



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

WYLLIAM EDUARDO ALVES SILVA

**ANÁLISE DE *CLUSTER* APLICADA AOS DADOS DE PREÇOS DE
COMBUSTÍVEIS NA CIDADE DE CAMPINA GRANDE - PB**

**CAMPINA GRANDE - PB
2021**

WYLLIAM EDUARDO ALVES SILVA

**ANÁLISE DE *CLUSTER* APLICADA AOS DADOS DE PREÇOS DE
COMBUSTÍVEIS NA CIDADE DE CAMPINA GRANDE - PB**

Trabalho de Conclusão de Curso 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. Me. Ednário Barbosa de Mendonça

**CAMPINA GRANDE - PB
2021**

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

S586a Silva, Wylliam Eduardo Alves.
Análise de cluster aplicada aos dados de preços de combustíveis na cidade de Campina Grande - PB [manuscrito]
/ Wylliam Eduardo Alves Silva. - 2021.
31 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2021.

"Orientação : Prof. Me. Ednário Barbosa de Mendonça, Departamento de Estatística - CCT."

1. Análise multivariada. 2. Análise de cluster. 3. Método hierárquico. 4. Preços de combustível. I. Título

21. ed. CDD 519.5

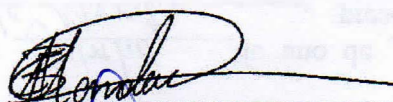
WYLLIAM EDUARDO ALVES SILVA

**ANÁLISE DE CLUSTER APLICADA AOS DADOS DE PREÇOS DE COMBUSTÍVEIS
NA CIDADE DE CAMPINA GRANDE - PB**

Trabalho de Conclusão de Curso 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.

Trabalho aprovado em 08 de Junho de 2021.

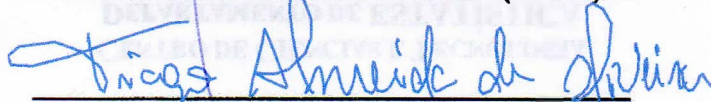
BANCA EXAMINADORA



Prof. Me. Ednário Barbosa de Mendonça
Universidade Estadual da Paraíba (UEPB)



Prof. Dr. Sílvio Fernando Alves Xavier Júnior
Universidade Estadual da Paraíba (UEPB)



Prof. Dr. Tiago Almeida de Oliveira
Universidade Estadual da Paraíba (UEPB)

Dedico este trabalho a minha mãe, Maria Cleoneide Alves da Silva, por todo apoio que tem me dado ao longo de minha vida e por todo esforço feito para fazer com que eu me tornasse uma pessoa melhor.

AGRADECIMENTOS

A Deus, por ter permitido que eu tivesse saúde e determinação para que meus objetivos fossem alcançados, durante todos esses anos de estudos.

Aos meus pais, Edmundo e Maria Cleoneide, por sempre querer o meu melhor, pelo empenho, esforço e dedicação para que eu estudasse em outro estado. Meus irmãos, familiares e todos os amigos que hoje estão distantes, obrigado pelo apoio, incentivo e por fazerem parte desta história de conquistas.

A Universidade Estadual da Paraíba pela oportunidade de realizar este curso. Aos professores, Ana Patricia, Diana, Geselly, Gustavo, Ricardo, Sílvio, Tiago e Vitoria, pela contribuição em minha vida acadêmica e profissional.

Ao Procon Municipal de Campina Grande - PB pela oportunidade de estagiar e colocar em prática todo conhecimento adquirido em sala de aula e por fornecer os dados para realização deste trabalho. Também agradeço o companheirismo de toda equipe, em especial Ana Cláudia, Antônio, Eufrásia, Jonathan, Miguel e Rivaldo (*in memoriam*).

Ao Instituto Brasileiro de Defesa do Consumidor (Idec) pela oportunidade de estagiar no programa de saúde, no qual, mais uma vez coloquei em prática todo o processo de aprendizagem durante minha formação. Agradeço a equipe do programa de saúde por todo apoio e contribuição com minha vida profissional, Ana, Laura, Mariana, Marina e Matheus, além de toda equipe do Idec.

Agradeço a minha namorada Mayara Macedo que sempre esteve ao meu lado, me incentivando e dando todo apoio que precisei. Muito obrigado por todo amor, carinho e paciência. Amo você!

Agradeço também a todos os amigos e colegas que o espetinho do Sr Antônio me deu, Alline, Andreza, Breno, Felipe, Jorge, Luiza, Michel e Niedja, sei a importância de cada um, foram muitos momentos felizes que compartilhamos juntos. Obrigado pelo apoio.

Aos amigos de curso Alvaro, Beatriz, Filype, Gilmar, Jeysianne, Mateus, Mayara, Rafella e Viviane por todos os momentos vividos durante todo curso. Também agradeço aos amigos que hoje não estão mais no curso, mas iniciamos juntos a turma de 2015.2, Allan, Augusto, Emanuel, Louhanne e Rayan. Juntos construímos uma brilhante trajetória de conhecimentos partilhados.

Por fim, e não menos importante, gostaria de deixar os meus agradecimentos ao meu orientador Ednário Barbosa, pois sem a sua orientação não teria conseguido realizar este trabalho, obrigado por todos os conselhos, pela paciência e por acreditar em mim.

*“Você não é definido pelo seu passado,
você é preparado por ele.”
(Joel Osteen)*

*“A estatística é a gramática da ciência.”
(Karl Pearson)*

RESUMO

O presente trabalho teve por objetivo verificar as similaridades dos preços, através da análise de *cluster* para os combustíveis, gasolina comum e etanol referente ao ano de 2019 no município de Campina Grande – PB. Postos de combustíveis são instalações que vendem combustível para veículos a motor, os tipos mais comuns vendidos são gasolina ou diesel, alguns postos fornecem combustível alternativos como álcool (etanol) e gás natural. A análise de *cluster*, também chamada de análise de agrupamento, trata-se de uma técnica estatística multivariada usada para classificar elementos em grupos, de forma que elementos dentro de um mesmo *cluster* sejam muito parecidos e os elementos em diferentes *clusters* sejam distintos entre si. Os métodos mais utilizados nas análises de *clusters* são, método hierárquico que consiste em organiza um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade dos indivíduos e o método não hierárquico que é caracterizados pela necessidade de definir uma partição inicial, dividem a base de dados em K -grupos, onde o número K é a quantidade de grupos definida previamente. Para tratamento e análise dos dados foi utilizado o programa computacional R, posteriormente foi realizada uma análise exploratória dos dados para observar algumas características importantes, também foi realizado o teste de Shapiro-Wilk para avaliar a normalidade dos dados, correlação de Spearman para medir a relação entre as variáveis. Partindo para análise de *cluster*, no método não hierárquico foi utilizado o método *k-means*, de modo que obteve-se os *clusters* e suas devidas características, como resultados, obtiveram-se 3 (três) *clusters* através de método de *Elbow* (cotovelo), *cluster* 1 contendo 15 postos de combustíveis, *cluster* 2, 37 postos e *cluster* 3 com 5 postos, em que por meio das análises exploratórias dos *clusters* observou-se que o *cluster* 1 teve a menor média de preço de combustível tanto para o etanol quanto para gasolina comum, ou seja, os postos que faziam parte deste *cluster* apresentavam os menores preços para estes dois combustíveis, já o *cluster* 3 apresentou a maior média para esses combustíveis. No método hierárquico foi utilizada a distância euclidiana, utilizando o mesmo critério do método anterior 3 *clusters* foram obtidos, através do gráfico do dendrograma verificou-se onde os postos se encontram em cada um dos *clusters*, assim identificando quais postos tem os melhores e os piores preços de etanol e gasolina comum. A partir do estudo proposto, pode-se concluir que a análise de *cluster* se mostra uma importante e muito útil para este estudo, no método hierárquico utilizou a distância euclidiana e a ligação completa, no não hierárquico utilizou-se o método *k-means*, ou seja, isso nos mostra que a análise de *cluster* foi bem empregada, pois em ambos os métodos não houve mudança dos postos de *clusters*.

Palavras-chaves: Análise multivariada. Análise de *cluster*. Método hierárquico. Preços de combustível.

ABSTRACT

The present work aimed to verify the similarities of prices, through cluster analysis for the fuels, regular gasoline and ethanol referring to the year of 2019 in the city of Campina Grande - PB. Gas stations are facilities that sell fuel for motor vehicles, the most common types sold are gasoline or diesel, some stations provide alternative fuels such as alcohol (ethanol) and natural gas. The cluster analysis, is a multivariate statistical technique used to classify elements into groups, so that elements within the same cluster are very similar and the elements in different clusters are distinct from each other. The most commonly used methods in cluster analysis are, hierarchical method that consists of organizing a data set into a hierarchical structure according to the proximity of individuals and the non-hierarchical method which is characterized by the need to define an initial partition, divide the database into K-groups, where the number K is the number of groups previously defined. For treatment and analysis of the data we used the computer program R, later an exploratory analysis of the data was performed to observe some important characteristics, the Shapiro-Wilk test was also performed to evaluate the normality of the data, Spearman's correlation to measure the relationship between the variables. Leaving for cluster analysis, in the non-hierarchical method the k-means method was used, so that obtained the clusters and their characteristics, as results, obtained 3 (three) clusters using the Elbow method, cluster 1 containing 15 gas stations, cluster 2, 37 and cluster 3 with 5 posts, where through the exploratory analysis of clusters it was observed that cluster 1 had the lowest average fuel price for both ethanol and common gasoline, i.e, the gas stations that were part of this cluster had the lowest prices for these two fuels, cluster 3 has the highest average for these fuels. In the hierarchical method used the Euclidean distance, using the same criteria as the previous method, we had 3 clusters, through the graph of the dendrogram it was verified where the stations are in each of the clusters, thus identifying which stations have the best and worst prices of ethanol and regular gasoline. From the proposed study, it can be concluded that cluster analysis proves to be an important and very useful for this study, in the hierarchical method used the Euclidean distance and the complete linkage, in the non-hierarchical one used the *k*-means method, i.e, this shows us that the cluster analysis was well employed, because in both methods there was no change of the cluster stations.

Keywords: Multivariate analysis. Cluster analysis. Hierarchical method. Fuel prices.

LISTA DE ILUSTRAÇÕES

Figura 1 – Dendrograma	15
Figura 2 – Boxplot para gasolina comum e etanol	23
Figura 3 – Diagrama de dispersão	23
Figura 4 – Números de <i>clusters</i> através do <i>scree plot</i>	24
Figura 5 – Frequência de postos por <i>cluster</i>	25
Figura 6 – Localização dos postos nos <i>clusters</i>	26
Figura 7 – Matriz de distância	27
Figura 8 – Dendrograma	28

LISTA DE TABELAS

Tabela 1 – Sumário das estatísticas para as variáveis	22
Tabela 2 – Resultados do teste de normalidade (Shapiro - Wilk)	24
Tabela 3 – Sumário das estatísticas para gasolina comum por <i>clusters</i>	25
Tabela 4 – Sumário das estatísticas para etanol por <i>clusters</i>	26

SUMÁRIO

1	INTRODUÇÃO	11
2	FUNDAMENTAÇÃO TEÓRICA	12
2.1	Análise de Cluster	12
2.2	Medidas de Semelhança e Distância	13
2.2.1	<i>Coefficientes de Correlação</i>	13
2.2.2	<i>Medidas de Distância</i>	13
2.2.2.1	<i>Distância Euclidiana</i>	14
2.2.2.2	<i>Distância Euclidiana Quadrática</i>	14
2.2.2.3	<i>Distância de Manhattan</i>	14
2.2.2.4	<i>Distância de Minkowski</i>	14
2.2.2.5	<i>Distância de Mahalanobis</i>	15
2.2.3	<i>Métodos Hierárquicos</i>	15
2.2.3.1	<i>Métodos Aglomerativos</i>	16
2.2.3.2	<i>Métodos Divisivos</i>	19
2.2.4	<i>Métodos não hierárquicos ou por Particionamento</i>	19
2.2.4.1	<i>Método k-means</i>	20
2.2.4.2	<i>Método k-medoid</i>	20
2.3	Material e Métodos	21
3	RESULTADOS E DISCUSSÕES	22
3.1	Análise Exploratória	22
3.2	Método não hierárquico	24
3.3	Método Hierárquico	27
4	CONCLUSÃO	29
	REFERÊNCIAS	30

1 INTRODUÇÃO

De acordo com (PETRÓLEO, 2017) (ANP), postos de combustíveis são instalações que vende combustível para veículos a motor, os tipos mais comuns vendidos são gasolina ou diesel, alguns postos fornecem combustível alternativos como álcool (etanol) e gás natural. Para que um posto de combustível seja instalado e entre em funcionamento deve passar por uma vistoria minuciosa da Agência Nacional do Petróleo, Gás Natural e Biocombustíveis (ANP).

Segundo (UCHÔA, 2008), toda a sociedade está prestando muita atenção à trajetória de alta dos preços do petróleo, seu impacto direto nos níveis de preço dos combustíveis também é um dos mais óbvios sem dúvida sobre a gasolina. Além da alta no preço da gasolina, podemos ver que o etanol vem acompanhando esse aumento.

Análise de *cluster* é o processo de separar o conjunto de dados em componentes que refletem padrões consistentes de comportamento, particionando o banco de dados de forma que cada partição ou grupo seja similar de acordo com algum critério ou métrica, (DONI, 2004). Uma vez que os padrões tenham sido estabelecidos, estes podem ser utilizados para “desmontar” os dados em subconjuntos mais compreensíveis para futuras análises.

Este método é muito utilizado quando contamos com um grande número de observações, onde um dos principais desafios do analista é resumir a informação coletada, pode ser de interesse criar grupos. O seu emprego em áreas tais como experimentos agrônômicos, medicina, economia, administração, entre outras, vem aumentando muito nos últimos anos (ALBUQUERQUE, 2005).

Os grupos são baseados em suas distâncias, medida de dissimilaridades (Euclidiana, Mahalanobis, etc.) ou similaridades. Um método de ligação entre os grupos refere-se à adoção de uma técnica para a formação dos grupos. Já existe um grande número de medidas de similaridade ou dissimilaridade propostas e usadas para análise de *cluster*, a seleção de uma delas fica de acordo com a preferência ou conveniência do pesquisador (BUSSAB; MIAZAKI; ANDRADE, 1990).

O objetivo deste trabalho é verificar as similaridades dos preços para os seguintes combustíveis, gasolina comum e etanol referente ao ano de 2019 no município de Campina Grande – PB. Para isto, foi utilizada a técnica multivariada análise de *cluster*, que consiste em formar grupos semelhantes, no caso do trabalho comparar os postos com preços mais parecidos.

2 FUNDAMENTAÇÃO TEÓRICA

Análise de *cluster* consiste em agrupar observações de modo que cada observação seja o mais semelhante possível segundo suas características, os grupos tem que ter elevada homogeneidade interna e heterogeneidade externa. Neste capítulo serão apresentados os principais métodos da análise, sendo eles, método hierárquico e o não hierárquico, que são métodos fundamentais para o estudo da análise de *cluster*. Primeiramente, destacam-se as matrizes de similaridade e o uso das medidas de dissimilaridade. Em seguida serão apresentados algumas definições e resultados relacionados a essa teoria.

2.1 Análise de Cluster

De acordo com (HAIR et al., 2009), análise de *cluster* (também conhecida como análise de agrupamentos) é um grupo de técnicas multivariadas cuja finalidade principal é agregar objetos com base nas características que eles possuem. Os objetos, em cada grupo, tendem a ser semelhante entre si, mas diferentes de objetos em outros grupos, (MALHOTRA, 2001).

De acordo com (REIS, 2001), análise de *cluster* pode ser classificada nas seguintes etapas:

- Seleção de indivíduos ou de uma amostra de indivíduos a serem agrupados;
- Definição de um conjunto de variáveis a partir das quais será obtida a informação necessária ao agrupamento dos indivíduos;
- Definir uma medida de similaridade ou dissimilaridade (parecença) entre cada dois indivíduos;
- Escolha de um critério de agregação ou desagregação dos indivíduos, isto é, a definição de um algoritmo de partição/ classificação;
- Por último, a validação dos resultados obtidos.

De acordo (ZAIANE et al., 2002), está ficando difícil acompanhar todas as novas estratégias de agrupamento. Uma análise de *cluster* bem acurada exige métodos que apresentem as seguintes características: trabalhar com grandes massas de dados, os métodos de agrupamentos devem se ajustar a todos os diferentes tipos de dados, ter o mínimo de conhecimento para determinação do parâmetro de entrada, ser robusto em relação a ruído, o método de agrupamento deve fornecer resultados consistentes, ser “escalável” com o número de dimensões e com a quantidade de elementos a serem agrupados, capacidade de descobrir *clusters* com diferentes formas. Não existe um algoritmo único que possa satisfazer totalmente todos os requisitos citados.

2.2 Medidas de Semelhança e Distância

Para (BARROSO; ARTES, 2003), há duas medidas de parecnça: 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).

2.2.1 Coeficientes de Correlação

Esses coeficientes são caracterizados por serem de fácil interpretação geométrica, são das medidas de semelhanças mais utilizadas nas ciências sociais, em particular o coeficiente de correlação de Pearson (REIS, 2001). É comum atribuir exclusivamente a Karl Pearson o desenvolvimento dessa estatística, no entanto, como bem lembrou (STANTON, 2001), a origem desse coeficiente remonta o trabalho conjunto de Karl Pearson e Francis Galton.

Este coeficiente é definido, em geral, para duas variáveis e mede o grau de associação linear entre elas.

Seja i e j dois objetos, caracterizados por um conjunto p de variáveis, temos:

$$r_{ij} = \frac{\sum_{v=1}^p (X_{iv} - \bar{X}_i)(X_{jv} - \bar{X}_j)}{\sqrt{\sum_{v=1}^p (X_{iv} - \bar{X}_i)^2 \sum_{v=1}^p (X_{jv} - \bar{X}_j)^2}} \quad (2.1)$$

sendo:

X_{iv} : valor da variável v para o objeto i , ($v = 1, \dots, p$);

X_{jv} : valor da variável v para o objeto j ;

\bar{X}_i : média de todas as variáveis para o objeto i ;

\bar{X}_j : média de todas as variáveis para o objeto j ;

p : número total de variáveis.

O valor do coeficiente varia entre -1 e 1, o valor zero significa que não existe correlação entre as variáveis.

Observação: Além da correlação de Pearson, também pode ser utilizada a correlação de Spearman.

2.2.2 Medidas de Distância

De acordo com (DONI, 2004), a maioria dos métodos de análise de *cluster* requer uma medida de dissimilaridade entre os elementos a serem agrupados, normalmente expressa como uma função distância ou métrica.

Seja M um conjunto, uma métrica em M é uma função $d : M \times M \rightarrow \mathbb{R}$, tal que para quaisquer $i, j, z \in M$, tenhamos:

1. $d_{ij} > 0 \forall i \neq j$ (positiva);
2. $d_{ij} = 0 \Leftrightarrow i = j$ (reflexiva);

3. $d_{ij} = d_{ji} \geq 0$ (simétrica);
4. $d_{ij} \leq d_{iz} + d_{zj}$ (desigualdade triangular).

Segundo (CORMACK, 1971), existem várias medidas que são utilizadas como medidas de distância. A seguir são apresentadas algumas distâncias enquadradas na definição de medidas de dissimilaridade.

2.2.2.1 Distância Euclidiana

A distância euclidiana é vastamente uma das medidas de dissimilaridade mais utilizada para a análise de *cluster* quando os indicadores ou variáveis são completamente quantitativos. De acordo com (CRISPIM; FERNANDES; ALBUQUERQUE, 2019), a distância euclidiana é definida pela expressão:

$$d_{ij} = \sqrt{\sum_{v=1}^p (X_{iv} - X_{jv})^2} \quad (2.2)$$

em que X_{iv} e X_{jv} representam as características dos indivíduos i e j para todas as variáveis ($v = 1, \dots, p$).

2.2.2.2 Distância Euclidiana Quadrática

A distância euclidiana quadrática é definida pela expressão:

$$d_{ij}^2 = \sum_{v=1}^p (X_{iv} - X_{jv})^2 \quad (2.3)$$

em que X_{iv} e X_{jv} representam as características dos indivíduos i e j para todas as variáveis ($v = 1, \dots, p$).

2.2.2.3 Distância de Manhattan

A Distância de *Manhattan* ou (*city block*) é definida pela expressão:

$$d_{ij} = \sum_{v=1}^p (|X_{iv} - X_{jv}|)^1 \quad (2.4)$$

em que X_{iv} e X_{jv} representam as características dos indivíduos i e j para todas as variáveis ($v = 1, \dots, p$).

2.2.2.4 Distância de Minkowski

Essa distância é uma generalização da distância euclidiana, que é definida pela expressão:

$$d_{ij} = \left(\sum_{v=1}^p |X_{iv} - X_{jv}|^n \right)^{\frac{1}{n}} \quad (2.5)$$

- n assume valores inteiros e positivos ($n = 1, 2, \dots$);
- $n = 2$ Representa a Distância Euclidiana;
- $n = 1$ Representa a Distância de Manhattan (city block).

2.2.2.5 Distância de Mahalanobis

A distância de *mahalanobis* ou distância generalizada foi introduzida pelo matemático indiano Prasanta Chandra Mahalanobis em 1936, é definida pela expressão:

$$d_{ij} = \sqrt{(\mathbf{X}_i - \mathbf{X}_j)' \Sigma^{-1} (\mathbf{X}_i - \mathbf{X}_j)} \quad (2.6)$$

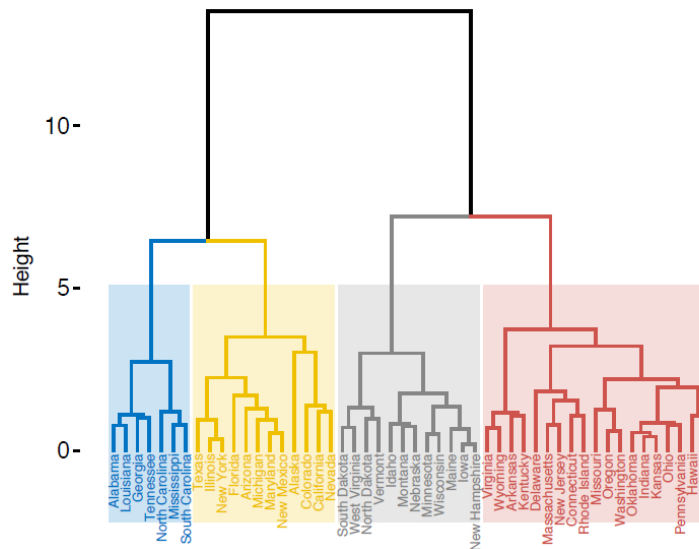
em que $\mathbf{X}'_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ e $\mathbf{X}'_j = (x_{j1}, x_{j2}, \dots, x_{jp})$ representam os vetores de médias dos grupos i e j , Σ^{-1} é a inversa da matriz de covariância.

2.2.3 Métodos Hierárquicos

Segundo (REIS, 2001) este método organiza um conjunto de dados em uma estrutura hierárquica de acordo com a proximidade dos indivíduos. Os métodos hierárquicos são compostos por duas classes naturais de algoritmos para formação dos agrupamentos, realizadas por uma série de fusões sucessivas (método aglomerativos) ou por uma série de sucessivas divisões (método divisos), (MINGOTI, 2007).

De acordo com (DONI, 2004), nos métodos hierárquicos os grupos são geralmente representados por um dendrograma ou diagrama de árvore. Neste diagrama, cada ramo representa um elemento, enquanto a raiz representa o grupo de todos os elementos.

Figura 1 – Dendrograma



Fonte: (KASSAMBARA, 2017, p. 12).

O resultado dos *clusters* pode ser obtido cortando-se o dendrograma em diferentes distâncias de acordo com o número de *clusters* n desejado. Neste trabalho essa decisão foi tomada através de método de *Elbow*, que geralmente é considerado como um indicador do número apropriado de *clusters*. Para (KODINARIYA; MAKWANA, 2013) esse é o método mais antigo para determinar o verdadeiro número de *clusters* em um conjunto de dados e é deslealmente chamado de método do cotovelo.

2.2.3.1 Métodos Aglomerativos

De acordo com (REIS, 2001) inicia-se com a quantidade de observações sendo a mesma quantidade de grupo, isto é, no início cada observação é um grupo. Inicialmente, as observações mais similares são agrupadas e fundidas, formando um único grupo, o processo é repetido com decréscimo da similaridade, formando um único grupo com todas as observações.

A seguir são apresentados diversos métodos de agrupamento que fazem parte do método aglomerativo e possuem muita aplicação prática, (AAKER; KUMAR; DAY, 2008).

- Métodos de ligação (*single linkage*, *complete linkage*, *average linkage*, *median linkage*);
- Método do centroide;
- Método da soma de erros quadráticos ou variância (método de *Ward*).

1. ***Single linkage*** ou (Ligação simples): Também denominado de método do vizinho mais próximo, é empregada a distância de valor mínimo:

$$d_{(IJ)W} = \min(d_{(IW)}, d_{(JW)}) \quad (2.7)$$

onde $d_{(IW)}$ e $d_{(JW)}$ são as distâncias entre os elementos IW e JW , respectivamente.

De acordo com (ANDERBERG, 2014), existem algumas características para este método:

- Em geral, grupos muito próximos podem não ser identificados;
- Permite detectar grupos de formas não-elípticas;
- Apresenta pouca tolerância a ruído, pois tem tendência a incorporar os ruídos em um grupo já existente;
- Apresenta bons resultados tanto para distâncias Euclidianas quanto para outras distâncias;
- Tendência a formar longas cadeias (encadeamento).

Para (ROMESBURG, 2004), encadeamento é um termo que descreve a situação onde há um primeiro grupo de um ou mais elementos que passa a incorporar, a cada iteração, um grupo de apenas um elemento. Assim, é formada uma longa cadeia, onde torna-se difícil definir um nível de corte para classificar os elementos em grupos.

2. **Complete linkage** ou (Ligação completa): Este método é bem parecido com o anterior, entretanto a distância entre os grupos é tomada como a máxima distância, sendo um dos métodos mais utilizados na análise de *cluster*, (GAMA, 1980). O método é definido pela expressão a seguir:

$$d_{(IJ)W} = \max(d_{(IW)}, d_{(JW)}) \quad (2.8)$$

onde $d_{(IW)}$ e $d_{(JW)}$ são as distâncias entre os elementos IW e JW , respectivamente. Segundo (KAUFMAN; ROUSSEEUW, 2009), este método possui as seguintes características:

- Apresenta bons resultados tanto para distâncias euclidianas quanto para outras distâncias;
 - Tendência a formar grupos compactos;
 - Os ruídos demoram a serem incorporados ao grupo.
3. **Average linkage** ou (ligação por média): Neste método a distância entre os grupos é tomada como sendo a média das distâncias entre os elementos, sua expressão é dada por:

$$d_{(IJ)W} = \frac{(N_I d_{(IW)} + N_J d_{(JW)})}{N_I + N_J} \quad (2.9)$$

onde: N_I e N_J são os números de elementos no grupo I e J , respectivamente; $d_{(IW)}$ e $d_{(JW)}$ são as distâncias entre os elementos IW e JW , respectivamente.

(KAUFMAN; ROUSSEEUW, 2009) destacam algumas características desse método:

- Menor sensibilidade à ruídos que o os métodos de ligação por vizinho mais próximo e por vizinho mais distante;
 - Apresenta bons resultados tanto para distâncias euclidianas quanto para outras distâncias;
 - Tendência a formar grupos com número de elementos similares.
4. **Median linkage** ou (Ligação por mediana): Com o método de ligação por mediana, a distância entre dois *clusters* é a distância mediana entre uma observação em um *cluster* e uma observação no outro *cluster*, sua expressão é dada por:

$$d_{(IJ)W} = \frac{(d_{(IW)} + d_{(JW)})}{2} - \frac{d_{(IJ)}}{4} \quad (2.10)$$

onde: $d_{(IW)}$, $d_{(JW)}$ e $d_{(IJ)}$ são as distâncias entre os elementos IW e JW e IJ , respectivamente.

De acordo com (DONI, 2004), esse método apresenta as seguintes características:

- Apresenta resultado satisfatório quando os grupos possuem tamanhos diferentes;
- Pode apresentar resultado diferente quando permutado os elementos na matriz de similaridade;
- Robustez na presença de outliers.

5. **Método da centroide:** Com o método da centroide a distância entre dois *clusters* é a distância entre os centroides ou meios do *cluster*, sua expressão é dada por:

$$d_{(IJ)W} = \frac{N_I d_{(IW)} + N_J d_{(JW)}}{N_I + N_J} - \frac{N_I N_J d_{(IJ)}}{(N_I + N_J)^2} \quad (2.11)$$

onde: N_I e N_J são os números de elementos no grupo I e J , respectivamente; $d_{(IW)}$, $d_{(JW)}$ e $d_{(IJ)}$ são as distâncias entre os elementos IW , JW e IJ , respectivamente.

Como características desse método, encontram-se, (KAUFMAN; ROUSSEEUW, 2009):

- Robustez à presença de ruídos;
- Fenômeno da reversão.

6. **Método de Ward** ou (Método da soma de erros quadráticos ou variância): No método ligação de *Ward*, a distância entre dois agrupamentos é a soma dos desvios quadrados dos pontos aos centroides. O objetivo da ligação de *Ward* é minimizar a soma dos quadrados dos catetos dentro do agrupamento. A função distância é dada por:

$$d_{(IJ)W} = \frac{((N_W + N_I)d_{(IW)} + (N_W + N_J)d_{(JW)} - N_W d_{(IJ)})}{N_W + N_I + N_J} \quad (2.12)$$

onde: N_I , N_J e N_W são os números de elementos no grupo I , J e W , respectivamente; $d_{(IW)}$, $d_{(JW)}$ e $d_{(IJ)}$ são as distâncias entre os elementos IW , JW e IJ , respectivamente.

(ROMESBURG, 2004) destaca as seguintes características desse método:

- Apresenta bons resultados tanto para distâncias euclidianas quanto para outras distâncias;

- Pode apresentar resultados insatisfatórios quando o número de elementos em cada grupo é praticamente igual;
- Tem tendência a combinar grupos com poucos elementos;
- Sensível à presença de outliers.

2.2.3.2 Métodos Divisivos

Os métodos divisivos são o inverso dos métodos aglomerativos, um único grupo é formado inicialmente com todas as observações e este grupo é subdividido em dois grupos de tal forma que exista o máximo de semelhança entre as observações do mesmo grupo e máxima dissimilaridade entre os subgrupos. O processo é repetido até que haja tantos grupos quanto a quantidade de observações (MALHOTRA, 2001).

Os métodos divisivos são pouco mencionados na literatura, pois exigem uma maior capacidade computacional que os métodos aglomerativos (KAUFMAN; ROUSSEEUW, 2009).

2.2.4 Métodos não hierárquicos ou por Particionamento

Os métodos não hierárquicos da análise de *cluster* são caracterizados pela necessidade de definir uma partição inicial, dividem a base de dados em K -grupos, onde o número K é a quantidade de grupos definida previamente. Em comparação com o método hierárquico, método não hierárquico é mais rápido, pois não precisa calcular e armazenar, durante o processamento, (DONI, 2004).

Para (REIS, 2001) as etapas usadas neste método são:

1. Começam por uma partição inicial dos indivíduos por um número de *cluster*, predefinido pelo analista; calculam, para cada *cluster*, o respectivo centroide;
2. Calculam as distâncias entre cada indivíduo e os centroides dos vários grupos; transferem cada indivíduo para o *cluster* relativamente ao qual se encontra a uma menor distância (por exemplo, distância euclidiana);
3. Calculam os novos centroides de cada *cluster*;
4. Repetem os passos 2 e 3 até que todos os indivíduos se encontrem em *clusters* estabilizados e não seja possível efetuar mais transferências de indivíduos de um *cluster* para outro.

A seguir tem-se os principais métodos por particionamento. Os métodos por particionamento mais conhecidos são o método *k-means* (k-médias) e o método *k-medoid* (k-medóides), que são descritos a seguir.

2.2.4.1 Método *k-means*

O agrupamento *K-means* é mais comumente usado para particionar um determinado conjunto de dados em um conjunto de k grupos (k *clusters*), em que k representa o número de grupos pré-especificado. Ele classifica objetos em vários grupos, de modo que objetos dentro do mesmo *cluster* sejam os mais similares possíveis, enquanto objetos de diferentes *clusters* são tão diferentes quanto possível (REIS, 2001).

Segundo (FUNG, 2001), o método *k-means* é um dos métodos mais utilizado das técnicas particionais. Diferentemente dos métodos hierárquicos, o *k-means* não cria uma estrutura em árvore para descrever o agrupamento dos dados e é mais adequado para uma grande quantidade de dados.

Algumas características desse método são (DONI, 2004):

- Sensibilidade a ruídos, uma vez que um elemento com um valor extremamente alto pode distorcer a distribuição dos dados;
- Tendência a formar grupos esféricos;
- O número de grupos é o mesmo durante todo o processo;
- Inadequado para descobrir grupos com formas não convexas ou de tamanhos muito diferentes.

2.2.4.2 Método *k-medoid*

O método *k-medoid* utiliza o valor médio dos elementos em um grupo como um ponto referência, chamado de medóide. Esse é o elemento mais centralmente localizado em um grupo (DONI, 2004). Os *medoids* são objetos representativos de cada agrupamento e contêm as características nas quais a dissimilaridade média dos objetos ou cisternas pertencentes a um dado grupo é mínima (VALE, 2005).

Para (DONI, 2004) algumas características desse método são:

- Independente da ordem os resultados serão os mesmos;
- Tendência a encontrar grupos esféricos;
- Processamento mais custoso que o *k-means*;
- Não aplicável à grandes bases de dados, pois o custo de processamento é alto;
- Mais robusto do que o *k-means* na presença de ruídos porque o medóide é menos influenciado pelos ruídos do que a média.

De acordo (KAUFMAN; ROUSSEEUW, 2009), uma forma de otimizar o método *k-medoid* para grandes bases de dados é considerar uma porção dos dados como uma

amostra representativa, e escolher os medóides dessa amostra. Se a amostra é selecionada aleatoriamente, ela deverá representar bem o conjunto de dados originais, apresentando bons resultados.

2.3 Material e Métodos

A metodologia utilizada para a obtenção dos dados foi a técnica presencial, foram utilizados os bancos de dados das pesquisas de preços de combustíveis que é realizada mensalmente pelo Fundo Municipal de Defesa dos Direitos Difusos (PROCON) de Campina Grande - PB, relativas ao ano de 2019. O conjunto de dados é composto por 57 postos de combustíveis distribuídos na cidade, suas respectivas variáveis são: posto de combustível, bandeira, bairro, gasolina comum e etanol.

Após o levantamento dos dados, iniciou-se uma minuciosa observação para verificar a presença de dados faltantes e supostos erros no armazenamento. Como em toda pesquisa antes de iniciar as análises deve-se sempre verificar o banco de dados. Assim, em seguida começou a extração de informações com análises descritivas e construções de gráficos para compreender os dados, depois deste processo iniciou-se as análises de *cluster* começando pelo método não hierárquico, utilizando o método *k-means*, depois dos *clusters* formados iniciou uma análise descritiva para observar o comportamento dos preços em cada um dos *clusters*. Já no método hierárquico utilizou-se a matriz de distância euclidiana para em seguida ser utilizada na construção do dendrograma, para construção do mesmo foi utilizada a *complete linkage* ou (ligação completa), assim visualizando os postos de combustíveis em cada *clusters*.

Os procedimentos estatísticos, cálculos e gráficos executados neste trabalho foram realizados no *software* R, que é atualmente uma das ferramentas mais utilizada para análises estatísticas o mesmo pode ser obtido gratuitamente em <https://www.r-project.org/>. As análises foram feitas utilizando vários pacotes, entre eles podemos citar: *sjstats* (2020) coleção de funções convenientes para cálculos estatísticos comuns, *cluster* (2019) para análise de cluster, *factoextra* (2019) visualizar a saída de análises de dados multivariadas, *ClustOfVar* (2017) é para análise de cluster de um conjunto de variáveis, *ggdendro* (2020) cria dendrogramas e diagramas de árvore usando 'ggplot2', entre outros.

3 RESULTADOS E DISCUSSÕES

Neste trabalho, primeiramente realizou-se um estudo exploratório para caracterizar o comportamento e um resumo dos preços no ano de 2019 para a gasolina comum e o etanol. Na sequência, foram aplicadas as técnicas de estatística multivariada (Análise de *Cluster* não Hierárquica e Hierárquica) que permitiram comparar os diferentes preços entre os postos de combustíveis. Em seguida, realizou-se uma análise descritiva para cada um dos grupos verificando o comportamento dos preços nos *clusters*, posteriormente aplicou-se o método hierárquico sendo representados através do dendrograma.

3.1 Análise Exploratória

A tabela a seguir apresenta algumas estatísticas descritivas obtidas para os preços das variáveis gasolina comum e etanol, no intuito de se fazer um levantamento sobre algumas características importantes.

Tabela 1 – Sumário das estatísticas para as variáveis

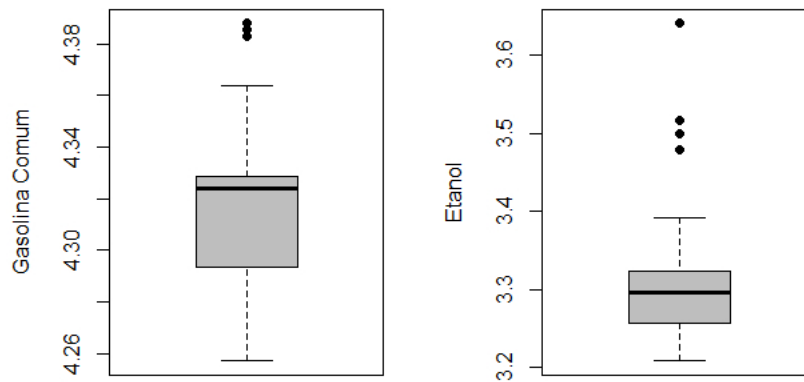
Estatísticas	Gasolina Comum	Etanol
Mínimo	4,257	3,209
1° Quartil	4,294	3,257
Mediana	4,324	3,297
Média	4,317	3,310
3° Quartil	4,328	3,324
Máximo	4,388	3,641

Fonte: Elaborada pelo autor, 2021.

De acordo com os boletins trimestrais da (PETRÓLEO, 2019) (ANP), a gasolina comum no país teve o preço médio de revenda de 4,379, enquanto na cidade de Campina Grande este preço teve média de 4,317. Assim, podemos observar que os valores não estão muito distantes um do outro, mas vale salientar que no município de Campina Grande - PB o preço esteve mais baixo do que o apurado pela ANP. Já para o etanol pode-se observar uma diferença entre os preços médios, onde de acordo com a (PETRÓLEO, 2019) o preço de revenda do combustível estava custando 2,901, enquanto no município estava 3,310, ou seja, uma diferença de 0,409 centavos.

Na figura a seguir, utilizou-se o gráfico de Boxplot, podendo observar o comportamento dos dados para a gasolina comum e o etanol.

Figura 2 – Boxplot para gasolina comum e etanol

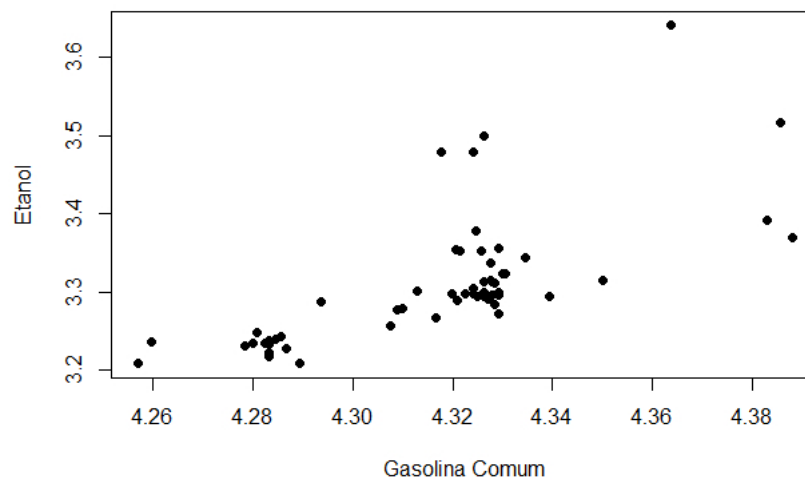


Fonte: Produzido pelo autor, 2021.

Visualmente, como o valor da mediana não está bem centrado em ambos os casos, tem-se indícios de falta de simetria em ambas as variáveis. É possível ainda distinguir alguns postos de combustíveis com o preço discrepante (pontos fora do boxplot) tanto para a gasolina comum quanto para o etanol, para gasolina são eles: JE, Maia e Dallas do Ligeiro. No caso do etanol o que chama atenção é o Vieira que está muito acima dos demais, esses valores discrepantes para o etanol podem ter interferido na diferença de preços entre a ANP e o preço da cidade.

A partir dos valores dos combustíveis, a Figura 3 apresenta o gráfico de dispersão entre estas duas variáveis.

Figura 3 – Diagrama de dispersão



Fonte: Produzido pelo autor, 2021.

Observando o gráfico há indício que se tem uma relação entre as variáveis, mas

para saber se isso realmente acontece e ver o quanto elas são correlacionadas é preciso calcular o coeficiente de correlação.

A Tabela 2 mostra o resultado do teste de normalidade de Shapiro - Wilk, afim de verificar a normalidade dos dados. Se o valor-p do teste for menor do que o nível de significância (0,05), conclui-se que a variável não segue distribuição normal.

Tabela 2 – Resultados do teste de normalidade (Shapiro - Wilk)

Combustível	W	Valor p
Gasolina Comum	0,90	< 0,01
Etanol	0,83	0,317

Fonte: Elaborada pelo autor, 2021.

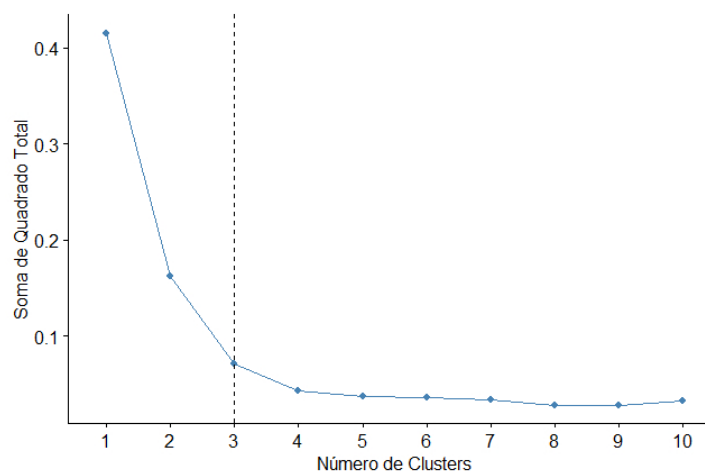
O teste de Shapiro - Wilk mostrou que apenas a variável etanol segue normalidade, já que ao nível de 5% de significância H_0 não é rejeitada. Tal fato não ocorre com a variável gasolina comum. Diante disso, como apenas o etanol segue distribuição normal, será calculado o coeficiente de correlação de Spearman.

Neste caso, o coeficiente ρ de Spearman é o mais apropriado para medir a intensidade da associação entre as variáveis. A correlação de Spearman foi estimada por $\hat{\rho} = 0,708$, ou seja, existe uma associação positiva entre a gasolina comum e o etanol.

3.2 Método não hierárquico

Obtendo o número de *clusters* através de método de *Elbow* (cotovelo), que se observa através do *scree plot*, por esse método a localização de uma curva (cotovelo) no gráfico é geralmente considerada como um indicador do número apropriado de *clusters*.

Figura 4 – Números de *clusters* através do *scree plot*

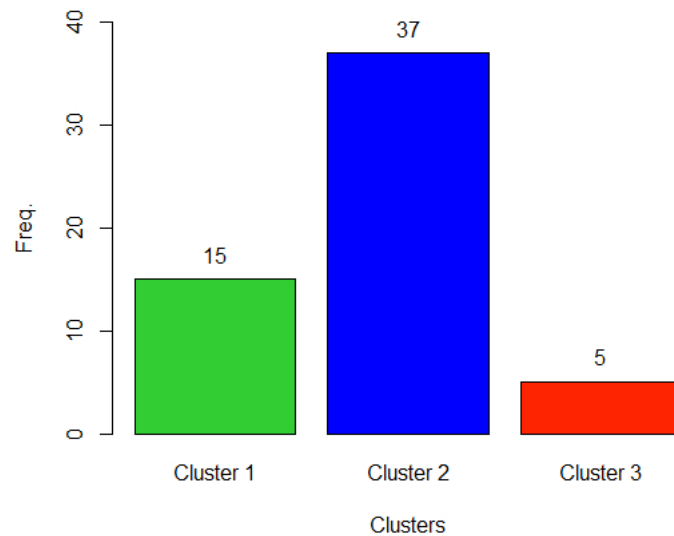


Fonte: Produzido pelo autor, 2021.

Observando o *scree plot* podemos verificar uma desaceleração no decaimento, assim o número ideal de *clusters* segundo o método de *Elbow* são 3 (três).

Utilizando o método *K-means* podemos observar a quantidade de postos de combustíveis em cada *cluster*, observe a Figura 5.

Figura 5 – Frequência de postos por *cluster*



Fonte: Produzido pelo autor, 2021.

Observa-se que o *cluster* 1 ficou um total de 15 postos de combustíveis, o *cluster* 2 com 37 e por último o *cluster* 3 com 5, totalizando 57 postos de combustíveis.

No intuito de se fazer um levantamento sobre algumas características importantes, a Tabela 3 apresenta algumas estatísticas descritivas obtidas para os *clusters* em relação a gasolina comum.

Tabela 3 – Sumário das estatísticas para gasolina comum por *clusters*

Estatísticas	cluster 1	cluster 2	cluster 3
Mínimo	4,257	4,294	4,318
Mediana	4,283	4,327	4,326
Média	4,282	4,328	4,343
Desvio padrão	0,012	0,017	0,030
CV	0,003	0,004	0,007
Máximo	4,308	4,388	4,386

Fonte: Elaborada pelo autor, 2021.

Observando as estatísticas para cada *cluster* o que chama atenção é o *cluster* 3 que tem um total de 5 postos de combustíveis e teve a maior média de preço, indicando que esses 5 postos vendiam a gasolina comum com os preços mais elevados durante o ano de

2019. Os valores da média indicam que os postos com os melhores preços são os do *cluster* 1 (um total de 15 postos), pois apresenta a menor média de preço para a gasolina comum, o *cluster* 2 com um total de 37 tem a média entre o *cluster* 1 e 3.

A Tabela 4 apresenta algumas estatísticas descritivas obtidas para os *clusters* em relação a o etanol, no intuito de se fazer um levantamento sobre algumas características importantes.

Tabela 4 – Sumário das estatísticas para etanol por *clusters*

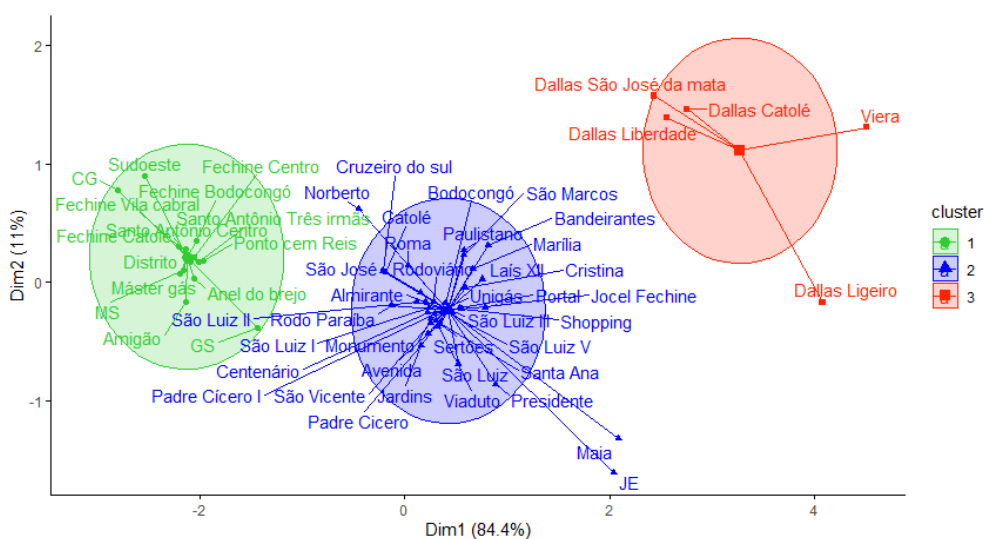
Estatísticas	cluster 1	cluster 2	cluster 3
Mínimo	3,209	3,268	3,478
Mediana	3,235	3,299	3,499
Média	3,232	3,312	3,523
Desvio padrão	0,013	0,031	0,068
CV	0,004	0,009	0,019
Máximo	3,257	3,391	3,641

Fonte: Elaborada pelo autor, 2021.

Portanto, o que foi observado para a gasolina comum acontece com o etanol, o *cluster* 3 tem o maior preço médio para o etanol, além dos maiores valores para todas as estatísticas apresentadas, como para a gasolina comum o menor preço médio se encontra no *cluster* 1.

A Figura 6 apresenta graficamente os *clusters* e onde está localizado cada um dos postos de combustíveis.

Figura 6 – Localização dos postos nos *clusters*



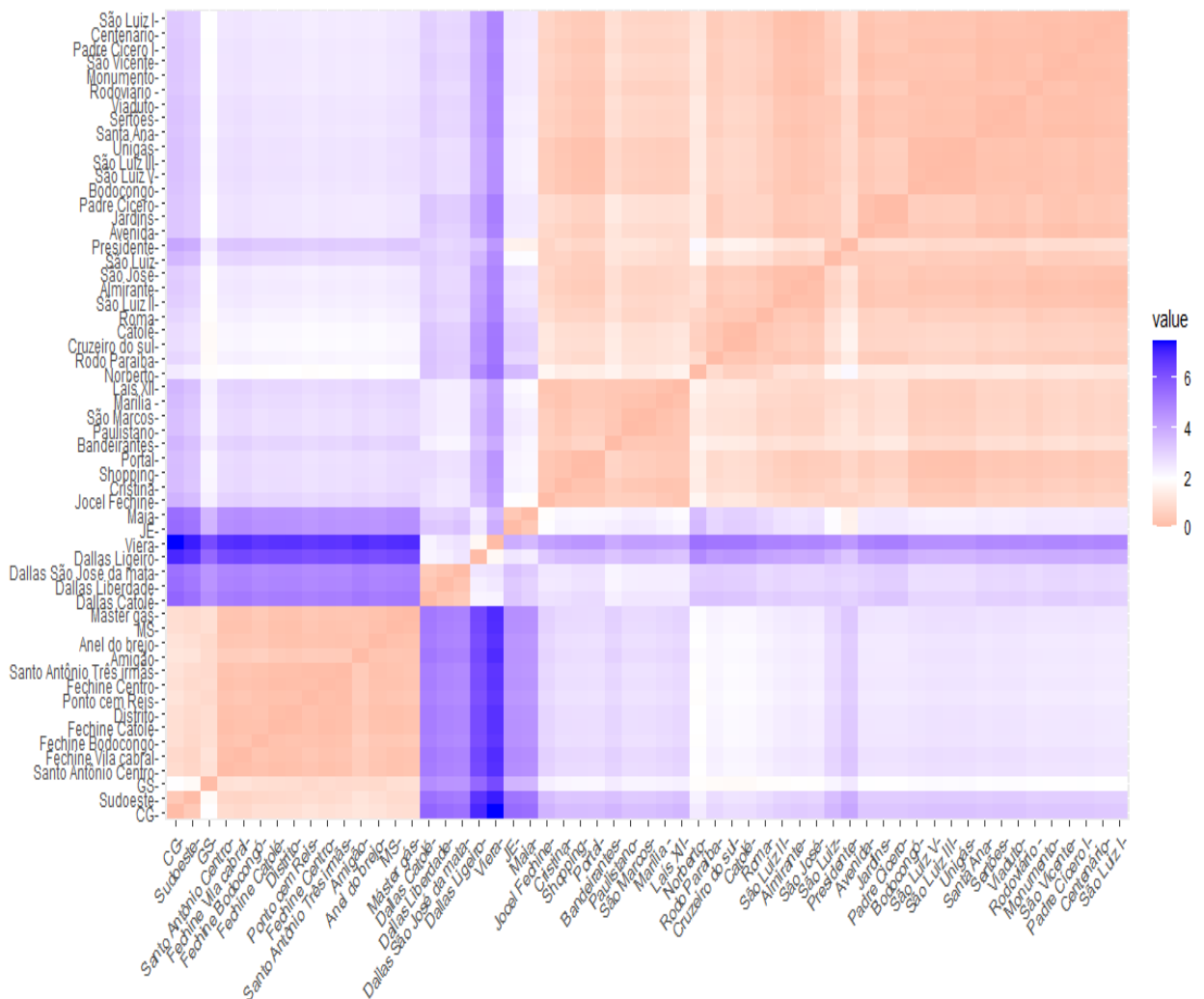
Fonte: Produzido pelo autor, 2021.

A Figura 6 indica a localização de cada posto de combustível nos *clusters*. Caso uma pessoa queira abastecer seu veículo com um desses dois combustíveis ela iria optar pelo os postos que estão no *cluster* 1 onde estão os preços mais acessíveis. Pode-se observar que tem alguns postos do *cluster* 2 que ficam bem próximo ao *cluster* 1.

3.3 Método Hierárquico

A Figura 7 apresenta graficamente a matriz de distância, usando o método da distância euclidiana para calcular a matriz de dissimilaridade.

Figura 7 – Matriz de distância



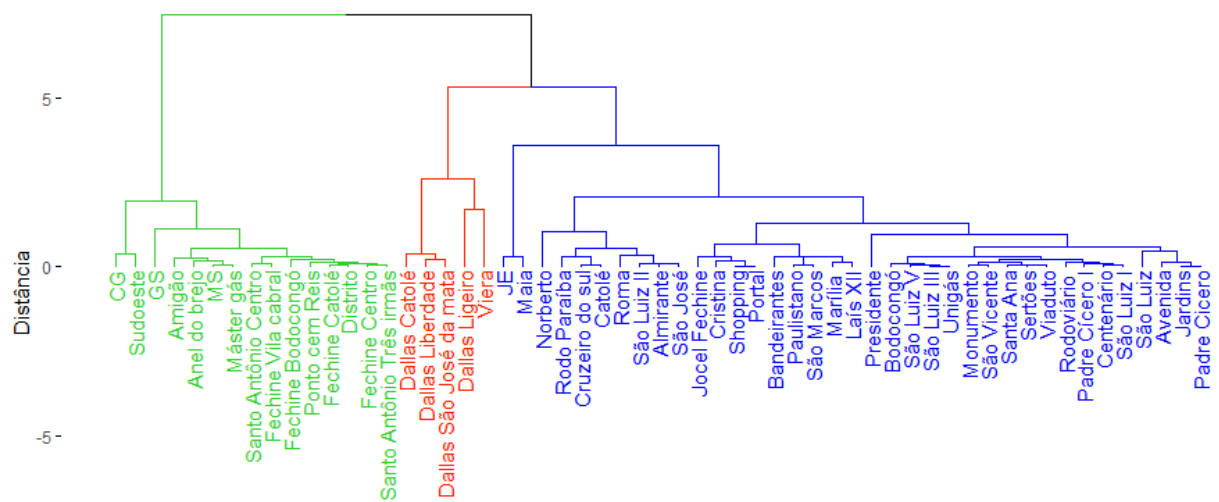
Fonte: Produzido pelo autor, 2021.

Analisando o gráfico é possível visualizarmos os contornos dos *clusters*, quanto mais perto de zero maiores similaridades terão os postos de combustíveis, observando a diagonal

podemos verificar isso, pois as distâncias estão sendo calculadas com os mesmos postos, por isso o valor 0 e quanto mais distante de zero mais dissimilares são os postos.

A Figura 8 apresenta o dendrograma evidenciando os *clusters* formados pelo método complete linkage (ligação completa), seguindo o mesmo critério utilizado para obter o número de *cluster* no método não hierárquico, ou seja, formando 3 *clusters*.

Figura 8 – Dendrograma



Fonte: Produzido pelo autor, 2021.

Através do dendrograma, verificou-se onde os postos se encontram em cada *clusters*, para além disso pode-se verificar que os *clusters* são formados pelos mesmos postos de combustíveis em ambos os métodos.

4 CONCLUSÃO

O objetivo deste trabalho foi identificar a similaridade dos preços da gasolina comum e etanol nos postos de combustíveis na cidade de Campina Grande - PB. Os resultados indicaram que existem 3 *clusters* alocando os postos com os preços mais similares. O método hierárquico utilizou a distância euclidiana e a ligação completa, no não hierárquico utilizou o método *k-means*, ou seja, isso indica que a análise de *cluster* foi bem empregada neste estudo, pois em ambos os métodos não houve mudança dos postos de *clusters*.

De maneira geral, após todas as análises feitas, pode-se concluir que o *cluster* 1 com um total de 15 postos de combustíveis tem os preços mais acessíveis para a gasolina comum e o etanol, já o *cluster* 3 tem os preços mais elevados, ou seja, os 5 postos que fazem parte deste *cluster* tinham os preços mais caros no ano de 2019, por fim o *cluster* 2 composto por 37 postos que esteve com os preços entre os *clusters* 1 e 3.

Acredita-se que o seguinte ponto é interessante de ser explorado, o Fundo Municipal de Defesa dos Direitos Difusos (PROCON) de Campina Grande - PB poderia implementar a análise de *cluster* nas suas pesquisas de preços de combustíveis que é realizada mensalmente, ajudando a população a identificar onde se encontra os postos com os preços mais acessíveis, assim fazendo-os economizar com os gastos no abastecimento da gasolina comum e do etanol.

REFERÊNCIAS

- AAKER, D. A.; KUMAR, V.; DAY, G. S. *Marketing research*. [S.l.]: John Wiley & Sons, 2008. Citado na página 16.
- ALBUQUERQUE, M. A. d. Estabilidade em análise de agrupamento. *Recife, PE*, 2005. Citado na página 11.
- ANDERBERG, M. R. *Cluster analysis for applications: probability and mathematical statistics: a series of monographs and textbooks*. [S.l.]: Academic press, 2014. v. 19. Citado na página 16.
- BARROSO, L. P.; ARTES, R. Análise multivariada. *Lavras: Ufla*, p. 151, 2003. Citado na página 13.
- BUSSAB, W. d. O.; MIAZAKI, E. S.; ANDRADE, D. F. *Introdução à análise de agrupamentos*. [S.l.]: Ime-Usp, 1990. Citado na página 11.
- CORMACK, R. M. A review of classification. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 134, n. 3, p. 321–353, 1971. Citado na página 14.
- CRISPIM, D. L.; FERNANDES, L. L.; ALBUQUERQUE, R. d. O. Aplicação de técnica estatística multivariada em indicadores de sustentabilidade nos municípios do marajó-pa. *Revista Principia-Divulgação Científica e Tecnológica do IFPB*, (46), p. 145–154, 2019. Citado na página 14.
- DONI, M. V. Análise de cluster: métodos hierárquicos e de particionamento. *Universidade Presbiteriana Mackenzie*, 2004. Citado 6 vezes nas páginas 11, 13, 15, 18, 19 e 20.
- FUNG, G. A comprehensive overview of basic clustering algorithms. 2001. Citado na página 20.
- GAMA, M. d. P. Bases da análise de agrupamentos (“cluster analysis”). *Brasília, UNV*, 1980. Citado na página 17.
- HAIR, J. F. et al. *Análise multivariada de dados*. [S.l.]: Bookman editora, 2009. Citado na página 12.
- KASSAMBARA, A. *Practical guide to cluster analysis in R: Unsupervised machine learning*. [S.l.]: Sthda, 2017. v. 1. Citado na página 15.
- KAUFMAN, L.; ROUSSEEUW, P. J. *Finding groups in data: an introduction to cluster analysis*. [S.l.]: John Wiley & Sons, 2009. v. 344. Citado 4 vezes nas páginas 17, 18, 19 e 20.
- KODINARIYA, T. M.; MAKWANA, P. R. Review on determining number of cluster in k-means clustering. *International Journal*, v. 1, n. 6, p. 90–95, 2013. Citado na página 16.
- MALHOTRA, N. K. *Pesquisa de Marketing-: Uma Orientação Aplicada*. [S.l.]: Bookman Editora, 2001. Citado 2 vezes nas páginas 12 e 19.
- MINGOTI, S. A. Análise de dados através de métodos estatística multivariada: uma abordagem aplicada. In: *Análise de dados através de métodos estatística multivariada: uma abordagem aplicada*. [S.l.: s.n.], 2007. p. 295–295. Citado na página 15.

- PETRÓLEO, G. N. e. B. Agência Nacional do. *Cartilha do posto revendedor de combustíveis*. 2017. Citado na página 11.
- PETRÓLEO, G. N. e. B. Agência Nacional do. Boletim trimestral de preços e volumes de combustíveis. 2019. Citado na página 22.
- REIS, E. Estatística multivariada aplicada. *Edições Sílabo*, 2001. Citado 6 vezes nas páginas 12, 13, 15, 16, 19 e 20.
- ROMESBURG, C. *Cluster analysis for researchers*. [S.l.]: Lulu. com, 2004. Citado 2 vezes nas páginas 17 e 18.
- STANTON, J. M. Galton, pearson, and the peas: A brief history of linear regression for statistics instructors. *Journal of Statistics Education*, Taylor & Francis, v. 9, n. 3, 2001. Citado na página 13.
- UCHÔA, C. F. A. Testando a assimetria nos preços da gasolina brasileira. *Revista Brasileira de Economia*, SciELO Brasil, v. 62, n. 1, p. 103–117, 2008. Citado na página 11.
- VALE, M. N. do. *Agrupamentos de dados: Avaliação de Métodos e Desenvolvimento de Aplicativo para Análise de Grupos*. Tese (Doutorado) — PUC-Rio, 2005. Citado na página 20.
- ZAIANE, O. R. et al. On data clustering analysis: Scalability, constraints, and validation. In: SPRINGER. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. [S.l.], 2002. p. 28–39. Citado na página 12.