Research, Society and Development, v. 11, n. 2, e18011225747, 2022
(CC BY 4.0) | ISSN 2525-3409 | DOI: http://dx.doi.org/10.33448/rsd-v11i2.25747

Cluster analysis applied to the Human Development Index (HDI) of Brazilian States

Análise de agrupamento aplicado no Índice de Desenvolvimento Humano (IDH) dos Estados Brasileiros

Análisis de conglomerados aplicado al Índice de Desarrollo Humano (IDH) de los Estados Brasileños

Received: 01/14/2022 | Reviewed: 01/18/2022 | Accept: 01/20/2022 | Published: 01/24/2022

Emanuela Rodrigues do Nascimento
ORCID: https://orcid.org/0000-0001-9750-734X
Universidade Estadual da Paraíba, Brazil
E-mail: nascimento.manu25@gmail.com
Mácio Augusto de Albuquerque
ORCID: https://orcid.org/0000-0002-0113-9130
Universidade Estadual da Paraíba, Brazil
E-mail: marcioaa@uepb.edu.br
Kleber Napoleão Nunes de Oliveira Barros
ORCID: https://orcid.org/0000-0003-2515-3292
Universidade Federal Rural de Pernambuco, Brazil
E-mail: kleber.barros@ufrpe.br
Patrícia Silva Nascimento Barros
ORCID: https://orcid.org/0000-0003-0681-2029
Universidade Federal da Paraíba, Brazil
E-mail: patricia@dcx.ufpb.br

Abstract
This study aims to compare the performance of each method (hierarchical and non-hierarchical) of the grouping formed by several HDI from the 27 brazilian states, through the cluster analysis technique. As well as determining how many states there are in each formed group, to thus specify which technique best represents the data. Data from Atlas Brasil 2013 were used in relation to the 2010 HDI. For cluster analysis, the Mahalanobins matrix was used with the hierarchical method, from the data obtained, we applied the simple linkage methods, complete, average, ward liaison and a non-hierarchical method through the K-means method, the conphenetic correlation coefficient was also applied to measure the degree of fit between the original similar matrices and the resulting matrix of simplification provided by the grouping method. However, the method that best represents the data was the complete link. When grouping the states, the similarity between the HDI-R, HDI-L and HDI-S variables was considered this relationship formed similar groups between the connections from different regions of Brazil.
Keywords: Mahalanobins; Methods; Brazilian States; Cluster.

Resumo
O presente artigo tem por objetivo compara o desempenho de cada método (hierárquico e não hierárquico) de agrupamento formado por vários IDH dos 27 estados brasileiros, por meio da técnica de análise de agrupamento. Bem como determina quantos estados tem em cada grupo formado, para assim especificar qual técnica melhor representa os dados. Utilizou-se dados do Atlas Brasil 2013 com relação ao IDH de 2010. Para a análise de agrupamento foi utilizado a matriz de Mahalanobins com o método hierárquico, a partir dos dados obtidos, 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 foi levado em consideração a semelhança entre as variáveis IDH-R, IDH-L e IDH-S está relação formou grupos semelhantes entre as ligações de diferentes regiões do Brasil.
Palavras-chave: Mahalanobins; Métodos; Estados brasileiros; Cluster.

Resumen
El objetivo de este artículo es comparar el desempeño de cada método (jerárquico y no jerárquico) de agrupamiento formado por varios IDH de los 27 estados brasileños, a través de la técnica de análisis de conglomerados. También determina cuántos estados hay en cada grupo formado, con el fin de especificar qué técnica representa mejor los datos. Se utilizaron datos del Atlas Brasil 2013 en relación al IDH de 2010. Para el análisis de conglomerados se utilizó la

matriz de Mahalanobins con el método jerárquico, a partir de los datos obtenidos, los métodos de simple, completo, promedio, ward binding y no -método 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, el método que mejor representa los datos es el método de enlace completo. Al agrupar los estados, se tuvo en cuenta la similitud entre las variables IDH-R, IDH-L e IDH-S, relación que formó grupos similares entre conexiones de diferentes regiones de Brasil.
Palabras clave: Mahalanobinas; Métodos; Estados Brasileños; Grupo.

1. Introduction

The Human Development Index (HDI) was created in 1998 by two economists, the Pakistani Mahbub Ul Haq and the Indian Amartya, at the United Nations Development Program (UNDP). The HDI is considered an average for summarizing the basic conditions of a population, focusing on education, income, and quality of life. Published in Brazil for the first time in the year 1990, the HDI has gradually become a reference in several places around the world (Braga, 2017).

The Human Development Index in Brazil can be consulted through the Atlas platform that refers to human development, covering the Atlas of the 27 States, the Municipal Human Development Index (MDI) and Metropolitan Regions. Demonstrating more 200 indicators of demography, education, income, work, housing, and vulnerability.

According to data from the 2013 atlas, the Brazilian states present a considerable discrepancy in values obtained by the HDI with a range of 0.631 to 0.824 within a probable range, and can assume values between 0 and 1, the closer the HDI of a country is to 1, the more developed it becomes. Therefore this difference can be partly explained by the varieties of characteristic that distinguish one state (or country) from another in its geographical, economic, and infrastructure aspects (Costa, 2019).

Cluster analysis, also known as cluster analysis is a multivariate technique with the goal of promoting the segmentation of data into categories or groups based on their homogeneous or heterogeneous characteristics by classifying into the same or distinct groups. This technique groups data for interpretation using some methods that look for excluding, ascending groups to thus repress the information of a set (Campos, 2019), when we compare states through this technique we can form new groups which can be significantly smaller when compared to the set of states provided. Because it is a widely used technique it can generate groups of states with similar characteristics, promoting a broad view regarding the state with similar HDI.

Based on the cluster analysis, the hierarchical and non-hierarchical methods can be used. The hierarchical method aims to form a hierarchical decomposition of the data sets, forming a structure of a hierarchical tree, already the non-hierarchical method the methodology, but used is the k-th (Nascimento, 2019).

Other important methods that can be used are the linking methods such as: simple, complete, medium and ward. The simple linkage method succeeds between two very similar elements; the complete linkage method occurs contrary to the simple method; the average linkage method uses the arithmetic mean of the dissimilarity measures which treats the distance between two conglomerates as the average of the distances between all pairs of elements that were formed with the elements of the two conglomerates being compared; the ward linkage has a different method of forming its group from maximizing the homogeneity between the groups or however the total minimization of the sum of squares between the groups (Costa, 2019).

Thus, this study aims to compare the performance of each method (hierarchical and non-hierarchical) of grouping formed by various HDI of the 27 Brazilian states through the cluster analysis technique. As well as determine how many states have in each group formed, to thus specify which technique best represents the data (Barroso & Artes (2003).

2. Materials and Method

Data regarding the HDI 2010, taken from Atlas Brazil 2013 (Table 1), were used, based on the Brazilian states. These data were calculated through three main aspects: Income, Longevity, and Education, and can vary between 1 and 0. The closer to 1, the more developed the state is, and the closer to 0, the less developed the state is.

Table 1: Data regarding the Human Development Index (HDI), 2013.

Posição	Estados	IDH	Renda	Longevidade	Educação
1 º	Distrito Federal	0.824	0.863	0.873	0.742
2 º	São Paulo	0.783	0.789	0.845	0.719
3 º	Santa Catarina	0.774	0.773	0.860	0.697
4 º	Rio de Janeiro	0.761	0.782	0.835	0.675
5 º	Paraná	0.749	0.757	0.830	0.668
6 º	Rio Grande do Sul	0.746	0.769	0.840	0.642
7 º	Espírito Santo	0.740	0.743	0.835	0.653
8 º	Goiás	0.735	0.742	0.827	0.646
9 º	Minas Gerais	0.731	0.730	0.838	0.638
10 º	Mato Grosso do Sul	0.729	0.740	0.833	0.629
11 º	Mato Grosso	0.725	0.732	0.821	0.635
12 º	Amapá	0.708	0.694	0.813	0.629
13 º	Roraima	0.707	0.695	0.809	0.628
14 º	Tocantins	0.699	0.690	0.793	0.624
15 º	Rondônia	0.690	0.712	0.800	0.577
16 º	Rio Grande do Norte	0.684	0.678	0.792	0.597
17 º	Ceará	0.682	0.651	0.793	0.615
18 º	Amazonas	0.674	0.677	0.805	0.561
19 º	Pernambuco	0.673	0.673	0.789	0.574
20 º	Sergipe	0.665	0.672	0.781	0.560
21 º	Acre	0.663	0.671	0.777	0.559
22 º	Bahia	0.660	0.663	0.783	0.555
23 º	Paraíba	0.658	0.656	0.783	0.555
24 º	Pará	0.646	0.646	0.789	0.528
24 º	Piauí	0.646	0.635	0.777	0.547
26 º	Maranhão	0.639	0.612	0.757	0.562
27 º	Alagoas	0.631	0.641	0.755	0.520

Source: Atlas Brasil (2019).

For the cluster analysis study we used the dissimilarity method based on Mahalanobis distance (D2), which is considered one of the most used distances ((Johnson & Wichern, 2002; Albuquerque & Barros, 2020)), and can be calculated according to the following expression:

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

where:

D2 exhibits characteristic of being invariant for any non-singular linear transformation,

Xi is the vector that belongs to plot i;

Xj is a vector that belongs to plot j;

∑1 is the inverse of the residual covariance matrix of X;

(Xi - Xj)'is the transposed vector of the difference between Xi and Xj.

Because it is a technique widely used in practice and easy to be found in some computer programs, the clustering algorithms used were: Simple linkage method which is defined by the two elements that are most similar to each other; Complete which is defined as the distance between the vectors of means; Average treats the distance between two conglomerates as the average of the distances between all pairs of elements that can be formed with the elements of the two conglomerates being compared and the Ward linkage method which can form the groups from the maximization of the homogeneity within the groups or the total minimization of the sum of squares within the groups (Costa 2019). It will be represented in the form of dendrograms.

According to Albuquerque, et al. (2016), the cohenetic correlation was used to measure the degree of fit between the original similar matrices and the matrix resulting from the simplification provided by the clustering method according to the expression:

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

where: Cij is the similarity value between individuals i and j, where will be obtained from the cohenetic matrix; Sij is the similarity value between individuals i and j, where will be obtained from the similarity matrix. Where:

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

It can be classified as: c = 0.10 to 0.39 (weak), c = 0.40 to 0.69 (moderate), and c = 0.70 to 1 (strong). The closer the correlation is to 1, the smaller the change caused in the dendrogram formed by grouping the elements with some chosen hierarchical method Oliveira, et al. (2016).

3. Results and Discussion

In the cluster analysis used, a difference was observed between the methods applied, both for the hierarchical method and for the non-hierarchical method, each method showing its advantage and disadvantage. The hierarchical method has the advantage of using different dissimilar measures, its disadvantage is to reduce the number of outliers. The non-hierarchical method has the advantage of using a very large data set, with a smaller presence of outliers, but the disadvantage is to use the centroid randomly, making the hierarchical method superior to this method.Ao analisar a matriz da distância de mahalanosbis com a aplicação dos métodos de ligação simples, ligação completa, média da distância e de ward,

A small change in the levels of the grouped elements has been observed, i.e. elements that are within the same group can be grouped in a different order when changing methods, where generally the grouping structures are quite similar.

In Table 2 obtained through the Mahalanobis distance, the most similar states are AM, TO, CE, PE, ES, RJ, PR, SC, MT, DF and the most distant were AL, MG, SP, that is, in the HDI rankings the state of Alagoas is in position 27 and the state of São Paulo in the second position as quite distinct HDI.

Table 2: Mahalanobis distance matrix applied to the HDI data of the states taken from Atlas Brazil 2013.

	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
AC	43.89																									
AM	23.42	43.89																								
RR	31.57	43.89	31.57																							
PA	43.89	1.28	43.89	43.89																						
A P	31.57	43.89	31.57	2.41	43.89																					
TO	23.42	43.89	3.38	31.57	43.89	31.57																				
MA	43.89	0.51	43.89	43.89	1.28	43.89	43.89																			
PI	31.57	43.89	31.57	2.41	43.89	0.64	31.57	43.89																		
CE	23.42	43.89	3.38	31.57	43.89	31.57	0.70	43.89	31.57																	
RN	43.89	2.17	43.89	43.89	2.17	43.89	43.89	2.1707	43.89	43.89																
PB	31.57	43.89	31.57	9.60	43.89	9.60	31.57	43.89	9.60	31.57	43.89															
PE	23.42	43.89	12.02	31.57	43.89	31.57	12.02	43.89	31.57	12.02	43.89	31.57														
AL	43.89	13.12	43.89	43.89	13.12	43.89	43.89	13.12	43.89	43.89	13.12	43.89	43.89													
SE	31.57	43.89	31.57	9.60	43.89	9.60	31.57	43.89	9.60	31.57	43.89	0.09	31.57	43.89												
BA	31.57	43.89	31.57	0.13	43.89	2.41	31.57	43.89	2.41	31.57	43.89	9.60	31.57	43.89	9.60											
MG	43.89	13.12	43.89	43.89	13.12	43.89	43.89	13.12	43.89	43.89	13.13	43.89	43.89	0.04	43.89	43.89										
ES	23.42	43.89	12.02	31.57	43.89	31.57	12.02	43.89	31.57	12.02	43.89	31.57	5.09	43.89	31.57	31.57	43.89									
RJ	23.42	43.89	12.02	31.57	43.89	31.57	12.02	43.89	31.57	12.02	43.89	31.57	0.29	43.89	31.57	31.57	43.89	5.09								
SP	43.89	13.12	43.89	43.89	13.12	43.89	43.89	13.12	43.89	43.89	13.12	43.89	43.89	0.11	43.89	43.890	0.11	43.89	43.89							
PR	23.42	43.89	12.02	31.57	43.85	31.57	12.02	43.89	31.57	12.02	43.89	31.57	5.09	43.89	31.57	31.57	43.89	0.18	5.09	43.89						
SC	23.42	43.89	3.38	31.57	43.89	31.57	0.70	43.89	31.57	0.29	43.89	31.57	12.02	43.89	31.57	31.57	43.89	12.02	12.02	43.89	12.02					
RS	43.89	2.17	43.89	43.89	2.17	43.89	43.89	2.17	43.89	43.89	0.34	43.89	43.89	13.12	43.89	43.89	13.12	43.89	43.89	13.12	43.89	43.89				
MS	31.57	43.89	31.57	9.60	43.89	9.60	31.57	43.89	9.60	31.57	43.89	2.40	31.57	43.89	2.40	9.60	43.89	31.57	31.57	43.89	31.57	31.57	43.89			
MT	23.42	43.89	12.02	31.57	43.89	31.57	12.02	43.89	31.57	12.02	43.895	31.57	0.55	43.89	31.57	31.57	43.89	5.094	0.55	43.89	5.09	12.02	43.89	31.57		
GO	43.89	1.28	43.89	43.89	0.361	43.89	43.89	1.28	43.89	43.89	2.17	43.89	43.89	13.12	43.89	43.89	13.12	43.89	43.89	13.12	43.89	43.89	2.17	43.89	43.89	
DF	23.42	43.89	12.02	31.57	43.89	31.57	12.02	43.89	31.57	12.02	43.89	31.57	5.09	43.89	31.57	31.57	43.89	1.33	5.09	43.89	1.33	12.02	43.89	31.57	5.09	43.89

Source: Authors.

When analyzing the dendograms, the presence of five groups was verified without using a cut-off criterion, each dendogram presented structures of groupings of homogeneous states to determine the groups that were formed, the dendograms represent different aspects for each group of states, some groups are identical, however the numbers of states obtained in each group are similar (figure 1 to 4).

The appropriate hierarchical method of grouping, was the complete linkage method obtained through the Mahalanobis Distance matrix, since the cohenetic correlation coefficient shows a better distinction of the groups when the coefficient is greater than 0.70 found in each hierarchical method (Table 3). From the dendrogram the complete linkage method its CCC was 0.728 which corresponds to 72.8% consistency in clustering, that is, the state that are within the same group can be grouped in other ways when changing the method.

Table 3: Co-phenetic correlation coefficient obtained.

Links	Correlation
Simple	0.676
Full	0.728
Mean	0.669
Ward	0.474

Source: Authors.

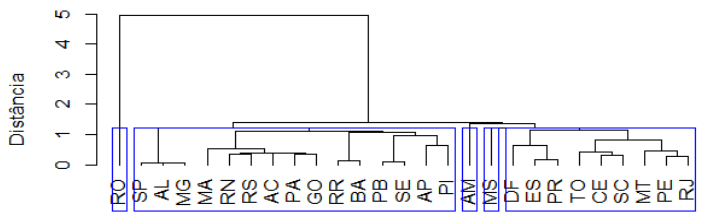
The method that best represents the solution to this problem is Complete Linkage with deferent numbers of states in each cluster. Although the overall structure of the cluster is quite similar, with an irrelevant change in the grouped states.

For a clearer result, the individual analysis of each method used will be performed for the groups, taking into account each group formed using the average of each variable.

Simple Connection: Composed by five groups, it presents three unitary groups and two groups with states from different regions of Brazil (Figure 1). In this method the states that stood out were:

- **Rondônia:** Located in the North region, with an HDI = 0.690, an estimated population of 1,815,278, the main economic activities are agriculture, cattle breeding, the food industry, and vegetal and mineral extraction.
- **Amazonia:** With 4,144,597 inhabitants and an HDI = 0.674, its economy is based on the primary sector, with the extractive activities (animals, minerals, and vegetables) as the highlight.
- **Mato Grosso do Sul:** located in the central region of Brazil with 2,778,986 inhabitants and an HDI = 0.729, the main economic activity is still agriculture and cattle raising.

Figure 1: Dendrogram obtained by means of the Simple linkage algorithm, based on the mahalanobis distance.

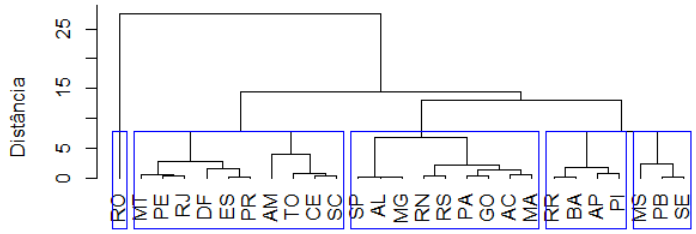


Source: Authors.

Complete linkage: Composed by five groups, it presents a unitary group, four groups with numbers of states from different regions of Brazil (Figure 2). In group two the states that stood out in this method were:

- **São Paulo:** It has an estimated population of 46,649,132 inhabitants, with the HDI = 0.783 the second highest among the states, located in the Southeast region of the country. The state has a diversified economy responsible for about one-third of Brazil's GDP.
- **Alagoas:** Located in the Northeast region with an estimated population of 3,365,351 inhabitants, it has an economy in several areas with agriculture (pineapple, coconut, sugar cane, beans, etc.), industry (construction, food, etc.) and tourism that has increased in recent years, but even with all this it has an HDI = 0.631 considered the lowest in the country.

Figure 2: Dendrograms obtained using the complete linkage algorithm, based on the Mahalanobis distance.



Source: Authors.

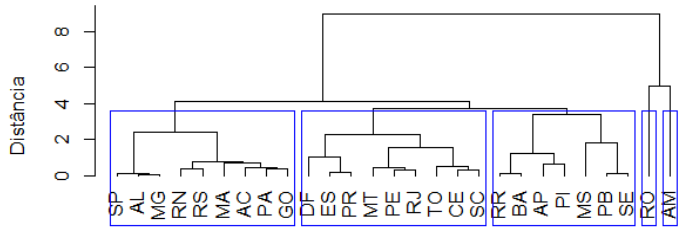
Medium link: Composed of five groups, it presents two unitary groups, three groups with states from different regions of Brazil (Figure 3). Group two stands out for having at least one state from each region of the country, the states are:

- **Tocantins:** Located in the North region of the country, with an HDI = 0.699 and a population estimated at 1,607,363 inhabitants, the economy of Tocantins is based on an aggressive expansionist model of agro-exports, in other words, agricultural products (rice, soy, pineapple, corn, and others).
- **Pernambuco:** With HDI = 0.673, located in the Northeast region of the country, its industrial production is among the largest in the North-Northeast, with the following sectors: naval, automobile, chemical, metallurgical, flat glass, electro-electronic, non-metallic minerals, textile, and food industries, but also known today as the largest producer of guava and acerola.
- **Ceará:** Also located in the Northeast region of the country, it has an HDI = 0.682, the Ceará economy has been growing, that is, it stands out in the agricultural activity (beans, corn, rice, herbaceous cotton, tree cotton, cashew nuts, sugar cane, cassava, castor beans, tomatoes, bananas, oranges, coconuts, and, more recently, grapes), in the industry sectors (clothing, food, metallurgy, textiles, chemicals, and footwear), with an estimated population of 9,240,580 inhabitants.
- **Distrito Federal:** Known as Brasilia the capital of Brazil has the highest HDI= 0.824, located in the central region of the country, it is an important economic center, the main economic activity of the federal capital results from its administrative function. The estimated population is 3,094,325 people.
- **Mato Grosso:** It is located in the Central region, its population estimated at 3,567,234 people, this state leads as the largest national producer of grains in the country, but also produces other crops (beans, sugarcane, corn, cotton, sunflower, cassava), that is, one of the main producers and exporters of soybeans in Brazil.
- **Rio de janeiro:** Situated in the southeast region of the country, a large part of the state's economy is based on services, with a significant part of industry and little influence in the agricultural sector, its population is estimated at 17,463,349 people.
- **Espírito Santo:** It is located in the southeast region, with an estimated population of 4,108,508 people, in the economy it has stood out in agriculture, cattle breeding, and mining, in agricultural production with sugar cane, oranges, and coffee. The

main industrial sectors are: Oil and Natural Gas Extraction, Extraction of metallic minerals, Construction, Metallurgy and Industrial Services of Public Utility, such as Electric Power and Water.

- **Santa Catarina:** Located in the southern region of the country with an estimated population of 7,338,473 people, its economy is based on the following activities: industry (especially agro-industry, textiles, ceramics, and metal-mechanics), extractivism (minerals), and cattle-raising.
- **Paraná**: The main economic activities are agriculture (sugar cane, corn, soy, wheat, coffee, tomatoes, manioc), industry (agro-industry, automobile, paper and cellulose), and vegetal extraction (wood and yerba mate). It has an estimated population of 11,597,484 people.

Figure 3: Dendrograms obtained using the average linkage algorithm, based on the Mahalanobis distance.

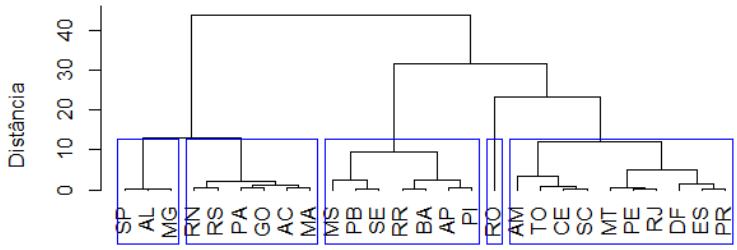


Source: Authors.

Ward Connection: Composed of five groups, it presents a unitary group, a group with three states and three groups with several states from different regions of Brazil (Figure 4), in this method the group two stood out with the following states:

- **Acre:** With an estimated population of 881,935 inhabitants, estimated located in the northern region of Brazil its economy is based on the exploitation of the resources found in the forest reserves, has the culture influenced by indigenous people who inhabit the region.
- **Pará:** Has an estimated population of approximately 8,602,865 inhabitants, the economy is based on the provision of services, commercial and agricultural activities also located in the northern region.
- **Maranhão:** Located in the northeast region, it has 7.075.181 inhabitants approximately, its main economy is based on the agricultural, industrial and mineral activities.
- **Rio Grande do Norte:** Also belonging to the Northeast region, with 3.506.853 inhabitants approximately, its HDI = 0,684 considers that it is the biggest of the region it is part of and has an economy based on commerce, on the textile industry, on agribusiness, tourism and on the extraction and processing of oil.
- **Rio grande do Sul:** Located in the southern region of Brazil, with a population of approximately 11.377.239 inhabitants, the main source of income is based on agribusiness and farming.
- **Goiás:** The estimated population of 7.018.354 inhabitants, its main economic activities are livestock farming, agriculture and agribusiness located in the central region of the country.

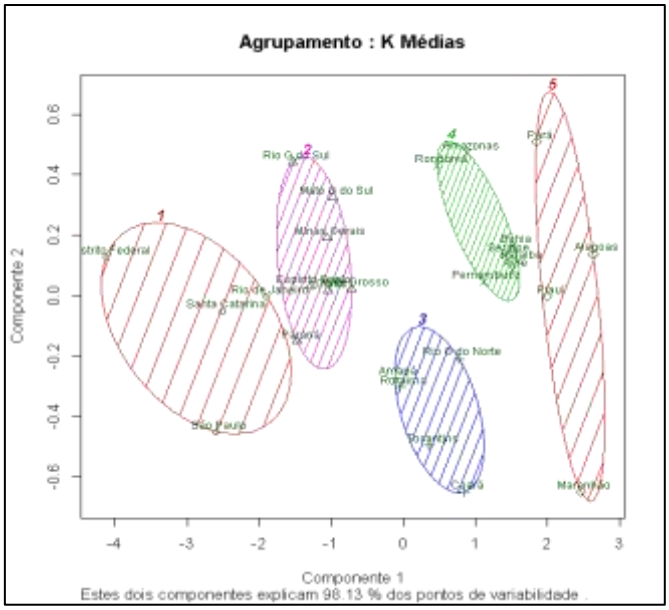
Figure 4: Dendrogram obtained using Ward's algorithm, based on the Mahalanobis distance.



Source: Authors.

In the non-hierarchical method as shown in Figure 5 the k-means method was applied to obtain new groups, 5 groups were obtained from this method.

Figure 5: Figure obtained using the non-hierarchical method.



Source: Authors.

Five new groups were found in this scatter diagram that will be used to explain the k-means method as shown in Table 4.

Table 4: groups formed in the dendogram obtained in figure 5.

Groups	States
1	Distrito Federal, São Paulo, Santa Catarina, Rio de janeiro
2	Paraná, Rio grande do Sul, Espírito Santo, Goiás, Minas Gerais, Mato Grosso do Sul, Mato Grosso
3	Amapá, Roraima, Tocantins, Rio Grande do Norte, Ceará
4	Rondônia, Amazonas, Pernambuco, Sergipe, Acre, Bahia, Paraíba
5	piauí, pará, maranhão, alagoas

Source: Authors.

When observing the k-means method, the groups obtained have no unit group, group 1 in it are the states with the highest HDI which are: Federal District, São Paulo, Santa Catarina and Rio de Janeiro of the Southeast, South and Midwest

regions of Brazil. Group 2 is composed of seven states with HDI between 0.725 and 0.749. Group 3 is composed of five states with an HDI between 0.682 and 0.708. Group 4 is composed of seven states with an HDI between 0.658 and 0.690, and Group 5 is composed of four states: Piauí, Pará, Maranhão and Alagoas.

4. Conclusion

In this work the use of cluster analysis was proposed because it is an important tool that allows the classification of individuals based on the observation of similarity between the variables being used.

In the study it was possible to evaluate the hierarchical and non-hierarchical methods observing the behavior of each Brazilian state according to the Human Development Index (HDI), available in Atlas Brazil 2013. A cluster analysis was performed considering the variables HDI-R, HDI-L and HDI-E. Thus, according to the methods used, five groups were identified both for the hierarchical technique and for the non-hierarchical technique in each of the methods referring to the Brazilian regions.

One similarity that can be observed in the hierarchical technique that between each method evaluated the state of Rondônia remained isolated in a cluster in all observed links, another point that also those of the states that regardless of their HDI the state of São Paulo and Acre remain in the same group. As for the non-hierarchical technique in the k-means method, what drew more attention was the states that have the highest HDI remain together in the same group, as also occurred with the states of lower HDI.

Support and Acknowledgments

Thanks 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. (2020). Determinação do número de grupos em análise de agrupamento via de raio de influência. *Brazilian Journal of Development*, *6*(6), 38342-38355.

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.http://eduepb.uepb.edu.br/download/introducao-a-analise-de-agrupamento-teoria-e-pratica-com-aplicacoes-em-r/?wpdmdl=997&masterkey=5e97904980fc9

Albuquerque, T. M., Araujo, G. A. B., Caminha, B. L., Albuquerque, M. L., & Albuquerque, M. A. (2016). Measures of association in epidemiological studies: smoking mothers and low birth weight children in the city of Campina Grande–PB. Acta Scientiarum. Health Sciences, 38(2), 179-84.

Alves, L. B., Belderrain, M. C., & Scarpel, R. A. (2007). Tratamento multivariado de dados por análise de correspondência e análise de agrupamentos. Anais do 13º encontro de iniciação científica e pós-graduação do ITA-XIII ENCITA. São José dos Campos-SP.

Braga, A. C., de Oliveira, M. A., Costa, J. C. Z., & Bueno, R. L. P. (2017). Estudo da Correlação entre o Índice de Desenvolvimento Humano (Idh) e os Tributos Arrecadados nos Estados Brasileiros. *Interfaces Científicas-Humanas e Sociais*, *5*(3), 69-84.

Barroso, L. P., & Artes, R. (2003). Análise multivariada. Lavras: Ufla, 151.

Campos, S. L S. **(**2019) *Busca não supervisionada de padrões por técnicas de agrupamento clássica e nebulosa*.

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

Gomes, L. M. (2020). Violência obstétrica: perspectiva de puérperas atendidas em um Hospital Universitário-HCU UFU.

Johnson, R. A., & Wichern, D. W. (2002). Applied multivariate statistical analysis (Vol. 5, No. 8). Upper Saddle River, NJ: Prentice hall.

Kaufman, L., & Rousseeuw, P. J. (2009). Finding groups in data: an introduction to cluster analysis (Vol. 344). John Wiley & Sons

Martin, N., & Maes, H. (1979). Multivariate analysis. London: Academic press.

Nascimento, J. M. P. (2019) *Análise de Agrupamento Aplicado aos Dados de Violência no Brasil*. UFU,

Oliveira Barros, K. N. N., de Albuquerque, M. A., dos Santos Gomes, A., & Dantas, D. R. G. (2020). Análise de agrupamentos exploratória dos usuários do Programa Multidisciplinar de Tratamento do Tabagismo do HUAC, Campina Grande–PB. Research, Society and Development, 9(8), e825986532-e825986532.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Viena, Austria. Disponível em: http://www.R-project.org/