Universidade Federal de Uberlândia
Faculdade de Matemática

Bacharelado em Estatística

ANÁLISE MULTIVARIADA DE PAÍSES DA AMÉRICA E EUROPA UTILIZANDO INDICADORES SOBRE A COVID-19 E DIETA DA POPULAÇÃO

Marcos Vinícius Vieira de Souza

Uberlândia-MG
2022

Marcos Vinícius Vieira de Souza

ANÁLISE MULTIVARIADA DE PAÍSES DA AMÉRICA E EUROPA UTILIZANDO INDICADORES SOBRE A COVID-19 E DIETA DA POPULAÇÃO

Trabalho de conclusão de curso apresentado à Coordenação do Curso de Bacharelado em Estatística como requisito parcial para obtenção do grau de Bacharel em Estatística.

Orientador: Prof. Dr. Lúcio Borges de Araújo

Uberlândia-MG

2022

Universidade Federal de Uberlândia
Faculdade de Matemática

Coordenação do Curso de Bacharelado em Estatística

A banca examinadora, conforme abaixo assinado, certifica a adequação deste trabalho de conclusão de curso para obtenção do grau de Bacharel em Estatística.

Uberlândia, ____ de _____ de 20____

BANCA EXAMINADORA

_____

Prof. Dr. Lúcio Borges de Araújo

_____

Profa. Dra. Patrícia Ferreira Paranaíba

_____

Prof. Dr. Leandro Alves Pereira

Uberlândia-MG

2022

Agradecimentos

Primeiramente, agradeço aos meus pais Vilma Vieira do Nascimento e Fábio Luiz Ferreira de Souza que me incentivaram, aconselharam e deram todo o apoio durante a minha graduação, todas as minhas conquistas sempre serão dedicadas a vocês. Agradeço aos meus irmãos Luiz Fernando e Maria Emanuelle, por acreditarem e me apoiarem. Agradeço a minha família, que sempre esteve ao meu lado seja na vitória ou na derrota.

Ao meu orientador Prof. Dr. Lúcio Borges de Araújo, pela sua disposição, paciência e compreensão. Obrigado pela sua confiança e por fazer parte do fim deste ciclo.

À todos os professores, técnicos e servidores da UFU que fizeram parte desta conquista, em especial ao Prof. Dr. Pedro Franklin e a Profa. Dra. Priscila Neves Faria que foram meus orientadores de Iniciação Científica, muito obrigado por todo o conhecimento que me foi repassado.

Aos tutores e colegas de curso com quem tive a oportunidade de conviver durante o Programa de Educação Tutorial (PET) Estatística. Obrigado pelas oportunidades e experiências que partilhamos juntos, com certeza me fizeram crescer como profissional e pessoa.

À Kyros Tecnologia, especialmente para o Roberto da Silva Perreira e para o Henrique de Castro Neto, que me deram a oportunidade de desenvolver e crescer como profissional, obrigado por acreditarem em mim.

Resumo

A alimentação é um fator primordial na rotina das pessoas e que interfere diretamente na saúde e bem estar do indivíduo. Dado a gravidade e os desdobramentos da pandemia do coronavírus (Covid-19) se faz necessário verificar como o perfil alimentar de uma população se relaciona com os desdobramentos daquela que, até o momento deste estudo, é a maior pandemia do século XXI. Assim, este trabalho teve como objetivo explorar indicadores populacionais referentes a dieta e Covid-19 de 80 países da América e Europa utilizando a Análise de *Cluster* e o Biplot, ambas técnicas da Estatística Multivariada. O presente estudo permitiu concluir que a Análise de *Cluster* mostrou-se eficaz na identificação de 6 grupos de países e que o Gráfico Biplot permitiu observar as associações que as variáveis relacionadas à Covid-19 possuem com as variáveis da dieta, que seria uma associação positiva com Leite, uma associação negativa com Raízes com amido e Frutas e uma aparente ausência de associação com Legumes, Cereais e Obesidade.

Palavras-chave: Agrupamento, Biplot, Pandemia.

Abstract

Food is a key factor in people's routine and directly interferes with the health and well-being of the individual. Given the severity and developments of the coronavirus (Covid-19) pandemic, it is necessary to verify how the population relates to the developments of a pandemic that, until the moment of this study, is the greatest pandemic of the 21st century. Thus, this work explored population indicators and Covid-19 references to 80 countries in America and Europe using Cluster Analysis and Biplot, both techniques of Multivariate Statistics. The present study allowed us to conclude that the Cluster Analysis proved to be effective in identifying 6 groups of countries and that the Biplot Chart allowed us to observe the associations that the variables related to Covid-19 have with the diet variables, which would be a positive association with Milk, a negative association with Starchy Roots and Fruits and an apparent lack of association with Vegetables, Cereals and Obesity.

Keywords: Grouping, Biplot, Pandemic.

Sumário

Lista de Figuras

Lista de Tabelas

1. Introdução

Segundo [16] o surto do Coronavírus (Covid-19) teve início no fim de 2019 em Wuhan, cidade localizada na República Popular da China. A primeira morte pela doença foi noticiada em 11 de janeiro de 2020 e o primeiro caso confirmado fora da China ocorreu em 20 de janeiro de 2020. Em 30 de janeiro de 2020, o Comitê de Emergência da OMS declarou uma emergência de saúde global com base nas crescentes taxas de notificação de casos em locais chineses e internacionais.

A maioria das pessoas infectadas com o vírus Sars-CoV-2 apresentam doença respiratória leve a moderada e se recuperam sem a necessidade de tratamento especial. Idosos e aqueles com problemas médicos subjacentes, como doenças cardiovasculares, diabetes, doenças respiratórias crônicas e câncer, têm maior probabilidade de desenvolver doenças graves [15].

Diante de tal situação, é de extrema importância conhecer as maneiras de se proteger dessa doença. Entre as medidas indicadas pelo Ministério da Saúde do Brasil, estão as não farmacológicas, como distanciamento social, etiqueta respiratória e de higienização das mãos, uso de máscaras, limpeza e desinfecção de ambientes, isolamento de casos suspeitos e confirmados e quarentena dos contatos dos casos de Covid-19, conforme orientações médicas. Ademais, o Ministério da Saúde recomenda ainda a vacinação contra a Covid-19 [4].

O fortalecimento do sistema imunitário auxilia no combate a infecções, e é uma medida necessária para que a recuperação após o contágio seja mais eficiente e cause menos danos possíveis à saúde. Para isso, é importante usar a nutrição e os bons hábitos de vida como aliados. Assim para enfrentar esta pandemia do coronavírus que assola a nossa sociedade, é necessário o reforço da imunidade por meio de uma alimentação saudável [3].

O Centro de Política e Promoção Nutricional do USDA (*United States Department of Agriculture*) recomenda uma alimentação que cumpre uma diretriz de ingestão diária seguindo as proporções de 30% de grãos, 40% de vegetais, 10% de frutas e 20% de proteína, porém seguir uma dieta adequada é um desafio, principalmente quando se tem poucas condições, informações ou opções. Ou dependendo do modo de vida do indivíduo, que acaba por não dar a devida atenção em seguir uma alimentação saudável. De acordo com a médica nutróloga do Hospital Universitário de Brasília, Ana de Oliveira Parada, essas situações ao extremo podem gerar a desnutrição e a obesidade. A obesidade e a desnutrição são doenças nutricionais que podem ou não ser relacionadas a problemas alimentares. Antigamente a obesidade e a desnutrição eram enxergadas como consequências. Se a pessoa tinha câncer, por exemplo, era normal ela estar mais magra. Se a pessoa era obesa, o problema era somente o sedentarismo. Hoje, os dois são percebidos como doenças [5].

O objetivo deste trabalho é contribuir com os estudos sobre a alimentação e suas implicações nas ocorrências relacionadas a Covid-19, utilizando as técnicas multivariadas da Análise de Agrupamentos e do gráfico Biplot. Espera-se que a Análise de Agrupamentos permita identificar grupos homogêneos utilizando os países e as variáveis selecionadas, assim possibilitando avaliar esses resultados através do gráfico Biplot, verificando os relacionamentos entre os grupos e as variáveis.

2. Metodologia

2.1 Dados

Este trabalho utilizou como fonte de informações dados referentes a dieta, população e situação da Covid-19 referentes a 80 países da América e Europa, onde essas informações foram originalmente reunidas pela pesquisadora Maria Ren e todo o processo pode ser encontrado em [2]. Vale destacar que todos as informações foram obtidas de portais oficiais e que disponibilizam dados para que sejam utilizados de maneira livre e sem restrição.

Os dados sobre a dieta foram retirados de um estudo publicado pela *Food and Agriculture Organization of the United Nations* (FAO) que analisou a alimentação populacional durante o período de 2016 a 2017, classificando em 21 os tipos de produtos alimentícios. A descrição dos alimentos presentes em cada categoria está presente no Apêndice A. Para cada categoria foi estimada a ingestão média de alimentos (kg) da população e escalando a proporção em um intervalo de 0 a 100. Dela também foram retiradas as taxas para obesidade e desnutrição, que assim como as variáveis citadas anteriormente, variam em um intervalo de 0 a 100. A FAO é uma agência especializada das Nações Unidas que lidera os esforços nacionais ao combate à fome. Ela têm como objetivo alcançar a segurança alimentar para todos e garantir que as pessoas tenham acesso regular a alimentos de alta qualidade em quantidade suficiente para levar uma vida ativa e saudável. Com mais de 194 estados membros, a FAO trabalha em mais de 130 países em todo o mundo. Os dados foram obtidos diretamente da página oficial de informações providas pela FAO, a FAOSTAT, que libera acesso livre a dados sobre alimentação e agricultura sobre mais de 245 países e territórios, contemplando dados de 1961 até o ano da elaboração desta pesquisa.

A população foi utilizada para definir a taxa das informações referentes à Covid-19, que foi definida como sendo as quantidades por 100 habitantes. A informação referente à população dos países foi retirada do site da *Population Reference Bureau* (PRB) e refere-se ao ano de 2020. A PRB é uma organização de pesquisa privada e sem fins lucrativos especializada em conduzir e utilizar da análise de dados demográficos para repassar ao público indicadores sobre população, saúde e o meio ambiente e para envolver pesquisadores, formuladores de políticas e defensores com evidências sobre essas questões. A PRB tem vários projetos e iniciativas em andamento nos Estados Unidos e internacionalmente com parceiros nos setores de governo, organizações sem fins lucrativos, pesquisa, negócios e filantropia.

Por fim os valores referentes à Covid-19 no ano de 2020 foram obtidos da página CO-

RONAVIRUS RESOURCE CENTER (CRC) que pertence à *Center for Systems Science and Engineering*, que é um grupo de pesquisa alojado na Universidade Norte Americana *Johns Hopkins*. A CRC é uma fonte de dados sobre a Covid-19 continuamente atualizada e sob orientação especializada, que coleta e analisa dados disponíveis sobre casos, mortes, testes, hospitalizações e vacinas para ajudar o público, legisladores e profissionais de saúde em todo o mundo a responder à pandemia.

2.2 Softwares Utilizados

A tratativa dos dados foi realizada pelos *sofwares* Microsoft Excel (2018) e KNIME *Analytics Platform*, enquanto as análises, tabelas e gráficos apresentados foram realizadas utilizando o *sofware* R (versão 4.1.3).

2.3 Análise de Agrupamentos

A técnica multivariada de análise de agrupamentos, também conhecida como análise de conglomerados, classificação ou *cluster*, tem como objetivo descobrir os agrupamentos naturais das variáveis, onde estes são feitos com base nas similaridades ou dissimilaridades (caracterizadas por diversas formas de cálculo de "distâncias").

De acordo com [10], a análise de agrupamentos é uma maneira de se obter grupos homogêneos, por um esquema que possibilite reunir os dados em um determinado número de grupos, de modo que exista grande homogeneidade dentro de cada grupo e heterogeneidade entre eles.

O primeiro passo ao se realizar a análise de agrupamentos consiste na formulação do problema, definindo as variáveis sobre as quais se baseará o agrupamento, seguida pela coleta dos dados que deverão ser reunidos em uma tabela com i colunas (variáveis) e j linhas (elementos). Escolhe-se, então, uma medida apropriada de distância, que irá determinar o quão similares, ou dissimilares, são os indivíduos que estão sendo agrupados. Cada problema irá apresentar suas peculiaridades, e cabe ao pesquisador analisar suas possibilidades e escolher as melhores ferramentas e métodos para o problema estudado.

Uma questão importante refere-se ao critério a ser utilizado para se decidir até que ponto dois indivíduos do conjunto de dados podem ser considerados como semelhantes ou não. Para responder essa questão é necessário considerar medidas que descrevam a similaridade entre elementos amostrais de acordo com as variáveis que neles foram medidas. Ao considerar que para cada elemento amostral têm-se informações de p-variáveis armazenadas em um vetor, a comparação de diferentes elementos amostrais poderá ser feita através de medidas matemáticas (métricas), que possibilitem a comparação de vetores, como as medidas de distância. Assim, pode-se calcular as distâncias entre os vetores, de observações dos elementos amostrais e agrupar aqueles com menor distância [14].

2.4 MEDIDAS DE DISSIMILARIDADE

As distâncias são medidas utilizadas para a representação dos pontos na estrutura de similaridade. Esta medida representa o menor espaço entre dois pontos, sendo uma extensão do teorema de Pitágoras para o caso multidimensional.

O termo dissimilaridade surgiu em função da relação da distância entre dois pontos P e Q, definida como d(P,Q), pois, à medida que ela cresce, diz-se que a divergência entre os pontos (observações) P e Q aumenta, ou seja, tornam-se cada vez mais dissimilares. Os valores de distâncias são geralmente obtidos a partir de informações de n observações, mensurados em relação a p variáveis.

Atualmente, diversas medidas de dissimilaridade são propostos na literatura, principalmente devido ao grande desenvolvimento e utilização das técnicas multivariadas [11].

Considere que em um conjunto de n elementos amostrais foi observado p-variáveis aleatórias para cada um. Com o objetivo de agrupar esses elementos em g grupos foram formados os seguintes vetores para cada elemento j:

$$\mathbf{X}_j = [X_{1j}\ X_{2j}\ \ldots\ X_{pj}],\ j = 1, 2, ..., n \tag{2.1}$$

onde X_{ij} representa o valor observado da variável i medida no elemento j.

1. Distância Euclidiana: A distância Euclidiana entre dois elementos $\mathbf{X}_l$ e $\mathbf{X}_k$, l $\neq$ k, corresponde à raiz quadrada da soma dos quadrados das diferenças entre os pares de observações, portanto é definida por:

$$d(\mathbf{X}_l, \mathbf{X}_k) = [(\mathbf{X}_l - \mathbf{X}_k)'(\mathbf{X}_l - \mathbf{X}_k)]^{\frac{1}{2}} = [\sum_{i=1}^{p}(X_{il} - X_{ik})^2]^{\frac{1}{2}} \tag{2.2}$$

2. Distância Euclidiana Padronizada: A distância Euclidiana Padronizada segue a mesma estrutura da Distância Euclidiana, porém os dados são padronizados antes da aplicação do cálculo das distâncias. Considere $\overline{X_i}$ e S_i a média e o desvio-padrão amostral da variável i respectivamente, a transformação da observação X_{ij} em Z_{ij} ocorre de acordo com a seguinte fórmula:

$$Z_{ij} = \frac{X_{ij} - \overline{X_i}}{S_i} \tag{2.3}$$

3. Distância de Minkowsky: A Distância de Minkowsky entre dois elementos $\mathbf{X}_l$ e $\mathbf{X}_k$, l $\neq$ k, é definida por:

$$d(\mathbf{X}_l, \mathbf{X}_k) = [\sum_{i=1}^{p} w_i |X_{il} - X_{ik}|^{\lambda}]^{\frac{1}{\lambda}} \tag{2.4}$$

com w_i's sendo os pesos de ponderação para as variáveis. Para $\lambda = 1$ essa distância é

conhecida como *city-block* ou *Manhattan*, e para $\lambda = 2$ tem-se a distância Euclidiana. A Distância de Minkowsky é menos afetada por *outliers* do que a distância Euclidiana.

2.5 Métodos Hierárquicos Aglomerativos

Os métodos hierárquicos aglomerativos partem do princípio de que, no início do processo de agrupamento, as n observações a serem agrupadas formam n grupos distintos, e através de sucessivas fusões, geram um único conglomerado de n observações ao final do processo.

Segundo [9], uma das principais características de um processo hierárquico é que os resultados de um estágio anterior são sempre aninhados com os resultados de um estágio posterior, análogo a estrutura de uma árvore.

De acordo com [14] os passos principais para aplicação das técnicas aglomerativas podem ser resumidas da seguinte forma:

1. Cada elemento constitui um *cluster* unitário. Portanto tem-se n *clusters*;

2. Em cada estágio do algorítimo de agrupamento, os pares de conglomerados mais "similares" são combinados e passam a constituir um único conglomerado. Apenas um único conglomerado pode ser formado em cada passo. Dessa forma, em cada estágio do processo, o número de conglomerados diminui;

3. *Propriedade de hierarquia*. Em cada estágio do algorítimo, cada novo conglomerado formado é um agrupamento de conglomerados formados nos estágios anteriores. Se duas observações aparecem juntas num mesmo *cluster* em algum estágio do processo de agrupamento, elas aparecerão juntas em todos os estágios subsequentes, ou seja, uma vez unidos estes elementos não poderão ser separados;

4. Devido à propriedade de hierarquia, pode-se construir um gráfico chamado dendrograma que tem como objetivo apresentar o arranjo entre os objetos em uma escala de distância. Esse gráfico tem a forma de árvore no qual a escala vertical indica o nível de similaridade (ou dissimilaridade). No eixo horizontal, são marcados os elementos amostrais numa ordem conveniente relacionada à história de agrupamento. As linhas verticais, partindo dos elementos amostrais agrupados, têm altura correspondente ao nível em que os elementos foram considerados semelhantes, isto é, a distância do agrupamento ou o nível de similaridade.

Não exite um critério objetivo para determinar o número de grupos em que o conjunto de dados deve ser repartido. O critério mais simples utilizado é pela análise gráfica dos *clusters* formados, o que torna esse procedimento naturalmente enviesado pelas necessidades e opiniões dos analistas e pesquisadores, porém existem ferramentas que auxiliam a encontrar um número ótimo de grupos, como a Estatística Pseudo-F que será apresentada posteriormente.

Existem vários métodos de agrupamentos hierárquicos. A seguir, serão apresentados os métodos mais comuns e disponíveis na grande maioria dos *softwares* estatísticos.

1. Método da Ligação Simples (*Single Linkage*): Neste método a distância entre dois grupos é determinada pela distância mínima entre os pares de elementos destes grupos, e aqueles que possuem a menor distância mínima são agrupados. Para exemplificar, sejam os grupos $\mathbf{C}_1 = \{\mathbf{X}_1, \mathbf{X}_3, \mathbf{X}_7\}$ e $\mathbf{C}_2 = \{\mathbf{X}_2, \mathbf{X}_6\}$, a distância entre os grupos será definida por:

$$d(\mathbf{C}_1, \mathbf{C}_2) = min\{d(\mathbf{X}_l, \mathbf{X}_k, l \neq k, l = 1, 3, 7 \; e \; k = 2, 6)\} \tag{2.5}$$

2. Método da Ligação Completa (*Complete Linkage*): Neste método a distância entre dois grupos é determinada pela distância máxima entre os pares de elementos destes grupos. O método busca agrupar elementos cuja distância entre os mais afastados seja a menor. Para exemplificar, sejam os grupos $\mathbf{C}_1 = \{\mathbf{X}_1, \mathbf{X}_3, \mathbf{X}_7\}$ e $\mathbf{C}_2 = \{\mathbf{X}_2, \mathbf{X}_6\}$, a distância entre os grupos será definida por:

$$d(\mathbf{C}_1, \mathbf{C}_2) = max\{d(\mathbf{X}_l, \mathbf{X}_k, l \neq k, l = 1, 3, 7 \; e \; k = 2, 6)\} \tag{2.6}$$

3. Método da Ligação Média (*Average Linkage* ou UPGMA): Este método utiliza a média da distância entre todos os pares que podem ser formados com os elementos dos dois grupos que estão sendo comparados e agrupa aqueles que possuem a menor distância média. Para exemplificar, sejam os grupos $\mathbf{C}_1$ com n_1 elementos e $\mathbf{C}_2$ com n_2, a distância entre eles será dada por:

$$d(\mathbf{C}_1, \mathbf{C}_2) = \sum_{l \in \mathbf{C}_1} \sum_{k \in \mathbf{C}_2} (\frac{1}{n_1 n_2}) d(\mathbf{X}_l, \mathbf{X}_k) \tag{2.7}$$

4. Método do Centroide (*Centroid Method*): No método do Centroide, a distância entre dois grupos é definida como sendo a distância entre os vetores de médias, ou centroides dos grupos que estão sendo comparados, priorizando a menor distância entre eles. Para exemplificar, sejam os grupos $\mathbf{C}_1 = \{\mathbf{X}_1, \mathbf{X}_3, \mathbf{X}_7\}$, com seu vetor de médias dado por $\overline{\mathbf{X}}_1 = \frac{1}{3}[\mathbf{X}_1 + \mathbf{X}_3 + \mathbf{X}_7]$ e $\mathbf{C}_2 = \{\mathbf{X}_2, \mathbf{X}_6\}$, com seu vetor de médias dodo por $\overline{\mathbf{X}}_2 = \frac{1}{2}[\mathbf{X}_2 + \mathbf{X}_6]$, a distância entre os grupos será definida por:

$$d(\mathbf{C}_1, \mathbf{C}_2) = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)'(\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) \tag{2.8}$$

5. Método de Ward (*Ward Method*): O método de Ward agrupa os elementos que possuem a menor soma de quadrados das distâncias. Trata-se de um método que tende a proporcionar agregados com aproximadamente o mesmo número de observações [8]. Inicialmente cada elemento é considerado um único agrupamento e em cada passo do algorítimo calcula-se a soma de quadrados dentro de cada conglomerado (quadrado da distância Euclidiana) de cada elemento pertencente ao conglomerado em relação ao correspondente vetor de médias do conglomerado, isto é,

$$\mathbf{SS}_i = \sum_{j=l}^{n_i} (\overline{\mathbf{X}}_{ij} - \overline{\mathbf{X}}_{i.})' (\overline{\mathbf{X}}_{ij} - \overline{\mathbf{X}}_{i.}) \tag{2.9}$$

em que n_i é o número de elementos no conglomerado $\mathbf{C}_i$ no passo k do processo, $\mathbf{X}_{ij}$ é o vetor de observações do j-*ésimo* elemento do i-*ésimo* conglomerado, $\overline{\mathbf{X}}_{i.}$ é o centroide do conglomerado $\mathbf{C}_i$, e $\mathbf{SS}_i$ é a soma de quadrados correspondente ao conglomerado $\mathbf{C}_i$. No passo k, a soma de quadrados total dentro dos grupos é definida como:

$$SSR = \sum_{i=l}^{g_k} \mathbf{SS}_i \tag{2.10}$$

sendo g_k o número de grupos existentes quando se está no passo k.
A distância entre os conglomerados $\mathbf{C}_l$ e $\mathbf{C}_i$ será definida por:

$$d(\mathbf{C}_l, \mathbf{C}_i) = [\frac{n_l n_i}{n_l + n_i}](\overline{\mathbf{X}}_{l.} - \overline{\mathbf{X}}_{i.})' (\overline{\mathbf{X}}_{l.} - \overline{\mathbf{X}}_{i.}) \tag{2.11}$$

que representa a soma de quadrados entre os *clusters* $\mathbf{C}_l$ e $\mathbf{C}_i$. Em cada passo do algorítimo de agrupamento, os dois conglomerados que minimizam a distância 2.11 são combinados.

2.6 Dendrograma

Depois de aplicado algum método hierárquico à matriz de distâncias, os agrupamentos podem ser representados de maneira bidimensional através de um Dendrograma (diagrama bidimensional em forma de árvore). Nele estão dispostos linhas ligadas segundo os níveis de similaridade, que agrupará pares de indivíduos ou de variáveis [7]. A Figura 2.1 considera um exemplo com 6 observações e apresenta a estrutura de um Dendrograma.

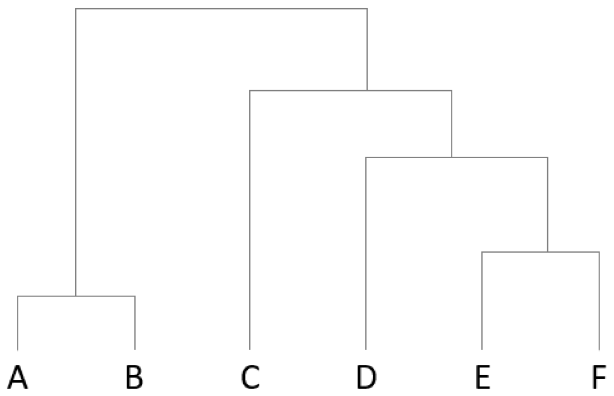

Figura 2.1: Exemplo de Dendrograma

O Dendrograma é um diagrama que apresenta a relação hierárquica entre os elementos em uma escala de distância, ilustrando as fusões ou partições efetuadas em cada estágio do

algorítimo de agrupamento, no qual o eixo das abscissas representa os indivíduos e o eixo das ordenadas as distâncias obtidas após a utilização de uma metodologia de agrupamento. Como esse processo constrói uma representação simplificada em duas dimensões de uma relação que originalmente é n-dimensional, ocorre o aparecimento de distorções quanto a similaridade, e para medir essa distorção deve-se calcular o Coeficiente de Correlação entre a matriz inicial de distâncias e a matriz derivada do dendrograma, esse cálculo é conhecido como Coeficiente de Correlação Cofenético.

2.7 Coeficiente de Correlação Cofenético (CCC)

Para a escolha do método de ligação mais adequado aos dados é utilizado o Coeficiente de Correlação Cofenético, uma medida equivalente à Correlação de Pearson, que é usado para medir o grau de ajuste entre a matriz original dos coeficientes de distâncias (matriz fenética **F**) e a matriz resultante do processo de agrupamento (matriz cofenética **C**). Vários autores recomendam um CCC acima de 70% e ele é calculado da seguinte forma:

$$CCC = \frac{\widehat{Cov}(\mathbf{F},\mathbf{C})}{\sqrt{\widehat{V}(\mathbf{F})\widehat{V}(\mathbf{C})}} = \frac{\sum_{i=1}^{n-1}\sum_{j=2}^{n}(f_{ij}-\overline{f})-(c_{ij}-\overline{c})}{\sqrt{[\sum_{i=1}^{n-1}\sum_{j=2}^{n}(f_{ij}-\overline{f})^2][\sum_{i=1}^{n-1}\sum_{j=2}^{n}(c_{ij}-\overline{c})^2]}} \tag{2.12}$$

em que f_{ij} é a medida de similaridade entre os indivíduos **i** e **j** obtidos da matriz fenética, c_{ij} é a medida de similaridade entre os indivíduos **i** e **j** obtidos da matriz cofenética, e

$$\overline{f} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=2}^{n} f_{ij}; \tag{2.13}$$

$$\overline{c} = \frac{2}{n(n-1)}\sum_{i=1}^{n-1}\sum_{j=2}^{n} c_{ij}. \tag{2.14}$$

2.8 Estatística Pseudo F

Segundo [13] o critério mais simples utilizado para decidir qual o número de grupos a adotar é o corte do dendrograma pela análise subjetiva dos diferentes níveis do mesmo, o que torna esse procedimento naturalmente enviesado pelas necessidades e opiniões dos analistas e pesquisadores. Com essa ideia convém a utilização de técnicas para encontrar um ótimo ponto de corte para o dendrograma.

Sugere-se que para cada passo do agrupamento, o cálculo da estatística Pseudo F, definida por [6]:

$$F = \frac{\frac{SSB}{(g^*-1)}}{\frac{SSR}{(n-g^*)}} = (\frac{n-g^*}{g^*-1})\frac{R^2}{1-R^2} \tag{2.15}$$

em que SSB é a soma de quadrados total entre os g^* grupos da partição, dada por:

$$SSB = \sum_{i=1}^{g^*} n_i(\overline{\mathbf{X}}_i - \overline{\mathbf{X}})'(\overline{\mathbf{X}}_i - \overline{\mathbf{X}}) \tag{2.16}$$

SSR é a soma de quadrados total dentro dos grupos da partição (Soma de Quadrados Residual), dada por:

$$SSR = \sum_{i=1}^{g^*} SS_i = \sum_{i=1}^{g^*}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i)'(\mathbf{X}_{ij} - \overline{\mathbf{X}}_i). \tag{2.17}$$

R^2 é o coeficiente da partição, dado por:

$$R^2 = \frac{SSB}{SST_c} \tag{2.18}$$

STT_c é a soma de quadrados total corrigida para a média global em cada variável, dado por:

$$STT_c = \sum_{i=1}^{g^*} SS_i = \sum_{i=1}^{g^*}\sum_{j=1}^{n_i}(\mathbf{X}_{ij} - \overline{\mathbf{X}})'(\mathbf{X}_{ij} - \overline{\mathbf{X}}). \tag{2.19}$$

g^* é o número de grupos relacionado com a partição do respectivo estágio de agrupamento; n é o número de elementos amostrais; $\mathbf{X}_{ij}^{'} = (X_{i1j}X_{i2j}...X_{ipj})$ é o vetor de medidas observadas para o j-ésimo elemento amostral do i-ésimo elemento amostral do i-ésimo grupo; $\overline{\mathbf{X}}_i^{'} = (\overline{\mathbf{X}}_{i1}\overline{\mathbf{X}}_{i2}...\overline{\mathbf{X}}_{ip})$ é o vetor de médias do i-ésimo grupo; $\overline{\mathbf{X}}^{'} = (\overline{\mathbf{X}}_1\overline{\mathbf{X}}_2...\overline{\mathbf{X}}_p)$ é o vetor de médias global, sem levar em conta qualquer posição; onde $\overline{\mathbf{X}}_l = \frac{1}{n}\sum_{i=1}^{g^*}\sum_{j=1}^{n_i} X_{ilj}, l = 1, 2, ..., p$.

Segundo os autores, se F é monotonicamente crescente com g^*, os dados sugerem que não existe qualquer estrutura natural de partição dos dados, porém, se isso não ocorrer e a função F apresentar um valor de máximo, o número de conglomerados e a partição referente a esse valor máximo corresponderão à partição ideal dos dados.

2.9 Biplot

O Biplot é uma representação gráfica de dados multivariados, onde os elementos de uma matriz de dados são representados de acordo com pontos e vetores associados às linhas e colunas de uma matriz. Foi criado por K. Ruben Gabriel em 1971 e segundo [12] pode ser usado para mostrar a relação existente entre variáveis, entre observações e entre variáveis e observações.

O Biplot é baseado nos mesmos princípios de técnicas de Análise de Componentes Principais para redução da dimensionalidade, onde a maior diferença é que, neste caso, serão representados as observações e as variáveis, obtendo uma representação entre os componentes principais e as coordenadas principais. Sua interpretação é baseada em conceitos geométricos intuitivos para o usuário, facilitando a sua compreensão.

Para a construção de um Biplot bidimensional, considere os dados organizados de acordo com a seguinte tabela de dupla entrada:

Tabela 2.1: Exemplo de Tabela de Dupla Entrada

		Variável Y						
		1	2	.	.	.	q	Total
	1	n_{11}	n_{12}	.	.	.	n_{1q}	n_1
	2	n_{21}	n_{22}	.	.	.	n_{2q}	n_2
Variável X	.	.						.
	.	.						.
	.	.						.
	p	n_{p1}	n_{p2}	.	.	.	n_{pq}	n_p
Total		n_1	n_2	.	.	.	n_q	n

Seja uma matriz $\mathbf{P}_{i\times j}$, tal que suas coordenadas são formadas por:

$$\mathbf{p}_{ij} = \frac{n_{ij}}{n} \tag{2.20}$$

defina a matriz de perfil das linhas como a matriz diagonal ($\mathbf{D}_r$) dada por:

$$\mathbf{r}^{'} = (\frac{n_1}{n}\ \frac{n_2}{n}\ \ldots\ \frac{n_p}{n}) \tag{2.21}$$

e defina a matriz de perfil das colunas como a matriz diagonal ($\mathbf{D}_c$) dada por:

$$\mathbf{c}^{'} = (\frac{n_1}{n}\ \frac{n_q}{n}\ \ldots\ \frac{n_p}{n}) \tag{2.22}$$

Considere a matriz $\mathbf{Q}$ = $\mathbf{P}$ - $\mathbf{rc}^{'}$. Os termos representam uma comparação da proporção observada com aquela esperada sob um modelo no qual as variáveis X e Y são independentes.

Com isso, pode-se decompor a matriz $\mathbf{Q}$ em seus autovalores e autovetores:

$$\mathbf{Q} = \mathbf{A}\Lambda\mathbf{B}^{'} \tag{2.23}$$

em que, $\mathbf{A} = \mathbf{D}_r^{\frac{1}{2}}\mathbf{U}$; $\mathbf{B} = \mathbf{D}_c^{\frac{1}{2}}\mathbf{V}$; Λ contêm os autovalores de $\mathbf{Q}$; $\mathbf{U}$ contêm os autovetores de $\mathbf{QQ}^{'}$ e $\mathbf{V}$ contém os autovetores de $\mathbf{Q}^{'}\mathbf{Q}$.

As coordenadas principais das linhas da matriz $\mathbf{P}$ são definidas como:

$$\mathbf{Y}_{p\times k} = \mathbf{D}_r^{-1}\mathbf{A}\Lambda \tag{2.24}$$

As coordenadas principais das colunas da matriz $\mathbf{P}$ são definidas como:

$$\mathbf{Z}_{q\times k} = \mathbf{D}_c^{-1}\mathbf{B}\Lambda \tag{2.25}$$

As coordenadas principais darão origem ao Gráfico Biplot, utilizando o vetor λ de autovalores de $\mathbf{P}$ pode-se calcular a proporção explicada pela i-ésima coordenada principal, dada por:

$$\mathbf{PE}_i = \frac{\lambda_i^2}{\sum_{i=1}^{k} \lambda_i^2} \tag{2.26}$$

A Figura 2.2 a seguir, considera um exemplo com 4 observações, 6 variáveis e uma proporção explicada de 86,6% para as duas primeiras dimensões, apresentando a visão de um Biplot. Sua forma de interpretação é através do cosseno do ângulo formado pelos vetores que ligam dois pontos e que parte da origem do gráfico, onde um ângulo agudo indica uma associação positiva, um ângulo obtuso indica uma associação negativa e um ângulo reto ou pontos muito próximos da origem indica ausência de associação.

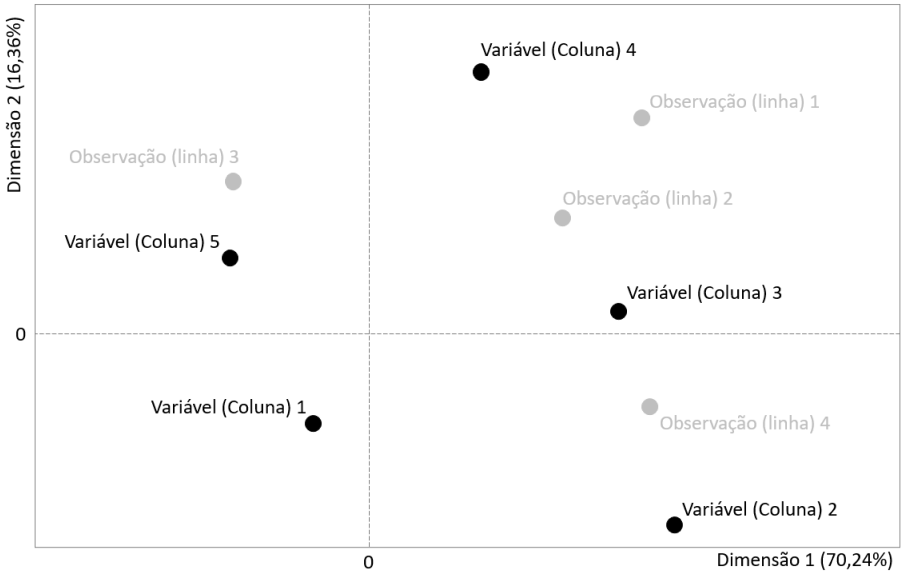

Figura 2.2: Exemplo de Biplot

3. RESULTADOS

Os elementos amostrais do estudo são 80 países da América e Europa, e as variáveis disponíveis são as proporções da dieta para 21 tipos de produtos alimentícios, taxas de obesidade e desnutrição da população, todas pertencentes ao intervalo de 0 a 100, juntamente com as taxas sobre a Covid-19 de Confirmados e Mortos e Recuperados/Ativos a cada 100 habitantes.

Durante a tratativa dos dados foi observado que a variável desnutrição possuía valores ausentes para muitos indivíduos, portanto a ação tomada foi retirar essa variável do estudo. A variável Recuperados/Ativos também foi removida, pois ela representa uma combinação linear entre as variáveis Confirmados e Mortes, portanto não haveria necessidade de mantê-la no modelo.

Para caracterizar as variáveis em estudo, foi realizado uma análise descritiva presente na Tabela 3.1. Analisando apenas as variáveis relacionadas aos produtos alimentícios é possível observar que as variáveis com maiores médias são Frutas (12,2182), Cereais (16,8635) e Leite (18,9774), além disso Legumes (6,1532), Frutas (6,4126), e Leite (8,9119) possuem os maiores para o desvio padrão, portanto elas também possuem uma maior dispersão dos dados em torno da média. Por outro lado, as variáveis com menores médias são Produtos aquáticos (0,0037), Culturas de açúcar (0,0340), e temperos (0,1371). Miúdos (0,2004) se junta com Produtos aquáticos (0,0097) e Temperos (0,2013) entre os produtos alimentícios que possuem os menores valores para o desvio padrão, logo, uma menor dispersão dos dados em torno da média.

Tabela 3.1: Análise descritiva das variáveis

Variável	Média	Desvio Padrão	Mínimo	Máximo
Açúcar e adoçantes	6,7575	2,7921	2,8250	16,2202
Bebidas alcoólicas	8,0296	3,6957	0,8028	19,7000
Carne	8,2827	2,7439	3,5079	16,3354
Cereais	16,8635	6,0754	6,8028	33,6033
Culturas de açúcar	0,0340	0,2363	0,0000	2,0222
Diversos	0,9747	1,4566	0,0000	7,3257
Estimulantes	0,4983	0,3931	0,0100	2,5640
Frutas	12,2182	6,4126	4,8010	38,6056
Gorduras animais	0,7052	0,6786	0,0124	2,7122
Legumes	12,2177	6,1532	3,8957	33,4037

Contínua na próxima página

Tabela 3.1 – *Continuação da página anterior*

Variável	Média	Desvio Padrão	Mínimo	Máximo
Leguminosas	0,7316	0,8271	0,0020	4,0133
Leite	18,9774	8,9119	1,7529	41,6755
Miúdos	0,3467	0,2004	0,0104	0,9348
Nozes	0,2253	0,2628	0,0000	1,4167
Oleaginosos	0,7043	0,7455	0,0232	3,6725
Óleos vegetais	1,6086	0,7059	0,1830	3,0471
Ovos	1,2051	0,4764	0,1050	2,9352
Peixes	2,3673	1,8401	0,4040	9,2240
Produtos aquáticos	0,0037	0,0097	0,0000	0,0656
Raízes com amido	7,1116	4,4379	1,3590	25,0560
Temperos	0,1371	0,2013	0,0016	1,2094
Obesidade	24,2825	3,7112	18,7000	37,3000
Confirmados	2,6702	1,9580	0,0593	7,8117
Mortos	0,0544	0,0418	0,0000	0,1700

Após a análise descritiva foi realizado um estudo para decidir quais variáveis, medida de distância e método de agrupamento serão utilizadas para a construção do dendrograma. Foi decidido que as variáveis relacionadas à Covid-19 não entrariam no modelo, mas sim seriam inseridas após a construção dos conglomerados, com o objetivo de analisar os grupos formados apenas com variáveis relacionadas a dieta e verificar se existe alguma relação entre elas e essas novas variáveis externas ao modelo.

Com as 21 variáveis relacionadas aos tipos de produtos alimentícios e a variável Obesidade, o próximo passo foi a utilização de um algorítimo que avaliaria quais são aquelas que mais contribuem para explicar a variabilidade dos dados originais, gerando, para cada medida de distância e método de agrupamento, o modelo com menor número de variáveis, que apresenta a maior correlação cofenética e alta similaridade com as distâncias originais. Mais informações sobre este algorítimo podem ser encontradas em [1].

A Tabela 3.2 apresenta o CCC resultante da combinação entre as medidas de distância (Euclidiana e Manhattan) e métodos de agrupamento (Ligação Simples, Ligação Completa, Ligação Média, Centroide e Ward). Nota-se que a Distância Euclidiana Padronizada não está presente, dado que todas as variáveis já estão na mesma escala não se provou necessário padronizar as variáveis antes da aplicação da Distância Euclidiana.

Tabela 3.2: CCC por Distância e Agrupamento

Combinação de Métodos	CCC
Distância Euclidiana e Método da Ligação Média	0,7532
Distância de Manhattan e Método da Ligação Média	0,7335

Contínua na próxima página

Tabela 3.2 – *Continuação da página anterior*

Combinação de Métodos	CCC
Distância Euclidiana e Método da Ligação Simples	0,7059
Distância de Manhattan e Método da Ligação Simples	0,6985
Distância Euclidiana e Método do Centróide	0,6919
Distância de Manhattan e Método do Centróide	0,6852
Distância Euclidiana e Método da Ligação Completa	0,6668
Distância de Manhattan e Método da Ligação Completa	0,6602
Distância de Manhattan e Método de Ward	0,6277
Distância Euclidiana e Método de Ward	0,6180

A combinação da Distância Euclidiana com o Método da Ligação Média obteve o maior valor para o CCC, que foi de 0,7532. Neste modelo as variáveis selecionadas foram Cereais, Frutas, Legumes, Leite, Raízes com amido e Obesidade.

Com a medida de distância, o método de agrupamento e as variáveis a serem utilizadas, o próximo passo seria obter o número ótimo de grupos. O critério utilizado para auxiliar a encontrar esse valor foi a Estatística Pseudo F, cujos resultados são apresentados na Figura 3.1. Foi calculado o valor do *Score* para cada número de grupos em um intervalo de 2 a 20, o número que obteve o valor máximo para o *Score* (16,6932) e por consequência é o número ótimo de grupos foi o valor 6, portanto esse será o valor utilizado neste estudo.

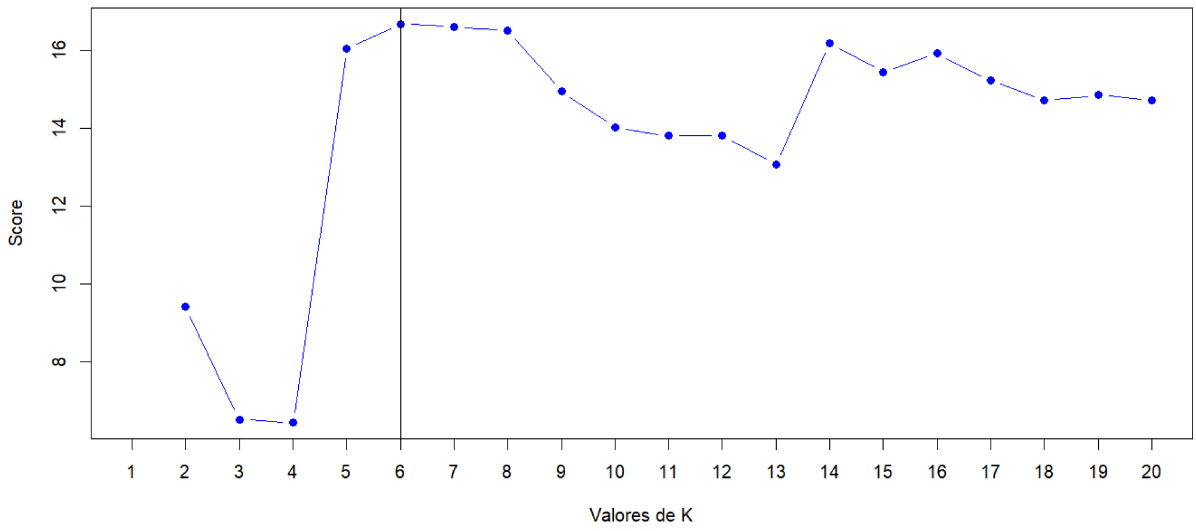

Figura 3.1: Estatística Pseudo F para o número ótimo de grupos

A Figura 3.2 apresenta o Dendrograma e a Tabela 3.3 caracteriza os Grupos formados pela combinação da Distância Euclidiana com o Método da Ligação Média. Foi adotada o Código ISO 3166-1 alfa-3 para se referenciar aos países, a relação entre código e descrição de cada país está presente no Apêndice B. O Grupo 1 é aquele com mais indivíduos, possuindo 47 países, o

Grupo 2 possui 15, o Grupo 3 possui 3, o Grupo 4 possui 11, o Grupo 5 possui 3 e o Grupo 6 é um grupo unitário.

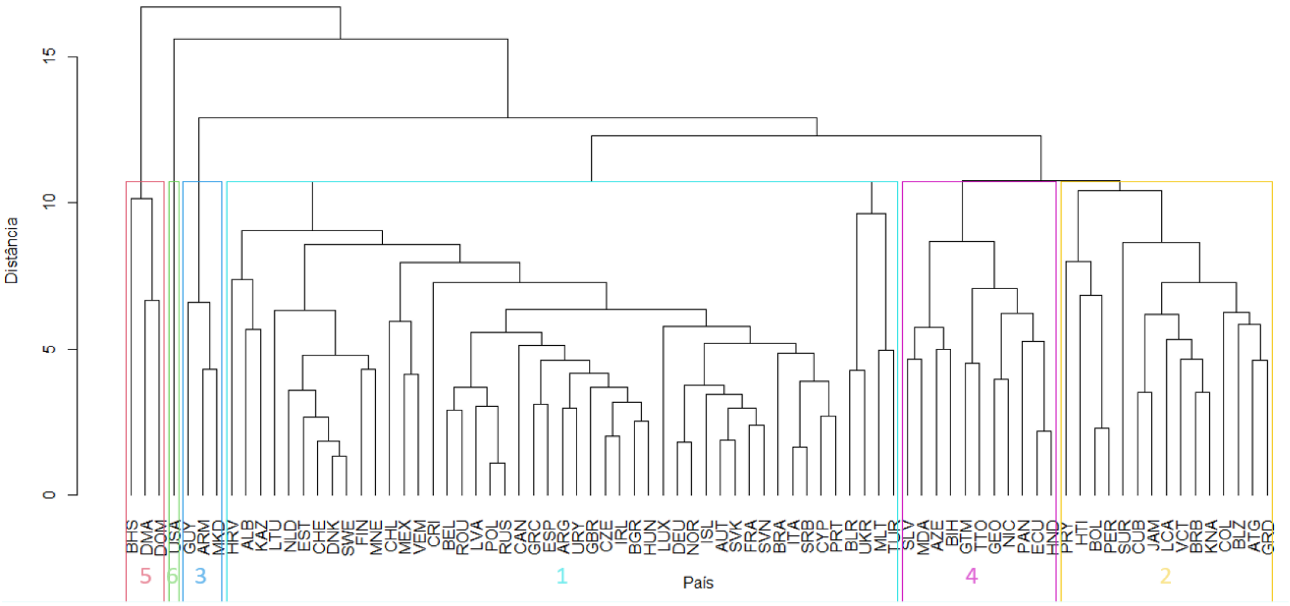

Figura 3.2: Dendrograma referente a aplicação da Distância Euclidiana em conjunto do Método da Ligação Média

Tabela 3.3: Grupos formados pelo Dendrograma da Figura 3.2

Grupo 1	ALB; ARG; AUT; BEL; BGR; BLR; BRA; CAN; CHE; CHL; CRI; CYP; CZE; DEU; DNK; ESP; EST; FIN; FRA; GBR; GRC; HRV; HUN; IRL; ISL; ITA; KAZ; LTU; LUX; LVA; MEX; MLT; MNE; NLD; NOR; POL; PRT; ROU; RUS; SRB; SVK; SVN; SWE; TUR; UKR; URY; VEM
Grupo 2	ATG; BLZ; BOL; BRB; COL; CUB; GRD; HTI; JAM; KNA; LCA; PER; PRY; SUR VCT
Grupo 3	ARM; GUY; MKD
Grupo 4	AZE; BIH; ECU; GEO; GTM; HND; MDA; NIC; PAN; SLV; TTO
Grupo 5	BHS; DMA; DOM
Grupo 6	USA

As médias das variáveis que deram origem ao Dendrograma e das variáveis externas referentes à Covid-19 para cada grupo são apresentados na Tabela 3.4. Pode-se destacar que os Grupos 2 e 5, que possuem as menores médias para as variáveis relacionas à Covid-19, também possuem as menores médias para Legumes e Leite, e as maiores médias para Frutas. O Grupo 6 formado pelos Estados Unidos (USA) possui as maiores médias para Leite (25,7283), Obesidade (37,3000) e para as variáveis da Covid - 19, em conjunto com a menor média para Frutas (9,0853). O Grupo 3 possui a segunda maior média para as variáveis da Covid-19 e a maior

média para Legumes (30,9697), além da menor média para Raízes com amido (5,0343). No Apêndice C é possível encontrar os resultados de uma análise descritiva contendo a quantidade, média, desvio padrão, mínimo e máximo de cada grupo formado.

Tabela 3.4: Médias para os grupo formados

Variável	$\overline{X}_{Grupo\ 1}$	$\overline{X}_{Grupo\ 2}$	$\overline{X}_{Grupo\ 3}$	$\overline{X}_{Grupo\ 4}$	$\overline{X}_{Grupo\ 5}$	$\overline{X}_{Grupo\ 6}$
Cereais	14,4857	18,7875	16,3158	27,1257	9,3658	11,0083
Frutas	9,4551	17,9893	11,3212	10,7406	34,0107	9,0853
Legumes	12,4265	9,3130	30,9697	10,8220	10,0931	11,4484
Leite	24,0320	7,4869	15,6474	16,2073	8,4782	25,7283
Raízes com amido	6,7280	10,1735	5,0343	5,4077	6,6147	5,6773
Obesidade	25,4681	22,0133	21,3333	20,6273	29,0667	37,3000
Confirmados	3,2538	0,9570	3,4122	2,3843	1,2564	6,1019
Mortos	0,0629	0,0278	0,0791	0,0518	0,0222	0,1054

Utilizando os grupos e as médias das variáveis apresentadas anteriormente foi realizado a construção do gráfico Biplot que se encontra na Figura 3.3, na qual a soma das duas primeiras dimensões (componentes principais) explicam 75,62% da variância total dos dados.

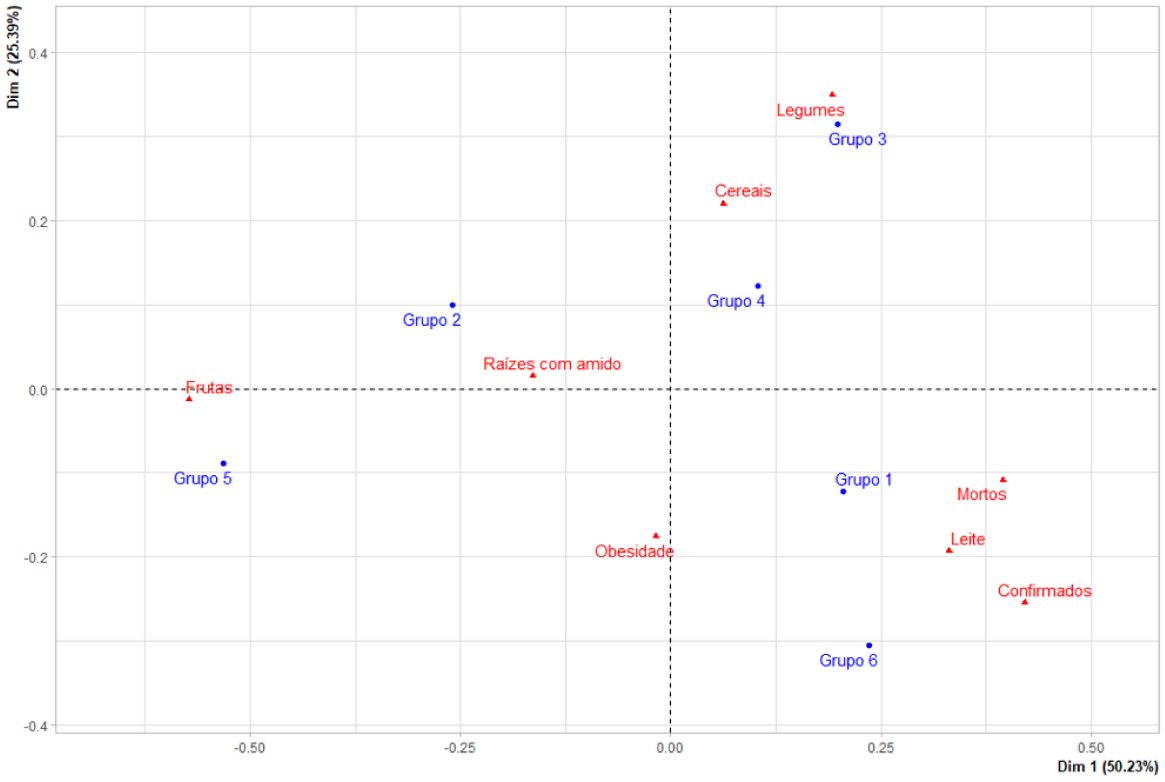

Figura 3.3: Biplot das médias das variáveis e dos grupos originados da clusterização

Através das possíveis análises que a Figura 3.3 apresenta, pode-se destacar que as variáveis Legumes e Cereais estão altamente associadas positivamente com os Grupos 3 e 4, as variáveis Confirmados, Mortos e Leite estão altamente associadas positivamente com os Grupos 1 e 6,

a variável Obesidade apresenta uma maior associação positiva com o Grupo 6 e as variáveis Raízes com amido e Frutas apresentam uma maior associação positiva com os Grupos 2 e 5. Também é possível observar uma associação negativa entre Legumes, Cereais, Grupo 1, Grupo 4 e a variável Obesidade. Confirmados, Mortos, Leite, Grupo 1 e Grupo 2 apresentam uma associação negativa com Frutas, Raízes com amido, Grupo 2 e Grupo 5.

Destacando-se as variáveis Confirmados e Mortos, que são os indicadores da Covid-19, observa-se que elas possuem uma alta associação positiva com a variável Leite, portanto elas propendem a assumir uma mesma tendência de aumento ou diminuição. As variáveis Frutas e Raízes com amido formam um ângulo obtuso com as variáveis Confirmados e Mortos, portanto sua relação é oposta. Por fim, as variáveis Obesidade, Cereais e Legumes apresentam uma ausência de associação com as variáveis sobre a Covid-19.

4. Conclusões

O presente estudo permitiu concluir que a Análise de Agrupamentos utilizando a Distância Euclidiana e o Método Hierárquico Aglomerativo da Ligação Média se mostrou eficaz em agrupar os 80 países da América e Europa em 6 grupos com respeito as variáveis da dieta. Os grupos 3 e 6 foram os grupos com os maiores valores para as variáveis sobre a Covid-19, enquanto os grupos 2 e 5 possuem os menores valores.

Foi possível verificar que dentre as 22 variáveis inicias relacionadas a dieta, apenas 6 se provaram necessárias para a construção do modelo utilizado para a Análise de Agrupamentos, onde a relação entre essas variáveis e aquelas relacionadas aos indicadores da Covid-19 se comportam da seguinte maneira: Leite possui associação positiva, Raízes com amido e Frutas possuem associação negativa e Legumes, Cereais e Obesidade não apresentam associação.

Referências Bibliográficas

[1] *clustering: Clustering analysis*. https://www.rdocumentation.org/packages/metan/versions/1.16.0/topics/clustering, acessado em 05/01/2022.

[2] *COVID-19 Dataset Challenge*. https://github.com/mariarencode/COVID_19_Dataset_Challenge, acessado em 01/09/2021.

[3] Alzerina, M.: *Nutrição e COVID-19*, 2020. https://covid19.cv/wp-content/uploads/2020/04/COVID-19-e-Nutri%C3%A7%C3%A3o-INSP.pdf, acessado em 10/08/2017.

[4] Brasil, G. F. do: *Como se proteger?*, 2021. https://www.gov.br/saude/pt-br/coronavirus/como-se-proteger, acessado em 03/08/2021.

[5] Brasil, S.: *Obesidade e desnutrição: nem tudo é o que parece*, 2018. https://saudebrasil.saude.gov.br/ter-peso-saudavel/obesidade-e-desnutricao-nem-tudo-e-o-que-parece, acessado em 04/08/2021.

[6] Calinski, T. e Harabasz, J.: *A Dendrite Method for Cluster Analysis*. Communications in Statistics, 3ª ed., 1974.

[7] Everitt, B.: *Cluster analysis*. London: Social Science Research Council/ Halsted Press, 2ª ed., 1993.

[8] Fávelo, L. P., Belfiore, P., Silva, F. L. e Chan, B. L.: *Análise de dados: modelagem multivariada para tomada de decisões*. Elsevier, 8ª ed., 2009.

[9] Hair, J. F.: *Análise multivariada de dados*. Bookman, 5ª ed., 2005.

[10] Johnson, R. A. e Wichern, D. W.: *Applied multivariate statistical analysis*. Englewood Cliffs: Prentice Hall, 4ª ed., 1998.

[11] Khattree, R. e Naik, D. N.: *Multivariate data reduction and discrimination with SAS software*. SAS Institute, 1ª ed., 2000.

[12] Lipkovish, I. A. e Smith, E. P.: *Biplot and Singular Value Decomposition Macros for Excel*. Journal of Statistica Software, [S. l.], v. 7, n. 5, p. 1–15, 2002.

[13] Martins, M. R., Pedro, S. e Rosa, S.: *Escolha do número de grupos e validação da solução em análise classificatória: da teoria à prática*. 2004. https://run.unl.pt/handle/10362/7686, acessado em 30/10/2021.

[14] Mingoti, S. A.: *Análise de dados através de métodos de estatística multivariada: uma abordagem aplicada*. Editora UFMG, 1ª ed., 2005.

[15] Organization, W. H.: *Coronavirus*, 2021. https://www.who.int/health-topics/coronavirus#tab=tab_1, acessado em 03/08/2021.

[16] Times, N. Y.: *A Timeline of the Coronavirus Pandemic*, 2021. https://www.nytimes.com/article/coronavirus-timeline.html, acessado em 03/08/2021.

A. Produtos Alimentícios

Tabela A.1: Descrição dos Produtos Alimentícios

Categoria	Itens
Açúcar e adoçantes	açúcar; açúcar não centrífugo; mel outros adoçantes
Bebidas Alcoólicas	Álcool não alimentar; bebidas alcoólicas; bebidas fermentadas; cerveja; vinho
Carne	Carne bovina; carne de aves; carne de carneiro e cabra; carne de porco
Cereais	Arroz; aveia; outros cereais; produtos de centeio; produtos de cevada; produtos de milho; produtos de painço; produtos de sorgo; produtos de trigo
Culturas de Açúcar	Beterraba sacarina; cana de açúcar
Diversos	Alimentos infantis; diversos
Estimulantes	Chá (incluindo mate); produtos de cacau; produtos de café
Frutas	Banana da terra; bananas; cítricos; outras frutas; produtos de laranjas, mandarinas e abacaxis; produtos de limões e limas; produtos de maçãs; produtos de toranja; produtos de uvas (exceto vinho)
Gorduras animais	Creme; gorduras de animais; manteiga; óleo corporal de peixe; óleo de fígado de peixe
Legumes	Cebolas; outros legumes; produtos de tomates
Leguminosas	Ervilhas; feijões; produtos de leguminosas
Leite	Leite (Excluindo Manteiga)
Miúdos	Miudezas comestíveis
Nozes	Produtos de nozes
Oleaginosas	Amendoins (sem casca); azeitonas (incluindo conservas); cocos; colza e semente de mostarda; grãos de soja; outras oleaginosas; semente de algodão; semente de gergelim; semente de girassol; sementes de palma

Contínua na próxima página

Tabela A.1 – *Continuação da página anterior*

Categoria	Itens
Óleos vegetais	Azeite; azeite de dendê; óleo de algodão; óleo de amendoim; óleo de côco; óleo de colza e mostarda; óleo de farelo de arroz; óleo de gergelim; óleo de gérmen de milho; óleo de girassol; óleo de palmiste; óleo de soja; outros óleos de oleaginosas
Ovos	Ovos
Peixes	Cefalópodes crustáceos; outros moluscos; outros peixes marinhos; peixe de água doce; peixes demersais; peixes pelágicos
Produtos Aquáticos	Carne de mamíferos aquáticos; outros animais aquáticos; plantas aquáticas
Raízes com amido	Batatas doces; inhame; outras raízes; produtos de batatas; produtos de mandioca
Temperos	Cravo-da-índia; outras especiarias; pimenta; pimento

B. Códigos ISO

Tabela B.1: Códigos ISO e Países Analisados

Código ISO 3166-1 alfa-3	País
ALB	Albânia
ARG	Argentina
ARM	Armênia
ATG	Antígua e Barbuda
AUS	Áustria
AZE	Azerbaijão
BEL	Bélgica
BGR	Bulgária
BHS	Bahamas
BIH	Bósnia e Herzegovina
BLR	Bielorrússia
BLZ	Belize
BOL	Bolívia
BRA	Brasil
BRB	Barbados
CAN	Canadá
CHE	Suíça
CHL	Chile
COL	Colômbia
CRI	Costa Rica
CUB	Cuba
CYP	Chipre
CZE	República Tcheca
DEU	Alemanha
DMA	Dominica
DNK	Dinamarca
DOM	República Dominicana
ECU	Equador

Contínua na próxima página

Tabela B.1 – *Continuação da página anterior*

Código ISO 3166-1 alfa-3	País
ESP	Espanha
EST	Estônia
FIN	Finlândia
FRA	França
GBR	Reino Unido
GEO	Geórgia
GRC	Grécia
GRD	Granada
GTM	Guatemala
GUY	Guiana
HND	Honduras
HRV	Croácia
HTI	Haiti
HUN	Hungria
IRL	Irlanda
ISL	Islândia
ITA	Itália
JAM	Jamaica
KAZ	Cazaquistão
KNA	São Cristóvão e Nevis
LCA	Santa Lúcia
LTU	Lituânia
LUX	Luxemburgo
LVA	Letônia
MDA	República da Moldávia
MEX	México
MKD	Macedônia do Norte
MLT	Malta
MNE	Montenegro
NIC	Nicarágua
NLD	Holanda
NOR	Noruega
PAN	Panamá
PER	Peru
POL	Polônia
PRT	Portugal
PRY	Paraguai
ROU	Romênia

Contínua na próxima página

Tabela B.1 – *Continuação da página anterior*

Código ISO 3166-1 alfa-3	País
RUS	Federação Russa
SLV	El Salvador
SRB	Sérvia
SUR	Suriname
SVK	Eslováquia
SVN	Eslovênia
SWE	Suécia
TTO	Trindade e Tobago
TUR	Peru
UKR	Ucrânia
URY	Uruguai
USA	Estados Unidos da América
VCT	São Vicente e Granadinas
VEN	Venezuela

C. Análise Descritiva

Tabela C.1: Análise descritiva dos grupos

Grupo	n	Variável	Média	Desvio Padrão	Mínimo	Máximo
		Cereais	14,4857	3,6673	8,2397	25,8497
		Frutas	9,4551	3,1083	4,8010	19,0409
		Legumes	12,4265	5,0454	5,9675	27,0563
1	47	Leite	24,0320	6,5669	12,4207	41,6755
		Raízes com amido	6,7280	3,3124	2,3287	18,3380
		Obesidade	25,4681	2,8512	21,2000	32,2000
		Confirmados	3,2538	1,7646	0,3976	7,8117
		Mortos	0,0629	0,0400	0,0036	0,1700
		Cereais	18,7875	5,0966	11,9883	27,8896
		Frutas	17,9893	4,6054	10,9045	25,2178
		Legumes	9,3130	3,0739	4,0603	16,7477
2	15	Leite	7,4869	4,0170	1,7529	15,4881
		Raízes com amido	10,1735	7,0690	2,1227	25,0560
		Obesidade	22,0133	2,7723	18,7000	26,7000
		Confirmados	0,9570	1,1713	0,0593	3,3470
		Mortos	0,0278	0,0384	0,0000	0,1148
		Cereais	16,3158	2,8152	14,5963	19,5647
		Frutas	11,3212	1,0360	10,1780	12,1978
		Legumes	30,9697	2,2064	29,1006	33,4037
3	3	Leite	15,6474	5,4836	9,4531	19,8813
		Raízes com amido	5,0343	2,0198	3,6759	7,3553
		Obesidade	21,3333	2,3798	19,2000	23,9000
		Confirmados	3,4122	2,3593	0,8066	5,4039
		Mortos	0,0791	0,0520	0,0208	0,1207
		Cereais	27,1257	4,5154	17,4292	33,6033
		Frutas	10,7406	3,1700	6,3468	16,0777
		Legumes	10,8220	6,3830	3,8957	23,2772
4	11	Leite	16,2073	5,4599	5,7576	22,7202

Contínua na próxima página

Tabela C.1 – *Continuação da página anterior*

Grupo	n	Variável	Média	Desvio Padrão	Mínimo	Máximo
		Raízes com amido	5,4077	2,6675	1,3590	9,4082
		Obesidade	20,6273	1,6156	18,8000	23,3000
		Confirmados	2,3843	2,1665	0,0917	6,1483
		Mortos	0,0518	0,0402	0,0025	0,1245
		Cereais	9,3658	3,8936	6,8028	13,8462
		Frutas	34,0107	6,5261	26,5408	38,6056
		Legumes	10,0931	5,0754	5,4769	15,5282
5	3	Leite	8,4782	3,9573	4,3509	12,2403
		Raízes com amido	6,6147	6,2076	2,1092	13,6954
		Obesidade	29,0667	2,7062	26,9000	32,1000
		Confirmados	1,2564	0,9992	0,1222	2,0069
		Mortos	0,0222	0,0218	0,0000	0,0435
		Cereais	11,0083	-	11,0083	11,0083
		Frutas	9,0853	-	9,0853	9,0853
		Legumes	11,4484	-	11,4484	11,4484
6	1	Leite	25,7283	-	25,7283	25,7283
		Raízes com amido	5,6773	-	5,6773	5,6773
		Obesidade	37,3000	-	37,3000	37,3000
		Confirmados	6,1019	-	6,1019	6,1019
		Mortos	0,1054	-	0,1054	0,1054