INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Autarquia Associada à Universidade de São Paulo

Algoritmos matemáticos aplicados em resultados experimentais

ANDRÉ LUIZ NOGUEIRA

Tese apresentada como parte dos requisitos para obtenção do Grau de Doutor em Ciências na Área de Tecnologia Nuclear - Aplicações

Orientador:
Prof. Dr. Casimiro Sepúlveda Munita

São Paulo
2023

INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Autarquia Associada à Universidade de São Paulo

Algoritmos matemáticos aplicados em resultados experimentais

Versão Corrigida
Versão Original Disponível no IPEN

ANDRÉ LUIZ NOGUEIRA

Tese apresentada como parte dos requisitos para obtenção do Grau de Doutor em Ciências na Área de Tecnologia Nuclear – Aplicações

Orientador:
Prof. Dr. Casimiro Sepúlveda Munita

São Paulo
2023

Como citar:

NOGUEIRA, A. L. ***Algoritmos matemáticos aplicados em resultados experimentais***. 2023. 147 f. Tese (Doutorado em Tecnologia Nuclear), Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN, São Paulo. Disponível em: ‹http://repositorio.ipen.br/› (data de consulta no formato: dd/mm/aaaa)

Ficha catalográfica elaborada pelo Sistema de geração automática da Biblioteca IPEN, com os dados fornecidos pelo(a) autor(a).

Nogueira, André Luiz
Algoritmos matemáticos aplicados em resultados experimentais / André Luiz Nogueira; orientador Casimiro Sepúlveda Munita. -- São Paulo, 2023.
147 f.

Tese (Doutorado) - Programa de Pós-Graduação em Tecnologia Nuclear (Aplicações) -- Instituto de Pesquisas Energéticas e Nucleares, São Paulo, 2023.

1. imputação. 2. outliers. 3. padronização. 4. análise de agrupamento. 5. software R. I. Munita, Casimiro Sepúlveda, orient. II. Título.

Janus

Universidade de São Paulo

ATA DE DEFESA

Aluno: 85131 - 5022077 - 1 / Página 1 de 1

Ata de defesa de Tese do(a) Senhor(a) André Luiz Nogueira no Programa: Tecnologia Nuclear, do(a) Instituto de Pesquisas Energéticas e Nucleares da Universidade de São Paulo.

Aos 23 dias do mês de março de 2023, no(a) Auditório Rui Ribeiro Franco realizou-se a Defesa da Tese do(a) Senhor(a) André Luiz Nogueira, apresentada para a obtenção do título de Doutor intitulada:

"Algoritmos matemáticos aplicados em resultados experimentais"

Após declarada aberta a sessão, o(a) Sr(a) Presidente passa a palavra ao candidato para exposição e a seguir aos examinadores para as devidas arguições que se desenvolvem nos termos regimentais. Em seguida, a Comissão Julgadora proclama o resultado:

Nome dos Participantes da Banca	Função	Sigla da CPG	Resultado
Casimiro Jaime Alfredo Sepúlveda Munita	Presidente	IPEN(IPEN)	Não Votante
Lucia Pereira Barroso	Titular	IME - USP	Aprovado
Mário Olímpio de Menezes	Titular	IPEN(IPEN)	Aprovado
José Osman dos Santos	Titular	Externo	Aprovado

Resultado Final: Aprovado

Eu, _____, lavrei a presente ata , que assino juntamente com os(as) Senhores(as) examinadores. São Paulo, aos 23 dias do mês de março de 2023.

Lucia Pereira Barroso

Mário Olímpio de Menezes

José Osman dos Santos

Casimiro Jaime Alfredo Sepúlveda Munita
Presidente da Comissão Julgadora

A defesa foi homologada pela Comissão de Pós-Graduação em ____/____/_____ e, portanto, o(a) aluno(a) faz jus ao título de Doutor em Ciências obtido no Programa Tecnologia Nuclear - Área de concentração: Tecnologia Nuclear - Aplicações.

_____
Presidente da Comissão de Pós-Graduação

AGRADECIMENTOS

Agradeço à minha esposa (Hamona), meus pais (Luiz Carlos e Alaide) e irmãos (Tiago e Aline) pelo apoio, incentivo e compreensão ao longo do doutoramento.

Agradeço ao Prof. Dr. Casimiro Sepúlveda Munita, pela orientação e dedicação durante toda a elaboração deste trabalho.

Aos colegas de trabalho do Instituto Federal de Sergipe, dentre eles destaco Prof. Dr. José Osman, Prof. Dr. Mauro José e Profa. Dra. Héstia, que me auxiliaram em momentos decisivos desta jornada.

Aos colegas do Grupo de Estudos Arqueológicos do IPEN-CNEN/SP, pela paciência e pelas conversas que, com certeza, me motivaram durante toda esta caminhada.

Aos professores do IPEN que cooperaram para a produção desse trabalho.

Aos funcionários do IPEN que contribuíram direta e indiretamente para produção deste trabalho.

São Paulo

2023

RESUMO

NOGUEIRA, A. L. ***Algoritmos matemáticos aplicados em resultados experimentais***. 2023. 147 p. Tese (Doutorado em Tecnologia Nuclear) – Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN/SP, São Paulo.

Este trabalho teve como objetivo avaliar o impacto da imputação de dados (métodos: média, *autoencoder*, análise de agrupamento e c-médias), avaliar os métodos de detecção de *outliers* (métodos: Mahalanobis e Mahalanobis robusta) e padronização de dados (transformadas z-score, mínimo-máximo, mínimo-máximo melhorada, logarítmica e Box-Cox) na análise de agrupamento, assim como identificar os métodos mais adequados para a base de amostras arqueológicas estudada. A base de dados foi fornecida pelo Grupo de Estudos Arqueológicos do IPEN-CNEN/SP, de modo que foram analisadas 140 amostras de fragmentos cerâmicos de três sítios arqueológicos. Para análise das amostras foram utilizados 13 elementos químicos: As, Na, Ce, Cr, Eu, Fe, Hf, La, Nd, Sc, Sm, Th e U. Os resultados mostraram que não houve impacto da imputação de dados nos métodos de agrupamento hierárquicos, particionais/crisp, c-médias e c-médias com polinômio fuzzificador. A exclusão dos *outliers* detectados pela distância Mahalanobis teve impacto no aumento da coesão entre as amostras dos sítios B e C. As transformadas utilizadas para padronização das amostras alteraram os valores da estatística de Hopkins, bem como as imagens VAT. As funções implementadas, desenvolvidas no software estatístico R, deram origem a uma aplicação web.

Palavras-chave: imputação; outliers; padronização; análise de agrupamento; software R; aplicação web.

São Paulo

2023

ABSTRACT

NOGUEIRA, A. L. **Mathematical algorithms applied to experimental results.** 2023. 147 p. Tese (Doutorado em Tecnologia Nuclear) – Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN/SP, São Paulo.

This thesis aimed to evaluate the impact of data imputation (methods: mean, autoencoder, clustering and c-means), evaluate the methods of outliers detection (methods: Mahalanobis and robust Mahalanobis), and data standardization (transforms: z-score, min-max, min-max improved, logarithmic and Box-Cox) in cluster analysis, as well as to identify the most suitable method to the test basis for the archaeological sampling researched. The basis was provided by the Group of Archaeological Studies from IPEN-CNEN/SP, so were analysed 140 samples of pottery fragments from three archaeological sites. For sample analysis, were used 13 chemical elements: As, Na, Ce, Cr, Eu, Fe, Hf, La, Nd, Sc, Sm, Th and U. The results showed that there was no impact of data imputation on the hierarchical clustering methods, crisp partitions, c-means and c-means with fuzzifier polynomial. The exclusion of outliers detected by Mahalanobis distance had the impact of increasing cohesion between the samples of sites B and C. The transforms used to standardize the samples changed the Hopkins coefficient value as well as VAT images. The implemented functions, developed in the R statistical software, resulted in a web application.

Keywords: imputation; outliers; standardization; cluster analysis; R software; web application.

LISTA DE SIGLAS

CCC – Coeficiente de Correlação Cofenética

CH – Calinski-Harabasz

CP – Compacidade

d – Número de características das amostras

DAE – Denoising Autoencoder

DM – Distância Mahalanobis

DMR – Distância Mahalanobis Robusta

EQ – Erros Quadrados

H – Estatística de Hopkins

k – Número de grupos da base de dados

La – Lantânio

log – transformada logarítmica

MD – Matriz Dissimilaridade

n – Número de amostras da base de dados

Na – Sódio

NRMSE – Normalized Root Mean Squared Error

PCA – Principal Component Analysis

VAT – Visual Assessment of Tendency

RL – Ratkowsky-Lance

SE – Separabilidade

Sm – Samário

$V(c_r)$ – Variação interna do grupo (c_r)

VT – Variação Total

WG – Wemmert-Gancarski

SUMÁRIO

1 INTRODUÇÃO

Muitos dos problemas que, hoje, chamamos de estatísticos surgiram e foram resolvidos nos primeiros períodos da história. Impostos, serviço militar, taxas alfandegárias deram origem à investigação e aos registros de caráter estatístico (MEITZEN e FALKNER, 1891).

Há mais de 2.000 anos, Confúcio relatou levantamentos feitos na China. No renascimento, houve o interesse da coleta de dados estatísticos por suas aplicações na administração pública. Entretanto, a primeira tentativa de se extrair conclusões a partir de dados numéricos ocorreu somente no século XVII, com a chamada Aritmética Política, que evoluiu para a chamada demografia (MEMÓRIA, 2004).

A Estatística tem evoluído de uma simples compilação de números até uma poderosa ferramenta para estudar relações de causalidade entre os fenômenos. Desse modo, a história da estatística pode ser dividida em três grandes períodos (CASTRO, 1975):

1º Período: vai desde o regime Feudal até meados do século XVII, em que é descrito pela organização de informação e cadastros do interesse do Estado, com objetivo fiscal ou de guerra;

2º Período: inicia-se em meados do século XVII e vai até a metade do século XIX, associado à preparação das teorias. Nesse período, ocorreu o estabelecimento da Estatística como uma disciplina autonôma;

3º Período: tem início em 1853 com o primeiro congresso de Estatística e se estende até hoje, um período de aperfeiçoamento, no qual o método estatístico vem sendo aplicado em diversas áreas, com grande intercâmbio de informações e ideias, além do desenvolvimento da Estatística como um método indicado para pesquisar relações de causa e efeito.

Nesse sentido, nos últimos anos houve um aumento significativo da quantidade de informação disponível, de maneira que a análise de grande parte dessa informação requer a utilização de técnicas estatísticas multivariadas. Contribuiu para esse aumento o crescente avanço das técnicas analíticas em

várias áreas da ciência. Para a interpretação desses resultados, faz-se necessário o uso de métodos estatísticos cada vez mais sofisticados, tais como as técnicas multivariadas. A análise multivariada abrange todas as técnicas estatísticas que analisam ao mesmo tempo múltiplas medidas associadas aos dados (HAIR et al., 2009).

As técnicas de estatística multivariada se dividem em dois tipos, a saber:

- técnicas exploratórias de simplificação, como análise de componentes principais, análise fatorial, análise de correlações canônicas, análise de agrupamentos (agrupamento), análise discriminante e análise de correspondência;
- técnicas de inferência estatística, por exemplo, testes de hipóteses, análise de variância, análise de covariância e regressão multivariada (MINGOTI, 2007).

A maioria das técnicas multivariadas, assim como, a análise de agrupamento envolve uma quantidade de cálculos considerável e sua utilização no dia a dia requer *softwares* ou pacotes. Muitas das análises de dados multivariados incorporam a construção de gráficos e diagramas apropriados, que podem ser feitos utilizando pacotes do *software* R (EVERITT e HOTHORN, 2011).

No tocante ao R (R CORE TEAM, 2022), este se trata de uma linguagem de programação e um ambiente de trabalho, ou seja, após a execução de comandos, os resultados são exibidos como texto ou gráficos. Além disso, o R é uma ferramenta poderosa e completa, bastante adaptada para métodos estatísticos (DE MICHEAUX et al., 2013), com milhares de pacotes produzidos por colaboradores, que auxiliam usuários nas mais diversas áreas (BEELEY, 2016; CARLSON, 2017). Entre as características do R, pode-se destacar: um ótimo sistema de documentação, uma vasta coleção de métodos para análise de dados e uma linguagem de programação simples e eficiente (DE MICHEAUX et al., 2013).

Muitas técnicas estatísticas clássicas e recentes estão implementadas em R. Métodos estatísticos mais avançados disponíveis através de pacotes podem ser instalados facilmente. Uma documentação detalhada dos pacotes está disponível em https://cran.r-project.org/. A comunidade de

usuários/desenvolvedores também adiciona regularmente métodos estatísticos recentes (DE MICHEAUX et al., 2013).

No final de 2012, o *RStudio* (RSTUDIO TEAM, 2021) lançou o *Shiny* (BEELEY, 2016; CHANG et al., 2022; GRANJON, 2022) para a comunidade R, depois disso, tornou-se um dos principais projetos da equipe do *RStudio* (FAY et al., 2021). O *Shiny* é um pacote para a criação de aplicações web usando a linguagem R. Além disso, tal pacote permite criar aplicativos web sem conhecimento de HTML, CSS ou JavaScritpt. Desde o seu lançamento, o número de usuários do *Shiny* cresceu rapidamente em quase todas as áreas (GRANJON, 2022).

No presente trabalho, realizou-se um estudo comparativo de quatro métodos de imputação de dados: média (LITTLE e RUBIN, 2019), rede neural autoencoder (AE) (ZHANG et al., 2020), agrupamento (SHI *et al*., 2020) e c-médias (NIKFALAZAR et al., 2017). Ademais, foram avaliados dois métodos de detecção de valores discrepantes (*outliers*) que utilizam as distâncias: Mahalanobis (DE MAESSCHALCK et al., 2000) e Mahalanobis robusta (LEYS et al., 2018). Por último, foram realizados testes com cinco transformadas para padronização dos dados: z-score (CASTRO e FERRARI, 2016), mínimo-máximo (MOHAMAD e USMAN, 2013), mínimo-máximo melhorada (KABIR. et al., 2016), logarítmica e Box-Cox (ATKINSON et al., 2021).

No decorrer deste trabalho, foram desenvolvidas funções do R para imputação, detecção de outliers e padronização de dados, além de funções de análise de agrupamento, responsáveis pela determinação dos grupos e sua visualização. Dessa forma, estas funções deram origem a uma aplicação *web*, interativa, de fácil uso, com uma interface leve e amigável, que proporciona a utilização de diversos métodos sem a necessidade de conhecimento em programação por parte do usuário. A aplicação foi desenvolvida utilizando o pacote *Shiny* no ambiente R, com intuito de operar como uma ferramenta de apoio para a análise de dados, podendo ser acessada a partir de qualquer navegador, sem a necessidade de instalação de *softwares* ou pacotes.

1.1 Estrutura da tese

Além da introdução, esta tese está organizada nas seguintes seções:

- Objetivos: informa o que se pretende desenvolver com o trabalho;
- Fundamentos teóricos dos métodos de agrupamento: trata dos métodos de agrupamento, tendência de agrupamento e índices de validação;
- Fundamentos teóricos dos métodos de imputação de dados: apresenta os métodos utilizados para imputação de dados;
- Fundamentos teóricos dos métodos de detecção de *outliers*: discorre sobre os métodos estudados para detecção de *outliers*;
- Fundamentos teóricos dos métodos de padronização de dados: trata das transformadas utilizadas para padronização dos dados;
- Resultados e discussão: expõe os resultados obtidos após aplicação dos métodos considerados. Discorre sobre os resultados obtidos, buscando determinar os métodos com melhor desempenho e em que situações isso ocorre;
- Conclusão: são destacados os principais pontos dos resultados e da discussão;
- Sugestões para trabalhos futuros: apresenta propostas de futuros desdobramentos sobre o assunto do trabalho.

Os tópicos abordados neste trabalho são expostos na Tabela 1.

Tabela 1: Tópicos abordados no trabalho

Análise de agrupamento	- métodos de agrupamento: hierárquicos (ligação simples, ligação média, ligação completa e Ward), particionais (k-médias, k-medóides e híbrido), baseados em lógica fuzzy (c-médias, c-medoides e c-médias com polinômio fuzzificador); - tendência de agrupamento: estatística de Hopkins e avaliação visual; - índices de validação: gama, Dunn, Calinski-Harabasz, PBM, Wemmert-Gancarski, Tau, Ratkowsky-Lance;
Métodos de imputação	- imputação pela média; - imputação por autoencoder; - imputação por agrupamento; - imputação pelo c-médias;
Métodos de detecção de outliers	- distância Mahalanobis; - distância Maralanobis robusta;
Métodos de padronização de dados	- transformada z-score - transformada mínimo-máximo - transformada mínimo-máximo melhorada - transformada logarítmica - transformada Box-Cox

Fonte: Próprio autor

2 OBJETIVOS

2.1 Objetivos gerais

Indicar os procedimentos de imputação, detecção de *outliers* e padronização de dados para serem aplicados em resultados experimentais.

Desenvolver uma aplicação *web* que permita utilização de métodos de agrupamento, visualização e pré-processamento de dados.

2.2 Objetivos específicos

Implementar funções na linguagem R:

- cálculo do erro produzido pelo método de agrupamento;
- visualização e análise dos dados;
- pré-processar de dados (imputação, detecção de outliers e padronização de dados);
- desenvolver uma aplicação web que integre as funções implementadas associadas ao pré-processamento de dados, análise de agrupamento e visualização de dados.

Na seção 3 são apresentados os métodos de agrupamento, as abordagens para avaliação da tendência de agrupamento e os índices de validação utilizados nesse trabalho.

3 FUNDAMENTOS TEÓRICOS DOS MÉTODOS DE AGRUPAMENTO

Entre as habilidades mais básicas dos organismos vivos, tem destaque a capacidade de gerar um agrupamento, ou seja, agrupar objetos (CASTRO e FERRARI, 2016). A classificação de Aristóteles dos seres vivos, foi um dos primeiros agrupamentos conhecidos (HENNIG et al.,2015). Um exemplo de técnica estatística exploratória bastante utilizada é a análise de agrupamentos (FÁVERO e BELFIORE, 2017), que consiste em determinar grupos em um conjunto de dados (amostras).

Estes grupos são constituídos de forma que amostras pertencentes a grupos distintos diferem uma da outra mais do que amostras que pertençam ao mesmo grupo (WIERZCHON e KLOPOTEK, 2018). Após a aplicação de um método de agrupamentos, deseja-se que a distância entre as amostras de um mesmo grupo seja minimizada e a distância entre os grupos maximizada (SILVA e BOSCARIOLI, 2016).

De outra forma, dado um conjunto de amostras $C = \{x_1, x_2, \dots, x_n\}$ após aplicação de métodos de análise de agrupamento, deseja-se obter um conjunto de subconjuntos de C, $\{c_1, c_2, \dots, c_k\}$, chamado partição (grupos) de C tal que $c_i \cap c_j = \emptyset, i, j = 1,2, \dots, k$, com $i \neq j$ e $c_1 \cup c_2 \cup \dots \cup c_k = C$ (XU e WUNSCHII, 2005).

3.1 Métodos de análise de agrupamento

Agrupamento é uma técnica comum na análise de amostras e pode ser utilizada em diferentes áreas, como aprendizado de máquina, mineração de dados, reconhecimento de padrões, processamento de imagens e bioinformática (MADHULATHA, 2012).

Os métodos de agrupamentos podem ser divididos em:

- Hierárquicos: localizam recursivamente grupos alinhados de modo aglomerativo ou divisivo (HENNIG et al*.,* 2015);
- Particionais: encontram todos os grupos simultaneamente como uma partição dos dados e não impõe uma estrutura hierárquica (JAIN, 2010);
- Baseados em modelo: tentam otimizar o ajuste entre as amostras a algum modelo matemático. Baseia-se na suposição de que os resultados serão gerados por uma mistura de distribuição de probabilidade (HENNING et al., 2015);
- Baseados em densidade: foram introduzidos para descobrir grupos de forma arbitrária. Utilizam o fato de que no interior de cada grupo há uma densidade típica de amostras e esta densidade é maior que fora do grupo. Regiões com baixa densidade são identificadas como ruído (NAGPAL et al., 2013);
- Baseados em grade: divide o espaço de amostras em uma estrutura de grade com um número finito de células. As operações são realizadas nas células como um todo, em vez das amostras individuais (BROWN et al., 2020);
- Baseados em redes neurais (conexionista): utiliza modelos matemáticos inspirados em processos cerebrais, que simulam o funcionamento do cérebro (SILVA e BOSCARIOLI, 2016).

Os métodos de agrupamentos também podem ser classificados em *hard* (*crisp*) ou *fuzzy* (*soft*) (HENNING et al., 2015). Em outras palavras, um agrupamento em que cada amostra x_i pertence a apenas um grupo c_j, é chamado agrupamento *hard*. O conceito de conjunto *fuzzy* forneceu uma estrutura mais rica para a especificação das relações entre as amostras e os grupos (RUSPINI, 1970). No caso de técnicas de agrupamento *fuzzy*, as amostras podem pertencer a mais de um grupo e ter diferentes graus de pertinência em cada um deles (BORA e GUPTA, 2014).

Os métodos de agrupamentos hierárquicos e particionais foram escolhidos por serem os mais utilizados em resultados experimentais como, por exemplo, a arqueometria, devido à sua facilidade de implementação e a inexistência de hipóteses que limitariam a sua aplicação (BAXTER, 2008). Já o método híbrido, foi escolhido por suprir uma deficiência do método k-médias, associada à escolha inicial dos centróides (ou protótipos) de cada

grupo. Por outro lado, os métodos de agrupamento *fuzzy* foram selecionados por permitirem que uma amostra possa pertencer a mais de um grupo. A Tabela 2 mostra os métodos de agrupamento estudados.

Tabela 2 – Classificação dos métodos de agrupamento

método	**Classe**
ligação simples	Hierárquico/*hard*
ligação média	Hierárquico/*hard*
ligação completa	Hierárquico/*hard*
Ward	Hierárquico/*hard*
k-médias	Particional/*hard*
k-medóides	Particional/*hard*
Híbrido	Particional/*hard*
c-médias	Particional/*soft*
c-médias com polinômio fuzzificador	Particional/*soft*
c-medóides	Particional/soft

Fonte: Próprio autor

Na análise de agrupamentos, dado um conjunto de amostras representado por uma matriz com n amostras e p variáveis, torna-se necessário mensurar a distância entre as amostras, isto é, a similaridade ou a dissimilaridade (HENNIG et al., 2015).

Os métodos de agrupamento hierárquicos executam uma série de sucessivas uniões (aglomerativos) ou divisões (divisivos) no conjunto de amostras (JOHNSON e WICHERN, 2007). Os métodos hierárquicos aglomerativos iniciam com cada amostra sendo um grupo e, em seguida, as amostras mais similares são agrupadas primeiro, de forma que estes grupos são unidos de acordo com sua similaridade. Assim, como a similaridade diminui, ao final do processo todos os grupos são fundidos em um único grupo.

Os métodos divisivos trabalham na direção oposta (JOHNSON e WICHERN, 2007), isto é, o grupo inicial é formado por todas as amostras, em seguida, é divido em dois grupos de acordo com a similaridade entre eles.

Estes grupos são, então, divididos em grupos dissimilares. Este processo continua até que cada grupo seja composto por apenas uma amostra.

Os passos de um método de agrupamento hierárquico aglomerativo, aplicado em um conjunto de n amostras (JOHNSON e WICHERN, 2007), são:

- passo 1: inicia-se com n grupos de amostras, cada um contendo uma amostra, e uma matriz simétrica das distâncias (ou matriz dissimilaridade), $MD = \left(d_{ij}\right)_{n \times n}$, sendo $d_{ij} = d\left(x_i, x_j\right)$ a distância entre as amostras x_i e x_j da base;
- passo 2: procura-se na matriz dissimilaridade pelos grupos mais similares, supondo que sejam os grupos c_i, c_j e a distância entre eles seja $d_{c_i c_j}$;
- passo 3: ocorre a fusão dos grupos c_i e c_j gerando um novo grupo $c_i c_j$. A matriz MD é atualizada, deletando as linhas e a colunas correspondentes aos grupos c_i e c_j, e adicionando à linha e coluna obtidas pelo cálculo das distâncias entre o grupo $c_i c_j$ e o restante dos grupos;
- os passos 2 e 3 são repetidos n vezes. O método termina quando restar apenas um grupo, constituído de n amostras.

A seguir serão descritos os métodos hierárquicos aglomerativos de ligação simples (distância mínima), ligação média, ligação completa (distância máxima) e Ward (KASSAMBARA, 2017).

3.1.1 Método hierárquico de ligação simples

Os dados de entrada do método de ligação simples podem ser as distâncias ou similaridades entre pares de amostras. Os grupos são formados a partir de amostras individuais, pela fusão com os vizinhos mais próximos, em que o termo vizinho mais próximo denota a menor distância ou maior similaridade.

Inicialmente, deve-se encontrar a menor distância na matriz MD, contendo a distância entre as amostras (grupos) de entrada. Se os grupos mais próximos são c_i e c_j, forma-se um novo grupo $c_i c_j$ obtido pela fusão dos dois grupos (JOHNSON e WICHERN, 2007). A distância entre os grupos $c_i c_j$ e qualquer outro grupo c_a é calculada por:

$$d_{(c_i c_j)c_a} = min\left\{d_{c_i c_a}, d_{c_j c_a}\right\}, \quad (1)$$

onde $d_{c_i c_a}$ e $d_{c_j c_a}$ são as distâncias entre os vizinhos mais próximos de c_i e c_a e de c_j e c_a repectivamente.

3.1.2 Método hierárquico de ligação média

O método de ligação média calcula a distância entre dois grupos como a média das distâncias entre todos os pares de amostras dos dois grupos.

Novamente, na matriz MD, o método procura os grupos mais próximos (ou similares), por exemplo c_i e c_j, unindo os dois, formando um novo grupo $c_i c_j$ (JOHNSON e WICHERN, 2007). A distância entre os grupos $c_i c_j$ e c_a é definida por:

$$\mathrm{d}_{(c_i c_j)c_a} = \frac{1}{|c_i c_j||\mathrm{c_a}|}\left(\sum_{x_i \in c_i c_j}\sum_{x_j \in c_a} d(x_i, x_j)\right), \quad (2)$$

x_i é uma amostra do grupo $c_i c_j$, x_j é uma amostra do grupo c_a, $|c_i c_j|$ e $|c_a|$ correspondem ao número de amostras dos grupos $c_i c_j$ e c_a.

3.1.3 Método hierárquico de ligação completa

O método de ligação completa é semelhante ao de ligação simples, com uma importante diferença: em cada estágio a distância entre os grupos é determinada pelas amostras mais distantes. Assim, o método garante que todas as amostras de um grupo estejam dentro de uma distância máxima do outro grupo (JOHNSON e WICHERN, 2007). Logo a distância entre $c_i c_j$ e c_a dada por:

$$d_{(c_i c_j)c_a} = max\left\{d_{c_i c_a}, d_{c_j c_a}\right\}. \quad (3)$$

onde $d_{c_i c_a}$ e $d_{c_j c_a}$ correspondem às distâncias entre c_i e c_a, c_j e c_a respectivamente.

3.1.4 Método hierárquico de Ward

O método de agrupamento hierárquico de Ward procura formar partições de modo a minimizar a perda da informação causada pela fusão de grupos (GAN et al., 2020). Usualmente, a perda de informação é quantificada em termos de um critério da soma dos erros quadrados (EQ), como mostrado na equação 4. Desse modo, o método de Ward é frequentemente chamado de método da variância mínima.

Dado um grupo $\mathrm{c_r}$ de amostras, o $EQ(c_r)$ associado é:

$$EQ(c_r) = \sum_{x_m \in c_r} \left|\left|x_m - \bar{c_r}\right|\right|^2, \tag{4}$$

onde x_m é uma amostra do grupo c_r e $\bar{c_r}$ é o centróide do grupo c_r.

Considerando, nesse sentido, que existam k grupos $\mathrm{c_1, c_2, \dots, c_k}$ em um nível de agrupamento. Então, a perda de informação será representada pela soma dos $EQ(c_r)'s$ dada por:

$$EQ = \sum_{i=1}^{k} EQ(c_i). \tag{5}$$

A cada passo do método de Ward, a união de todos os pares de grupos é considerada e os dois grupos de amostras que apresentarem a menor perda de informação são unidos.

Além dos métodos hierárquicos, existem os particionais, que não geram uma estrutura hierárquica de grupos e possuem como parâmetro de entrada o número de grupos da base estudada. Dito isto, os métodos de agrupamento por partição dividem o conjunto de dados em grupos, que são formados de acordo com um critério de particionamento. Os métodos mais usados são o k-médias e o k-medoides (CASTRO; FERRARI, 2016).

3.1.5 Método particional k-médias

A ideia do método k-médias consiste em definir grupos, de modo que a variação interna dos grupos seja minimizada (HARTIGAN e WONG, 1979; KASSAMBARA, 2017). O método define a variação interna como a soma dos quadrados das distâncias entre as amostras e o centróide correspondente:

$$V(c_r) = \sum_{x_m \in c_r} \left|\left|x_m - \overline{c_r}\right|\right|^2, \quad (6)$$

onde x_m é uma amostra que pertence ao grupo c_r, $\overline{c_r}$ é centroide do mesmo grupo e $V(c_r)$ é a variação interna associada ao grupo c_r.

Cada amostra x_m é atribuída a um grupo, de forma que a soma dos quadrados das distâncias das amostras aos seus respectivos centróides, seja mínima.

A variação total interna é dada pela equação:

$$VT = \sum_{i=1}^{k} V(c_r), \quad (7)$$

onde $V(c_r)$ é a variação interna no grupo c_r.

3.1.6 Método particional k-medoides

O método k-medoides pode ser visto como adaptação do método k-médias. Assim, no lugar de calcular a média de cada grupo, o medoide é obtido em cada iteração e é uma amostra do grupo (REYNOLDS et al., 2006). O medoide para cada grupo é a amostra γ do grupo que minimiza a equação:

$$Z = \sum_{x_m \in c_r} ||x_m - \gamma||, \quad (8)$$

onde x_m é uma amostra do grupo c_r.

Uma das vantagens do método k-medoides, deve-se salientar, é que não há necessidade de repetir o cálculo das distâncias em cada iteração, já que o método k-medoides pode simplesmente procurar a distância na matriz dissimilaridade.

O método inicia com a escolha de k amostras aleatoriamente, que serão os medoides iniciais de cada grupo, a seguir cada amostra é atribuída ao grupo que possui o medoide mais próximo. Após a seleção dos medoides iniciais, todo o processo descrito anteriormente se repete até que o conjunto dos medoides não se altere (CASTRO e FERARRI, 2016).

3.1.7 Método particional híbrido

O método de agrupamento k-médias é um dos mais populares. Entretanto, possui algumas limitações, tais como a necessidade de se conhecer o número de grupos da base e selecionar inicialmente os centroides aleatoriamente.

A solução final do k-médias é sensível à seleção aleatória inicial dos centroides dos grupos (CHEN et al., 2005). Diante disso, o resultado pode ser diferente cada vez que o método é executado. Essa limitação é superada pelo método híbrido, uma vez que os centros dos grupos são determinados por um método de agrupamento hierárquico (Ward), aplicado antes do k-médias.

Em síntese, todos os métodos de agrupamento apresentados até aqui são *crisp*, ou seja, uma amostra pode ou não pertencer a um grupo. Já quanto aos métodos *fuzzy*, uma amostra pode pertencer a mais de um grupo com diferentes graus de pertinência.

A teoria de conjuntos *fuzzy* proposta por ZADEH (1965) é uma alternativa para representação e manipulação de informações imprecisas. Esses conjuntos *fuzzy* podem ser interpretados como uma generalização da teoria de conjuntos.

Na lógica clássica, as informações podem ser classificadas em verdadeiras ou falsas (0 ou 1). Por exemplo, um elemento x pode pertencer ou não a um conjunto A. Desse modo, associado a um conjunto A se tem uma função de pertinência:

$$\mu_A: A \rightarrow \{0,1\}. \tag{9}$$

Os valores que a função assume obedecem à seguinte regra: se o elemento b pertence ao conjunto A, então $\mu_A(b) = 1$, caso contrário, $\mu_A(b) = 0$.

No caso de um conjunto fuzzy, a função de pertinência assume valores no intervalo [0,1]:

$$\mu_F: F \rightarrow [0,1], \tag{10}$$

sendo F um conjunto fuzzy em um universo de discurso U. O conjunto F em U pode ser representado por:

$$F = \{(a, \mu(a)) | a \in U\}, \tag{11}$$

onde a é um elemento de F e $\mu(a)$ a função de pertinência associada (LIMA et al., 2014).

A propriedade de uma amostra pertencer a mais de um grupo, com graus de pertinência entre 0 e 1, é útil quando há sobreposição entre grupos, o que pode revelar características das amostras, não reveladas por outros métodos de agrupamento *crisp* (BAXTER, 2009). Com isso, a seguir serão apresentados três métodos de agrupamentos *fuzzy*: c-médias, c-medoides e c-médias com polinômio fuzzyficador.

3.1.8 Método particional c-médias

O método fuzzy c-médias é utilizado para determinar grupos e seus centros. A relação entre a amostra e grupo é feita por uma função de pertinência que pode variar em [0,1]. Quanto mais próximo do centro estiver a amostra, maior será o valor da função de pertinência (BEZDEK et al., 1984).

As amostras $\{x_t, t = 1,2, \dots, n\}$ são classificadas minimizando a função objetivo baseada na norma euclidiana e nos centros (protótipos) do grupo. A função objetivo é uma soma ponderada dos erros quadráticos entre os grupos, que é definida pela equação:

$$J_m(X, U, V) = \sum_{t=1}^{n}\sum_{i=1}^{k}(\mu_{it})^m \left|\left|x_t - v_i\right|\right|^2, \tag{12}$$

sendo $1 < m < \infty$ e $V = \{v_1, v_2, \dots, v_k\}$, $v_i \in R^n$, $i = 1,2, \dots, k$ é o conjunto dos protótipos dos grupos mostrados na equação 13 e U é uma matriz em que cada elemento μ_{it} representa o valor da função de pertinência de x_t no grupo i, conforme mostra a equação 14.

Se $\left|\left|x_t - v\right|\right|^2 > 0$, $\forall$ i, t, então U e V podem minimizar $\mathrm{J_m}$ apenas quando $\mathrm{m} > 1$ e

$$v_i = \left(\sum_{i=1}^{k}(\mu_{it})^2 x_t\right)\left(\sum_{i=1}^{k}(\mu_{it})^2\right)^{-1}, \tag{13}$$

$$\mu_{it} = \left(\sum_{j=1}^{k}\frac{\left|\left|x_t - v_i\right|\right|^2}{\left|\left|x_t - v_j\right|\right|^2}\right)^{-\frac{1}{m-1}}. \tag{14}$$

3.1.9 Método particional c-medóides

Sejam $\{v_1, v_2, \dots, v_k\} \subset \{x_t, t = 1,2, \dots, n\} = C$. Se $\mathrm{C_s}$ representa todos os subconjuntos V de C. O método fuzzy c-medóides minimiza a equação 15:

$$J_m(V, X) = \sum_{t=1}^{n} \sum_{i=1}^{k} u_{ij}^2 \left|\left|x_j - v_i\right|\right|, \tag{15}$$

onde a minimização é executada sobre todos V em C. Na equação 15, u_{ij} representa a função de pertinência fuzzy de x_j associada ao grupo i (KRISHNAPURAM *et al.*, 2001) e é dada pela equação:

$$\mu_{it} = \left(\sum_{s=1}^{k} \frac{1}{\left|\left|x_t - v_s\right|\right|^2} \right)^{-\frac{1}{m-1}} \left(\frac{1}{\left|\left|x_j - v_i\right|\right|} \right)^{\frac{1}{m-1}}, \tag{16}$$

onde $m \in [1, \infty)$. A equação 16 gera uma partição fuzzy de X, na qual a soma das funções de pertinência de x_j em todos os grupos deve ser 1.

3.1.10 Método particional c-médias com polinômio fuzzificador

O método c-médias com polinômio fuzzificador (WINKLER et al., 2011), utiliza ideias de agrupamentos *crisp* e *fuzzy*. O conceito de polinômio fuzzificador foi alterado por KLAWONN E HÖPPNER (2003), de maneira a substituir a função exponencial μ_{ij}^2 da equação 15 por uma função polinomial de grau 2, da equação 17. O polinômio fuzzificador é dado pela equação:

.

$$\mathrm{f}(\mu_{\mathrm{ik}}) = \frac{1-\beta}{1+\beta}(\mu_{\mathrm{ik}})^2 + \frac{2\beta}{1+\beta}\mu_{\mathrm{ik}}, \tag{17}$$

β ($\beta = 0{,}5$ em todos os testes realizados) controla quanto a função de pertinência torna-se *crisp*. Assim, substituindo o polinômio acima na equação 15, temos a seguinte função objetivo dada pela equação:

$$\mathrm{J_m(X, U, V)} = \sum_{\mathrm{w=1}}^{\mathrm{n}} \sum_{\mathrm{i=1}}^{k} \left(\frac{1-\beta}{1+\beta}(\mu_{\mathrm{iw}})^2 + \frac{2\beta}{1+\beta}\mu_{\mathrm{iw}} \right) \left|\left|\mathrm{x}_w - \mathrm{v_i}\right|\right|^2. \tag{18}$$

O polinômio fuzzificador cria áreas de funções de pertinência *crisp* ao redor dos protótipos, enquanto fora dessas regiões as funções de petinência

fuzzy são atribuídas. Dessa maneira, o polinômio fuzzificador produz graus de pertinência iguais a 1 para amostras próximas aos protótipos e graus de pertinência entre 0 e 1 para amostras distantes dos protótipos.

O processo de atualização dos valores da função de pertinência do método *fuzzy* c-médias com polinômio fuzzificador é descrito pela equação 19:

$$\mu_{iw} = \begin{cases} \left(\frac{1}{1-\beta}\right)(1+(\hat{c}_w-1)\beta)\left(\sum_{w=1}^{\hat{c}_w}\frac{||x_w-v_i||}{||x_w-v_{\varphi(p)}||}\right), & \varphi(i) \le \hat{c}_j \\ 0, \text{ caso contrário} \end{cases}, \quad (19)$$

onde $\hat{c}_w$ é o número de protótipos que sofrerão atualizações e φ é a permutação que armazena os protótipos sorteados.

A seguir são apresentadas duas técnicas para avaliação da tendência de agrupamento, a saber: uma estatística e a outra visual.

3.2 Tendência de Agrupamento

A determinação da estrutura de dados multidimensionais é uma questão muito discutida em análise de dados, em que métodos de agrupamento têm sido usados. Contudo, esses métodos localizam e especificam grupos de amostras mesmo se estes não estiverem presentes. Dessa forma, é apropriado medir a tendência de agrupamento ou aleatoriedade das amostras antes de aplicar um método de agrupamento (CROSS e JAIN, 1982).

A questão de determinar se os grupos estão presentes antes da aplicação de um método de agrupamento é chamada avaliação de tendência de agrupamento. As técnicas de avaliação de tendência de agrupamento podem ser divididas em duas categorias: estatística (estatística Hopkins) (LAWSON e JURS, 1990) e visual (avaliação de tendência visual – VAT – Visual Assessment of Tendency) (KUMAR e BEZDEK, 2020). Nessa perspectiva, embora as abordagens estatísticas determinem se vale a pena procurar grupos em um determinado conjunto de amostras, elas promovem pouca informação sobre o número de grupos, que é um parâmetro de alguns métodos de agrupamento. A classe de abordagens de avaliação de tendência visual de agrupamento utiliza técnicas visuais para indicar se o conjunto de amostras possui ou não grupos, e em alguns casos até sua quantidade.

3.2.1 Estatística de Hopkins

Para determinar se há uma tendência de agrupamento para prosseguir com os métodos de agrupamentos, o conjunto de amostras que está sendo investigado é comparado com uma estrutura de dados conhecida. Em termos estatísticos, isto envolve a construção de uma hipótese nula H_0 (o conjunto de amostras em questão não tem mais tendência de agrupamento do que amostras aleatórias distribuídas uniformemente), calculando a estatística de teste e, em seguida, comparando com o valor padrão para determinar o grau de rejeição da hipótese nula. Em seguida, será descrita uma variação da estatística de Hopkins (LAWSON e JURS, 1990).

Em linhas gerais, a ideia é comparar as amostras com números aleatórios que têm distribuição idêntica às das variáveis. Uma maneira simples de se obter isto é embaralhar a ordem dos valores das variáveis. As distribuições individuais são as mesmas, mas organizadas em ordem aleatória. Desse modo, caso haja uma estrutura significativa no conjunto de amostras, devido a uma relação entre as variáveis, essa estrutura seria destruída no conjunto de amostras embaralhadas. Se a única estrutura no conjunto de amostras veio da natureza das distribuições das variáveis individuais, o conjunto de amostras embaralhadas teriam o mesmo grau de estrutura que o conjunto de amostras.

O cálculo da estatística de Hopkins (LAWSON e JURS, 1990) envolve basicamente duas distâncias denotadas por $d(x_i, x_j)$ e $d(x_i, x'_j)$. As distâncias $d(x_i, x_j)$ são obtidas calculando a distância entre amostras e suas vizinhas mais próximas. Já $d(x_i, x'_j)$ é calculada em duas etapas, primeiro gerando uma pseudo-amostra (x'_j) selecionando um valor da variável aleatoriamente, o que se repete para cada uma das variáveis. Em seguida, é calculada a distância entre a pseudo-amostra e sua vizinha mais próxima, uma amostra. As distâncias calculadas anteriormente, são utilizadas para determinar o valor da estatística de Hopkins e é dada por:

$$\mathrm{H} = \frac{\sum d(x_i, x'_j)}{\sum d(x_i, x_j) + \sum d(x_i, x'_j)}, \tag{20}$$

onde H representa o valor da estatística de Hopkins.

Se as amostras não possuem uma estrutura de grupos, então a distância $d(x_i, x_j)$ de uma amostra a outra será aproximadamente a mesma, na média, à distância $d(x_i, x'_j)$, de uma pseudo-amostra a uma amostra. O valor de H, nesse caso será de aproximadamente 0,5, portanto, não há evidências para se rejeitar a hipótese nula, ou seja, as amostras não são mais agrupadas do que números distribuídos aleatoriamente.

Por outro lado, se as amostras estão arranjadas em grupos, as distâncias $d(x_i, x_j)$ serão relativamente bem menores que $d(x_i, x'_j)$, consequentemente H será aproximadamente 1 e a hipótese nula poderá ser rejeitada. De acordo com (LAWSON e JURS, 1990), para $\mathrm{H} > 0{,}75$ a hipótese nula H_0 pode ser rejeitada.

Com efeito, uma vez verificada a tendência de agrupamento das amostras através da estatística de Hopkins, pode-se determinar o número de grupos da base. A seguir é apresentada a avaliação visual de tendência de agrupamento.

3.2.2 Avaliação visual de tendência de agrupamento (VAT)

No caso da avaliação visual de tendência de agrupamento, seja MD (n por n) uma matriz dissimilaridade entre amostras, MD é:

$$MD = \begin{pmatrix} d(x_1, x_1) & \cdots & d(x_1, x_n) \\ \vdots & \ddots & \vdots \\ d(x_n, x_1) & \cdots & d(x_n, x_n) \end{pmatrix}, \tag{21}$$

onde $d(x_i, x_j)$, corresponde a distância entre as amostras x_i e x_j e MD, satisfaz as seguintes condições (BEZDEK et al., 2002):

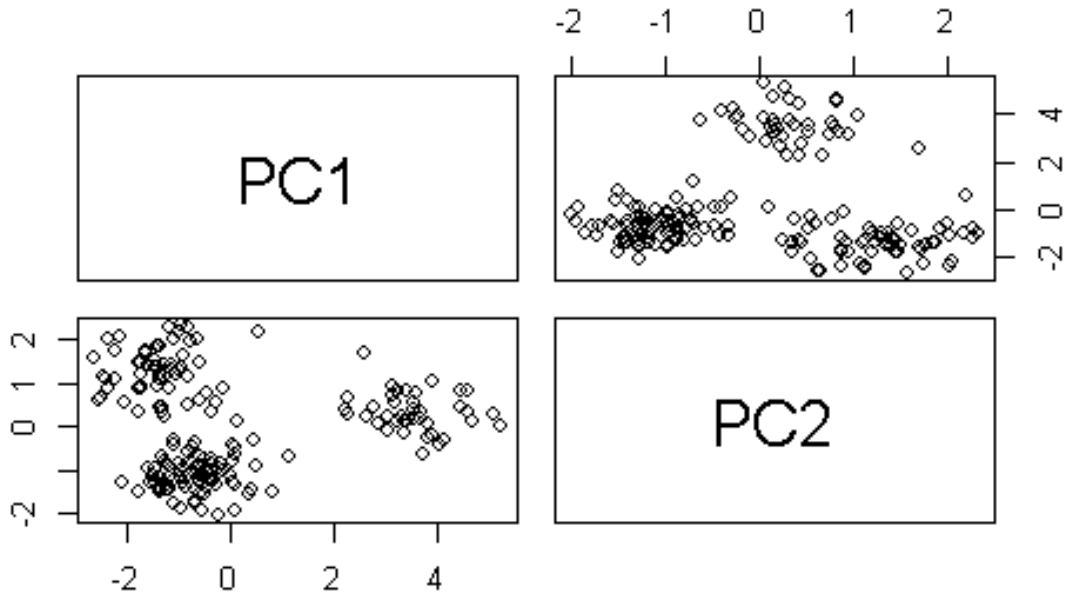
- $d(x_i, x_j) \geq 0$
- $d(x_j, x_i) = d(x_i, x_j)$
- $d(x_i, x_i) = 0, \forall\, i, j = 1,2, \ldots, n.$

A matriz MD está associada a uma imagem, as cores de seus *pixels* correspondem aos valores das distâncias entre as amostras.

A ideia para se obter a avaliação de tendência visual é ordenar o conjunto de amostras $\{x_1, x_2, \ldots, x_n\}$ como $\{x_{k_1}, x_{k_2}, \ldots x_{k_n}\}$, de modo que se k_i é próximo de k_j então x_{k_i} também está próximo de x_{k_j}. Neste caso, a imagem

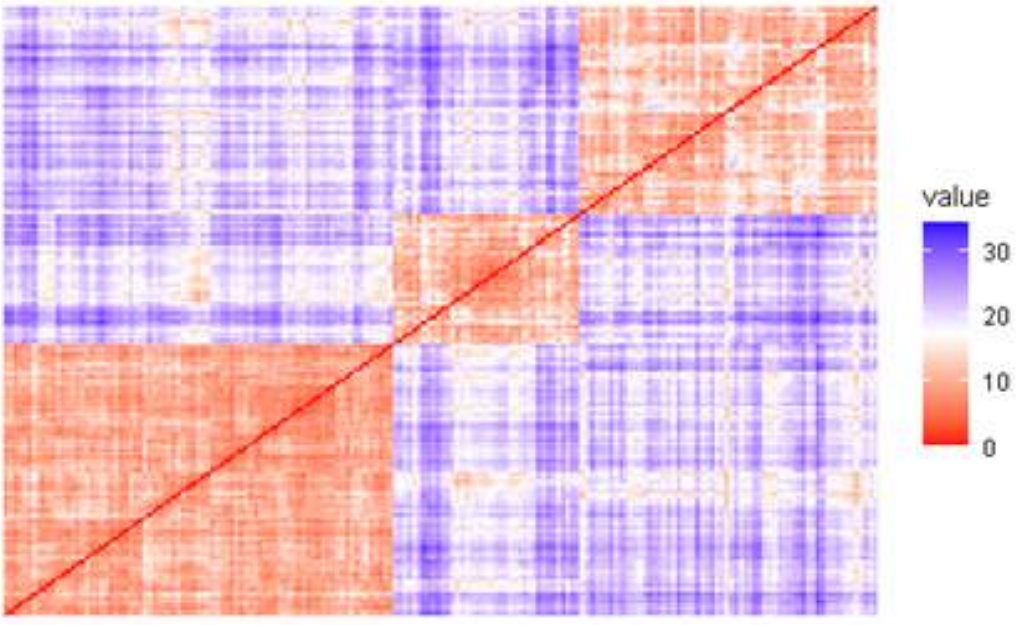
correspondente às amostras ordenadas indica a tendência de agrupamento das amostras, por meio de blocos ao longo da diagonal principal, por exemplo, conforme pode ser observado nas Figuras 1 e 2. Na figura 1, observa-se uma base de dados com três grupos, por outro lado, a Figura 2 (VAT) mostra três blocos (retângulos). Assim sendo, a base de amostras da Figura 1 apresenta tendência de agrupamento e o número de grupos é três. A legenda ao lado da Figura 2 indica a distância entre as amostras, observa-se que a distância entre amostras no interior dos três retângulos (grupos) é menor do que a distância entre as amostras fora deles.

Figura 1 - Gráfico de dispersão obtido através da análise de componentes principais.



Fonte: Próprio autor

Figura 2 – Avaliação Visual de tendência de agrupamento da base de amostras correspondente à Figura 1



Fonte: Próprio autor

3.3 Índices de validação de Agrupamento

Depois de verificada a tendência de agrupamento, e após aplicação dos métodos de agrupamento, faz-se necessário avaliar os resultados obtidos. Nesse sentido, uma das formas de ser fazer isto é através de índices de validação, os índices utilizados nos testes foram: gama, Dunn, Calinski-Harabasz, PBM, Wemmert-Gancarski, tau e Ratkowsky-Lance.

As técnicas de validação de agrupamentos dividem-se em três classes: interna, externa e relativa (BRUN et al., 2007). A validação interna é baseada em propriedades dos agrupamentos como compacidade e separabilidade, logo, não há necessidade de informações adicionais sobre a base de amostras. A compacidade (CP) é uma medida de proximidade entre amostras de um mesmo grupo e pode ser dada por (IAM-ON e GARRETT, 2010):

$$CP = \frac{1}{n}\sum_{i=1}^{k} |c_i| \left(\sum_{x_m, x_j \in c_i} d(x_m, x_j) \left(\frac{|c_i|(|c_i|-1)}{2} \right)^{-1} \right), \tag{22}$$

onde $d(x_m, x_j)$ correponde a distância entre as amostras x_m e x_j do grupo c_i e $|c_i|$ indica o número de amostras do grupo c_i. Uma das definições de separabilidade é (HALKIDI et al., 2001):

$$SE = \frac{D_M}{D_m} \sum_{i=1}^{k} \left(\sum_{j=1}^{k} d(\bar{c_i}, \bar{c_j}) \right)^{-1}, \tag{23}$$

com $D_M = \max \{d(\bar{c_l}, \bar{c_j})\}$, $D_m = \min \{ d(\bar{c_l}, \bar{c_j})\}$, $i, j = 1,2, \dots k$ e $\bar{c_i}$ o centróide do grupo i.

Os índices de validação interna também podem ser utilizados para determinar o número de grupos de uma base de amostras. Como muitos métodos de agrupamento têm como parâmetro de entrada o número de grupos k, ou seja, antes de sua aplicação faz-se necessário determinar o valor de k. Uma abordagem usual para se obter k consiste em executar o(s) método(s) de agrupamento várias vezes com diferentes k's, e escolher k de acordo com os valores dos índices de validação (ARBELAITZ et al., 2013).

Por outro lado, a validação relativa é baseada na comparação de partições geradas pelo mesmo método com parâmetros diferentes, o que também não requer informações adicionais.

A terceira técnica é a validação externa, sendo esta também baseada na comparação entre duas partições, isto é, uma obtida após a aplicação de um método de agrupamento e a outra partição com informações da base de amostras. A validação externa corresponde a uma medida de erro.

Entre todos os índices de validação interna do pacote (do software R), *clusterCrit* (DESGRAUPES, 2018), sete índices descritos foram selecionados da seguinte maneira:

Realizou-se testes com todos os índices de validação interna do pacote clusterCrit (DESGRAUPES, 2018), para determinar o número de grupos da base de amostras estudada. Os métodos de agrupamento utilizados nos testes foram os hierárquicos e os particionais estudados. Eles foram aplicados nas amostras não padronizadas e padronizadas pelas transformadas z-score, mínimo-máximo, mínimo-máximo melhorada,

logarítmica e Box-Cox. Os sete índices com os melhores desempenhos foram selecionados.

Os índices de validação interna estudados foram: gama, Dunn, Calinski-Harabasz, PBM, Wemmert-Gancarski, tau e Ratkowsky-Lance (ARBELAITZ *et al*., 2013; DESGRAUPES, 2018).

3.3.1 Índice gama de Baker-Hubert (Γ)

O índice gama de Baker-Hubert é uma adaptação do índice gama de correlação entre dois vetores, $a = (a_1, a_2, \dots, a_d)$ e $b = (b_1, b_2, \dots, b_d)$. Na análise de agrupamento, o primeiro vetor a corresponde ao conjunto de distâncias entre as amostras, já o segundo b vetor é binário, cujos componentes assumem o valor 0 se duas amostras pertencerem ao mesmo grupo e 1 caso contrário.

Há concordância entre i e j se $a_i > a_j$ implicar em $b_i > b_j$, caso contrário i e j são discordantes. O número de pares concordantes é s^+ e discordantes s^-, o índice gama é:

$$\Gamma = \frac{s^+ - s^-}{s^+ + s^-}. \tag{24}$$

O índice gama de Baker-Hubert, utiliza a equação 24, sendo s^+ o número de vezes em que a distância entre dados que pertencem ao mesmo grupo é menor do que a distância entre dois dados de grupos diferentes, s^- é o número de vezes que a distância entre dados que pertencem ao mesmo grupo é maior que a distância entre dados de grupos diferentes (DESGRAUPES, 2018).

3.3.2 Índice Dunn (D)

O índice Dunn é calculado utilizando a distância entre os grupos e a distância entre amostras de um mesmo grupo:

$$D = \frac{D_{min}}{D_{max}}, \tag{25}$$

onde $D_{max} = \max_{1 \le i \le k} D_i$, $D_{min} = \min_{i \neq j} d_{ij}$, $D_i = \max_{x_i, x_{i'} \in c_i, i \neq i'} \{d(x_i, x_{i'})\}$ e $d_{ij} = \min_{x_i \in c_i, x_j \in c_j} \{d(x_i, x_j)\}$ (DESGRAUPES, 2018).

3.3.3 Índice Calinski-Harabasz (CH)

O índice Calinski-Harabasz é definido por (CHARRAD et al., 2014):

$$CH = \frac{B(n-k)}{W(k-1)}, \tag{26}$$

onde W é a soma das distâncias internas dos grupos, n é o número de amostras e k é o número de grupos:

$$W = \sum_{i=1}^{k} \sum_{j=1}^{|c_j|} d(x_j, \overline{c_\iota}), \tag{27}$$

onde $\overline{c_k}$ e x_j são o centróide e a j-ésima amostra do grupo k e a distância entre os grupos é dada pela equação:

$$B = \sum_{i=1}^{k} |c_i| d(\overline{c_\iota}, \bar{c}), \tag{28}$$

onde $|c_i|$ é o número de amostras do grupo c_i e $\bar{c}$ é o centróide de todo conjunto de amostras.

3.3.4 Índice PBM

O índice PBM é calculado a partir das distâncias entre as amostras e seus respectivos centróides, assim como a distância entre os centróides, de modo que é definido por (DESGRAUPES, 2018):

$$PBM = \left(\frac{E_T D_M}{kW}\right)^2. \tag{29}$$

Sendo:

$$D_M = \max_{i,j=1\ldots k} \{d(\overline{c_\iota}, \overline{c_j})\}, \tag{30}$$

$$E_T = \sum_{i=1}^{n} d(x_i, \overline{c}), \tag{31}$$

onde D_M é a distância máxima entre os centróides e E_T soma das distâncias entre as amostras e o centróide do conjunto de amostras.

3.3.5 Índice Wemmert-Gancarski (WG)

O cálculo do índice Wemmert-Gancarski utiliza um quociente de distâncias entre as amostras e os centróides dos grupos. Para cada amostra x_i pertencente ao grupo c_j, $R(x_i)$ é o quociente entre as distâncias $d(x_i, \overline{c_j})$ e $\min_{j \neq j'}\{d(x_i, \overline{c_{j'}})\}$ e é dado por:

$$R(x_i) = \frac{d(x_i, \overline{c_j})}{\min_{j \neq j'}\{d(x_i, \overline{c_{j'}})\}}, \tag{32}$$

onde $d(x_i, \overline{c_j})$ corresponde à distância entre a amostra x_i e o centróide $\overline{c_j}$, e $\min_{j \neq j'}\{d(x_i, \overline{c_{j'}})\}$, a menor distância da amostra x_i e o centróide dos outros grupos. O índice Wemmert-Gancarski é (DESGRAUPES, 2018):

$$WG = \frac{1}{n}\sum_{j=1}^{k} max\left\{0, |c_j| - \sum_{x_i \in c_j} R(x_i)\right\}, \tag{33}$$

onde n é o número de amostras, k é o número de grupos e $|c_j|$ o número de amostras do grupo c_j.

3.3.6 Índice Tau (τ)

Ao utilizar a mesma notação do índice gama, o índice tau é definido por (DESGRAUPES, 2013):

$$\tau = \frac{s^+ - s^-}{\left[\left(\frac{N_t(N_t-1)}{2} - t\right)\left(\frac{N_t(N_t-1)}{2}\right)\right]^{\frac{1}{2}}}, \tag{34}$$

onde N_t é o número total de distância$s$ e t é o número de comparações de amostras quando ambas pertencem ao mesmo grupo ou estão em grupos diferentes.

3.3.7 Índice Ratkowsky-Lance (RL)

O índice Ratkowsky é calculado a partir da utilização da média dos quocientes entre B^j e T^j (DESGRAUPES, 2018):

$$RL = \sqrt{\frac{R}{k}} \tag{35}$$

onde

$$R = \frac{1}{d}\sum_{j=1}^{d}\frac{B^j}{T^j}, \tag{36}$$

$$B^j = \sum_{m=1}^{k} |c_m|\left(c_m^j - c^j\right)^2, \tag{37}$$

$$T^j = \sum_{i=1}^{n}\left(x_{ij} - c^j\right)^2, \tag{38}$$

onde k é o número de grupos, d é número de variáveis das amostras, n é o número de amostras da base de dados, $\bar{c}_m^j$ é o j-ésimo componente do centróide do grupo c_m ($\bar{c}_m = (\bar{c}_m^1, \bar{c}_m^2, \dots, \bar{c}_m^p)$), $\bar{c}^j$é o j-ésimo componente do centróide do conjunto de amostras ($\bar{c} = (\bar{c}^1, \bar{c}^2, \dots, \bar{c}^p)$) e x_{ij} é o valor da j-ésima variável associado à i-ésima amostra.

Diante do exposto, a seguir é apresentado o erro cometido pelos métodos de agrupamento.

3.4 Erro dos métodos de agrupamento

Os métodos de agrupamento foram avaliados através do cálculo do erro (BRUN et al., 2007), haja vista o conhecimento dos rótulos da base de amostras estudada.

Seja C o conjunto de amostras a ser particionado em k grupos, o particionamento pode ser descrito por uma função (rótulo) $\phi_C: C \to \{1,2,\dots,k\}$, que indica a qual grupo cada amostra pertence. Os grupos são denominados por $\phi_C^{-1}(1),\ \phi_C^{-1}(2), \dots, \phi_C^{-1}(k)$.

O conjunto C^{β} denota os rótulos de C obtidos pelo método de agrupamento β e C^{σ} os rótulos conhecidos de C. Assim, seja $I_{\beta}(C; x_i)$ e $I_{\sigma}(C; x_i)$ os rótulos da amostra x_i em C^{β} e C^{σ} respectivamente, então a divergência entre os dois rótulos consiste em:

$$\epsilon\left(C^{\beta}, C^{\sigma}\right) = \frac{\left|\left\{x_i : I_{\beta}(C; x_i) \neq I_{\sigma}(C; x_i)\right\}\right|}{|C|}, \tag{39}$$

onde $|C|$ denota o número de amostras do conjunto C. Desse modo, desde que a divergência entre duas partições não depende dos índices usados para rotular os grupos, o erro da partição é definido por:

$$\epsilon^{*}\left(C^{\beta}, C^{\sigma}\right) = \min \epsilon\left(C^{\beta}, C^{\sigma}\right), \tag{40}$$

no qual o mínimo é tomado sobre todas as possíveis permutações, πC^{σ} de k conjuntos em C^{σ}.

Mais adiante, para o cálculo do erro cometido pelos métodos de agrupamento baseados em lógica fuzzy, a cada amostra foi associada ao grupo que corresponde ao maior valor da função de pertinência. Na seção 4 são apresentados os métodos de imputação estudados.

4 FUNDAMENTOS TEÓRICOS DOS MÉTODOS DE IMPUTAÇÃO DE DADOS

Em resultados experimentais, a presença de valores faltantes (ausentes) é frequente. Esse foi e continua sendo um grande problema no estudo dos resultados utilizando métodos estatísticos (ALLISON, 2001; RUBIN, 1987). Os problemas associados com valores ausentes são: perda de eficiência, complicações na análise dos dados, viés resultante de diferenças entre resultados ausentes e completos, redução do desempenho estatístico, entre outros. (HAWTHORNE e ELLIOTT, 2005). As dificuldades surgem porque os métodos estatísticos consideram que em todas as amostras as mesmas variáveis foram determinadas e incluídas na matriz de amostras.

Com frequência, as variáveis ou amostras com valores ausentes, são excluídas da matriz de amostras, de forma que este é o procedimento mais comum usado pelos programas estatísticos, haja vista que tais pacotes estatísticos não são capazes de processar amostras com valores ausentes (MISZTAL, 2018). Por exemplo, eles não podem executar a classificação nem a regressão, se uma das variáveis possui valores ausentes. Nesse sentido, pesquisadores têm estudado estratégias para substituição dos valores faltantes por um valor plausível, isto é, um processo geralmente chamado imputação de dados (LITTLE et al., 2014).

Os valores ausentes ocorrem na fase de coleta de dados, por diversas razões. É importante classificar os valores ausentes com base nos mecanismos que os produzem, para determinar melhor abordagem para seu tratamento (KEERIN, 2021).

Os mecanismos de ausência de dados podem ser classificados em (MISZTAL, 2018):

- completamente aleatório (MCAR-*Missing Completely At Random*), os valores ausentes ocorrem de forma aleatória, não estão relacionados com a variável nem a amostra. Geralmente ocorrem por erros na transferência de dados;
- aleatório (MAR – *Missing At Random*), o mecanismo do valor ausente está relacionado com a amostra;

- não aleatório (MNAR – *Not Missing at Random*), o valor ausente da variável não é aleatório, está relacionado a um motivo específico. É crucial que o pesquisador tenha uma compreensão clara do mecanismo de ausência de resultados e do processo de coleta de dados.

Diante disso, ao levar em conta os mecanismos de ausência de resultados, a exclusão de amostras pode ser realizada somente se a suposição MCAR for verificada. Entretanto, mesmo nesse caso, a eliminação da amostra pode levar a um viés, principalmente, se a quantidade de amostras com valores ausentes for grande (MISZTAL, 2018). O mecanismo MNAR é a situação mais comum na prática (LAAKSONEN, 2018) e nesse caso não existe um método universal para tratar valores ausentes adequadamente (NORAZIAN et al., 2013). Na maioria das vezes é impossível classificar inequivocamente os valores ausentes em um dos três mecanismos de ausência de dados (GRAHAM, 2009).

As abordagens para o tratamento de valores ausentes são:

- exclusão da amostra ou variável, é geralmente padrão em pacotes estatísticos. Essa abordagem tem a vantagem de poder ser aplicada em qualquer tipo de análise estatística e ser simples. Entretanto, pode excluir uma grande quantidade de amostras ou variáveis, comprometendo a aplicação de métodos estatísticos;
- métodos de imputação, que substituem o valor ausente por uma estimativa, em seguida a análise é executada como se não houvesse valores ausentes. Contudo, acarreta uma subestimação do erro, visto que os valores imputados são determinados pelos valores não ausentes.

Os métodos de imputação são classificados em dois tipos: simples e múltipla (SINHARAY et al., 2001). A imputação simples refere-se à substituição de um valor faltante por um valor plausível de uma variável em um conjunto de amostras. Em métodos de imputação simples, assume-se que o valor imputado é único e está correto. No entanto, nunca há certeza sobre a validade dos valores imputados. Dessa maneira, a incerteza em torno desses valores imputados deve ser incorporada aos métodos que tratam valores ausentes (LITTLE e RUBIN, 1989). A imputação múltipla, em vez de substituir um único valor para cada amostra ou variável ausente, ela substitui vários valores plausíveis, e então determina a incerteza sobre os valores a serem imputados. Todavia, se a proporção de valores ausentes for pequena, inferior a 5%, então a imputação simples pode ser bastante precisa (SCHAFER, 1999).

Em geral os métodos de imputação de valores ausentes também podem ser classificados em: estatísticos e de aprendizado de máquina. Entre as técnicas estatísticas a imputação pela média e pela mediana são as mais simples (MIT CRITICAL DATA, 2016). Entre as ferramentas de aprendizado de máquina, os métodos de agrupamento têm sido amplamente utilizados para imputação de dados ausentes (WAN et al., 2021). Outra ferramenta do aprendizado de máquina empregada na imputação de dados é a rede neural autoencoder (AE), em que seu uso é motivado pelo seu sucesso na extração de características úteis da base de dados (ABIRI et al., 2019).

Existem vários procedimentos para imputar os valores faltantes, em uma base de amostras. Porém, aqui serão apresentados quatro métodos de imputação simples: média, rede neural autoencoder, agrupamento e c-médias.

4.1 Imputação pela média

A abordagem mais comumente utilizada é a imputação pela média (MALARVIZHI e THANAMANI, 2012). Esta assume que a média dos valores da variável é a melhor estimativa para qualquer amostra que tenha valores ausentes na mesma variável (SONG e SHEPPERD, 2007), de forma que é definida por (LITTLE e RUBIN, 2019):

$$\tilde{x}_j = \frac{1}{n-1} \sum_{m=1, m \neq i}^{n} x_{mj}, \tag{41}$$

onde $\tilde{x}_j$ é o valor imputado da j-ésima variável e x_{mj} é o valor da matriz de amostras correspontende a amostra m e variável j.

Este método é facilmente implementável e simples, mas possui algumas desvantagens. Se os valores ausentes ocorrem em grande quantidade, então todos eles são substituídos por um mesmo valor de imputação, o que leva a mudança na forma da distribuição das amostras (JADHAV et al., 2019). Consequentemente a imputação média atenua o desvio padrão e a variância.

4.2 Imputação por autoencoder

Redes neurais artificiais (RNA) são sistemas constituídos por unidades de processamento (neurônios artificiais), que lembram a estrutura do cérebro humano. Os principais atrativos das RNAs na solução de problemas são: a capacidade de aprender através de exemplos e generalizar o que foi aprendido (BRAGA et al., 2007).

Muitas técnicas de aprendizado profundo têm sido utilizadas em problemas de imputação de dados, entre elas o *autoencoder* e suas variantes. O *Autoencoder* tornou-se uma poderosa ferramenta na captura das principais características dos dados. É um tipo de rede neural para tarefas de aprendizado não supervisionado, por exemplo, redução de dimensão de dados, extração de características e reconstrução de dados (ZHANG et al., 2020). Além disso o AE é treinado para copiar sua entrada na saída, e é constituído de duas partes: o codificador e o decodificador. O codificador é uma função $h = f(x) = s_f(W_x + b_h)$, onde s_f é uma função de ativação, o codificador é parametrizado por uma matriz peso W, de ordem $d_h \times d_x$ e um vetor viés $b_h \in R^{d_h}$. Por outro lado, o decodificador é uma função g que mapeia a representação h na reconstrução y, onde $y = g(h) = s_g(W'h + b_y)$, s_g é a função de ativação do decodificador (SALAH et al., 2011).

O treinamento da rede *autoencoder* consiste em determinar os parâmetros W, b_h e b_y que minimizam o erro de reconstrução nos exemplos do conjunto de treinamento D_m, que corresponde a minimizar a seguinte função objetivo (SALAH et al., 2011):

$$J_{AE} = \sum_{z \in D_n} L(x, g(f(x)), \tag{42}$$

onde L é o erro de reconstrução, uma escolha típica é $L(x, y) = \left|\left|x - y\right|\right|^2$ e uma função decodificadora $y = g(h)$. Assim, se um *autoencoder* tivesse sucesso simplesmente copiando a entrada na saída, $x = g(f(x))$, para todo x, a rede funcionaria como uma função identidade, o que não seria útil. Razão pela qual estas redes foram projetadas para não serem capazes de apreender a copiar perfeitamente a entrada. Dessa forma ela é forçada a priorizar aspectos da entrada que devem ser mantidos (GOODFELLOW et al., 2016).

As redes *autoencoder* possuem uma estrutura simétrica (CHARTE et al., 2018), e são um caso especial das redes *feedforward* e que podem ser treinadas com as mesmas técnicas (GOODFELLOW et al., 2016).

Um caso especial da rede *autoencoder* é a rede DAE (*Denoising Autoencoder*), que é treinada para reconstruir uma entrada sem ruído x, a partir de uma versão corrompida x_c. A entrada x_c, é codificada através de uma função $f(x) = s_f(W_x + b_h) = y$, obtendo-se uma representação intermediária h. Em seguida y é decodificada, utilizando uma função g, tal qual, $g(h) = z$, sendo z uma reconstrução de x. Os parâmetros das funções f e g são obtidos de forma a minimizar o erro médio de reconstrução sobre o conjunto de treinamento, isto é, tornar z o mais próximo possível de x (VINCENT et al., 2010). Nos testes realizados o valor faltante é estimado inicialmente através da imputação pela média, gerando um valor que servirá de entrada para a rede DAE. A saída desta rede fornecerá o valor imputado.

4.3 Imputação por agrupamento

Um dos métodos de agrupamento mais conhecido é o k-médias. O método de imputação de dados descrito a seguir utiliza o k-médias para estimar o valor a ser imputado (SHI et al., 2020). Inicialmente ele divide o conjunto de amostras C, em dois subconjuntos: um somente de amostras com valores ausentes (I_1) e outro com amostras completas (I_2), com $I_1 \cup I_2 = C$. Primeiro aplica-se o método k-médias no conjunto I_2. O valor faltante é substituído pelo valor médio da variável correspondente, em seguida a análise de perturbação do centróide é feita, para determinar o melhor valor de imputação. O método é composto pelos seguintes passos:

- passo 1: o conjunto C é dividido em dois subconjuntos I_1 e I_2;
- passo 2: aplica-se o método k-médias em I_2, obtendo-se os grupos $\{c_1, c_2, \dots, c_k\}$;
- passo 3: para cada amostra $x = (x_1, \dots, x_l, \dots x_d) \in I_1$, é calculado um conjunto de valores:

$$\left\{x_l^{(1)}, x_l^{(2)}, \dots, x_l^{(k)}\right\}, \tag{43}$$

utilizando a equação:

$$x_l^{(i)} = \frac{1}{|c_i|} \sum_{x \in c_i} x_l \, , i = 1, \dots, k, \tag{44}$$

onde $|c_i|$ indica o número de elementos de c_i.

- passo 4: em seguida os centróides de I_2 são recalculados, adicionando os valores imputados do passo anterior, $\{c_1^*, c_2^* \dots, c_k^*\}$;

- passo 5: calcula-se:

$$\min_{1 \le i \le k} \left|| c_i - c_i^* || \right|, i = 1,2, \dots k, \tag{45}$$

o menor valor de (45), indica o valor de (43) que deve ser utilizado na imputação.

4.4 Imputação pelo c-médias

No método de agrupamento k-médias, cada amostra pertence a um único grupo (NIKFALAZAR et al., 2017). Dessa forma, não é possível avaliar a sobreposição entre os grupos.

Para suprir essa desvantagem do método k-médias, foi desenvolvido o método c-médias, que nada mais é que uma versão fuzzy do método k-médias, em que os dados podem pertencer a mais de um grupo. A seguir é apresentado uma versão de imputação simples do método proposto por NIKFALAZAR et al. (2017) e que pode ser descrito pelos seguintes passos:

- passo 1: aplicação do método de imputação média, utilizado para produzir uma estimativa inicial para o valor ausente;

- passo 2: execução do método c-médias na base de amostras, gerando: os graus de pertinência entre 0 e 1 para cada um das amostras e os centróides;

- passo 3: os valores imputados pela média são atualizados utilizando a expressão:

$$v_{ij} = \sum_{g=1}^{k} \mu_{ig} x_{gj}, \forall \, (i,j) \in M, \tag{46}$$

onde M é o conjunto das coordenadas dos valores faltantes, μ_{ig} é a função de pertinência da amostra i no grupo g e x_{gj} o valor da j-ésima variável no grupo g.

4.5 Erro quadrático médio normalizado associado à imputação de dados

Para avaliação das estimativas obtidas pelos métodos de imputação foi utilizado o cálculo do erro quadrático médio normalizado (NRMSE-Normalized Root Mean Squared Error), que calcula o erro entre o valor real (y_j) e valor estimado (imputado) ($\widehat{y_j}$) da seguinte forma:

$$NRMSE = \frac{1}{\sigma_y}\sqrt{\frac{\sum_{j=1}^{N}\left(y_j - \widehat{y_j}\right)^2}{N}}, \tag{47}$$

onde σ_y é o desvio padrão dos N valores reais correspondentes a todos os valores faltantes (BRÁS e MENEZES, 2007). O NRMSE leva em conta a escala dos valores, indicando a proximidade destes.

Na seção 5 são descritos os métodos de detecção de outliers analisados.

5 FUNDAMENTOS TEÓRICOS DOS MÉTODOS DE DETECÇÃO DE VALORES DISCREPANTES (OUTLIERS)

Uma das principais tarefas antes da aplicação de técnicas de classificação das amostras, é a detecção de *outliers* em estudos de resultados experimentais. Um *outlier* é definido como uma amostra muito diferente do restante da base. Muitos artigos e estudos têm demonstrado que bases sem *outliers* são uma exceção. Desse modo, *outliers* podem influenciar modelos estatísticos, e não é desejável que os parâmetros estimados sofram influência deles. Este problema pode ser evitado utilizando métodos robustos para o ajuste do modelo ou detectando *outliers*, de forma, que depois da exclusão, sejam aplicados métodos estatísticos (FILZMOSER, 2016). Do ponto de vista da análise de agrupamento, os *outliers* são amostras que não pertencem aos grupos da base (AGGARWAL e YU, 2005).

Entre as possíveis causas para a ocorrência de *outliers* pode-se destacar: erros de medidas e erros por considerar uma ou mais amostras que não pertencem à população de interesse. Nesse sentido, a importância do estudo de *outliers* reside no fato de que sua presença pode conduzir a falsas alternativas e interpretações (BARNETT e LEWIS, 1994).

Nas últimas décadas, muitos métodos para detecção de *outliers* têm sido propostos. Eles podem ser divididos em quatro tipos (JIN et al., 2001; ZHANG et al., 1997): baseados em distribuição, baseados em distância (BARNETT e LEWIS, 1994), baseados em agrupamentos e baseados em densidade.

Nos métodos baseados em distribuição, uma amostra é considerada um *outlier* caso desvie de uma distribuição (BARNETT e LEWIS, 1994). O problema deste tipo de método, consiste na distribuição das amostras, usualmente desconhecida ou nas amostras que podem não seguir uma distribuição.

Já os métodos que utilizam o cálculo da distância, estes detectam *outliers* calculando a distância entre as amostras. Assim, as amostras mais distantes dos grupos são classificadas como *outliers*. Esta é a abordagem mais comum de detecção de *outliers* (MANDHARE e IDATE, 2017). Além disso esses métodos são simples de implementar e não fazem suposições prévias sobre o modelo de distribuição das amostras (HODGE e AUSTIN, 2004).

Métodos que operam utilizando a análise de agrupamento detectam *outliers* no processo de busca por grupos. Dito isto, a amostra que não pertence aos grupos é considerada *outlier* (ZHANG et al., 1997).

Por outra via, em métodos baseados em densidade, um *outlier* é detectável quando sua densidade local se difere da densidade de seus vizinhos.

Os métodos de detecção de *outliers* possuem dois tipos de saída (AGGARWAL, 2017) :

- baseada em pontuação (escore): a saída do método atribui uma pontuação a cada uma das amostras da base;
- a saída associa a cada uma das amostras um rótulo, indicando se ela é um *outlier* ou não.

A maioria dos métodos de detecção de *outliers* gera uma pontuação e um limiar dessa pontuação é utilizado para converter as pontuações em rótulos. Com isso, se o limiar for selecionado de forma muito restritiva, minimizando o número de *outliers* declarados, o método perderá *outliers*. Em contrapartida, se o método declarar muitos *outliers*, isso acarretará muitos falsos *outliers* (AGGARWAL, 2017).

Neste trabalho foram testados métodos de detecção de outliers, baseados nas distâncias: Mahalanobis e Mahalanobis robusta.

5.1 Distância de Mahalanobis

A distância de Mahalanobis (D_i), é bastante utilizada na detecção de *outliers*, uma vez que leva em consideração a correlação das amostras, de modo que é calculada usando a inversa da matriz covariância do conjunto das amostras (DE MAESSCHALCK et al., 2000). A distância Mahalanobis quadrática é dada pela equação:

$$D_i^2 = (x_i - \overline{x})S^{-1}(x_i - \overline{x})^T \quad i = 1,2,\dots,n, \tag{48}$$

S é a estimativa da matriz covariância dentro dos grupos, x_i é a i-ésima amostra e $\bar{x}$ é o centróide do grupo.

Para cada uma das n amostras com p variáveis, é calculada a D_i utilizando a equação 48. Os maiores valores da D_i^2 indicam a presença de *outliers*. Um limiar D_c é calculado através da equação (PENNY, 1996):

$$D_c = \frac{p(n-1)^2 F_{p,n-p-1,\frac{\alpha}{n}}}{n(n-p-1)pF_{p,n-p-1,\frac{\alpha}{n}}}, \tag{49}$$

onde F corresponde à distribuição de Fisher, n é o número de amostras e p o número de variáveis. As amostras com valores da distância D_i^2 maiores que D_c são classificadas como *outliers***.**

5.2 Distância Mahalanobis robusta

Embora seja fácil detectar um único *outlier* por meio da distância de Mahalanobis, essa abordagem pode sofrer o efeito do mascaramento, ou seja, a base de dados pode possuir vários *outliers* e nem todos possuírem valores elevados de D_i^2 (ROUSSEEUW e DRIESSEN, 1999). Uma maneira de contornar esse problema é usar distâncias baseadas em estimadores robustos de localização e dispersão. O estimador robusto utilizado é a covariância de determinante mínimo, cuja ideia é encontrar uma fração ζ ($\approx 0{,}75n$) de amostras, que não possuam *outliers*. Feito isso, é calculada a média e a matriz covariância correspondente às ζ amostras. Valores da distância Mahalanobis robusta (DMR_i) maiores que D_c indicam outliers. O procedimento se repete para todos os conjuntos com ζ amostras, de maneira que o conjunto com menor determinante é selecionado. A distância de Mahalanobis robusta é (LEYS et al., 2018):

$$DMR_i = [(x_i - \mu_E)^T \Sigma_E^{-1} (x_i - \mu_E)], \tag{50}$$

onde μ_E e Σ_E são média e a matriz covariância estimada respectivamente. Uma amostra é classificada como *outlier* se a DMR_i correspondente for maior que D_c.

Na seção 6 são apresentados os métodos de padronização dos dados analisados.

6 FUNDAMENTOS TEÓRICOS DOS MÉTODOS DE PADRONIZAÇÃO DOS DADOS

Os métodos estatísticos, como análise de componentes principais, análise de agrupamento e análise discriminante são amplamente utilizados em resultados experimentais, como por exemplo, em arqueometria (BAXTER, 2006). Neste caso, o foco principal é análise química de artefatos arqueológicos, a saber, como objetos de vidro e cerâmica. As amostras são analisadas em relação à sua composição química, com muitas variáveis (elementos químicos) que podem ter escalas diferentes (MUCHA et al., 2008).

Por outro lado, em muitas aplicações de análise estatística multivariada, as amostras, ou medições reais, não são usadas diretamente (JAIN e DUBES, 1988). Assim, um problema que surge na análise estatística multivariada envolve a decisão de padronizar ou não as variáveis, antes da aplicação de algum método de estatística multivariada. É necessário, portanto, padronizar as variáveis se a medida de dissimilaridade for sensível às diferenças nas magnitudes ou escalas das variáveis (MILLIGAN e COOPER, 1988). Assim, se algumas variáveis têm uma grande escala ou variabilidade, estas variáveis afetam o desempenho dos métodos de estatística (TANIOKA e YADOHISA, 2012) e de redes neurais (AL SHALABI et al., 2006). Estas variáveis podem ser quantitativas ou qualitativas e podem ser subdivididas em (JAIN et al., 1999):

-variável quantitativa:

a) contínua

b) discreta

-variável qualitativa:

a) nominal

b) ordinal

Todas as transformadas descritas neste trabalho foram aplicadas em dados com variáveis quantitativas contínuas. A padronização é usada para

garantir que as variáveis tenham igual contribuição no cálculo das distâncias (CHU et al., 2009).

A existência de uma grande quantidade de transformadas complica essa decisão de padronizar ou não as amostras. Portanto, torna-se importante avaliar os efeitos da padronização na análise de agrupamento, uma vez que os métodos de agrupamento utilizam o cálculo de distâncias. As abordagens de padronização de variáveis são essencialmente de dois tipos: padronização global e padronização interna do grupo. Quanto à padronização global, esta é aplicada em todas as variáveis, já no tocante à padronização interna do grupo, esta é aplicada somente ao grupo em questão.

As transformadas para padronização apresentadas a seguir utilizaram a abordagem global. Por conveniência, seja $C = \{x_1, x_2, \dots, x_n\}$ um conjunto de amostras com p variáveis. A matriz das amostras $n \times p$ é dada por:

$$(x_1, x_2, \dots, x_n)^T = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \tag{51}$$

onde x_{ij} corresponde à j-ésima variável da i-ésima amostra.

A seguir, são apresentadas as técnicas de padronização abordadas neste trabalho: z-score, mínimo-máximo, mínimo-máximo melhorada, logarítmica e Box-Cox. As transformadas são aplicadas nas colunas da matriz da equação 51.

6.1 Transformada z-score

A transformada z-score (CASTRO e FERRARI, 2016) é uma forma de padronização dos dados em que após a sua aplicação, a variável passa a ter média 0 e desvio padrão 1. Informações de localização e escala das variáveis são perdidas e a transformada é definida por:

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x_j}}{\sigma_j}, \tag{52}$$

onde $\overline{x_j}$ e σ_j são a média e o desvio padrão da j-ésima coluna da matriz de amostras.

6.2 Transformada mínimo-máximo

A próxima padronização utiliza a variação da variável como divisor (MOHAMAD e USMAN, 2013):

$$x_{ij}^{*} = \frac{x_{ij} - \min\,(x_j)}{\max\big(x_j\big) - \min\,(x_j)}, \tag{53}$$

onde x_j corresponde a j-ésima coluna da matriz da equação 51. Após a sua aplicação os valores de x_{ij}^{*} estão contidos no intervalo [0,1], facilitando a comparação de valores que estão em diferentes escalas ou unidades.

6.3 Transformada mínimo-máximo melhorada

A transformada mínimo-máximo melhorada (KABIR et al., 2016) pode ser descrita pelos seguintes passos:

- passo 1: selecionar os valores da matriz da equação 51 que ocorrem mais de uma vez, para construir o conjunto $R_k = \{R_{k1}, R_{k2}, \ldots\}$;
- passo 2: calcular a média ($m(R_k)$) e o desvio padrão ($sd(R_k)$) de R_k;
- passo 3: calcular $R_{kA} = m(R_k) + sd(R_k)$;

$$x_{ij}^{*} = \begin{cases} \dfrac{x_{ij} - \min\,(x_j)}{2R_{kA} - 2\min\,(x_j)}, x_{ij} \leq R_{kA} \\ 0.5 + \dfrac{x_{ij} - R_{kA}}{\max\big(x_j\big) - R_{kA}}, x_{ij} > R_{kA}. \end{cases} \tag{54}$$

6.4 Transformada logarítmica

A transformada logarítmica é dada por:

$$x_{ij}^{*} = \log_{10}(x_{ij}). \tag{55}$$

A aplicação da transformação logarítmica em dados de concentração compensa as diferenças das unidades entre as variáveis.

6.5 Transformada Box-Cox

A transformada Box-Cox (ATKINSON et al., 2021) é uma família de transformadas de potência que, após a sua aplicação deseja-se obter uma distribuição aproximadamente normal multivariada (RODE e CHINCHILLI, 1988) e é definida por:

$$x_{ij}^{*(\lambda)} = \begin{cases} \dfrac{\left(x_{ij}\right)^{\lambda} - 1}{\lambda}, \lambda \neq 0 \\ \ln\left(x_{ij}\right), \lambda = 0, j = 1, \ldots, n. \end{cases} \quad (56)$$

onde $\lambda \in [-1, 2]$. O valor de λ foi estimado por máxima verossimilhança (OSBORNE, 2010).

7 BASE DE DADOS

Neste estudo foi utilizada uma base de dados de 140 amostras de fragmentos cerâmicos de três sítios arqueológicos: Prado, Água Limpa e Rezende, chamados de A, B e C. Os 3 sítios são superficiais localizados na parte intermediária de colinas com riachos na parte inferior. As cerâmicas nesses locais estão associadas à preparação de alimentos, urnas funerárias e uso decorativo.

O sítio Prado está localizado na fazenda Engenho, na cidade de Perdizes, no estado de Minas Gerais. É um sítio a céu aberto situado em área de cultivo permanente e temporário. Com ocupação em relevo de colina e com um único nível arqueológico: lito-cerâmico. Os recipientes cerâmicos foram parcialmente reconstruídos, em campo ou em laboratório. Os recipientes coletados são lisos sem decoração plástica, com predominância de granularidade média e grande, com má seleção dos grãos (ALVES, 1991; 1994). A datação revelou que o sítio foi habitado no ano 493±74 antes do presente (AP).

O sítio Água Limpa está localizado na confluência de três sítios na cidade de Monte Alto, no norte do estado de São Paulo. Foram encontrados locais de sepultamento primários de jovens e adultos estendidos e semi-flexionados. As cerâmicas recolhidas neste local são de dois tipos: pintadas e lisas. A pintura é em vermelho e branco, de maneira que os poucos fragmentos pintados e inteiros que foram coletados não têm forma. A seleção de grãos é boa com predominância de grãos finos e médios. O sítio foi habitado no ano 1524±70 AP.

O sítio de Rezende está localizado na fazenda Paiolão, em Piedade, no Vale do Paranaíba, 7 km da cidade de Centralina, no estado de Minas Gerais. Os estudos arqueológicos indicam duas ocupações: a mais recente é representada por uma ocupação ceramista no ano 1190±60 AP. Desse modo, estudos mostraram que a população vivia em cabanas ovais formando aldeias e usava fogo para iluminar, cozinhar e como fonte de calor. A cerâmica produzida

era simples, utilitária e funerária. Em síntese, representam os primeiros e mais antigos habitantes de Minas Gerais (ALVES, 1991; 1994).

Para os estudos foram utilizadas 140 amostras, 34 amostras do sítio A, 76 do sítio B e 30 do sítio C. As amostras foram analisadas por análise por ativação com nêutrons instrumental, INAA, nas quais foram determinados 13 elementos químicos: As, Ce, Cr, Eu, Fe, Hf, La, Na, Nd, Sc, Sm, Th e U. Na Tabela 3 estão apresentados os resultados das frações de massa (CARVALHO, 2018).

Tabela 3 – Resultados das frações de massa das amostras de fragmentos cerâmicos em $\mu g\ g^{-1}$, quando não indicado, dos sítios A, B e C

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U
A													
1	2,52	113,22	123,43	1,51	3,81	8,81	31,53	302,12	35,23	31,51	7,74	17,81	4,61
2	1,53	108,13	109,22	1,47	3,26	9,52	41,12	766,44	40,45	26,16	7,81	18,01	4,34
3	1,47	103,33	114,81	1,36	3,61	8,73	40,43	644,12	38,22	27,61	7,84	17,16	4,38
4	2,01	86,51	116,23	1,24	3,47	9,44	32,93	643,10	37,32	27,38	7,74	16,56	3,45
5	1,87	117,15	134,81	1,52	3,35	8,77	33,74	484,33	37,47	30,61	7,91	17,24	3,71
6	1,69	115,39	124,77	1,68	3,84	8,41	30,45	328,49	43,44	32,59	7,43	17,73	3,92
7	2,13	112,12	117,20	1,43	3,33	8,27	29,78	500,50	32,84	30,21	7,27	16,72	3,53
8	1,74	120,39	115,23	1,72	3,60	9,97	32,65	377,64	40,31	30,75	8,09	16,65	4,94
9	2,12	121,20	121,81	1,61	3,73	9,11	33,53	493,38	34,21	31,86	6,63	17,68	5,25
10	1,56	104,82	110,41	1,43	3,06	9,65	36,95	731,35	48,22	27,47	7,92	16,78	3,76
11	1,13	120,40	130,21	1,45	2,74	8,56	35,84	526,58	51,44	28,88	8,12	18,02	3,97
12	1,39	110,43	138,31	1,36	2,89	8,41	30,77	552,36	34,10	29,59	6,81	17,78	1,88
13	1,87	105,28	142,47	1,16	2,66	9,32	27,26	543,42	26,13	27,95	6,35	16,41	3,32
14	1,84	108,30	157,18	1,26	3,07	9,24	29,33	552,23	36,58	31,41	6,75	17,95	6,33
15	1,36	130,01	136,19	1,42	2,80	9,12	34,34	533,43	38,11	28,84	7,37	17,86	5,34
16	1,83	118,14	156,22	1,43	2,98	8,84	33,01	590,82	32,38	30,25	7,43	18,78	3,57
17	1,62	114,74	164,01	1,36	2,99	9,14	29,54	555,10	26,85	31,46	6,82	17,64	4,18
18	1,47	121,37	152,06	1,42	2,96	9,23	33,51	621,08	39,77	30,47	7,76	18,52	5,48
19	1,83	114,54	170,13	1,27	2,99	9,53	30,10	635,45	27,38	31,37	7,80	17,24	4,39
20	1,64	119,33	151,29	1,37	2,83	8,26	33,23	590,98	34,19	28,91	7,49	18,08	5,01
21	1,26	113,61	138,10	1,33	2,80	8,54	31,45	557,71	29,91	28,63	7,02	15,81	4,81

(continua)

Tabela 3 – (continuação)

A	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U
22	1,57	116,49	136,81	1,37	2,87	9,15	32,46	548,12	41,34	29,54	7,39	16,76	4,63
23	1,58	104,38	136,71	1,31	2,63	8,47	29,38	579,48	38,62	27,68	6,83	16,14	3,54
24	1,53	118,28	144,66	1,44	2,82	8,38	32,13	656,83	31,41	29,31	7,27	17,31	4,26
25	1,83	131,37	140,56	1,64	2,65	8,99	35,37	593,23	46,88	29,16	6,51	16,57	5,77
26	1,41	113,34	140,41	1,34	2,7	8,30	31,65	558,29	36,71	28,56	7,15	17,06	2,61
27	1,85	118,11	175,81	1,01	1,73	10,11	38,54	786,29	57,65	26,71	7,75	19,21	4,53
28	2,37	106,81	186,71	0,97	2,09	10,23	31,83	678,24	52,44	27,28	6,73	18,29	4,23
29	0,79	117,54	129,20	1,62	2,04	8,31	32,82	535,81	40,41	33,26	8,45	17,15	4,55
30	0,97	138,64	126,10	1,61	2,34	7,63	45,62	532,72	44,45	30,31	9,26	20,21	5,45
31	1,10	116,54	134,17	1,46	2,24	9,12	30,16	476,62	49,77	33,69	8,03	16,91	4,67
32	1,36	122,43	135,22	1,57	2,17	8,41	33,43	516,66	36,81	33,91	8,08	18,67	4,18
33	1,54	109,64	115,18	1,39	2,16	7,61	28,12	507,44	29,55	29,93	6,82	16,41	3,39
34	1,60	137,23	186,01	1,34	1,72	11,05	38,97	727,34	45,23	27,09	8,12	19,53	4,73
B													
35	4,41	134,81	148,38	2,77	4,55	8,47	86,64	2487,44	66,61	14,95	11,77	12,89	1,90
36	2,46	98,87	128,81	1,98	3,95	7,54	56,38	1065,33	52,21	13,16	7,77	10,73	1,05
37	2,28	111,55	155,34	2,72	4,45	7,97	75,73	1940,43	69,21	14,77	10,37	10,90	1,38
38	4,47	120,39	159,19	2,36	4,20	9,16	70,21	2535,38	58,21	15,28	10,38	12,41	1,23
39	2,13	124,19	142,27	2,62	3,88	8,30	73,21	2414,12	66,13	14,89	9,65	12,16	1,27
40	3,94	124,10	175,16	2,65	4,39	9,11	72,51	2254,01	63,83	16,87	10,21	15,10	1,30
41	2,28	97,82	130,41	2,10	3,92	7,35	67,29	2765,10	41,98	12,27	8,69	10,88	1,14
42	3,33	123,71	151,51	2,61	4,08	7,87	66,81	1702,88	54,77	16,31	9,04	14,45	1,11
43	4,52	148,22	173,71	2,37	4,49	9,10	68,61	2435,23	66,10	16,62	9,42	12,78	1,21

(continua)

Tabela 3 – (continuação)

B	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U
44	5,81	125,28	177,41	2,73	4,83	8,81	74,27	2156,54	68,23	17,41	10,33	13,14	1,35
45	3,37	109,34	127,16	2,08	4,97	8,78	64,57	2340,55	55,71	10,92	8,46	10,23	1,56
46	2,79	104,21	129,37	2,35	4,01	8,60	60,71	2310,10	60,14	13,44	9,13	11,61	1,89
47	3,10	96,17	145,10	2,24	4,57	7,88	61,29	2599,12	49,29	13,19	8,18	11,68	1,14
48	2,10	121,32	152,19	2,71	4,11	8,77	89,11	2119,39	64,33	15,81	10,89	12,43	1,70
49	3,21	127,18	166,37	2,63	4,10	9,99	80,93	2223,81	72,91	17,18	11,22	14,13	1,23
50	2,01	142,28	166,41	3,07	4,13	8,35	86,41	2376,91	72,17	16,91	11,61	13,94	1,42
51	1,54	108,12	134,33	2,52	3,20	7,87	64,19	1961,61	63,23	12,95	8,89	9,81	1,31
52	1,41	111,44	156,81	2,37	3,39	9,69	71,43	2938,12	60,81	14,51	9,79	12,15	2,54
53	2,23	84,91	125,67	1,86	3,12	6,81	59,67	3151,34	49,06	10,23	7,54	9,06	1,46
54	3,64	99,92	139,84	2,18	3,50	9,39	66,83	2514,23	47,12	12,81	8,93	10,89	1,38
55	2,73	122,15	133,28	2,57	3,86	6,37	83,41	1487,33	64,18	15,21	10,14	12,67	1,21
56	2,51	134,01	182,44	2,28	3,28	9,86	70,78	1699,11	57,29	18,29	9,83	17,25	1,67
57	1,82	103,80	118,54	2,23	3,49	5,74	72,91	2260,81	46,18	12,61	9,24	9,71	1,24
58	2,12	112,59	138,71	2,31	3,78	8,43	62,79	2254,21	49,27	12,64	8,43	12,19	0,96
59	2,13	107,31	122,66	2,58	3,35	5,42	76,01	2129,13	63,61	13,66	10,18	9,75	1,18
60	1,61	99,21	148,43	2,28	3,51	7,87	67,68	1982,88	52,51	12,68	8,94	10,23	1,22
61	2,28	127,78	143,74	2,39	3,44	8,19	71,41	2052,61	62,18	14,81	9,28	1244	1,27
62	1,23	126,77	150,19	2,67	3,44	9,34	83,46	1617,12	51,19	17,29	11,39	13,54	1,32
63	2,41	133,21	148,01	3,06	3,78	8,16	80,94	2957,32	64,61	15,38	11,79	10,55	1,72
64	2,21	100,08	145,30	2,07	3,72	9,97	58,23	1919,32	54,51	13,21	8,14	11,48	1,23
65	2,41	117,45	154,65	1,98	3,53	8,28	61,12	1496,39	49,23	13,65	7,51	11,89	1,84

(continua)

Tabela 3 – (continuação)

B	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U
66	2,19	98,23	130,10	2,07	3,26	7,89	61,17	2484,19	37,31	12,44	8,07	9,90	1,45
67	2,10	127,61	157,78	2,78	3,51	8,50	77,31	1526,13	62,38	17,71	10,83	15,44	0,81
68	1,61	114,10	180,53	1,95	3,04	10,01	54,95	1604,22	47,15	15,33	7,03	12,21	1,42
69	2,75	115,12	145,91	2,48	3,15	7,41	70,30	2166,17	61,66	13,95	9,36	11,66	1,51
70	1,61	105,78	150,12	2,42	3,09	7,73	61,81	2437,34	47,44	12,81	8,73	11,77	1,35
71	1,05	121,71	141,82	2,84	3,26	7,67	87,15	1408,54	59,33	14,96	11,28	12,29	1,56
72	2,78	115,88	155,95	3,04	3,58	7,60	79,26	1725,71	62,11	15,77	10,79	12,32	1,38
73	2,79	123,19	186,11	2,72	3,32	8,61	71,67	2367,81	59,24	17,61	8,95	13,67	1,57
74	2,10	127,03	166,39	2,45	3,59	8,24	65,68	1693,17	59,61	16,33	9,62	12,75	1,54
75	2,41	120,10	141,52	2,19	3,32	7,36	59,93	1665,81	52,62	15,76	8,76	12,90	2,05
76	2,23	118,64	184,61	2,45	3,39	9,27	69,53	2151,99	57,71	17,98	9,81	14,61	2,10
77	1,81	139,34	192,72	2,67	3,21	9,38	78,24	2183,12	57,84	19,71	10,53	15,56	1,78
78	1,29	125,62	158,77	2,79	2,98	9,20	71,38	1176,13	58,45	17,74	9,89	13,12	1,42
79	2,01	132,81	169,26	2,98	3,49	9,30	77,67	1037,82	60,67	17,81	10,37	14,43	1,73
80	1,49	113,01	137,08	2,33	3,16	8,18	69,42	1536,91	50,87	13,17	8,99	11,61	1,50
81	3,61	111,10	156,11	2,33	3,45	8,47	64,94	2140,23	48,33	14,73	8,56	12,27	1,39
82	0,79	106,22	184,32	2,48	2,99	8,91	73,91	1512,19	60,01	19,71	9,84	12,85	1,21
83	1,14	138,31	173,01	2,60	2,55	9,62	73,75	1638,61	80,12	17,24	9,97	14,17	1,37
84	2,21	116,81	130,10	2,24	2,80	9,23	70,16	1660,52	60,17	12,21	9,68	10,51	1,55
85	0,91	138,12	195,31	2,24	2,14	8,18	54,26	1928,13	39,35	18,33	7,34	13,29	1,64
86	1,88	152,10	150,61	2,56	2,90	7,63	79,34	2188,08	59,71	14,66	11,19	12,15	1,13
87	2,44	159,07	215,66	2,63	2,06	9,18	72,55	1384,71	55,92	18,29	10,51	18,19	2,05

(continua)

Tabela 3 – (continuação)

B	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U
88	1,91	119,71	185,29	2,83	2,56	8,82	66,87	1941,23	64,43	20,33	10,12	15,58	1,06
89	0,91	133,81	188,61	2,11	1,97	7,29	57,85	1729,01	49,88	16,91	7,88	12,93	1,21
90	1,51	114,34	145,33	2,06	2,55	8,13	58,45	1625,21	57,29	15,82	7,86	13,41	1,11
91	1,67	149,81	142,51	2,31	2,71	7,83	70,35	1851,92	57,41	17,18	10,57	16,10	1,13
92	1,81	143,71	168,27	2,52	2,67	8,21	78,74	1322,77	53,49	16,71	10,69	14,66	1,38
93	1,21	137,61	144,69	2,56	2,83	8,49	72,61	1706,51	59,21	15,18	10,18	12,73	1,14
94	1,56	105,54	135,33	2,12	2,45	9,29	60,73	1015,01	46,41	14,91	8,16	13,79	1,32
95	1,05	127,41	171,99	2,19	1,85	6,51	57,97	841,09	54,01	17,28	8,11	12,51	1,23
96	2,10	117,23	183,12	2,27	2,65	8,17	61,44	1690,49	59,12	17,76	8,21	14,18	1,46
97	2,31	105,18	143,31	2,09	2,23	8,51	62,56	1250,40	61,08	14,41	8,83	15,19	1,68
98	1,19	120,14	184,08	2,57	2,72	9,66	68,42	1835,20	50,71	16,55	9,47	13,81	1,40
99	2,36	105,65	161,47	2,19	2,91	9,26	63,65	2499,13	50,44	15,18	8,26	12,88	1,26
100	1,91	85,55	147,39	2,33	2,88	10,22	61,51	1480,31	44,69	14,12	9,28	11,72	1,62
101	2,73	117,56	187,18	2,18	2,74	11,01	67,33	2627,23	57,49	16,19	9,05	14,94	2,41
102	1,89	122,14	160,91	2,55	2,93	8,63	72,41	1712,97	63,21	16,46	9,88	11,15	1,24
103	1,31	152,29	158,01	2,57	2,48	7,45	80,73	1985,76	68,11	15,30	10,17	12,61	1,15
104	1,13	126,71	182,15	2,21	1,81	9,86	68,91	1284,51	50,14	17,51	9,31	14,87	1,57
105	1,47	115,23	147,18	2,34	2,75	7,77	65,34	2181,33	47,20	14,60	8,91	11,56	1,24
106	1,91	116,11	130,21	2,13	2,60	7,84	66,58	1353,12	44,32	12,77	8,15	11,26	1,28
107	4,81	139,66	222,67	2,52	2,49	8,75	64,71	150,39	57,60	19,91	9,03	15,74	1,79

(continua)

Tabela 3 – (continuação)

B	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U
108	2,13	146,34	242,04	3,12	3,79	10,29	84,71	4068,06	81,03	20,40	12,18	14,41	1,15
109	1,51	180,66	275,03	3,58	2,25	8,29	91,43	700,01	80,05	21,93	14,34	19,15	1,83
110	1,71	227,03	275,12	3,83	2,29	9,81	116,6	1187,34	93,01	16,15	13,85	14,04	1,64
C													
111	2,47	85,28	214,55	3,38	1,62	11,89	37,69	155,91	53,11	43,96	10,80	5,23	1,25
112	1,84	102,81	230,13	3,49	1,39	12,03	45,56	144,22	51,19	45,30	11,41	7,73	1,33
113	3,10	65,38	212,20	2,85	1,29	11,10	33,51	138,55	50,10	42,61	9,73	6,81	1,61
114	2,59	54,56	203,09	2,95	1,25	11,20	34,16	138,34	44,13	44,73	9,61	6,79	1,24
115	1,41	93,24	243,11	3,44	1,25	13,64	40,97	189,87	54,50	45,81	11,41	6,11	1,26
116	1,69	110,77	260,20	3,81	1,26	12,09	48,38	159,45	59,70	44,16	13,25	5,83	0,97
117	1,71	95,27	204,85	3,42	1,34	13,08	43,52	192,91	48,15	50,11	11,17	6,75	1,25
118	1,35	89,23	249,77	3,42	1,44	12,19	39,51	165,61	62,18	48,96	11,18	5,75	1,44
119	1,80	92,71	253,19	3,60	1,49	13,65	44,24	125,23	63,19	48,31	11,70	6,43	1,24
120	1,73	75,81	205,05	2,94	0,85	13,87	31,86	121,34	45,39	41,85	8,98	6,96	1,67
121	1,61	56,43	183,81	2,39	0,81	11,92	28,12	120,81	35,34	43,40	7,45	6,44	1,58
122	2,01	61,71	212,53	2,98	0,86	11,53	34,17	125,32	48,40	47,45	9,17	6,95	1,79
123	2,28	62,54	195,44	2,82	0,91	11,81	29,33	92,91	46,11	42,51	9,21	7,16	1,33
124	1,91	52,51	195,13	2,68	0,92	12,91	26,21	136,71	43,99	43,22	8,46	7,37	1,42
125	1,94	110,01	218,17	3,29	0,75	12,14	37,85	181,41	60,94	39,41	10,30	5,24	1,15
126	1,61	78,93	230,22	3,23	0,86	11,19	41,16	189,32	69,34	40,03	11,31	5,13	1,11

(continua)

Tabela 3 – (continuação)

C	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U
127	1,21	68,94	204,91	2,92	0,84	11,20	32,88	191,23	51,23	44,31	10,21	6,82	1,63
128	1,48	70,91	192,55	3,01	0,83	12,10	36,15	117,17	61,13	46,11	10,32	7,46	1,52
129	1,84	97,52	238,71	3,27	0,80	12,39	38,18	167,81	52,10	42,38	10,46	6,27	1,87
130	2,61	67,83	212,61	2,94	1,12	11,34	31,81	132,77	41,13	39,91	9,43	6,45	1,38
131	1,90	61,12	215,10	2,91	1,03	11,66	30,83	176,91	45,34	46,20	9,12	7,34	1,56
132	2,83	68,65	215,40	2,95	1,16	12,98	34,43	145,34	48,09	45,12	9,26	6,33	1,44
133	1,61	82,11	187,14	3,21	1,11	11,39	37,21	260,44	47,13	37,22	9,81	4,83	1,23
134	1,57	90,8	303,49	3,23	1,21	11,20	39,51	266,57	52,48	41,71	10,27	5,62	1,12
135	1,44	95,24	245,10	3,55	1,14	12,10	44,81	187,82	57,10	43,12	11,38	5,81	1,46
136	2,87	67,55	205,32	2,83	1,05	11,22	36,60	93,12	42,13	43,51	10,14	7,25	1,97
137	1,29	63,44	183,33	2,85	0,98	11,67	33,91	130,08	44,34	40,73	9,57	6,79	1,78
138	2,76	67,82	236,34	3,02	1,10	11,92	33,87	139,11	55,49	41,20	9,99	6,32	1,49
139	1,75	87,81	241,10	3,31	1,13	11,75	40,81	200,88	71,21	45,61	11,67	7,53	1,35
140	1,15	245,61	162,04	4,72	0,70	11,64	53,56	172,34	72,20	48,43	15,62	5,37	1,09

Fonte: (MUNITA, et al., 2003; CARVALHO, 2018)

8 RESULTADOS E DISCUSSÃO

A ausência de valores em uma base de amostras é um problema, pois a maioria dos métodos de estatística assume que a base está completa. Uma abordagem simples, muitas vezes utilizada para resolução desse problema, consiste na exclusão das amostras ou das variáveis que apresentam valores ausentes. Entretanto, tais amostras excluídas podem conter informação importante que não deve ser ignorada na análise dos resultados. Nesse sentido, uma outra forma de tratar o problema é substituir os valores faltantes, utilizando métodos de imputação de dados.

Os métodos de imputação: média, *autoencoder*, agrupamento e c-médias foram aplicados na base de amostras mostrada na Tabela 3. Dos elementos La, Na e Sm, selecionou-se duas amostras de cada sítio (A, B e C), pois ao serem irradiados com nêutrons térmicos dão origem à Na^{24}, La^{140} e Sm^{153} cujas meias vidas são de 15,0 h, 40,3 h e 47,3 h, respectivamente. Diante disso, dependendo da concentração desses elementos nas amostras e se a medida não é realizada no tempo de decaimento certo, a precisão da análise é prejudicada. Ademais, pode ocorrer, também, que no espectro não apareça o pico da energia do radioisótopo como consequência do tempo de decaimento. Nesses casos, o analista elimina o elemento ou a amostra da base, o que pode prejudicar a interpretação dos resultados. Uma maneira de contornar esse problema é através da imputação de dados. Para estudar os métodos de imputação, os resultados das amostras selecionadas foram excluídos da base. A Tabela 4 apresenta as amostras dos sítios A, B e C, e as frações de massa das variáveis La, Na e Sm que foram excluídas da base.

Para a análise do desempenho de cada um dos métodos de imputação, utilizou-se:

1 – Análise de componentes principais, (PCA - Principal Component Analysis);

2 – Raiz do erro médio quadrático normalizado (NRMSE);

3 – Tempo de processamento de cada método;

4 – Erro cometido pelos métodos de agrupamento após a imputação de dados.

Os métodos de imputação foram aplicados de duas maneiras:

1 – Utilizando os resultados dos três sítios, simultaneamente, para o cálculo dos valores imputados (Figuras 4, 5, 6, 7, 9, 10, 11,12, 14, 15, 16 e 17, Tabelas 5, 7 e 9);

2 – Empregando para o cálculo dos valores imputados, apenas, as amostras do sítio que possuem valores ausentes (Tabelas 6, 8 e 10).

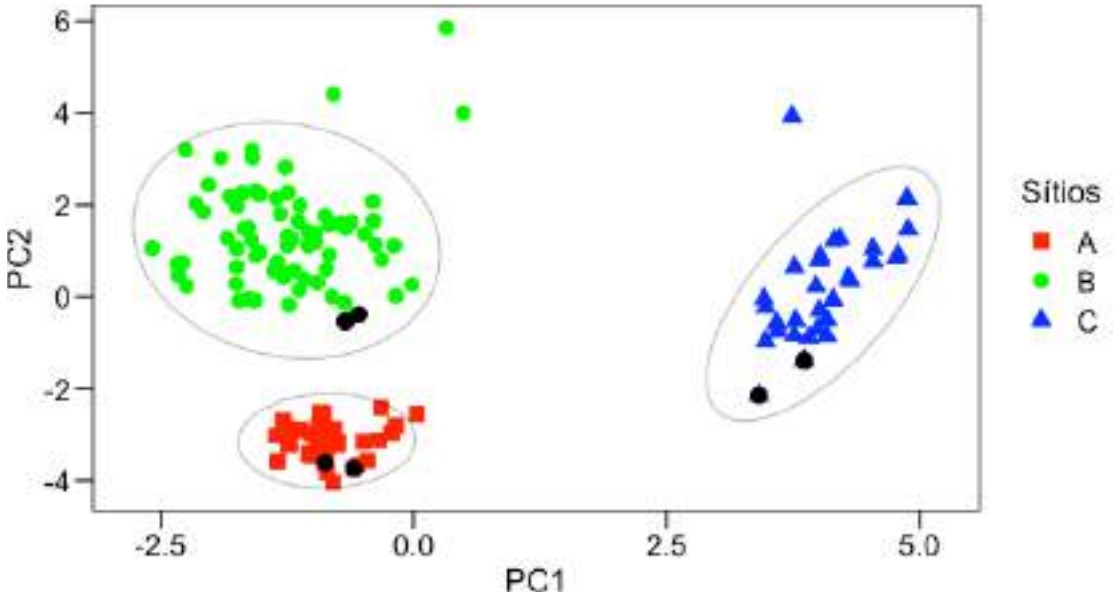
Tabela 4 – Sítio, amostras e resultados, em $\mu g\ g^{-1}$, das frações de massa que foram excluídas

	Elemento					
	La		**Na**		**Sm**	
Sítios	amostra	valor excluído	amostra	valor excluído	amostra	valor excluído
A	13	27,26	1	302,12	13	6,35
	33	28,12	6	328,49	25	6,51
B	68	54,95	95	841,09	68	7,03
	85	54,26	109	700,01	85	7,34
C	121	28,12	123	92,91	121	7,45
	124	26,21	136	93,12	124	8,46

Fonte: Próprio autor

Inicialmente, as frações de massa das amostras, Tabela 3, foram padronizadas utilizando a transformada z-score. Em seguida, foi aplicada a PCA para gerar o gráfico de dispersão da base completa (sem valores ausentes), apresentado na Figura 3, as amostras dos 3 sítios formam 3 grupos composicionais distintos. As elipses representam um nível de confiança 98%. Os pontos, destacados em negrito, em cada grupo, correspondem às frações de massa do La, selecionadas para serem imputadas (amostras:13, 33, 68, 85, 121 e 124).

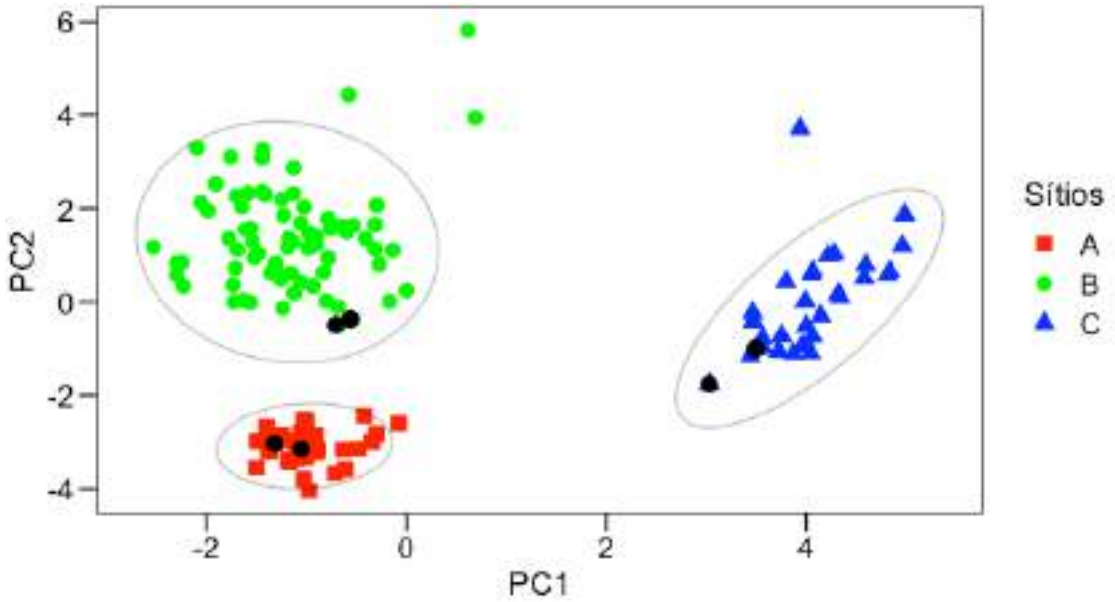
Figura 3 – Componente principal 1 vs componente principal 2 para todas as amostras, $n = 140$. As elipses representam um nível de confiança de 98%.



Fonte: Próprio autor

Posteriormente, foram calculados os valores imputados para os sítios: A (amostras 13 e 33), B (amostras 68 e 85) e C (amostras 121 e 124), utilizando o método de imputação pela média (equação 41). A base com valores imputados foi padronizada com a transformada z-score, em seguida utilizou-se a PCA para visualização dos grupos conforme exibe a Figura 4. Os pontos, destacados em negrito, em cada grupo, correspondem às frações de massa do La, dos valores imputados. A Figura 4 mostra que as amostras com valores imputados permaneceram nos mesmos grupos composicionais e no interior das elipses. Dessa forma, não houve alteração na estrutura dos grupos.

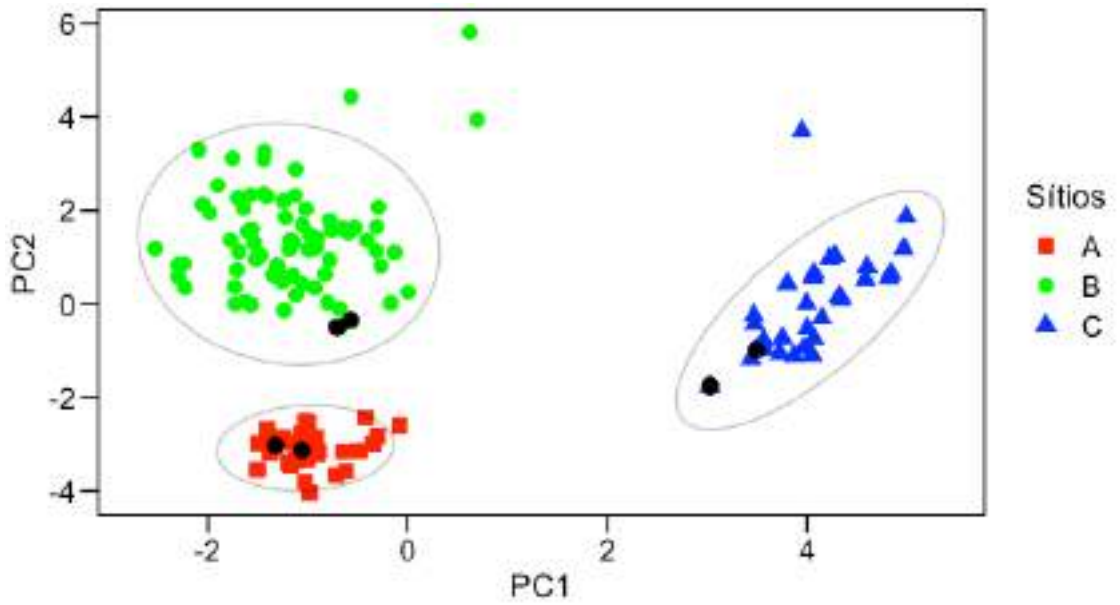
Figura 4 – Resultado da imputação pela média para La, $n = 140$. As elipses representam um nível de confiança de 98 %



Fonte: Próprio autor

Em seguida, foram calculados os valores imputados para os sítios: A (amostras 13 e 33), B (amostras 68 e 85) e C (amostras 121 e 124), utilizando o método de imputação por *autoencoder* discutido na subseção 4.2. Após padronização da base com a transformada z-score, utilizou-se a PCA para visualização dos grupos, de acordo com a Figura 5. Em cada grupo, os pontos, destacados em negrito, correspondem às frações de massa do La, dos valores imputados. A Figura 5 revela que as amostras com valores imputados formaram parte dos mesmos grupos composicionais e no interior das elipses, não ocorrendo alteração na estrutura dos grupos.

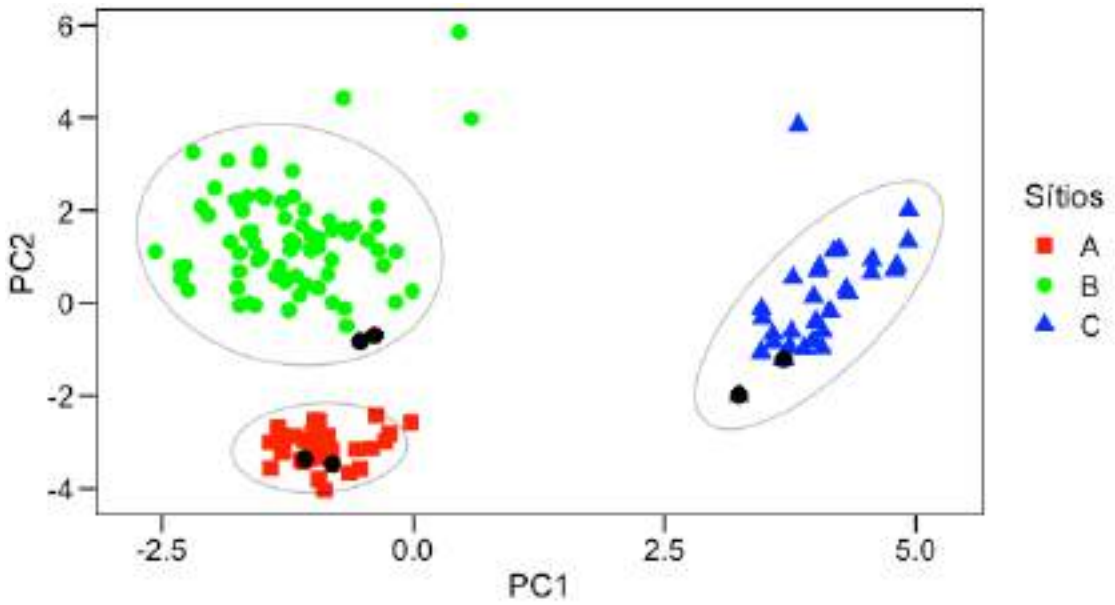
Figura 5 – Resultado da imputação por autoencoder para La, $n = 140$. As elipses respresentam um nível de confiança de 98%.



Fonte: Próprio autor

Utilizando o método de imputação por agrupamento, discutido na subseção 4.3, foram imputados os valores para os sítios: A (amostras 13 e 33), B (amostras 68 e 85) e C (amostras 121 e 124) inseridos na base. Depois da padronização da base com a transformada z-score, utilizou-se a PCA para visualização dos grupos conforme apresenta a Figura 6. Os pontos destacados em negrito, em cada grupo, correspondem às frações de massa do La, dos valores imputados. As amostras com valores imputados permaneceram nos mesmos grupos composicionais e no interior das elipses. Logo, não houve alteração na estrutura dos grupos.

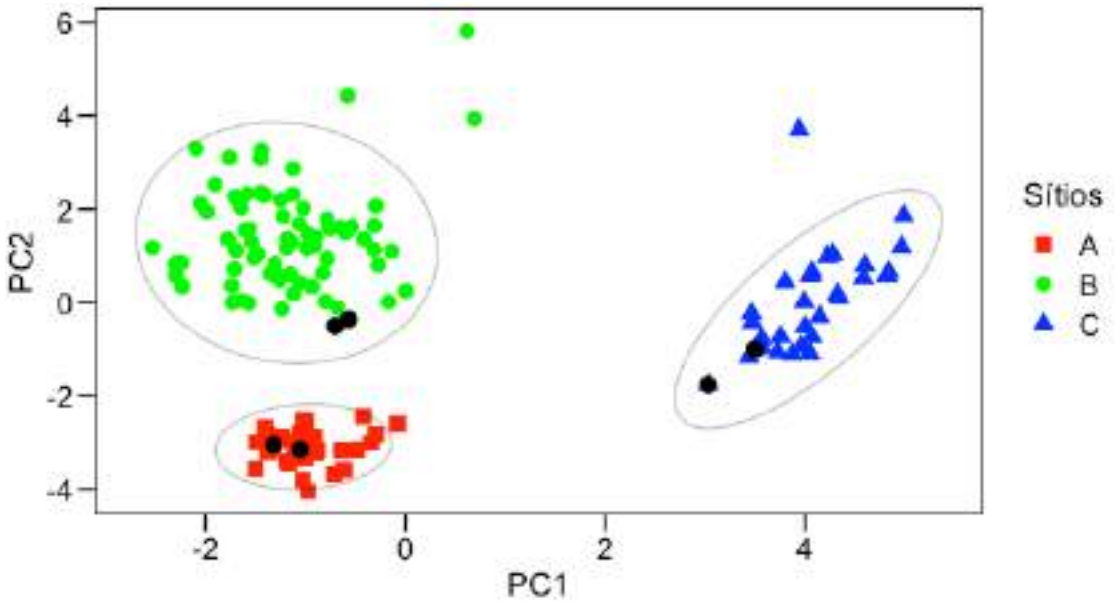
Figura 6 – Resultado da imputação por agrupamento para La, $n = 140$. As elipses representam um nível de confiança de 98%.



Fonte: Próprio autor

A seguir, foram calculados os valores imputados para os sítios: A (amostras 13 e 33), B (amostras 68 e 85) e C (amostras 121 e 124), utilizando o método de imputação pelo c-médias (equação 46). As concentrações foram padronizadas com a transformada z-score, utilizou-se a PCA para visualização dos grupos conforme mostra o gráfico da Figura 7. Os pontos destacados em negrito, em cada grupo, correspondem às frações de massa do La, dos valores imputados. Na Figura 7 observou-se que as amostras com valores imputados permaneceram nos mesmos grupos composicionais e no interior das elipses, ou seja não havendo dessa maneira alteração na estrutura dos grupos.

Figura 7 – Resultado da imputação pelo c-médias para La, $n = 140$. As elipses representam um nível de confiança de 98%.



Fonte: Próprio autor

A Tabela 5 apresenta o valor do NRMSE para La para cada um dos métodos de imputação do elemento La, utilizando a equação 47. Os valores imputados foram calculados utilizando as amostras dos 3 sítios, simultaneamente. O método baseado em agrupamento (0,93) foi superior aos métodos: média (1,57), *autoencoder* (1,59) e c-médias (1,57).

Tabela 5 – Método e valor do NRMSE para La, $n = 140$

Método	**NRMSE**
Média	1,57
Autoencoder	1,59
Agrupamento	0,93
C-Médias	1,57

Fonte: Próprio autor

Ao aplicar os métodos de imputação de dados em cada sítio, individualmente, calculou-se os valores do NRMSE para La e depois somou-se, de modo que os resultados são mostrados na Tabela 6. O pior desempenho ocorreu com o método baseado em agrupamento (0,83) e o melhor com o método baseado em lógica *fuzzy* (0,23). Ao comparar, ainda as Tabelas 5 e 6 observou-se uma diminuição do NRMSE.

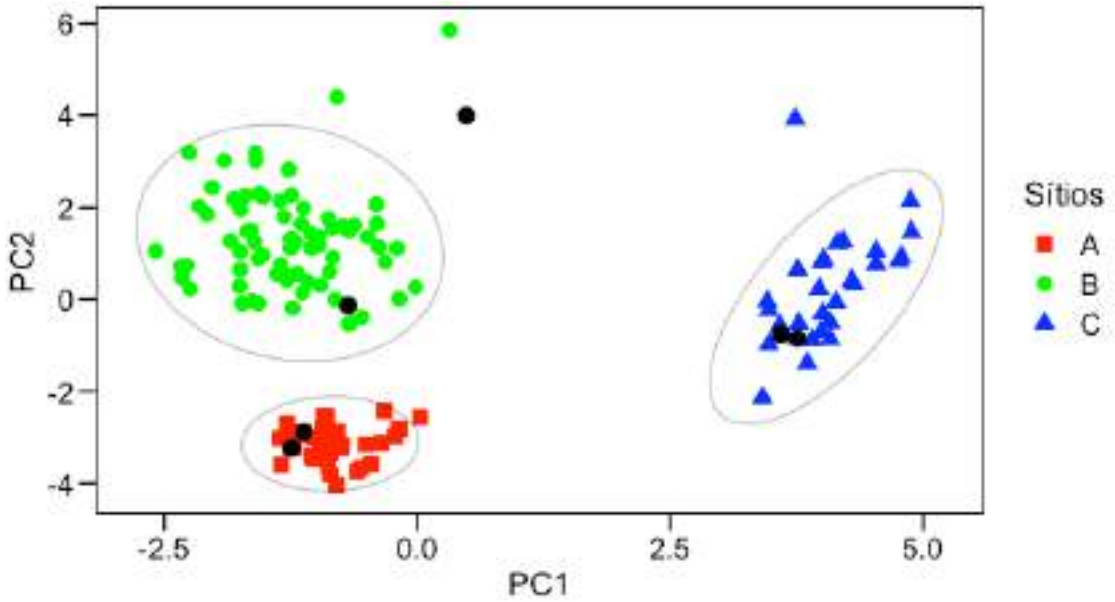
Tabela 6 - Método e valor do NRMSE para La (n =34, 76 e 30)

Método	**NRMSE**
Média	0,74
Autoencoder	0,73
Agrupamento	0,83
C-Médias	0,23

Fonte: Próprio autor

Para Na, primeiramente, padronizaram-se as frações de massa das amostras utilizando a transformada z-score. Logo depois foi aplicada a PCA para gerar o gráfico de dispersão da base completa (sem valores ausentes), apresentado na Figura 8. A Figura 8 mostra que as amostras dos 3 sítios formam 3 grupos composicionais distintos, as elipses representam um nível de confiança 98%. Os pontos, destacados em negrito, em cada grupo, referem-se às frações de massa do Na, selecionados para serem imputados (amostras : 1, 6, 95, 109, 123 e 136).

Figura 8 – Componente principal 1 vs componente principal 2 para todas as amostras, n=140. As elipses representam um nível de confiança de 98%.

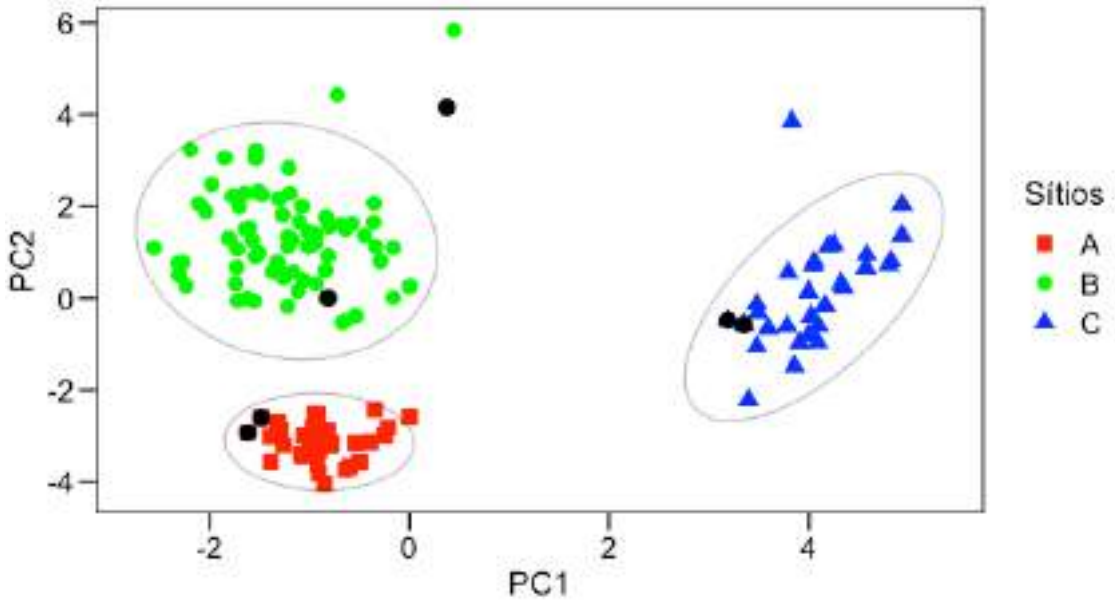


Fonte: Próprio autor

Os valores imputados para os sítios: A (amostras 1 e 6), B (amostras 95 e 109) e C (amostras 123 e 136), foram calculados utilizando o método de imputação pela média (equação 41). A base com valores imputados foi

padronizada com a transformada z-score, em seguida utilizou-se a PCA para visualização dos grupos conforme mostra a Figura 9. Os pontos destacados em negrito, em cada cluster, dizem respeito às frações de massa do Na, dos valores imputados. A Figura 9 mostra que cinco amostras com valores imputados permaneceram nos mesmos grupos composicionais e no interior das elipses e uma amostra com valor imputado fora da elipse. Dessa forma, não houve alteração na estrutura dos grupos.

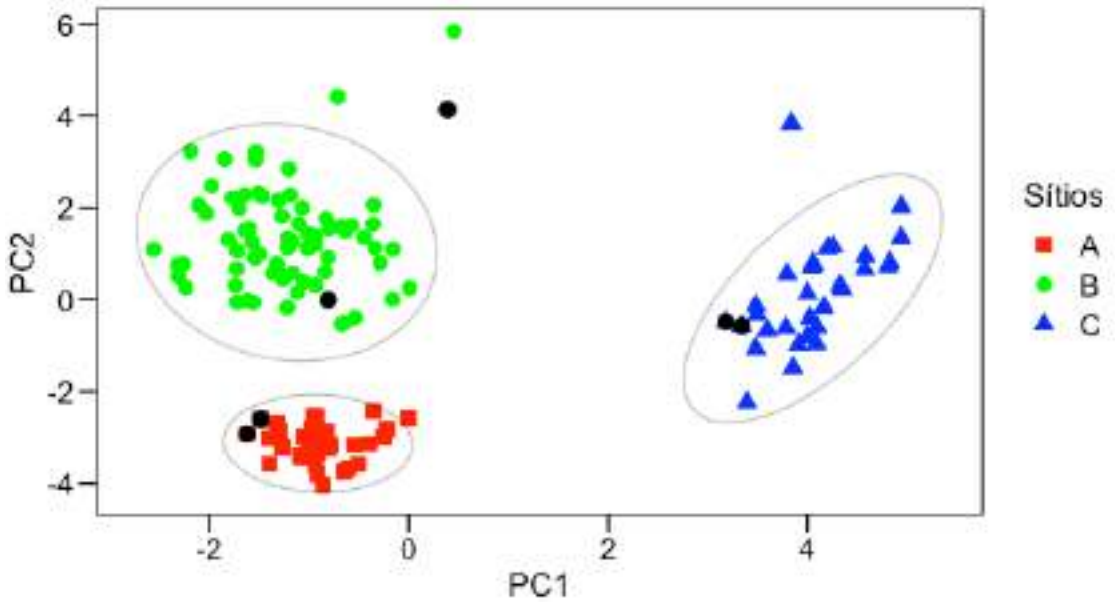
Figura 9 – Resultado da imputação pela média para Na, $n = 140$. As elipses representam um nível de confiança de 98%



Fonte: Próprio autor

Mais adiante, foram calculados os valores imputados para os sítios: A (amostras 1 e 6), B (amostras 95 e 109) e C (amostras 123 e 136), utilizando o método de imputação por *autoencoder*, tratado na seção 4.2. Após padronização da base com a transformada z-score, utilizou-se a PCA para visualização dos grupos como mostra Figura 10. Em cada grupo, os pontos, destacados em negrito, correspondem às frações de massa do Na, dos valores imputados. A Figura 10 mostra que cinco amostras com valores imputados permaneceram nos mesmos grupos composicionais e no interior das elipses, além de ainda, uma amostra fora das elipses. Assim, não ocorreu mudança na disposição dos grupos.

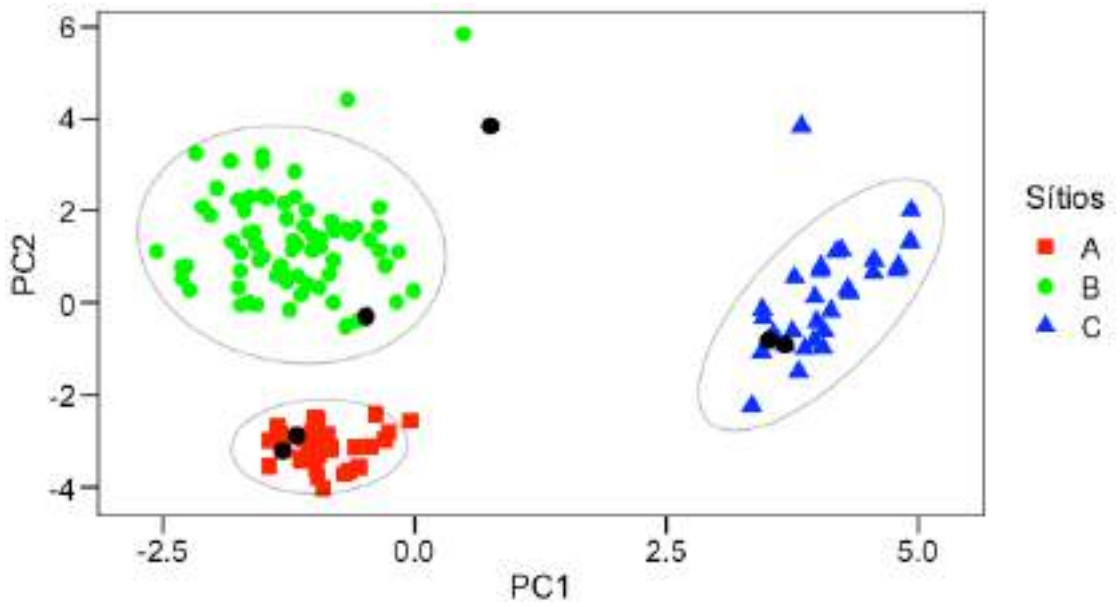
Figura 10 – Resultado da imputação por autoencoder para Na, $n = 140$. As elipses respresentam um nível de confiança de 98%.



Fonte: Próprio autor

Nessa pespectiva, ao aplicar o método de imputação por agrupamento, tratado na subseção 4.3 foram imputados os valores para os sítios: A (amostras 1 e 6), B (amostras 95 e 109) e C (amostras 123 e 136), e inseridos na base. Depois da padronização dos resultados com a transformada z-score, utilizou-se a PCA para visualização dos grupos composicionais conforme revela o gráfico da Figura 11. Os pontos destacados em negrito, em cada grupo, correspondem às frações de massa do Na, dos valores imputados. As amostras com valores imputados permaneceram nos mesmos grupos composicionais e no interior das elipses, há também uma amostra com valor imputado fora das elipses, como mostra a Figura 11. Logo, não houve alteração na estrutura dos grupos.

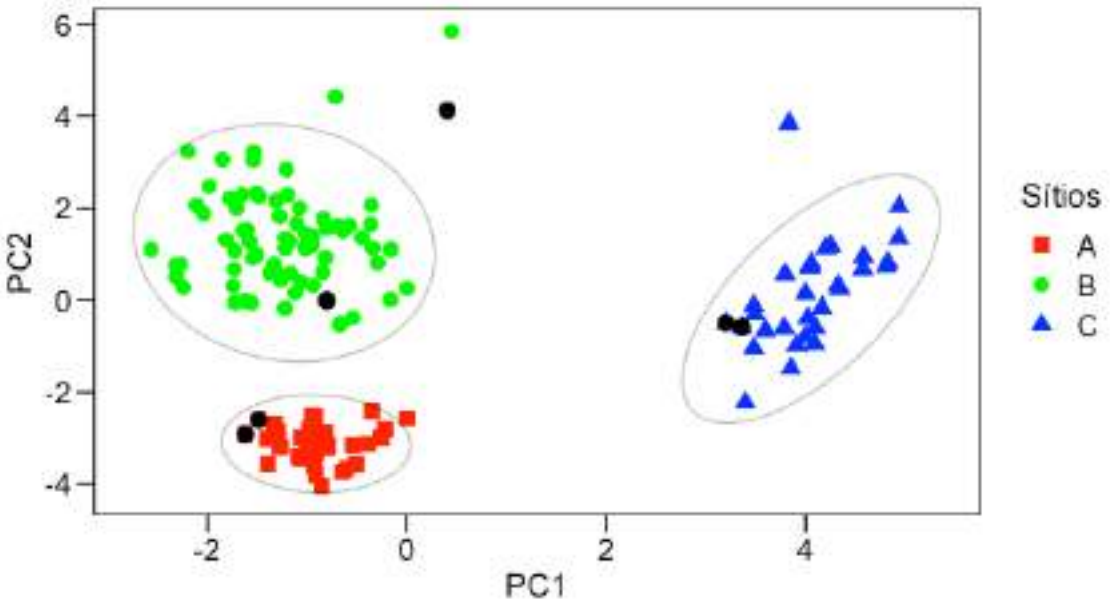
Figura 11 – Resultado da imputação por agrupamento para Na, $n = 140$. As elipses representam um nível de confiança de 98%.



Fonte: Próprio autor

Foram calculados os valores imputados para os sítios: A (amostras 1 e 6), B (amostras 95 e 109) e C (amostras 123 e 136), utilizando o método de imputação pelo c-médias (equação 46). Com isso, realizada a padronização da base com a transformada z-score, utilizou-se a PCA para visualização dos grupos conforme mostra Figura 12. Os pontos destacados em negrito, em cada grupo, referem-se às frações de massa do Na, dos valores imputados. Na Figura 12 observou-se que cinco amostras com valores imputados permaneceram nos mesmos grupos composicionais e no interior das elipses, de modo que também houve uma amostra com valor imputado fora das elipeses. Não havendo dessa maneira alteração na disposição dos grupos.

Figura 12 – Resultado da imputação pelo c-médias para Na, $n = 140$. As elipses representam um nível de confiança de 98%.



Fonte: Próprio autor

A Tabela 7 mostra o valor do NRMSE para Na dos quatro métodos, calculado ao utilizar as frações de massa dos três sítios. O desempenho foi obtido com o método baseado em agrupamento (1,02), foi superior aos métodos: média (2,80), autoencoder (2,83) e c-médias (2,71).

Tabela 7 - Método e valor do NRMSE para Na (n =140)

Método	**NRMSE**
Média	2,80
Autoencoder	2,83
Agrupamento	1,02
C-Médias	2,71

Fonte: Próprio autor

Assim, aplicando os métodos de imputação em cada um dos sítios individualmente, e calculando o NRMSE para Na, somando os valores, obtêm-se os valores mostrados na Tabela 8. A Tabela 8 mostra que os métodos estudados: média (0,49), *autoencoder* (0,49), agrupamento (0,48) e c-médias (0,49) apresentaram diminuição nos valores do NRMSE, quando comparados com os valores da Tabela 7.

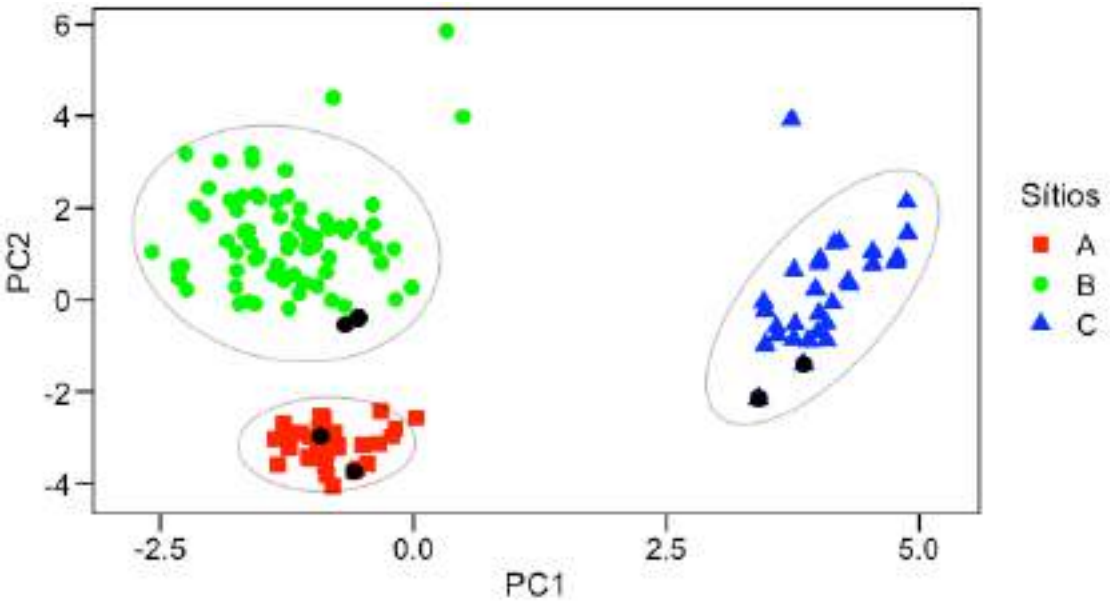
Tabela 8 - Método e valor do NRMSE para Na (n = 34, 76, 30)

Método	**NRMSE**
Média	0,49
Autoencoder	0,49
Agrupamento	0,48
C-Médias	0,49

Fonte: Próprio autor

Para o Sm, preliminarmente, aplicou-se a transformada z-score para padronização da base. Em seguida utilizou-se a PCA para gerar o gráfico de dispersão da base completa (sem valores ausentes), apresentado na Figura 13. Na Figura 13 as amostras dos 3 sítios formam 3 grupos composicionais distintos. As elipses representam um nível de confiança 98%. Os pontos, destacados em negrito, em cada grupo, correspondem às amostras com valores do Sm selecionados para serem imputados (amostras:13, 25, 68, 85, 121 e 124).

Figura 13 – Componente principal 1 vs componente principal 2 para todas as amostras, $n = 140$. As elipses representam um nível de confiança de 98%.

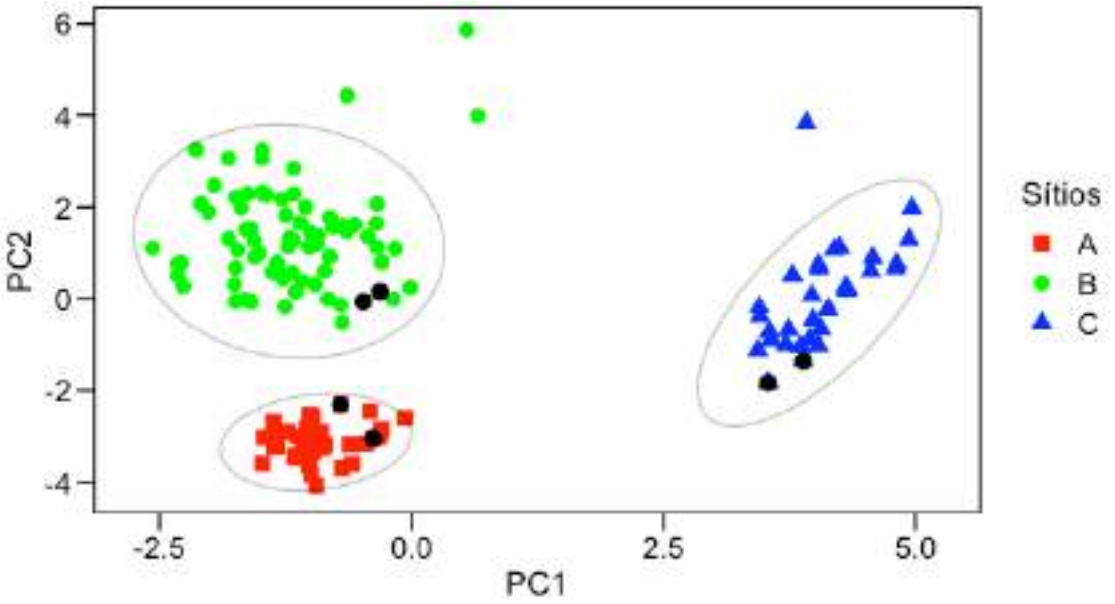


Fonte: Próprio autor

Posteriormente, foram calculados os valores imputados para os sítios: A (amostras 13 e 25), B (amostras 68 e 85) e C (amostras 121 e 124), utilizando o método de imputação da média (equação 41). A base com valores imputados foi padronizada com a transformada z-score, em seguida utilizou-se a PCA para

visualização dos grupos como mostra a Figura 14. Os pontos, destacados em negrito, em cada grupo, dizem respeito às frações de massa do Sm, dos valores imputados. A Figura 14 mostra que as amostras com valores imputados permaneceram nos mesmos grupos composicionais, em cada sítio, e no interior das elipses. Dessa forma, não houve alteração na estrutura dos grupos.

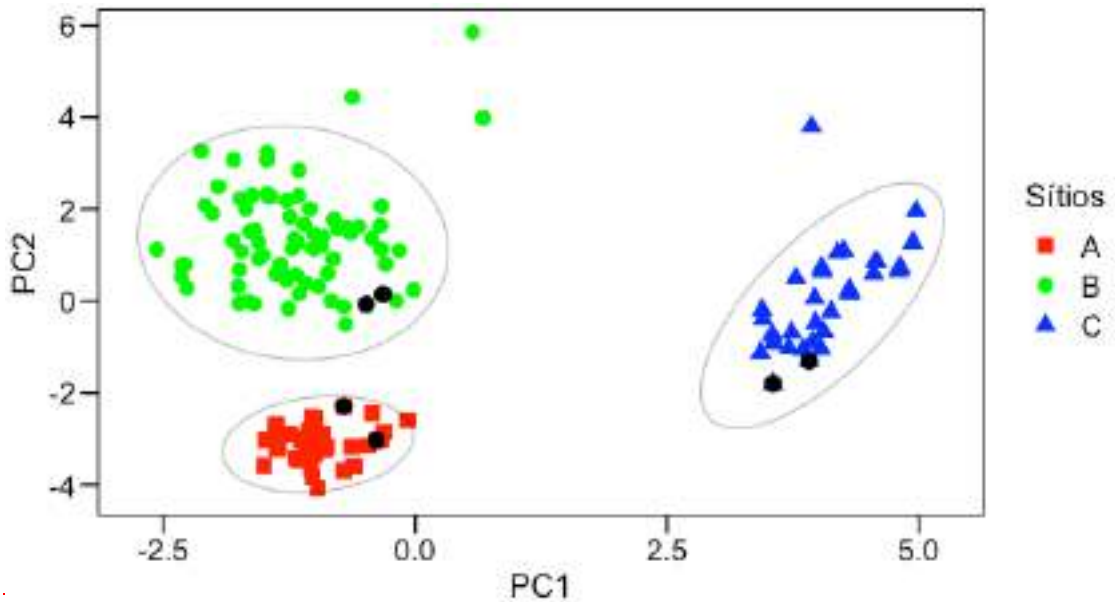
Figura 14 – Resultado da imputação pela média para Sm, $n = 140$. As elipses representam um nível de confiança de 98%.



Fonte: Próprio autor

Seguidamente, foram calculados os valores imputados para os sítios: A (amostras 13 e 25), B (amostras 68 e 85) e C (amostras 121 e 124), utilizando o método de imputação por *autoencoder* abordado na subseção 4.2. Após padronização da base com a transformada z-score, utilizou-se a PCA para visualização dos grupos de acordo com a Figura 15. Em cada grupo, os pontos, destacados em negrito, correspondem às frações de massa do Sm, dos valores imputados. A Figura 15 mostra que as amostras com valores imputados permaneceram nos mesmos grupos composicionais, em cada sítio, e no interior das elipses. Assim, não houve alteração na composição dos grupos.

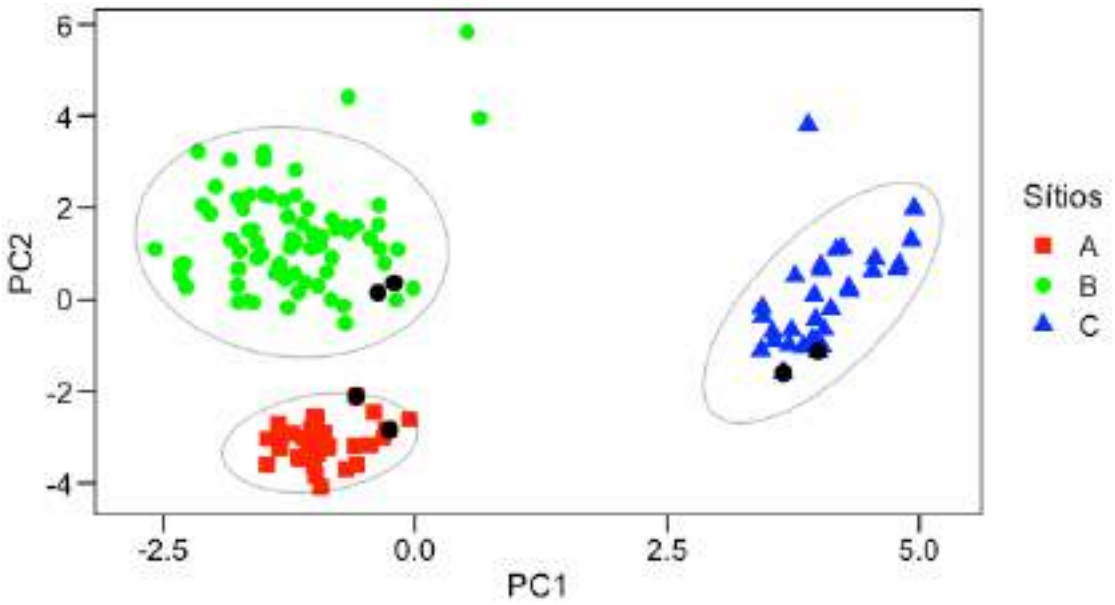
Figura 15 – Resultado da imputação por autoencoder para Sm, $n = 140$. As elipses representam um nível de confiança de 98%



Fonte: Próprio autor

Diante disso, e com a utilização do método de imputação por agrupamento abordado na subseção 4.3 foram imputados os valores para os sítios: A (amostras 13 e 25), B (amostras 68 e 85) e C (amostras 121 e 124), e inseridos na base. Depois da padronização dos resultados com a transformada z-score, utilizou-se a PCA para visualização dos grupos conforme mostra na Figura 16. Os pontos destacados em negrito, em cada grupo, correspondem às frações de massa do Sm, dos valores imputados. As amostras com valores imputados permaneceram nos mesmos grupos composicionais, em cada sítio, e no interior das elipses, como é possível identificar na Figura 16. Desse modo, não houve alteração na estrutura dos grupos.

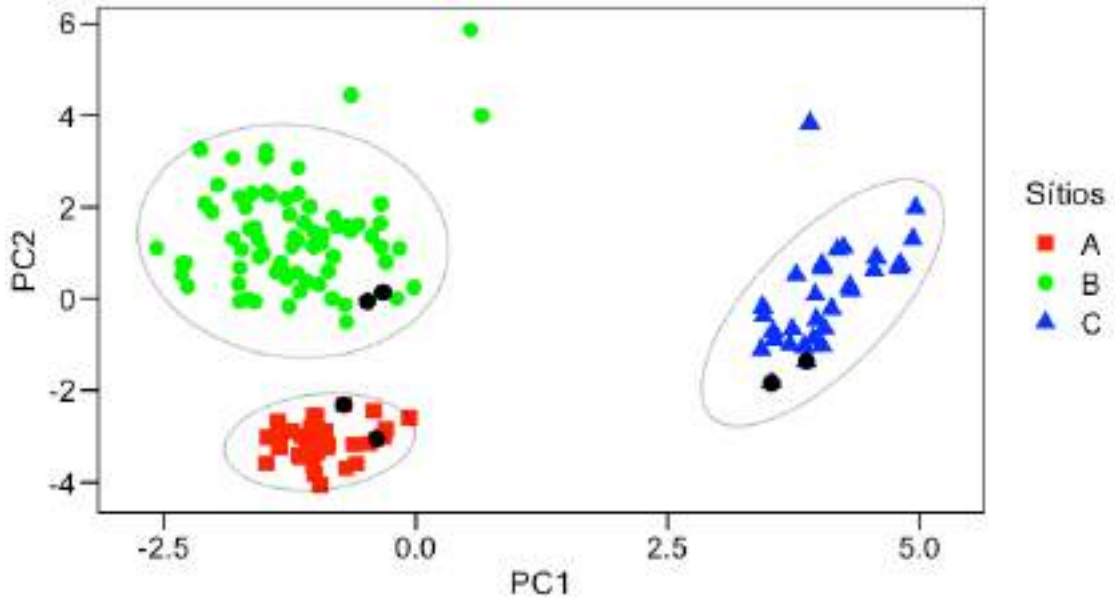
Figura 16 – Resultado da imputação por agrupamento para Sm, $n = 140$. As elipses representam um nível de confiança de 98%.



Fonte: Próprio

Foram calculados os valores imputados para os sítios: A (amostras 13 e 25), B (amostras 68 e 85) e C (amostras 121 e 124), utilizando o método de imputação pelo c-médias (equação 46). Realizada a padronização da base com a transformada z-score, utilizou-se a PCA para visualização dos grupos conforme mostra a Figura 17. Os pontos destacados em negrito, em cada grupo, correspondem às frações de massa do Sm, dos valores imputados. Na Figura 17 observou-se que não houve alteração na composição dos grupos.

Figura 17 – Resultado da imputação pelo c-médias para Sm, $n = 140$. As elipses representam um nível de confiança de 98%.



Fonte: Próprio

A Tabela 9 mostra os valores do NRMSE para Sm, calculados utilizando as frações de massa dos três sítios ao mesmo tempo. Diferente do que ocorreu nas Tabelas 5 e 7, em que o melhor desempenho foi alcançado pelo método agrupamento, a Tabela 9, mostrou que os métodos: média (2,78), autoencoder (2,86) e c-média (2,72) tiverem um desempenho superior ao método baseado em agrupamento (3,93).

Tabela 9 – Método e valor do NRMSE para Sm (n = 140)

Método	**NRMSE**
média	2,78
autoencoder	2,86
agrupamento	3,93
c-médias	2,72

Fonte: Próprio autor

Ao aplicar os métodos de imputação de dados em cada sítio, individualmente, calculou-se os valores do NRMSE para Sm e somou-se, de maneira que os resultados dos três sítios são mostrados na Tabela 10. O melhor desempenho foi alcançado com o método baseado em lógica *fuzzy* (1,06) e o pior com o agrupamento (1,18). Ao comparar as Tabelas 9 e 10 observa-se uma diminuição do NRMSE dos métodos estudados.

Tabela 10 – Método e valor do NRMSE para Sm (n =34, 76 e 30)

Método	**NRMSE**
média	1,07
autoencoder	1,08
agrupamento	1,18
c-médias	1,06

Fonte: Próprio autor

O tempo de processamento de cada um dos métodos de imputação, aplicados às variáveis La, Na e Sm, é mostrado na Tabela 11. O menor tempo de processamento, foi alcançado com o método baseado em média (0,01s/La; 0,01s/Na e 0,01s/Sm). Seguido dos métodos baseados em agrupamento (0,11s/La; 0,11s/Na e 0,11s/Sm) e autoencoder (0,51s/La; 0,46s/Na e 0,49s/Sm). No tocante ao método mais lento foi o c-médias (3,49s/La; 3,78s/Na e 3,45s/Sm).

Tabela 11 – Tempo de processamento

Método	**Elemento La**	**Elemento Na**	**Elemento Sm**
média	0,01s	0,01s	0,01s
autoencoder	0,51s	0,46s	0,49s
agrupamento	0,11s	0,11s	0,11s
c-médias	3,49s	3,78s	3,45s

Fonte: Prórprio autor

Também foi avaliado o impacto da imputação na determinação dos grupos pelos métodos de agrupamento, através da taxa de erro (equação 40), após a aplicação dos métodos de imputação no elemento La. Antes da análise de agrupamento foi aplicada a transformada logarítmica tanto na base completa como nas bases com valores imputados. A Tabela 12 mostra o desempenho dos métodos de agrupamento, comparando os resultados com valores imputados e a base completa. A tabela mostrou que não houve impacto na determinação dos grupos pelos métodos de agrupamento, após a imputação dos valores ausentes, pois não ocorreu alteração nas taxas dos erros desses métodos.

Tabela 12 – Erro cometido pelos métodos de análise de agrupamentos após a aplicação dos métodos de imputação dos valores ausente do La

Métodos	**Base**	**Métodos de imputação**			
de agrupamento	**completa**	**média**	**autoencoder**	**agrupamento**	**c-médias**
ligação simples	0 %	0 %	0 %	0 %	0 %
ligação média	0 %	0 %	0 %	0 %	0 %
ligação completa	0 %	0 %	0 %	0 %	0 %
Ward	0 %	0 %	0 %	0 %	0 %
k-médias	0 %	0 %	0 %	0 %	0 %
k-medóides	0 %	0 %	0 %	0 %	0 %
híbrido	0 %	0 %	0 %	0 %	0 %
c-médias	45,71 %	45,71	45,71 %	45,71 %	45,71
c-medóides	50,00 %	50,00	50,00 %	50,00 %	50,00
c-médias-pf	45,71 %	45,71	45,71 %	45,71 %	45,71

Fonte: Próprio Autor

A Tabela 13 exibe as taxas de erro da base com valores não imputados e imputados para Na. Os métodos de agrupamento mantiveram a taxa de erro na base completa, com frações de massa originais e com frações de massa calculadas pelos métodos de imputação.

Tabela 13 – Erro cometido pelos métodos de análise de agrupamentos após a aplicação dos métodos de imputação dos valores ausentes do Na

Métodos	**Base**	**Métodos de imputação**			
de agrupamento	**completa**	**média**	**autoencoder**	**agrupamento**	**c-médias**
ligação simples	0 %	0 %	0 %	0 %	0 %
ligação média	0 %	0 %	0 %	0 %	0 %
ligação completa	0 %	0 %	0 %	0 %	0 %
Ward	0 %	0 %	0 %	0 %	0 %
k-médias	0 %	0 %	0 %	0 %	0 %
k-medóides	0 %	0 %	0 %	0 %	0 %
híbrido	0 %	0 %	0 %	0 %	0 %
c-médias	45,71 %	45,71	45,71 %	45,71 %	45,71
c-medóides	50,00 %	50,00	50,00 %	50,00 %	50,00
c-médias-pf	45,71 %	45,71	45,71 %	45,71 %	45,71

Fonte: Próprio Autor

A Tabela 14 avalia o desempenho dos métodos de agrupamento da base completa e das bases com valores imputados, para Sm. A Tabela 14 revela, portanto, não haver diferença de desempenho entre a base completa e a base com valores imputados, quando estes foram utilizados os métodos de agrupamento hierárquicos, particionais/hard, c-médias e c-médias com polinômio fuzzificador.

Tabela 14 – Erro cometido pelos métodos de análise de agrupamentos após a aplicação dos métodos de imputação de dos valores ausentes do Sm

Métodos de agrupamento	**Base completa**	**Métodos de imputação**			
		média	**autoencoder**	**agrupamento**	**c-médias**
ligação simples	0 %	0 %	0 %	0 %	0 %
ligação média	0 %	0 %	0 %	0 %	0 %
ligação completa	0 %	0 %	0 %	0 %	0 %
Ward	0 %	0 %	0 %	0 %	0 %
k-médias	0 %	0 %	0 %	0 %	0 %
k-medóides	0 %	0 %	0 %	0 %	0 %
híbrido	0 %	0 %	0 %	0 %	0 %
c-médias	45,71 %	45,71	45,71 %	45,71 %	45,71
c-medóides	50,00 %	31,42	31,42 %	31,42 %	31,42
c-médias-pf	45,71 %	45,71	45,71 %	45,71 %	45,71

Fonte: Próprio Autor

Em seguida são apresentados os estudos com métodos de detecção de outliers.

Todo conjunto de resultados pode apresentar amostras (*outliers*) que se distinguem da maior parte das amostras. Dito isto, entre as suas causas mais prováveis de ocorrência estão: erros gerados nos procedimentos experimentais, tais como, manipulação de materiais e medidas obtidas a partir de equipamentos.

A detecção de *outliers* em um conjunto de amostras é uma etapa fundamental do pré-processamento de dados, sem a qual a análise pode ser prejudicada. Uma vez que os *outliers* afetam, por exemplo, estimativas como a média e o desvio padrão, resultando em valores superestimados ou subestimados.

Além de impactar na separação das amostras, podendo afetar o desempenho de métodos estatísticos, como métodos de agrupamento. Os

outliers multivariados não são necessariamente caracterizados por valores extremamente altos ou baixos em resultados experimentais, o que pode dificultar e inviabilizar a utilização de métodos estatísticos na sua detecção. A eficiência da detecção de *outliers* é uma questão bastante importante, infelizmente a avaliação dos métodos é uma tarefa difícil.

Nessa perspectiva, embora a detecção de *outliers* enfrente desafios, muitos métodos multivariados para identificação de *outliers* têm sido desenvolvidos, utilizando diferentes metodologias. Os métodos baseados em distância identificam os outliers em um conjunto de amostras através do cálculo de distâncias e, de modo geral, esses métodos comparam as distâncias calculadas, com um valor crítico.

Os estudos, a seguir, mostram o desempenho de dois métodos de detecção de outliers utilizando o cálculo de distâncias: Mahalanobis e Mahalanobis robusta, em cada um dos sítios, separadamente.

O reconhecimento de *outliers* foi feito após a padronização pela transformada logarítmica da base de amostras, utilizando as distâncias Mahalanobis (DM), Mahalanobis robusta (DMR) e o critério de lambda Wilks (equação 49). As distâncias que ultrapassam o valor crítico do critério de lambda Wilks, foram classificadas como *outliers*. Nas Tabelas 15, 16 e 17 são mostrados os resultados obtidos da detecção de outliers, as amostras identificadas como outliers são destacadas em negrito. Os outliers são eliminados, e novamente são calculadas as distâncias e o valor crítico D_c, de modo que o procedimento anterior se repete, até que não sejam mais detectados outliers.

A Tabela 15 mostra que para 34 amostras os valores das distâncias (Mahalanobis e Mahalanobis robusta) foram inferiores ao valor crítico $D_c = 23{,}79$. Logo no sítio A não foram identificados *outliers*.

Já o conjunto de 76 amostras do sítio B (Tabela 16), analisados com o valor crítico de Wilks ($D_c = 30{,}52$), duas amostras (109 e 110) foram apontadas como outliers pela distância Mahalanobis. Após a exclusão das duas amostras, calculadas as distâncias e o valor crítico ($D_c = 30{,}24$) nenhum *outlier* foi encontrado. Por outro lado, utilizando a distância Mahalanobis robusta, nenhuma amostra foi classificada como outlier, uma vez que os valores de todas as distâncias foram inferiores à $D_c = 30{,}52$.

Por fim, no sítio C (Tabela 17) uma amostra (140) foi identificada como outlier, depois de sua exclusão nenhum outro outlier foi encontrado, utilizando a distância Mahalanobis. A distância Mahalanobis robusta não apontou nenhum *outlier*, de maneira que os valores das distâncias foram inferiores à $D_c = 22{,}46$. Assim entre as 140 amostras analisadas (que não possuíam valores ausentes) foram identificados três outliers, sendo necessárias duas iterações.

Tabela 15 – Frações de massa das amostras de fragmentos cerâmicos em $\mu g\ g^{-1}$, identificados como outliers pelas distâncias Mahalanobis e Mahalanobis robusta do sítio A

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U	DM	DMR
1	2,5	113,4	123,1	1,5	3,8	8,8	31,5	302,1	35,6	31,5	7,7	17,8	4,6	16,57	3,18
2	1,5	108,2	109,3	1,5	3,2	9,5	41,1	766,3	40,5	26,1	7,8	18,1	4,3	15,77	15,16
3	1,4	102,9	114,4	1,4	3,6	8,7	40,4	644,3	38,7	27,6	7,8	17,0	4,3	13,52	11,98
4	2,5	86,5	116,9	1,2	3,4	9,4	32,9	643,8	37,3	27,3	7,7	16,5	3,4	13,75	13,12
5	1,8	116,7	134,8	1,5	3,3	8,7	33,7	484,3	37,1	30,6	7,9	17,2	3,7	7,19	2,14
6	1,6	115,4	124,3	1,7	3,8	8,4	30,4	328,4	43,1	32,5	7,4	17,7	3,9	13,03	2,78
7	2,1	111,7	117,1	1,4	3,3	8,2	29,7	500,5	32,3	30,2	7,3	16,7	3,5	13,24	2,69
8	1,7	120,3	115,3	1,7	3,6	9,7	32,6	377,1	40,2	30,7	8,1	16,6	4,9	11,16	2,59
9	2,1	121,0	121,4	1,6	3,7	9,1	33,5	493,5	34,1	31,8	6,6	17,6	5,2	16,05	9,63
10	1,5	103,8	110,9	1,4	3,0	9,6	36,9	731,1	48,2	27,4	7,9	16,7	3,7	10,84	11,89
11	1,1	119,6	130,4	1,4	2,7	8,5	35,8	526,2	51,5	28,8	8,1	18,0	3,9	8,13	2,24
12	1,3	110,1	138,0	1,4	2,8	8,4	30,7	552,3	34,1	29,5	6,8	17,7	1,8	18,9	3,14
13	1,8	105,4	142,1	1,2	2,6	9,3	27,2	543,8	26,3	27,9	6,3	16,4	3,3	12,58	2,87
14	1,8	108,2	157,4	1,3	3,0	9,2	29,3	552,1	36,1	31,4	6,8	17,9	6,3	12,78	2,79
15	1,3	129,7	136,5	1,4	2,8	9,1	34,3	533,3	38,9	28,8	7,3	17,8	5,3	12,34	3,14
16	1,8	117,6	156,9	1,4	2,9	8,8	33	590,5	32,8	30,2	7,4	18,7	3,5	6,25	1,75
17	1,6	113,5	164,8	1,4	2,9	9,1	29,5	555,6	26,1	31,4	6,8	17,6	4,1	8,37	2,07
18	1,4	120,9	152,1	1,4	2,9	9,3	33,5	621,7	39,2	30,4	7,8	18,5	5,4	8,57	2,38

(continua)

Tabela 15 – (continuação)

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U	DM	DMR
19	1,8	113,5	170,1	1,3	2,9	9,5	30	635,1	27,9	31,3	7,0	17,2	4,3	10,84	2,34
20	1,6	119,4	151,4	1,4	2,8	8,2	33,2	590,2	34,8	28,9	7,5	18,0	5,0	9,96	2,7
21	1,2	113,2	138,5	1,3	2,8	8,5	31,4	557,3	29,7	28,6	7,1	15,8	4,8	14,71	2,9
22	1,5	115,7	136,8	1,4	2,8	9,1	32,4	548,3	41,9	29,4	7,4	16,7	4,6	2,51	2,03
23	1,5	104,2	136,1	1,3	2,6	8,4	29,3	579,1	38,8	27,6	6,8	16,1	3,5	9,19	2,48
24	1,5	117,6	144,3	1,4	2,8	8,3	32,1	656,3	31,7	29,3	7,3	17,3	4,2	6,13	1,74
25	1,8	131,8	140,2	1,6	2,6	8,9	35,3	593,4	46,6	29,1	6,5	16,5	5,0	22,59	11,95
26	1,4	112,5	140,5	1,3	2,7	8,3	31,6	558,5	36,8	28,4	7,2	17,3	2,6	7,34	2,47
27	1,8	117,5	175,7	1,0	1,7	10,1	38,5	786,8	57,2	26,7	7,8	19,2	4,5	15,4	2,79
28	2,2	106,1	186,8	1,0	2,0	10,1	31,8	678,1	52,0	27,1	6,7	18,2	4,2	14,08	2,71
29	0,7	117,4	129,4	1,6	2,0	8,3	32,8	535,3	40,1	33,2	8,4	17,1	4,5	13,95	2,88
30	0,9	137,7	126,1	1,6	2,3	7,6	45,6	532,6	44,5	30,3	9,3	20,0	5,4	18,61	3,27
31	1,2	116,1	134,1	1,5	2,2	9,3	30,1	476,7	49,6	33,6	8,0	16,9	4,6	12,36	2,79
32	1,3	121,5	135,3	1,5	2,1	8,4	33	516,1	36,2	33,9	8,1	18,0	4,1	11,37	2,56
33	1,5	108,8	115,7	1,4	2,1	7,6	28,1	507,3	29,3	29,9	6,8	16,4	3,3	18,62	3,07
34	1,6	137,2	186,6	1,3	1,7	11,3	38,9	727,9	d45,1	27	8,1	19,5	4,7	22,28	3,47
															$D_c = 23,79$

Fonte: Próprio autor

Tabela 16 – Frações de massa das amostras de fragmentos cerâmicos em $\mu g\ g^{-1}$ identificados como outliers pelas distâncias Mahalanobis e Mahalanobis robusta do grupo B

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U	DM	DMR	DM
35	4,4	133,5	148,1	2,8	4,5	8,4	86,6	2487,1	66,9	14,9	11,7	12,8	1,9	10,25	2,97	10,05
36	2,4	98,8	128,3	2	3,9	7,5	56,3	1065,3	52,1	13,1	7,8	10,7	1,5	16,9	5,03	17,22
37	2,2	110,9	155,5	2,7	4,4	7,9	75,7	1940,1	69,3	14,7	10,3	10,9	1,3	7,64	2,92	8,95
38	4,4	120,1	159,9	2,4	4,2	9,1	70,2	2535,1	58,8	15,2	10,2	12,4	1,2	9,54	2,8	9,37
39	2,1	123,9	141,8	2,6	3,8	8,3	73,0	2414,6	66,3	14,8	9,7	12,1	1,2	4,92	2,21	5,74
40	3,9	123,8	175,1	2,6	4,3	9,1	72,5	2254,5	63,3	16,8	10,2	15,0	1,3	6,51	2,16	6,83
41	2,2	97,8	130,2	2	3,9	7,3	67,2	2765,3	41,3	12,1	8,7	10,8	1,1	14,22	3,07	14,28
42	3,3	123,4	151,3	2,6	4,0	7,8	66,8	1702,2	54,8	16,3	9,1	14,1	1,0	11,15	2,88	11,13
43	4,5	148,2	173,1	2,4	4,4	9,3	68,0	2435,1	66,6	16,6	9,4	12,7	1,2	16,55	4,99	16,48
44	5,8	124,8	177,2	2,7	4,8	8,8	74,2	2156,8	68,8	17,4	10,3	13,1	1,3	10,64	2,78	10,42
45	3,3	109,1	127,3	2,1	4,9	8,4	64,5	2340,1	55,0	10,9	8,5	10,2	1,5	12,91	3,02	13,05
46	2,7	104,0	129,2	2,4	4,0	8,6	60,7	2310,2	60,1	13,4	9,1	11,6	1,8	13,99	5,55	16,48
47	3,1	96,0	145,1	2,2	4,5	7,8	61,2	2599,9	49,8	13,1	8,3	11,6	1,1	10,54	2,51	10,4
48	2,8	121,2	152,0	2,7	4,1	8,7	89,0	2119,1	64,4	15,8	10,8	12,4	1,7	12,22	3,2	12,55
49	3,3	127,3	166,7	2,6	4,1	9,9	80,9	2223,3	72,1	17,0	11,2	14,3	1,2	8,82	2,77	8,57
50	2,1	142,0	166,6	3,1	4,1	8,3	86,4	2376,5	72,1	16,9	11,6	13,9	1,4	7,77	2,84	8,14
51	1,5	108,3	134,2	2,5	3,2	7,8	64,1	1961,1	63,8	12,9	8,9	9,8	1,3	10,88	2,65	10,81
52	1,4	110,6	156,1	2,3	3,3	9,6	71,4	2938,2	60,1	14,5	9,8	12,0	2,0	12,71	3,08	12,99

(continua)

Tabela 16 – (continuação)

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U	DM	DMR	DM
53	2,2	84,9	125,2	1,9	3,1	6,8	59,6	3151,2	49,8	10,2	7,5	9,1	1,4	17,95	3,39	17,75
54	3,6	99,9	139,3	2,2	3,5	9,3	66,8	2514,7	47,3	12,8	8,9	10,8	1,3	10,68	3,13	11,69
55	2,7	122,3	133,3	2,6	3,8	6,3	83,4	1487,8	64,2	15,2	10,1	12,6	1,0	15,55	3,25	15,14
56	2,5	133,6	182,5	2,3	3,2	9,8	70,7	1699,2	57,2	18,1	9,8	17,2	1,6	9,49	2,82	9,74
57	1,8	102,6	118,1	2,2	3,4	5,7	72,9	2260,5	46,8	12,6	9,2	9,7	1,2	14,41	3,30	15,46
58	2,5	111,9	138,3	2,3	3,7	8,4	62,7	2254,8	49,3	12,6	8,4	12,1	0,9	14,18	3,07	14,21
59	2,3	107,4	122,2	2,6	3,3	5,4	76,0	2129,1	63,4	13,6	10,1	9,7	1,1	16,97	3,36	17,06
60	1,6	99,2	148,1	2,3	3,5	7,8	67,6	1982,1	52,8	12,6	8,9	10,2	1,2	7,32	2,96	9,25
61	2,2	127,1	143,1	2,4	3,4	8,1	71,4	2052,2	62,4	14,8	9,3	12,4	1,2	3,34	2,00	3,99
62	1,2	125,6	150,5	2,7	3,4	9,3	83,4	1617,3	51,3	17,2	11,3	13,5	1,3	13,18	3,18	13,26
63	2,4	132,5	148,0	3,1	3,7	8,1	80,9	2957,5	64,1	15,3	11,7	10,5	1,7	18,27	3,7	17,92
64	2,2	100	145,0	2,1	3,7	9,9	58,2	1919,8	54,5	13,2	8,1	11,4	1,2	8,42	2,82	8,31
65	2,4	116,5	154,1	2,0	3,5	8,2	61,0	1496,1	49,9	13,6	7,5	11,8	1,8	13,38	3,21	13,44
66	2,1	98,2	130,5	2,1	3,2	7,8	61,1	2484,5	37,4	12,4	8,1	9,9	1,4	10,92	3,11	12,71
67	2,1	126,5	157,9	2,8	3,5	8,5	77,3	1526,6	62,3	17,7	10,8	15,4	0,8	12,99	2,93	12,89
68	1,6	113,8	180,7	2,0	3,0	10,1	54,9	1604,7	47,3	15,2	7,0	12,2	1,4	13,71	3,54	14,22
69	2,7	115,2	145,8	2,5	3,1	7,4	70,0	2166,9	61,3	13,9	9,4	11,6	1,5	5,13	2,11	5,1
70	1,6	104,5	150,4	2,4	3,0	7,7	61,8	2437,1	47,4	12,8	8,7	11,1	1,3	11,08	2,91	11,73

(continua)

Tabela 16 – (continuação)

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U	DM	DMR	DM
71	1,0	120,9	141,1	2,8	3,2	7,7	87,1	1408,1	59,1	14,9	11,2	12,2	1,5	12,34	3,14	14,02
72	2,7	115,1	155,3	3	3,5	7,6	79,2	1725,4	62,2	15,7	10,7	12,0	1,3	8,51	2,94	9,4
73	2,7	123,1	186,8	2,7	3,3	8,6	71,6	2367,1	59,1	17,6	8,9	13,1	1,5	12,15	3,45	13,16
74	2,1	126,8	166,0	2,5	3,5	8,2	65,6	1693,6	59,1	16,3	9,6	12,7	1,5	5,8	2,97	6,86
75	2,4	120,1	141,7	2,2	3,3	7,3	59,9	1665,5	52,9	15,0	8,8	12,9	2,0	14,05	5,73	17,5
76	2,1	117,5	184,6	2,5	3,3	9,2	69,5	2151,7	57,7	17,0	9,8	14,6	2,0	7,52	2,76	8,40
77	1,8	138,5	192,5	2,7	3,2	9,3	78,2	2183,9	57,5	19,7	10,5	15,5	1,7	7,47	2,60	7,53
78	1,2	125,2	158,4	2,8	2,9	9,2	71,3	1176,1	58,3	17,7	9,9	13,3	1,4	10,8	3,46	10,49
79	2,2	131,9	169,1	3,2	3,4	9,3	77,6	1037,1	60,2	17,8	10,3	14,4	1,7	14,79	4,43	15,91
80	1,4	112,7	137,0	2,3	3,1	8,1	69,4	1536,1	50,1	13,1	9,0	11,6	1,5	4,36	2,05	4,70
81	3,6	111	156,1	2,3	3,4	8,4	64,9	2140,8	48,3	14,7	8,6	12,2	1,3	5,37	1,86	5,52
82	0,7	105,7	184,3	2,5	2,9	8,9	73,9	1512,5	60,9	19,7	9,8	12,8	1,2	23,95	5,73	24,50
83	1,1	137,5	172,6	2,6	2,5	9,6	73,7	1638,8	80,5	17,2	10,0	14,1	1,3	13,66	3,28	14,25
84	2,2	116,4	130,1	2,2	2,8	9,1	70,1	1660,2	60,6	12,2	9,7	10,5	1,5	14,94	3,56	15,91
85	0,9	137,9	195,9	2,2	2,1	8,1	54,0	1928,8	39,8	18,3	7,3	13,2	1,6	24,69	6,26	25,06
86	1,8	151,5	150,0	2,6	2,9	7,6	79,3	2188,1	59,1	14,6	11,1	12,2	1,1	11,51	5,11	13,03
87	2,4	158,8	215,1	2,6	2,0	9,1	72,5	1384,3	55,2	18,2	10,5	18,1	2,0	15,00	5,65	24,46
88	1,9	118,9	185,2	2,8	2,5	8,8	66,8	1941,4	64,4	20,3	10,0	15,5	1,0	16,26	3,40	17,18
89	0,9	133	188,3	2,1	1,9	7,2	57,8	1729,8	49,1	16,9	7,9	12,9	1,2	13,55	5,79	13,64

(continua)

Tabela 16 – (continuação)

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U	DM	DMR	DM
90	1,5	114	145,5	2,1	2,5	8,3	58,4	1625,6	57,9	15,8	7,9	13,4	1,0	8,82	3,32	11,31
91	1,6	148,7	142,5	2,3	2,7	7,8	70,3	1851,5	57,5	17,0	10,5	16,0	1,1	19,85	5,54	22,59
92	1,8	142,7	168,3	2,5	2,6	8,2	78,7	1322,1	53,3	16,7	10,6	14,6	1,3	6,69	3,13	8,94
93	1,2	137,2	144,1	2,6	2,8	8,4	72,6	1706,3	59,1	15	10,1	12,7	1,1	7,52	2,70	7,69
94	1,5	104,6	135,3	2,1	2,4	9,2	60,7	1015,8	46,2	14,9	8,2	13,7	1,3	11,19	5,40	11,91
95	1,2	127,4	171,1	2,1	1,8	6,5	57,9	841,1	54,6	17,2	8,1	12,5	1,2	19,60	5,68	21,81
96	2,1	116,8	183,2	2,3	2,6	8,1	61,4	1690,1	59,8	17,7	8,2	14,4	1,4	7,11	2,46	7,07
97	2,3	105,1	142,5	2,1	2,2	8,5	62,5	1250,3	61,3	14,4	8,8	15,2	1,6	20,36	6,51	20,25
98	1,1	119,5	184,8	2,6	2,7	9,6	68,4	1835,5	50,1	16,5	9,5	13,8	1,4	8,04	2,80	9,14
99	2,3	104,7	161,9	2,2	2,9	9,3	63,0	2499,9	50,7	14,9	8,2	12,8	1,2	7,00	2,37	8,11
100	1,9	85,5	147,1	2,3	2,8	10,4	61,5	1480,1	44,8	14,0,	9,3	11,7	1,6	21,90	6,81	22,27
101	2,6	117,3	187,1	2,2	2,7	10,5	67,3	2627,0	57,9	16,1	9,1	14,9	2,4	15,08	3,35	14,80
102	1,8	121,6	160,3	2,5	2,9	8,6	72,4	1712,3	63,3	16,4	9,9	11,1	1,2	10,62	3,09	10,53
103	1,3	152,1	158,4	2,6	2,4	7,4	80,7	1985,0	68,1	15,3	10,1	12,6	1,1	10,88	5,47	14,37
104	1,1	125,5	182,1	2,2	1,8	9,8	68,9	1284,1	50,9	17,5	9,3	14,8	1,5	11,97	3,19	12,00

(continua)

Tabela 16 – (continuação)

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U	DM	DMR	DM
105	1,4	115,1	147,2	2,3	2,7	7,7	65,3	2181,0	47,7	14,6	8,9	11,5	1,2	3,9	2,18	4,21
106	1,1	116,3	130,8	2,1	2,6	7,8	66,5	1353,3	44,9	12,7	8,2	11,2	1,2	9,09	2,74	10,03
107	4,8	138,8	222,8	2,5	2,4	8,7	64,7	1500,5	57,5	19,9	9,2	15,7	1,7	18,7	3,98	19,68
108	2,1	146,3	242,3	3,1	3,7	10,2	84,7	4068,8	81,3	20,4	12,1	14,4	1,3	21,98	7,42	23,48
109	**1,5**	**180,6**	**275,0**	**3,5**	**2,2**	**8,2**	**91,4**	**700,7**	**80,2**	**21,9**	**14,3**	**19,1**	**1,8**	**36,26**	10,14	
110	**1,7**	**227**	**275,0**	**3,8**	**2,2**	**9,8**	**116,6**	**1187,1**	**93,1**	**16,1**	**13,8**	**14,0**	**1,6**	**46,6**	16,25	
													D_c	**30,52**		**30,34**

Fonte: Próprio autor

Tabela 17 – Frações de massa das amostras de fragmentos cerâmicos em $\mu g\ g^{-1}$ identificados como outliers pelas distâncias Mahalanobis e Mahalanobis robusta do grupo C

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U	DM	DMR	DM
111	2,4	85,2	214,1	3,3	1,6	10,8	37,6	155,1	53,2	43,9	10,8	5,2	1,2	12,7	2,84	15,42
112	1,8	101,5	230,4	3,4	1,3	11,7	45,5	144,4	51,1	45	11,4	7,7	1,3	14,34	2,77	13,86
113	3,1	65,3	212,3	2,9	1,2	10,5	33,5	138,1	50,5	42,6	9,7	6,8	1,6	9,08	2,33	9,38
114	2,5	54,5	203,5	3,0	1,2	10,9	34,1	138,4	44,0	44,7	9,6	6,8	1,2	11,02	2,35	11,35
115	1,4	93,2	243,3	3,4	1,2	12,8	40,9	189,3	54,2	45,8	11,4	6,1	1,2	6,56	1,93	6,55
116	1,6	110	260,1	3,8	1,2	12,3	48,3	159,5	59,3	44,1	13,2	5,8	0,9	13,11	2,84	14,47
117	1,7	95,2	204,1	3,4	1,3	12,5	43,5	192,8	48,6	50,1	11,1	6,8	1,2	15,17	2,87	15,81
118	1,3	89,2	249,9	3,4	1,4	12,3	39,5	165,1	62,1	48,9	11,1	5,7	1,4	14,64	2,76	14,12
119	1,8	92,7	253,8	3,6	1,4	12,8	44,2	125,4	63,3	48,3	11,7	6,4	1,2	9,95	2,4	10,03
120	1,7	75,8	205,3	2,9	0,8	12,5	31,8	121,1	45,2	41,8	9,0	6,9	1,6	9,5	2,37	9,96
121	1,6	56,4	183,1	2,4	0,8	10,8	28	120,8	35,1	43,4	7,4	6,4	1,5	20,15	12,98	19,78
122	2,2	61,7	212,7	3	0,8	10,8	34	125,6	48,0	47,4	9,2	6,9	1,7	15,75	2,92	16,38
123	2,2	62,5	195,3	2,8	0,9	11,3	29,3	92,5	46,5	42,5	9,2	7,1	1,3	8,97	2,2	9,55
124	1,9	52,5	195,7	2,7	0,9	11,6	26,2	136,1	43,8	43,2	8,5	7,3	1,4	10,2	2,73	10,13
125	1,9	109,7	218,5	3,3	0,7	11,7	37,8	181,5	60,5	39,4	10,3	5,2	1,1	10,92	2,93	15,28
126	1,6	78,9	230,4	3,2	0,8	10,9	41,1	189,2	69,8	40	11,3	5,1	1,1	16,04	5,98	15,98
127	1,2	68,9	204,3	2,9	0,8	11,4	32,8	191,1	51,0	44,3	10,2	6,8	1,6	14,16	2,74	13,97
128	1,4	70,9	192,1	3	0,8	11,9	36,1	117,1	61,1	46,1	10,3	7,4	1,5	10,55	2,77	10,17
129	1,8	97,5	238,1	3,3	0,8	11,9	38,0	167,2	52,4	42,3	10,4	6,2	1,8	13,83	3,17	14,4
130	2,6	67,8	212,2	2,9	1,1	10,8	31,8	132,4	41,0	39,9	9,4	6,4	1,3	6,53	2,11	6,27

(continua)

Tabela 17 –
(continuação)

Amostra	As	Ce	Cr	Eu	Fe (%)	Hf	La	Na	Nd	Sc	Sm	Th	U	DM	DMR	DM
131	1,9	61,1	215,8	2,9	1,0	10,9	30,8	176,5	45,9	46,2	9,1	7,3	1,5	8,00	2,29	8,08
132	2,8	68,6	215,1	3,0	1,1	11,8	34,0	145,6	48,1	45,0	9,3	6,3	1,4	7,43	2,23	8,79
133	1,6	82,1	187,2	3,2	1,1	10,8	37,2	260,1	47,3	37,2	9,8	4,8	1,2	18,00	2,99	17,35
134	1,5	90,8	303,1	3,2	1,2	11,0	39,5	266,9	52,1	41,7	10,2	5,6	1,1	17,35	10,20	18,5
135	1,4	95,2	245,4	3,5	1,1	12,1	44,0	187,1	57,2	43,0	11,3	5,8	1,4	6,22	1,92	6,7
136	2,2	67,5	205,5	2,8	1,0	10,6	36,6	93,0	42,8	43,5	10,1	7,1	1,9	16,55	6,82	15,99
137	1,2	63,4	183,3	2,9	0,9	10,5	33,9	130,3	44,1	40,7	9,6	6,7	1,7	14,81	2,73	14,44
138	2,7	67,8	236,1	3,0	1,1	11,0	33,8	139,2	55,3	41,2	10	6,3	1,4	5,40	2,15	6,86
139	1,7	87,8	241,3	3,3	1,1	10,9	40,8	200,1	71,2	45,6	11	7,3	1,3	14,74	2,91	14,42
140	**1,1**	**245,3**	**162,1**	**4,7**	**0,7**	**11,6**	**53,5**	**172,9**	**72,1**	**48,4**	**15,6**	**5,3**	**1,0**	25,35	19,43	
													D_c	**22,46**		**22,09**

Fonte: Próprio autor

Para avaliação do impacto da exclusão das amostras apontadas como *outliers* (pela distância Mahalanobis), foi calculada a compacidade (equação 22) das amostras de cada um dos sítios.

A Tabela 18 mostra a compacidade entre as amostras dos sítios A, B e C, antes e depois da remoção dos outliers. Nota-se, diante disso, e comparando à compacidade dos sítios que B e C, com e sem outliers, uma diminuição da distância médias entre as amostras, ou seja, uma maior coesão entre os grupos correspondentes, após exclusão dos outliers.

Tabela 18 – Compacidade entre as amostras dos sítios A, B e C

Compacidade dos sítios (com *outliers*)			**Compacidade os sítios (sem *outliers*)**		
A	B	C	A	B	C
6448,77	8317,43	2775,19	6448,77	8186,64	2690,87

Fonte: Próprio autor

A Tabela 19 apresenta o desempenho dos métodos de agrupamento aplicados na base completa (com todas as 140 amostras) e sem os outliers detectados pela distância Mahalanobis. Antes da aplicação de cada um dos métodos de agrupamento foi aplicada a transformada logarítmica.

Tabela 19 – Erro cometido pelos métodos de agrupamentos após a aplicação dos métodos de detecção de outliers

Métodos agrupamento	**Base completa**	**sem *outliers***
ligação simples	0 %	0 %
ligação média	0 %	0 %
ligação completa	0 %	0 %
ward	0 %	0 %
k-médias	0 %	0 %

(continua)

Tabela 1– (continuação)

Métodos agrupamento	**Base completa**	**sem *outliers***
k-medóides	0 %	0 %
híbrido	0 %	0 %
c-médias	50,00%	0 %
c-médias-pf	50,00 %	0 %
c-medóides	45,71 %	52,55 %

Fonte: Próprio Autor

A Tabela 19 mostra que após a exclusão dos outliers detectados pela distância Mahalanobis, os métodos de agrupamento c-médias, c-médias com polinômio fuzzyficador, foram capazes de determinar corretamente todos os grupos. Foi observado também que os métodos hierárquicos e particionais/crisp foram capazes de determinar todos os grupos perfeitamente, tanto na base completa, quanto na base sem os outliers determinados pela distância Mahalanobis.

Em seguida são apresentados os efeitos da padronização das amostras: no cálculo da estatística de Hopkins, nas imagens VAT, na determinação do número de grupos e no desempenho dos métodos de agrupamento, aplicados à base de amostras.

A análise de agrupamento e a PCA são métodos estatísticos multivariados dependentes da escala das variáveis. Desse modo, a aplicação de uma transformada para padronização das amostras é necessária, para garantir que todas as variáveis de uma amostra contribuam igualmente, no cálculo da similaridade.

Uma das formas de analisar a relação entre as amostras de uma base de amostras é através da análise de agrupamento. A padronização de amostras desempenha um papel importante na análise de agrupamento, uma vez que pode corrigir violações de suposições estatísticas e/ou melhorar relações entre variáveis. As variáveis podem estar em diferentes escalas ou apresentarem diferentes magnitudes, que podem impactar negativamente na determinação dos grupos pelos métodos de agrupamento. Variáveis padronizadas podem gerar

bons agrupamentos (compactos e separados) e consequentemente melhorar o desempenho de métodos de agrupamento.

A primeira etapa antes da aplicação dos métodos de agrupamento consiste na avaliação da existência de uma estrutura de grupos na base de amostras ou se as amostras estão dispostas de forma aleatória. Para esse propósito foi utilizada a estatística de Hopkins e a as imagens VAT.

A Tabela 20 apresenta a estatística de Hopkins (H) calculada para a base de amostras não padronizadas e para as amostras padronizadas pelas transformadas: z-score, min-max, min-max-m, logarítmica e Box-Cox. Após aplicação das transformadas estudadas, o valor de H foi maior que 0,75, o que indica uma tendência de agrupamento, porém, o mesmo não ocorreu com a base de amostras não padronizadas, visto que H=0,68.

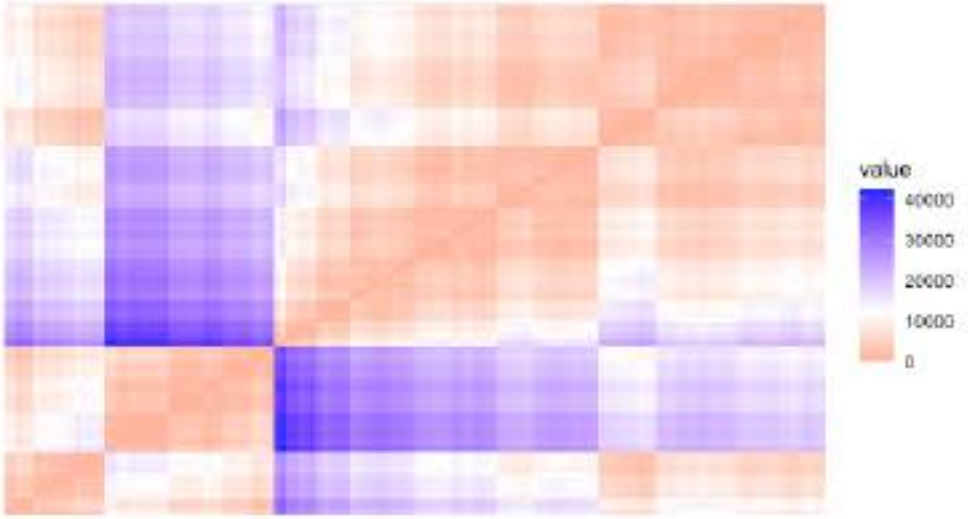
Tabela 20 - Impacto da padronização das amostras no cálculo da estatística de Hopkins

	Estatística de Hopkins
não padronizadas	0,68
z-score	0,78
min-max	0,79
min-max-m	0,96
logarítmica	0,79
Box-Cox	0,81

Fonte: Próprio Autor

Outra técnica para avaliação da tendência de agrupamento utilizada foi a imagem VAT. Inicialmente foi gerada a imagem VAT associada às amostras não padronizadas da Tabela 3, como mostra a Figura 18. A imagem apresenta retângulos ao longo da diagonal, indicando uma tendência de agrupamento, contudo não é possível determinar o número de grupos da base.

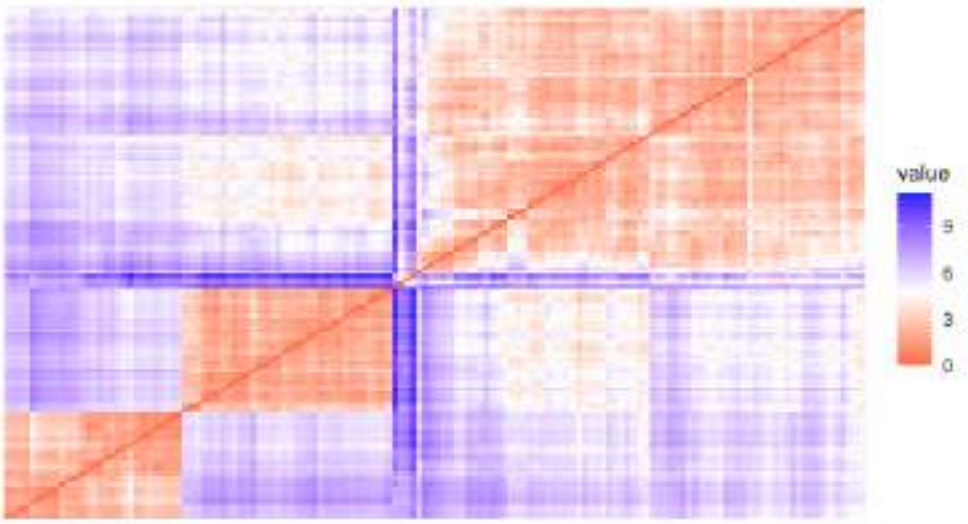
Figura 18 – Imagem VAT (amostras não padronizadas)



Fonte: Próprio autor

Por outro lado, a imagem VAT apresentada na Figura 19, obtida a partir da aplicação da transformada z-score (equação 52) na base, além de indicar uma tendência de agrupamento, aponta que o número de grupos da base é três.

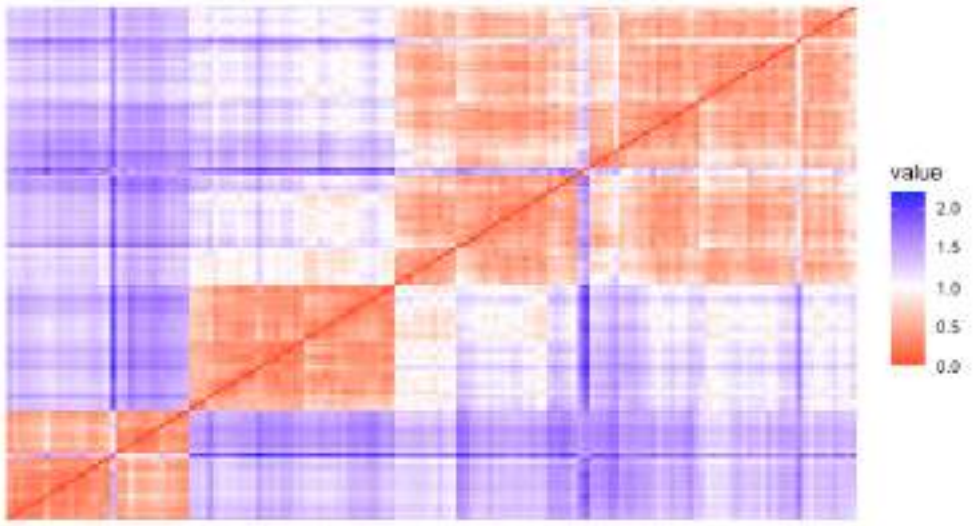
Figura 19 – Imagem VAT (transformada z-score)



Fonte: Próprio autor

Já a imagem VAT mostra na Figura 20, que é relacionada com a base, padronizada com a transformada min-max (equação 53). Novamente a imagem aponta para tendência de agrupamento e três grupos.

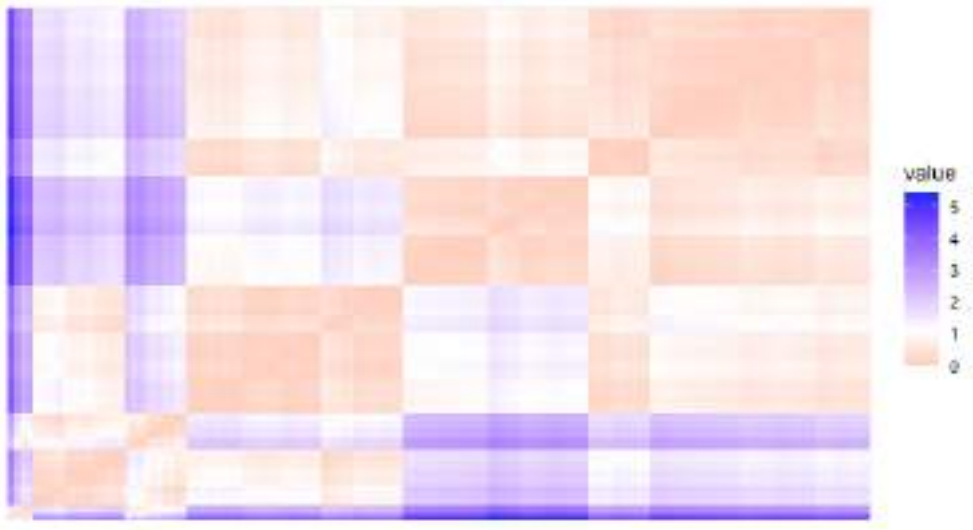
Figura 20 – Imagem VAT (transformada min-max)



Fonte: Próprio autor

A imagem VAT, da Figura 21, correspondente à base, obtida depois da aplicação da transformada min-max-m (equação 54), que indicou apenas a tendência de agrupamento.

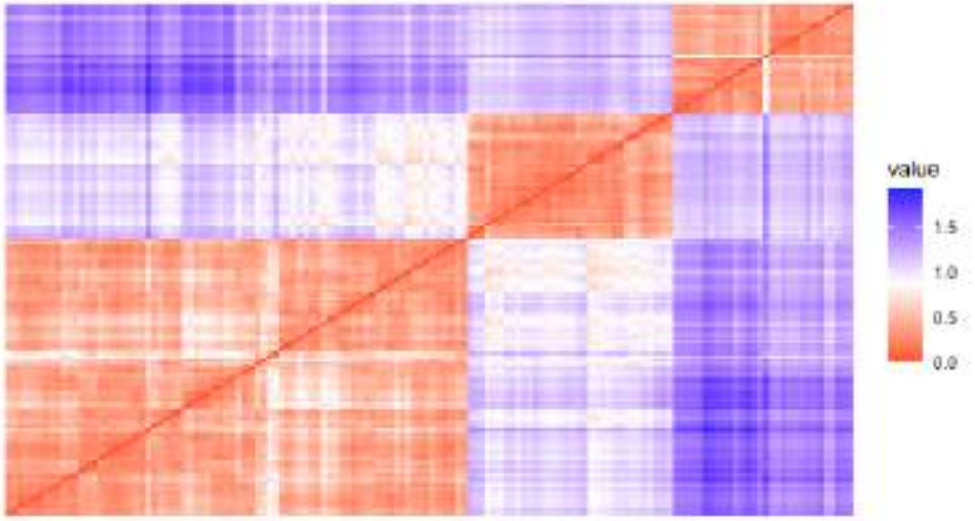
Figura 21 – Imagem VAT (transformada min-max-m)



Fonte: Próprio autor

Em seguida, a imagem da Figura 22, associada à base, padronizada pela transformada logarítmica (equação 55), mostrou que a base apresenta tendência de agrupamento e o número de grupos é três.

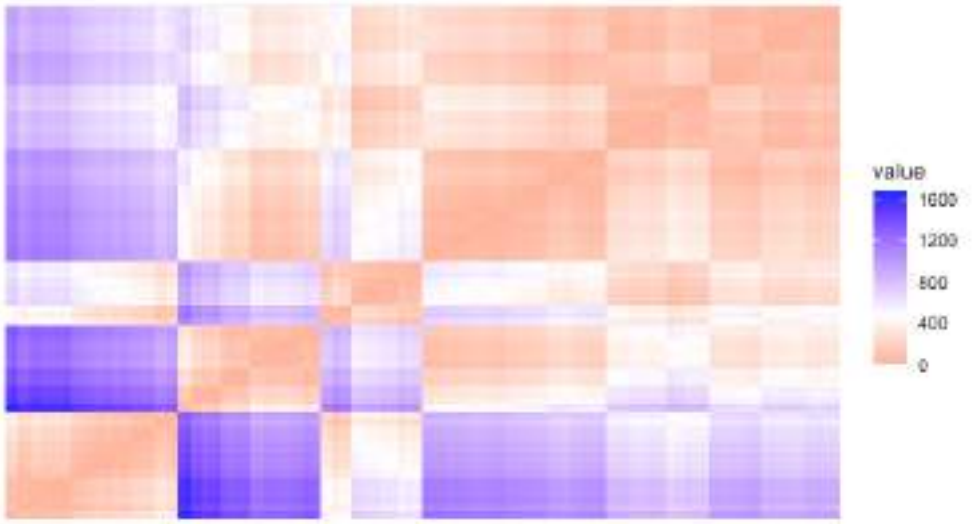
Figura 22 – Imagem VAT (transformada logarítmica)



Fonte: Próprio autor

Por último a imagem VAT exibida na figura 23, associada à padronização da base, pela transformada Box-Cox (equação 56), indicou apenas a tendência de agrupamento da base.

Figura 23 – Imagem VAT (transformada Box-Cox)



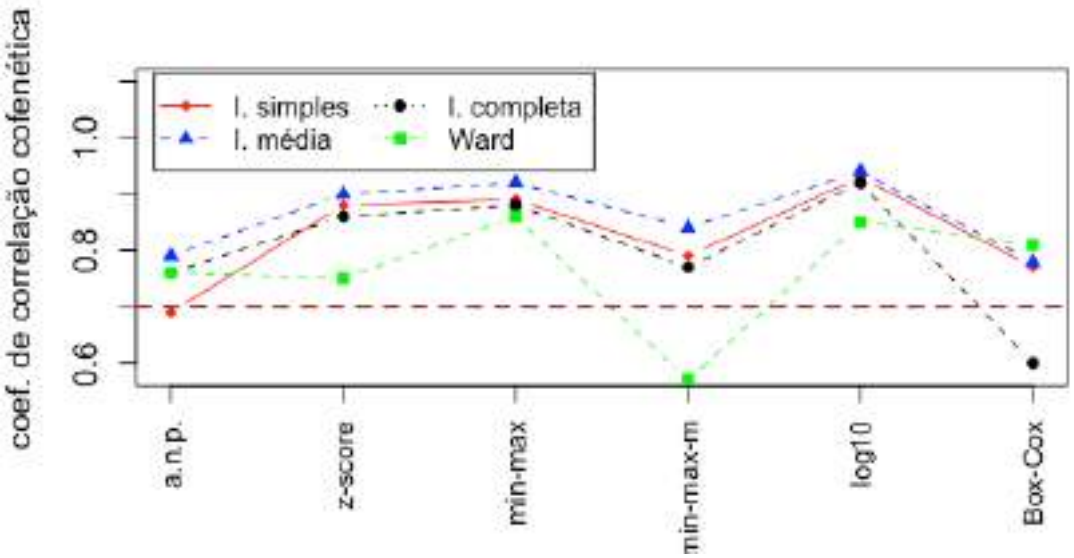
Fonte: Próprio autor

Diante do exposto, uma vez determinado que a base de amostras possui tendência de agrupamento, o passo seguinte é determinar o número de grupos da base. Nos testes realizados, o número de grupos foi determinado pelo

coeficiente de correlação cofenética (CCC) e por sete índices de validação interna.

Mais adiante, na Figura 24, são apresentados os valores do CCC para avaliação dos métodos de agrupamento hierárquicos estudados: ligação simples, ligação média, ligação completa e Ward. Os métodos de agrupamento são aplicados em amostras padronizadas e amostras não padronizadas (a.n.p.). O CCC auxiliará na determinação do número de grupos.

Figura 24 - Gráfico dos métodos de agrupamento hierárquicos versus coeficiente de correlação cofenética

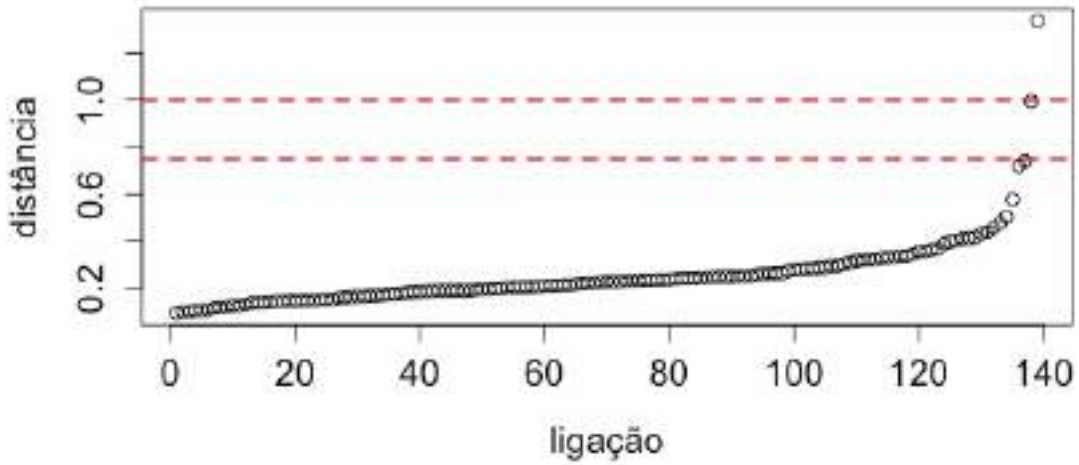


Fonte: Próprio autor

Os valores do CCC no gráfico da Figura 24 abaixo de 0,70 ocorreram com as amostras não padronizadas e transformadas min-max-m e Box-Cox, e os métodos de ligação simples, Ward e de ligação completa, respectivamente. Assim sendo, indicando a inadequação do método de agrupamento em resumir a informação da base de amostras (ROHLF, 1970).

No caso da utilização da transformada logarítmica e do método de ligação média, o CCC foi de $0{,}94$, o que indica uma maior consistência dos agrupamentos formados. O gráfico da ligação em função da distância (Figura 25) foi feito com intuito de apontar “pontos de salto”. Tais pontos indicam o momento de parada do método de agrupamento e consequentemente indicam o número de grupos. No gráfico da Figura 25, o maior salto ocorre entre $0{,}77$ e 1.

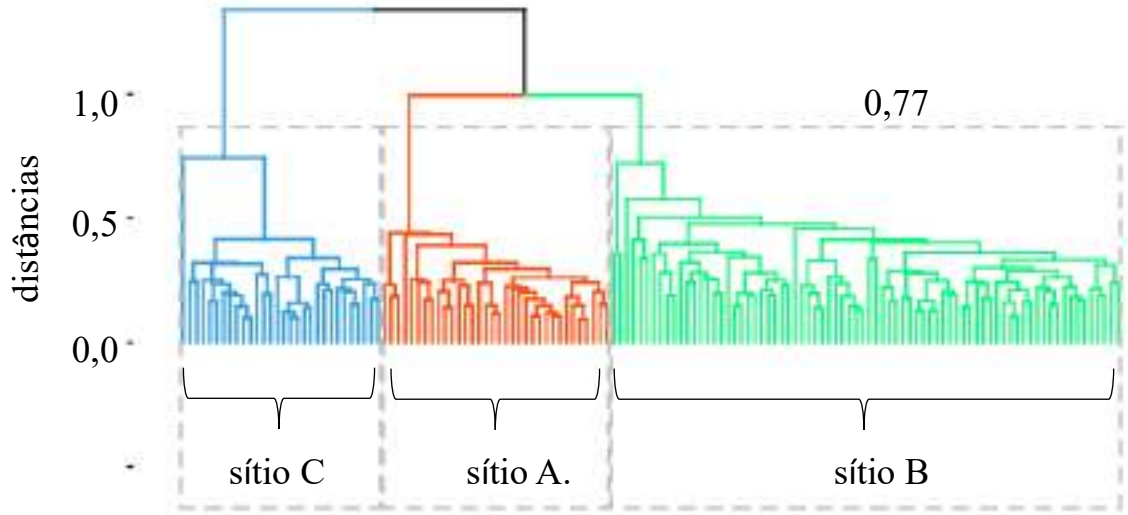
Figura 25 - Gráfico da ligação em função da distância, utilizando o método de ligação média aplicado nas amostras após aplicação da transformada logarítmica.



Fonte: Próprio autor

O dendrograma da Figura 26 foi gerado aplicando: a transformada logarítmica na base dada pela Tabela 3 e o método de agrupamento hierárquico de ligação média. O corte foi feito com a distância igual a 0,77 conforme observado no dendrograma, indicando a existência de três grupos.

Figura 26 - Dendrograma obtido com aplicação do método de ligação média nas amostras após aplicação da transformada logarítmica.



Fonte: Próprio autor

Com isso, tendo em vista que determinado o número de grupos da base de amostras, foi analisada a influência da padronização das amostras, na determinação do número de grupos através de índices de validação. Nesse

sentido, para determinação do número de grupos utilizando índices de validação interna, foram realizados testes aplicando os métodos agrupamento da Tabela 2 (na base com 140 amostras), variando o número de grupos ($k = 2,3,\dots,8$) e calculando os índices de validação correspondentes. O número de grupos k associado ao maior valor de cada índice é selecionado e se $k = 3$, o número de grupos está correto e será indicado nas Tabelas a seguir com 'X'. Os índices analisados foram: Dunn, Calinski-Harabasz (CH), Gama, PBM, Wemmert-Gancarski (WG), Tau e Ratkowsky-Lance (RL).

A Tabela 21 apresenta os acertos obtidos na determinação dos grupos pelos índices de validação, utilizando os métodos de agrupamento (ligação simples (l. simples), ligação média (l. média) e ligação completa (l. completa), Ward, k-médias, k-medóides, híbrido, c-médias, c-medóides e c-médias com polinômio fuzzificador (c-médias-pf)) aplicados na base com amostras não padronizadas. Os índices com maiores quantidades de acerto foram RL e Tau, já o índice Gama não foi capaz de determinar o número correto de grupos.

Tabela 21 – Acertos na determinação do número de grupos utilizando os índices de validação (amostras não padronizadas)

	Índices de validação						
Métodos	**Dunn**	**CH**	**Gama**	**PBM**	**WG**	**Tau**	**RL**
l. simples							
l. media						X	
l. completa							
Ward						X	
k-medias							X
k-medóides							
híbrido							X
c-médias		X					X
c-medóides							
c-médias-pf	X	X		X	X	X	X

Fonte: Próprio autor

Nota-se na Tabela 22 que após padronização das amostras pela transformada z-score, os índices CH, Gama e RL determinaram corretamente o número de grupos na maioria dos métodos de agrupamento, e os métodos de agrupamento que não foram capazes de determinar o número de grupos

corretamente são: hierárquicos de ligação simples, média, completa e c-medóides. Os piores desempenhos foram dos índices Dunn e PBM.

Tabela 22 - Acertos na determinação do número de grupos utilizando os índices de validação (transformada z-score)

	Índices de validação						
Métodos	**Dunn**	**CH**	**Gama**	**PBM**	**WG**	**Tau**	**RL**
l. simples							
l. media							
l. completa							
Ward		X	X		X	X	X
k-medias		X	X				X
k-medóides	X	X	X	X	X	X	X
híbrido		X	X				X
c-médias	X	X	X	X	X	X	X
c-medóides							
c-médias-pf	X	X	X	X	X	X	X

Fonte: Próprio autor

Quanto à Tabela 23, esta apresenta o desempenho dos índices de validação após a aplicação da transformada min-max. Observou-se uma superioridade dos índices CH, Gama, PBM, WG, Tau e RL, em relação ao índice de Dunn.

Tabela 23 - Acertos na determinação do número de grupos utilizando os índices de validação (transformada min-max)

	Índices de validação						
Métodos	**Dunn**	**CH**	**Gama**	**PBM**	**WG**	**Tau**	**RL**
l. simples							
l. média							
l. completa							
Ward		X	X	X	X	X	X
k-medias		X		X	X	X	X
k-medóides		X	X	X	X	X	X
híbrido		X	X	X	X	X	X
c-médias	X	X	X	X	X	X	X
c-medóides							
c-médias-pf		X	X	X	X	X	X

Fonte: Próprio

Na Tabela 24, tem-se o desempenho dos índices de validação após a padronização das amostras pela transformada min-max-m. Os resultados mostraram, portanto, que o melhor desempenho foi alcançado com o índice Tau e os índices CH, Gama, PBM e WG, não foram capazes de determinar corretamente o número de grupos na base estudada.

Tabela 24 - Acertos na determinação do número de grupos utilizando os índices de validação (transformada min-max-m)

	Índices de validação						
Métodos	**Dunn**	**CH**	**Gama**	**PBM**	**WG**	**Tau**	**RL**
l. simples							
l. media							X
l. completa							
Ward	X					X	
k-medias							
k-medóides						X	
híbrido							
c-médias						X	
c-medóides	X					X	
c-médias-pf							

Fonte: Próprio autor

A transformada logarítmica aplicada às amostras, conforme mostram os resultados da Tabela 25, acarretou o mesmo desempenho para os CH, PBM, WG, Tau e RL, somente o método c-medóides não foi capaz de determinar corretamente o número de grupos. O índice Dunn apresentou o pior desempenho.

Tabela 25 – Acertos na determinação do número de grupos utilizando os índices de validação (transformada logarítmica)

	Índices de validação						
Métodos	**Dunn**	**CH**	**Gama**	**PBM**	**WG**	**Tau**	**RL**
l. simples		X		X	X	X	X
l. media		X		X	X	X	X
l. completa		X	X	X	X	X	X
Ward	X	X	X	X	X	X	X
k-medias		X	X	X	X	X	X

(continua)

Tabela 25 – (continuação)

	Índices de validação						
Métodos	**Dunn**	**CH**	**Gama**	**PBM**	**WG**	**Tau**	**RL**
k-medóides	X	X	X	X	X	X	X
híbrido	X	X	X	X	X	X	X
c-médias		X	X	X	X	X	X
c-medóides							
c-médias-pf	X	X	X	X	X	X	X

Fonte: Próprio autor

Por fim, a Tabela 26 apresenta o desempenho dos índices após padronização dos dados pela transformada Box-Cox. O índice com o melhor desempenho foi o RL. Os métodos de ligação média, Ward, k-medóides e híbrido não foram capazes de determinar corretamente o número de grupos.

Tabela 26 – Acertos na determinação do número de grupos utilizando os índices de validação (transformada Box-Cox)

	Índices de validação						
Métodos	**Dunn**	**CH**	**Gama**	**PBM**	**WG**	**Tau**	**RL**
l. simples			X	X	X		X
l. media							
l. completa							X
Ward							
k-medias							X
k-medóides							
híbrido							
c-médias							X
c-medóides				X			X
c-médias-pf	X	X		X	X		X

Fonte: Próprio autor

Nas Tabelas 21, 22, 23, 24, 25 e 26 considera-se acerto quando o número de grupos indicado foi igual a três, ou seja, quando o maior valor de cada índice ocorreu para k igual a 3. A taxa de acerto na determinação do número de grupos com a utilização de:

- amostras não padronizadas: 17,14%;
- transformada z-score: 45,71%;
- transformada min-max: 51,42%;
- transformada min-max-m: 10%;

- transformada logarítmica: 80%;
- transformada Box-Cox: 20%.

Os melhores desempenhos entre as transformadas foram obtidos com as transformadas min-max e logarítmica, com taxa de acerto de 51,42% e 80% respectivamente. Estas duas transformadas apresentaram taxas superiores à taxa obtida com as amostras não padronizadas, que foi de 17,14%. Os piores resultados entre as transformadas ocorreram com as transformadas min-max-m e Box-Cox, com taxas de acerto de 10% e 20% respectivamente. Por outro lado, avaliando os métodos de agrupamento as taxas de acerto foram:

- l. simples: 21,42%;
- l. média: 16,66%;
- l. completa: 16,66%
- hward: 50%;
- k-médias: 38,09%;
- k-medóides: 50%;
- híbrido: 40,47%;
- c-médias: 57,14%;
- c-medóides: 9,52%;
- c-medias c/ pf: 73,80%;

Entre os métodos de agrupamento os que apresentaram os melhores desempenhos foram c-médias e c-médias com polinômio fuzzificador com taxas de 57,14% e 73,80%. Já o pior resultado foi obtido com o método c-medóides com taxa de 9,52%. No tocante aos índices de validação que foram utilizados para determinação do número de grupos na base real, estes tiveram a seguinte taxa de acerto:

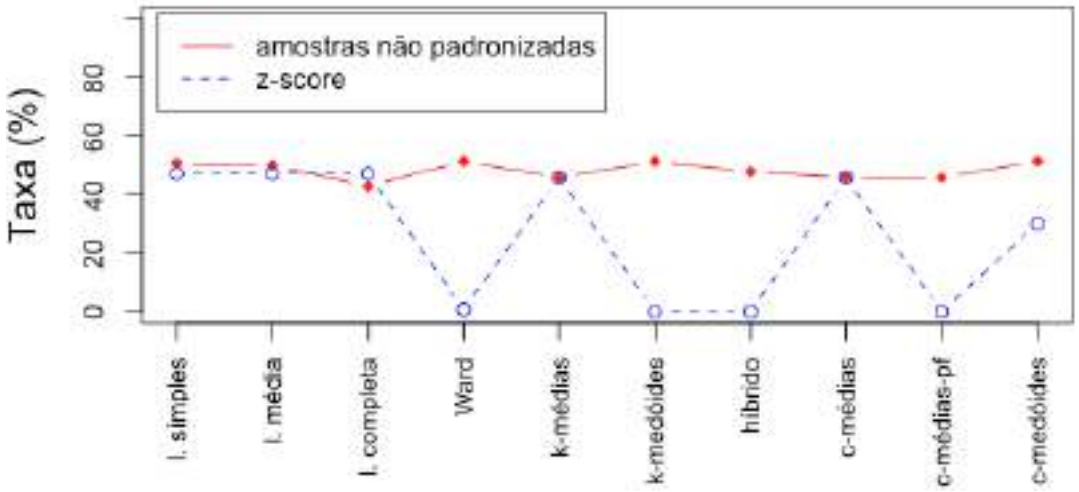
- Dunn: 20%;
- CH: 40%;
- gama: 31,66%;
- pbm: 36,66%;
- WG: 36,66%;
- Tau: 43,33%;
- RL: 53,33%.

Os índices de validação interna com as maiores taxas de acerto foram Tau e RL com taxa de 43,33% e 53,33%.

Em contrapartida, as Figuras 27, 28, 29, 30 e 31 mostraram o impacto da padronização dos dados na habilidade dos métodos de agrupamento em determinar os grupos da base de 140 amostras, através do cálculo do erro conforme a equação 40.

A Figura 27 compara o desempenho dos métodos de agrupamento aplicados às amostras não padronizadas e padronizadas pela transformada z-score. Observou-se, nesse sentido, que os métodos Ward, k-medóides, híbrido e c-médias com polinômio fuzzificador, foram capazes de determinar a partição correta das amostras padronizadas, isto é, a taxa de erro foi zero.

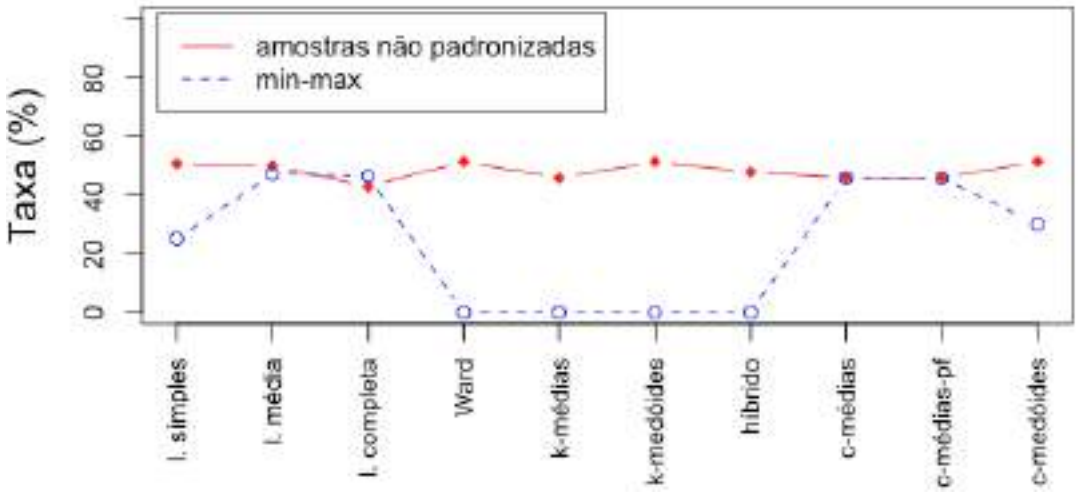
Figura 27 – Gráfico que compara o desempenho dos métodos de agrupamento utilizando amostras não padronizadas e padronizadas pela transformada z-score



Fonte: Próprio autor

O impacto da transformada min-max levou a uma diminuição da taxa de erro ao se comparar os resultados obtidos com as amostras não padronizadas e padronizadas, como mostra a Figura 28. Mais uma vez, os métodos Ward, k-medóides e híbrido, aplicados nas amostras padronizadas, apresentaram uma taxa de erro zero, além do método k-médias.

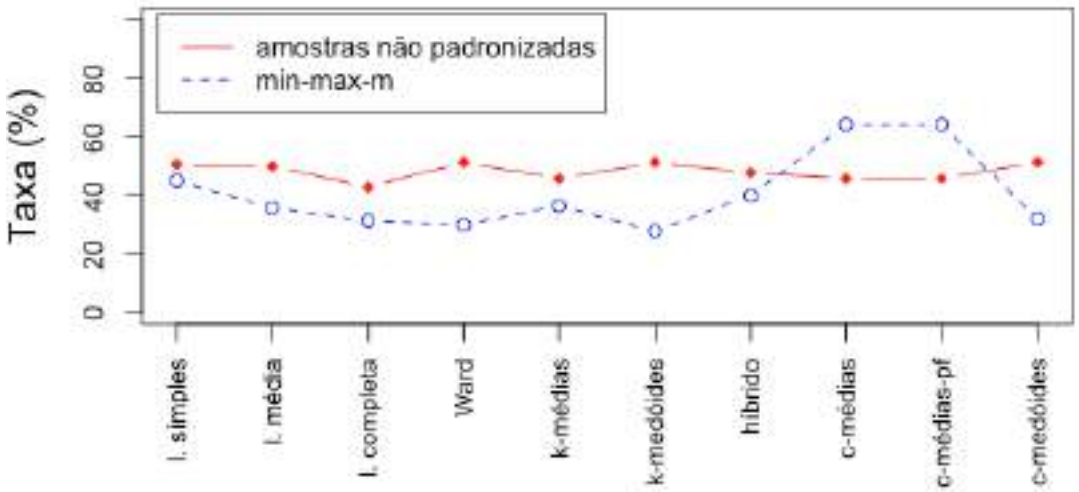
Figura 28 - Gráfico que compara o desempenho dos métodos de agrupamento utilizando amostras não padronizadas e padronizadas pela transformada min-max



Fonte: Próprio autor

Na Figura 29, observa-se após a aplicação da transformada min-max-m os métodos de agrupamento não conseguiram determinar corretamente os grupos da base de amostras.

Figura 29 - Gráfico que compara o desempenho dos métodos de agrupamento utilizando amostras não padronizadas e padronizadas pela transformada min-max-m



Fonte: Próprio autor

A transformada logarítmica aplicada à base de amostras impactou positivamente na determinação dos grupos pelos métodos de agrupamento hierárquicos, k-médias, k-medóides e híbrido, visto que, a taxa de erro foi zero.

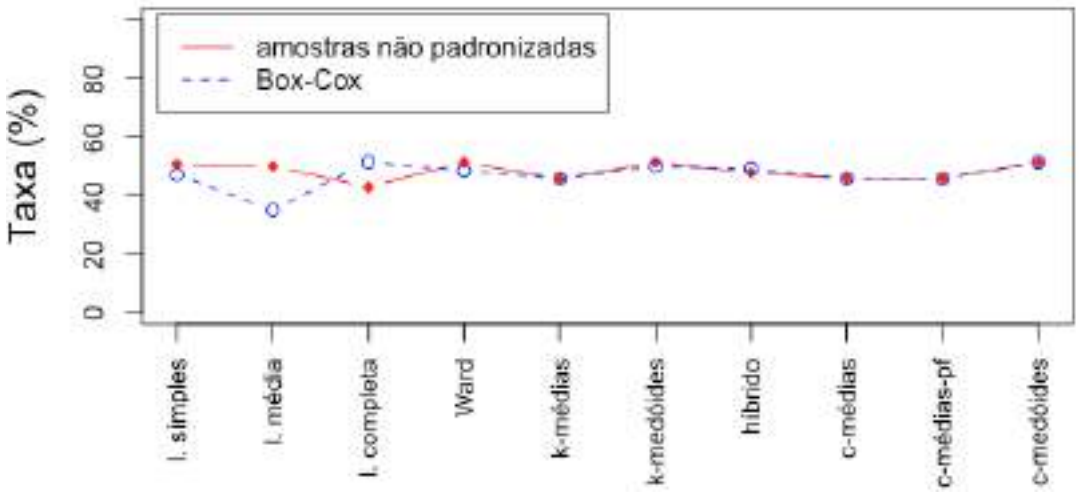
O mesmo não ocorreu com as amostras não padronizadas, com taxa de erro superior à 40%, como mostra a Figura 30.

Figura 30 - Gráfico que compara o desempenho dos métodos de agrupamento utilizando amostras não padronizadas e as padronizadas pela transformada logarítmica

Fonte: Próprio autor

Finalmente, no caso da transformadas Box-Cox os únicos métodos de agrupamento que apresentaram desempenho superior das amostras padronizadas com relação às amostras não padronizadas foram: ligação simples, média e Ward. Como pode ser observado na Figura 31.

Figura 31 - Gráfico que compara o desempenho dos métodos de agrupamento utilizando amostras não padronizadas e as padronizadas pela transformada Box-Cox



Fonte: Próprio autor

Na próxima seção, é apresentada a aplicação web, que incorpora os métodos de pré-processamento estudados, além de funções desenvolvidas para visualização de gráficos e imagens.

9 APLICAÇÃO WEB

Durante o desenvolvimento deste trabalho, foi elaborada uma aplicação web que incorpora técnicas de pré-processamento de dados, tais como: imputação, detecção de outliers e padronização de dados. Além disso, fazem parte da aplicação, técnicas que vão desde a verificação de tendência de agrupamento (estatística de Hopkins e imagens VAT), até a determinação dos grupos por métodos de agrupamento. Sendo possível a visualização dos grupos formados através: da análise de componentes principais e da análise de discriminantes lineares. Desse modo, permitem, analisar a repercussão de técnicas de pré-processamento dos dados, na análise de agrupamentos, uma vez que a alteração de parâmetros, leva a uma atualização dos resultados visualizados.

Diferente da ferramenta disponível no pacote visxhclust (HENKIN e BARNES, 2022), que utiliza apenas métodos de agrupamento hierárquicos e para padronização somente a transformada z-score. A aplicação descrita neste trabalho utiliza métodos de agrupamento hierárquicos, particionais/crisp, baseados em lógica fuzzy, além de cinco transformadas para padronização de dados.

Por outro lado, a aplicação web ClustVis (METSALU e VILO, 2015), permite a elaboração de gráficos de dispersão através da PCA, todavia, esta, não apresenta recursos de análise de agrupamento, tampouco a análise de discriminantes lineares. Para o pré-processamento de dados possui apenas a transformada logarítmica.

A aplicação web exposta neste trabalho, é interativa, de fácil uso, com uma interface leve e amigável, de forma que proporciona a utilização de diversos métodos sem a necessidade de conhecimento em programação por parte do usuário. Ela foi implementada utilizando o pacote Shiny (CHANG et al., 2022) no ambiente R e se propõe a operar como uma ferramenta de apoio à análise de dados, podendo ser acessada a partir de qualquer navegador, sem a necessidade de instalação de softwares ou pacotes.

O pacote Shiny permite a desenvolvedores implementar todas as ferramentas estatísticas do R em um app, sem a necessidade de conhecimento de estruturas de linguagens de desenvolvimento web. Isto ocorre tendo em vista que o Shiny faz uma conversão de linguagem do R para HTM

No pacote Shiny um app possui três componentes, mostrado no script da Figura 32:

- objeto *user* interface (*ui*): controla o *layout* e aparência do app (Figura 32);
- função *server*: contém instruções necessárias para construir o app (Figura 32);
- função *shinyApp*: cria o app a partir do par *ui/server* (Figura 32).

Figura 32 - Script de app

```
> library(shiny)
> ui <- fluidPage("Hello, world!")
> server <- function(input, output, session) { }
> shinyApp(ui, server)
```

Fonte: https://mastering-shiny.org/basic-app.html

A seguir são descritos os recursos da aplicação desenvolvida. Ao acessar o link https://webaplication.shinyapps.io/app_v2/ é aberta uma nova aba do navegador padrão, conforme mostra a Figura 33. Seu *layout* é dividido em duas partes: do lado esquerdo é possível carregar bases de dados e alterar configurações, do lado direito, torna visível a base carregada, além de permitir a visualização de gráficos e imagens.

A aplicação importa arquivos no formato txt, de forma que para carregar o arquivo basta clicar abaixo de “CHOOSE txt FILE” no canto superior esquerdo (Figura 34) e selecionar o arquivo da base de dados (Figura 35). Em seguida a base poderá ser visualizada clicando-se na aba *Data e* utilizando-se as barras de rolagens, (figura 36). Caso a base possua valores ausentes, eles devem ser indicados por “NA” e o separador decimal dos valores das amostras deve ser o ponto.

Figura 33 – Aplicação desenvolvida

Fonte: Próprio autor

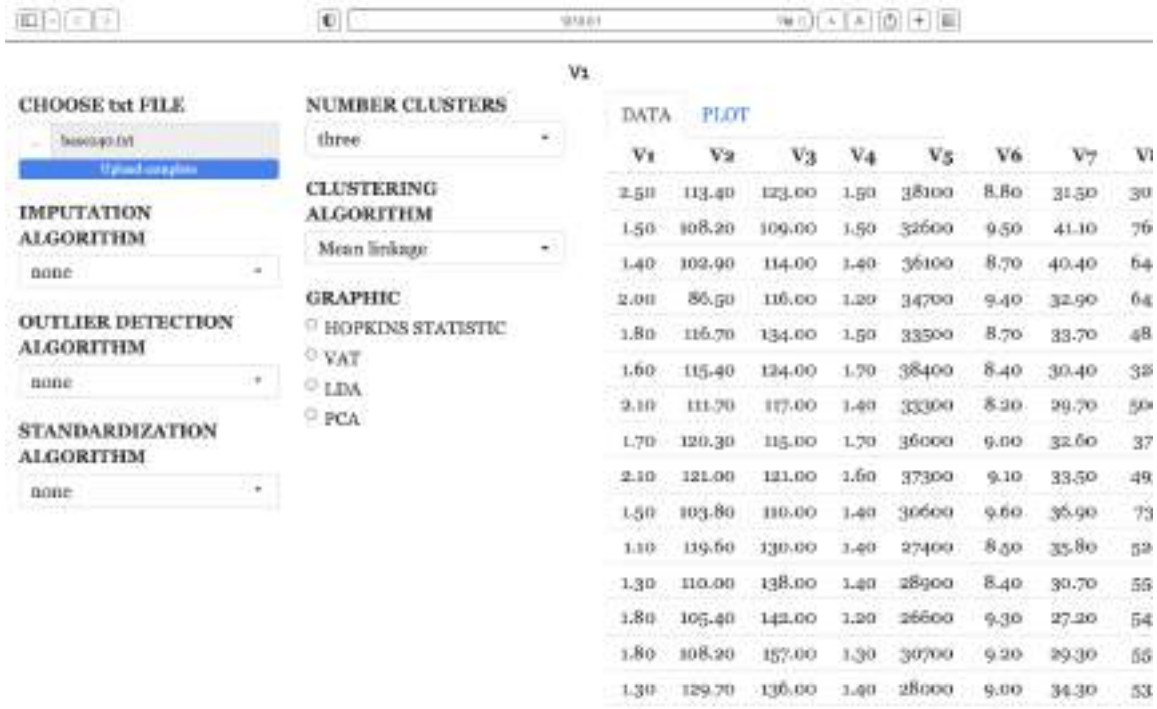
Figura 34 – Como selecionar um arquivo txt

Fonte: Próprio autor

Figura 35 – Selecionando um arquivo txt

Fonte: Próprio autor

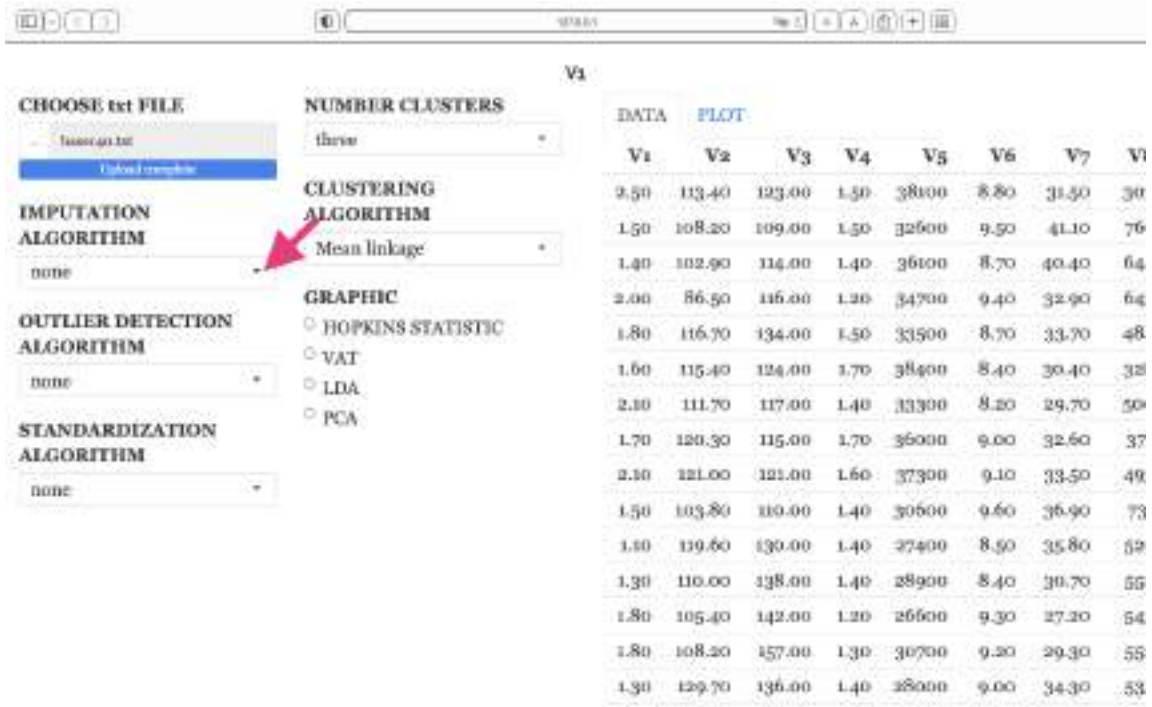
Figura 36 – Visualização da base selecionada



Fonte: Próprio autor

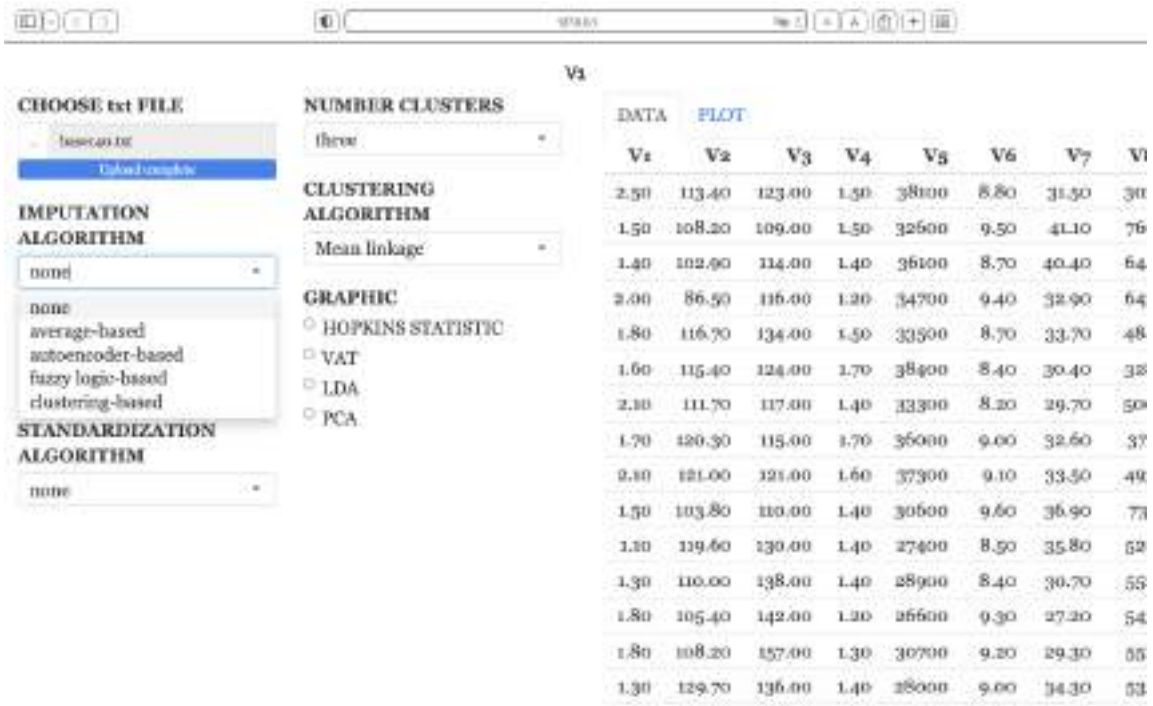
Ademais, se a base de dados possui valores ausentes é possível escolher entre quatro métodos de imputação de dados (média, rede autoencoder, c-médias e agrupamento), bastando para isso clicar abaixo de Imputation algorithm (Figuras 37) e selecionar o método desejado (Figura 38).

Figura 37 – Como selecionar o método de imputação de dados



Fonte: Próprio autor

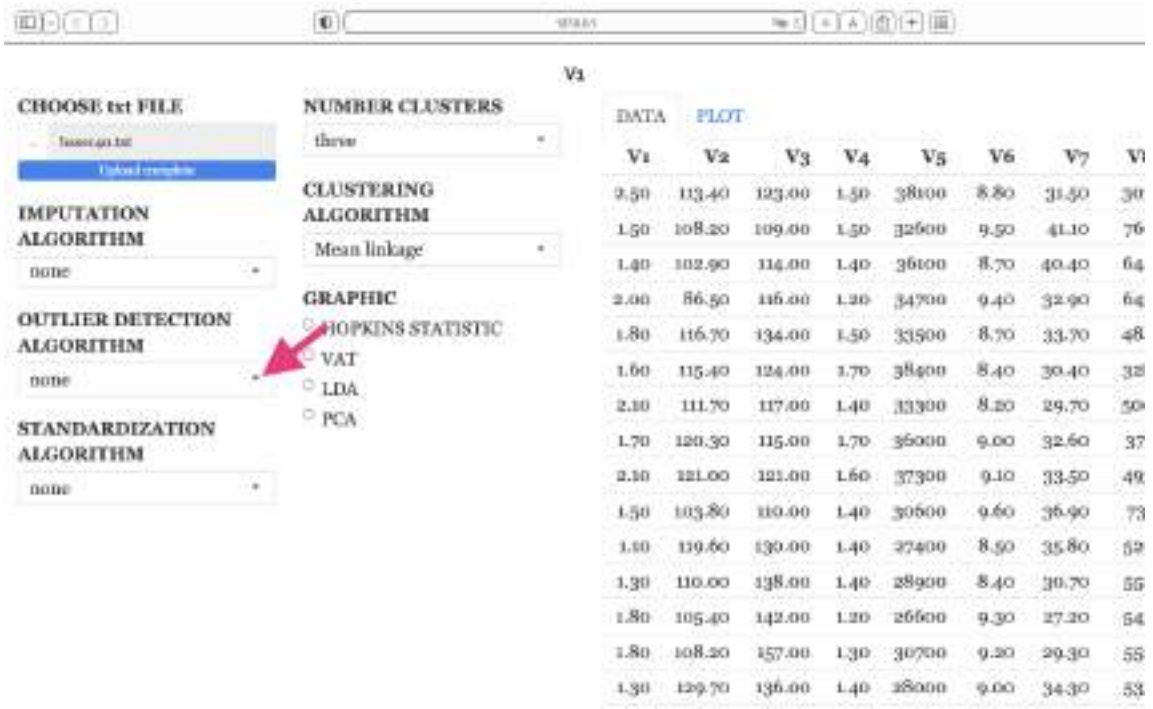
Figura 38 – Selecionado o método de imputação de dados



Fonte: Próprio autor

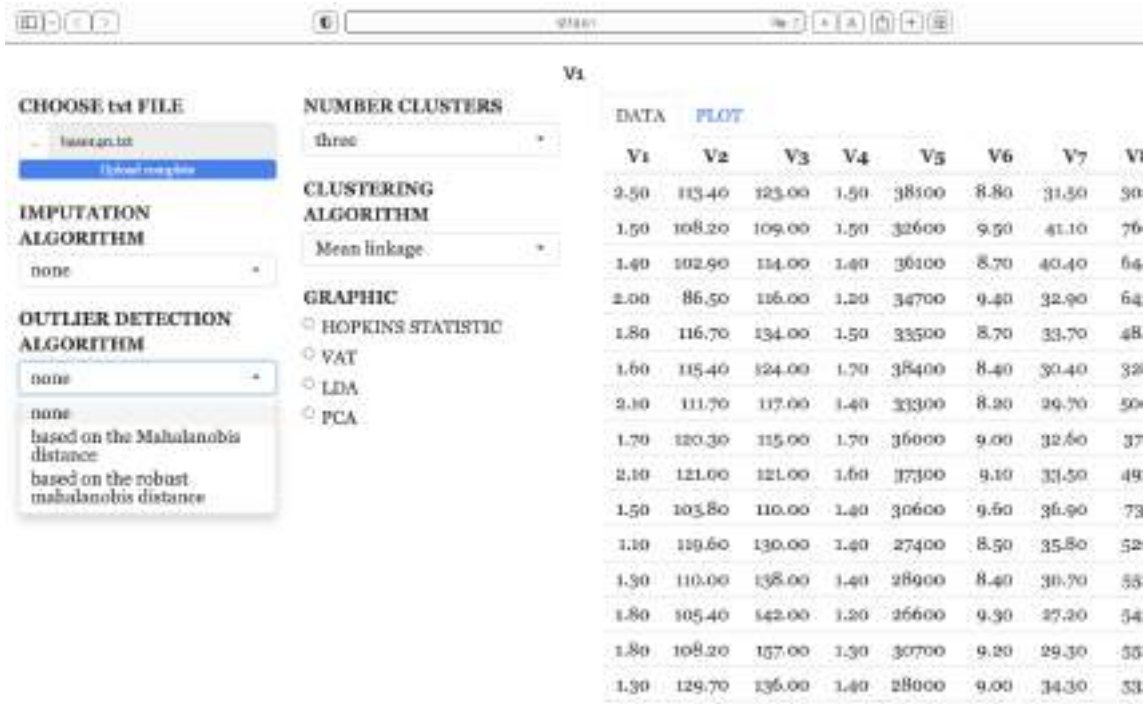
Para detecção de outliers a aplicação possui dois métodos, que utilizam as distâncias: Mahalanobis e Mahalanobis Robusta, de maneira que basta clicar abaixo de *Outlier detection algorithm*, (Figura 39). Depois, pode-se escolher o método pretendido (Figura 40).

Figura 39 – Como selecionar o método de detecção de outliers



Fonte: Próprio autor

Figura 40 – Selecionando método de detecção de outliers

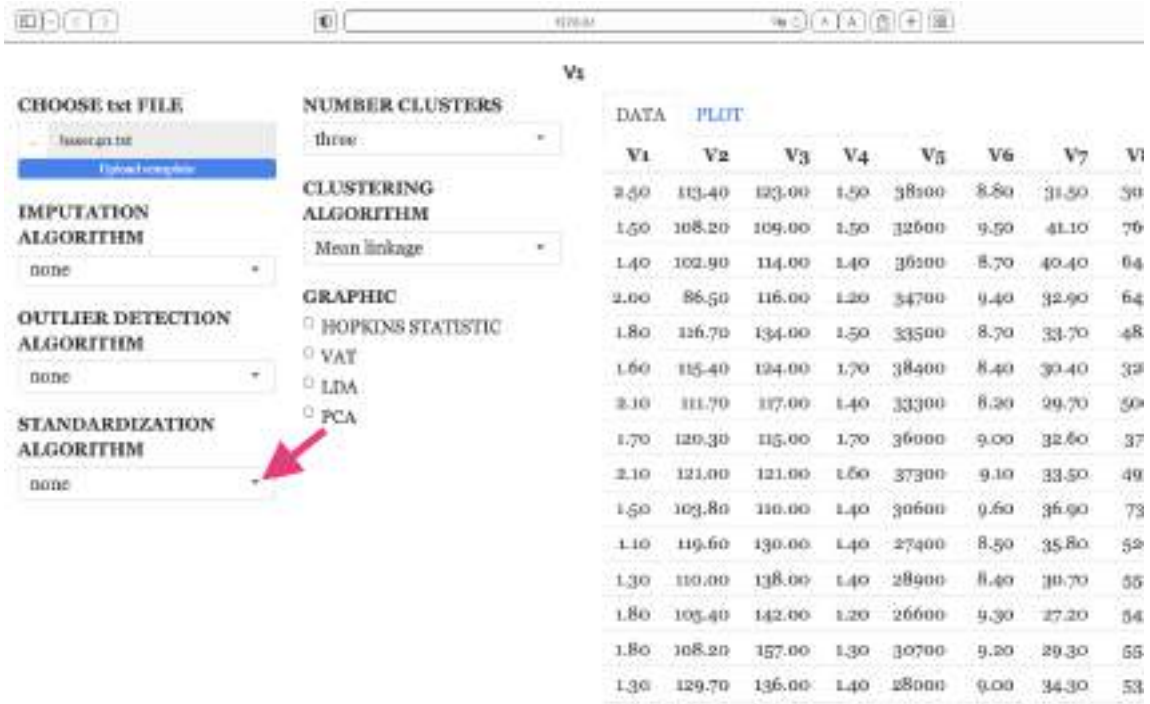


V1	V2	V3	V4	V5	V6	V7	V
2.50	113.40	123.00	1.50	38100	8.80	31.50	50
1.50	108.20	109.00	1.50	32600	9.50	41.10	76
1.40	102.90	114.00	1.40	36100	8.70	40.40	64
2.00	86.50	116.00	1.20	34700	9.40	32.90	64
1.80	116.70	134.00	1.50	33500	8.70	33.70	48
1.60	115.40	124.00	1.70	38400	8.40	30.40	32
2.10	111.70	117.00	1.40	33300	8.20	29.70	50
1.70	120.30	115.00	1.70	36000	9.00	32.60	37
2.10	121.00	121.00	1.60	37300	9.10	33.50	49
1.50	103.80	110.00	1.40	30600	9.60	36.90	73
1.10	119.60	130.00	1.40	27400	8.50	35.80	52
1.30	110.00	138.00	1.40	28900	8.40	30.70	55
1.80	105.40	142.00	1.20	26600	9.30	27.20	54
1.80	108.20	157.00	1.30	30700	9.20	29.30	55
1.30	129.70	136.00	1.40	28000	9.00	34.30	53

Fonte: Próprio autor

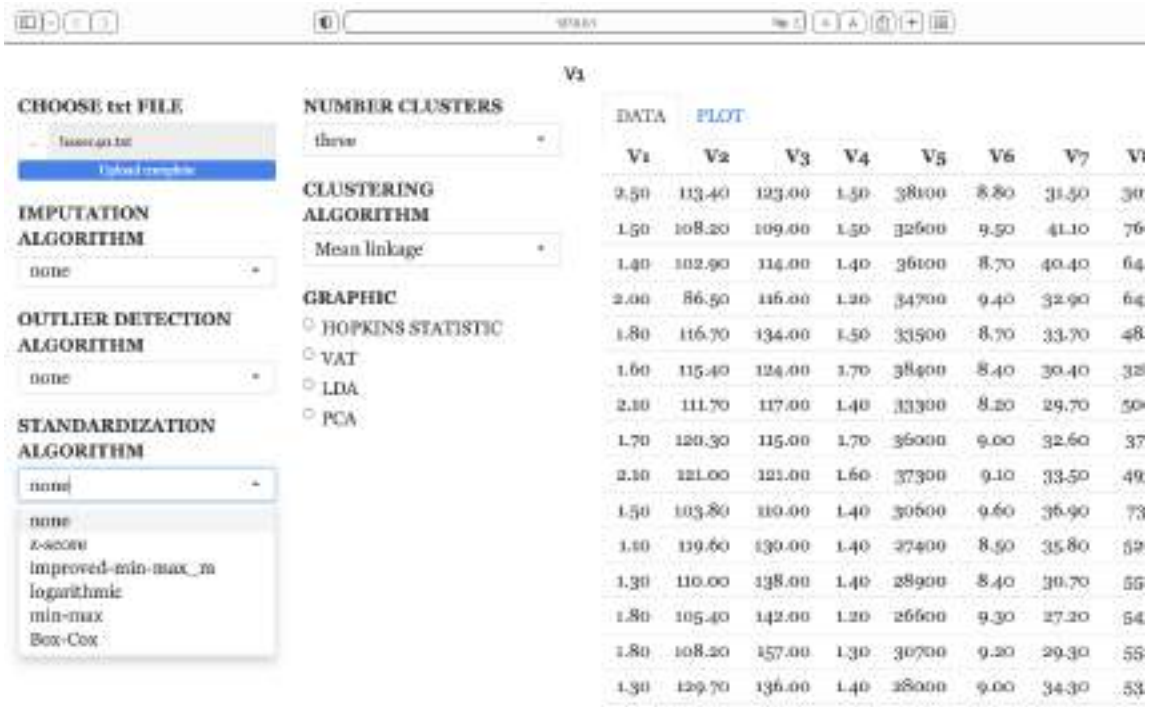
Cinco transformadas (z-score, min-max, min-max-m, logarítmica e Box-Cox) estão disponíveis para padronização dos dados, clicando abaixo de *Standardization algorithm*, (Figura 41), apenas é necessário escolher uma das transformadas (Figura 42).

Figura 41: Como selecionar transformada para padronização dos dados



Fonte: Próprio autor

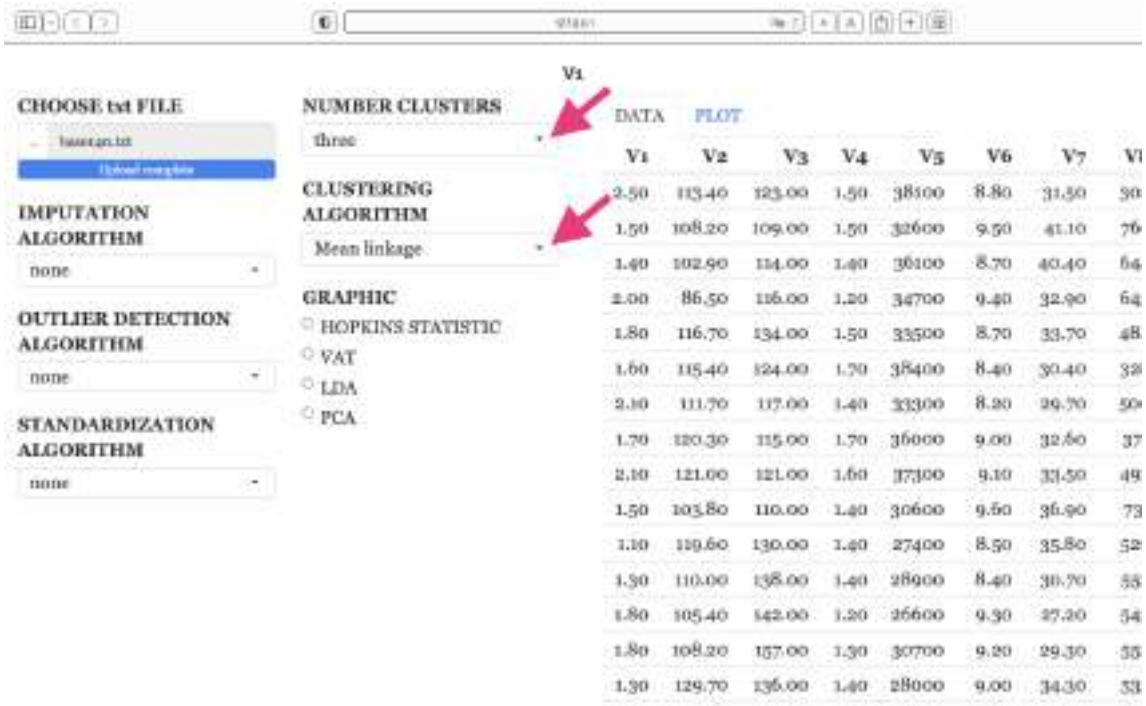
Figura 42 – Selecionando a transformada para padronização dos dados



Fonte: Próprio autor

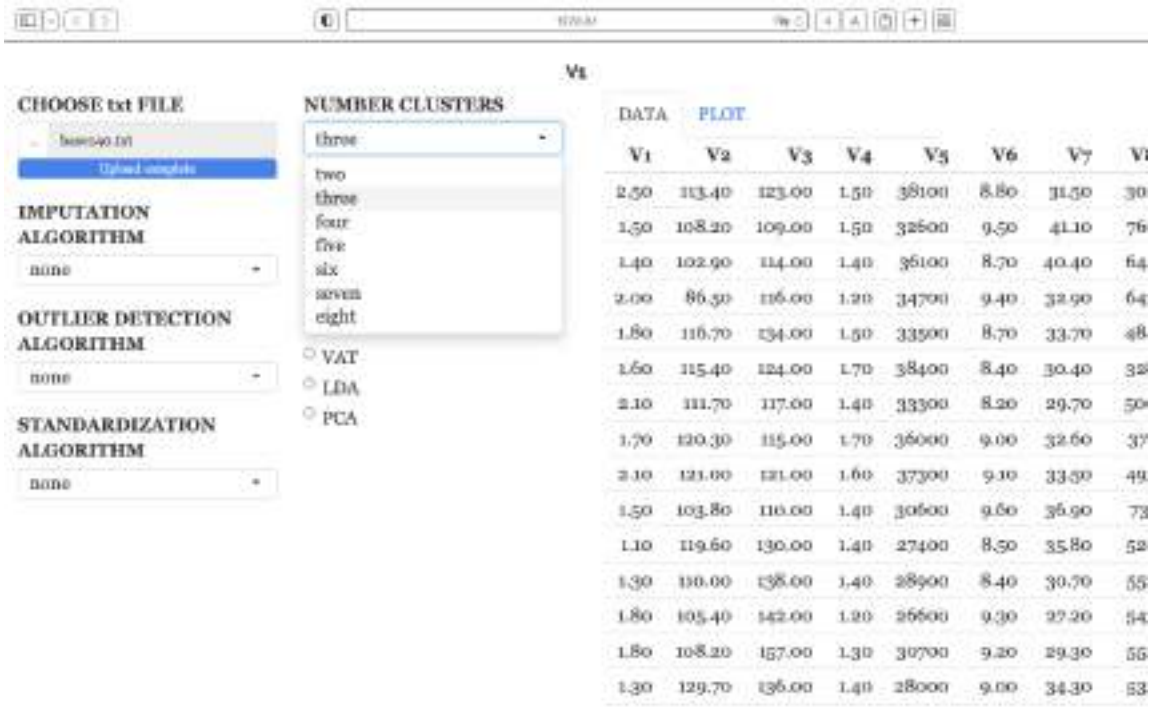
Em *number clusters* (figuras 43 e 44) indica-se o número de grupos (com valores entre 2 e 8) que será utilizado nos métodos de agrupamento (de ligação simples, ligação média, ligação completa, Ward, k-médias, k-medóides, híbrido, c-médias, c-medóides, c-médias-pf) que pode ser selecionado em *Clustering algorithm* (Figuras 43 e 45).

Figura 43 - Como selecionar o número de grupos e o método de agrupamento



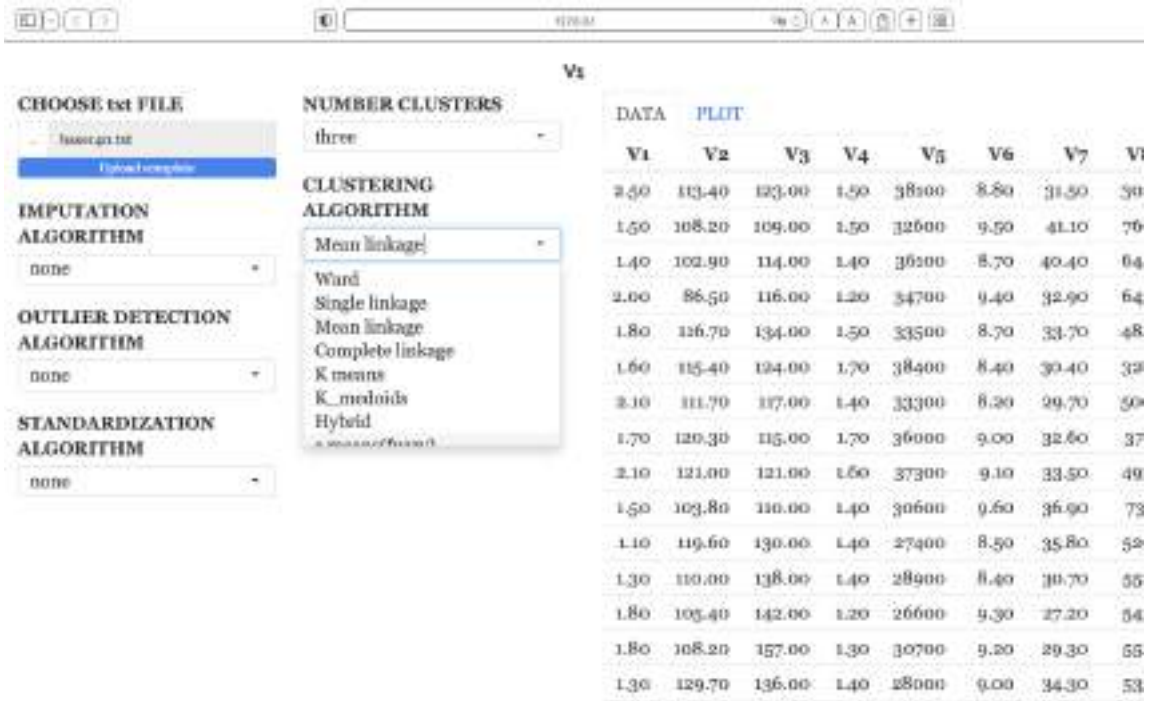
Fonte: Próprio autor

Figura 44 – Selecionando o número de grupos utilizado nos métodos de agrupamento



Fonte: Próprio autor

Figura 45 – Selecionando o método de agrupamento

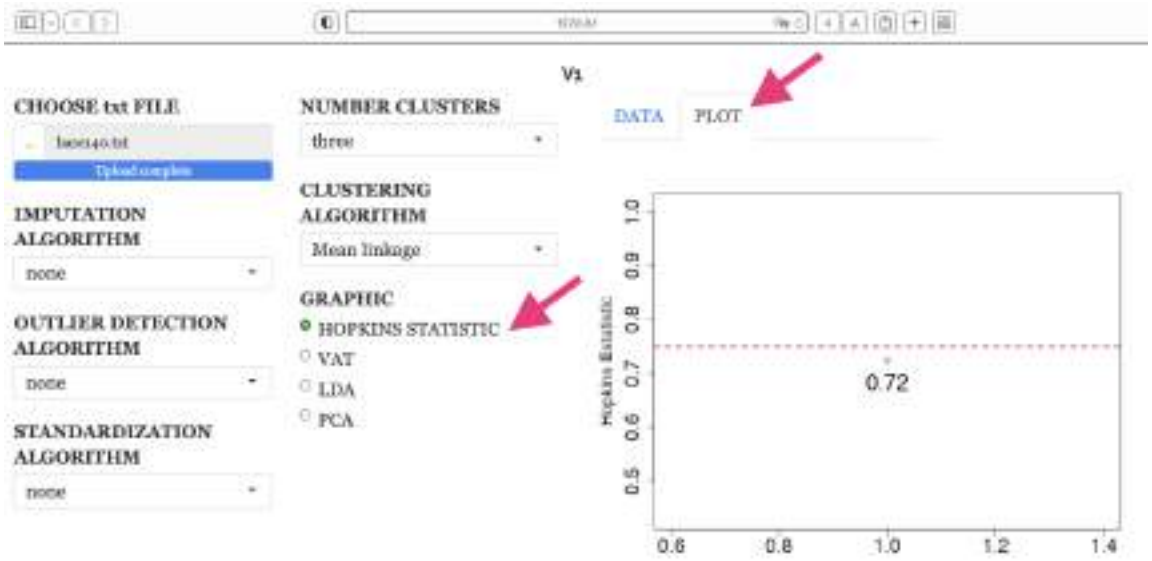


Fonte: Próprio autor

A aplicação permite a visualização de gráficos, somente é preciso clicar na aba *Plot*, para visualização dos gráficos:

-*Hopkins Statistic* (gráfico que mostra o valor da estatística de Hopkins), Figura 46;

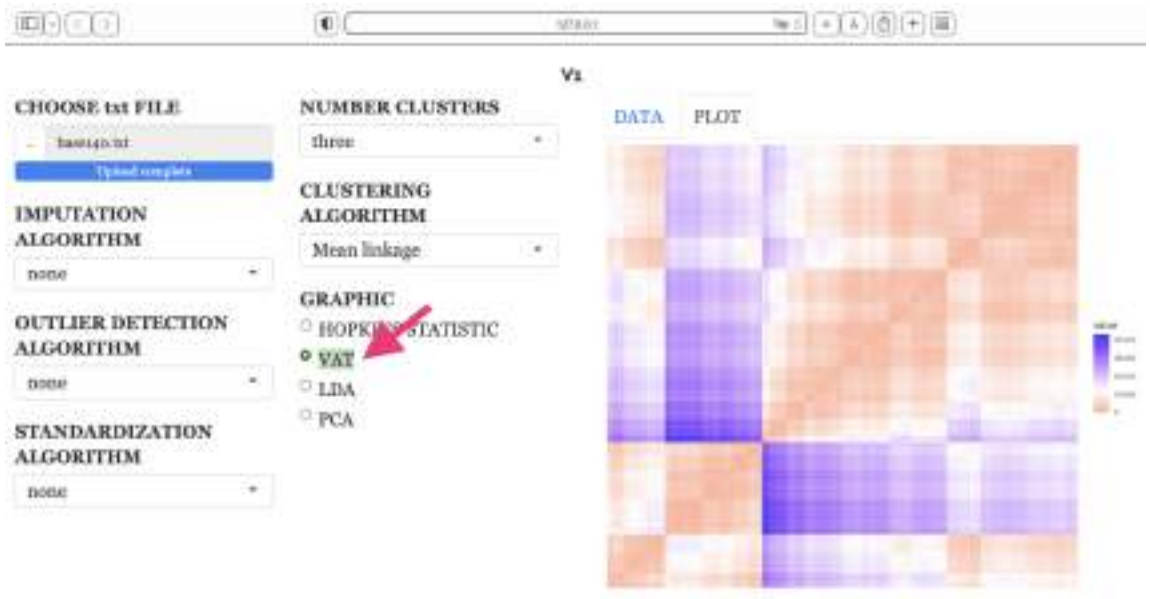
Figura 46 – Selecionando o gráfico que indica a estatística de Hopkins



Fonte: Próprio autor

- *VAT* (Avalidação Visual de tendência de agrupamento), Figura 47;

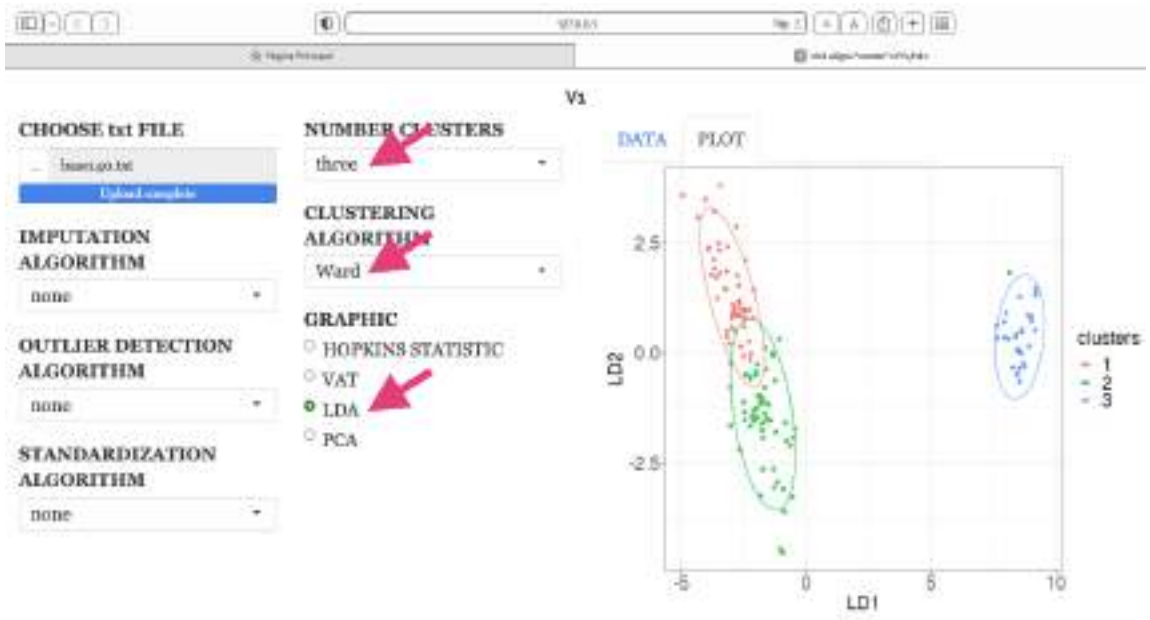
Figura 47 – Selecionado a imagem VAT



Fonte: Próprio autor

- *LDA* (gráfico gerado utilizando análise discriminante) ,Figura 48;

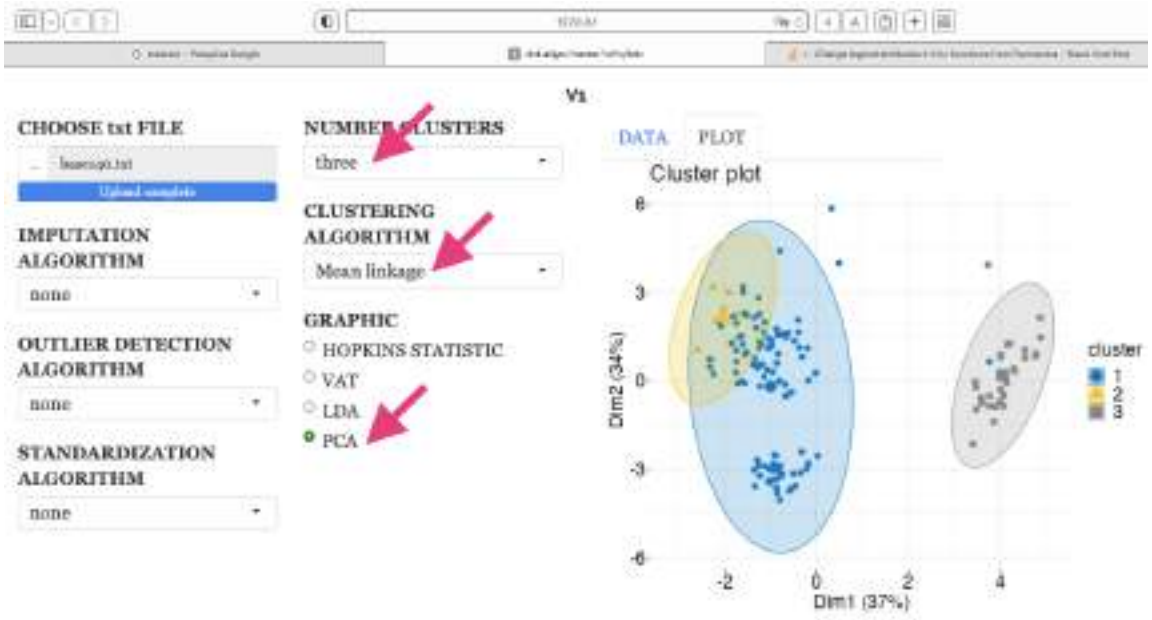
Figura 48 – Selecionando o gráfico que mostra estrutura dos grupos utilizando análise discriminante linear



Fonte: Próprio autor

- *cluster-pca* (gráfico em duas dimensões utilizando PCA), Figura 49;

Figura 49 – Selecionando o gráfico que mostra os grupos utilizando análise de componentes principais



Fonte: Próprio autor

As Tabelas 27 e 28 apresentaram os pacotes e funções utilizados para realização dos testes e desenvolvimento do app.

Tabela 27 – Pacotes utilizados para desenvolvimento do app web

	Pacotes utilizados
Métodos de imputação	*ANN2* (LAMMERS, 2020), *ppclust* (CEBECI, 2019), *neighbr* (BOLOTOV, 2020), *fclust* (FERRARO et al., 2019)
Métodos de detecção de outliers	*stats* (R CORE TEAM, 2022), *robust* (WANG et al., 2022), ISLR (JAMES, et al., 2021)
Métodos de padronização	*DescTools* (SIGNORELLI, 2022)
Métodos de agrupamento	*stats, factoextra* (KASSAMBARA e MUNDT, 2020), *cluster* (MAECHLEAR et al., 2022), *ppclust*
Gráficos	*factoextra, clustertend* (WRIGHTet al., 2022), *GGally* (SCHLOERKE et al., 2021), *MASS* (VENABLES e RIPLEY, 2002), *ggplot2* (WICKHAM, 2016)
Aplicação web	*Shiny* (CHANG et al., 2022), *shinythemes* (CHANG, 2020)

Fonte: Próprio autor

Tabela 28 – Funções utilizadas nos testes e no app web

	Funcões
Métodos de imputação	*Mean, is.na, dim, as.data.frame, autoencoder, fcm, ppclust2*
Métodos de detecção de outliers	*mahalanobis, cov, covRob, cook.distance, qf, qchisq*,
Métodos de padronização	*log10, BoxCoxLambda, BoxCox*
Métodos de agrupamento	*Dist, hclust, cuttree, hkmeans, kmeans, pam, FKM, FKM.med, FKM.fp*
Gráficos	*ggplot, fviz_cluster, get_clust_tendency, plot, pairs, prcomp, fviz_dist, ggpairs, fviz_dend, lda*
Aplicação web	*titlePanel, fluidRow, selectInput, radioButtons, mainPanel, renderTable, renderPlot, read.table, validate,*

Fonte: Próprio autor

10 CONCLUSÃO

Os testes realizados com a base de 140 amostras mostraram que os métodos de imputação apresentaram desempenho semelhante. Ao aplicar os métodos de imputação em cada sítio individualmente, ocorreu diminuição do erro quadrático médio normalizado, pois há uma diminuição da amplitude das variáveis. Não houve impacto da imputação de dados na determinação dos grupos pelos métodos de agrupamento: hierárquicos, particionais/crisp, c-médias e c-médias com polinômio fuzzificador. Deste modo, analisando o tempo de processamento, destaca-se o método de imputação baseado na média. Ademais, levando-se em conta o tudo o que foi exposto, o método de imputação pela média torna-se a melhor escolha.

A exclusão dos outliers detectados pela distância Mahalanobis acarretou um aumento da coesão dos sítios B e C, bem como diminuiu a taxa de erro dos métodos de agrupamento c-médias e c-médias com polinômio fuzzificador.

Após a aplicação das transformadas, houve alteração significativa na estatística de Hopkins, uma vez que depois da aplicação das transformadas concluiu-se que há tendência de agrupamento, o que não ocorreu com as amostras não padronizadas.

A padronização das amostras realizadas por meio das transformadas z-score, min-max e logarítmica permitiu às imagens VAT, indicarem tendência de agrupamento, e o número de grupos da base de amostras. No caso das imagens VAT correspondentes às amostras não padronizadas e as transformadas min-max-m e Box-Cox, verificou-se apenas a tendência de agrupamento.

Na determinação do número de grupos utilizando os índices de validação interna, a transformada com melhor desempenho foi a logarítmica, já o método de agrupamento com melhor desempenho foi o c-médias com polinômio fuzzificador. O índice de validação com a maior taxa de acerto foi o RL.

Por fim, na determinação dos grupos pelos métodos de agrupamento estudados, a transformada com o melhor desempenho foi a logarítmica, visto que permitiu que os métodos de agrupamento hierárquicos e particionais/crisp determinassem corretamente os três grupos.

Neste trabalho foi desenvolvida uma aplicação web, para ajudar pesquisadores no pré-processamento de dados, análise de agrupamentos e visualização dos resultados. A aplicação proporciona a utilização de diversos métodos sem a necessidade de conhecimento de programação por parte do usuário.

TRABALHOS FUTUROS

Avaliação do impacto da padronização de amostras na imputação e na detecção de outliers.

Atualizar o aplicativo com: outras transformadas para padronização das amostras e técnicas de imputação múltipla.

Estudo de técnicas de seleção de variáveis.

Classificação de amostras utilizando redes neurais, PCA, transformada wavelet e lógica fuzzy.

REFERÊNCIAS BIBLIOGRÁFICAS

ABIRI, N.; LINSE, B.; EDÉN, P.; OHLSSON, M. Establishing strong imputation performance of a denoising autoencoder in a wide range of missing data problems. ***Neurocomputing***, v. 365, p. 137–147, 2019. Disponível em : <https://www.sciencedirect.com> Acesso em : 15 abr. 2020. < https://doi.org/10.1016/j.neucom.2019.07.065.>

AGGARWAL, C. C. ***Outlier analysis***. 2nd ed. New York: Springer, 2017.

AGGARWAL, C. C.; YU, P. S. An effective and efficient algorithm for high-dimensional outlier detection. ***International Journal on Very Large Data Bases Journal***, v. 14, n. 2, p. 211-221, 2005. Disponível em : < https://link.springer.com/> Acesso em : 20 maio 2019. <https://doi.org/10.1007/s00778-004-0125-5>

ALLISON, P. D. ***Missing data***. Thousand Oaks: Sage, 2001.

ALVES, M. A. Ceramists cultures of São Paulo and Minas Gerais: technical typological study. ***Revista do Museu de Arqueologia e Etnologia***, n· 1, p. 71-96, 1991. Disponível em : <https://www.revistas.usp.br/revmae/> Acesso em : 10 dez. 2020. https://doi.org/10.11606/issn.2448-1750.revmae.1991.107946.

ALVES, M. A. Technical study on prehistoric ceramic of Brazil. ***Revista do Museu de Arqueologia e Etnologia***, n.4, p. 39-70, 1994. Disponível em : <https://www.revistas.usp.br/revmae/> Acesso em : 10 dez. 2020. https://doi.org/10.11606/issn.2448-1750.revmae.1994.109194.

AL SHALABI, L.; SHAABAN, Z.; KASASBEH, B. Data mining: a preprocessing engine. ***Journal of Computer Science***, v. 2, n. 9, p. 735-739, 2006.

ARBELAITZ, O.; GURRUTXAGA, I.; MUGUERZA, J.; PÉREZ, J. M.; PERONA, I. An extensive comparative study of cluster validity indices. ***Pattern Recognition***, v. 46, n. 1, p. 243-256, 2013. Disponível em : < https://www.sciencedirect.com/> Acesso em : 3 nov. 2019. < https://doi.org/10.1016/j.patcog.2012.07.021>

ATKINSON, A. C.; RIANI, M.; CORBELLINI, A. The box–cox transformation: review and extensions. ***Statistical Science***, v. 36, n. 2, p. 239-255, 2021. Disponível em < https://projecteuclid.org/ > Acesso em : 10 jun. 2019. < https://doi.org/10.1214/20-STS778 >

BARNETT, V.; LEWIS, T. ***Outliers in statistical data.*** 3rd ed. California: Wiley, 1994.

BAXTER, M. J. A review of supervised and unsupervised pattern recognition in archaeometry. ***Archaeometry***, v. 48, n. 4, p. 671-694, 2006. Disponível em: < https://onlinelibrary.wiley.com/> Acesso em 4 nov. 2018. < https://doi.org/10.1111/j.1475-4754.2006.00280.x >

BAXTER, M. J. Archaeological data analysis and fuzzy clustering. ***Archaeometry***, v. 51, n. 6, p. 1035-1054, 2009. Disponível em : < https://onlinelibrary.wiley.com> Acesso em : 3 out. 2019.< https://doi.org/10.1111/j.1475-4754.2008.00449.x.>

BAXTER, M. J. Mathematics, statistics and archaeometry: the past 50 years or so. ***Archaeometry***, v. 50, n.6, p. 968-982, 2008. Disponível em : < https://onlinelibrary.wiley.com> Acesso em : 7 out. 2019. <https://doi.org/10.1111/j.1475-4754.2008.00427.x.>

BEELEY, C. ***Web aplication development with R using shiny***. 2nd ed. Birbmingham: Packt, 2016.

BEZDEK, J. C.; HATHAWAY, R. J. VAT: a tool for visual assessment of (cluster) tendency. In: PROCEEDINGS OF THE INTERNATIONAL JOINT CONFERENCE ON NEURAL NETWORKS, May 12-17, 2002, Honolulu, USA. ***Proceedings…*** p. 2225-2230. Diponível em : < https://ieeexplore.ieee.org/ > Acesso em : 7 maio 2019. https://doi.org/10.1109/IJCNN.2002.1007487.

BEZDEK, James C.; EHRLICH, R.; FULL, W. FCM: the fuzzy c-means clustering algorithm. ***Computers and Geosciences***, v. 10, p. 191-203, 1984. Disponível em : < https://www.sciencedirect.com/> Acesso em : 15 jun. 2019. https://doi.org/10.1016/0098-3004(84)90020-7.

BOLOTOV, D. ***neighbr: classification, regression, clustering with k nearest neighbors***. 2020. R package version 1.0.3. Disponível em : < https://CRAN.R-project.org/package=neighbr>.

BORA, D. J.; GUPTA, A. K. A Comparative study between fuzzy clustering algorithm and hard clustering algorithm. ***International Journal of Computer Trends and Technology***, v. 10, n. 2, p. 108–113, 2014. Disponível em : <https://arxiv.org/> Acesso em : 15 fev. 2019. < https://doi.org/10.14445/22312803/IJCTT-V10P119.>

BRAGA, A. de P.; CARVALHO, A. P. de L. F. de; LUDERMIR, T. B. ***Redes neurais artificiais: Teoria e aplicações***. 2. ed. Rio de Janeiro: LTC, 2007.

BRÁS, L. P.; MENEZES, J. C. Improving cluster-based missing value estimation of DNA microarray data. ***Biomolecular Engineering***, v. 24, n. 2, p.273-282, 2007. Disponível em : <https://www.sciencedirect.com/> Acesso em : 14 out. 2019. https://doi.org/10.1016/j.bioeng.2007.04.003.

BROWN, D.; JAPA, A.; SHI, Y. A fast density-grid based clustering method. In: IEEE 9TH ANNUAL COMPUTING AND COMMUNICATION WORKSHOP AND CONFERENCE, Jan. 6-8, 2020, Las Vegas, United States, ***Proceedings…*** p. 48-54. Diponível em : <https://ieeexplore.ieee.org > Acesso em : 13 nov. 2021. https://doi.org/10.1109/CCWC.2019.8666548.

BRUN, M.; SIMA, C.; HUA, J.; LOWEY, J.; CARROLL, B.; SUH, E.; DOUGHERTY, E. R. Model-based evaluation of clustering validation measures. ***Pattern Recognition***, v. 40, n. 3, p. 807-824, 2007. Disponível em : <https://www.sciencedirect.com> Acesso em : 20 mar. 2020. https://doi.org/10.1016/j.patcog.2006.06.026.

CARLSON, D. L. ***Quantitative methods in archaeology using R***. 1st ed. Cambridge University, 2017.

CARVALHO, P. R. ***Estudo comparativo dos algoritmos hierárquicos de análise de agrupamento em resultados experimentais***. 2018. 140p. Dissertação (Mestrado em Tecnologia Nuclear) – Instituto de Pesquisas Energéticas e Nucleares, IPEN-CNEN/SP, São Paulo. Disponível em <http://www.teses.usp.br> Acesso em: 15 mar. 2018

CASTRO, L. N. de; FERRARI, D. G**.** ***Introdução à mineração de dados - conceitos básicos, métodos e aplicações***. 1. ed. São Paulo: Saraiva, 2016.

CASTRO, L. S. V. de. ***Pontos de estatística***. 17. ed. Rio de Janeiro: Científica, 1975.

CEBECI, Z. Comparison of internal validity indices for fuzzy clustering. ***Journal of Agricultural Informatics,*** v. 10, n. 2, p. 1-14, 2019. Disponível em : < https://www.cabdirect.org/> Acesso em : out. 2020.

CHANG, W.; CHENG, J.; ALLAIRE, J.; ALAINE, J. J.; CIEVERT, C.; SCHLOERKE, B; XIE, Y.; ALLEN, J.; MCPHERSON, J.; DIPERT, A.; BORGES, B. ***shiny: Web application framework for R***, 2022. R package version 1.7.2. Disponível em : < https://CRAN.R-project.org/package=shiny>

CHANG, W. ***shinythemes: themes for shiny***. 2020. package version 1.2.0. Disponível em : https://CRAN.R-project.org/package=shinythemes

CHARRAD, M.; GHAZZALI, N.; BOITEAU, V.; NIKNAFS, A. Nbclust: an R package for determining the relevant number of clusters in a data set. ***Journal of Statistical Software***, v. 61, p.1-36, 2014. Disponível em : < https://www.jstatsoft.org/> Acesso em : 5 fev. 2020.< https://doi.org/10.18637/jss.v061.i06.>

CHARTE, D.; CHARTE, F.; GARCÍA, S.; DEL JESUS, M. J.; HERRERA, F. A practical tutorial on autoencoders for nonlinear feature fusion: taxonomy, models, software and guidelines. ***Information Fusion***, vol. 44, p. 78-96 ,2018. Disponível em : < https://www.sciencedirect.com/> Acesso em : 10 jun. 2020. https://doi.org/10.1016/j.inffus.2017.12.007.

CHEN, B.; TAI, P. C.; HARRISON, R.; PAN, Y. Novel hybrid hierarchical-k-means clustering method (h-k-means) for microarray analysis. In: 2005 IEEE COMPUTATIONAL SYSTEMS BIOINFORMATICS CONFERENCE, WORKSHOPS AND POSTER ABSTRACTS, Ago. 8-12, 2005, Stanford, United States, ***Proceedings…***, 2005. p. 105-108. Disponível em : < https://ieeexplore.ieee.org/> Acesso em : 14 maio 2019 https://doi.org/10.1109/CSBW.2005.98.

CROSS, G. R.; JAIN, A. K. Measurement of clustering tendency. ***Theory and Application of Digital Control***, Elsevier, 1982. p. 315–320. Disponível em : < https://www.sciencedirect.com/> Acesso em : 3 jan. 2020. https://doi.org/10.1016/B978-0-08-027618-2.50054-1.

CHU, C. W.; HOLLIDAY, J. D.; WILLETT, P. Effect of data standardization on chemical clustering and similarity searching. ***Journal of Chemical Information and Modeling***, v. 49, n. 2, p. 155-161, 2009. Disponível em : <https://pubs.acs.org/> Acesse em : 8 fev. 2019. < https://doi.org/10.1021/ci800224h>

DE MAESSCHALCK, R.; JOUAN-RIMBAUD, D.; MASSART, D. L. The Mahalanobis distance. ***Chemometrics and Intelligent Laboratory Systems***, v. 50, p. 1-18, 2000. Disponível em : <https://www.sciencedirect.com/> Acesso em : 23 out. 2019. https://doi.org/10.1016/S0169-7439(99)00047-7.

DE MICHEAUX, P. L.; DROUILHET, R.; LIQUET, B. ***The R software***. Springer, 2013.

DESGRAUPES, B. ***clusterCrit: Clustering indices***. 2018. package version 1.2.8. Disponível em : < https://CRAN.R-project.org/package=clusterCrit>

EVERITT, B.; HOTHORN, T. **An introduction to applied multivariate analysis with R**. New York: Springer, 2011.

FÁVERO, L. P.; BELFIORE, P. ***Análise de dados: técnicas multivariadas exploratórias com SPSS e STATA*** . 1. ed. Rio de Janeiro: Elsevier, 2017.

FAY, C.; ROCHETTE, S.; GUYADER, V.; GIRARD, C. ***Engineering production-grade shiny apps***. 1st ed. New York: Chapman and Hall/CRC, 2021.

FERRARO, M. B.; GIORDANI, P.; SERAFINI, A. fclust: an R package for fuzzy clustering. ***R Journal***, v. 11, n. 1, p. 1-18, 2019. Disponível em : < https://journal.r-project.org/> Acesso em : 19 jan. 2020. https://doi.org/10.32614/rj-2019-017.

FILZMOSER, P. Identification of multivariate outliers: a performance study. ***Austrian Journal of Statistics***, v. 34, n. 2, p. 127-138, 2016. Disponível em : <https://www.ajs.or.at/> Acesso em : 20 fev. 2019. < https://doi.org/10.17713/ajs.v34i2.406.>

GAN, G.; MA, C.; WU, J. ***Data clustering: theory, algorithms, and applications*****.** 2nd ed. Philadelphia: Society for Industrial and Applied Mathematics, 2020.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A.; BENGIO, Y. ***Deep learning***. Cambridge: MIT, 2016.

GRAHAM, J. W. Missing data analysis: making it work in the real world. **Annual *review of psychology***, v. 60, n. 1, p. 549-576, 2009. Disponível em : < https://www.annualreviews.org/> Acesso em : 17 julho 2022. https://doi.org/10.1146/annurev.psych.58.110405.085530.

GRANJON, D. ***Outstanding user interfaces with shiny***. 1st ed. New York: Chapman and Hall/CRC, 2022.

HAIR, J. F.; BABIN, B. J.; TATHAM, R.; ANDERSON, R.; BLACK, W. **Análise *multivariada de dados***. 6. ed. Porto Alegre: Bookman, 2009.

HALKIDI, M.; BATISTAKIS, Y.; VAZIRGIANNIS, M. On clustering validation techniques. ***Journal of Intelligent Information Systems***, v. 17, n. 2, p. 107-145, 2001. Disponível em : < https://link.springer.com> Acesso em : 22 set. 2019. https://doi.org/10.1023/A:1012801612483.

HARTIGAN, J. A.; WONG, M. A. Algorithm AS 136: A k-Means clustering algorithm. ***Applied Statistics***, v. 28, n.1, p. 100-108, 1979. Disponível em : <https://www.jstor.org/> Acesso em : 12 abr. 2019. https://doi.org/10.2307/2346830.

HAWTHORNE, G.; ELLIOTT, P. Imputing cross-sectional missing data: comparison of common techniques. ***Australian and New Zealand Journal of Psychiatry***, v. 39, n. 7, p. 583-590, 2005. Disponível em : < https://journals.sagepub.com> Acesso em : 18 out. 2019. https://doi.org/10.1111/j.1440-1614.2005.01630.x.

HENKIN, R.; BARNES, M. visxhclust: An R shiny package for visual exploration of hierarchical clustering. ***Journal of Open Source Software***, v. 7, n.70, p. 1-4, 2022. Disponível em : < https://joss.theoj.org/> Acesso em : 8 out. 2020. https://doi.org/10.21105/joss.04074.

HENNIG, C.; MEILA, M.; MURTAGH, F.; ROCCI, R. ***Handbook of cluster analysis***. New York: CRC, 2015.

HODGE, V. J.; AUSTIN, J. A survey of outlier detection methodologies. ***Artificial Intelligence Review***, v. 22, n.2, p. 85-126, 2004. Disponível em : < https://link.springer.com/> Acesso em : 10 jan 2019. https://doi.org/10.1023/B:AIRE.0000045502.10941.a9.

IAM-ON, N.; GARRETT, S. Linkclue: A matlab package for link-based cluster ensembles. ***Journal of Statistical Software***, v. 36, n. 9, p. 1-36, 2010. Disponível em : < https://www.jstatsoft.org> Acesso em : 9 fev. 2019 <https://doi.org/10.18637/jss.v036.i09.>

JADHAV, A.; PRAMOD, D.; RAMANATHAN, K. Comparison of performance of data imputation methods for numeric dataset. ***Applied Artificial Intelligence***, v. 33, n. 10, p. 913-933, 2019. Disponível em: <https://www.tandfonline.com> Acesso em 20 out. 2019. <https://doi.org/10.1080/08839514.2019.1637138>

JAIN, A. K. Data clustering: 50 years beyond K-means. ***Pattern Recognition Letters***, v. 31, n. 8, p. 651-666, 2010. Disponível em : < https://www.sciencedirect.com/ > Acesso em : 10 out. 2019. https://doi.org/10.1016/j.patrec.2009.09.011.

JAIN, A. K.; DUBES, R. C. ***Algorithms for clustering data***. New Jersey: Prentice-Hall, 1988.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. J. Data clustering: a review. ***ACM computing surveys (CSUR)***, v. 31, n. 3, p. 264-323, 1999. Disponível em : < https://dl.acm.org/> Acesso em 4 jan. 2019. < https://doi.org/10.1145/331499.331504>

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. ***ISLR: data for an Introduction to Statistical Learning with Applications in R***. 2021. R package version 1.4. Disponível em : <https://CRAN.R-project.org/package=ISLR>

JIN, W.; TUNG, A. K. H.; HAN, J. Mining top-n local outliers in large databases. In: SEVENTH ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, Aug. 26-29, 2001, San Francisco, United States. ***Proceesing…*** p. 293-298. Disponível em : < https://dl.acm.org/ > Acesso em : 22 maio 2019. https://doi.org/10.1145/502512.502554.

JOHNSON, R. A.; WICHERN, D. W. **Appl*ied multivariate statistical analysis***. 3rd ed. New Jersey: Prentice Hall, 2007.

KABIR, W.; AHMAD, M. O.; SWAMY, M. N. S. A new anchored normalization technique for score-level fusion in multimodal biometrie systems. 2016 IEEE INTERNATIONAL SYMPOSIUM ON CIRCUITS AND SYSTEMS, May 22-25, 2016, Montreal, Canada. ***Proceedings…*** p. 93-96. Disponível em : < https://ieeexplore.ieee.org/> Acesso em : 27 jun. 2018 < https://doi.org/10.1109/ISCAS.2016.7527178>

KASSAMBARA, A. ***Practical guide to cluster analysis in R*****:** Unsupervised machine learning. 1st ed. Sthda, 2017.

KASSAMBARA, A.; MUNDT, F. ***factoextra: extract and visualize the results of multivariate data analyses***. 2020. R package version 1.0.7. Disponível em : https://CRAN.R-project.org/package=factoextra.

KEERIN, P. A Comparative study of missing value imputation methods for education data. In: 29TH INTERNATIONAL CONFERENCE ON COMPUTERS IN EDUCATION, Nov. 22-26, 2021, Taoyuan, Taiwan. ***Proceedings***…p. 109-117. Disponível em : < https://icce2021.apsce.net/> Acesso em : 11 maio 2022.

KLAWONN, F.; HÖPPNER, F. What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier. INTERNATIONAL SYMPOSIUM ON INTELLIGENT DATA ANALYSIS, Aug. 28-30, 2003, Berlin, Gernany. ***Proceedings***…, Springer, p. 254-264. Disponível em : < https://link.springer.com/> Acesso em : 3 out. 2020. < https://doi.org/10.1007/978-3-540-45231-7_24>

KRISHNAPURAM, R.; JOSHI, A.; NASRAOUI, O.; YI, L. Low-complexity fuzzy relational clustering algorithms for web mining. ***IEEE Transactions on Fuzzy Systems***, v.9, n. 4, p. 595-607, 2001. Disponível em : < https://ieeexplore.ieee.org/ > Acesso em : 10 ago. 2019. https://doi.org/10.1109/91.940971.

KUMAR, D.; BEZDEK, J. C. Visual approaches for exploratory data analysis: a survey of the visual assessment of clustering tendency (VAT) family of algorithms. ***IEEE Systems, Man, and Cybernetics Magazine***, v.6, n.2, p. 10-48, 2020. Disponível em : < https://ieeexplore.ieee.org/> Acesso em: 10 out. 2021. https://doi.org/10.1109/msmc.2019.2961163.

LAAKSONEN, S. ***Survey methodology and missing data***. Helsinki: Springer, 2018.

LAMMERS, B. ***NN2: artificial neural networks for anomaly detection***. 2020. R package version 2.3.4. Disponível em : https://CRAN.R-project.org/package=ANN2.

LAWSON, R. G.; JURS, P. C. New index for clustering tendency and its application to chemical problems. ***Journal of Chemical Information and Computer Sciences***, v. 30, n. 1, p. 36-41, 1990. Disponível em : < https://pubs.acs.org/> Acesso em : 10 abr. 2019.< https://doi.org/10.1021/ci00065a010.>

LEYS, C.; KLEIN, O.; DOMINICY, Y.; LEY, C. Detecting multivariate outliers: use a robust variant of the Mahalanobis distance. ***Journal of Experimental Social Psychology***, v. 74, p. 150-156, 2018. Disponível em : < https://www.sciencedirect.com/> Acesso em: 10 out. 2019. < https://doi.org/10.1016/j.jesp.2017.09.011.>

LIMA, I.; PINHEIRO, C. A.; SANTOS, F. A. O. ***Inteligência artificial***. 1. ed. Rio de Janeiro: Elsevier, 2014.

LITTLE, R. J. A.; RUBIN, D. B. ***Statistical analysis with missing data*** 3rd ed. New Delhi: John Wiley & Sons, 2019.

LITTLE, R. J. A.; RUBIN, D. B. The analysis of social science data with missing values. ***Sociological Methods & Research***, v. 18, n. 2–3, p. 292-326, 1989. Disponível em : < https://journals.sagepub.com/> Acesso em : 10 mar. 2019. https://doi.org/10.1177/0049124189018002004.

LITTLE, T. D.; JORGENSEN, T. D.; LANG, K. M.; MOORE, E. W. G. On the joys of missing data. ***Journal of Pediatric Psychology***, v. 39, n.2, p. 151-162, 2014. Disponível em : < https://academic.oup.com/ > Acesso em : 22 set. 2021. https://doi.org/10.1093/jpepsy/jst048.

MADHULATHA, T. S. An overview on clustering methods. ***IOSR Journal of Engineering***, v. 2, n. 4, p. 719-725, 2012. Disponível em : < https://arxiv.org/> https://doi.org/10.9790/3021-0204719725. Acesso em : 10 ago. 2020.

MAECHLER, M.; ROUSSEEUW, P.; STRUYF, A.; HUBERT, M.; HORNIK, K. ***cluster: cluster analysis basics and extensions***. 2022. package version 2.1.4. Disponível em : < https://CRAN.R-project.org/package=cluster>

MALARVIZHI, R.; THANAMANI, A. K-nearest neighbor in missing data imputation. ***International Journal of Engineering Research***, v. 5, n.1, p. 5-7, 2012.

MANDHARE, H. C.; IDATE, S. R. A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques. 2017 INTERNATIONAL CONFERENCE ON INTELLIGENT COMPUTING AND CONTROL SYSTEMS, June 15-16, 2017, Madurai, India ***Proceedings…*** IEEE, p. 931-935. Disponível em **:** <https://ieeexplore.ieee.org/**>** < https://doi.org/10.1109/ICCONS.2017.8250601> Acesso em : 20 set. 2019.

MEITZEN, A.; FALKNER, R. P. History, theory, and technique of statistics. Part First: History of Statistics. ***The Annals of the American Academy of Political and Social Science***, v. 1, n· 2, p. 101-237, Sage, 1891.

MEMÓRIA, J. M. P. Breve história da Estatística. 1. ed. Brasília: ***Embrapa Informação Tecnológica***, 2004.

METSALU, T.; VILO, J. ClustVis: A web tool for visualizing clustering of multivariate data using principal component analysis and heatmap. ***Nucleic Acids Research***, v. 43, p. 566-570, 2015. Disponível em : < https://academic.oup.com/> Acesso em : 11 out. 2021.< https://doi.org/10.1093/nar/gkv468.>

MILLIGAN, G.W.;COOPER, M.C. A study of standardization of variables in cluster analysis. ***Journal of Classification***, v. 5, p. 181–204, 1988. <https://doi.org/10.1007/BF01897163>

MINGOTI, S. A. ***Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada***. Belo Horizonte: UFMG, 2007.

MISZTAL, M. Comparison of selected multiple imputation methods for continuous variables-preliminary simulation study results. ***Acta Universitatis Lodziensis. Folia Oeconomica***, v. 6, p. 73-98, 2018. Disponível em : <http://www.czasopisma.uni.lodz.pl/ > Acesso em : 2 set. 2021. <https://doi.org/10.18778/0208-6018.339.05.>

MIT CRITICAL DATA. ***Secondary analysis of electronic health records***. Cambridge: Springer, 2016.

MOHAMAD, I. B.; USMAN, D. Standardization and its effects on K-means clustering algorithm. ***Research Journal of Applied Sciences, Engineering and Technology***, v. 6, n. 17, p. 3299-3303, 2013. https://doi.org/10.19026/rjaset.6.3638.

MUCHA, H. J.; BARTEL, H. G.; DOLATA, J. Effects of data transformation on cluster analysis of archaeometric data. In: DATA ANALYSIS, MACHINE LEARNING AND APPLICATIONS. Mar. 7-9, 200, Berlin, Heidelberg. **Proceedings…** Springer , 2008. p. 681-688. Disponível em : < https://link.springer.com> Acesso em : 3 set. 2020.

MUNITA, C. S.; PAIVA, R. P.; ALVES, M. A.; DE OLIVEIRA, P. M. S.; MOMOSE, E. F. Provenance study of archaeological ceramic. ***Journal of Trace and Microprobe Techniques***. v. 21, n. 4, p. 697-706, 2003.

NAGPAL, A.; JATAIN, A.; GAUR, D. Review based on data clustering algorithms. In: 2013 IEEE CONFERENCE ON INFORMATION AND COMMUNICATION TECHNOLOGIES**,** Apr. 11-12, 2013, Thuckalay, India. ***Proceedings…,*** IEEE, p. 298-303. Disponível em : < https://ieeexplore.ieee.org/> Acesso em : 9 maio 2019. <https://doi.org/10.1109/CICT.2013.6558109.>

NIKFALAZAR, S.; YEH, C. H.; BEDINGFIELD, S.; KHORSHIDI, H. A. A new iterative fuzzy clustering algorithm for multiple imputation of missing data. In : 2017 IEEE INTERNATIONAL CONFERENCE ON FUZZY SYSTEMS, July 9-12, 2017, Naples, Italy. ***Proceeding…***, IEEE, p. 1-6, 2017. Disponivel em : <https://ieeexplore.ieee.org/> Acesso em : 2 jan. 2019. <https://doi.org/10.1109/FUZZ-IEEE.2017.8015560>

NORAZIAN RAMLI, M. N.; YAHAYA, A. S.; RAMLI, N. A.; YUSOF, N. F. F. M.; ABDULLAH, M. M. A. Roles of imputation methods for filling the missing values: A review. ***Advances in Environmental Biology***, v. 7, n. 12, p. 3861-3870, 2013. Disponível em : < https://go.gale.com/ps/> Acesso em: 10 nov. 2020.

OSBORNE, J. W. Improving your data transformations: applying the Box-Cox transformation. **Practical Assessment, Research, and Evaluation**, v. 15, n. 1, p. 1-9, 2010. Disponível em : < https://scholarworks.umass.edu/ > Acesso em: 3 nov. 2018.

PENNY, K. I. Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance. ***Journal of the Royal Statistical Society: Series C (Applied Statistics)***, v. 45, n. 1, p. 73-81, 1996. < https://doi.org/10.2307/2986224>

R CORE TEAM. ***R: A language and environment for statistical computing*****.** R Foundation for Statistical Computing, 2022. Disponível em : < https://www.R-project.org/>

REYNOLDS, A. P.; RICHARDS, G.; DE LA IGLESIA, B.; RAYWARD-SMITH, V. J. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. ***Journal of Mathematical Modelling and Algorithms***, v. 5, n.4, p. 475-504, 2006. Disponível em : < https://link.springer.com/> Acesso em : 21 maio 2019. https://doi.org/10.1007/s10852-005-9022-1.

ROUSSEEUW, P. J.; DRIESSEN, K. V. A aast Algorithm for the minimum covariance determinant estimator. ***Technometrics***, v. 41, p. 212-223, 1999. Disponível em : <https://www.tandfonline.com/> Acesso em : 25 de maio 2019.

ROHLF, F. J. Adaptive hierarchical clustering schemes. ***Systematic Biology***, v. 19, n. 1, p. 58-82, 1970. Disponível em : < https://academic.oup.com/> Acesso em : 14 de jun. 2019. < https://doi.org/10.1093/sysbio/19.1.58>

RSTUDIO TEAM. ***RStudio: integrated development for R***. *RStudio, Inc., Boston, MA*, 2021. Disponível em : < http://www.rstudio.com/ > Acesso em: 12 de fev. 2018.

RODE, R. A.; CHINCHILLI, V. M. The use of Box-Cox transformations in the development of multivariate tolerance regions with applications to clinical chemistry. ***The American Statistician***, v. 42, n. 1, p. 23-30, 1988. Disponível em : < https://www.tandfonline.com/> Acesso em : 25 fev. 2019. < https://doi.org/10.2307/2685257>

RUBIN, D. B. ***Multiple imputation for nonresponse in surveys***. Wiley Series in Probability and Statistics New Jersey: John Wiley & Sons, 1987

RUSPINI, E. H. Numerical methods for fuzzy clustering. ***Information Sciences***, v. 2, n.3, p. 319–350, 1970. Disponível em : < https://www.sciencedirect.com/> Acesso em : 22 maio 2019.<https://doi.org/10.1016/S0020-0255(70)80056-1.>

SALAH, R.; VINCENT, P.; MULLER, X. Contractive auto-encoders: explicit invariance during feature extraction. In: 28TH INTERNATIONAL CONFERENCE ON MACHINE LEARNING, June 28 – July 2, 2011, Bellevue, United States. **Proceesing…** p. 833-840. Disponível em : <http://www.icml-2011.org/> Acesso em: 21 jan. 2020.

SCHAFER, J. L. Multiple imputation: a primer. ***Statistical Methods in Medical Research***, v.8, n. 1, p. 3-15, 1999. Diponível em : < https://journals.sagepub.com/>
Acesso em : 12 out. 2020.< https://doi.org/10.1177/096228029900800102>

SCHLOERKE, B.; COOK, D.; LARMARANGE, J.; BRIATTE, F.; MARBACH, M.; THOEN, E.; ELBERG, A.; CROWLEY, J. ***GGally: extension to 'ggplot2'*.** 2021, R package version 2.1.2. Disponível em : <https://CRAN.R-project.org/package=GGally>

SHI, H.; WANG, P.; YANG, X.; YU, H. An improved mean imputation clustering algorithm for incomplete data. ***Neural Processing Letters***, p. 1-14, 2020. Disponível em : < https://link.springer.com> Acesso em : 22 jan. 2021. https://doi.org/10.1007/s11063-020-10298-5.

SIGNORELLI, A. et al. ***DescTools: tools for descriptive statistics***. 2022. R package version 0.99.46. Disponível em : < https://cran.r-project.org/package=DescTools>

SILVA, L. A. da; M., S.; BOSCARIOLI, Clodis. ***Introdução à mineração de dados com aplicações em R***. 1. ed. Rio de Janeiro: Elsevier, 2016.

SINHARAY, S.; STERN, H. S.; RUSSELL, D. The use of multiple imputation for the analysis of missing data. ***Psychological Methods***, v. 6, n. 3, p. 317-329, 2001. Disponível em : < https://psycnet.apa.org/ > Acesso em : 18 jun. 2020. https://doi.org/10.1037/1082-989x.6.4.317.

SONG, Q.; SHEPPERD, M. Missing data imputation techniques. **International *Journal of Business Intelligence and Data Mining***, v. 2, n. 3, p. 261–291, 2007. Disponível em : < https://dl.acm.org/> Acesso em : 2 set. 2019. https://doi.org/10.1504/IJBIDM.2007.015485.

TANIOKA, K., YADOHISA, H. Effect of data standardization on the result of k-means clustering. In: 34TH ANNUAL CONFERENCE OF THE GESELLSCHAFT FÜR KLASSIFIKATION E. V., KARLSRUHE. July 21-23, 2010, Berlin, Germany. ***Proceedings…*** p. 59-67, 2012. Disponível em : < https://link.springer.com> Acesso em : 11 maio 2019. <https://doi.org/10.1007/978-3-642-24466-7_7>

VENABLES, W. N.; RIPLEY, B. D. ***Modern applied statistics with S*.** 4th ed. New York: Springer, 2002.

VINCENT, P.; LAROCHELLE, H.; LAJOIE, I.; BENGIO, Y.; MANZAGOL, P. A. Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion. ***Journal of Machine Learning Research***, v. 11, n. 12, p. 3371-3308, 2010. Disponível em : < https://dl.acm.org/journal/> Acesso em : 17 set 2019.

WAN, D.; RAZAVI-FAR, R.; SAIF, M.; MOZAFARI, N. COLI: Collaborative clustering missing data imputation. ***Pattern Recognition Letters***, v. 152, p. 420-427, 2021. Disponível em : < https://www.sciencedirect.com/> Acesso em : 7 maio. 2022. <https://doi.org/10.1016/j.patrec.2021.11.011.>

WANG, J.; ZAMAR, R.; MARAZZI, A.; YOHAI, V.; SALIBIAN-BARRERA, M.; MARONNA, R.; ZIVOT, E.; ROCKE, D.; MARTIN, D.; MAECHLER, M.; KONIS, K. ***robust: port of the S+ "Robust Library"****.* 2022. R package version 0.7-1. Disponível em : https://CRAN.R-project.org/package=robust

WICKHAM, H. ***ggplot2: elegant graphics for data analysis***. New York: Springer-Verlag, 2016.

WIERZCHOŃ, S.; KŁOPOTEK, M. ***Modern algorithms of cluster analysis. studies in big data***. Springer, 2018.

WINKLER, R.; KLAWONN, F.; KRUSE, R. Fuzzy clustering with polynomial fuzzifier function in connection with m-estimators. ***Applied and Computational Mathematics***, v. 10, n. 1, p. 146-163, 2011.

WRIGHT, K.; YILAN, L.; RUTONG, Z. ***clustertend: check the clustering tendency***, 2022. R package version 1.6. Disponível em : https://CRAN.R-project.org/package=clustertend

XU, R.; WUNSCHII, D. Survey of clustering algorithms. ***IEEE Transactions on Neural Networks***, v. 16, n. 3, p. 645–678, 2005. Disponível em : <https://ieeexplore.ieee.org> Acesso em : 5 dez. 2019. https://doi.org/10.1109/TNN.2005.845141.

ZADEH, L. A. Fuzzy sets. ***Information and Control***, v. 8, n.3, p. 338-353, 1965. https://doi.org/10.1016/S0019-9958(65)90241-X.

ZHANG, G.; LIU, Y.; JIN, X. A survey of autoencoder-based recommender systems. ***Frontiers of Computer Science***, v. 14, n. 2, p. 430–450, 2020. Disponível em : < https://link.springer.com/> Acesso em: 12 out. 2021. < https://doi.org/10.1007/s11704-018-8052-6.>

ZHANG, T.; RAMAKRISHNAN, R.; LIVNY, M. BIRCH: A new data clustering algorithm and its applications. ***Data Mining and Knowledge Discovery***, v. 1, n. 2, p. 141-182, 1997. Disponível em: < https://link.springer.com/ > Acesso em: jun. 2019. https://doi.org/10.1023/A:1009783824328.

INSTITUTO DE PESQUISAS ENERGÉTICAS E NUCLEARES
Diretoria de Pesquisa, Desenvolvimento e Ensino
Av. Prof. Lineu Prestes, 2242 – Cidade Universitária CEP: 05508-000
Fone/Fax(0XX11) 3133-8908
SÃO PAULO – São Paulo – Brasil
http://www.ipen.br

O IPEN é uma Autarquia vinculada à Secretaria de Desenvolvimento, associada à Universidade de São Paulo e gerida técnica e administrativamente pela Comissão Nacional de Energia Nuclear, órgão do Ministério da Ciência, Tecnologia, Inovações e Comunicações.