Research, Society and Development, v. 11, n. 10, e203111032625, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i10.32625

Use of grouping techniques applied to the numbers of infected by Covid-19 in the Brazilian States comparing with the HDI of each State

Uso das técnicas de agrupamento aplicado aos números de infectados por Covid-19 nos Estados brasileiros comparando com o IDH de cada Estado

Uso de Técnicas de Agrupación Aplicadas al Número de Contagiados por Covid-19 en los Estados Brasileños en comparación con el IDH de cada Estado

Received: 07/07/2022 | Reviewed: 07/19/2022 | Accept: 07/21/2022 | Published: 07/27/2022

Mácio Augusto de Albuquerque
ORCID: https://orcid.org/0000-0002-0113-9130
Universidade Estadual da Paraíba, Brazil
E-mail: marcioaa@uepb.edu.br
Emanuela Rodrigues do Nascimento
ORCID: https://orcid.org/0000-0001-9750-734X
Universidade Estadual da Paraíba, Brazil
E-mail: nascimento.manu25@gmail.com
Cleanderson Romualdo Fidelis
ORCID: https://orcid.org/0000-0002-3726-5833
Universidade de São Paulo, Brazil
E-mail: cleanderson@usp.br

Abstract
This article aims to show how a cluster analysis can be carried out, through the technical hierarchy and the non-hierarchy in the rate of infected by Covid-19 of the Brazilian states through the numbers of infected states in order to identify a similarity between the states. through the numbers of infected, offering a counterpoint to the criterion used to analyze the number of infected in the states, based on the size of the population and comparing with their Human Development Index (HDI). public and free platforms called Coronavirus//Brasil and Atlas Brasil 2013 in relation to the 2010 HDI. For the cluster analysis, the Mahalanobins matrix was used with the hierarchical method, the simple link, complete, average, ward link and a non-hierarchical method using the K-means method were also applied, the coefficient of confenetic correlation to measure the degree of fit between the original similar matrices and the matrix resulting from the simplification provided by the clustering method. However, the method that best represents the data was found to be the complete linkage method. When grouping the states of both data, it took into account the similarity between the variables of the data and the correlation where it can be observed that the data are correlated.
Keywords: Methods; Brazilian states; Covid-19; Cluster.

Resumo
O presente artigo tem por objetivo mostrar como pode ser feita 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).Utilizou-se dados da Covid-19 retirado de uma plataformas públicas e gratuitas 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 confénetica 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: Métodos; Estados brasileiros; Covid-19; Cluster.

Resumen
Este artículo tiene como objetivo mostrar cómo se puede hacer el análisis de conglomerados, a través de la técnica jerárquica y no la jerarquía en la tasa de infectados por Covid-19 de los estados brasileños a través de los números de

infectados en cada estado, para identificar las similitudes entre los estados a través de las cifras de infectados, ofreciendo un contrapunto al criterio utilizado para analizar el número de infectados en los estados, con base en el tamaño de la población y comparando con su Índice de Desarrollo Humano (IDH), plataformas públicas y gratuitas denominadas Coronavirus //Brasil y Atlas Brasil 2013 en relación al IDH 2010. Para el análisis de conglomerados se utilizó la matriz de Mahalanobins con el método jerárquico, el simple, completo, promedio, ward binding y un método no jerárquico a través del método K-means, también se aplicó el coeficiente de correlación confenético para medir el grado de ajuste entre las matrices similares originales y la matriz resultante de la simplificación proporcionada por el método de agrupamiento. Sin embargo, se encontró que el método que mejor representa los datos es el método de enlace completo. Al agrupar los estados de ambos datos se tomó en cuenta la similitud entre las variables de los datos y la correlación donde se puede observar que los datos están correlacionados.

Palabras clave: Métodos; Estados brasileños; Covid-19; Grupo.

1. Introduction

In December 2019, the Chinese government announced in Wuhan (Hubei, China) a discovery of a new coronavirus, named SARS-CoV-2 (Covid-19), With this announcement, the World Health Organization (WHO) was alerted, declaring that the infection caused by this virus contaminated humans with high transmission potential (Alves, 2020).

WHO recommendations to slow down transmission, its main decision was social isolation, Covid-19 has spread throughout countries, in Brazil it reached 27 federative units, this fact occurred due to the challenges regarding the conditions of social vulnerability, housing and precarious sanitation, in addition to household overpopulation, however, due to the heterogeneity of the population in each of the states, the pandemic spread differently (Alves, 2020).

As the pandemic did not affect all people uniformly, affecting the most vulnerable population, that is, considering the poorest populations are more likely to have chronic conditions, this puts them at greater risk of virus-associated mortality. Since the pandemic can generate an economic crisis, if we take into account the unemployment rate that grew in the states (Dourado, 2021).

In this way, the HDI, an index that guides on factors that influence human development, can be a tool to assess this vulnerability, since factors such as lack of sanitary infrastructure impair preventive care for infection by the virus. On the other hand, states with the highest HDI have more indispensable conditions for an advanced economy (Domingo, 2021).

The Human Development Index (HDI) was created to assess the development of a country, state or municipality, not just economic growth. The HDI is measured from the geometric mean between indices that measure each of the following factors, considering some points in human development, such as having a long and healthy life, acquiring knowledge and having a decent standard of living (UNDP, 2020).

Cluster analysis is a multivariate technique (Fávero & Belfiore, 2019) with the objective of promoting the segmentation of data into categories or groups based on their homogeneous or heterogeneous characteristics, classifying them in the same or different groups. This technique groups data for interpretation using some methods that look for excluding groups, ascending to suppress the information of a set (Campos, 2019), there are several types of grouping techniques found in the literature according to (Albuquerque & Barros, 2020), of which the researcher has to decide which one is most suitable for his/her purpose, since the different techniques can lead to different solutions. As técnicas de agrupamento podem ser classificadas em hierárquicas e não-hierárquicas (Albuquerque et al., 2009).

The hierarchical technique consists of a series of successive groupings or successive divisions of elements, in which the elements are aggregated or disaggregated. The non-hierarchical technique was developed to group elements into K groups, where K is the number of groups defined previously.

A large number of similarity or dissimilarity measures have been proposed and used in cluster analysis, the choice between them being based on the preference and/or convenience of the researcher (Albuquerque & Barros, 2020).

With the definition of the dissimilarity measure to be used, the next step is the adoption of a clustering technique to form the groups. To carry out this task, there are a large number of methods available, from which the researcher has to decide which is the most suitable for his purpose, since the different techniques can lead to different solutions, in general, the Non-hierarchical method is to find the number "k" of clusters that can perform the division of observations in a satisfactory way, that is, that can identify similarities and differences between the observations.

Cluster analysis techniques require users to make a series of independent decisions, which require knowledge of the properties of the various algorithms available and that can represent different clusters. In addition, the result of the clusters can be influenced by the choice of technique used as well as by the dissimilarity measure, as well as by the definition of the number of groups.

In view of the above, the objective of this work is to show how the cluster analysis can be done, through the hierarchical technique and not hierarchy in the rate of infected by Covid-19 of the Brazilian states through the numbers of infected in each state, in order to identify the similarities between states through the numbers of infected, offering a counterpoint to the criterion used to analyze the number of infected in the states, based on the size of the population and comparing it with their Human Development Index (HDI).

2. Methodology

Data from Covid-19 taken on December 9, 2021 were used, this study focused on all Brazilian states. The data used come from public and free platforms called Coronavirus//Brazil, another data used was from the HDI (2010) of Brazilian states taken from Atlas Brasil 2013. To perform the cluster analysis, the hierarchy and non-hierarchy technique was used as a measure of dissimilarity to the Mahalanobis distance (D^2), with the most common clustering methods to determine the distance between clusters are: simple connection, complete connection, averages of distances and Ward method, and the cophenetic correlation, using the RStudio Software as a tool to carry out the entire analysis.

2.1 Coronavirus (Covid-19)

Covid-19 disease is a contagious disease caused by acute respiratory syndrome. The first case was identified in Wuhan, China, in December 2019. At that time, the disease spread around the world, leading to a pandemic. Transmission of Covid-19 occurs when people are exposed to respiratory droplets containing virus, that is, infected people can transmit the virus to another person up to two days before showing symptoms, as well as people who do not show symptoms as a result. The number of people infected by Covid-19 has increased rapidly in several parts of the world, including Brazil (Tizotte, 2021).

Preventive measures were social distancing, quarantine, ventilation of indoor spaces, covering coughs and sneezes, the use of face masks in public environments to minimize the risk of transmission (Who, 2020). Symptoms were diverse, ranging from mild to severe symptoms, so various measures were used to quantify mortality, these numbers vary by region.

With the arrival of Covid-19 in Brazil, the health authorities together with the Federal, State and Municipal bodies adopted several measures to control and prevent the disease for the Brazilian states, these measures differed from region to region, however the most announced measure by the authorities was the practice of social distancing (Bezerra, 2020). In some states, the isolation measures adopted by the population vary depending on the income, sex and education of the population, that is, the perception and behavior of Brazilians regarding the adoption of self-isolation and compliance with quarantine decrees varied from state to state. state, because even with the advance of the pandemic, part of the population began to have difficulties to remain isolated, even with an increasing number of cases.

2.2 Human Development Index (HDI)

The Human Development Index (HDI) was created in 1998 by two economists, Pakistani Mahbub Ul Haq and Indian Amartya, at the United Nations Development Program (UNDP). The HDI is considered an average to summarize the basic conditions of a population, centered on education, income, and quality of life. Published in Brazil for the first time in 1990, the HDI gradually became a reference in several parts of the world, and can be calculated through three main aspects: Income, Longevity and Education, with a variation between 1 and 0, the closer to 1 the more developed state is and the closer to 0 the less developed state is, that is, from these aspects it is possible to observe the improvements provided to the state through the HDI (Costa, 2019).

2.3 Cluster Analysis

The multivariate technique of cluster analysis allows an analysis of interdependence between the variables, causing them to be aggregated based on their common characteristics. According to Albuquerque & Barros (2020), for the study of cluster analysis, use the dissimilarity method based on the Mahalanobis distance (D^2), considered one of the most used distances, which can be calculated according to the following expression:

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

on what: D^2 has the characteristic of being invariant for any non-singular linear transformation, X_i is the vector that belongs to the parcel *i*, X_j will be a vector that belongs to the parcel *j*, $\sum^{1}$ is the inverse of the residual covariance matrix of *X* and $(X_i - X_j)'$ is the transposed vector of the difference between X_i and X_j.

2. 4 Hierarchical Method

In the hierarchical method, the focus is not on the exact number of clusters, but on the cluster to be analyzed, whose construction is based on a larger cluster and dividing the observations into smaller clusters, or one of each. the observation is a conglomerate and will be grouped into larger groups in the following steps, with the criteria for these groupings varying according to the technique (Duarte, 2021).

Because it is a widely used technique and easy to find in some computer programs, the algorithm techniques used according to Costa (2019) are: Simple connection method, which is defined by the two elements most similar to each other; Complete, which is defined as the distance between the vectors of means; Average treats the distance between two clusters as the average of the distances between all pairs of elements that can be formed with the elements of the two clusters being compared and the Ward Linkage Method that can form the clusters by maximizing the homogeneity within of the groups or the total minimization of the sum of squares within the groups.

a) **Simple linkage method:** As it is one of the oldest and simplest algorithms used in the literature, called “nearest neighbor method”, this is an agglomerative hierarchy technique and has, as one of its characteristics, not requiring the number of clusters is fixed a priori. According to Souza (2022) the distance between two groups is determined by the minimum distance between the pairs of elements of these groups and the one with the smallest minimum distance is grouped, that is, if two groups $C_1 = \{X_1, X_3, X_7\}$ e $C_2 = \{X_2, X_6\}$, the distance between the groups is defined by:

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

b) **Full connection method:** In this method, the distance between two groups is determined by the maximum distance between pairs of elements in these groups. The method tries to group the elements that have the smallest distance among the most distant ones. be two groups $C_1 = \{X_1, X_3, X_7\}$ e $C_2 = \{X_2, X_6\}$ the distance between the groups is defined by:

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

c) **Average connection method:** This method was originally proposed by Sokal and Michener (1958) and is a balance between single-link and full-link methods between all pairs found. It can be formed with the elements of the two groups to be compared and groups those with the smallest average distance (Souza, 2022). For example, if groups C_1 have n_1 elements and C_2 has n_2 elements, the distance between the groups is given by:

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

d) **Ward connection method:** According to Souza (2022) the Ward method groups the elements that have the smallest sum of the squares of the distances, it is a method that tends to provide aggregates with approximately the same number of observations, initially each element is considered a single grouping and at each step of the algorithm, the algorithm calculates the sum of squares within each cluster of each element belonging to the cluster, with respect to the corresponding mean vector of the cluster. The distance between C_l and C_i represents the square sum between the clusters that can be defined by:

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

2.5 Non-Hierarchical Method

The non-hierarchical methods in general, the focus is to find the number “k” of clusters that can perform the division of observations in a satisfactory way, the method that can identify similarities and differences between the observations (Duarte, 2021). As it is a more dynamic and interactive grouping process, the non-hierarchical method, the number of groups is specified before the grouping process, the criterion, but used by this method is that of K-means, according to Alves (2020) this method has some conditions such as the previous information of the numbers of clusters k, where their observations are grouped in these k clusters using a function with objective and criterion. Being of simple application and fast processing according to Duarte (2021), the algorithm logic follows 4 steps:

1. Chosen the number of groups, called k
2. Within the data, some random observation is assigned to a cluster, then using some distance measure, the closest element is assigned the same cluster and the average of the distances is calculated, forming the center of the cluster.
3. Recalculate the centroids for each K clusters, calculate the average of all elements in the groups.
4. Step 2 is repeated, and the center of the cluster is recalculated given the new object that has entered, this is repeated until all data have their respective clusters.

2.6 Cophenetic correlation

To measure the degree of fit between the original similar matrices and the matrix resulting from the simplification, the cophenetic correlation will be used, provided by the grouping method according to the expression according to Albuquerque, et al. (2016):

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

Where: C_{ij} is the similarity value between individuals *i* and *j*, where will be obtained from the cophenetic matrix; S_{ij} is the similarity value between individuals *i* and *j*, where they will be obtained from the similarity matrix.

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

It is observed that this correlation corresponds to Pearson's correlation between the original similarity matrix and the one obtained after the construction of the dendrogram, that is, it uses a scale from 0 to 1, if $c \leq 0,39$ is considered weak, if $0,40 \geq c \leq 0,69$ is considered moderate and if $c \leq 0,70$ is considered strong, so the closer to 1, the smaller the dendrogram distortion caused by grouping individuals with some chosen hierarchical method.

3. Results and Discussion

After using different grouping methods and analyzing their figures, it was verified that the constructed groups differ from each other, from the hierarchical method and application of the Mahalanobis distance matrix, the following hierarchical agglomerative methods were applied: nearest neighbor, neighbor furthest distances, averages of distances and Ward. In a later analysis, the non-hierarchical method was also evaluated, each method has its advantages and disadvantages, the hierarchical method has the advantage of using several different measures, its disadvantage is to reduce the number of outliers. The advantage of non-hierarchical is to use a very large dataset with fewer outliers, but the disadvantage is to randomly use the centroid, which makes the hierarchical method superior to this method.

Analyzing the Mahalanosbis distance matrix by the linkage methods, a small change in the levels of the grouped elements was observed, the elements located within each group their structure is usually quite similar in relation to each method used.

1) Analysis of COVID-19 data

Using data from the 27 states of Brazil regarding Covid-19 data and the available variables are Confirmed Cases, Deaths, Incidence and Mortality, to characterize the variables under study, a descriptive analysis was carried out in Table 1, and it was possible to observe that the variable with the highest average was for confirmed cases and the lowest average was for Mortality. Then, a study of distance measurement and clustering method used for the construction of the dendrogram (Figures 1 to 4) was carried out with the variable Confirmed cases with the combination of the Mahalanobis (D^2) distance and clustering methods (Single Link, Full Link, Media Link, Ward Link) and Euclidean distance with the K-means method.

Table 1: Descriptive analysis of Covid-19 data variables.

Variables	Mean	Standard Deviation	Min	Max
cases	821660	898479,3	88264	4449552
Deaths	22842	310776,68	1849	154691
Incidence	11465	4155,891	1727	21254
Mortality	282,8	72,859	146,7	401,1

Source: Authors.

With this combination of distance from the Mahalanobis distance (D^2) and clustering methods, the cophenetic correlation coefficient (CCC) was obtained in order to measure the degree of fit between the matrices formed (Table 2), that is, the Mahalanobis distance (D^2) and complete binding methods obtained the highest value for the CCC which was equal to 0.850 as the value being close to 1 the CCC is considered strong.

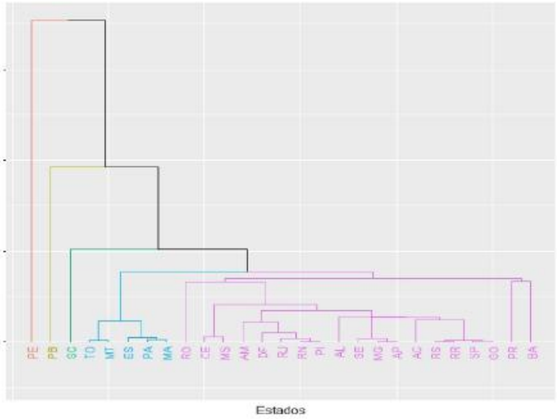
Table 2: Cophenetic correlation coefficient obtained for the Covid-19 data.

links	Correlation
Simple	0,742
complete	0,850
Average	0,814
Ward	0,570

Source: Authors.

In the Simple connection method denoted as "nearest neighbor method", the 5 clusters formed by this method can be observed in the dendrogram (Figure 1). In group 1, 2 and 3 were formed by only one state Pernambuco, Paraíba and Santa Catarina, that is, one state for each group. In relation to group 4, formed by 5 states (Tocantins, Mato Grosso, Espírito Santo, Pará and Maranhão), and group 5 was the one that had the largest number of states possible with 19 (Rondônia, Ceará, Mato Grosso do Sul , Amazonas, Distrito Federal, Rio de Janeiro, Rio Grande do Norte, Piauí, Alagoas, Sergipe, Minas Gerais, Amapá, Acre, Rio Grande do Sul, Roraima, São Paulo, Goiás, Paraná and Bahia), in this group we can see that in it is the state with the highest number of confirmed cases is São Paulo with 4 449 552 inhabitants and the state with the lowest number of confirmed cases is Amapá = 125 336 inhabitants, that is, the greater the population of the state, the greater the number of confirmed cases.

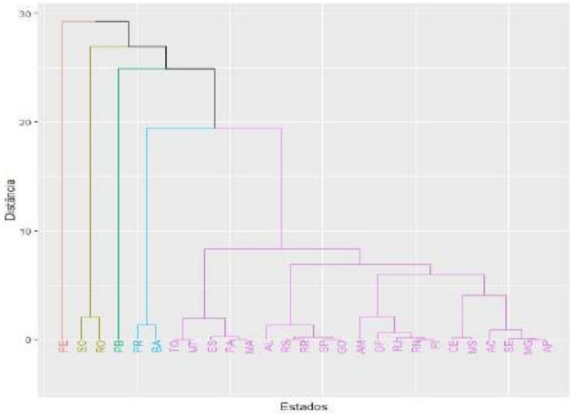
Figure 1: Dendrogram obtained through the variable Confirmed Cases based on the mahalanobis distance and the Simple linkage method.



Source: Authors.

With the Complete linkage method that grouping the elements with the smallest distance between the most distant can be observed in the dendrogram (Figure 2), the 5 clusters formed, in clusters 1 and 3 were formed by only one state Pernambuco and Paraíba, that is, a state for each cluster. In relation to cluster 2 and 4, formed by two states each, cluster 2 (Santa Catarina and Rondônia), cluster 4 (Paraná and Bahia), and cluster 5 again obtained the highest number of states (Tocantins, Mato Grosso, Espírito Santo). Santo, Pará, Maranhão, Ceará, Mato Grosso do Sul, Amazon, Federal District, Rio de Janeiro, Rio Grande do Norte, Piauí, Alagoas, Sergipe, Minas Gerais, Amapá, Acre, Rio Grande do Sul, Roraima, São Paulo and Goiás), in this method, what caught the attention was cluster 4, because the states that compose it have a close rate of confirmed cases, even though they are from different regions, Bahia located in the Northeast region with an estimated population of 14.985,284 inhabitants and Paraná with 11,597,484 inhabitants located in the southern region of Brazil.

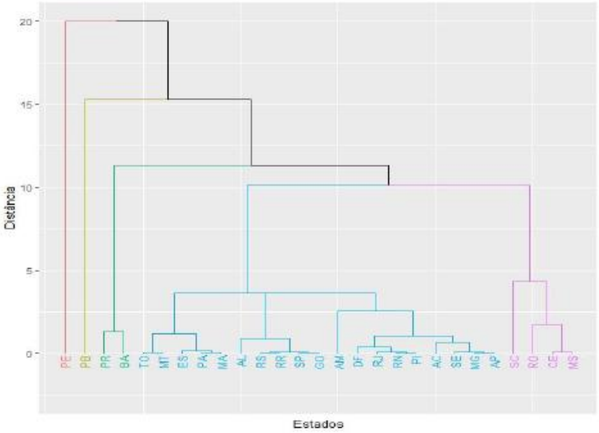
Figure 2: Dendrogram obtained using the variable Confirmed Cases based on the mahalanobis distance and the Complete linkage method.



Source: Authors.

Regarding the average linkage method of the 5 clusters formed, it can be seen in the dendrogram (Figure 3), cluster 1 and 2 remained with only one state (Pernambuco and Paraíba), cluster 2 with Paraná and Bahia, cluster 3 was the which obtained the highest concentration of states (Tocantins, Mato Grosso, Espírito Santo, Pará, Maranhão, Alagoas, Rio Grande do Sul, Roraima, São Paulo, Goiás, Amazonas, Distrito Federal, Rio de Janeiro, Rio Grande do Norte, Piauí, Acre, Sergipe, Mina Gerais and Amapá) and cluster 5 with only four states (Santa Catarina, Rondônia, Ceara and Mato Grosso do Sul), in this method groups those with the lowest average distance, that is, we can highlight cluster 5 that was formed with states from different regions and the one with the highest number of confirmed cases is Santa Catarina and the one with the lowest number of confirmed cases was Rondônia.

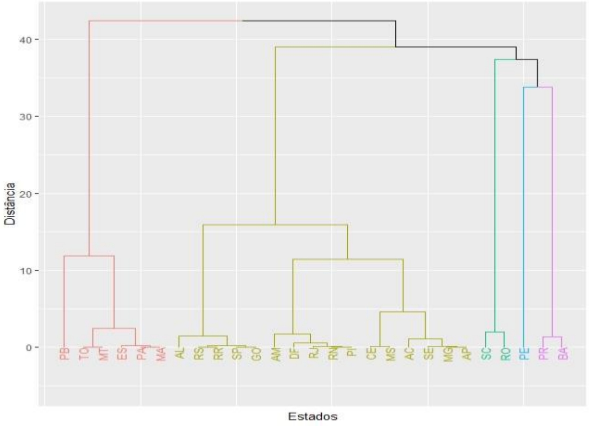
Figure 3: Dendrogram obtained through the variable Confirmed Cases based on the mahalanobis distance and the Mean linkage method.



Source: Authors.

The Ward method of the 5 clusters formed can be observed in the dendrogram (Figure 4), cluster 1 formed by six states (Paraíba, Tocantins, Mato Grosso, Espírito Santo, Pará and Maranhão) in cluster 2 with the presence of sixteen states (Alagoas, Rio Grande do Sul, Roraima, São Paulo, Goiás, Amazon, Federal District, Rio de Janeiro, Rio Grande do Norte, Piauí, Ceara, Mato Grosso do Sul, Acre, Sergipe, Minas Gerais and Amapá), the cluster 3 composed of two states (Santa Catarina and Rondônia), cluster 4 is unitary with only one state (Pernambuco) and cluster 5 with two states (Pará and Bahia) in this method groups the states that have the smallest sum of the squares of the distances Cluster 4 with the state of Pernambuco can be highlighted in it, which has a number of confirmed cases equal to 643,307 and remained isolated in all other methods used.

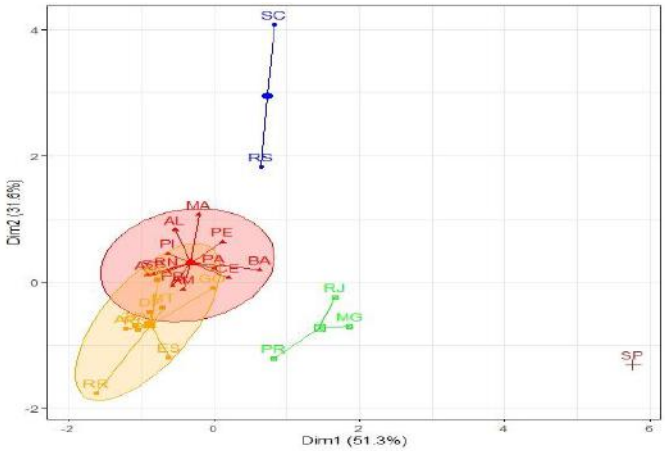
Figure 4: Dendrogram obtained using the variable Confirmed Cases based on the mahalanobis distance and the Ward linkage method.



Source: Authors.

When analyzing the 5 clusters of the Euclidean distance with the K-means method (Figure 5), one can observe in cluster 1 formed by two states (RS and SC), cluster 2 by twelve states (PA, AM, AC, BA, PB, PE, CE, RN, MA, AL, SE and PI), cluster 3 with nine states (ES, RR, AP, TO, RO, MT, GO, MS and DF), in cluster 4 only one state (SP) and cluster 5 with three states (PR, MG and RJ), in this method it is possible to verify the characteristics of each agglomeration. Based on the average K-means of each cluster, it was observed that the state with the most confirmed cases of Covid-19 on average is in cluster 4 along with the highest average number of deaths, whereas cluster 1 has the number mean higher mortality, cluster 2 has the lowest mean in all variables and cluster 3 has a higher mean in relation to the other clusters.

Figure 5: Graph obtained using the K-maens method for Covid-19 data.



Source: Authors.

2) **Analysis of HDI data**

The data used will refer to the Human Development Index of the 27 Brazilian states and the available variables are HDI, HDI-R, HDI-L and HDI-E, for a better understanding of the variables under study, a descriptive analysis was performed (Table 3) being able to observe that the variable with the highest average was the HDI-L and the lowest average was the HDI-E. A study of distance measurement and clustering method used for the construction of the dendrogram (Figures 6 to 9) was also carried out with the HDI variable using the combination of the Mahalanobis distance (D^2) with the clustering methods and the Euclidean distance with the K-means method.

Table 3: Descriptive analysis of HDI data variables.

Variables	Mean	Standard Deviation	Min	Max
HDI	0,7045	0,0492	0,6310	0, 8240
HDI-R	0,7069	0,0582	0,6120	0,8630
HDI-L	0,8086	0.0304	0,7550	0,8730
HDI-E	0,6124	0,0582	0,5200	0,7420

Source: Authors.

This combination of the Mahalanobis distance (D^2) and clustering methods (Single Link, Full Link, Media Link, Ward and K-means), the cophenetic correlation coefficient (CCC) was obtained in order to measure the degree of fit between the formed matrices (Table 4), that is, according to Nascimento(2022) the Mahalanobis distance (D^2) and complete linkage methods obtained the highest value for the CCC which was equal to 0.728, as this value is close to 1 the CCC is considering strong, so the states that are within the same group can be grouped in other ways when changing the method.

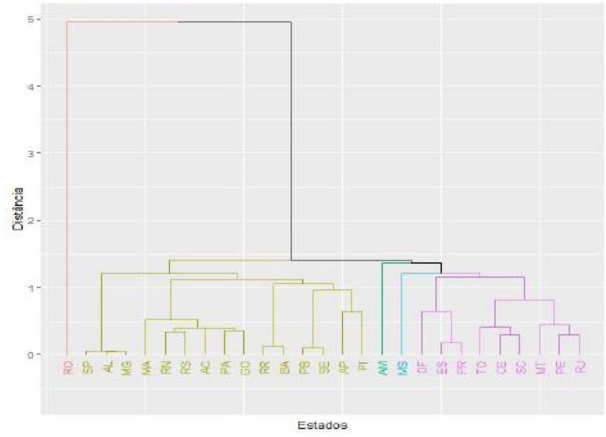
Tabela 4: Coeficiente de correlação cofenético obtidos para os dados do IDH.

links	Correlation
Simple	0,676
complete	0,728
Average	0,669
Ward	0,474

Source: Authors.

Composed of five clusters, the simple linkage method can be observed in the dendrogram (Figure 6), it presents three unitary clusters cluster 1 with Rondônia, cluster 3 with Amazonia and cluster 4 with Mato Grosso do Sul and two groups with a greater number of states from different states. regions of Brazil (Figure 6), that is, cluster 2 (Acre, Roraima, Pará, Amapá, Maranhão, Piauí, Rio Grande do Norte, Paraíba, Alagoas, Sergipe, Bahia, Minas Gerais, São Paulo, Rio Grande do Sul and Goiás) and cluster 5 (Tocantins, Ceará, Pernambuco, Espírito Santo, Rio de Janeiro, Paraná, Santa Catarina, Mato Grosso and Federal District). This method shows that there is a variation of states in cluster 2, regardless of the HDI of the states.

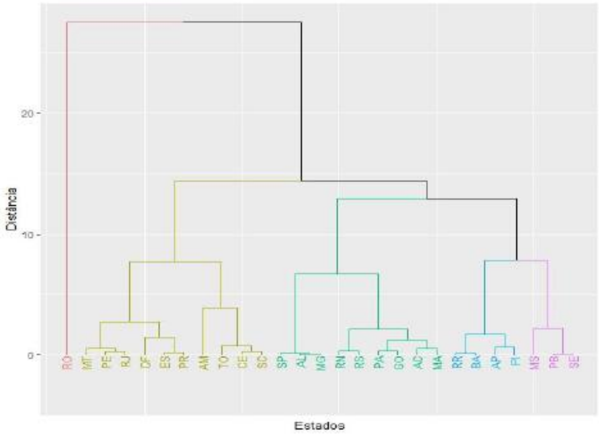
Figure 6: Dendrogram obtained using the HDI variable based on the mahalanobis distance and the Simple linkage method.



Source: Authors.

With the complete linkage method (Figure 7) you can observe the presence of five clusters, cluster 1 with one state (Rondônia), cluster 2 with ten states (Amazona, Tocantins, Ceará, Pernambuco, Espírito Santo, Rio de Janeiro , Santa Catarina, Paraná, Mato Grosso and Distrito Federal), cluster 4 with nine states (Acre, Pará, Maranhão, Rio Grande do Norte, Alagoas, Minas Gerais, São Paulo, Rio Grande do Sul and Goiás), cluster 4 with four states (Roraima, Amapá, Piauí and Bahia) and cluster 5 with three states (Paraíba, Mato Grosso do Sul and Sergipe) because the states that compose it have a similar HDI, even though they are from different regions.

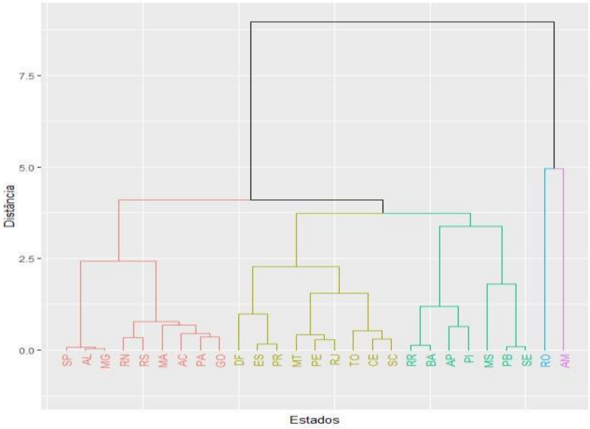
Figure 7: Dendrogram obtained using the HDI variable based on the mahalanobis distance and the Complete linkage method.



Source: Authors.

The average linkage method of the 5 clusters formed to be observed in the dendrogram (Figure 8), cluster 1 has nine states (São Paulo, Alagoas, Minas Gerais, Rio Grande do Norte, Rio Grande do Sul, Acre, Maranhão, Pará, Goiás), cluster 2 with nine states (Federal District, Espírito Santo, Tocantins, Paraná, Mato Grosso, Pernambuco, Rio de Janeiro, Ceará and Santa Catarina,) cluster 3 had seven states (Roraima, Amapá, Piauí, Paraíba, Sergipe, Bahia, Mato Grosso do Sul) and cluster 4 and 5 with only one states (Rondônia and Amazona), in this method groups those with the lowest average distance, that is, we can highlight cluster 1 and 2 formed with states from regions and the one with the highest HDI index is São Paulo and the one with the lowest HDI index was Alagoas.

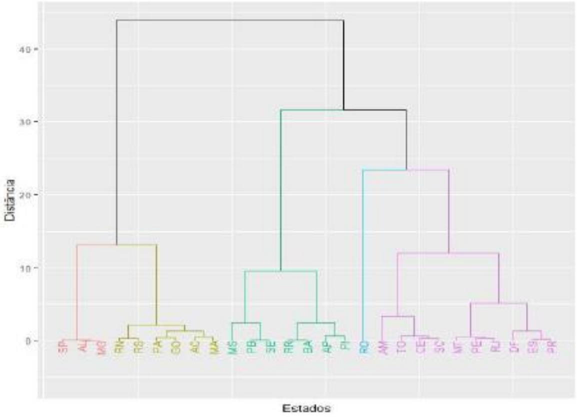
Figure 8: Dendrogram obtained using the HDI variable based on the mahalanobis distance and the Mean linkage method.



Source: Authors.

Ward's method can be observed in the dendrogram (Figure 9) of the 5 clusters formed, cluster 1 formed by three states (São Paulo, Alagoas and Minas Gerais) in cluster 2 with the presence of six states (Rio Grande do Norte, Rio grande do Sul, Pará, Goiás, Acre and Maranhão) in cluster 3 with seven states (Mato Grosso do Sul, Paraíba, Sergipe, Roraima, Bahia, Amapá, Piauí), in cluster 4 with one state in Rondônia and cluster 5 the largest with ten states (Amazona, Tocantins, Ceará, Santa Catarina, Mato Grosso, Pernambuco, Rio de Janeiro, Distrito Federal, Espírito Santo and Paraná), in this method cluster 4 with the state of Rondônia stands out because it remained isolated .

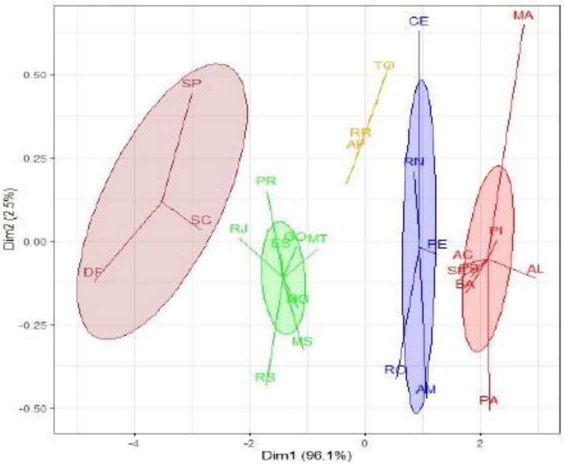
Figure 9: Dendrogram obtained using the HDI variable based on the mahalanobis distance and the Ward binding method.



Source: Authors.

Analyzing the Euclidean distance with the K-means method. (Figure 10) and the 5 clusters, cluster 1 formed by five states (RO, RN, CE, AM and PE), cluster 2 by eight states (SE , AC, BA, PB, PI, MA and AL), cluster 3 with three states (AP, RR, and TO), cluster 4 also with three states (DF, SP and SC) and cluster 5 with eight states (RJ, PR, RS, ES, GO, MG, MS and MT), in this method it is possible to verify the characteristics of each agglomeration. Based on the average of the K-means of each cluster, it was observed that the state with the highest average HDI is found in clusters 4 and 5, together with the highest average also of HDI-L, HDI-R and HDI-E, since cluster 1 has the lowest mean in all variables.

Figure 10: Graph obtained using the K-maens method for HDI data.

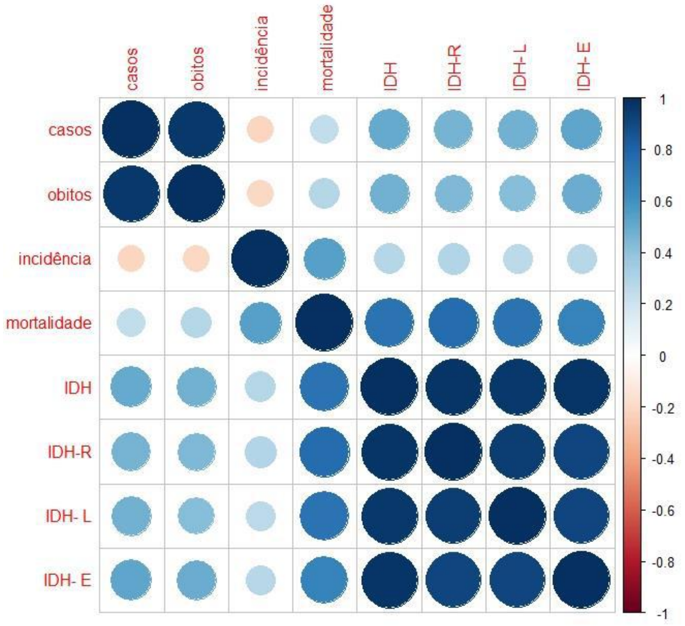


Source: Authors.

Correlation Analysis between the data

Correlation coefficients are statistical methods to measure the relationships between variables and what they represent, that is, they seek to understand how a variable behaves in a scenario where another is varying, in order to identify if there is any relationship between the variability of both, will be applied to data from the 27 states of Brazil referring to COVID-19 data and its variables (Confirmed cases, Deaths, Incidence and Mortality) and data referring to the Human Development Index with the variables (HDI, HDI-R, HDI- L and HDI-E. You can see in Figure 11 the variables that have a strong and positive correlation is mortality and HDI with 0.730, in relation to HDI and Cases they had a positive correlation of 0.50, that is, there is a correlation positive among the data.

Figure 11: Graph referring to the Correlation between Covdi-19 data and the HDI.



Source: Authors.

4. Conclusion

In this study we can conclude that the method that best represented both data was the complete linkage according to the cophenetic correlation coefficient (CCC), the state of Pernambuco remained in an isolated cluster in all methods for the Covid-19 data. and the state of Rondônia for HDI data. There is a positive correlation between the variables of the Covid-19 data in relation to the HDI, that is, the mortality variable had the highest correlation with HDI, HDI-R, HDI-L and HDI-E, therefore, to a correlation between the Dice.

Acknowledgments

Acknowledgments to the Institutional Scientific Initiation Scholarship Program (PIBIC) UEPB that provided an initiation scholarship for this work.

References

Albuquerque, M. A, de Oliveira Barros, K. N. N., Gouveia, J. F., & Ferreira, R. L. C. (2016). < b> Determinação e validação de números de grupo em uma análise de cluster: Um estudo de caso aplicado à ciência florestal. *Acta Scientiarum. Tecnologia*, *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]. *EDUEPB. http://eduepb. uepb. edu. br/download/introducao-a-analise-de-agrupamento-teoria-e-pratica-com-aplicacoes-em-r*.

Albuquerque, M. A., & de Oliveira Barros, K. N. N. (2020). Determinação do número de grupos em análise de agrupamento via de raio de influência. *Brazilian Journal of Development*, *6*(6), 38342-38355.

Bezerra, A. C. V., Silva, C. E. M. D., Soares, F. R. G., & Silva, J. A. M. D. (2020). Fatores associados ao comportamento da população durante o isolamento social na pandemia de COVID-19. *Ciência & Saúde Coletiva*, *25*(suppl 1), 2411-2421.

Costa, G. D. (2019). Análise multivariada de países da América do Sul por meio de Indicadores socioeconômicos.

Dourado. P. B. M. (2021). *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, S. R. N. (2021). *Um guia para agrupamento com pacote cluster do R utilizando dados do Spotify* (Bachelor's thesis, Universidade Federal do Rio Grande do Norte).

Fávero, L. P, & Belfiore, P. (2019). *Ciência de dados para negócios e tomada de decisões*. Imprensa Acadêmica.

Nascimento, ER, de Albuquerque, MA, de Oliveira Barros, KNN, & Barros, PSN (2022). Análise de cluster aplicada ao Índice de Desenvolvimento Humano (IDH) dos estados brasileiros. *Pesquisa, Sociedade e Desenvolvimento*, *11* (2), e18011225747-e18011225747.

Neto, D. M. F., Morbeck, N. B. M., Welter, Á., & Panontin, J. F. (2021). Relação entre índice de desenvolvimento humano e número de casos de covid-19 em cidades do tocantins. *Singular. Saúde e Biológicas*, *1*(2), 23-27.

Olivatto, T. F., & Lollo, J. A. D. (2022). Urban Sustainable Development Index: a geospatial approach to add Tree Cover to HDI in São Paulo City. *Sociedade & Natureza*, *34*.

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

Paula Alves, H. J, Fernandes, F. A, de Lima, K. P, de Oliveira Batista, B. D, & Fernandes, T. J. (2020). A pandemia da COVID-19 no Brasil: uma aplicação do método de clusterização k-means. *Pesquisa, Sociedade e Desenvolvimento*, *9* (10), e5829109059-e5829109059.

Silva Campos, S. L. (2019). Busca nao supervisionada de padroes por técnicas de agrupamento clássica e nebulosa.

Souza, M. V. V. D. (2022). Análise multivariada de países da América e Europa utilizando indicadores sobre a Covid-19 e dieta da população.

Tizotte, T. R. L., Thesing, N. J., & Gomes, F. B. M. (2021). Análise bibliométrica dos artigos da base de dados da Scopus sobre a Produção Científica Brasileira da Covid-19 Bibliometric analysis of articles from the Scopus Database on the Brazilian Scientific Production of Covid-19. *Brazilian Journal of Development*, *7*(7), 73457-73474.