UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

KAORI TOBARA FELIPPE POLIS

COMPARAÇÃO DE MÉTODOS DE ANÁLISE DE CLUSTER PARA AGRUPAR PAÍSES DE ACORDO COM OS DADOS DO RELATÓRIO MUNDIAL DA FELICIDADE

PRESIDENTE PRUDENTE
2023

KAORI TOBARA FELIPPE POLIS

COMPARAÇÃO DE MÉTODOS DE ANÁLISE DE CLUSTER PARA AGRUPAR PAÍSES DE ACORDO COM OS DADOS DO RELATÓRIO MUNDIAL DA FELICIDADE

Relatório Final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp.
Orientadora: Profa. Dra. Miriam Rodrigues Silvestre.

PRESIDENTE PRUDENTE
2023

P768c

Polis, Kaori Tobara Felippe
Comparação de métodos de análise de cluster para agrupar países de acordo com os dados do Relatório Mundial da Felicidade / Kaori Tobara Felippe Polis. -- Presidente Prudente, 2023
58 p.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientadora: Miriam Rodrigues Silvestre

1. Estatística. 2. Análise multivariada. 3. Análise por agrupamento. 4. Felicidade. 5. Países. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

KAORI TOBARA FELIPPE POLIS

COMPARAÇÃO DE MÉTODOS DE ANÁLISE DE CLUSTER PARA AGRUPAR PAÍSES DE ACORDO COM OS DADOS DO RELATÓRIO MUNDIAL DA FELICIDADE

Relatório Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientadora: _____

Profa. Dra. Miriam Rodrigues Silvestre

Departamento de Estatística

_____

Prof. Dr. Guilherme Aparecido Santos Aguilar

Departamento de Estatística

_____

Prof. Dr. Edilson Ferreira Flores

Departamento de Estatística

Presidente Prudente, 4 de dezembro de 2023.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer minha orientadora, Profa. Miriam Rodrigues Silvestre, pela orientação e confiança depositada ao longo de todo o desenvolvimento deste trabalho. Mesmo diante da distância física, sua prontidão e precisão ao oferecer suporte foram inestimáveis.

Agradeço, igualmente, os professores que compuseram minha banca, o Prof. Guilherme Aparecido Santos Aguilar e o Prof. Edilson Ferreira Flores, por toda contribuição para o aprimoramento deste trabalho.

Desejo expressar minha profunda gratidão à minha família, cujo papel foi fundamental para meu progresso até aqui. Sempre me incentivaram, proporcionando apoio e suporte contínuo ao longo desta jornada acadêmica.

Por último, gostaria de agradecer aos meus colegas da faculdade Allan Delfino, Dayane Sedano e Jorge Crestani, que estiveram sempre presentes ao longo da graduação. A nossa colaboração mútua, foi uma fonte constante de auxílio e apoio.

Muito obrigada a todos!

RESUMO

Este trabalho de conclusão de curso tem como objetivo comparar diversas técnicas de agrupamento (ligação simples, média, completa, centroide e Ward) na tarefa de categorizar países com base nos dados do Relatório Mundial da Felicidade de 2019. O principal objetivo foi identificar e avaliar os padrões e semelhanças entre os países, considerando as variáveis disponíveis. Após a aplicação de todas as técnicas e a escolha do número ideal de grupos, elas foram minuciosamente comparadas. A técnica de ligação completa se destacou, obtendo os melhores resultados nas análises do dendrograma, medidas de silhueta e correlação cofenética. Como resultado, quatro grupos distintos foram formados. Os grupos 1 e 2 apresentaram os menores valores nas variáveis consideradas positivas, enquanto os grupos 3 e 4 mostraram os maiores valores nessas mesmas variáveis. É relevante notar que a maioria dos países pertencentes aos grupos 3 e 4 está em regiões mais desenvolvidas do mundo e, de acordo com a variável "Score da Felicidade," são considerados os grupos mais felizes. Esse estudo proporcionou uma análise aprofundada sobre como as técnicas de agrupamento podem ser aplicadas para revelar padrões e semelhanças entre países com base nos dados do Relatório Mundial da Felicidade de 2019, enriquecendo a compreensão dos níveis de felicidade em escala global.

Palavras-chave: análise multivariada; análise de agrupamento hierárquico; Relatório Mundial da Felicidade; países; padrões.

ABSTRACT

This undergraduate thesis aims to compare various clustering techniques (single linkage, average linkage, complete linkage, centroid, and Ward) in the task of categorizing countries based on the data from the 2019 World Happiness Report. The primary objective was to identify and evaluate patterns and similarities among countries, taking into account the available variables. After applying all the techniques and selecting the ideal number of groups, they were thoroughly compared. The complete linkage technique stood out, achieving the best results in dendrogram analysis, silhouette measures, and cophenetic correlation. As a result, four distinct groups were formed. Groups 1 and 2 showed the lowest values in the considered positive variables, while groups 3 and 4 exhibited the highest values in those same variables. It is noteworthy that most countries in groups 3 and 4 are from more developed regions of the world and, according to the "Happiness Score" variable, are considered the happiest groups. This study provided an in-depth analysis of how clustering techniques can be applied to reveal patterns and similarities among countries based on the data from the 2019 World Happiness Report, enriching the understanding of happiness levels on a global scale.

Keywords: multivariate analysis; hierarchical clustering analysis; World Happiness Report; countries; variables; patterns.

SUMÁRIO

1 INTRODUÇÃO 10

2 RELATÓRIO MUNDIAL DA FELICIDADE 12

3 REFERENCIAL TEÓRICO 13

4 METODOLOGIA 14

5 ANÁLISE MULTIVARIADA 15

6 ANÁLISE DE AGRUPAMENTOS 16

7 MEDIDAS DE SIMILARIDADES E DISSIMILARIDADES 19

7.1 DISTÂNCIA EUCLIDIANA 20
7.2 DISTÂNCIA DE *CITY-BLOCK* 21
7.3 DISTÂNCIA DE MINKOWSKY 21
7.4 DISTÂNCIA DE CANBERRA 21

8 ANÁLISE DE AGRUPAMENTOS HIERÁRQUICOS 21

8.1 LIGAÇÃO SIMPLES 22
8.2 LIGAÇÃO COMPLETA 22
8.3 LIGAÇÃO MÉDIA 23
8.4 LIGAÇÃO CENTROIDE 23
8.5 LIGAÇÃO DE WARD 24

9 AVALIAÇÃO DA MELHOR TÉCNICA DE AGRUPAMENTO 26

9.1 GRÁFICO DA SILHUETA 26
9.2 CORRELAÇÃO COFENÉTICA 27

10 MÉTODOS DE ESCOLHA DO NÚMERO DE GRUPOS 28

10.1 DENDROGRAMA 28
10.2 FUNÇÃO NBCLUST 29

11 APRESENTAÇÃO E ANÁLISE DESCRITIVA DOS DADOS 31

12 APLICAÇÃO DOS AGRUPAMENTOS HIERÁRQUICOS 38

12.1 RESULTADO DO AGRUPAMENTO COM LIGAÇÃO WARD 38
12.2 RESULTADO DO AGRUPAMENTO COM LIGAÇÃO SIMPLES 41
12.3 RESULTADO DO AGRUPAMENTO COM LIGAÇÃO MÉDIA 43
12.4 RESULTADO DO AGRUPAMENTO COM LIGAÇÃO COMPLETA 45
12.5 RESULTADO DO AGRUPAMENTO COM LIGAÇÃO CENTRÓIDE 47

13 COMPARAÇÃO DOS AGRUPAMENTOS HIERÁRQUICOS 49

14 CONSIDERAÇÕES FINAIS 51

REFERÊNCIAS..... 57

1 Introdução

Segundo o dicionário, a felicidade é um substantivo abstrato definido como “sensação real de satisfação plena; estado de contentamento, de satisfação.” (Felicidade, 2022). Diversos filósofos também buscaram entender e definir a felicidade, para o Aristóteles se consagra através do equilíbrio e harmonia praticando o bem, de forma similar, Confúcio dizia que era através da harmonia entre as pessoas, já para Epicuro, ocorre através da satisfação dos desejos.

A felicidade, e como ela pode ser maximizada, se tornou assunto de interesse de muitos pesquisadores. É compreensivo que o estado de felicidade seja um objetivo fundamental na vida da maioria das pessoas, portanto, se a fonte e causa da felicidade forem compreendidas, não há dúvidas que toda a sociedade se beneficiará. Desta forma, os líderes políticos poderão conhecer os fatores que exercem uma influência positiva nesta causa e investir neles (Tan, Singchapreecha e Yamaka, 2020).

Como afirma a reportagem *Sustainable Development Solutions Network* (SDSN) *Tracks the World's Happiness*, do site *Gallup Poll* (Sustainable, 2022), as estatísticas mais famosas e usadas para medir a evolução de um país, como PIB, renda familiar, desemprego, etc., são métricas tradicionais que não nos dizem nada sobre a felicidade das pessoas. Têm-se como exemplo alguns países árabes onde o PIB estava aumentando, mas as avaliações dos residentes sobre suas vidas diminuíram antes mesmo dos conflitos.

Países ricos estão diminuindo de forma significativa a pobreza, fome e doenças de sua nação, mas após alcançar estas metas e as necessidades econômicas básicas não serem mais os principais fatores da mudança social, os líderes deveriam se questionar se a felicidade e bem-estar da população foi alcançada. Inclusive, seguindo esta ideia, a primeira ministra da Nova Zelândia, Jacinda Ardern, em 2018, declarou estar planejando incluir o bem-estar dos neozelandeses como uma medida de sucesso econômico de seu país (Helliwel, Lavard e Sachs, 2012).

Para atender a essa carência de informações sobre a medição e avaliação do índice de felicidade nos países, é possível recorrer aos dados fornecidos pelo Relatório Mundial de Felicidade. Essa publicação anual, produzida pela Rede de Soluções para o Desenvolvimento Sustentável das Nações Unidas, realiza uma avaliação do bem-estar subjetivo e da felicidade em diversas nações ao redor do mundo. Utilizando uma variedade de indicadores, o relatório busca mensurar o nível de felicidade presente em cada país.

Objetivo Geral

Este trabalho tem como objetivo geral comparar diferentes métodos de análise de cluster hierárquicos para agrupar países de acordo com os dados do Relatório Mundial da Felicidade de 2019. Desta forma, o intuito é utilizar a análise multivariada de forma exploratória, buscando reconhecer padrões entre os países através das variáveis disponíveis no estudo.

Para atingir este objetivo, após a aplicação dos diferentes métodos de agrupamento, será definido um número ideal de grupos e o melhor modelo testado através de critérios decisivos, e por fim, será analisado individualmente cada cluster a fim de apontar as variáveis mais influentes, chegando então em conclusões úteis a respeito dos grupos.

Objetivos Específicos

Para atingir o objetivo geral, alguns objetivos específicos a serem alcançados são:

1. Buscar conhecer melhor as variáveis disponíveis no estudo através da análise exploratória básica e também verificar a qualidade do conjunto de dados;
2. Aplicar diferentes métodos hierárquicos aglomerativos para classificar as observações (países) semelhantes;
3. Utilizar critérios decisivos na escolha do melhor método hierárquico aglomerativo e do número ideal de clusters;
4. Analisar individualmente cada cluster afim de conhecer os indicadores mais influentes;

5. Analisar os grupos criados e obter as conclusões finais.

Ao completar os objetivos do trabalho será possível identificar os países que se classificam no mesmo nível de felicidade, além de identificar também quais são os principais fatores que contribuem para essa classificação, o que fornecerá uma compreensão geral das condições socioeconômicas relacionadas à felicidade dos países individualmente.

Ao longo deste estudo, o capítulo 2 discorrerá sobre a história do Relatório Mundial da Felicidade. No capítulo 3, será apresentado o Referencial Teórico, que abrange discussões sobre publicações relacionadas ao tema. Nos capítulos 4 a 10, serão detalhadas e explicadas as metodologias aplicadas, incluindo tipos de agrupamentos, técnicas de ligação, medidas de dissimilaridade e métodos para determinar o número ideal de clusters, bem como a escolha da melhor técnica de ligação. Os capítulos 11 a 13 abordarão os resultados gerados neste estudo, abrangendo desde a análise exploratória até as aplicações de agrupamento, encerrando-se com as considerações finais apresentadas no capítulo 14.

2 Relatório Mundial da Felicidade

Com o intuito de compreender melhor sobre o bem-estar e usar os dados para acompanhar a satisfação e qualidade de vida ao redor do mundo, criou-se o Relatório Mundial da Felicidade, publicado pela Rede de Soluções para o Desenvolvimento Sustentável da ONU (SDSN, na sigla em inglês). O primeiro Relatório foi preparado para apoiar as decisões na Assembleia Geral das Nações Unidas, em abril de 2012. No encontro chamado *"*Felicidade e bem-estar: definindo um novo paradigma econômico*"*, os especialistas uniram dados globais disponíveis sobre felicidade com os avanços da ciência. E desde então, anualmente os relatórios unem especialistas de diversas áreas para descreverem como as medições de bem-estar podem ser efetivamente usadas para avaliar o progresso das nações.

O Relatório Mundial de Felicidade busca oferecer uma visão abrangente do estado de felicidade e bem-estar em diferentes países, considerando tanto aspectos subjetivos quanto objetivos. Ele combina dados subjetivos, como as avaliações de felicidade autodeclaradas pelas pessoas, com dados objetivos, como indicadores

socioeconômicos. Essa abordagem visa fornecer uma compreensão mais completa e abrangente da felicidade em cada país participante. Ao analisar tanto a perspectiva pessoal das pessoas quanto os fatores socioeconômicos, o relatório busca fornecer uma visão mais ampla e informada sobre o bem-estar da população em geral.

Os dados são fornecidos pela empresa global de análise e consultoria *Gallup Poll.* Anualmente a pesquisa é feita em uma amostra de 1.000 pessoas por país, como em cada Relatório é usado informações dos três últimos anos, em cada publicação são analisadas 3.000 respostas de perguntas específicas aos indivíduos.

Através das publicações anuais do Relatório Mundial da Felicidade, é possível analisar que a felicidade difere entre as sociedades e ao longo do tempo por razões que se constataram identificáveis e alteráveis pela maneira que as políticas públicas são administradas. Portanto, faz sentido o aumento da felicidade e bem-estar da nação se tornar um importante fator para o desenvolvimento do país assim como o aumento da economia.

3 Referencial Teórico

Diversos materiais foram estudados para auxiliar na formulação deste trabalho, sendo artigos e reportagens com assuntos similares e ideias parecidas. O principal artigo de referência foi o *A Real Data-Driven Clustering Approach for Countries Based on Happiness Score* (Chakraborty e Tsokos, 2021), em que o principal objetivo é, de forma similar a este trabalho, aplicar diferentes algoritmos de agrupamento para classificar observações semelhantes (países) com alto nível de precisão com base nos indicadores no conjunto de dados do Relatório Mundial da Felicidade de 2019, após isso, os autores realizam análises individuais para escolher os indicadores que são mais influentes e finalizam com algoritmos de classificação de aprendizado de máquina muito populares, com a finalidade de validar o agrupamento. Ao realizar todas estas etapas, o artigo traz resultados interessantes sobre a felicidade e também outras condições socioeconômicas relacionadas à felicidade nos países estudados. O que difere este artigo com o trabalho que está sendo proposto, é que no artigo os autores escolheram seguir pelo método

aglomerativo não-hierárquico ao utilizar o algoritmo k-means, resultando na formação de três grupos, já neste trabalho, serão abordados diferentes métodos aglomerativos hierárquicos.

Outra publicação que traz assuntos interessantes que agregam aos objetivos deste trabalho é o texto “Estatística da Felicidade”, publicado na revista Valor-P do Conre-3 em 2007. Diferentemente deste trabalho, as análises são feitas a partir de um questionário aplicado apenas no estado de São Paulo, e as técnicas estatísticas usadas foram a média (para calcular o grau de felicidade), o desvio-padrão (para medir a desigualdade da felicidade), tabelas de frequência simples e cruzadas a fim de apontar os fatores apontados pelos entrevistados como motivo da sua felicidade e por fim, a técnica CHAID com o intuito de entender as ligações entre a felicidade e os fatores apontados como causais. Mesmo que a base de dados e metodologias do estudo “Estatística da Felicidade” sejam diferentes deste trabalho que está sendo escrito, as análises e conclusões são muito relevantes e ajudam na geração de ideias e colaboram com diferentes percepções sobre o estudo da felicidade.

4 Metodologia

A metodologia a ser usada neste trabalho será baseada em técnicas de análise multivariada. A aplicação escolhida foi a análise de agrupamento, que basicamente tem o objetivo de agrupar um conjunto de dados em subgrupos homogêneos, de forma que elementos dentro de um mesmo grupo sejam parecidos, e os elementos em diferentes grupos sejam distintos entre si.

A análise de agrupamento pode ser dividida em outros dois métodos, sendo o método não hierárquico e o hierárquico e este último será explorado neste trabalho. O método hierárquico consiste em agrupamentos formados por nós, em que as observações mais semelhantes vão sendo agrupadas em cada etapa até serem fundidos em um único cluster. Seguindo esta técnica, serão abordados os métodos de ligação simples, completa, média, centroide e Ward.

5 Análise Multivariada

A análise multivariada é um campo da estatística que se dedica ao estudo de conjuntos de dados que envolvem múltiplas variáveis. Seu objetivo principal é compreender as relações existentes entre essas variáveis e como elas se influenciam mutuamente. Para isso, a análise multivariada utiliza diversas técnicas estatísticas e matemáticas, que são aplicadas com o intuito de extrair informações relevantes e embasar decisões fundamentadas nos dados disponíveis.

Assim como Johnson Wichern (2002) apresenta em seu livro *Applied Multivariate Statistical Analysis*, os principais objetivos que a análise multivariada alcança com as suas aplicações são:

- Redução ou simplificação dos dados com a menor perda de informações possível, a fim de facilitar a interpretação e/ou reduzir a dimensionalidade dos dados;

- Investigação da dependência entre variáveis, natureza de suas relações e como se relacionam;

- Previsões feitas a partir de uma ou mais variáveis com base nas observações das outras variáveis, consequentes do estudo das relações entre as variáveis;

- Construção e teste de hipóteses, baseadas em parâmetros de populações multivariadas, que são posteriormente testadas. Esse processo pode ser realizado para validar suposições ou fortalecer crenças pré-existentes.

- Ordenação e agrupamento através da criação de grupos de observações ou variáveis semelhantes por suas características medidas. Com isso, é possível criar regras para classificar objetos em grupos que foram bem definidos;

Será este último objetivo que será abordado neste trabalho.

6 Análise de Agrupamentos

Segundo Luiz Paulo Fávero e Patrícia Belfiore (2017), a análise de agrupamentos é um conjunto de técnicas estatísticas amplamente utilizadas para identificar padrões de comportamento semelhantes entre observações, podendo ser utilizadas para agrupar indivíduos, produtos, municípios, países, entre outros exemplos, com base em variáveis específicas. O objetivo principal é criar grupos ou clusters nos quais haja homogeneidade interna, ou seja, as observações dentro de um mesmo grupo são semelhantes entre si, mas também são consideravelmente diferentes das observações de outros grupos.

Essas técnicas, também conhecidas como análise de conglomerados ou análise de clusters, visam alocar as observações em um número reduzido de grupos que representem o comportamento coletivo com base nas variáveis consideradas. No entanto, é importante ressaltar que a análise de agrupamentos é exploratória e não tem um caráter preditivo, ou seja, caso novas observações sejam incluídas no conjunto de dados, é necessário refazer a modelagem para potencialmente gerar novos agrupamentos. Além disso, a introdução de uma nova variável também pode resultar em uma nova organização das observações nos grupos (Fávero e Belfiore, 2017).

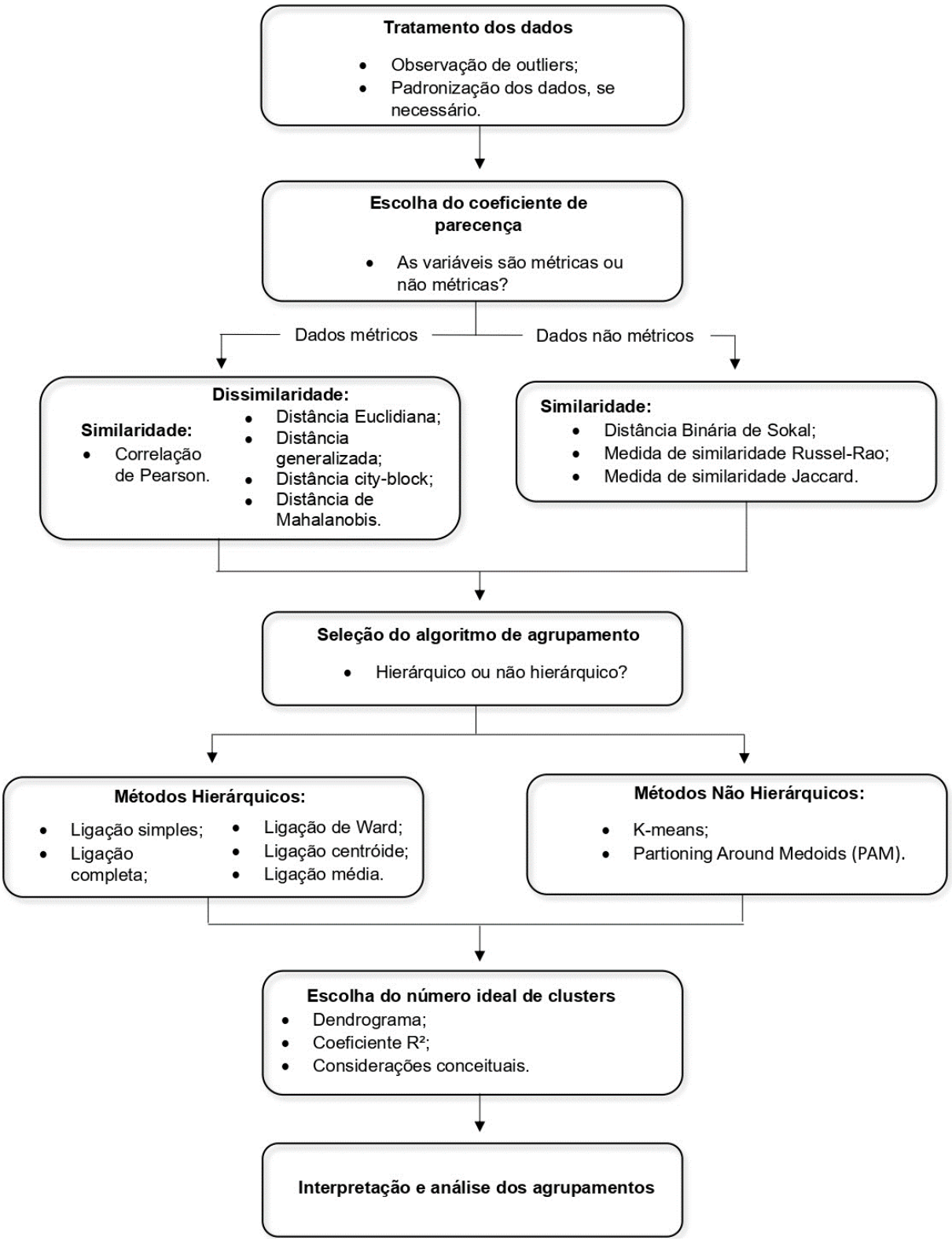
Ao utilizar a análise de cluster, há dois métodos distintos usados para identificar padrões e estruturas em conjuntos de dados: o hierárquico e não hierárquico. No agrupamento não hierárquico, é necessário a definição prévia do número de clusters desejados, em seguida, é realizada a alocação das observações nesses clusters através do algoritmo de agrupamento escolhido, o mais comumente usado é o K-means, que procura dividir o conjunto de dados em K clusters, otimizando a minimização da variância dentro dos clusters e a maximização da diferença entre eles. Outro algoritmo popular é o *Partitioning Around Medoids* (PAM), que busca formar grupos a partir de pontos representativos chamados de medoides e basicamente funciona atribuindo cada objeto do conjunto de dados ao medoide mais próximo, formando assim os grupos.

Por outro lado, o agrupamento hierárquico, que será explorado neste trabalho, cria uma estrutura hierárquica de clusters, na qual as observações são agrupadas

em níveis sucessivos de similaridade ou dissimilaridade, oferecendo possibilidades para que o pesquisador estude, avalie e decida sobre a quantidade de agrupamentos formados.

As etapas da análise de agrupamento podem ser visualizadas no diagrama da Figura 1 abaixo, destacando cada fase e exemplos de técnicas que podem ser aplicadas em cada uma delas. No entanto, é fundamental enfatizar que essas etapas não seguem uma sequência linear e unidirecional. É responsabilidade do pesquisador analisar os resultados de cada fase e, se necessário, voltar a etapas anteriores para garantir uma análise precisa e assertiva. A análise de agrupamento é um processo iterativo que requer uma abordagem flexível e adaptativa para melhor compreender os dados e extrair informações significativas.

Figura 1: Diagrama das etapas da análise de agrupamento.



Fonte: Elaborado pela autora (2023).

7 Medidas de Similaridades e Dissimilaridades

Antes de realizar qualquer agrupamento, é necessário estabelecer um critério para avaliar a distância ou similaridade entre as observações ou elementos em um conjunto de dados, que servirá de base para que cada observação seja alocada em determinado grupo. Essa medida é conhecida como coeficiente de parecença e pode ser classificada em duas categorias:

- Medida de similaridade: quanto maior seu valor, mais os objetos se parecem;
- Medida de dissimilaridade: quanto maior seu valor, mais diferentes são os objetos.

Existem algumas regras para que a medida seja definida como de similaridade ou de dissimilaridade, para apresentar estas características para que uma medida seja considerada um coeficiente de parecença, considere $\boldsymbol{x}_i = (X_{i1}\ X_{i2}\ \ldots\ X_{ip})'$ como sendo o vetor de observações do objeto $i, i = 1, 2, \ldots, n,$ Tal que X_{ij} representa o valor assumido na variável i na posição j e d_{ik} represente a função que avalia a distância entre dois objetos i e $k,$ com relação às variáveis $j, j = 1, 2, \ldots, p$ (Silvestre, 2022).

Dado essas informações, uma função real d_{ik} é chamada função de distância se satisfaz as seguintes condições:

i) $d_{ik} \geq 0, \forall\ X_i\ e\ X_k$;

ii) $d_{ik} = 0$ se e somente se $X_i = X_k$;

iii) $d_{ik} = d_{ki}$;

iv) $d_{il} \leq d_{ik} + d_{kl}$.

Logo, as medidas de distâncias calculadas entre *n* objetos, podem ser representadas pela seguinte matriz simétrica:

$$\boldsymbol{D}_{nxn} = \begin{bmatrix} 0 & d_{12} & \cdots & d_{1n} \\ d_{21} & 0 & \cdots & d_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ d_{n1} & d_{n2} & \cdots & 0 \end{bmatrix}.$$

Do outro lado, uma função real s_{ik} é chamada função de similaridade se satisfaz as seguintes condições:

i) $0 \leq s_{ik} \leq 1, se\ X_i \neq X_k;$

ii) $s_{ii} = 1;$

iii) $s_{ik} = s_{ki}.$

Logo, as medidas de similaridades calculadas entre *n* objetos, podem ser representadas pela seguinte matriz simétrica:

$$\boldsymbol{S}_{nxn} = \begin{bmatrix} 1 & s_{12} & \cdots & s_{1n} \\ s_{21} & 1 & \cdots & s_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ s_{n1} & s_{n2} & \ldots & 1 \end{bmatrix}.$$

As medidas de distância são comumente usadas quando as variáveis são métricas, pois a diferença entre os valores dessas variáveis entre duas observações indica a dissimilaridade entre elas. Já as medidas de semelhança são frequentemente utilizadas para variáveis binárias, onde a frequência de pares convergentes 1-1 ou 0-0 indica a similaridade entre as observações. A medida de correlação de Pearson é uma exceção, pois é calculada a partir de variáveis métricas, mas possui características de similaridade.

Os coeficientes de parecença variam de acordo com o tipo de variáveis em estudo, podendo ser variáveis quantitativas, variáveis qualitativas nominais, variáveis qualitativas ordinais ou variáveis mistas. Nas seções seguintes, serão abordadas as principais medidas utilizadas para variáveis quantitativas, que é o tipo das variáveis da base de dados que será estudada neste trabalho.

7.1 Distância Euclidiana

A distância euclidiana entre os objetos *p* e *q* é estabelecida com base no conceito de distância entre dois pontos no espaço. Essa medida calcula a distância direta entre os objetos considerando as coordenadas das variáveis em análise, proporcionando uma noção de proximidade ou afastamento entre eles. E é dada por:

$$d_{pq} = \sqrt{(\boldsymbol{x_p} - \boldsymbol{x_q})'(\boldsymbol{x_p} - \boldsymbol{x_q})} = \sqrt{\sum_{j=1}^{k}(X_{pj} - X_{qj})^2} \text{ ou } \left[\sum_{j=1}^{k}(X_{pj} - X_{qj})^2\right]^{1/2}$$

7.2 Distância de *City-block*

Diferentemente da distância euclidiana, na distância de *city-block* é feita a soma das diferenças absolutas das variáveis. Este procedimento é o mais simples de calcular, mas pode conduzir a agrupamentos inválidos se as variáveis forem altamente correlacionadas (Hair, Black, Babin, Anderson e Tatham, 2009).

$$d_{pq} = \sum_{j=1}^{k} \left| X_{jp} - X_{jq} \right| .$$

7.3 Distância de Minkowsky

A distância de Minkowsky possui uma expressão de medida de dissimilaridade mais geral a partir da qual outras derivam. Uma vantagem desta técnica é que ela ser menos afetada pela presença de outliers na amostra do que a distância euclidiana. E é dada por:

$$d_{pq} = \left[\sum_{j=1}^{k} \left(\left| X_{jp} - X_{jq} \right| \right)^{m} \right]^{1/m} .$$

7.4 Distância de Canberra

A distância de Canberra é um exemplo utilizado quando as variáveis apresentam apenas valores positivos. Sua expressão é dada por:

$$d_{pq} = \sum_{j=1}^{k} \frac{\left| X_{jp} - X_{jq} \right|}{\left(X_{jp} + X_{jq} \right)} .$$

8 Análise de Agrupamentos Hierárquicos

Um dos principais objetivos deste trabalho é comparar diferentes métodos de análise de agrupamentos hierárquico. Conforme afirmam Johnson e Wichern (2002), estas técnicas envolvem uma série de fusões ou divisões sucessivas. Os métodos aglomerativos começam com cada objeto formando seu próprio cluster, resultando em tantos clusters quanto objetos iniciais. Os clusters mais similares, que serão

definidos a partir do coeficiente de parecença, são então mesclados com base em suas semelhanças, até que todos os subgrupos sejam fundidos em um único cluster.

Por outro lado, os métodos divisivos iniciam com um único grupo contendo todos os objetos, que é dividido em subgrupos de forma que os objetos de um subgrupo sejam "distintos" dos objetos dos outros subgrupos. Esse processo de divisão continua até que cada objeto se torne um grupo individual, resultando em tantos grupos quanto objetos.

Neste trabalho serão exploradas as diferentes técnicas de agrupamento aglomerativas, e os métodos de encadeamento em esquemas hierárquicos aglomerativos que serão estudados são: Ligação Simples, Ligação Média, Ligação Completa, Centroide e Ward.

8.1 Ligação Simples

De acordo com Joseph Hair, et al. (2009), o método de ligação simples, também conhecido como método do vizinho mais próximo, determina a similaridade entre agrupamentos como a menor distância entre qualquer objeto de um agrupamento e qualquer objeto do outro. Esse método permite utilizar a matriz original de distância entre observações sem a necessidade de calcular novas medidas. Basta encontrar todas as distâncias entre observações nos dois agrupamentos e escolher a menor como medida de similaridade entre os agrupamentos. Tal abordagem é provavelmente o algoritmo aglomerativo mais versátil, pois ele pode definir uma vasta gama de padrões de aglomeração.

Ao considerar dois grupos de objetos como sendo G_1 e G_2, a menor distância entre um elemento dos dois grupos é dada por:

$$d[G_1, G_2] = \min_{i \in G_1} \min_{k \in G_2} d_{ik} .$$

8.2 Ligação Completa

O método de ligação completa, também conhecido como o método do vizinho mais distante, é semelhante ao método de ligação simples, mas utiliza a distância máxima entre observações em cada agrupamento como medida de similaridade.

Dessa forma, ele favorece as maiores distâncias entre as observações ou grupos para formar novos agrupamentos.

Esse método é denominado "ligação completa" porque todos os objetos dentro de um agrupamento estão conectados uns aos outros a uma distância máxima. A similaridade entre agrupamentos é determinada pela menor esfera (diâmetro mínimo) capaz de incluir todas as observações de ambos os agrupamentos (Hair, Black, Babin, Anderson e Tatham, 2009).

A maior distância entre um elemento do grupo G_1 e um elemento de G_2 é dada por:

$$d[G_1, G_2] = \max_{i \in G_1} \max_{k \in G_2} d_{ik} .$$

8.3 Ligação Média

O método de ligação média difere dos métodos de ligação simples e completa, pois a similaridade entre dois agrupamentos é calculada como a média da similaridade entre todos os indivíduos de um agrupamento e todos os indivíduos do outro agrupamento. Ao contrário dos métodos de ligação simples e completa, o método de ligação média não depende de valores extremos (pares mais próximos ou mais distantes) para determinar a similaridade. Em vez disso, a similaridade é calculada com base em todos os elementos dos agrupamentos, o que torna o método menos influenciado por observações atípicas.

A média da distância entre os elementos de G_1 e um elemento de G_2 é dada por:

$$d[G_1, G_2] = \sum_{i \in G_1} \sum_{k \in G_2} \frac{d_{ik}}{g_1 g_2}.$$

8.4 Ligação Centroide

No método do centroide, a similaridade entre dois agrupamentos é determinada pela distância entre seus centroides. O centroide representa o valor médio das observações em relação às variáveis consideradas na análise de agrupamento. Nesse método, sempre que indivíduos são agrupados, um novo

centroide é calculado. Os centroides se deslocam à medida que ocorrem fusões de agrupamentos. Em outras palavras, o centroide de um agrupamento é atualizado sempre que um novo indivíduo ou grupo de indivíduos é adicionado a um agrupamento existente (Hair, Black, Babin, Anderson e Tatham, 2009).

O centroide do grupo G_1 é definido como a média entre seus elementos representados por o_i:

$$\overline{C_1} = \frac{1}{g_1} \sum_{i \in G_1} o_i$$

O centroide do grupo G_2 é definido como a média entre seus elementos representados por o_k:

$$\overline{C_2} = \frac{1}{g_2} \sum_{k \in G_2} o_k$$

Após calcular os centroides, a distância entre os grupos é obtida através do cálculo das distâncias entre os centroides:

$$d[G_1, G_2] = \sqrt{(\overline{C_1} - \overline{C_2})'(\overline{C_1} - \overline{C_2})}$$

8.5 Ligação de Ward

Conforme afirma Joseph Hair, et al. (2009), o método de Ward se diferencia das técnicas anteriores ao considerar a similaridade entre dois agrupamentos como a soma dos quadrados das diferenças dentro dos agrupamentos, levando em conta todas as variáveis. O método de Ward, também conhecido como "mínima variância", é baseado na mudança da variação entre os grupos e dentro dos grupos que estão sendo formados em cada etapa do agrupamento. Ele busca minimizar a variação total dentro dos agrupamentos, promovendo a formação de agrupamentos mais compactos e homogêneos.

O método de Ward é definido da seguinte forma:

No início do algoritmo, cada elemento é tratado como um único grupo. A cada passo do algoritmo, é calculada a soma dos quadrados dentro de cada grupo. Essa soma é obtida através do cálculo da distância euclidiana ao quadrado entre cada elemento do grupo e o vetor médio desse grupo, que é dado por:

$$SS_i = \sum_{j=1}^{n_i} (X_{ij} - \overline{\boldsymbol{X}_i})'(X_{ij} - \overline{\boldsymbol{X}_i}) \, .$$

Tal que:

- n_i: Representa o número de elementos no grupo G_i quando se está no passo *k* do processo de agrupamento;
- X_{ij}: Corresponde ao vetor de observações do *j*-ésimo elemento que pertence ao *i*-ésimo grupo;
- $\overline{\boldsymbol{X}_i}$: É o centroide do grupo G_i;
- SS_i: Indica a soma de quadrados correspondente ao grupo G_i.

No passo *k*, a soma de quadrados total dentro dos grupos é definida como:

$$SSR = \sum_{i=1}^{g_k} SS_i \, .$$

Sendo que g_k representa o número de grupos existentes no momento do passo *k*.

Por fim, a distância dos dois grupos é definida por:

$$d[G_p, G_s] = \frac{n_p n_s}{n_p + n_s} (\overline{\boldsymbol{X}}_p - \overline{\boldsymbol{X}}_s)' (\overline{\boldsymbol{X}}_p - \overline{\boldsymbol{X}}_s) \, .$$

A distância utilizada é a soma dos quadrados entre os grupos G_p e G_s. A cada iteração do algoritmo, os dois grupos que resultam na menor distância são fundidos. Essa medida representa a diferença entre o valor de SSR após a fusão dos grupos G_p e G_s em um único grupo em relação ao valor anterior. Dessa forma, em cada etapa do processo de agrupamento, o método de Ward combina os dois grupos que levam ao menor valor de SSR (Silvestre, 2022).

9 Avaliação da Melhor Técnica de Agrupamento

9.1 Gráfico da Silhueta

A avaliação da qualidade dos agrupamentos pode ser realizada utilizando o Gráfico da Silhueta, um método proposto por Rousseeuw em 1987 apud Vale (2005).

Segundo Vale (2005), ao considerar i um objeto que pertence ao grupo A, a dissimilaridade média de i em relação aos restantes dos objetos do grupo A é dada por:

$$a(i) = \frac{1}{|A| - 1} \sum_{j \in A, j \neq i} d(i,j) \, .$$

Onde:

- |A|: Representa o número total de objetos existentes no grupo A;
- d(i,j): Indica a dissimilaridades entre os objetos i e j.

Ao considerarmos um grupo adicional C, é possível calcular a dissimilaridade média do objeto i em relação a todos os objetos do grupo A da seguinte maneira:

$$d(i, C) = \frac{1}{|C|} \sum_{j \in C} d(i,j) \, .$$

Tal que:

- |C|: Representa o número total de objetos existentes no grupo C;
- d(i, j): Indica a dissimilaridade entre os objetos i e j.

Desta forma, a menor distância de dissimilaridade entre o objeto i e um grupo A será dada por:

$$b(i) = \min d(i, C) \, .$$

Suponha que o grupo C seja o grupo com a menor distância definida como $b(i)$. Nesse caso, consideramos o grupo C como o vizinho do objeto i e o segundo

melhor grupo para esse objeto. Isso significa que, entre todos os grupos, o grupo C é o mais próximo do objeto i em termos de dissimilaridade, após o grupo principal ao qual o objeto i pertence. Desta forma, o valor da silhueta do objeto i é dada por:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}.$$

A interpretação de s(i) é dada por:

- s(i) ~ 1: Significa que o objeto *i* está bem classificado no grupo A;
- s(i) ~ 0: Significa que o objeto *i* está entre os grupos A e C;
- s(i) ~ -1: Significa que o objeto está mal classificado no grupo A e que na verdade está mais próximo do grupo C do que do A.

A silhueta do grupo A é calculada a partir das silhuetas de todos os objetos pertencentes a esse grupo, ordenados em ordem decrescente. Quanto mais próximo o valor da silhueta estiver de 1, melhor será a qualidade do agrupamento para o grupo A, indicando que os objetos estão bem agrupados e distantes dos objetos de outros grupos (SILVESTRE, 2022).

9.2 Correlação Cofenética

Em sua tese Albuquerque (2005) afirma que uma maneira de determinar qual é o melhor método de agrupamento é através da correlação cofenética. O seu objetivo é medir o grau de ajuste entre a matriz de dissimilaridade original e a matriz cofenética, que é obtida após aplicar a técnica de agrupamento desejado na matriz de distâncias entre todos os pares de objetos.

A correlação cofenética calculada entre a matriz original D e a matriz cofenética C é dada por:

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

Onde:

- c_{ij}: Indica o valor de dissimilaridade entre os indivíduos *i* e *j*, obtidos a partir da matriz cofenética;
- d_{ij}: Indica o valor de dissimilaridade entre os indivíduos *i* e *j*, obtidos a partir da matriz de dissimilaridade

E as medidas $\bar{c}$ e $\bar{d}$ são calculadas da seguinte forma:

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

$$\bar{d} = \frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=j+1}^{n} d_{ij} \, .$$

Valores mais próximos de 1 indicam que o dendrograma preserva de forma satisfatória a estrutura original dos dados de entrada. A correlação cofenética é uma medida que pode ser utilizada para comparar os resultados de diferentes métodos de agrupamento e auxiliar na escolha dos mais adequados, aqueles que conseguem preservar melhor a estrutura original dos dados após o agrupamento. Com base na correlação cofenética, é possível identificar um ou mais métodos que apresentem uma boa concordância com a estrutura original dos dados, tornando-os opções mais promissoras para análise (Albuquerque, 2005).

10 Métodos de Escolha do Número de Grupos

10.1 Dendrograma

Dendrograma oferece uma representação visual do processo de agrupamento, mostrando a estrutura em forma de árvore que representa cada etapa desse processo. Normalmente, o gráfico é ajustado de forma que distâncias mais curtas entre as combinações indiquem maior semelhança (Hair, Black, Babin, Anderson e Tatham, 2009).

Os nós no dendrograma nos grupos ou objetos (grupos formados por um único objeto) vinculados a eles. Ao cortar o dendrograma em um nível de distância específico, é possível obter uma classificação dos grupos que existem nesse nível,

bem como dos indivíduos que compõem esses grupos. A determinação dos números de grupo dos indivíduos é alcançada ao cortar o dendrograma no nível de distância desejado, resultando na formação de grupos a partir de cada componente conectada (Albuquerque, 2005).

10.2 Função NbClust

Com o objetivo de automatizar a seleção do número adequado de clusters em uma análise de clusterização, foi desenvolvida a função NbClust, do pacote NbCust, que pode ser utilizada no software R, sendo uma ferramenta valiosa amplamente utilizada na comunidade de análise de dados.

Essa função funciona por meio da avaliação de cerca de 30 métricas de validação, aplicadas a diferentes números de clusters. Ela realiza uma análise sistemática das variações nas métricas à medida que o número de clusters aumenta, identificando o ponto em que as métricas atingem valores ótimos. Portanto, sua utilização é de grande relevância em estudos como este, onde a identificação do número ideal de clusters desempenha um papel crucial na análise.

Conforme indica no artigo de referência na documentação da função, *NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set* (Charrad, Ghazzali e Niknafs, 2014), essas são todas as 30 métricas disponíveis na função, sendo que são calculadas apenas as que se adequam ao tipo de agrupamento que está sendo realizado:

Tabela 1: Visão geral dos índices implementados no pacote NbClust.

Nome do índice no NcClust	Número ideal de clusters
1. "ch" (Calinski e Harabasz 1974	Valor máximo do índice.
2. "duda" (Duda e Hart 1973)	Menor número de clusters, tal que índice > valor crítico.
3. "pseudot2" (Duda e Hart 1973)	Menor número de clusters, tal que índice < valor crítico.
4. "cindex" (Hubert e Levin 1976)	Valor mínimo do índice.
5. "gamma" (Baker e Hubert 1975)	Valor máximo do índice.
6. “beale” (Beale 1969)	Número de clusters tal que valor crítico >= alfa.

7. "ccc" (Sarle 1983)	Valor máximo do índice.
8. "ptbiserial" (Milligan 1980, 1981)	Valor máximo do índice.
9. "gplus" (Rohlf 1974; Milligan 1981)	Valor mínimo do índice.
10. "db" (Davies e Bouldin 1979)	Valor mínimo do índice.
11. "frey" (Frey and Van Groenewoud 1972)	Nível do cluster antes do valor do índice <1,00.
12. "hartigan" (Hartigan 1975)	Diferença máxima entre os níveis de hierarquia do índice.
13. "tau" (Rohlf 1974; Milligan 1981)	Valor máximo do índice.
14. "ratkowsky" (Ratkowsky e Lance 1978)	Valor máximo do índice.
15. "scott" (Scott e Symons 1971)	Diferença máxima entre os níveis de hierarquia do índice.
16. "marriot" (Marriot 1971)	Máx. valor das segundas diferenças entre os níveis do índice.
17. "ball" (Ball e Hall 1965)	Diferença máxima entre os níveis de hierarquia do índice.
18. "trcovw" (Milligan e Cooper 1985)	Diferença máxima entre os níveis de hierarquia do índice.
19. "tracew" (Milligan e Cooper 1985)	Máx. valor das segundas diferenças entre os níveis.
20. "friedman" (Friedman e Rubin 1967)	Diferença máxima entre os níveis de hierarquia do índice.
21. "mcclain" (McClain e Rao 1975)	Valor mínimo do índice.
22. "rubin" (Friedman e Rubin 1967)	Valor mínimo das segundas diferenças entre níveis.
23. "kl" (Krzanowski e Lai 1988)	Valor máximo do índice.
24. "silhouette" (Rousseeuw 1987)	Valor máximo do índice.
25. "gap" (Tibshirani et al. 2001)	Menor número de clusters tal que valor crítico >= 0.
26. "dindex" (Lebart et al. 2000)	Método gráfico.
27. "dunn" (Dunn 1974)	Valor máximo do índice.
28. "hubert" (Hubert e Arabie 1985)	Método gráfico.
29. "sdindex" (Halkidi et al. 2000)	Valor mínimo do índice.
30. "sdbw" (Halkidi e Vazirgiannis 2001)	Valor mínimo do índice.

Fonte: Tabela traduzida do artigo NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set (Charrad, Ghazzali e Nikafs, 2014).

11 Apresentação e Análise Descritiva dos Dados

Os dados foram obtidos do site *World Happiness Report*, onde são publicados anualmente os relatórios e fornecidos alguns dados que dão suporte aos estudos das publicações.

Os dados foram coletados de um total de 156 países. Os indivíduos respondem através de questionários aplicados pelo *Gallup Poll* a perguntas específicas, e como resultado são produzidas as variáveis do estudo. Assim como explica o artigo *A Real Data-Driven Analytical Model to Predict Happiness* (Chakraborty e Tsokos, 2021) as variáveis calculadas que serão utilizadas neste trabalho são:

- **Log PIB Per Capita:** Produto Interno Bruto per capita (em escala logarítmica) em Paridade do Poder de Compra (PPC), ou seja, é o resultado do cálculo do Produto Interno Bruto para refletir a verdadeira capacidade econômica de um país, eliminando distorções causadas por taxas de câmbio, custos de vida e rendimentos variados;
- **Suporte Social:** É definido através da média nacional das respostas binárias (0 ou 1) à pergunta: "Se você estava com problemas, você tem parentes ou amigos com quem pode contar para ajudá-lo sempre que você precisar deles, ou não?";
- **Expectativa de Vida:** As expectativas de vida saudável ao nascer baseiam-se nos dados fornecidos pelo Observatório Global de Saúde da Organização Mundial de Saúde (OMS), em anos;
- **Liberdade:** É a média nacional das respostas binárias para a pergunta GWP: "Você está satisfeito ou insatisfeito com sua liberdade de escolher o que fazer com sua vida?"
- **Generosidade:** Obtido através do resíduo da média nacional regressiva de resposta à pergunta: "Você doou dinheiro para uma instituição de caridade no último mês?" sobre o PIB per capita;
- **Percepção de Corrupção:** Resultado da média nacional das respostas 0

ou 1 das duas questões: "A corrupção está disseminada em todo o governo ou não?" e "A corrupção está disseminada nas empresas ou não?";

- **Afeto Positivo:** Definido como a média de três medidas de afeto positivo, isto é, felicidade, riso e prazer na pesquisa *Gallup World*;
- **Afeto Negativo:** De forma contrária ao anterior, esta medida é definida como a média das três medidas de afeto negativo, isto é, preocupação, tristeza e raiva, respectivamente;
- **Confiança no Governo:** Variável criada através das respostas a pergunta: Quanta confiança alguém tem no governo quando se trata de lidar com problemas internacionais/problemas domésticos – muita, bastante, pouca ou nenhuma?
- **Qualidade Democrática:** Média nacional das duas primeiras dimensões do Índice de Governança Mundial, ou seja, voz e responsabilidade e estabilidade política e ausência de violência/terrorismo;
- **Qualidade da Entrega:** Média nacional das duas últimas dimensões do Índice de Governança Mundial, ou seja, eficácia Governamental, Qualidade Regulatória, Estado de Direito e Controle da Corrupção.

Para a realização deste trabalho, será utilizado o software R, a fim de realizar as análises, construir gráficos, aplicar os métodos de agrupamento, calcular as medidas que indicarão o número ideal de grupos e a melhor técnica e verificar quais indicadores são mais significativos para o estudo. Ao concluir todos os passos, será possível obter as conclusões finais e atingir o objetivo do trabalho.

Para a aplicação das técnicas de análise de agrupamento, é imprescindível que cada elemento a ser agrupado possua apenas uma observação correspondente na base de dados, além de ser necessário que não haja valores faltantes, exigindo tratamento prévio. Levando essas condições em consideração, os valores ausentes foram removidos. Ao analisar os dados do presente estudo, os quais se referem ao Relatório Mundial da Felicidade de 2019, foi selecionado o ano com maior quantidade de valores preenchidos, que no caso é o ano de 2017. Portanto, os

dados utilizados neste trabalho serão do ano de 2017, correspondente à publicação do Relatório feita em 2019. Como resultado, um total de 129 países serão analisados e agrupados com base nessas informações.

Ao explorar as estatísticas descritivas das variáveis, têm-se como resultado o resumo descritivo apresentado na Tabela 1:

Tabela 2: Estatísticas Descritivas das Variáveis.

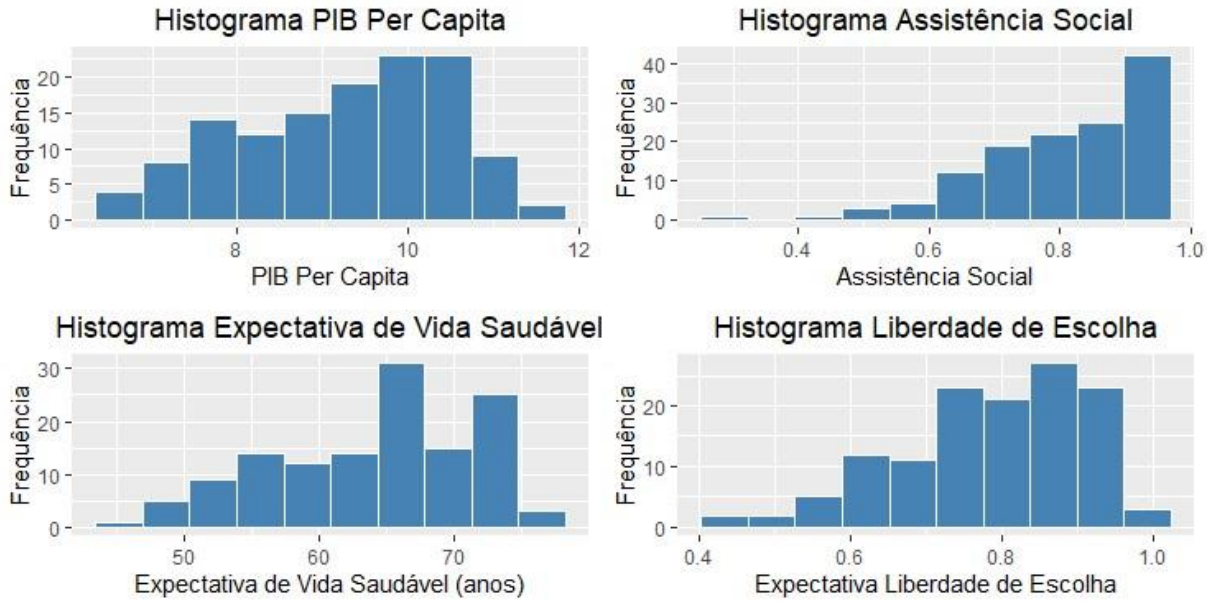
Variáveis	Mínimo	Média	Mediana	Máximo	Variância	Desvio Padrão
PIB Per Capita	6,49	9,23	9,41	11,45	1,49	1,22
Assistência Social	0,31	0,8	0,83	0,96	0,01	0,12
Expectativa de Vida Saudável	45,2	64,17	66,1	76,5	54,01	7,34
Liberdade de Escolha	0,42	0,78	0,81	0,98	0,01	0,12
Generosidade	-0,29	0	-0,02	0,62	0,02	0,15
Percepções de Corrupção	0,16	0,72	0,78	0,95	0,03	0,18
Afeto Positivo	0,42	0,71	0,73	0,9	0,01	0,1
Afeto Negativo	0,14	0,28	0,27	0,59	0,007	0,08
Confiança no Governo	0,11	0,5	0,49	0,98	0,04	0,2
Qualidade Democrática	-1,88	-0,01	-0,09	1,57	0,7	0,83
Qualidade de Entrega	-1,74	0,04	-0,16	2,06	0,96	0,98

Fonte: Elaborado pela autora (2023).

Através dessas estatísticas descritivas, é possível obter informações importantes sobre a distribuição e a variabilidade dos dados, o que auxilia na compreensão do comportamento geral dos dados. É evidente que a padronização dos dados será necessária, dado que as variáveis estão em diferentes escalas. Além disso, nota-se que a maioria das variáveis apresenta baixa variabilidade, ou seja, os valores entre os países nessas variáveis são próximos, com exceção daquela que indica a Expectativa de Vida Saudável. Além disso, é importante relembrar que a variável PIB Per Capita está sofrendo o ajuste da Paridade do Poder de Compra e está em escala logarítmica, por conta disso ela possui os resultados apresentados acima.

Além dessas informações, também é possível analisar a distribuição de cada variável numérica e contínua através dos seus histogramas, dispostos da Figura 2.

Figura 2: Gráficos de Histograma das Variáveis: PIB Per Capita, Assistência Social. Expectativa de Vida Saudável e Liberdade de Escolha.



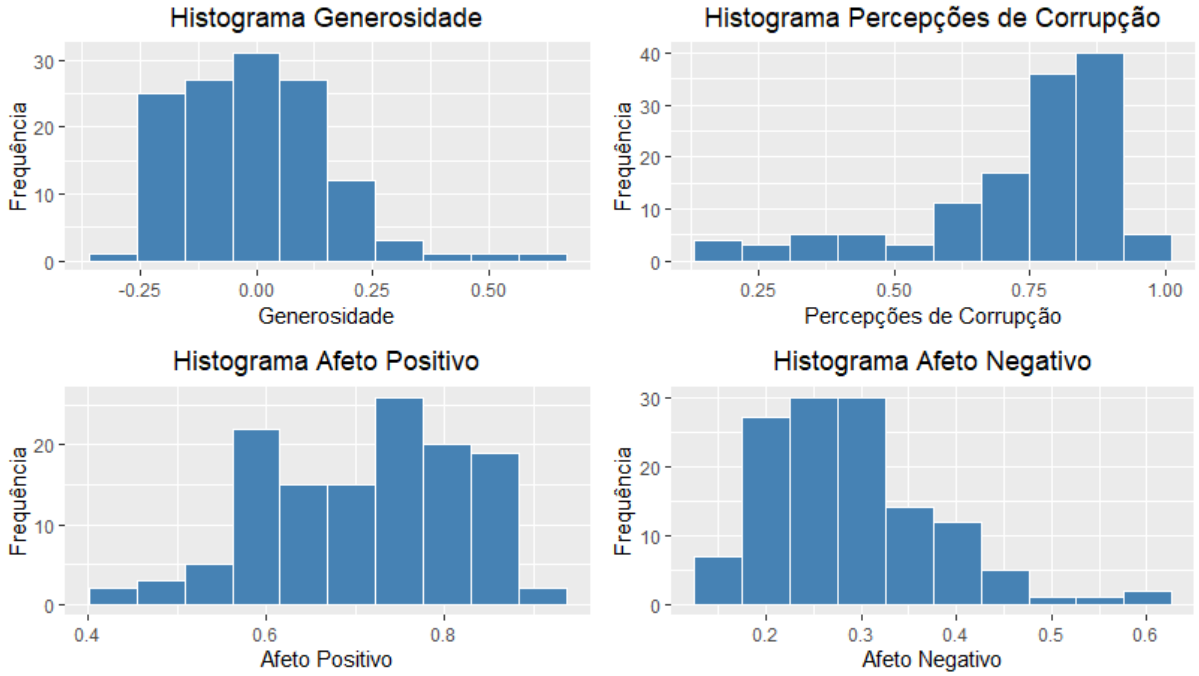
Fonte: Elaborado pela autora (2023).

Ao observar os histogramas das variáveis PIB Per Capita e Expectativa de Vida Saudável, em anos, é evidente a falta de simetria, mas apresentando uma distribuição assimétrica à direita. Em vez de exibir uma distribuição uniforme de valores em torno de um ponto central, os dados parecem estar alternando-se de maneira desordenada.

Já ao observar as variáveis Assistência Social e Liberdade de Escolha, seus respectivos histogramas apresentam assimetria e pouca variabilidade dos dados. No parâmetro Assistência Social, também é possível notar que a distribuição dos valores é caracterizada por um formato crescente, indicando uma tendência clara de aumento nos valores à medida que se move para a direita do eixo horizontal, logo existe uma maior quantidade de países com valores mais altos.

Seguindo com a análise gráfica, na Figura 3 apresenta a sequência dos histogramas de mais quatro variáveis.

Figura 3: Gráficos de Histograma das Variáveis: Generosidade, Percepções de Corrupção, Afeto Positivo e Afeto Negativo.



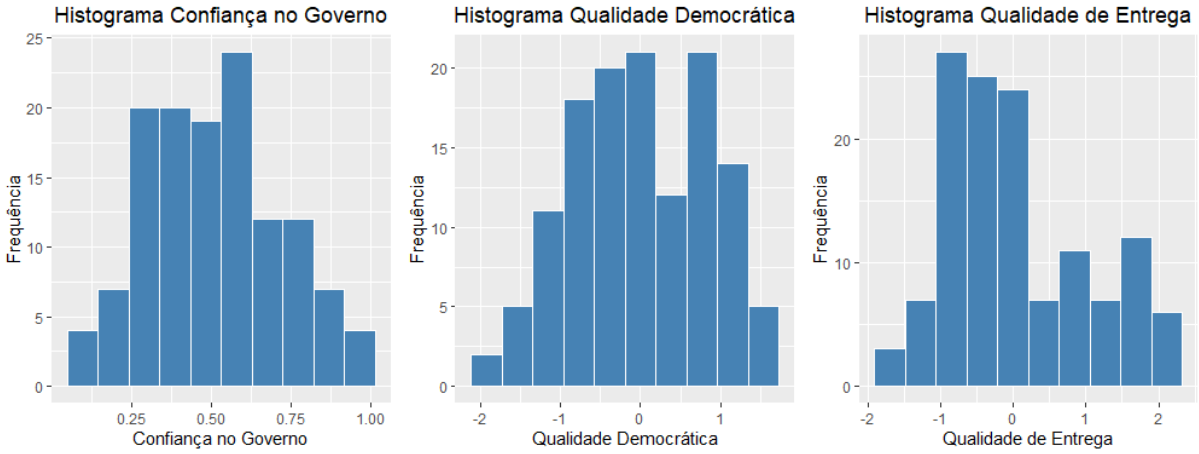
Fonte: Elaborado pela autora (2023).

A interpretação dos histogramas na Figura 3 para as variáveis Generosidade e Afeto Positivo é semelhante à dos dois primeiros histogramas mencionados anteriormente. Ao analisar as Percepções de Corrupção e o Afeto Negativo, observamos uma tendência de crescimento, porém em direções opostas. Conclui-se, assim, que há uma maior quantidade de países com valores menores de Afeto Negativo e valores maiores em Percepções de Corrupção. Além disso, ambas as variáveis apresentam pouca variabilidade.

A observação de poucos valores mínimos e poucos valores máximos dos gráficos sugere uma concentração dos dados em faixas específicas, com uma escassez de pontos extremos.

Em seguida, serão apresentados na Figura 4 os histogramas das variáveis Confiança no Governo, Qualidade Democrática e Qualidade de Entrega.

Figura 4: Gráficos de Histograma das Variáveis: Confiança no Governo, Qualidade Democrática e Qualidade de Entrega.

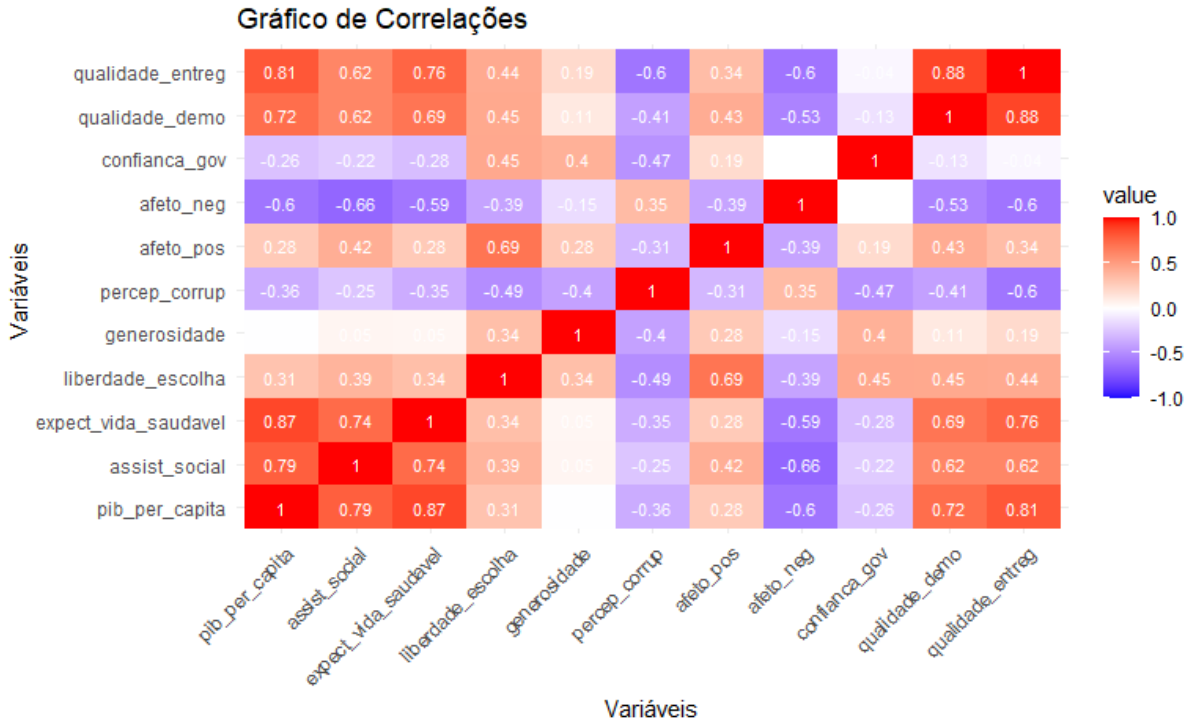


Fonte: Elaborado pela autora (2023).

A análise dos gráficos na Figura 4 assemelha-se aos previamente mencionados, uma vez que não apresentam simetria e os dados alternam-se de maneira não padronizada. Há, assim, indícios de que a maioria dos países possui valores próximos, corroborando com as conclusões obtidas na análise das estatísticas descritivas.

Em seguida, foi feito o gráfico de correlações entre todas as variáveis apresentado na Figura 5.

Figura 5: Gráfico de Correlação entre as Variáveis.



Fonte: Elaborado pela autora (2023).

Ao examinar o gráfico de correlações entre as variáveis, ilustrado na Figura 5, é perceptível a existência de relações fortes entre algumas variáveis. Por exemplo, a variável "Qualidade de Entrega" demonstra uma alta correlação com "PIB Per Capita", "Expectativa de Vida Saudável" e "Qualidade Democrática". Além disso, a variável "Qualidade Democrática" também apresenta uma correlação significativa com "PIB Per Capita". Nota-se, ainda, que as variáveis "Assistência Social", "PIB Per Capita" e "Expectativa de Vida Saudável" também possuem uma alta correlação entre si. Dado que todas essas variáveis mencionadas apresentam altas correlações positivas, isso indica uma tendência de que, quando uma variável aumenta, a outra também tende a aumentar, e vice-versa.

Esses altos valores podem causar um impacto na análise de agrupamento hierárquico. É importante ressaltar que a existência de correlações fortes entre variáveis não inviabiliza a aplicação dessa técnica, porém é necessário lidar adequadamente com esse aspecto para evitar viés e interpretações equivocadas.

Uma abordagem apropriada para lidar com variáveis altamente correlacionadas envolve a cuidadosa seleção das variáveis a serem incluídas na

análise, a normalização dos dados e a utilização de medidas de dissimilaridade adequadas. Neste trabalho, optou-se por adotar todos esses procedimentos, com exceção da seleção de variáveis, uma vez que a quantidade de variáveis neste estudo já é relativamente baixa.

12 Aplicação dos Agrupamentos Hierárquicos

Após a introdução das técnicas a serem avaliadas e a análise exploratória dos dados, os resultados serão apresentados, juntamente com comparações e as conclusões finais referentes à aplicação das técnicas de agrupamento hierárquico.

Uma vez concluídos todos os procedimentos de pré-processamento dos dados, foi gerada uma matriz de distâncias euclidianas. A escolha dessa medida de dissimilaridade baseou-se na simplicidade e adequação aos dados da base, que consistem em valores numéricos contínuos e não apresentam valores discrepantes (outliers).

Para cada par de países, é feito o cálculo da distância euclidiana entre eles com base nas coordenadas das variáveis. A distância euclidiana é calculada usando a fórmula da distância euclidiana, que é a raiz quadrada da soma dos quadrados das diferenças entre os valores das variáveis em cada dimensão. O resultado desses cálculos forma uma matriz de distâncias. Esta matriz é simétrica e quadrada, onde cada elemento (i, j) da matriz representa a distância euclidiana entre o país i e o país j com base nas variáveis escolhidas. Quanto menor a distância entre dois países na matriz, mais próximos eles estão em relação às variáveis selecionadas, quanto maior a distância, mais distantes eles estão.

Com a matriz de distâncias calculada, agora serão aplicadas e analisadas as técnicas de agrupamentos. Conforme explicado no capítulo 10, serão utilizados o dendrograma e a função NbClust para definir o número ideal de grupos.

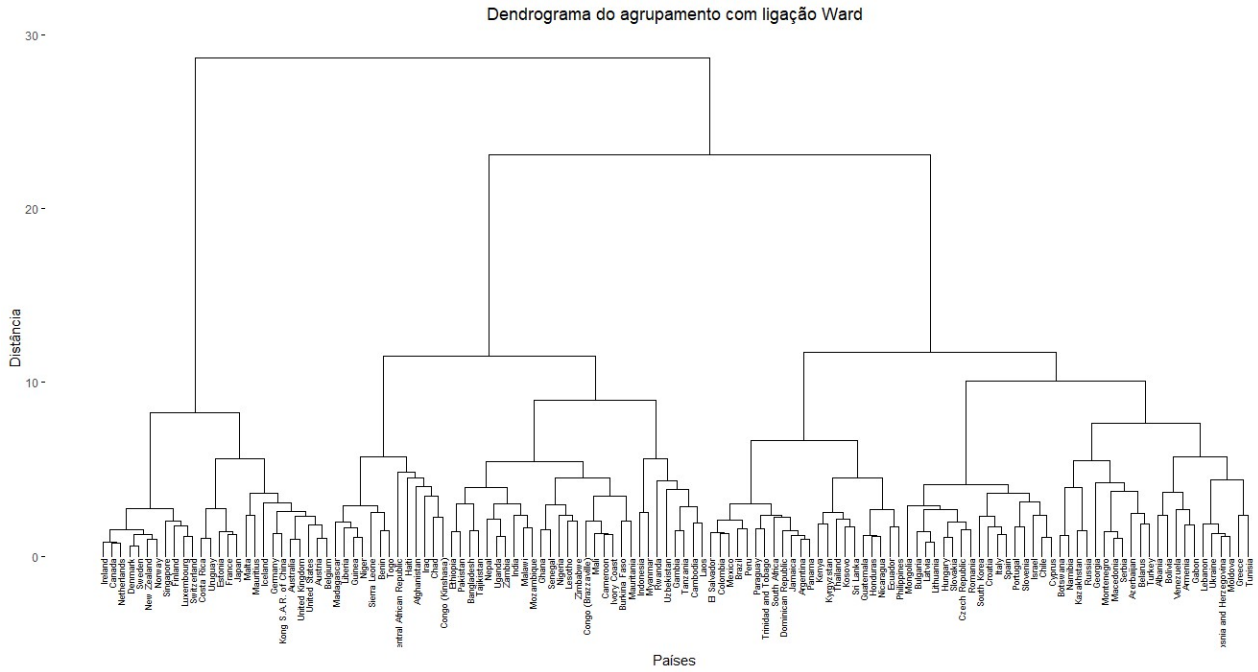
12.1 Resultado do Agrupamento com Ligação Ward

Considerando inicialmente cada país como seu próprio cluster, a técnica de ligação Ward combina gradualmente os clusters de forma a minimizar a variância dentro de cada novo cluster formado. Para cada possível par de clusters, o algoritmo

calcula como a variabilidade dentro do novo cluster formado difere da variabilidade nos clusters originais separados. O algoritmo de ligação Ward opta por combinar os dois clusters que resultam na menor adição à variabilidade total. Esse processo de cálculo de variabilidade, escolha da combinação de clusters e criação de novos clusters é repetido até que todos os países estejam agrupados em um único cluster.

É possível observar o resultado final através do dendrograma apresentado na Figura 6.

Figura 6: Dendrograma do Resultado do Agrupamento com Ligação Ward.



Fonte: Elaborado pela autora (2023).

Pelo dendrograma acima, pode-se considerar que o número ideal de clusters seja 3, uma vez que se observam ligações mais próximas antes de ocorrer um grande aumento na distância.

Ao aplicar a função NbClust, foi obtido o resultado apresentado na Figura 7 abaixo.

Figura 7: Resultado da Função NbClust para Ligação Ward.

```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 4 proposed 2 as the best number of clusters
* 14 proposed 3 as the best number of clusters
* 1 proposed 4 as the best number of clusters
* 1 proposed 13 as the best number of clusters
* 2 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3

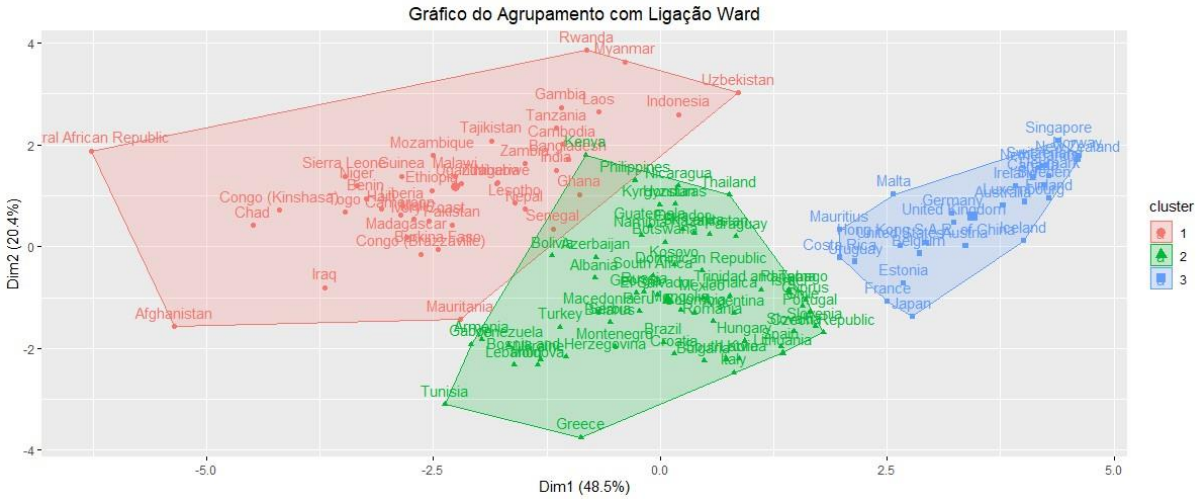

*******************************************************************
```

Fonte: Elaborado pela autora (2023).

Logo, foi indicado por 14 índices que 3 é o número ideal. Desta forma, foi determinado que o número ideal de clusters é 3, resultando em 42 países no grupo 1, 61 no grupo 2 e 26 no grupo 3.

Desta forma, a divisão dos grupos pode ser observada pelo gráfico abaixo, onde cada cor representa um grupo.

Figura 8: Gráfico do Agrupamento com Ligação Ward.



Fonte: Elaborado pela autora (2023).

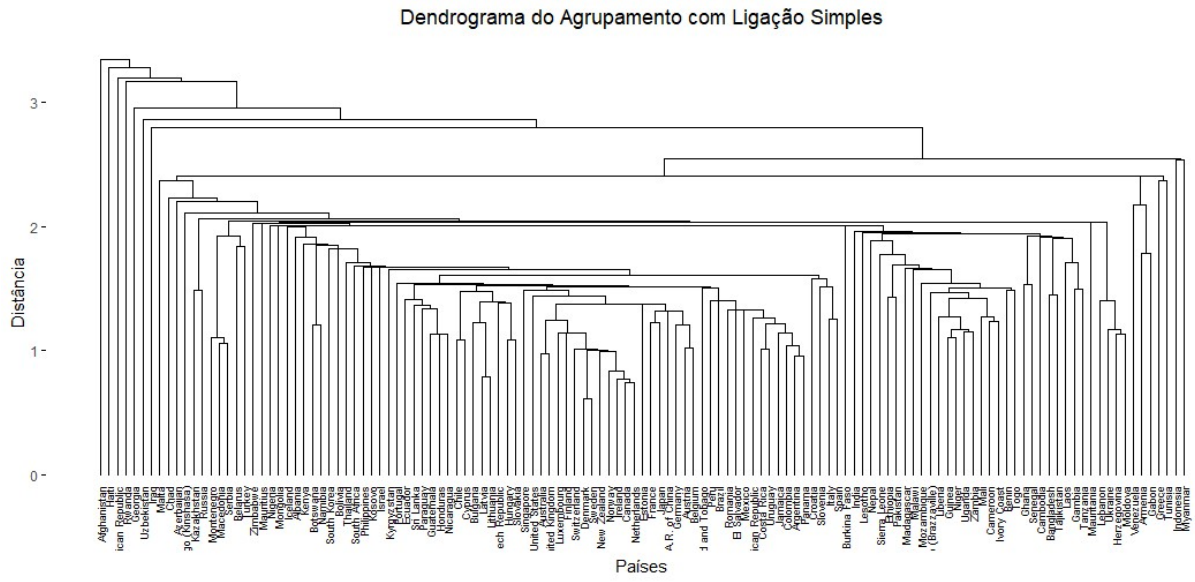
A imagem acima revela grupos claramente distintos, isso sugere que as observações foram agrupadas com base em semelhanças evidentes. Essa clareza nos grupos destaca a eficácia do método de agrupamento.

12.2 Resultado do Agrupamento com Ligação Simples

Na aplicação do agrupamento hierárquico com ligação simples, o algoritmo iterativamente vai selecionando os dois clusters cujos países estão mais próximos um do outro em termos de distância e combinando esses dois clusters em um novo cluster, repetindo até que todos os países estejam agrupados em um único cluster.

E como resultado, tem-se o seguinte dendrograma:

Figura 9: Dendrograma do Resultado do Agrupamento com Ligação Simples.



Fonte: Elaborado pela autora (2023).

Ao observar o dendrograma para ligação simples, nota-se que não há grupos bem definidos. Isso sugere que os dados não têm uma estrutura bem organizada que possa ser identificada por esse método, que se baseia em encontrar as distâncias mais curtas entre os pontos.

Como pelo dendrograma não é possível determinar o número de grupos, então através do resultado da função NbClust disposto na Figura 10, será definido o número ideal.

Figura 10: Resultado da Função NbClust para Ligação Simples.

```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 8 proposed 2 as the best number of clusters
* 2 proposed 3 as the best number of clusters
* 3 proposed 4 as the best number of clusters
* 2 proposed 5 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 3 proposed 9 as the best number of clusters
* 1 proposed 10 as the best number of clusters
* 3 proposed 11 as the best number of clusters
* 1 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  2

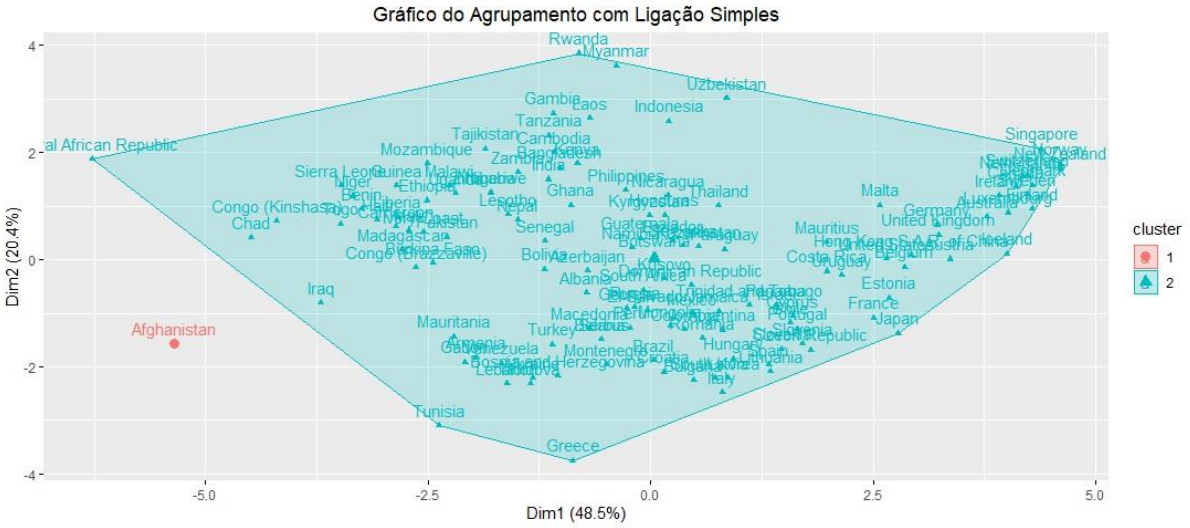

*******************************************************************
```

Fonte: Elaborado pela autora (2023).

Desta forma, foi definido por 8 índices que o número ideal seja 2, resultando no grupo 1 com 1 observação e no grupo 2 com 128 observações. Essa divisão não fornece informações úteis para a análise, isso torna difícil tirar conclusões com base nesse cluster de apenas 1 elemento.

Com isto, é possível observar os dois grupos no gráfico a seguir.

Figura 11: Gráfico do Agrupamento com Ligação Simples.



Fonte: Elaborado pela autora (2023).

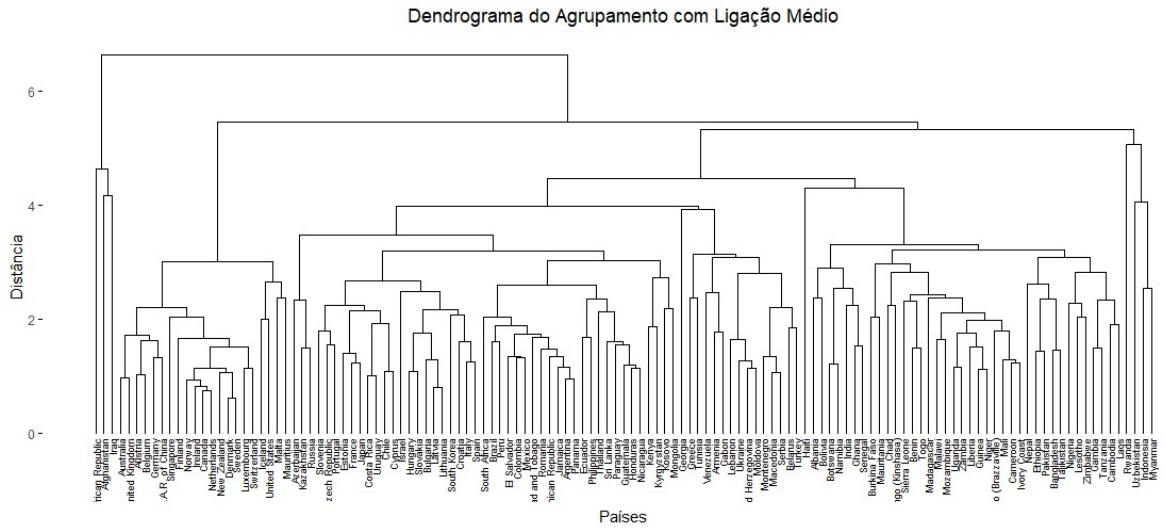
Como já comentado anteriormente, esta técnica de ligação não se adequou bem aos dados e não houve uma boa separação de grupos.

12.3 Resultado do Agrupamento com Ligação Média

Ao aplicar o agrupamento com a técnica de ligação média, os países iniciam como grupos individuais, em seguida, é calculado a distância média entre todos os pares de países e une os dois grupos que têm a menor distância média. E então, o processo é repetido até que todos os países estejam agrupados em um único grupo.

Como resultado, foi gerado o seguinte dendrograma:

Figura 12: Dendrograma do Resultado do Agrupamento com Ligação Média.



Fonte: Elaborado pela autora (2023).

Com a imagem acima, é possível considerar a existência de 3 possíveis grupos. O resultado da função NbClust para a técnica de ligação média será apresentado na Figura 13.

Figura 13: Resultado da Função NbClust para Ligação Média.

```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 5 proposed 2 as the best number of clusters
* 9 proposed 3 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 4 proposed 7 as the best number of clusters
* 1 proposed 9 as the best number of clusters
* 2 proposed 10 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 1 proposed 15 as the best number of clusters

                   ***** Conclusion *****

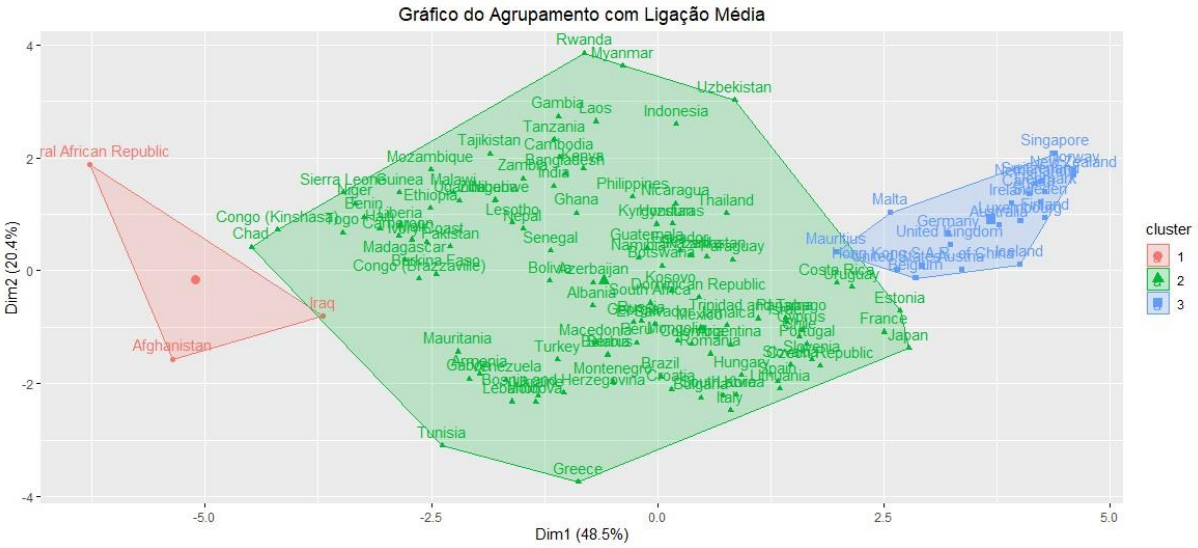
* According to the majority rule, the best number of clusters is  3


*******************************************************************
```

Fonte: Elaborado pela autora (2023).

Logo, comprovou-se o número ideal 3 através de 9 índices. Desta forma, o grupo 1 ficou com 3 países, o grupo 2 com 105 e o grupo 3 com 21 países, e é possível observá-los através do gráfico abaixo.

Figura 14: Gráfico do Agrupamento com Ligação Média.



Fonte: Elaborado pela autora (2023).

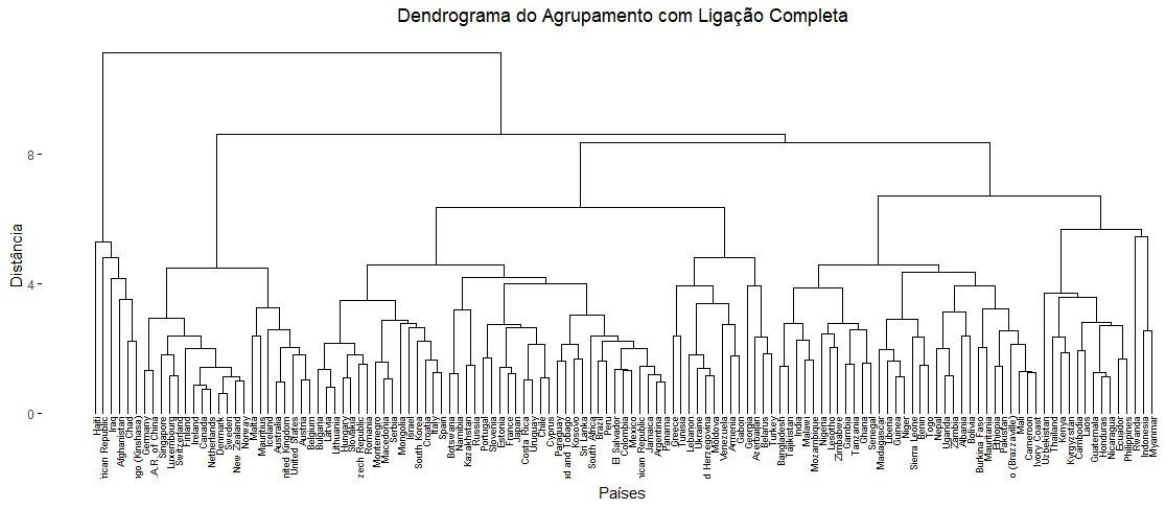
A imagem acima mostra que os grupos estão bem separados, porém pela pouca quantidade de países no grupo 1, a interpretação final dos agrupamentos talvez não seja muito significativa.

12.4 Resultado do Agrupamento com Ligação Completa

No método de agrupamento hierárquico com ligação completa, é avaliada a dissimilaridade entre dois clusters observando a maior distância possível (a distância máxima) entre qualquer par de pontos que pertencem a esses clusters. Quando são encontrados os dois clusters com a maior dissimilaridade máxima entre seus pontos, eles são combinados para formar um novo cluster.

Para essa técnica, obteve-se o seguinte dendrograma:

Figura 15: Dendrograma do Resultado do Agrupamento com Ligação Completa.



Fonte: Elaborado pela autora (2023).

Levando em consideração o dendrograma gerado, é possível considerar que existem 4 grupos para o agrupamento com ligação completa.

Em seguida, será exibido o resultado da função NbClust.

Figura 16: Resultado da Função NbClust para Ligação Completa.

```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 3 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 10 proposed 4 as the best number of clusters
* 2 proposed 6 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 1 proposed 10 as the best number of clusters
* 1 proposed 11 as the best number of clusters
* 1 proposed 13 as the best number of clusters
* 1 proposed 15 as the best number of clusters

                   ***** Conclusion *****

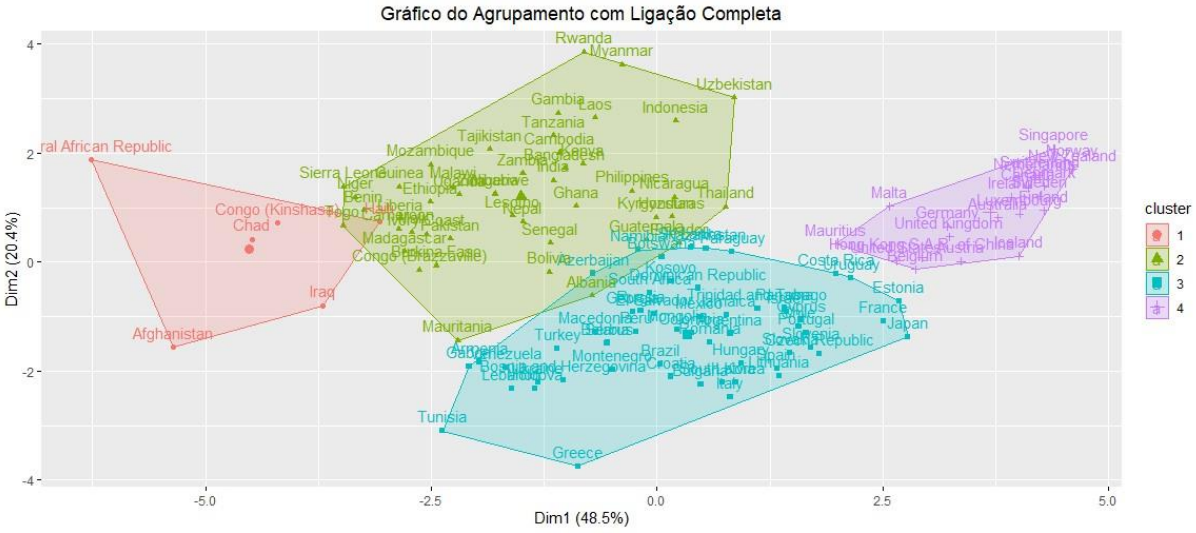
* According to the majority rule, the best number of clusters is  4


*******************************************************************
```

Fonte: Elaborado pela autora (2023).

Com o resultado da Figura 16, foi confirmada a análise do dendrograma, sendo que foi indicado por 10 índices que 4 é o número ideal de clusters. Resultando em 6 países no grupo 1, 46 países no grupo 2, 56 no grupo 3 e 21 países no grupo 4. É possível verificar este resultado no gráfico abaixo:

Figura 17: Gráfico do Agrupamento com Ligação Completa.



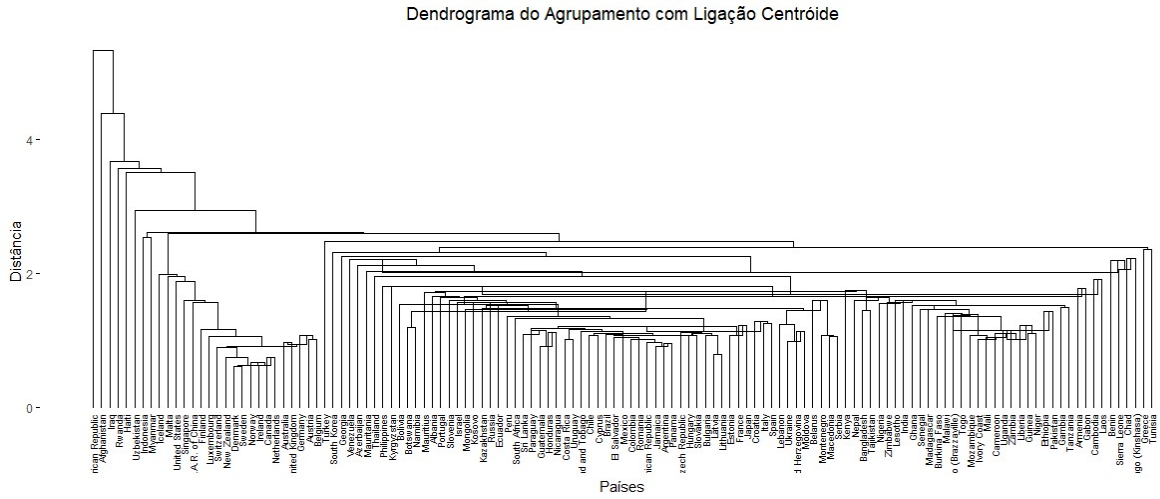
Fonte: Elaborado pela autora (2023).

O gráfico gerado mostra uma boa divisão dos países, sendo que foi a técnica que gerou o maior número de grupos, esse ponto pode ser levado em consideração no momento de escolher qual a melhor técnica de agrupamento.

12.5 Resultado do Agrupamento com Ligação Centróide

No método de ligação centroide, a dissimilaridade entre dois clusters é medida pela distância entre seus centróides. O centróide de um cluster é calculado como o ponto médio (a média) das coordenadas dos pontos pertencentes a esse cluster. Os dois clusters que tiverem a menor distância entre seus centróides serão unidos em um novo cluster. E assim segue iterativamente até todos os países serem agrupados em um único cluster, resultando no dendrograma abaixo.

Figura 18: Dendrograma do Agrupamento com Ligação Centróide.



Fonte: Elaborado pela autora (2023).

Assim como no caso do agrupamento com ligação simples, o dendrograma não mostra uma clara divisão de grupos.

Agora, será exposto na Figura 19 o resultado da função NbClust para a ligação centróide.

Figura 19: Resultado da Função NbClust para Ligação Centróide.

```
*** : The Hubert index is a graphical method of determining the number of clusters.
                In the plot of Hubert index, we seek a significant knee that corresponds to a
                significant increase of the value of the measure i.e the significant peak in Hubert
                index second differences plot.

*** : The D index is a graphical method of determining the number of clusters.
                In the plot of D index, we seek a significant knee (the significant peak in Dindex
                second differences plot) that corresponds to a significant increase of the value of
                the measure.

*******************************************************************
* Among all indices:
* 8 proposed 2 as the best number of clusters
* 2 proposed 3 as the best number of clusters
* 1 proposed 6 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 10 proposed 9 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 1 proposed 15 as the best number of clusters

                   ***** Conclusion *****

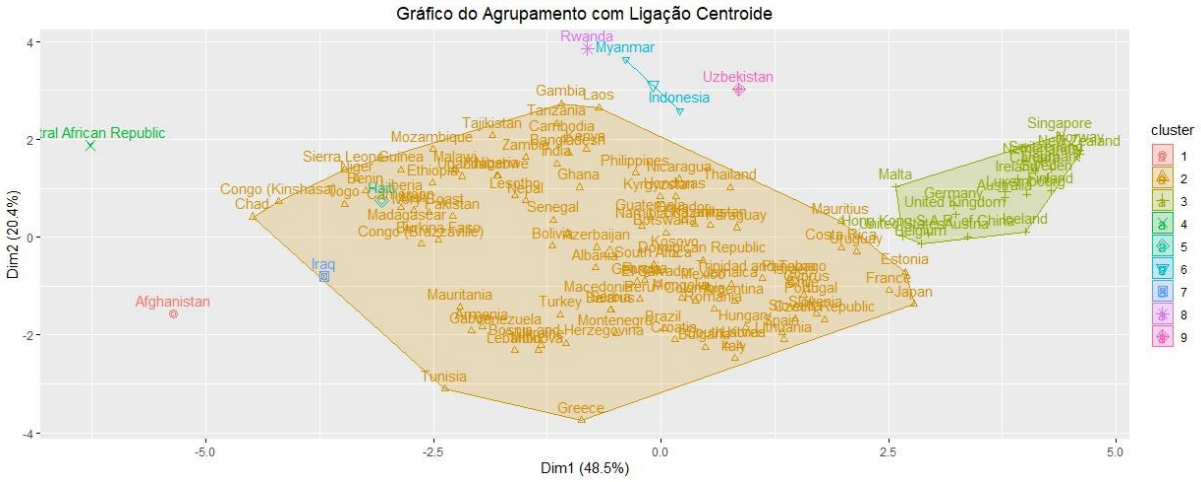
* According to the majority rule, the best number of clusters is  9


*******************************************************************
```

Fonte: Elaborado pela autora (2023).

Por meio da função NbClust, 10 índices apontaram que o número ideal de grupos é 9. Isso resultou na formação de 6 grupos contendo 1 apenas 1 país, outro 1 grupo contendo 2 países, um grupo com 101 países e um último grupo com 20 países. Sendo possível visualizá-los no gráfico abaixo.

Figura 20: Gráfico de Agrupamento com Ligação Centróide.



Fonte: Elaborado pela autora (2023).

Assim como na ligação simples, o fato de existir muitos grupos com poucos países dificulta na interpretação do resultado, o que influenciará negativamente na comparação entre técnicas de ligação.

13 Comparação dos agrupamentos hierárquicos

Como apresentado anteriormente, para definir qual a melhor técnica de ligação para este estudo, além das análises feitas no capítulo anterior, será analisada a correlação cofenética e o gráfico de silhueta.

A seguir, será apresentada uma tabela com o valor da correlação cofenética para cada técnica.

Tabela 3: Correlação Cofenética Para Cada Técnica de Ligação.

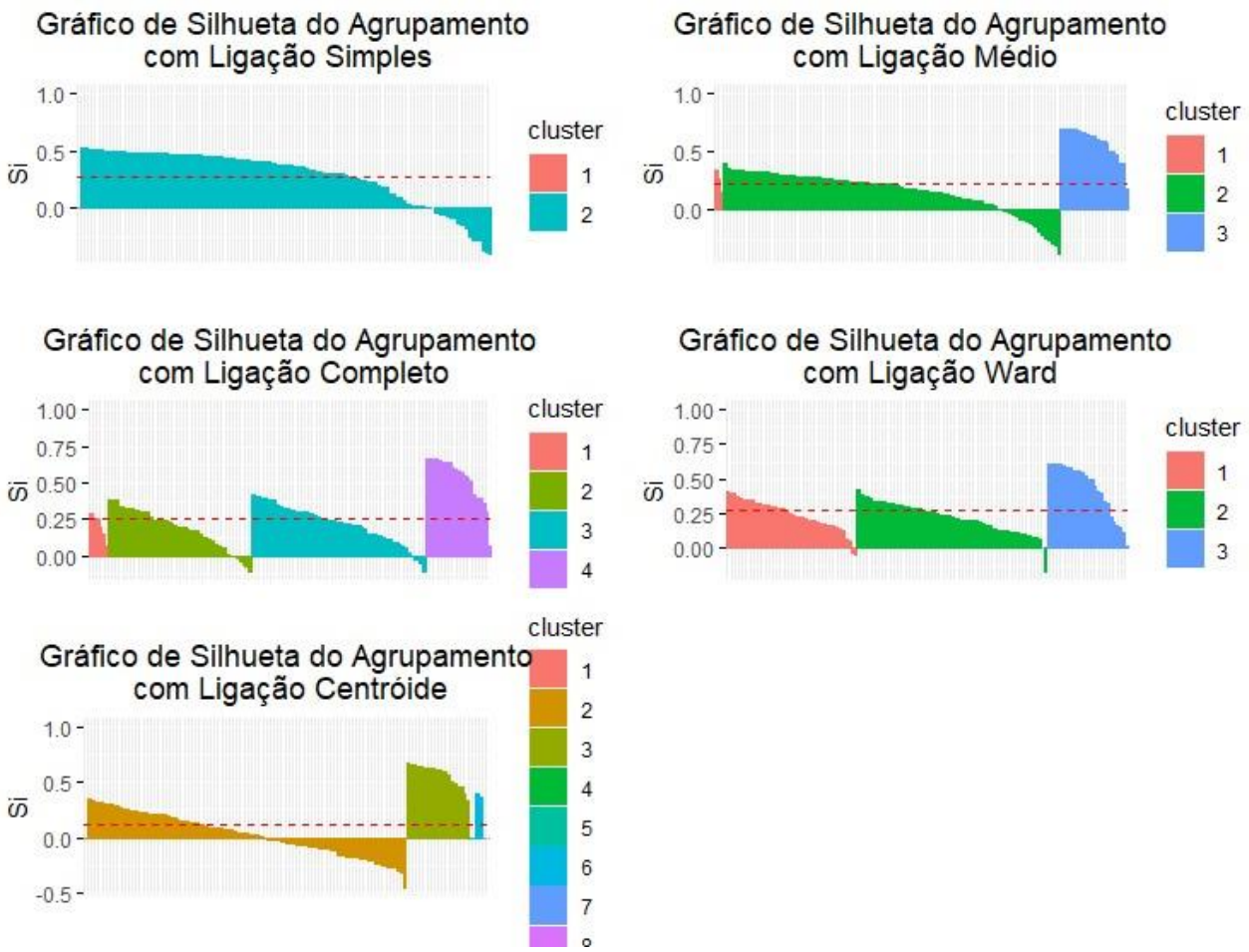
Tipo de ligação	**Correlação Cofenética**
Simples	0,4811
Média	0,7235
Completa	0,6752
Centróide	0,6485
Ward	0,6187

Fonte: Elaborado pela autora (2023).

Com os resultados acima, é possível concluir que a análise de agrupamento com ligação média exibe uma correlação cofenética mais elevada, isso significa que a estrutura de agrupamento hierárquico é uma representação confiável da verdadeira dissimilaridade entre as observações.

Agora analisando os gráficos de silhueta de cada técnica de ligação apresentados abaixo.

Figura 21: Gráficos de Silhueta para Cada Técnica de Ligação.



Fonte: Elaborado pela autora (2023).

Primeiro, ao olhar para a média das métricas de silhueta de todos os pontos em seus grupos, que é basicamente uma média das medidas individuais de cada ponto, representada por uma linha vermelha tracejada em cada gráfico, é possível perceber que todas as técnicas de ligação têm valores próximos e positivos. Isso sugere que os dados estão razoavelmente bem agrupados dentro de seus clusters.

Agora, ao analisar a métrica de silhueta para cada elemento em cada técnica de ligação, nota-se que nas técnicas de ligação Ward e completa, a maioria dos valores são positivos e mais próximos de 1. Isso indica que os dados estão bem agrupados em seus respectivos clusters. Por outro lado, nas outras técnicas, há uma maior quantidade de valores negativos e mais próximos de -1, o que sugere que alguns elementos (países) podem ter sido atribuídos aos clusters errados.

Com base em todas as análises realizadas até agora, é possível concluir que as técnicas de ligação que se destacaram neste estudo foram a Ward e a completa. Embora não tenham alcançado os melhores resultados de correlação cofenética, ainda assim apresentaram valores satisfatórios. Ao examinar os dendrogramas e a forma como os grupos foram divididos visualmente, essas duas técnicas se destacaram por oferecerem as divisões mais claras e informativas sobre os grupos e os países em cada grupo. Além disso, elas apresentaram um desempenho superior na análise do gráfico de silhueta.

Em contraste, a técnica de ligação média, apesar de ter a maior correlação cofenética, não obteve bons resultados nas métricas de silhueta e não conseguiu uma separação eficaz dos grupos no dendrograma. O mesmo se aplica às técnicas de ligação centróide e simples, que desde a construção do dendrograma já demonstraram uma clara dificuldade na separação dos grupos.

Portanto, com base em várias análises, as técnicas de ligação Ward e completa se destacaram como as mais adequadas para este estudo. No entanto, a ligação completa será a escolhida para a análise dos grupos devido à sua correlação cofenética ligeiramente superior e à sua capacidade de gerar um maior número de grupos, o que pode proporcionar *insights* mais nítidos e informativos.

14 Considerações finais

Conforme planejado nos objetivos deste trabalho, agora serão apresentados e analisados os grupos e os países obtidos por meio do agrupamento utilizando a técnica de ligação completa.

Os países em cada grupo serão apresentados nos quadros abaixo.

Quadro 1: Países Agrupados no Grupo 1.

GRUPO 1					
Afeganistão	República Centro-Africana	Chade	República Democrática do Congo	Haiti	Iraque

Fonte: Elaborado pela autora (2023).

No grupo 1, observa-se que 50% dos países são da África, 33,33% são da Ásia e 16,67% da América do Norte.

Quadro 2: Países Agrupados no Grupo 2.

GRUPO 2					
Albânia	Bangladesh	Benim	Bolívia	Burquina Fasso	Camboja
Camarões	República do Congo	Equador	Etiópia	Gâmbia	Gana
Guatemala	Guiné	Honduras	Índia	Indonésia	Costa do Marfim
Quênia	Quirguistão	Laos	Lesoto	Libéria	Madagascar
Malawi	Mali	Mauritânia	Moçambique	Myanmar	Nepal
Nicarágua	Níger	Nigéria	Paquistão	Filipinas	Ruanda
Senegal	Serra Leoa	Tajiquistão	Tanzânia	Tailândia	Togo
Uganda	Uzbequistão	Zâmbia	Zimbábue		

Fonte: Elaborado pela autora (2023).

Já no grupo 2, 58,70% são da África, 28,27% são da Ásia, 6,52% são da América do Norte, 4,34% são da América do Sul, e apenas 2,17% são da Europa.

Quadro 3: Países Agrupados no Grupo 3.

GRUPO 3					
Argentina	Armênia	Azerbaijão	Bielorrússia	Bósnia e Herzegovina	Botsuana
Brasil	Bulgária	Chile	Colômbia	Costa Rica	Croácia
Chipre	República Tcheca	República Dominicana	El Salvador	Estônia	França
Gabão	Geórgia	Grécia	Hungria	Israel	Itália
Jamaica	Japão	Cazaquistão	Kosovo	Letônia	Líbano
Lituânia	Macedônia	México	Moldávia	Mongólia	Montenegro
Namíbia	Panamá	Paraguai	Peru	Portugal	Romênia

Rússia	Sérvia	Eslováquia	Eslovênia	África do Sul	Coréia do Sul
Espanha	Sri Lanka	Trindade e Tobago	Tunísia	Turquia	Ucrânia
Uruguai	Venezuela				

Fonte: Elaborado pela autora (2023).

Agora no grupo 3, 46,43% dos países são da Europa, 17,85% são da Ásia, 14,30% são da América do Sul, 12,50% são da América do Norte e 8,92% são da África.

Quadro 4: Países Agrupados no Grupo 4.

GRUPO 4					
Austrália	Áustria	Bélgica	Canadá	Dinamarca	Finlândia
Alemanha	China	Islândia	Irlanda	Luxemburgo	Malta
Ilhas Maurício	Países Baixos	Nova Zelândia	Noruega	Singapura	Suécia
Suíça	Reino Unido	Estados Unidos			

Fonte: Elaborado pela autora (2023).

Por fim, observando o grupo 4, 63,64% dos países estão localizados na Europa, 13,64% estão na Oceania, e por fim, 9,09% estão na América do Norte e Ásia.

Com base nas análises, pode-se observar uma distribuição dos países por grupos da seguinte forma:

- A maioria dos países da África está presente nos grupos 1 e 2, sendo que o grupo 2 abriga a maior parte, com um total de 27 países;
- Os países da América do Norte estão dispersos em diferentes grupos, com a maioria deles no grupo 3, totalizando 7 países;
- A América do Sul segue um padrão semelhante à América do Norte, com a maioria dos seus países também no grupo 3, totalizando 8 países;
- A Ásia concentra-se principalmente nos grupos 2 e 3, enquanto nos grupos 1

e 4 têm apenas 2 países cada;

- A Europa predomina nos grupos 3 e 4;
- Finalmente, a Oceania está exclusivamente no grupo 4.

Agora, ao examinar a média de cada variável dentro de cada grupo, têm-se os seguintes resultados:

Tabela 4: Médias das Variáveis em Cada Grupo.

Variável	Médias no Grupo 1	Médias no Grupo 2	Médias no Grupo 3	Médias no Grupo 4
Log PIB Per Capita	7,54	8,08	9,78	10,78
Assistência Social	0,58	0,72	0,86	0,93
Expectativa de Vida	52,28	58,23	67,24	72,41
Liberdade	0,58	0,80	0,75	0,91
Generosidade	0,07	0,04	-0,11	0,17
Percepção de Corrupção	0,81	0,75	0,82	0,41
Afeto Positivo	0,56	0,71	0,69	0,79
Afeto Negativo	0,47	0,33	0,27	0,20
Confiança no Governo	0,45	0,66	0,36	0,60
Qualidade Democrática	-1,51	-0,55	0,15	1,13
Qualidade da Entrega	-1,45	-0,64	0,16	1,69
Score da Felicidade	3,85	4,75	5,66	7,05

Fonte: Elaborado pela autora (2023).

Observa-se que, para as variáveis em que um valor mais alto indica condições melhores, as médias aumentam à medida que se avança de grupos, com os valores menores nos grupos 1 e 2 e os maiores nos grupos 3 e 4. Por outro lado,

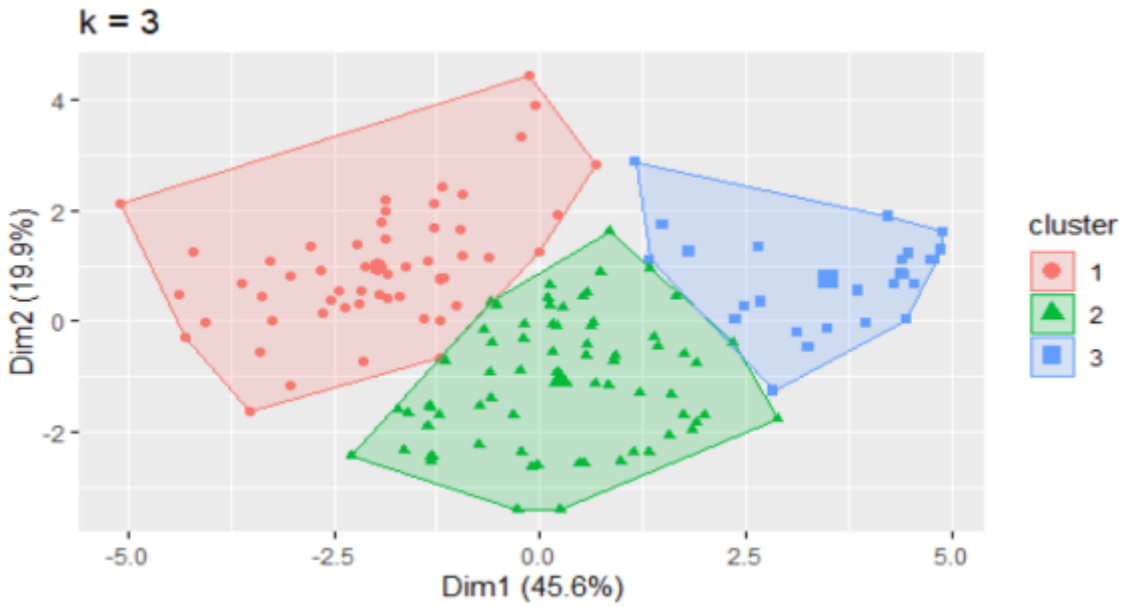
nas variáveis em que um valor mais baixo indica melhores condições, ocorre o oposto. Portanto, os países nos grupos 3 e 4 claramente demonstram um nível mais elevado de desenvolvimento e melhores condições, enquanto a maioria dos países nos grupos 1 e 2 se encontra em desenvolvimento.

Na análise final, foi incorporada a variável "Score de Felicidade," que é um resultado anualmente divulgado no Relatório Mundial da Felicidade. Essa variável não foi incluída durante o desenvolvimento do modelo, a fim de evitar influências nos resultados. Essa abordagem permitiu que se observasse se houve uma diferença notável nos escores de felicidade entre os grupos.

Os resultados claramente indicam que essa variável segue um padrão semelhante às outras variáveis, onde os países mais felizes estão agrupados nos grupos 3 e 4, enquanto os países menos felizes estão predominantemente nos grupos 1 e 2.

Ao comparar com o trabalho mencionado no referencial teórico, *A Real Data-Driven Clustering Approach for Countries Based on Happiness Score* (Chakraborty e Tsokos, 2021), no qual foram empregadas técnicas de agrupamento não hierárquico, é possível observar resultados semelhantes. Através do algoritmo K-means, os países foram divididos em 3 grupos, resultando no seguinte gráfico:

Figura 22: Os Três Grupos Obtidos Pelo Agrupamento K-Means.



Fonte: Figura do Artigo A Real Data-Driven Clustering Approach for Countries Based on Happiness Score (Chakraborty e Tsokos, 2021).

Nota-se, a partir dos resultados do agrupamento mencionado anteriormente, que a técnica utilizada conseguiu uma divisão mais eficaz entre os países, visto que há pouca intersecção entre os grupos. Contudo, é importante ressaltar que essa conclusão se baseia apenas na análise visual do gráfico, pois não é viável comparar numericamente a eficácia dos métodos, uma vez que métricas de qualidade diferentes foram utilizadas nos dois estudos.

Dessa forma, todos os objetivos da pesquisa foram integralmente alcançados. As cinco técnicas de ligação foram aplicadas, analisadas e comparadas. A técnica mais apropriada foi selecionada, e uma análise aprofundada de todos os grupos foi conduzida, destacando-se as influências das variáveis em cada grupo e as notáveis disparidades socioeconômicas entre eles. Vale ressaltar que o tema abordado neste estudo é versátil e suscetível a múltiplas abordagens, e os resultados obtidos podem oferecer diversas outras conclusões quando combinados com diferentes visões além das análises estatísticas.

Referências

ALBUQUERQUE, Mácio Augusto de. **Estabilidade em análise de agrupamento (cluster analysis)**. 2005. 64 f. Dissertação (Programa de Pós-Graduação em Biometria e Estatística Aplicada) - Universidade Federal Rural de Pernambuco, Recife.

BITTENCOURT, Ubiratan. **Relatório Mundial da Felicidade** - ranking dos países mais felizes. Canal da Felicidade, 3 de abr. de 2020. Disponível em: https://canaldafelicidade.com.br/relatorio-mundial-da-felicidade/. Acesso em: 18 dez. 2022.

CHAKRABORTY, Aditya; TSOKOS, Chris. **A Real Data-Driven Analytical Model to Predict Happiness.** 2021.

CHAKRABORTY, Aditya; TSOKOS, Chris. **A Real Data-Driven Clustering Approach for Countries Based on Happiness Score**. Amfiteatru Economic, v. 23, n. 15, p. 1031-1045. 2021.

ESTATÍSTICA da felicidade. **Valor-P**, São Paulo: Conre-3, n. 5, p. 3-6, jan./fev./mar. 2007.

FÁVERO, Luiz; BELFIORE, Patrícia. **Manual de análise de dados:** estatística e modelagem multivariada com Excel®, SPSS® e Stata®. Elsevier Brasil, 2017.

FELICIDADE. *In:* DICIO, Dicionário Online de Português. Porto: 7Graus, 2022. Disponível em: https://www.dicio.com.br/felicidade/. Acesso em: 10 dez. 2022.

FERREIRA, Rogério; PAIM, Fernando; RODRIGUES, Valéria; CASTRO, Gustavo. et al. **Análise de cluster não supervisionado em R**: agrupamento hierárquico. 2020.

HAIR, Joseph; BLACK, Willian; BABIN, Barry; ANDERSON, Rolph; TATHAM, Ronald. **Análise multivariada de dados**. Bookman editora, 2009.

HELLIWEL, John; LAVARD, Richard; SACHS, Jeffrey. World Happiness Report 2012. World Happiness Report, 2012. Disponível em: https://worldhappiness.report/ed/2012/. Acesso em: 10 dez. 2022.

JOHNSON, Richard; WICHERN, Dean. **Applied multivariate statistical analysis**. 5.ed. New Jersey: Prentice Hall, 2002. 767p.

KASSAMBARA, Alboukadel. **Practical guide to cluster analysis in R**: Unsupervised machine learning. Sthda. 2017.

MINGOTI, Sueli. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada**.** Belo Horizonte: UFMG, 2005. 295p.

OLIVEIRA, Natália Lima de. **Análise de agrupamento hierárquicos.** 2022. 50 f.

Monografia (Graduação em Estatística) – Centro de Ciências, Universidade Federal do Ceará, Fortaleza, 2022.

SILVESTRE, M. R. **Notas de aula de Análise Multivariada II**. 2022. Presidente Prudente: [s. n.], 2022. Apostila disponível no ambiente virtual de apoio disciplina Análise Multivariada II da FCT/Unesp.

SUSTAINABLE Development Solutions Network (SDSN) Tracks the World's Happiness. Gallup, c2022. Disponível em: https://www.gallup.com/analytics/247355/gallup-world-happiness-report.aspx. Acesso em: 11 dez. 2022.

TAN, Yu; SINGHAPREECHA, Charuk; YAMAKA, Woraphon. **Applying Machine Learning to Predict Happiness**: A case study of 20 Countries. Mukht Shabd Journal, v. 9, n. 6, p. 3433-3437, 2020.

VALE, M. N. **Agrupamentos de dados: avaliação de métodos e desenvolvimento de aplicativo para análise de grupos**. 2005. 120f. Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Engenharia Elétrica, Rio de Janeiro, 2005.