UNIVERSIDADE ESTADUAL DA PARAÍBA
CAMPUS I
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA
CURSO DE BACHARELADO EM ESTATÍSTICA

EMANUELA RODRIGUES DO NASCIMENTO

USO DA TÉCNICA DE AGRUPAMENTO APLICADO AO NÚMEROS DE INFECTADOS POR COVID-19 EM COMPARAÇÃO COM O IDH DE CADA ESTADOS DO BRASIL

CAMPINA GRANDE- PB
2022

EMANUELA RODRIGUES DO NASCIMENTO

USO DA TÉCNICA DE AGRUPAMENTO APLICADO AO NÚMEROS DE INFECTADOS POR COVID-19 EM COMPARAÇÃO COM O IDH DE CADA ESTADOS DO BRASIL

Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr° Mácio Augusto de Albuquerque

CAMPINA GRANDE- PB
2022

N244u Nascimento, Emanuela Rodrigues do.

Uso da técnica de agrupamento aplicado ao número de infectados por covid-19 em comparação com o IDH de cada estados do Brasil [manuscrito] / Emanuela Rodrigues do Nascimento. - 2022.

43 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2022.

"Orientação : Prof. Dr. Mácio Augusto de Albuquerque , Coordenação do Curso de Estatística - CCT."

1. Método de Agrupamento. 2. Covid-19. 3. Índice de Desenvolvimento Humano - IDH. 4. Estados brasileiros. I. Título

21. ed. CDD 616.2

Elaborada por Deivid A. M. Neves - CRB - 15/899 BC/UEPB

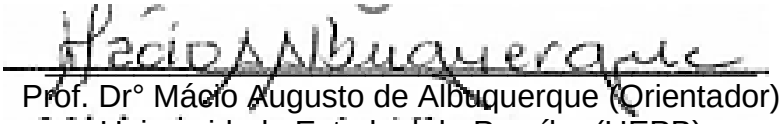
EMANUELA RODRIGUES DO NASCIMENTO

USO DA TÉCNICA DE AGRUPAMENTO APLICADO AO NÚMEROS DE INFECTADOS POR COVID-19 EM COMPARAÇÃO COM O IDH DE CADA ESTADOS DO BRASIL

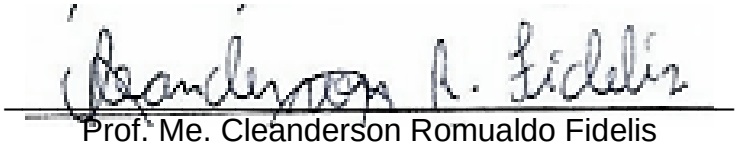
Trabalho de Conclusão de Curso (Artigo) apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba, como requisito parcial à obtenção do título de bacharel em Estatística.

Aprovada em 18 de novembro de 2022.

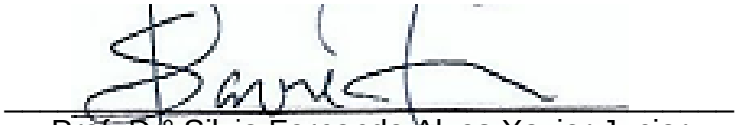
BANCA EXAMINADORA



Prof. Dr° Mácio Augusto de Albuquerque (Orientador)
Universidade Estadual da Paraíba (UEPB)



Prof. Me. Cleanderson Romualdo Fidelis
Universidade Estadual da Paraíba (UEPB)



Prof. Dr° Silvio Fernando Alves Xavier Junior
Universidade Estadual da Paraíba (UEPB)

Dedico este trabalho ao meu pai, mãe, irmãos e irmãs, que me deram apoio e me incentivaram a chegar até aqui. Muito obrigada!

“Para ir da oportunidade ao êxito é preciso enfrentar os medos de mudança, romper com o memos e ter a capacidade de se antecipar.”

Mario Sergio Cortella

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

USO DA TÉCNICA DE AGRUPAMENTO APLICADO AO NÚMEROS DE INFECTADOS POR COVID-19 EM COMPARAÇÃO COM O IDH DE CADA ESTADOS DO BRASIL

USE OF THE GROUPING TECHNIQUE APPLIED TO THE NUMBERS OF INFECTED BY COVID-19 COMPARED TO THE HDI OF EACH STATE OF BRAZIL.

Emanuela Rodrigues do Nascimento*

Resumo

O presente trabalho tem por objetivo mostrar como pode ser feita a análise de cluster, usando a técnica hierárquica e não hierarquia na taxa de infectados por Covid-19 dos estados brasileiros com base no números de infectados de cada estado para assim identificar a similaridades entre os estados e os números de infectados, oferecendo um contraponto ao critério utilizado de análise do número de infectados dos estados, baseando-se no tamanho da população e comparando com seu Índice de Desenvolvimento Humano (IDH). Utilizou-se dados da Covid-19 retirado de uma plataforma pública e gratuita chamado Coronavírus//Brasil e o Atlas Brasil 2013 com relação ao IDH de 2010. Para a análise de agrupamento foi utilizado a matriz de Mahalanobins 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 o coeficiente de correlação conféneticа para medir o grau de ajuste entre as matrizes similares originais e a matriz resultante da simplificação proporcionada pelo método de agrupamento. No entanto, foi verificado o método que melhor representa os dados é o de ligação completa. 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: distância; método de agrupamento; dendrograma; grupos.

Abstract

The present work aims to show how the cluster analysis can be carried out, using the hierarchical and non-hierarchical technique in the rate of infected by Covid-19 in the Brazilian states based on the number of infected in each state, in order to identify the similarities between the states and the number of infected people, offering a counterpoint to the criterion used to analyze the number of infected people in the states, based on the size of the population and comparing it with its Human Development Index (HDI). Covid-19 data taken from a public and free platform called Coronavirus//Brasil and Atlas Brasil 2013 were used in relation to the 2010 HDI. If the simple, complete, average, ward link and a non-hierarchical method through the K-means method were used, the confenetic correlation coefficient was also applied to measure the degree of fit between the original similar matrices and the resulting matrix the simplification provided by the grouping method. However, the method that best represents the data was found to be complete linkage. When grouping the states of both data, it took into account the similarity between the data variables and the correlation where it can be seen that the data are correlated.

Keywords: distance; clustering method; dendrogram; groups.

* Aluna do curso Estatística, Depto Estatística, UEPB, Campina Grande, PB, emanuela.nascimento@aluno.uepb.edu.br

1 INTRODUÇÃO

As técnicas de análise multivariada possibilitam avaliar um conjunto de características, levando em consideração as correlações existentes, que permitem que inferências sobre o conjunto de variáveis sejam feitas em um nível de significância conhecido.

Nas diversas áreas do conhecimento uma das técnicas multivariadas mais utilizadas é a análise de agrupamento. O seu emprego em áreas tais como engenharia, educação, saúde, sociologia, economia, administração, entre outras, vem aumentando muito nos últimos anos.

A análise de agrupamento tem por finalidade reunir, por algum critério de classificação, as unidades amostrais em grupos, de tal forma que exista homogeneidade dentro do grupo e heterogeneidade entre grupos (JOHNSON & WICHERN, 1992; CRUZ & REGAZZI, 1994).

O processo de agrupamento envolve basicamente duas etapas. A primeira se refere à estimação de uma medida de dissimilaridade entre os indivíduos e a segunda, refere-se à adoção de uma técnica de formação de grupos.

Um grande número de medidas de similaridade ou de dissimilaridade tem sido proposto e utilizado em análise de agrupamento, sendo a escolha entre elas baseada na preferência e/ou na conveniência do pesquisador (ALBUQUERQUE, 2005).

Com a definição da medida de dissimilaridade a ser utilizada, a etapa seguinte é a adoção de uma técnica de agrupamento para formação dos grupos. Para realização desta tarefa, existe um grande número de métodos disponíveis, dos quais o pesquisador tem de decidir qual o mais adequado ao seu propósito, uma vez que as diferentes técnicas podem levar a diferentes soluções (ALBUQUERQUE ET AL., 2006).

As técnicas de análise de agrupamento exigem de seus usuários a tomada de uma série de decisões independentes, que requerem o conhecimento das propriedades dos diversos algoritmos à disposição e que podem representar diferentes agrupamentos. Além disso, o resultado dos agrupamentos pode ser influenciado pela escolha da medida de dissimilaridade, bem como pela definição do número de grupos (GOWER & LEGENDRE, 1986; JACKSON ET AL., 1989; DUARTE ET AL., 1999).

A análise de agrupamento é uma técnica de análise multivariada (FÁVERO E BELFIORE, 2019) com o objetivo de promover a segmentação dos dados em categorias ou grupos com base nas suas características homogêneas ou heterogêneas classificando em mesmos grupos ou distintos. Essa técnica agrupa dados para interpretação utilizando alguns métodos que procuram por grupos excludentes, ascendentes para assim reprimir as informações de um conjunto (CAMPOS, 2019), vários são os tipos de técnicas de agrupamento encontradas na literatura conforme (ALBUQUERQUE & BARROS, 2020), dos quais o pesquisador tem de decidir qual o mais adequado ao seu propósito, uma vez que as diferentes técnicas podem levar a diferentes soluções.

As técnicas de agrupamento podem ser classificadas em hierárquicas e não-hierárquicas (ALBUQUERQUE ET AL., 2006). A técnica hierárquica consiste em uma série de sucessivos agrupamentos ou sucessivas divisões de elementos, em que os elementos são agregados ou desagregados. A técnica não-hierárquica foi desenvolvida para agrupar elementos em K grupos, em que K é a quantidade de grupos definida previamente.

Um grande número de medida de similaridade ou de dissimilaridade tem sido proposto e utilizado em análise de agrupamento, sendo a escolha entre elas baseada na preferência e, ou, na conveniência do pesquisador (ALBUQUERQUE & BARROS, 2020).

Com a definição da medida de dissimilaridade a ser utilizada, a etapa seguinte é a adoção de uma técnica de agrupamento para formação dos grupos. Para realização dessa tarefa, existe grande números de métodos disponíveis, dos quais o pesquisador tem de decidir qual o mais adequado ao seu propósito, uma vez que as diferentes técnicas podem levar a diferentes soluções, no geral, o método Não-hierárquicos é encontrar o número “k” de clusters que consiga realizar a divisão das observações de maneira satisfatória, ou seja, que consiga identificar semelhanças e diferenças entre as observações.

As técnicas de análise de agrupamento exigem de seus usuários a tomada de uma série de decisões independentes, que requerem o conhecimento das propriedades dos diversos algoritmos à disposição e que podem representar diferentes agrupamentos. Além disso, o resultado dos agrupamentos pode ser influenciado pela escolha da técnica utilizada como também pela medida de dissimilaridade, bem como pela definição do número de grupos.

Em dezembro de 2019 foi anunciado pelo governo chinês em Wuhan (Hubei, China) uma descoberta de um novo coronavírus, denominada SARS-CoV-2 (COVID-19), com esse anúncio deixou em alerta a Organização Mundial da Saúde (OMS), declarando que a infecção causada por esse vírus contaminava humanos com potencial de transmissão alto (ALVES, 2020).

As recomendações da OMS para diminuir a velocidade de transmissão, sua principal decisão foi o isolamento social, a COVID-19 se espalhou por todo os países, no Brasil atingiu as 27 unidades federativas, esse fato ocorreu devido aos desafios quanto as condições de vulnerabilidade social, de moradia e saneamentos precários, além de superpopulação domiciliar, no entanto devido a heterogeneidade da população em cada um dos estados, a pandemia se difundiu de forma distinta (ALVES,2020).

Como a pandemia não afetou todas as pessoas de forma uniforme, afetando as parcelas da população mais vulneráveis, ou seja, considerando as populações mais pobres são mais propensas a ter condições crônicas, isso as coloca em maior risco de mortalidade associada ao vírus. Já que a pandemia pode gerar uma crise econômica, se levamos em conta a taxa de desemprego que cresceu nos estados (DOURADO,2021).

Desta forma, o IDH, índice que orienta sobre fatores que influenciam no desenvolvimento humano, pode ser uma ferramenta para avaliar esta vulnerabilidade, uma vez que fatores como falta de infraestrutura sanitária, prejudica cuidados preventivos para a infecção pelo vírus. Por outro lado, os estados com o IDH mais elevado possuem maior condições indispensáveis de uma economia avançada (DOMINGO, 2021).

O Índice de Desenvolvimento Humano (IDH) foi criado para avaliar o desenvolvimento de um país, estado ou município, não apenas crescimento econômico. O IDH é aferido a partir da média geométrica entre índices que medem cada um dos seguintes fatores, considerados alguns pontos no desenvolvimento humano, como ter uma vida longa e saudável, adquirir conhecimentos e ter um padrão de vida decente (UNDP, 2020).

Diante do exposto, o objetivo deste é utiliza a análise de cluster, através da técnica hierárquica e não hierarquia na taxa de infectados por Covid-19 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 seu Índice de Desenvolvimento Humano (IDH).

2 MÉTODOLOGIA

2.1 Coronavírus (COVID-19) metodologia

SARS-COVID19 é uma doença contagiosa causada pela síndrome respiratória aguda. O primeiro caso foi identificado em Wuhan, China, em dezembro de 2019.

A doença se espalhou pelo mundo, levando a uma pandemia. A transmissão de COVID-19 ocorre quando as pessoas são expostas a gotículas respiratórias contendo vírus, ou seja, pessoas infectadas podem transmitir o vírus a outra pessoa até dois dias antes de apresentarem os sintomas, assim como as pessoas que não apresentam sintomas com isso o número de infectados por COVID-19 aumentou rapidamente em vários lugares do mundo inclusive no Brasil (TIZOTTE,2021).

As medidas preventivas eram distanciamento social, quarentena, ventilação de espaços internos, cobertura de tosses e espirros, a utilização de máscaras faciais em ambientes públicos para minimizar o risco de transmissões (OMS, 2020). Os sintomas eram diversos, variando de sintomas leves a graves, várias medidas foram utilizadas para quantificar a mortalidade, esses números variam de acordo com a região onde foram calculadas.

Com a chegada do COVID-19 no Brasil as autoridades sanitárias juntamente com os órgãos Federais, Estaduais e Municipais adotaram diversas medidas de controle e prevenção da doença para os estados brasileiros, essas medidas se diferenciaram de uma região entre região, entretanto a medida mais anunciada pelas autoridades foi a prática do distanciamento social (BEZERRA, 2020). Em alguns estados, as medidas de isolamento adotadas pela população possuem variações em função da renda, sexo e escolaridade da população, ou seja, a percepção e o comportamento dos brasileiros em relação à adoção de autoisolamento e respeito aos decretos de quarentena variaram de estados para estado, pois mesmo com o avanço da pandemia, parte da população começou a ter dificuldades de se manter isolada, mesmo com um número crescente de casos. Os dados do COVID-19 retirado em 09 dezembro de 2021, provêm de plataformas públicas e gratuitas chamado Coronavírus//Brasil (https://covid.saude.gov.br).

2.2 Índice de Desenvolvimento Humano (IDH)

O Índice de Desenvolvimento Humano (IDH) foi criado em 1998 por dois economistas, um paquistanês Mahbub Ul Haq e um indiano Amartya. O IDH é considerado uma média para resumir as condições básica de uma população, centrada na educação, renda, e qualidade de vida. Publicado no Brasil pela primeira vez no ano de 1990, o IDH aos poucos foi se tornando referência em vários lugares

do mundo, podendo ser calculado através de três aspectos principais: Renda, Longevidade e Educação, com variação entre 0 e 1, quanto mais próxima de 1 é o estado mais desenvolvido e quanto mais próxima do 0 é o estado menos desenvolvido, ou seja, a partir desses aspectos é possível observar as melhorias fornecida ao estado por meio do IDH (COSTA, 2019). Esses dados foram retirados do Atlas Brasil 2013.

2.3 Análise de Agrupamento

A análise de agrupamento é uma técnica multivariada que tem por objetivo proporcionar uma ou várias partições na massa de dados, em grupos, por algum critério de classificação, de tal forma que exista homogeneidade dentro e heterogeneidade entre grupos (SNEATH & SOKAL, 1973; MARDIA ET AL., 1997).

Essa técnica sumariza dados para interpretação e utiliza métodos que procuram grupos excludentes, ascendentes, reduzindo as informações de um conjunto de n indivíduos para informações de um novo conjunto de grupos, onde g é significativamente menor que n, resultando um dendrograma de exclusão (MARDIA ET AL., 1997).

Conforme Albuquerque (2005), de modo sintético, a técnica pode ser descrita como se segue: dado um conjunto de n indivíduos para os quais existe informação sobre a forma de p variáveis, o método de análise de agrupamento procede ao agrupamento dos indivíduos em função da informação existente, de tal modo que os indivíduos pertencentes a um mesmo grupo sejam tão semelhantes quanto possível e sempre mais semelhante aos elementos do mesmo grupo do que aos elementos dos grupos restantes. Essa técnica é também chamada de técnica de partição, classificação ou taxonomia, embora o termo partição seja mais utilizado para uma das técnicas especificas da análise: aquela em que os indivíduos são divididos por um número preestabelecido de grupos.

Segundo Aaker et al. (2001), a premissa mais importante da análise de agrupamento é a de que a medida de similaridade ou dissimilaridade na qual o processo de agrupamento se baseia é uma medida válida de similaridade ou dissimilaridade entre os indivíduos. A segunda premissa mais importante é a de que existe uma justificativa teórica para estruturar os indivíduos em grupos. Como em outras técnicas multivariadas, também há teoria e lógica guiando e dando base à análise de agrupamento.

Geralmente, é difícil avaliar a qualidade do processo de agrupamento. Não existem testes estatísticos padrões para garantir que o resultado seja aleatório. O valor do critério medida, legitimidade do resultado, aparência de uma hierarquia natural (quando for empregado um método não hierárquico) e confiabilidade de testes de divisão de amostra, oferecem informações úteis (ALBUQUERQUE & BARROS, 2020). Entretanto, é difícil saber, exatamente, quais os grupos são muito parecidos e quais objetos são difíceis de serem inseridos. Geralmente, não é fácil selecionar um critério e programa de agrupamento por meio de outra referência que não a disponibilidade.

Na análise de agrupamento, é fundamental ter particular cuidado na seleção das variáveis de partida que vão caracterizar cada indivíduo, e determinar, em última instância, qual o grupo em que deve ser inscrito. Nesta análise não existe qualquer tipo de dependência entre as variáveis, isto é, os grupos se configuram por si mesmo sem necessidade de ser definida uma relação causal entre as variáveis utilizadas. Essa análise não faz uso de modelos aleatórios, mas é útil por fornecer um sumário

bem justificado de um conjunto de dados. As técnicas são exploratórias e a ideia é, sobretudo gerar hipóteses, mais do que testá-las, sendo necessária a validação posterior dos resultados encontrados através da aplicação de outros métodos estatísticos (REIS, 1997).

Genericamente, a análise de agrupamento compreende cinco etapas (AAKER ET AL., 2001):

1. A seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
2. A definição de um conjunto de variáveis a partir das quais serão obtidas informações necessárias ao agrupamento dos indivíduos;
3. A definição de uma medida de semelhança ou distância entre os indivíduos;
4. A escolha de um algoritmo de partição/classificação;
5. Por último, a validação dos resultados encontrados.

2.4 Distâncias

As medidas distâncias são usadas para representar os pontos na estrutura de similaridade. Essa medida representa a menor distância entre dois pontos e é uma extensão do teorema de Pitágoras para o caso multidimensional. O termo dissimilaridade surgiu em função da distância entre dois pontos P e Q, definida como d(P,Q), pois à medida que a divergência entre os pontos P e Q aumenta, eles se tornam cada vez mais diferentes. Os valores das distâncias são geralmente obtidos a partir de informações de n observações, medidas em relação a p variáveis. Várias medidas de dissimilaridade são atualmente propostas, principalmente devido ao grande desenvolvimento e uso de técnicas multivariadas (SOUZA, 2022).

Segundo Albuquerque (2005), há dois tipos de medidas de parecença: medidas de similaridade (quanto maior o valor, maior a semelhança entre os objetos) e medidas de dissimilaridade (quanto maior o valor, menor a semelhança entre os objetos).

De um modo geral, as medidas de similaridade e de dissimilaridade são interrelacionadas e, facilmente, transformáveis entre si (ALBUQUERQUE & BARROS, 2020). Há um grande número de coeficientes de similaridade e/ou de dissimilaridade para caracteres binários disponíveis na literatura. Segundo Clifford & Stephenson (1975), tais coeficientes podem ser, facilmente, convertidos para coeficientes de dissimilaridade: se a similaridade for denominada, a medida de dissimilaridade será o seu complementar (1 – s).

A maioria dos métodos de análise de agrupamento requer uma medida de similaridade ou dissimilaridade entre os elementos a serem agrupados, normalmente expressa como uma função distância ou métrica (ALBUQUERUQE, 2005).

Seja M um conjunto, uma métrica em M é uma função d: MxM $\rightarrow \Re$, tal que para quaisquer i, j, z $\in$M, tenhamos: $\Re$

1. d (i, j) = d (j, i) (simétrica);
2. d (i, j) > 0, se i$\neq$ j;
3. d (i, j) = 0, se e somente se, i = j; e
4. d (i, j) $\leq$ d (i, z) + d (z, j) (desigualdade triangular).

Além disso, é esperado que d (i, j) aumente quando a dissimilaridade entre i e j aumentar.

Existem várias medidas que podem ser utilizadas como medidas de distâncias ou dissimilaridade entre elementos de uma matriz de dados.

2.4.1 Distâncias Euclidiana

A distância euclidiana é a medida mais conhecida e utilizada entre as métricas. Isso representa a menor distância entre os dois objetos comparados, pois é dada pela distância em linha reta entre dois pontos de dados no espaço euclidiano e é deduzida do teorema de Pitágoras, segundo Marcondes (2020), se existe pontos $p_1, p_2, \ldots, p_n, \in R^r$ calcula-se a distância em linha reta entre os pontos. É importante perceber que cada medição faz uma contribuição igual para a fórmula e dependendo da quantidade, isso pode levar a problemas, ou seja, com isso possibilita o uso de outras métricas de distância que aplicam pesos e ajustes dependendo da situação da variável (AGUDELO 2021), se as observações de n indivíduos para p variáveis, e o indivíduo i é representado como um ponto observações de n indivíduos para p variáveis, e o indivíduo i é representado como um ponto $x_i \in R^p$. A distância Euclidiana entre *i* e *j* é dada por:

$$d(i,j) = \sqrt{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^2} \quad (1)$$

2.4.2 Distâncias Manhattan

Outra distância utilizada é a de Manhattan, também é conhecida como a distância máxima ou city block que é a soma das distâncias de todos os atributos, para os dois pontos de dados x, y nas dimensões do espaço d. Segundo Agudelo (2021), se utilizamos dois pontos A = (x_i; y_i) e B = (x_j; y_j), então a distância de Manhattan entre A e B é calculada da seguinte forma:

$$d_{ij}(A,B) = |x_i - y_i| + \left| x_j - y_j \right| \quad (2)$$

Se utilizamos n-pontos, por exemplo A = (x_1, x_2, x_3, ..., x_n) e B = (y_1, y_2, y_3, ..., y_n), então calcular a distância de Manhattan entre A e B da seguinte forma:

$$d_{ij}(A,B) = \sum_{i=1}^{n} |x_i - y_i| \quad (3)$$

Por ser mais fácil de calcular do que a distância Euclidiana a distância de Manhattan permite caminhar em quatro direções para ir de um ponto a outro, ou seja, nem sempre fornece a mesma distância em linha reta entre dois pontos.

2.4.3 Distâncias Mahalanobis (D^2)

A distância de Mahalanobis foi introduzida por Mahalanobis em 1936, sendo que sua utilidade decorre do fato de fornecer uma maneira de determinar a semelhança entre duas variáveis aleatórias multidimensionais. Segundo Albuquerque & Barros (2020), para análise de agrupamento utilizar-se o método de dissimilaridade baseado na distância de Mahalanobis (D^2), considerada uma das distâncias mais utilizadas, podendo ser calculada de acordo com a expressão a seguir:

$$D^2 = (X_i - X_j)'. \sum^{-1} (X_i - X_j) \quad (4)$$

Em que: D^2 apresenta característica de ser a-1 para qualquer transformação linear não-singular, X_i é o vetor que pertence à parcela i, X_j será um vetor que pertence a parcela j, $\sum^{-1}$ é a inversa da matriz de covariância residual de X e $(X_i - X_j)'$ é o vetor transposto da diferença entre X_i e X_j.

2.5 Método Hierárquico

Nestes métodos os indivíduos são classificados em grupos em diferentes etapas de modo ordenado (hierárquico), produzindo ao final da análise um gráfico em forma de árvore (dendrograma). Estes agrupamentos podem ser utilizados tanto para agrupar indivíduos quanto para agrupar variáveis. Quando o dendrograma construído é das variáveis, a similaridade entre duas variáveis aponta forte correlação entre elas. Os métodos hierárquicos podem ser subdivididos em métodos aglomerativos e métodos divisivos (SARTORIO, 2008).

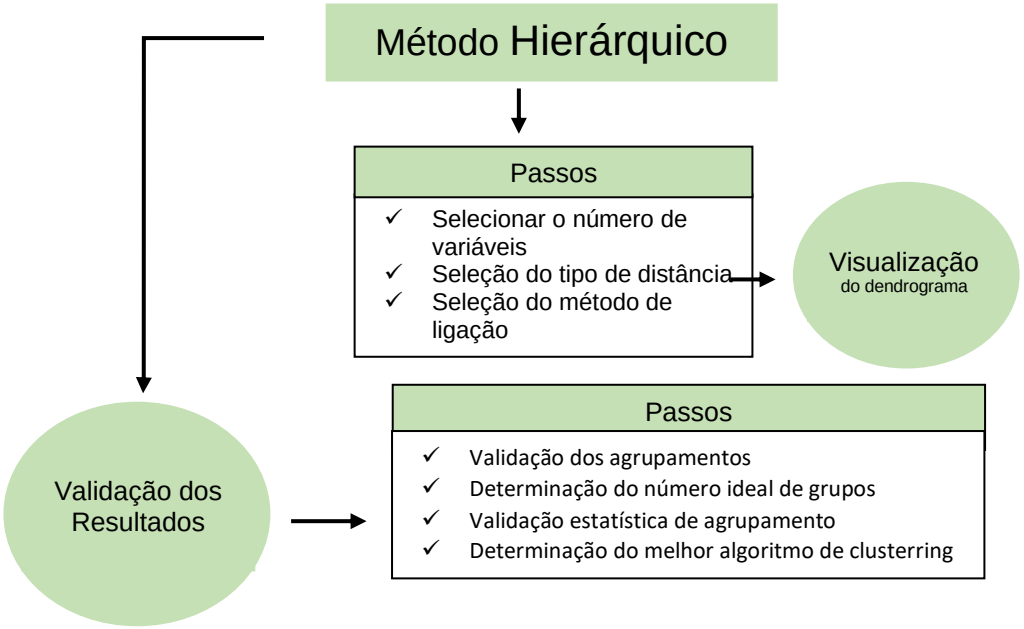


Figura 1: Algoritmo aplicado no Método Hierárquico.
Fonte: Autoria própria

2.5.1 Métodos aglomerativos

São os mais comuns de se encontrar em trabalhos científicos. Estes métodos são baseados em uma medida de dissimilaridade escolhida a priori, reduzindo a um único grupo ao final. Em geral, o algoritmo para n objetos é:

1. Inicia-se com n grupos, cada um com um único elemento e com uma matriz de distâncias simétrica de ordem n;
2. Busca-se na matriz de distâncias o par de grupos com a menor distância (mais similar);
3. Fundir dois grupos x e y e nomeá-lo, por exemplo, como xy. Calcular novamente e rearranjar as distâncias na matriz de distâncias utilizando a medida de parecença e o método aglomerativo escolhido;
4. Eliminar as linhas e a as colunas correspondentes aos grupos x e y;
5. Acrescentar uma nova linha e uma nova coluna com as distâncias entre o grupo xy e os demais grupos;
6. Repetir os passos 2 e 3 n − 1 vezes até que todos os objetos estejam agrupados em um único grupo.

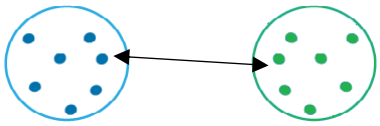
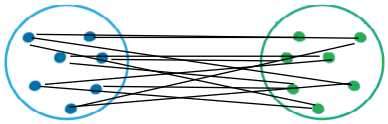
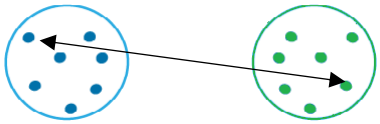
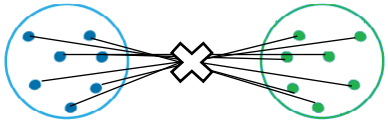
2.5.2 Métodos divisivos

Estes métodos partem de um único grupo com os n elementos e por divisões sucessivas são divididos em 2, 3, ou mais grupos, de tal modo que os indivíduos em um subgrupo estarão distantes dos indivíduos de outro subgrupo. Esses indivíduos são divididos novamente em subgrupos e o processo se repete até que cada indivíduo forme um grupo (SARTORIO, 2008). Por este método não ser tão abrangente em pesquisas científicas, não abordaremos o mesmo neste trabalho.

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 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 utilizada e fácil de se encontra em alguns softwares, 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. Na Tabela 1 observa-se a representação esquemática de cada um dos tipos de ligação, facilitando assim a sua interpretação

Tabela 1:Representação dos tipos de ligação entre agrupamentos.

Tipos de ligações	Representação
Simples	
Média	
Completa	
Ward	

Fonte: Autoria própria

• 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:

$$d(C_1, C_2) = \min\{d(X_l, X_k, l \neq k, l = 1{,}3{,}7 \; e \; k = 2{,}6)\} \quad (5)$$

• Método de ligação completa: Este 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:

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

• 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 C_1 possuem n_1 elementos e C_2 com n_2 elementos, a distância entre os grupos é dada por:

$$d(C_1, C_2) = \sum_{l \in C_1} \sum_{k \in C_2} \left(\frac{1}{n_1 n_2}\right) d(X_l, X_k) \quad (7)$$

• 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 C_l e C_i representa a soma quadrada entre os cluster que pode ser definida por:

$$d(C_l, C_i) = \left[\frac{n_l n_i}{n_l + n_i}\right] (\overline{X_l} - \overline{X_l})'(\overline{X_l} - \overline{X_l}) \quad (8)$$

2.6 Método Não-Hierárquico

Para os métodos não hierárquicos de Análise de Cluster, definir uma partição inicial o número “k” de clusters, quando todos os elementos podem ser trocados de grupo durante a execução do algoritmo, o método consegue identificar semelhanças e as 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 ideal 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. Recalcula os centróides 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.

2.7 Dendrograma

Dendrograma é uma representação gráfica bidimensional em forma de árvore utilizada para ilustrar a análise de agrupamentos feita sobre um conjunto de dados. Essa representação ilustra todo o procedimento de agrupamentos por meio de uma estrutura de árvore. Se for feito um corte em um determinado nível do gráfico, este corte representará o número de grupos existentes nesse nível e dos indivíduos que os formam (ALBUQUERQUE, 2005). A interpretação de um dendrograma de similaridade entre amostras fundamenta-se na intuição: duas amostras próximas devem ter também valores semelhantes para as variáveis medidas, portanto, quanto maior a proximidade entre as medidas relativas às amostras, maior a similaridade entre elas. O gráfico denominado por dendrograma hierarquiza esta similaridade de modo que podemos ter uma visão bidimensional da similaridade de todo o conjunto de amostras utilizado no estudo. Um caso particular é verificado quando o dendrograma construído é das variáveis, a similaridade entre duas variáveis aponta forte correlação entre estas variáveis do conjunto de dados estudado. Vale salientar que o dendrograma ilustra as partições feitas em cada nível sucessivo do processo de agrupamento. Neste, um eixo representa os indivíduos e o outro eixo representa a variabilidade ou 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 apresenta a primeira ligação, o segundo nível apresenta a segunda ligação, e assim sucessivamente. Por outro lado, a falta de critérios objetivos para determinar o número ótimo de grupos é uma dificuldade encontrada em estudos que utilizam análise de agrupamentos

2.8 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):

$$r_{cof} = \frac{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})(s_{ij}-\bar{s})}{\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(c_{ij}-\bar{c})^2}\sqrt{\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(s_{ij}-\bar{s})^2}} \quad (9)$$

onde: C_{ij} é o valor de similaridade entre os indivíduos i e j, onde será obtido a partir da matriz cofenética; S_{ij} é o valor de similaridade entre os indivíduos i e j, onde serão obtidos a partir da matriz de similaridade.

$$\bar{c} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} c_{ij} \quad (10)$$

$$\bar{s} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} s_{ij} \quad (11)$$

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 0 a 1, se $c \leq 0,39$ é considerado fraco, se $0,40 \leq c \leq 0,69$ é considerado moderado e se $c \geq 0,70$ 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.

2.9 Software R

O software R é uma linguagem utilizada por pesquisadores para apresentar modelos simples e sofisticados por meio de suas bibliotecas. Os métodos de agrupamento particionamento, agrupamento hierárquico e não hierárquico, (Anexo I e II) podem ser calculados usando os pacotes R *"cluster"* (LOPES,2021), tal como representa abaixo.

```
> install.packages(c("ecodist", "factoextra", "cluster"))
```

cluster: Para computação de algoritmos de cluster; *factoextra:* Para análise e visualização de clusters, incluindo os resultados de agrupamento; *ecodist:* Esta função calcula uma variedade de métricas de dissimilaridade ou distância. Embora duplique a funcionalidade de *dist()* e *bcdist()*, ele é escrito de tal forma que novas métricas podem ser facilmente adicionadas *distance()*.

Para calcular automaticamente o número de cluster da classificação utiliza-se função *fviz_nbclust()* do pacote *factoextra* (LOPES,2021). Em seguida utiliza-se o comando *K-means* classifica o objeto dentro de múltiplos grupos, com o resultado verifica a similaridade e usa o comando *ggplot2* com a função *fviz_nbclust()* para visualizar graficamente os resultados utilizando a técnica de componentes principais.

Utilizando a função *distance()* para calcula a distância entre os pares de dados que será o objeto de análise para a formação dos primeiros pares de dados similares com a distância de mahalanobis . A função *hclust()* utiliza os dados das distancias encontrada anteriormente aplicando os métodos a ser utilizado: *ward.D, single*, *complete*, *averge*, *median* ou *centroid* (KASSAMBARA, 2017) .

```
dist1<-distance(x,"mahalanobis")
clust1<-hclust(dist1,"ward.D")
```

A representação gráfica do método hierárquico é feita através de um dendrograma criado com o comando *plot(clust1).*

```
plot(clust1, main = "Dendograma Ward")
```

Algumas funções tem seus parâmetros e características mais específicas, podendo ser consultada a documentação utilizando a função *help('nome da função')* no próprio R.

3 RESULTADO E DISCUSSÕES

Foram utilizados os dados do COVID-19 coletado em 09 dezembro de 2021 de todos os estados brasileiros. Os dados utilizados provêm de plataformas públicas e gratuitas, chamada Coronavírus//Brasil, outro dado utilizado foi do IDH (2010) dos estados brasileiros retirado do Atlas Brasil 2013. Para realizar a análise de agrupamento foi utilizada a técnica hierárquico e não hierárquico como medida de dissimilaridade a distância de Mahalanobis (D^2), com os métodos mais comuns de agrupamento para determinar a distância entre agrupamentos são: ligação simples, ligação completa, médias das distâncias e método Ward, e a correlação cofenética.

Com o uso de diferentes métodos de agrupamento e análise de suas Figuras foi verificado que os grupos construídos diferem entre si. A partir do método hierárquico e aplicação da matriz de distância Mahalanobis 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 Mahalanosbis pelos métodos de ligação, observou-se uma pequena 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.

3.1 Análise dos dados de COVID-19

Utilizando os dados dos 27 estados do Brasil referente os dados da COVID-19 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 2 podendo observar que a variável dos Casos confirmados apresenta um desvio padrão alto em relação à média. A variável mortalidade está entre 146,7 e 401,1, possui um desvio padrão menor que a média, ou seja, quanto menor o desvio padrão, mais homogêneo são os dados.

Tabela 2: Análise descritiva das variáveis dos dados COVID-19.

Variáveis	Média	Desvio Padrão	Min	Max
Casos	821660	898479,3	88264	4449552
Óbitos	22842	310776,68	1849	154691
Incidência	11465	4155,891	1727	21254
Mortalidade	282,8	72,859	146,7	401,1

Fonte: Autoria própria

A determinação do número ideal de clusters pode ser feita usando dois tipos de métodos: métodos diretos e métodos indiretos. Nesses dados foram usados o

Métodos diretos para otimizar dois critérios: a soma dos quadrados do cluster usando o método do cotovelo. Segundo Matos (2021), esse método é frequentemente representado através de um gráfico (Figura 2) da soma dos quadrados do clusters em função do número de agrupamentos, sendo que para esses dados o número ideal dado foram de 5 clusters.

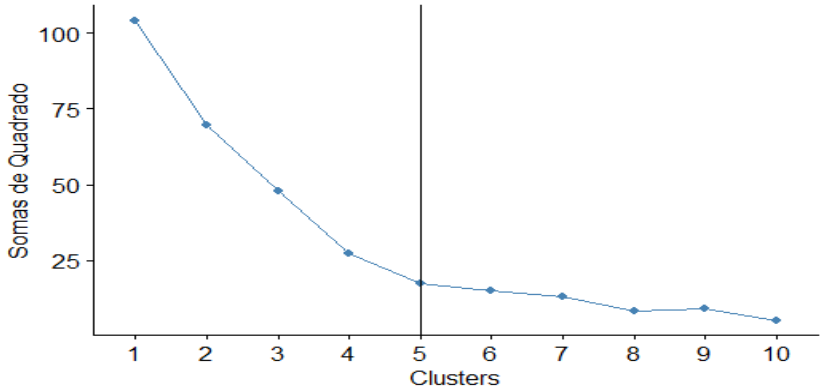


Figura 2: Referente ao número de clusters obtido para os dados do Covid-19.

Utilizando os 5 clusters obtidos anteriormente e aplicando a distância Euclidiana com o método K-means (Figura 3), observa-se no clusters 1 formando por nove estados (RO,RR, AP, TO, ES, MS, MT,GO e DF), no clusters 2 só apenas um estado (SP), o clusters 4 por doze estados (PA, AM, AC, BA, PB, PE, CE, RN, MA, AL, SE e PI) e o clusters 5 com três estados (PR, MG e RJ), 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 confirmados de Covid-19 em média se encontra no clusters 4 juntamente com a maior média de números de óbitos, já o clusters 1 possui o número médio de mortalidade maior, o clusters 2 possui a média menor em todas as variáveis e o clusters 3 possui uma média maior em relação aos outros clusters.

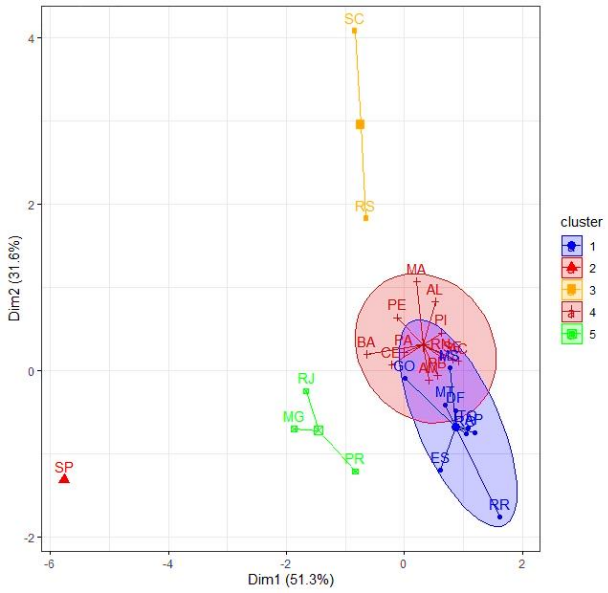


Figura 3: Gráfico obtido por meio do método de K-means para os dados do Covid-19

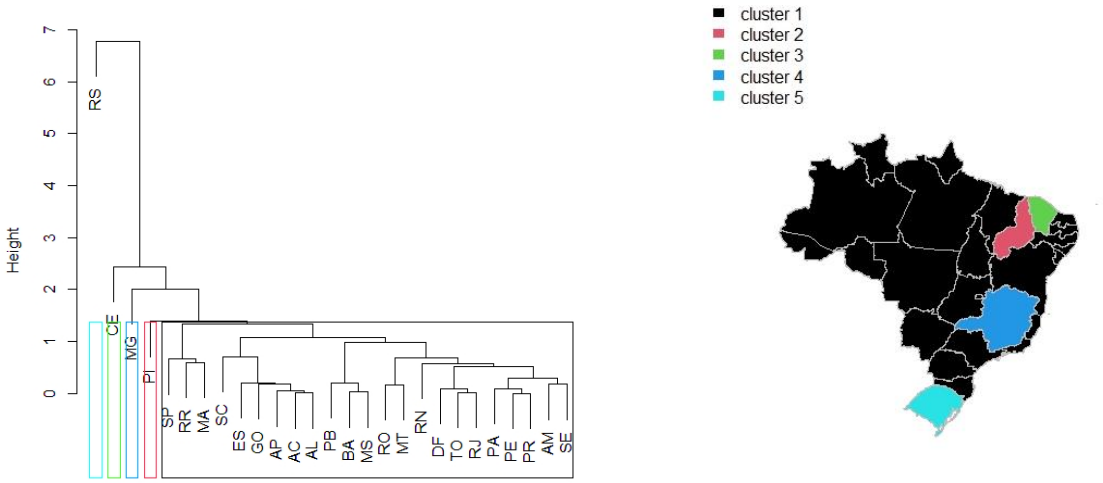
Em seguida foi realizado um estudo de medida de distância e método de agrupamento utilizando para a construção do dendrograma (Figura 4 a 7) com a variável Casos confirmados com a combinação da distância de Mahalanobis (D^2) e métodos de agrupamento (Ligação Simples, Ligação Completa, Ligação Median e Ligação Ward). Com essa combinação de distância da distância de Mahalanobis (D^2) e métodos de agrupamento foram obtidos o coeficiente de correlação cofenética (CCC) com intuito de medir o grau de ajuste entre as matrizes formadas (Tabela 3), ou seja, a distância de Mahalanobis (D^2) 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 3: Coeficiente de correlação cofenético para os dados da Covid-19.

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

Fonte: Autoria própria.

No método de ligação Simples denotado como “método do vizinho mais próximo”, na Figura 4 (a) é mostrado agrupamento dos casos de Covid-19 em cinco grupos por meio de um dendrograma. Verifica-se na Figura 4 (b) a distribuição dos clusters no mapa do Brasil. O grupo 1 com maior quantidade de estados de diferentes regiões do país, onde podemos perceber que nele se encontra o estado com o maior número de casos confirmados é São Paulo com 4.449.552 habitantes, ou seja, quanto maior for a população do estado maior era o número de casos confirmados e o estado. com menor número de casos confirmados é o Amapá = 125.336 habitantes. No entanto os outros grupos (2 a 5) com apenas um estado cada.

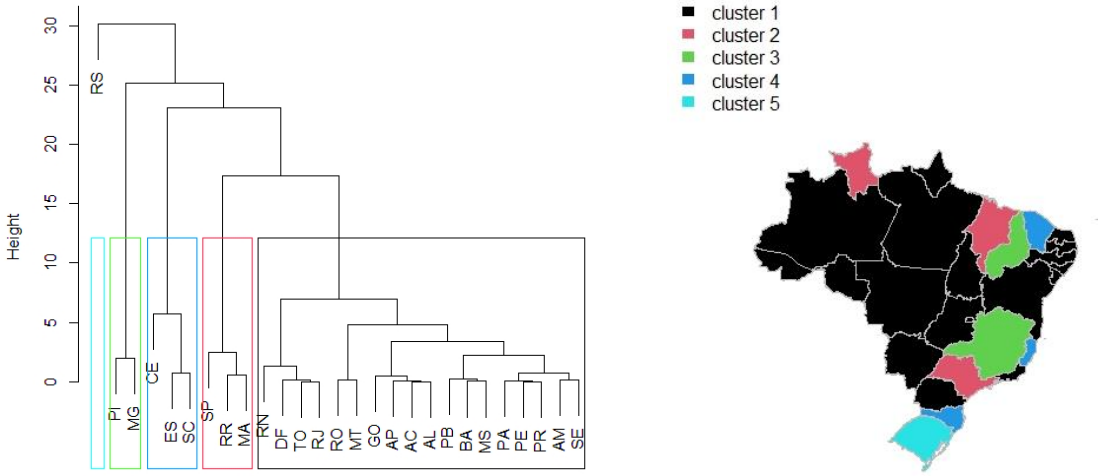


(a) (b)

Figura 4: Agrupamento dos Casos Confirmados da Covid-19 com base na distância mahalanobis e o método de ligação Simples.

Com o método de ligação Completa por definição agrupa os elementos com menor distância entre os mais distantes, na Figura 5 (a) mostra o agrupamento dos casos de covid-19 em cinco grupos por meio de um dendrograma, na Figura 5 (b) verifica-se a distribuição dos grupos no mapa do Brasil. O grupo 1 com a maior quantidade de estados de diferentes regiões, nele se encontra os estados que um

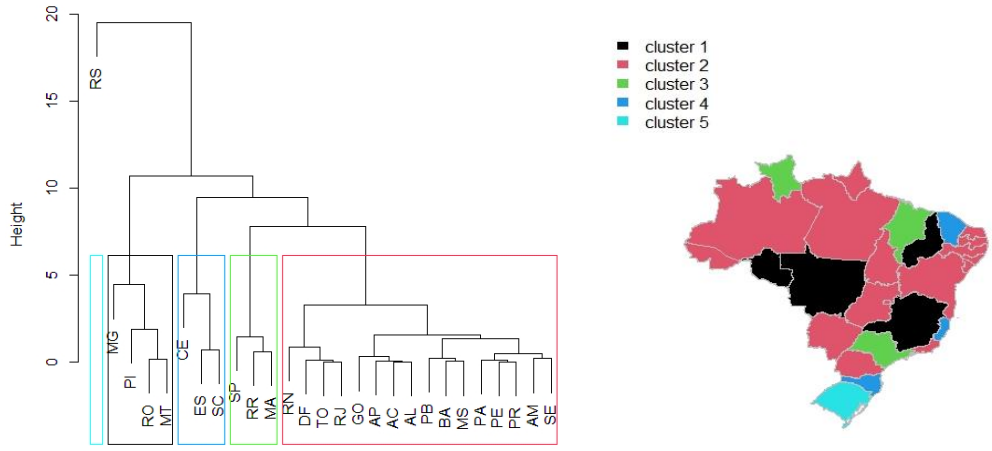
índice de casos confirmados próximos, a Bahia localizado na região Nordeste com uma população estimada em 14.985.284 habitantes e o Paraná com 11.597.484 habitantes localizado na região do Sul do Brasil. O grupo 2 e 4 com apenas três estados cada, o grupo 3 com dois estados Piauí e Mina Gerais e o grupo 5 com um estado Rio Grande do Sul.



(a) (b)

Figura 5: Agrupamento dos Casos Confirmados da Covid-19 com base na distância mahalanobis e o método de ligação Completa

Em relação ao método de ligação média na Figura 6 (a), mostra o agrupamento dos casos de Covid-19 em cinco grupos por meio de um dendrograma. A distribuição dos grupos pode ser verificada na Figura 6 (b) com o mapa do Brasil. O grupo 1 composto por 4 estados (Rondônia, Piauí, Mato Grosso e Mina Gerais), o grupo 2 com a maior quantidade de estados, o grupo 3 e 4 com três estados cada, o grupo 5 com um estado.

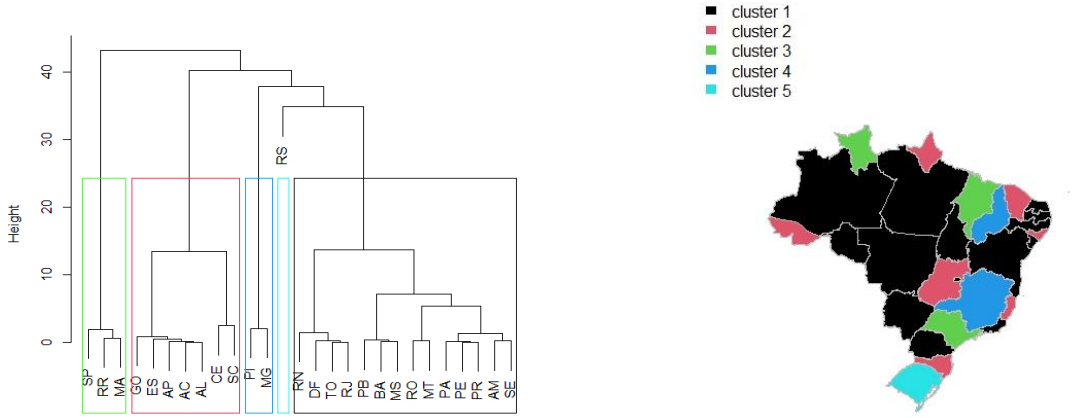


(a) (b)

Figura 6: Agrupamento dos Casos Confirmados da Covid-19 com base na distância mahalanobis e o método de ligação Média.

Já o método de Ward na Figura 7 (a), mostra o agrupamento dos casos de Covid-19 em cinco grupos por meio de um dendrograma. A distribuição dos grupos pode ser verificada na Figura 7 (b) com o mapa do Brasil. O grupo 1 é o que possui a

maior quantidade de estados, o grupo 2 com sete estados de todas as regiões do país. O grupo 3 com três estados, o grupo 4 com dois estados e o grupo 5 com um estado. Nesse método agrupa os estados que possuem a menor soma dos quadrados das distâncias, nele pode ser destacado o grupo 5 com estado de Rio Grande do sul que possui o número de casos confirmados igual a 1498577 e se manteve isolado em todos os outros métodos utilizados.



(a) (b)

Figura 7: Agrupamento dos Casos Confirmados da Covid-19 com base na distância mahalanobis e o método de ligação Ward.

3.2 Análise dos dados IDH

Utilizando dados referente ao Índice de Desenvolvimento Humano dos 27 estados brasileiros e as variáveis disponíveis são IDH, IDH-R, IDH-L e IDH-E, para uma melhor compreensão das variáveis em estudo, foi realizado uma análise descritiva (Tabela 4) podendo observar que a variável com maior média foi as dos IDH-L e a menor média foi a de IDH-E, o desvio padrão das variáveis foram menores em relação à média.

Tabela 4: Análise descritiva das variáveis dos dados IDH

Variáveis	Média	Desvio Padrão	Min	Max
IDH	0,7045	0,0492	0,6310	0, 8240
IDH-R	0,7069	0,0582	0,6120	0,8630
IDH-L	0,8086	0.0304	0,7550	0,8730
IDH-E	0,6124	0,0582	0,5200	0,7420

Fonte: Autoria própria.

Nesses dados também foram utilizados o Método direto para otimizar que é a soma dos quadrados do cluster usando o método do cotovelo. Para determinar o número ideal de cluster, esse método é representado através de um gráfico (Figura 8) da soma dos quadrados do cluster em função do número de agrupamentos, logo para esses dados o número ideal de cluster é 5.

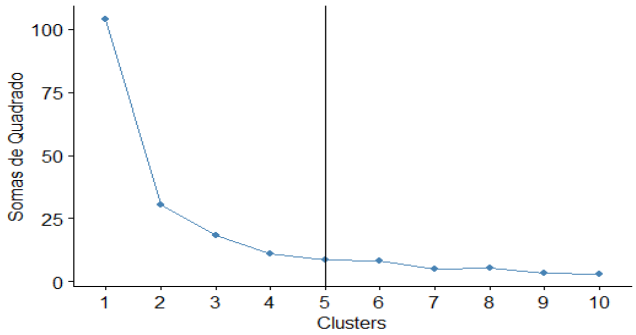


Figura 8: Referente ao número de cluster obtido para os dados do IDH.

Aplicando os 5 clusters na distância Euclidiana com o método de K-means na Figura 9, observou-se no clusters 1 formando por cinco estados (RO, RN, CE, AM e PE), clusters 2 por oito estados (SE, AC, BA, PB, PI, MA e AL), o cluster 3 com três estados (AP, RR, e TO), no cluster 4 também com três estados (DF, SP e SC) e o clusters 5 com oito estados (RJ, PR, RS, ES, GO, MG, MS e MT), nesse método é possível verificar as características de cada aglomeração. Com base na média do K-means de cada clusters observou-se que o estado com o maior IDH em média se encontra no cluster 4 e 5 juntamente com a maior média também de IDH-L, IDH-R e IDH-E, já o clusters 1 possui a média menor em todas as variáveis.

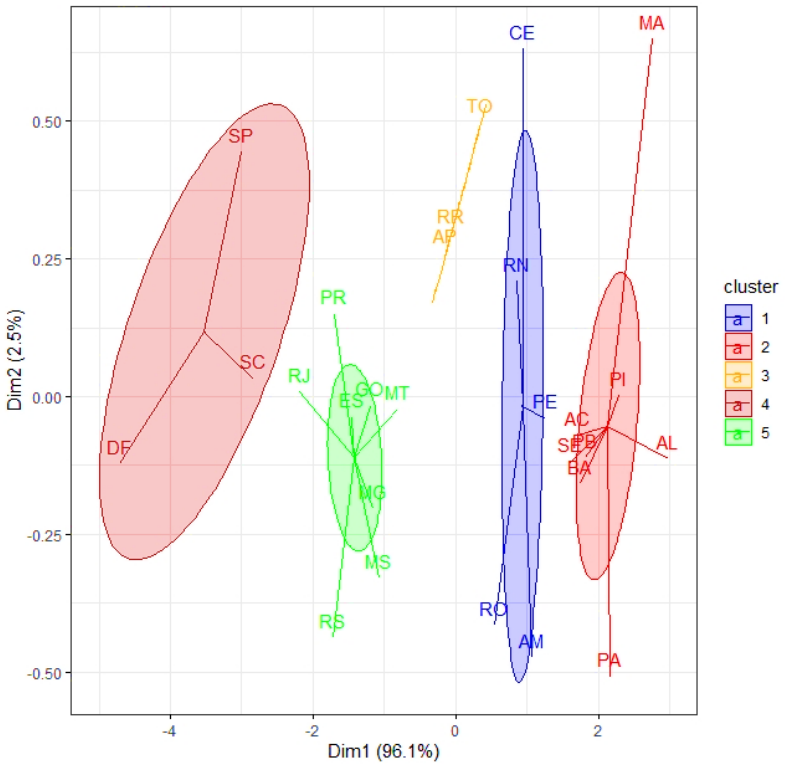


Figura 9: Gráfico obtido através do método de K-maens para os dados do IDH.

Também foi realizado um estudo de medida de distância e método de agrupamento utilizados para a construção do dendrograma (Figura 10 a 13) com a variável IDH utilizando a combinação da distância de Mahalanobis (D^2) com os métodos de agrupamento.

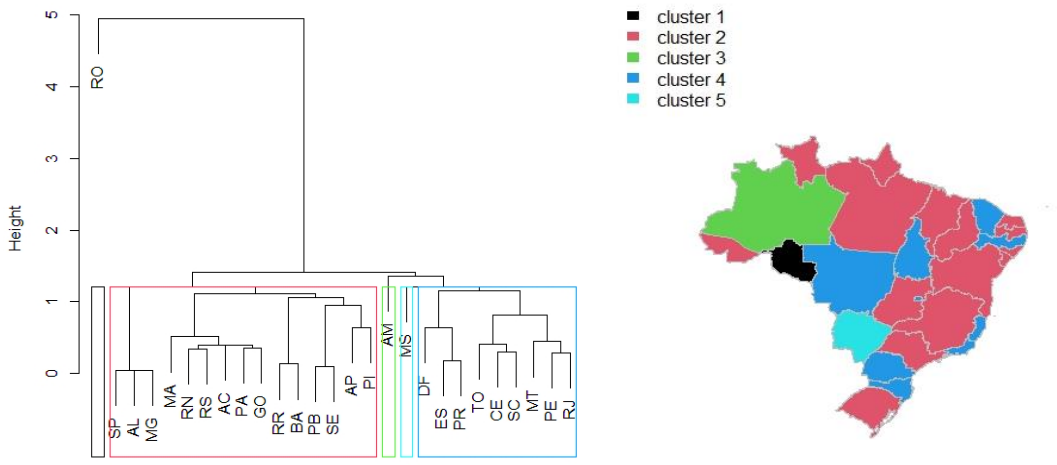
Essa combinação da distância de Mahalanobis (D^2) e métodos de agrupamento (Ligação Simples, Ligação Completa, Ligação Media e Ward), foram obtidos o coeficiente de correlação cofenética (CCC) como intuito de medir o grau de ajuste entre as matrizes formadas (Tabela 5), ou seja, segundo Nascimento (2022) a distância de Mahalanobis (D^2) e métodos de ligação completa obtiveram o maior valor para o CCC que foi igual a 0,728, como este valor estar próximo de 1 a CCC é considerando forte, logo o estado que estão dentro do mesmo grupo podem ser agrupados de outras maneiras quando muda o método.

Tabela 5: Coeficiente de correlação cofenético obtidos para os dados do IDH.

Ligações	Simples	Completa	Média	Ward
Correlação	0,676	0,728	0,669	0,474

Fonte: Autoria própria.

Composto por cinco clusters o método de ligação simples na Figura 10 (a), mostra o agrupamento do IDH em cinco clusters por meio de um dendrograma. Verifica-se na Figura 10 (b), a distribuição dos clusters em um mapa do Brasil, onde apresenta três cluster unitários clusters 1 com Rondônia, clusters 3 com Amazonia e clusters 5 Mato grosso do Sul com e dois grupos com maior número de estados de diferentes regiões, o clusters 2 (São Paulo, Alagoas, Mina Gerais, Maranhão, Rio Grande do Norte, Rio Grande do Sul, Acre, Pará, Goiás, Roraima, Bahia, Paraíba, Sergipe, Amapá e Piauí) e o clusters 5 (Tocantins, Ceará, Pernambuco, Espírito Santo, Rio de Janeiro, Paraná, Santa Catarina, Mato Grosso e Distrito Federal). Esse método mostra que há uma variação de estados no cluster 2, independente do IDH dos estados.

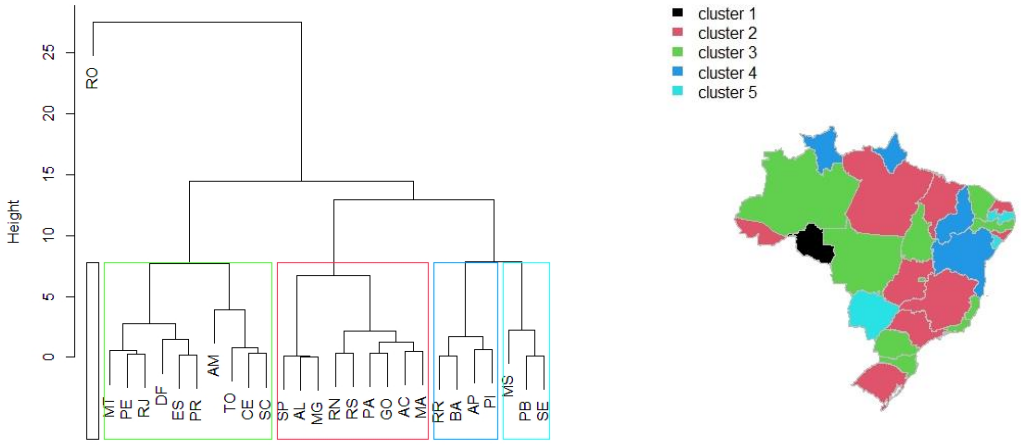


(a) (b)

Figura 10: Agrupamento da variável IDH com base na distância mahalanobis e o método de ligação Simples.

Com o método de ligação completa mostra na Figura11 (a) o agrupamento do IDH em cinco clusters por meio de um dendrograma. Verifica-se na Figura 11 (b), a distribuição dos clusters em um mapa do Brasil, observa-se o clusters 1 composto por um estado (Rondônia), o clusters 2 com nove estados (Acre, Pará, Maranhão, Rio Grande do Norte, Alagoas, Minas Gerais, São Paulo, Rio grande do Sul e Goiás), o

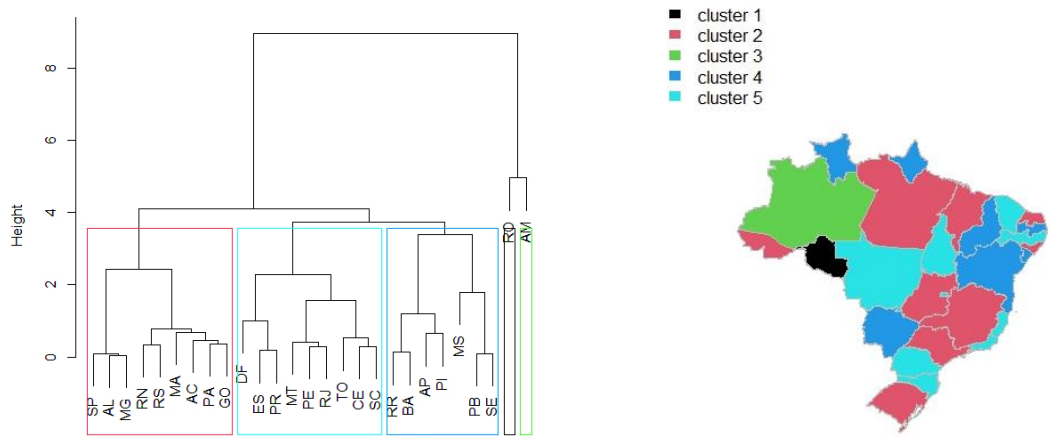
clusters 3 com dez estados (Amazona, Tocantins, Ceará, Pernambuco, Espírito Santo, Rio de Janeiro, Santa Catarina, Paraná, Mato Grosso e Distrito Federal), o clusters 4 com quatro estados (Roraima, Amapá, Piauí e Bahia) e clusters 5 com três estados (Paraíba, Mato grosso do Sul e Sergipe) pois os estados que compõem tem um IDH próximos, mesmo sendo de regiões diferentes.



(a) (b)

Figura 11: Agrupamento da variável IDH com base na distância mahalanobis e o método de ligação Completa.

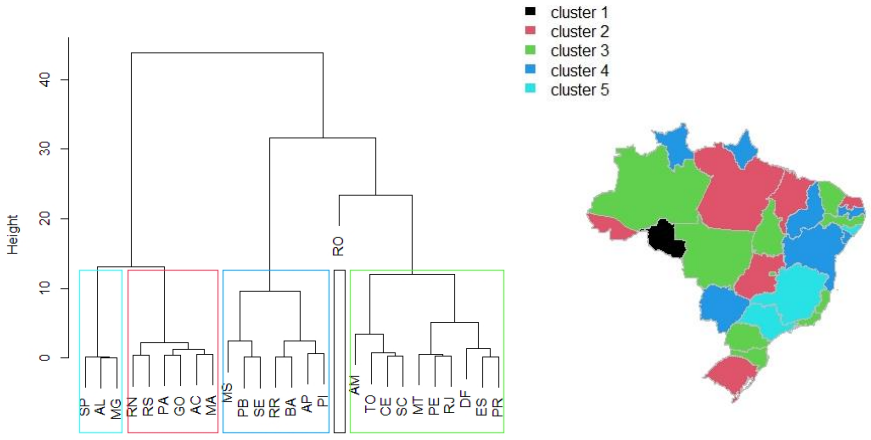
O método de ligação média mostra na Figura 12 (a) o agrupamento do IDH em cinco clusters por meio de um dendrograma. Na Figura 12 (b) verifica, a distribuição dos clusters em um mapa do Brasil, o clusters 2 tem nove estados (São Paulo, Alagoas, Minas Gerais, Rio Grande do Norte, Rio grande do Sul, Acre, Maranhão, Pará, Goiás), o clusters 4 obteve sete estados (Roraima, Amapá, Piauí, Paraíba, Sergipe, Bahia, Mato grosso do Sul), o clusters 5 com nove estados (Distrito Federal, Espírito Santo, Tocantins, Paraná, Mato Grosso, Pernambuco, Rio de Janeiro, Ceará e Santa Catarina) e o clusters 1 e 3 com apenas um estados (Rondônia e Amazona), nesse método agrupa aqueles com a menor distância média, ou seja, pode-se destaca o clusters 1 e 2 formado com estados de regiões diferente e o que possui o maior índice de IDH é São Paulo e o que teve a o menor índice de IDH foi Alagoas.



(a) (b)

Figura 12: Agrupamento da variável IDH com base na distância mahalanobis e o método de ligação Média.

Já o método de Ward pode ser observar no dendrograma Figura 13 (a) o agrupamento do IDH em cinco clusters. Na Figura 13 (b) verifica, a distribuição dos cluster em um mapa do Brasil, o clusters 1 com um estado o de Rondônia, clusters 2 com a presença de seis estados (Rio Grande do Norte, Rio grande do Sul, Pará, Goiás, Acre e Maranhão), o clusters 3 o maior com dez estados (Amazona, Tocantins, Ceará, Santa Catarina, Mato Grosso, Pernambuco, Rio de Janeiro, Distrito Federal, Espírito Santo e Paraná), no clusters 4 com sete estados (Mato grosso do Sul, Paraíba, Sergipe, Roraima, Bahia, Amapá, Piauí), e o cluster 5 formado por três estados (São Paulo, Alagoas e Minas Gerais), nesse método se destaca o clusters 1 com o estado de Rondônia pois ele se manteve isolado.



(a) (b)

Figura 13: Agrupamento da variável IDH com base na distância mahalanobis e o método de ligação Ward.

3.3 Análise de Correlação entre os dados

Os coeficientes de correlação são métodos estatísticos para se medir as relações entre variáveis e o que elas representam, ou seja, procura entender como uma variável se comporta em um cenário onde outra está variando, visando identificar se existe alguma relação entre a variabilidade de ambas, serão aplicados aos dados dos 27 estados do Brasil referente os dados da COVID-19 e suas variáveis (Casos confirmados, Óbitos, Incidência e Mortalidade) e os dados referente ao Índice de Desenvolvimento Humano com as variáveis IDH, IDH-R, IDH-L e IDH-E. Pode observar na Figura 14 as variáveis que possui uma correlação forte e positiva é a mortalidade e IDH com 0,730, já em relação ao IDH e Casos tiveram uma correlação positiva de 0,50, ou seja, há uma correlação positiva entre os dados.

Figura 14: Gráfico referente a Correlação entre os dados da Covdi-19 e o IDH.

4 CONCLUSÃOS

As técnicas de análise de agrupamento têm grande importância, devido a complexidade de melhor interpretação dos dados. Nesse estudo podemos destacar as técnicas de análise de agrupamento hierárquicos e não hierárquicos. Ao aplicar essas técnicas em alguns conjuntos de dados, verificou-se sua aplicabilidade, onde foi possível a extração de informações sobre os dados, embora a análise seja mais complexa, uma vez que os agrupamentos não são determinados previamente.

Na primeira análise, aplicamos a técnica não-hierárquica do k-means, com a função de similaridade da distância Euclidiana, nessa análise os resultados foram bastantes satisfeitos para os cinco grupos formados de ambos os dados. Na segunda análise empregamos diversas técnicas hierárquica de agrupamento, com a distância de mahalanobins como medida de similaridade. Em seguida utilizou o método de ligação do vizinho mais próximo, ligação do vizinho de menor distância, ligação do vizinho de menor distância média e ligação de vizinho da menor soma de quadrado das distâncias, logo os resultados foram satisfatório, o método que melhor representou ambos os dados foi o de ligação completa com maior coeficiente de correlação cofenética (CCC), o estado de Rio Grande do Sul se manteve em um cluster para os dados da Covid-19 e o estado de Rondônia para os dados do IDH em todos os métodos de ligação.

Em uma última análise foi verificado que há uma correlação positiva entre as variáveis, dos dados do Covid-19 em relação ao IDH, ou seja, verificou-se que a variável mortalidade teve a maior correlação com IDH, IDH-R, IDH-L e IDH-E, portanto, a uma correlação entre os dados.

REFERÊNCIAS

Aaker, D. A., et al. (2001) - Pesquisa de Marketing. *Tradutor Reynaldo Cavalheiro Marcondes*. Atlas. São Paulo.

Agudelo, S.C.; Detección de Outliers usando Métricas de Distancia y Análisis Cluster, *Ingeniería Matemática*, Universidad EAFIT, 2021.

Albuquerque, M.A.; Estabilidade em análise de agrupamento: *Estudo de caso em ciência florestal1*, Viçosa-MG, v.30, n.2, p.257-265, 2005.

Albuquerque, M. A.; Ferreira, R. L. C.; Silva, J. A. A.; Santos, E. S.; Stosic, B.; Souza, A. L. *Estabilidade em análise de agrupamento*: estudo de caso em ciência florestal. Revista Árvore, Viçosa-MG, v. 30, n. 2, p. 257-265,2006.

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-teoria-e-pratica-com-aplicacoes-em-r/?wpdmdl=997&masterkey=5e97904980fc9.*

Alves. H. J. P.; A pandemia da COVID-19 no Brasil: *Uma aplicação de método de clusterização k-means*, Researd, Society and Development, v.9, n.10, e58109059, 2020.

Bezerra, A.C.V.; *Fatores associados ao comportamento da população durante o isolamento social na pandemia de COVID-19*, https://doi.org/10.1590/1413-81232020256.1.10792020.

Campos, S.LS.; *Busca não supervisionada de padrões por técnicas de agrupamento clássica e nebulosa*, Juiz de Fora, 2019.

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

Clifford, H.T.; Stephenson, W. An introduction to numerical taxonomy. London: *Academic Press*, 1975.

Cruz, C.D.; REGAZZI, A.J. *Modelos biométricos aplicados ao melhoramento genético*. Viçosa: UFV, 1994.

Domingos. M. F. N.; Relação entre Índice de Desenvolvimento Humano e número de casos de COVID-19 em cidades do Tocantins, v. 1 n. 2, Singular, Saúde e Biológicas CEULP/ULBRA, 2021.

Dourado. P. B. M.; *Relação da COVID-19 com o Índice de Desenvolvimento Humano – IDH.* Síntese de Evidências e Análise Exploratória, Subsecretaria de Saúde, Gerência de Informações Estratégicas em Saúde CONECTA-SUS, 2021.

Duarte, M. C.; santos, J. B.; MELO, L. C. Comparison of similarity coefficients based on RAPD markers in the common bean. Genetics e Molecular Biology, v. 22, n. 3, p. 427- 432, 1999.

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.

Fávero, L. P., & Belfiore, P.; Data Science for Business and Decision Making. *Academic Press*, Cambridge, MA, USA, 2019.

Gower, J. C.; Legendre, P. Metric e euclidean properties of dissimilarity coefficients, *Journal of Classification*, v. 3, p. 5 - 48, 1986.

Jackson, A. A.; Somers, K. M.; Harvery, H. H. Similarity coefficients: measures for co-occurrence e association or simply measures of occurrence. *American Naturalist*, v.133, p. 436 - 453, 1989.

Johnson, R.A. e Wichern, D. W.; *Applied multivariate statistical analysis*. Englewood Cliffs: Prentice Hall, 4a ed., 1998.

Johnson, RA e Wichern, DW (1992) Applied Multivariate Statistical Analysis. *Prentice Hall*, Penhascos de Englewood.

Kassambara, A.; *Practical Guide To Cluster Analysis in R Unsupervised Machine Learning*, sthda.com Edition 1, Copyright ©2017.

Lopes, H. E.G.; Gosling, M. S.; Cluster Analysis in Practice: Dealing with Outliers in Managerial Research, *Revista de Administração Contemporânea*, v. 25, n. 1, e-200081, 2021.

Marcondes, D.M.S.V.; Completamente de matrizes de Distância Euclidiana, Instituto de Matemática e Estatística, *Universidade de São Paulo*, 2020.

Mardia, A.K.V.; Kent, J.T.; Bibby, J.M. Multivariate analysis London*: Academic Press*, 1997. 518p.

Matos, A. P. C.; Aplicação de Ferramentas Quimiométricas para Categorização de Vinhos através dos seus Componentes Químicos, Faculdade de Ciências e Tecnologia da *Universidade de Coimbra*, 2021.

Nascimento, E.R.; Cluster analysis applied to the Human Development Index (HDI) of Brazilian States, Research, Society and Development 11.2, 2022: e18011225747-e18011225747.

OMS. Organização Mundial de Saúde. Coronavirus disease (COVID-19) pandemic. Disponivel em*: https://www.who.int/emergencies/diseases/novel-coronavirus-2019. Acessado em 2022*.

REIS, E. Estatística multivariada aplicada. *Lisboa*, 1997.

Sartorio, S. D.; *Aplicações de técnicas de análise multivariada em experimentos agropecuários usando o software R*, Universidade de São Paulo, Escola Superior de Agricultura “Luiz de Queiroz”, Piracicaba, 2008.

Souza, M. V. V.; Análise 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.

Sneath, P. H. A; Sokal, R. R. Numeric taxonomy: the principles e practice of numerical classification. *San Francisco: W. H. Freeman*, 1973.

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 (HDI)*, acessado em 2022.

Anexo A

```
##############          Análise dados Covid-19      ###################
library(cluster)
library(ecodist)

## Método de K-Means

DDCovid = read.csv("CO-19.csv",sep = ";")
DDCovid

rownames(DDCovid) <- DDCovid$ï..Estados
View(DDCovid)

dados <- DDCovid[,2:5]

View(dados)
str(dados)

as.numeric(dados$casos)
as.numeric(dados$Obitos)
as.numeric(dados$Incidencia)
as.numeric(dados$Mortalidade)

summary(dados)
sd(dados$casos)
sd(dados$obitos)
sd(dados$Incidencia)
sd(dados$Mortalidade)

dados_z = scale(dados)
View(dados_z)

library(cluster)
library(factoextra)

### Obtendo o número óptimo de clusters

set.seed(123) # Fixando a semente

fviz_nbclust(dados_z, kmeans, method = "wss") +
  ggtitle("") +
  xlab("Clusters") + ylab("Somas de Quadrado")

# OBS: 5 clusters é o ideal

fviz_nbclust(dados_z, kmeans, method = "wss") +
  ggtitle("") +
  xlab("Clusters") + ylab("Somas de Quadrado") +
  geom_vline(xintercept = 5)
```

```
### Obtendo os clusters

set.seed(123) # Fixando a semente

km <- kmeans(dados_z, centers = 5)

cbind(km$size) # Tamanho dos clusters

km$centers # Médias dos clusters

cbind(km$cluster)

### Gráfico dos clusters

library(ggplot2)
library(factoextra)
windows()
fviz_cluster(km, data = dados_z,
             palette = c("blue", "red", "orange","brown", "green" ),
             ellipse.type ="t",
             star.plot=TRUE,
             ggtheme = theme_bw()
)


## Método Hierárquico

x = c ( 280495, 6671, 15782.8,  375.4,
        88264,  1849, 10008.0, 209.7,
        431263,13812, 10405.4, 333.5,
        128751,2063,  21254.4, 340.6,
        614815,16970, 7146.6, 197.3,
        125336,2007,  14819.8, 237.3,
        234113,3927,  14884.5, 249.7,
        367204,10335, 5190.0, 146.1,
        333573,7225,  10191.0, 220.7,
        953019,24717,10435.9, 270.7,
        384333,7518, 10959.5, 214.4,
        462567,9556,  11512.0, 237.8,
        643307,20310,6731.2,212.5,
        241780,6367,  7244.7, 190.8,
        278374,6050,  12110.1, 263.2,
        1264804, 27384, 8504.0, 184.1,
        2214356, 56443,10460.5, 266.6,
        624425,13237, 15538.2, 329.4,
        1349461, 69299, 7816.2, 401.1,
        4449552, 154691, 9690.0, 336.9,
        1586917, 40828, 13879.0, 357.1,
        1237533, 20076, 1727.4, 280.2,
```

```
        1498577, 36273, 13171.7, 318.8,
        379502,9704,  13656.1, 349.2,
        551056,13761, 15814.6, 394.9,
        943102,24604, 13437.7, 350.6,
        518345,11058, 17190.7, 366.7)

x=matrix(x,nrow=4,ncol=27,byrow=TRUE)

y=c("RO","AC","AM","RR","PA","AP","TO","MA","PI","CE",
    "RN","PB","PE","AL","SE","BA","MG","ES","RJ","SP","PR",
    "SC","RS","MS","MT","GO","DF")

# Método Ward

colnames(x) <- y
x=t(x)
dist1<-distance(x,"mahalanobis")
clust1<-hclust(dist1,"ward.D")

Cluster<- cutree(clust1,5)
city_label <- as.vector(mapa$"abbrev_state ")
names(Cluster) <- city_label


plot(clust1,
     main = "dendograma Ward")
rect.hclust(clust1 ,k = 5, border = c(3,2,4,5,1))
legend("topright", legend = paste("cluster",1:5),
       fill=1:5,bty= "n", border = "white")


plot(mapa[,2], border = "grey",xlim=c(-90,-10), col = Cluster,
     main = "método Ward")
legend("topleft",legend = paste("cluster",1:5),
       fill = 1:5, bty = "n", border = "white")

#Matriz Cofenética
options(digits=3)
cophenetic(clust1)
cor(dist1,cophenetic(clust1))
var(cophenetic(clust1))
colMeans(x)

# Método simples

dist2<-distance(x,"mahalanobis")
clust2<-hclust(dist2,"single")

Cluster2<- cutree(clust2,5)
city_label <- as.vector(mapa$"abbrev_state")
```

```
names(Cluster2) <- city_label


plot(clust2,
     main = "dendograma simples")
rect.hclust(clust2 ,k = 5, border = c(5,3,4,2,1))
legend("topright", legend = paste("cluster",1:5),
       fill=1:5,bty= "n", border = "white")


plot(mapa[,2], border = "grey",xlim=c(-90,-10), col = Cluster2,
     main = "método Simples")
legend("topleft",legend = paste("cluster",1:5),
       fill = 1:5, bty = "n", border = "white")

#Matriz Cofenética
options(digits=3)
cophenetic(clust2)
cor(dist2,cophenetic(clust2))
var(cophenetic(clust2))
colMeans(x)

# Método completa

dist3<-distance(x,"mahalanobis")
clust3<-hclust(dist3,"complete")

Cluster3<- cutree(clust3,5)
city_label <- as.vector(mapa$"abbrev_state")
names(Cluster3) <- city_label


plot(clust3,
     main = "dendograma Completa")
rect.hclust(clust3 ,k = 5, border = c(5,3,4,2,1))
legend("topright", legend = paste("cluster",1:5),
       fill=1:5,bty= "n", border = "white")

plot(mapa[,2], border = "grey",xlim=c(-90,-10), col = Cluster3,
     main = "método Completa")
legend("topleft",legend = paste("cluster",1:5),
       fill = 1:5, bty = "n", border = "white")

#Matriz Cofenética
options(digits=3)
cophenetic(clust3)
cor(dist3,cophenetic(clust3))
var(cophenetic(clust3))
colMeans(x)
```

```
# Método de Média

dist4<-distance(x,"mahalanobis")
clust4<-hclust(dist4,"median")

Cluster4<- cutree(clust4,5)
city_label <- as.vector(mapa$"abbrev_state")
names(Cluster4) <- city_label

plot(clust4,
     main = "dendograma média")
rect.hclust(clust4 ,k = 5, border = c(5,1,4,3,2))
legend("topright", legend = paste("cluster",1:5),
       fill=1:5,bty= "n", border = "white")

plot(mapa[,2], border = "grey",xlim=c(-90,-10), col = Cluster4,
     main = "método média")
legend("topleft",legend = paste("cluster",1:5),
       fill = 1:5, bty = "n", border = "white")

#Matriz Cofenética
options(digits=3)
cophenetic(clust4)
cor(dist4,cophenetic(clust4))
var(cophenetic(clust4))
colMeans(x)
```

Anexo B

```
##############     Análise dados  IDH     ###################
library(cluster)
library(ecodist)

# Método de K-Means

IDH1 = read.csv("IDH.csv",sep = ";")
IDH1

rownames(IDH1) <- IDH1$ï..Estados
View(IDH1)

dados <- IDH1[,2:5]
View(dados)
str(dados)
as.numeric(dados$IDH)
as.numeric(dados$IDH.R)
as.numeric(dados$IDH..L)
as.numeric(dados$IDH..E)

summary(dados)
sd(dados$IDH)
sd(dados$IDH.R)
sd(dados$IDH..L)
sd(dados$IDH..E)

dados_z = scale(dados)
View(dados_z)

library(cluster)
library(factoextra)

### Obtendo o número ótimo de clusters

set.seed(123) # Fixando a semente

fviz_nbclust(dados_z, kmeans, method = "wss") +
  ggtitle("") +
  xlab("Clusters") + ylab("Somas de Quadrado")

fviz_nbclust(dados_z, kmeans, method = "wss") +
  ggtitle("") +
  xlab("Clusters") + ylab("Somas de Quadrado") +
  geom_vline(xintercept = 5)

### Obtendo os clusters

set.seed(123) # Fixando a semente
```

```
km <- kmeans(dados_z, centers = 5)

cbind(km$size) # Tamanho dos clusters

km$centers # Médias dos clusters

cbind(km$cluster)
km

### Gráfico dos clusters

library(ggplot2)
library(factoextra)
windows()
fviz_cluster(km, data = dados_z,
             palette = c("blue", "red", "orange","brown", "green" ),
             geom = "text",
             ellipse.type = "confidence",
             star.plot=TRUE,
             ggtheme = theme_bw()
)


## Método Hierárquico

x=c( 0.863,   0.873, 0.742,
     0.789,   0.845, 0.719,
     0.773,   0.860, 0.697,
     0.782,   0.835, 0.675,
     0.757,   0.830, 0.668,
     0.769,   0.840, 0.642,
     0.743,   0.835, 0.653,
     0.742,   0.827, 0.646,
     0.730,   0.838, 0.638,
     0.740,   0.833, 0.629,
     0.732,   0.821, 0.635,
     0.694,   0.813, 0.629,
     0.695,   0.809, 0.628,
     0.690,   0.793, 0.624,
     0.712,   0.800, 0.577,
     0.678,   0.792, 0.597,
     0.651,   0.793, 0.615,
     0.677,   0.805, 0.561,
     0.673,   0.789, 0.574,
     0.672,   0.781, 0.560,
     0.671,   0.777, 0.559,
     0.663,   0.783, 0.555,
     0.656,   0.783, 0.555,
     0.646,   0.789, 0.528,
```

```
      0.635,     0.777, 0.547,
      0.612,     0.757, 0.562,
      0.641,     0.755, 0.520)

x=matrix(x,nrow=3,ncol=27,byrow=TRUE)

y=c("RO","AC","AM","RR","PA","AP","TO","MA","PI","CE",
    "RN","PB","PE","AL","SE","BA","MG","ES","RJ","SP","PR",
    "SC","RS","MS","MT","GO","DF")

# Método Ward

colnames(x) <- y
x=t(x)
dist1<-distance(x,"mahalanobis")
clust1<-hclust(dist1,"ward.D")

Cluster<- cutree(clust1,5)
city_label <- as.vector(mapa$"abbrev_state ")
names(Cluster) <- city_label

plot(clust1,
     main = "dendograma Ward")
rect.hclust(clust1 ,k = 5, border = c(3,2,4,5,1))
legend("topright", legend = paste("cluster",1:5),
       fill=1:5,bty= "n", border = "white")


plot(mapa[,2], border = "grey",xlim=c(-90,-10), col = Cluster,
     main = "método Ward")
legend("topleft",legend = paste("cluster",1:5),
       fill = 1:5, bty = "n", border = "white")

#Matriz Cofenética
options(digits=3)
cophenetic(clust1)
cor(dist1,cophenetic(clust1))
var(cophenetic(clust1))
colMeans(x)

# Método simples

dist2<-distance(x,"mahalanobis")
clust2<-hclust(dist2,"single")

Cluster2<- cutree(clust2,5)
city_label <- as.vector(mapa$"abbrev_state")
names(Cluster2) <- city_label
```

```
plot(clust2,
     main = "dendograma simples")
rect.hclust(clust2 ,k = 5, border = c(5,3,4,2,1))
legend("topright", legend = paste("cluster",1:5),
       fill=1:5,bty= "n", border = "white")


plot(mapa[,2], border = "grey",xlim=c(-90,-10), col = Cluster2,
     main = "método Simples")
legend("topleft",legend = paste("cluster",1:5),
       fill = 1:5, bty = "n", border = "white")

#Matriz Cofenética
options(digits=3)
cophenetic(clust2)
cor(dist2,cophenetic(clust2))
var(cophenetic(clust2))
colMeans(x)

# Método completa

dist3<-distance(x,"mahalanobis")
clust3<-hclust(dist3,"complete")

Cluster3<- cutree(clust3,5)
city_label <- as.vector(mapa$"abbrev_state")
names(Cluster3) <- city_label

plot(clust3,
     main = "dendograma Completa")
rect.hclust(clust3 ,k = 5, border = c(5,3,4,2,1))
legend("topright", legend = paste("cluster",1:5),
       fill=1:5,bty= "n", border = "white")

plot(mapa[,2], border = "grey",xlim=c(-90,-10), col = Cluster3,
     main = "método Completa")
legend("topleft",legend = paste("cluster",1:5),
       fill = 1:5, bty = "n", border = "white")

#Matriz Cofenética
options(digits=3)
cophenetic(clust3)
cor(dist3,cophenetic(clust3))
var(cophenetic(clust3))
colMeans(x)

# Método de Média

dist4<-distance(x,"mahalanobis")
clust4<-hclust(dist4,"median")
```

```
Cluster4<- cutree(clust4,5)
city_label <- as.vector(mapa$"abbrev_state")
names(Cluster4) <- city_label

plot(clust4,
     main = "dendograma média")
rect.hclust(clust4 ,k = 5, border = c(5,1,4,3,2))
legend("topright", legend = paste("cluster",1:5),
       fill=1:5,bty= "n", border = "white")

plot(mapa[,2], border = "grey",xlim=c(-90,-10), col = Cluster4,
     main = "método média")
legend("topleft",legend = paste("cluster",1:5),
       fill = 1:5, bty = "n", border = "white")

#Matriz Cofenética
options(digits=3)
cophenetic(clust4)
cor(dist4,cophenetic(clust4))
var(cophenetic(clust4))
colMeans(x)
```

AGRADECIMENTOS

Primeiramente agradeço a Deus, que por sua infinita bondade tem me sustentado ao longo desses anos de graduação, sem sua presença seria impossível o meu crescimento.

Aos meus pais, Maria Hilda e João Felix, aos meus irmãos: Maria das Graças, Manuel, Christiana e Ezaquiel. Os maiores incentivadores dos meus sonhos.

À UEPB e ao Departamento de Estatística por oferecerem a estrutura necessária para as aulas e os estudos e também ao Programa Institucional de Bolsas de Iniciação Científica (PIBIC).

Ao professor Mácio Augusto de Albuquerque por me orientar durante o desenvolvimento deste trabalho, incluindo os momentos de dificuldades, sempre disposto a ajudar e tirar dúvidas.

Aos professores do Departamento de Estatística por todo o conhecimento compartilhado durante as aulas do curso.

A todos os amigos e colegas que caminharam até aqui comigo.