcelas/objetos em parte de população, com é o caso de amostras sem interesse de inferir a priori resultados para universo populacional, ou o nosso caso em especial onde trabalhamos com não com uma parte restrita, mas números relativos ao total da população com é o caso do presente trabalho. Procedimentos como esses exigem decisões de forma independente do pesquisador, pois a, escolhas de medidas, forma de associar unidades e validação acerca da

qualidade da reunião dos mesmos, irá consequentemente levar a tipos de agrupamentos distintos, também a seleção das variáveis visto que essas podem influenciar diretamente no estudo, e outliers caso o pesquisador deixar ou não a presença delas (DUARTE, 2021). A presença desses valores destoantes/discrepantes, onde basicamente são valores não muito comuns se comparado aos demais, por possuírem características distintas dos demais pode levar a distorção dos resultados (ALBUQUERQUE & BARROS, 2020). Portanto, a experiência carregada frente a escolhas corretas levará algo que represente bem uma realidade, como consequência da escolha do melhor modelo a ser representado.

Tendo em vista a grande relevância do monitoramento das doenças preexistentes e as principais causas que levam à evolução ou não das mesmas, quais os principais fatores que tendem a aumentar, diminuir determinado espectro de uma doença. Um exemplo mais recente a ser citado seria a pandemia do covid-19 e sua vacinação, nesse contexto houve a necessidade de mensuração de grande volume de dados para que pudéssemos estudar quais características comportamentais diminuiria ou não a proliferação do vírus no meio social, com isso pude concluir que se mediante a tais estudos, o uso de máscaras, álcool em gel e consequente imunização massiva auxiliaria no "quase fim" do vírus ou não mais no perigo iminente do grau elevado do número de mortes. Tendo isso em conta, estudo acerca da saúde vem de encontro a qualidade de vida das pessoas, logicamente não apenas benefícios numa dada minoria estudada na pesquisa, mas toda uma população.

Usando como campo de pesquisa a rede de saúde pública, os dados aqui expostos foram colhidos do Departamento de Informática do Sistema Único de Saúde (DATASUS), mais especificamente usando como base referencial consideração casos de óbitos, e imunização de crianças numa dada faixa etária de idade de 28 dias até no máximo um ano de vida em um recorte histórico que vai de 2011 a 2021, ao nível de observacional serão estudados os estados e unidade federativa Distrito Federal.

Neste trabalho apresentamos as técnicas hierárquica e não hierárquica, os pressupostos de como cada uma funciona e as suas diferenças, a características de cada parecença, quando é preferível utilizar ou destacar, ou seja, que se adapte melhor ao contexto das informações que se queira aplicar, além de estratégias para determinação e avaliação dos números de grupos que norteiam a escolha da melhor estratégia ou técnica. A parte de programação envolvida no apêndice, os pacotes usados como *ggplot2*, *factoextra*, *cluster* dentre outros, especificamente discutindo cada um, os parâmetros para aplicabilidade.

MATERIAL E MÉTODOS

Análise de Agrupamento

A técnica multivariada de análise de agrupamento permite uma análise de interdependência entre as variáveis, fazendo com que elas sejam agregadas a partir de características comuns que possui. Segundo Albuquerque & Barros (2020), para o estudo de analisar de agrupamento utilizar o método de dissimilaridade baseado na distância de Euclidiana), considerada uma das distâncias mais utilizadas, podendo ser calculada de acordo com a expressão a seguir:

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

em que:em que X_{iv} representa a característica do indivíduo i, X_{jv} representa a característica do individuo j, p é o número de parcelas na amostra, v é o número indivíduo na amostra.

Método Hierárquico

No método hierárquico, o foco não está no número exato de clusters, mas sim no agrupamento a ser analisado, cuja construção se baseia em um cluster maior e dividindo as observações em clusters menores, ou um de cada. a observação é um conglomerado e será agrupada em grupos maiores nas etapas seguintes, com os critérios para esses agrupamentos variando de acordo com a técnica (DUARTE,2021).

Por ser um técnica bastante utilizado e fácil de ser encontrado em alguns programas computacional, as técnicas de algoritmos utilizado segundo Costa (2019) são: Método de ligação Simples que é definida pelos dois elementos mais parecidos entre si; Completa que é definida como sendo a distância entre os vetores de médias; Média 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 e o Método de ligação Ward que pode formar os grupos a partir da maximização da homogeneidade dentro dos grupos ou a minimização total da soma de quadrados dentro dos grupos.

Método de ligação simples: Por ser um dos algoritmos, mas antigo e mais simples de utilizado na literatura, denominado "método do vizinho mais próximo" esse é uma

técnica de hierarquização aglomerativa e tem, como uma de suas características, não exigir que o número de agrupamentos seja fixado a priori. Segundo Souza (2022) a distância entre dois grupos é determinada pela distância mínima entre os pares de elementos desses grupos e aquele com a menor distância mínima é agrupado, ou seja, se dois grupos $C_1 = \{X_1, X_3, X_7\}$ e $C_2 = \{X_2, X_6\}$, a distância entre os grupos é definida por:

$$d(C_1, C_2) = \min\{d(X_l, X_k, l \neq k, l = 1,3,7 \; e \; k = 2,6)\}$$

Método de ligação completa: Neste método, a distância entre dois grupos é determinada pela distância máxima entre os pares de elementos desses grupos. O método tenta agrupar os elementos que possuem a menor distância entre os mais distantes. Sejam dois grupos $C_1 = \{X_1, X_3, X_7\}$ e $C_2 = \{X_2, X_6\}$ a distância entre os grupos é definida por:

$$d(C_1, C_2) = \max\{d(X_l, X_k, l \neq k, l = 1,3,7 \; e \; k = 2,6)\}$$

Método de ligação Média: Este método, foi originalmente proposto por Sokal e Michener (1958) e é uma ponderação entre os métodos de ligação simples e ligação completa entre todos os pares encontrados. Pode ser formado com os elementos dos dois grupos a serem comparados e agrupa aqueles com a menor distância média (SOUZA, 2022). Por exemplo, se os grupos C1 possuem n1 elementos e C2 com n2 elementos, a distância entre os grupos é dada por:

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

Método de ligação Ward: Segundo Souza (2022) o método de Ward agrupa os elementos que possuem a menor soma dos quadrados das distâncias, é um método que tende a fornecer agregados com aproximadamente o mesmo número de observações, inicialmente cada elemento é considerado um único agrupamento e a cada passo de o algoritmo, o algoritmo calcula a soma dos quadrados dentro de cada cluster de cada elemento pertencente ao cluster, em relação ao vetor médio correspondente do cluster. A distância entre Cl e Ci representa a soma quadrada entre os cluster que pode ser definida por:

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i}\right] (\overline{X}_l - \overline{X}_i)'(\overline{X}_l - \overline{X}_i)$$

Método Não-Hierárquico

O métodos Não-hierárquicos no geral, o foco é encontrar o número "k" de clusters que consiga realizar a divisão das observações de maneira satisfatória, o método que

consiga identificar semelhanças e diferenças entre as observações (DUARTE,2021). Por ser um processo de agrupamento mais dinâmico e interativo o método não hierárquico o número de grupos é especificado antes do processo de agrupamento, o critério, mas utilizado por esse método é o de K-means, segundo Alves (2020) esse método possui algumas condições como a informação prévia dos números de clusters k, onde suas observações são agrupadas nesse k clusters utilizando uma função com objetivo e critério. Sendo de simples aplicação e rápido processamento segundo Duarte (2021), a lógica do algoritmo segue 4 passos:

1. Escolhido o número de grupos, denominado k
2. Dentro dos dados atribui-se a alguma observação aleatória a um cluster, depois utilizando alguma medida de distância, se atribui ao elemento mais próximo o mesmo cluster e calcula-se a média das distâncias, formando o centro do cluster.
3. Recalcular os centroides para cada K clusters, calcula a média de todos os elementos dos grupos.
4. Repete-se o passo 2, e se recalcula o centro do cluster dado o novo objeto que entrou, isso se repete até todos os dados tenham seus respectivos clusters.

Determinação do número de grupos ideal

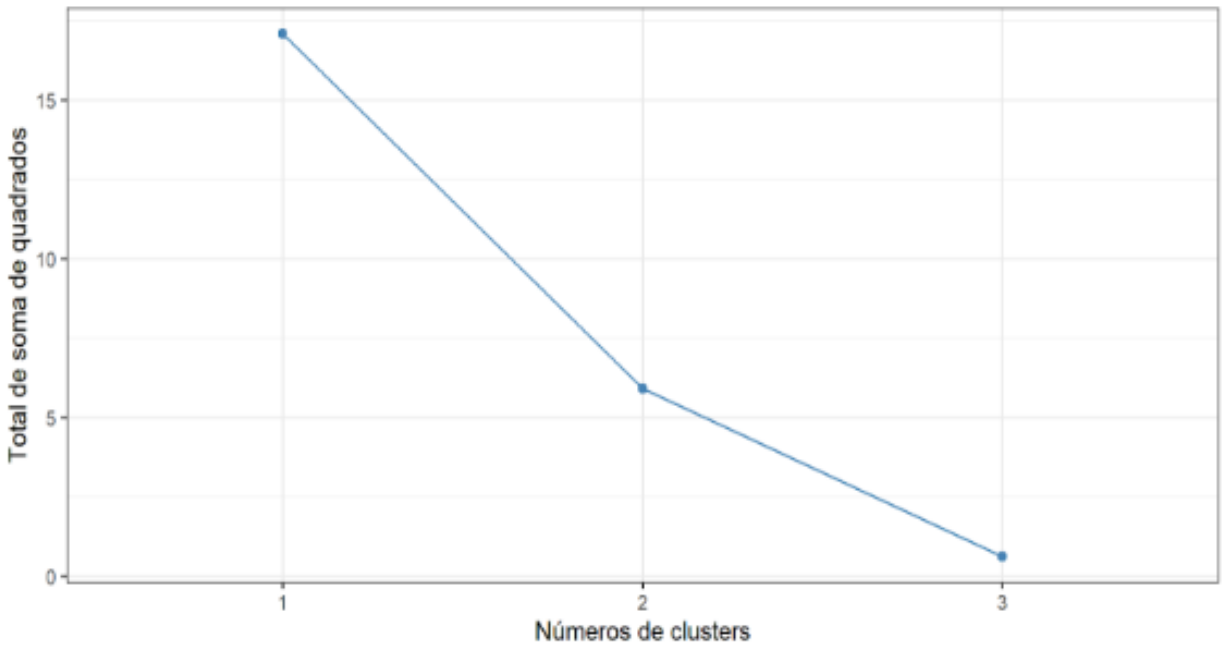
Nem sempre é viável a escolha da quantidade de grupos iniciais de forma subjetiva, sendo sempre preferível partir de estratégias metodológicas convenientes para se chegar a essa quantidade ideal, pois isso vai influenciar o agrupamento posterior, com vimos no K-means e consequente versões alternativas do mesmo como PAM e FCM, como veremos adiante, caso elemento representativo do seu cluster ou mesmo centroide que serve de parâmetro para cada um dos grupos depende não apenas dos elementos, como também desse valor amplamente discutido de K. Além disso, técnicas avaliativas nos ajudam a identificar distorções impostas ou mesmo sobre a qualidade acerca de cada unidade em seu respectivo grupo, como estão localizados e se de fato deveriam ou não está em outros grupos pela baixa dissimilaridade entre os seus elementos do cluster atual.

Método do cotovelo

Discutimos anteriormente que a determinação do número de K-grupos pode ocorrer de maneira a analisar esses “pulos de distâncias” ao longo que visualizamos, por exemplo, o dendrograma, representação dos métodos hierárquicos mais comumente usada nesse tipo

de análise, entretanto dentro da literatura temos outras alternativas para o mesmo fim como o critério do método do cotovelo, aqui vamos considerar agrupamentos com diferentes números de K, onde podemos visualizar ao decorrer que aumentamos o número de K diminuímos consequentemente os valores dos erros quadráticos dos grupos. A melhor partição é justamente onde o gráfico tem um formato de cotovelo, ou seja, onde teve maior um decréscimo considerável a soma erros quadráticos total, observamos também a medida que vamos aumentando o valor de K o erro tende a um valor próximo de zero.

Figura 1 : método do cotovelo

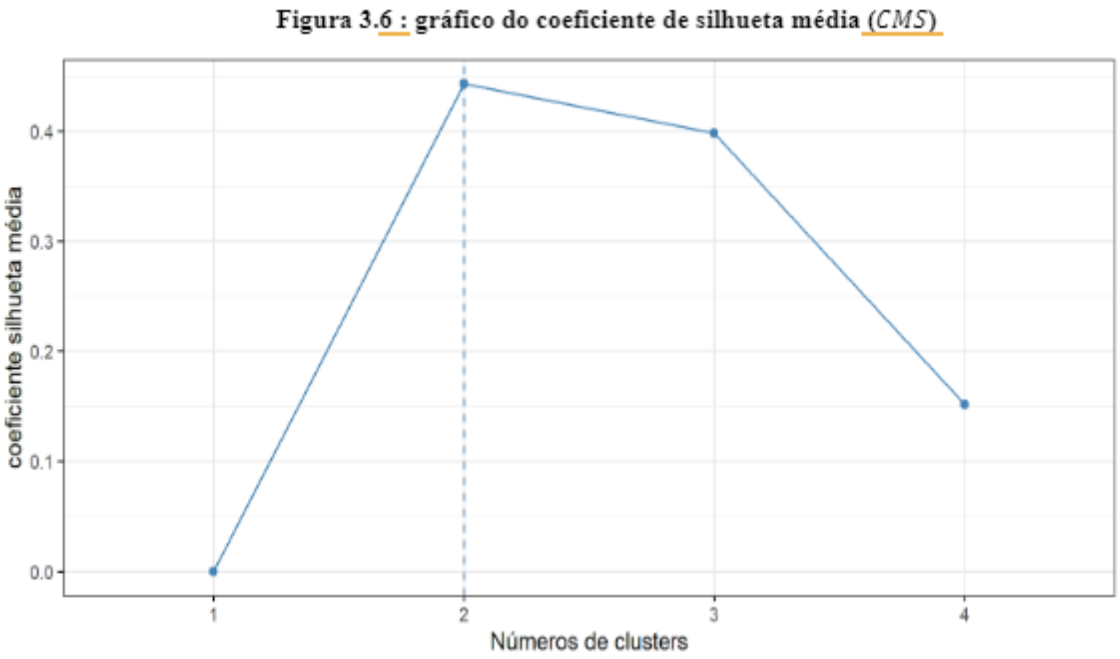


Usando os dados da análise anterior, podemos aplicar e visualizar a quantidade ideal de k, sendo 2 a quantidade ideal na Figura 1, neste ponto o gráfico buscamos a forma de "cotovelo"

Método da silhueta

Outra forma de ir encontro ao valor ideal de k é analisar a Silhueta que apresenta também a ideia de expressar essa quantidade ideal. Essa medida que expressa o quanto que um determinado objeto qualquer é similar a sua conjunção (SAMUEL, 2021). Assim, esse coeficiente varia de numa escala de [-1,1] indicando a divisão de boa qualidade quando s(i) tende ao valor 1.

Figura 2 : gráfico do coeficiente de silhueta



O maior valor índice de silhueta CSM na Figura 2, melhor indica, nesse caso o valor para k = 2.

Índice de Rand ajustado

O índice de Rand permite comparar duas partições com número de grupos não necessariamente iguais. Basicamente, este índice baseia-se no número de pares de parcelas que foram atribuídos da mesma maneira em cada uma das partições, ou seja, baseia-se no número de pares de parcelas concordantes (tipo I e II). Assim, temos o índice de Rand, designado por IR é dado por (RAND, 1971):

$$IR = \frac{a}{a+b+c+d} = \frac{A}{\binom{n}{2}}$$

De um modo mais detalhada tem-se:

$$IR = \frac{\binom{n}{2} + \sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}^2 - \frac{1}{2}\left[\sum_{i=1}^{R} n_{i.}^2 + \sum_{j=1}^{C} n_{.j}^2\right]}{\binom{n}{2}}$$

Verifica-se $0 \leq IR \leq 1$, tomando o valor o quando as duas partições não têm qualquer semelhança (ou seja, quando uma partição é constituída por uma só grupo com todos as parcelas, e a outra é constituída por n grupos com 1 parcela cada) e o valor 1 quando o acordo entre as duas partições é completo.

Correlação cofenética

Para medir o grau de ajuste entre as matrizes similares originais e a matriz resultantes da simplificação será utilizado a correlação cofenética, proporcionada pelo método de agrupamento conforme a expressão segundo Albuquerque, M. et al. (2016):

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

Onde:

Cij é o valor de similaridade entre os indivíduos i e j, onde será obtido a partir da matriz cofenética;

Sij é o valor de similaridade entre os indivíduos i e j, onde serão obtidos a partir da matriz de similaridade.

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

Observa-se que essa correlação corresponde à correlação de Pearson entre a matriz de similaridade original e aquela obtida após a construção do dendrograma, ou seja, utiliza uma escala de o há 1, se c ≤ 0,39 é considerado fraco, se 0,40 ≥c ≤ 0,69 é considerado moderado e se c ≥ 0,70 é considerado forte, de modo que quanto mais próximo de 1, menor a distorção do dendrograma causada pelo agrupamento de indivíduos com algum método hierárquico escolhida.

RESULTADOS E DISCUSSÃO

Após o uso de diferentes métodos de agrupamento, e análise de suas figuras foi verificado de que os grupos construídos diferem entre si, a partir do método hierárquico e aplicação da matriz de distância Euclidiana, precisamos definir a quantidade ideal de k, para maior confiabilidade utilizaremos as técnicas Método do coeficiente silhueta, Método do cotovelo e de Rand e ver se de fato coincidem ou não. foram aplicados os seguintes métodos hierárquicos aglomerativos: vizinho mais próximo, vizinho mais distantes, médias das distâncias e Ward. Numa análise posterior foi também avaliado o método não hierárquico, cada método apresenta suas vantagens e desvantagens, o método hierárquico tem a vantagem de utilizar várias medidas diferentes, sua desvantagem é reduzir o número

de outliers. A vantagem do não hierárquica é usar um conjunto de dados muito grande com menos outliers, mas a desvantagem é usar aleatoriamente o centroide, o que torna o método hierárquico superior a esse método.

Analisando a matriz de distância de euclidiana pelos métodos de ligação, observou-se uma alteração nos níveis dos elementos agrupados, os elementos localizados dentro de cada grupo sua estrutura geralmente é bastante semelhante em relação a cada método utilizado.

Análise dos dados

Utilizados os dados dos 27 estados e do distrito Federal do Brasil referente os dados de informações, possui números acerca de imunização e mortalidade infantil de crianças com idade variando entre 30 dias de vida a um ano, o nível de observação é estadual, logo temos números de 10 anos decorrentes de 2011-2021 e as variáveis disponíveis são Casos confirmados, Óbitos, Incidência e Mortalidade, para caracterizar as variáveis em estudo, foi realizado uma análise descritiva presente

na Tabela 1 medidas descritivas para ambas as variáveis, em todos os estados e distrito federal durante 2011-2021 houve um o maior valor máximo 22197 casos de mortalidade, porém, um total máximo de vacinação de 587562, a média possui, em contrapartida, alto desvio padrão para as variáveis não sendo uma boa opção desrespeito a representatividade, por este motivo optamos por usar a mediana onde cerca de 50% dos casos de vacinas infantis foram menores ou iguais ao valor de 25501, assim como a 50% mortes no mesmo período de idade são menores 2706, ou 50% dos valores estão acimas dessas duas quantidades populacionais.

Tabela 1 : Medidas descritivas

MEDIDAS RESUMO	IMUNIZAÇÃO	ÓBITOS
Média	62214	4471
Mediana	25501	2706
Desvio padrão	114187	4454
Máximo	587562	22197
Mínimo	5160	966

(Figura 3 a 6) com a variável Casos confirmados com a combinação da distância de Euclidiana e métodos de agrupamento (Ligação Simples, Ligação Completa, Ligação Media, Ligação Ward) e a distância Euclidiana com o método de K-means.

Com essa combinação de distância da distância de euclidiana) e métodos de agrupamento foram obtidos o coeficiente de correlação cofenética (CCC) como intuito de medir o grau de ajuste entre as matrizes formadas (Tabela 2), ou seja, a distância de euclidiana e métodos de ligação completa obtiveram o maior valor para o CCC que foi igual a 0,850 como o valor estar próximo de 1 a CCC é considerando forte.

Tabela 2: Coeficiente de correlação cofenético

Ligações	Correlação
Simples	0,742
Completa	0,850
Média	0,814
Ward	0,570

No método de ligação Simples denotado como "método do vizinho mais próximo", podendo observar no dendograma (Figura 3) os 3 cluster formados por este método. No grupo 1 e 2 foram formados por apenas um estadoSão Paulo, nesse grupo podemos perceber que nele se encontra o estado com o maior vacinação, Rio Grande Sul, o terceiro grupo formado pelos estados, Paraíba, Santa Catarina, Tocantins, Mato Grosso, Espírito Santo, Pará e Maranhão, Rondônia, Ceará, Mato Grosso do Sul, Amazona, Distrito Federal, Rio de Janeiro, Rio Grande do Norte, Piauí, Alagoas, Sergipe, Mina Gerais, Amapá, Acre, Roraima, Goiás, Paraná e Bahia, quanto maior for a população do estado maior era o número de vacinação.

Figura 3: Dendrograma na distância euclidiana e o método de ligação Simples.

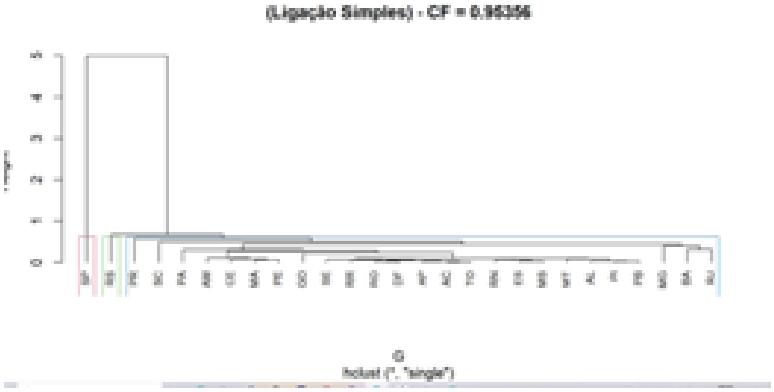
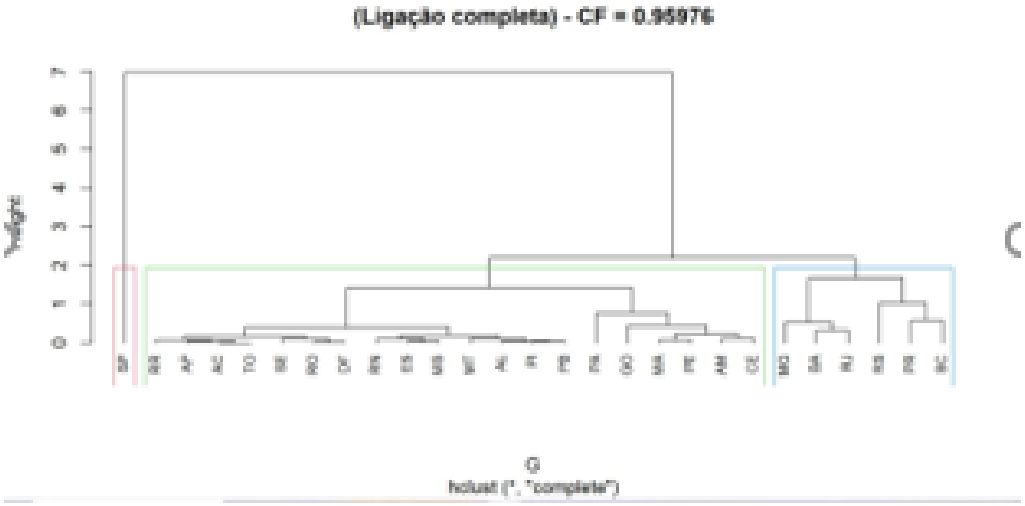
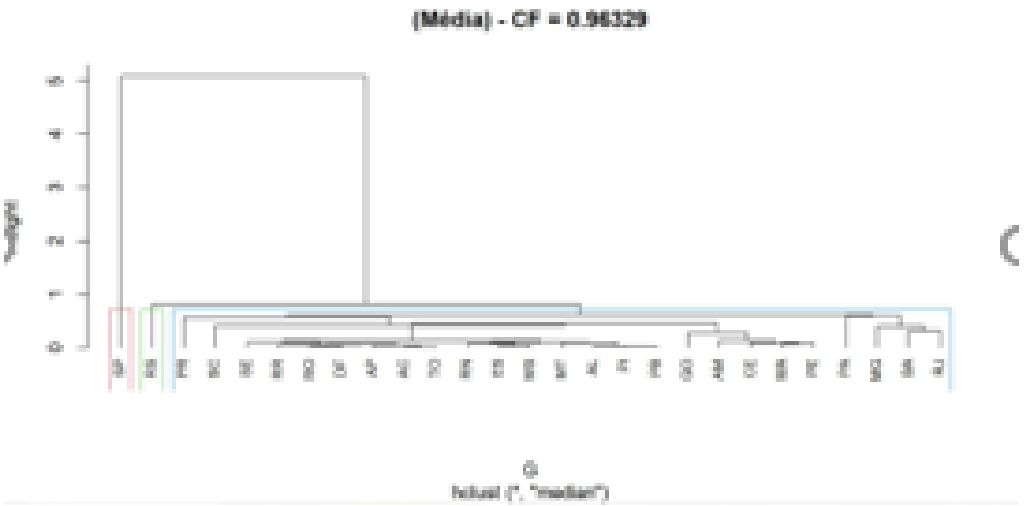


Figura 4: Dendrograma na distância euclidiana e o método de ligação Completa.



Com o método de ligação Completa que agrupar os elementos com menor distância entre os mais distantes pode ser observar no dendrograma (Figura 4), os 3 cluster formado pelo estado de São Paulo, no cluster 2 foi formado por seis estado Minas Gerais, Bahia, Rio de Janeiro, Rio Grande do Sul, Paraná e Santa Catarina. Já em relação ao cluster 3 formado por 19 estados, Rondônia, Tocantins, Mato Grosso, Espírito Santo, Pará, Maranhão, Ceará, Mato Grosso do Sul, Amazona, Distrito Federal, Rio Grande do Norte, Piauí, Alagoas, Sergipe, Amapá, Acre, Roraima, Goiás, Pernambuco, Para, Paraíba, Mato grosso e Distrito Federal.

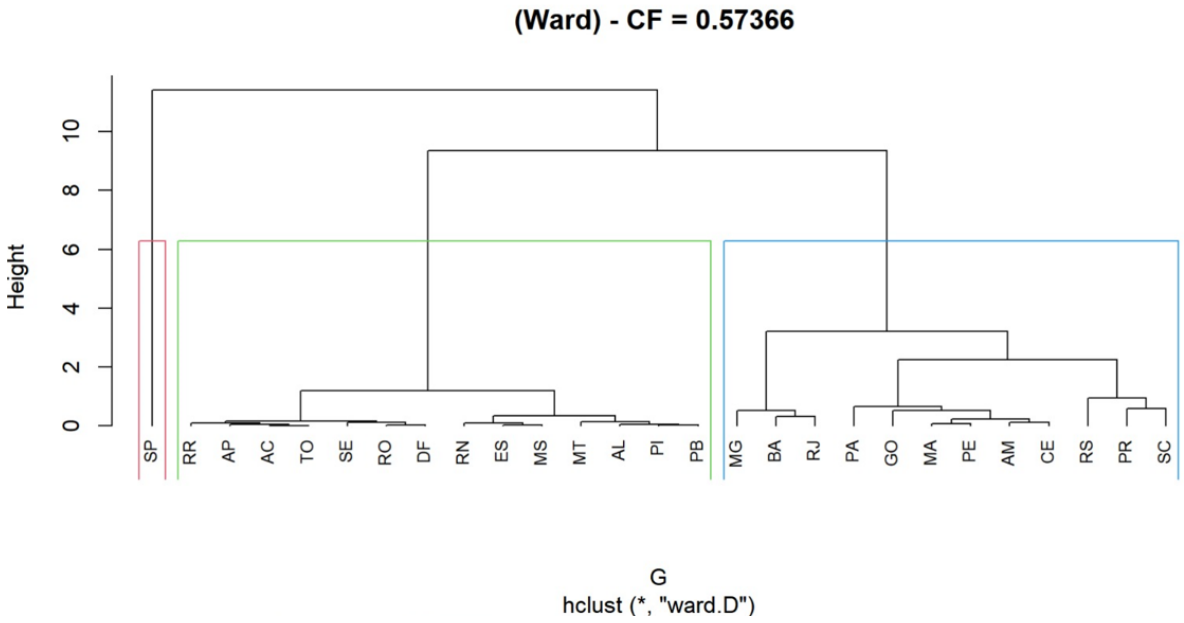
Figura 5: Dendrograma na distância euclidiana e o método de ligação média



No método de ligação média, podendo observar no dendograma (Figura 5) os 3 cluster formados por este método. No grupo 1 e 2 foram formados por apenas um estado São Paulo, e Rio Grande Sul, no grupo 3 o terceiro grupo formado pelos estados, Paraíba, Santa Catarina, Tocantins, Mato Grosso, Espírito Santo, Pará e Maranhão, Rondônia, Ceará, Mato Grosso do Sul, Amazona, Distrito Federal, Rio de Janeiro, Rio Grande do Norte, Piauí, Alagoas, Sergipe, Mina Gerais, Amapá, Acre, Roraima, Goiás, Minas Gerais, Bahia, Rio de Janeiro, Rio Grande do Sul, Paraná e Santa Catarina Paraná e Bahia, quanto maior for a população do estado maior era o número de vacinação.

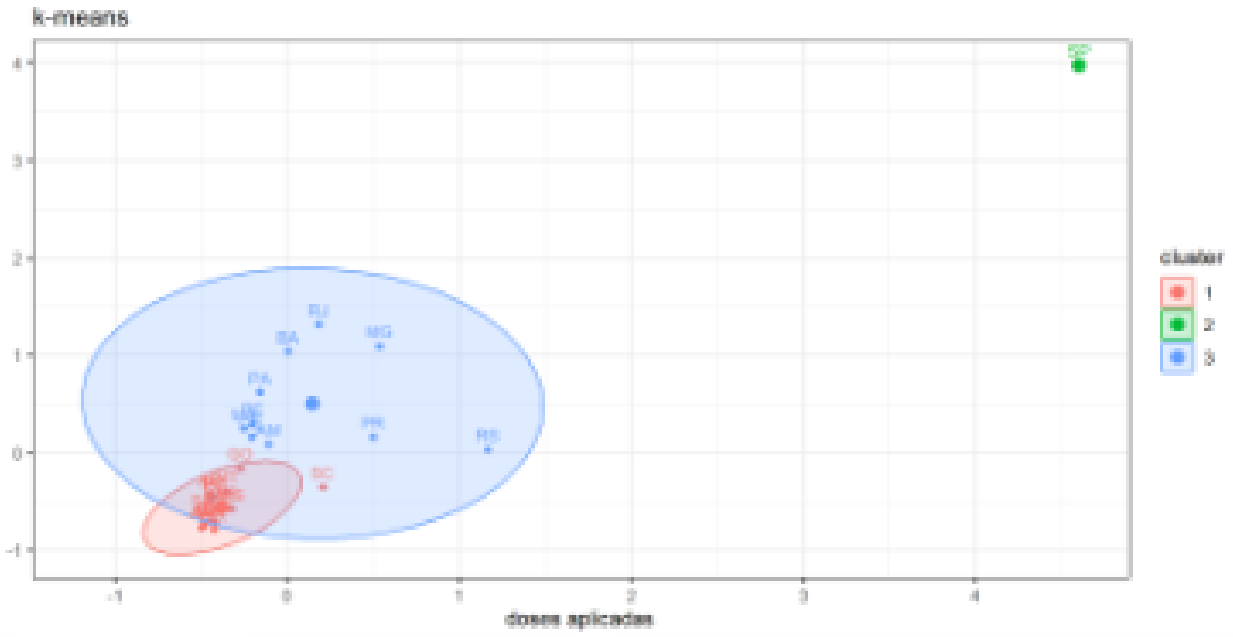
Já o método de Ward dos 3 cluster formados pode ser observar no dendrograma (Figura 6)

Figura 6: Dendrograma na distância euclidiana e o método de ligação Ward.



Ao analisar os 3 cluster da distância Euclidiana com o método K-means (Figura 6), pode observar no cluster 1 formando por o estados de São Paulo cluster 2 por dez estados (PA, AM, BA, PE, CE, MA, RS, PR, CE e MG), o cluster 3 com 15 estados e Distrito Federal (ES, RR, AP, TO, RO, MT, GO, MS, PB, AC, RG, AL, SE, PI e DF), no nesse método é possível verificar as características de cada aglomeração. Com base na média do K-means de cada cluster observou-se que o estado com mais caso de vacinação e maior média de números de óbitos.

Figura 6: Gráfico obtido por meio do método de K-maens para os dados do Covid-19.



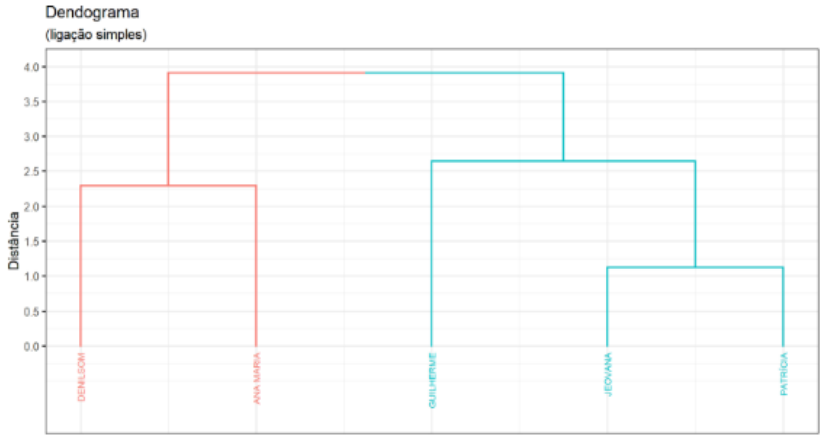
Nosso interesse é avaliar a estrutura do dendrograma e a qualidade desse possível agrupamento recém formado. A partir daqui podemos ter uma noção melhor sobre a qualidade dos grupos e indicar o corte ao longo da árvore, deixando explícito, para k=3.

Tabela 3: medidas avaliativas acerca da qualidade

	Divisão de k=3
Correlação Cofenética	0,9308
Coeficiente de Silhueta Média	0,4430977
Índice de Dunn	0,6666667

Tanto silhueta, correlação cofenética e Índice de Dunn observado na Tabela 3 designa a melhor divisão é k=3. Poderíamos escolher a métrica de euclidiana, Mahalanobis assim com as demais, porém por não possuir correlação significativa, valores atípicos seriam escolhas pouco convenientes a elaboração do agrupamento. As distâncias assim como os critérios de ligações devem ser escolhidos de modo a alcançar os critérios de heterogeneidade entre grupos e homogeneidade inter grupos.

Figura 7 : Dendrograma representações dos grupos



A correlação cofenética indica valor próximo a 1 significa um bom desempenho aceitável do agrupamento, na Figura 7 acima, expõe o dendrograma onde os elementos PJ possuem o menor salto, seguido de G que foi fundido a PJ, DA são os mais distantes . Ressaltamos que o agrupamento hierárquico pode preceder o não-hierárquico e que os demais métodos de ligação seguem os mesmos pressupostos onde consideramos os objetos mais similares.

CONCLUSÕES

A partir das principais técnicas de análise de cluster hierárquicas e não hierárquicas, o método de ligação Ward como representação final da clusterização gerada e representada por meio do dendrograma, observou-se que em relação a todos os outros agrupamentos soube lidar melhor em relação à especificidade das variáveis, análises complementares com silhueta e correlação cofenética nos deram uma noção do melhor rendimento se comparado às técnicas K-means, PAM e FCM, os critérios de homogeneidade também foram melhor satisfeitos, em comparação às não hierárquicas. Comparando os grupos formados pelos métodos Ward e K-means, conservou partes das unidades encontradas no grupo formado para os estados de Minas Gerais, Bahia, Rio de Janeiro, Rio Grande do Sul, Paraná e ademais os estados como Santa Catarina e São Paulo houve-se uma divergência entre os métodos por considerar grupo unitário ou realocação de Santa Catarina. Estados com baixos índices de mortalidade e vacinação também em ambas as vertentes foram similares, no entanto, com a diferença que Amazonas, Pará, Ceará, Maranhão e Pernambuco formam um único grupo pelo método Ward. PAM em termos de visualização gráfica teve um desempenho bem similar ao K-means com estados já ressaltados acima, com a exceção de Goiás, Santa Catarina, Pernambuco entre outros divergindo entre os grupos, no entanto, a prevalência de Minas Gerais, Bahia, Rio de Janeiro, Rio Grande do Sul e Paraná. O

algoritmo FCM não considera independentemente da espécie pertencente a um único grupo, divergiu amplamente se comparado a todas as metodologias.

Verificamos por meio método de ligação Ward, as classes formadas que os estados que forneceram mais vacinas foram de Santa Catarina, Paraná, Rio Grande do Sul, São Paulo, Bahia, Minas Gerais e Rio de Janeiro em decorrência também de maior volume de vidas perdidas sendo o maior de todas outras classes, em segundo lugar nas duas frentes apresentando maiores números vêm Amazonas, Pará, Ceará, Maranhão e Pernambuco sendo o menor em quantidade de unidades, porém reunindo taxas altas e por último os estados com os menores números de vacinação dado o menor volume de mortes que são os estados de Roraima, Rondônia, Sergipe, Amapá, Acre, Tocantins, Goiás, Rio grande do Norte, Espírito Santo, Mato Grosso, Mato Grosso do Sul, Alagoas, Piauí, Paraíba e o Distrito Federal como unidade federativa, muitas similaridades entre si, ou seja, sendo a classe menor variância.

AGRADECIMENTOS

"Agradecimentos ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC) UEPB que propiciou uma bolsa de iniciação para estre trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

Albuquerque, M.A.; Barros, K.N.O.; **Determinação do número de grupos em análise de agrupamento via de raio de influência**, Braz. J. of Develop., Curitiba, v. 6, n.6, p.38342-38355 jun. 2020.

Albuquerque, M. A., Barros, K. N. N. O., Gouveia, J. F., & Ferreira, R. L. C. (2016). **Determination and validation of group numbers in a cluster analysis: A case study applied to forestry science. Acta Scientiarum.** Technology, 38(3), 339-344.

Albuquerque, M. A. & Barros, K. N. N. O. (2020). **Introdução à Análise de Agrupamento: teoria e prática com aplicações em R. [e-book].** Campina Grande. Ed. EDUEPB. Disponível em: http://eduepb.uepb.edu.br/download/introducao-a-analise-de-agrupamento-teoriaepraticacomaplicacoesemr/?wpdmdl=997&masterkey=5e9790498ofc9.

Costa, G.D.; **Análise Multivariada de Países da América do Sul por Meio de Indicadores Socioeconômicos,** Uberlândia: UFU, 2019.

Duarte. S. R. N.; **Um guia para agrupamento com pacote cluster do Rutilizando dados do spotify, Universidade Federal do Rio Grande do Norte** – UFRN, 2021.

RAND, William M. Objective criteria for the evaluation of clustering methods. **Journal of the American Statistical association**, v. 66, n. 336, p. 846-850, 1971.

Souza, M. V. V.; **Analise Multivariada De Países Da América E Europa Utilizando Indicadores Sobre A Covid-19 E Dieta Da População**. Universidade Federal de Uberlândia Faculdade de Matemática, MG, 2022.

Tizotte, T. R. L.; **Análise bibliométrica dos artigos da base de dados da Scopus sobre a Produção Científica Brasileira da Covid-19.** Brazilian Journal of Development, Curitiba, v.7, n.7, p.73457-73474 jul. 2021.

UNDP.; **United Nations Development Program. Human Development Index** (H), acessad.