Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

DEPARTAMENTO DE ESTATÍSTICA
E-mail: dest@mat.ufrgs.br

Trabalho de Conclusão de Curso

Análise Não Supervisionada das Séries Temporais de COVID-19 nos Municípios Brasileiros e seu Envolvimento com Fatores Socioeconômicos e Políticos

João Lucas Simon

30 de agosto de 2024

João Lucas Simon

Análise Não Supervisionada das Séries Temporais de COVID-19 nos Municípios Brasileiros e seu Envolvimento com Fatores Socioeconômicos e Políticos

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. Marcio Valk

Porto Alegre
Agosto de 2024

João Lucas Simon

Análise Não Supervisionada das Séries Temporais de COVID-19 nos Municípios Brasileiros e seu Envolvimento com Fatores Socioeconômicos e Políticos

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador e pela Banca Examinadora.

Orientador: _____
Prof. Dr. Marcio Valk, UFRGS
Doutor(a) pela Universidade Federal do Rio Grande do Sul, Porto Alegre, RS

Banca Examinadora:

Prof. Dr. Gabriela Bettella Cybis, UFRGS
Doutora pela University of California, Los Angeles

Prof. Dr. Luciana Neves Nunes, UFRGS
Doutora pela Universidade Federal do Rio Grande do Sul – Porto Alegre, RS

Porto Alegre
Agosto de 2024

"A primeira regra é manter o espírito tranquilo. A segunda é enfrentar as coisas de frente e tomá-las pelo que realmente são.". (Marco Aurélio)

Agradecimentos

Este trabalho representa o fim de mais um ciclo. Um dos mais importantes da minha vida até então, e que já existia como sonho muito antes de se concretizar. Minha felicidade não está apenas em ter chegado até aqui, mas sim em cada experiência, em cada aprendizado, e em cada troca com todas as pessoas que passaram pela minha vida durante esse período. Quero agradecer a todos os ótimos professores que tive a felicidade de conviver e aprender, e em especial ao meu estimado orientador, Prof. Dr. Marcio Valk, por todo o apoio, empenho, sugestões e tempo dedicados para que este trabalho se torne uma realidade. Sou grato à minha família que sempre me apoiou durante essa jornada e me deu o suporte necessário para que eu alcançasse minhas metas. Por fim, agradeço a todos os amigos que estão ao meu lado e que deixam a minha vida mais feliz.

Resumo

O avanço humano tem sido marcado por impactos ambientais negativos, contribuindo para o surgimento de doenças emergentes e reemergentes, como a COVID-19. Este estudo analisa as séries temporais de novos casos e óbitos por COVID-19 em municípios brasileiros com mais de 100 mil habitantes entre 2020 e 2022, utilizando métodos de agrupamento de séries temporais com a medida de distância Dynamic Time Warping (DTW). Os resultados foram comparados com os de outras pesquisas, que analisaram a mortalidade por COVID-19 no Brasil considerando variáveis como o Índice de Desenvolvimento Humano Municipal (IDHM), o partidarismo político no segundo turno das eleições presidenciais de 2018, além da proporção de indivíduos vulneráveis à pobreza. Os resultados indicam padrões distintos na evolução dos casos e óbitos em diferentes municípios, com variáveis socioeconômicas apresentando diferenças estatisticamente significativas entre os grupos de novos óbitos. Essas variáveis mostram-se candidatas a influenciar diretamente a dinâmica das curvas pandêmicas, embora novos métodos estatísticos e colaborações interdisciplinares sejam necessários para validação. Concluímos que a clusterização de séries temporais é eficaz na análise da pandemia no Brasil e que nossos achados fornecem uma base para futuras pesquisas e políticas públicas, contribuindo para o entendimento da propagação e mortalidade do vírus e a formulação de estratégias de saúde pública mais eficazes.

Palavras-Chave: COVID-19, Agrupamento de Séries Temporais, Dynamic Time Warping, Política, Desigualdade Social, Mineração de Dados, Brasil.

Abstract

Human advancement has been marked by negative environmental impacts, contributing to the emergence and re-emergence of diseases such as COVID-19. This study analyzes the time series of new cases and deaths from COVID-19 in Brazilian municipalities with more than 100,000 inhabitants between 2020 and 2022, using time series clustering methods with the Dynamic Time Warping (DTW) distance measure. The results were compared with those of other studies, which analyzed COVID-19 mortality in Brazil considering variables such as the municipal-level human development index (HDI) and the proportion of votes for Bolsonaro in the 2018 elections, as well as the proportion of individuals vulnerable to poverty. The results indicate distinct patterns in the evolution of cases and deaths in different municipalities, with socioeconomic variables showing statistically significant differences between the groups of new deaths. These variables are candidates to directly influence the dynamics of the pandemic curves, although new statistical methods and interdisciplinary collaborations are needed for validation. We conclude that time series clustering is effective in analyzing the pandemic in Brazil and that our findings provide a basis for future research and public policies, contributing to the understanding of the spread and mortality of the virus and the formulation of more effective public health strategies.

Keywords: COVID-19, Time Series Clustering, Dynamic Time Warping, Politics, Social Inequality, Data Mining, Brazil.

Sumário

Lista de Figuras

Lista de Tabelas

1 Introdução

O avanço humano tem sido marcado por impactos negativos ao meio ambiente e à biodiversidade, contribuindo para o surgimento de diversas doenças emergentes e reemergentes nas últimas décadas. A mais recente e devastadora dessas doenças é a COVID-19, identificada pela primeira vez em 2019 em Wuhan, China. Causada pelo SARS-CoV-2, a COVID-19 foi declarada uma pandemia pela Organização Mundial da Saúde (OMS) em 11 de março de 2020 e, em 2022, já havia resultado em mais de 5 milhões de mortes em todo o mundo (Xavier et al., 2022).

No Brasil, a severidade da pandemia de COVID-19 pode ser atribuída a fatores como desigualdades socioeconômicas, a falta de uma resposta coordenada e eficaz, atrasos na campanha de vacinação e a disseminação de desinformação, promovendo o uso de substâncias ineficazes como cloroquina e ivermectina (Ferrante et al., 2021).

Este estudo analisa as séries temporais de novos casos e óbitos por COVID-19 em municípios brasileiros com mais de 100 mil habitantes entre 2020 e 2022. A metodologia baseia-se no estudo de Luo et al. (2023), que utilizou agrupamento de séries temporais com a medida de distância Dynamic Time Warping (DTW) para identificar padrões de evolução de novos casos e óbitos em mais de 100 países. Os resultados deste estudo foram comparados com os de Xavier et al. (2022), que utilizou árvores de regressão para analisar a mortalidade por COVID-19 no Brasil, considerando variáveis como o Índice de Desenvolvimento Humano Municipal (IDHM) e a proporção de votos para Bolsonaro no segundo turno das eleições presidenciais de 2018 (partidarismo político). Além dessas variáveis, este estudo introduz a proporção de indivíduos vulneráveis à pobreza em cada município. Os resultados encontrados foram condizentes com os de Xavier et al. (2022) ao compararmos as variáveis comuns.

A pandemia gerou impactos significativos em várias áreas, como a precarização da educação, a superlotação de hospitais e o aumento de doenças psicológicas devido ao isolamento social (Malta et al., 2021). Esses impactos revelaram a necessidade de entender melhor os padrões de propagação do vírus e a mortalidade associada a ele. Este estudo visa fornecer insights sobre os padrões de comportamento das curvas de novos casos e óbitos por COVID-19, ajudando a compreender a propagação e a mortalidade do vírus. Além disso, busca avaliar a eficácia dos métodos de agrupamento de séries temporais e identificar variáveis significativas na dinâmica dessas curvas, contribuindo para o desenvolvimento de estratégias eficazes de prevenção e controle em futuras crises.

O objetivo principal é aprofundar a compreensão da evolução da pandemia no Brasil e avaliar as decisões tomadas ao longo do tempo, oferecendo embasamento

para pesquisadores e autoridades governamentais na formulação de estratégias de saúde pública mais eficazes. Para isso, serão analisadas as séries temporais de novos casos e óbitos de COVID-19 nos municípios brasileiros de grande porte entre 2020 e 2022, utilizando métodos de clusterização e os grupos encontrados de novos óbitos serão analisados utilizando diferentes estratégias, como testes de hipótese não-paramétricos para comparar os resultados encontrados com outras pesquisas relacionadas.

Este estudo utilizará dados de casos e óbitos por município, obtidos através de informações oficiais do Ministério da Saúde e das Secretarias Estaduais de Saúde via Brasil.IO (Cota, 2020), complementados por dados socioeconômicos do Atlas Brasil (Fundação João Pinheiro, 2022) e dados demográficos do IBGE.

O trabalho está organizado da seguinte forma: a Seção 2 mostra o plano de pesquisa e revisa a literatura relevante; a Seção 3 descreve as metodologias utilizadas; a Seção 4 apresenta e discute os resultados das análises; e a Seção 5 conclui o estudo.

2 Planejamento e Revisão Bibliográfica

2.1 Planejamento

Os dados utilizados neste estudo consistem em novos casos e óbitos semanais por COVID-19 nos municípios de grande porte do Brasil, conforme definido pelo Censo de 2010. A escolha dos municípios de grande porte se deu pela maior disponibilidade de dados e pela relevância desses centros urbanos na disseminação da pandemia. Para garantir a comparabilidade, os dados foram normalizados, ajustando as séries temporais para uma mesma escala.

Inicialmente, foi realizada uma análise exploratória dos dados para entender suas características e identificar possíveis anomalias. A normalização das séries temporais foi crucial para assegurar que as comparações entre diferentes municípios não fossem enviesadas por diferenças de escala. Para a formação dos grupos, utilizamos a medida de distância Dynamic Time Warping (DTW), que permite comparar séries temporais de diferentes durações e identificar alinhamentos não lineares entre elas. Em seguida, aplicamos o algoritmo de agrupamento hierárquico com o método de Ward, que minimiza a variância dentro de cada grupo formado.

Para definir a quantidade ideal de grupos, empregamos medidas de validação interna, como o índice de silhueta, que avalia a coesão e separação dos clusters formados. Primeiramente, os dados de novos casos foram agrupados, permitindo a análise da distribuição dos municípios por região do Brasil e a identificação de padrões distintos de evolução da pandemia. Os dados de novos óbitos foram então agrupados, resultando na identificação de quatro padrões distintos de evolução. A análise revelou uma predominância de determinadas regiões do Brasil em grupos específicos, indicando possíveis diferenças regionais na evolução de novos casos e mortalidade por COVID-19.

Para comparar os grupos formados de novos óbitos, utilizamos o teste de Kruskal-Wallis, avaliando três variáveis: o Índice de Desenvolvimento Humano Municipal (IDHM), a proporção de votos em Bolsonaro no segundo turno das eleições presidenciais em 2018 (partidarismo político) e a proporção de indivíduos vulneráveis à pobreza em cada município. Estudos anteriores já haviam indicado que essas variáveis são significativas na mortalidade por COVID-19, relacionando o IDHM a fatores socioeconômicos e o negacionismo da pandemia a certos posicionamentos políticos. Todos os códigos foram desenvolvidos em R, utilizando as bibliotecas dtwclust e tsclust para o agrupamento de séries temporais, e tidyverse e ggstatplot para a transformação, visualização dos dados e realização dos testes de hipótese.

A seguir, a próxima seção abordará os principais estudos e abordagens que tam-

bém utilizaram métodos de agrupamento e influenciaram a condução desta pesquisa, destacando as contribuições relevantes para o entendimento da dinâmica da pandemia e das metodologias utilizadas.

2.2 Referencial Teórico

O artigo de Luo et al. (2023) foi a principal fonte de referência para a metodologia utilizada neste trabalho. Nele, é utilizada a distância Dynamic Time Warping (DTW) e clusterização hierárquica para analisar séries temporais de casos e mortes diárias por COVID-19 em 100 países. Identificaram-se quatro padrões distintos associados à distribuição geográfica e estrutura demográfica, destacando-se a alta mortalidade na Europa Ocidental no início da pandemia. Os resultados reforçaram a viabilidade da clusterização de séries temporais para extrair informações relevantes, apesar de suas limitações.

Em Cassão et al. (2022) foi utilizado o agrupamento de séries temporais usando uma variação do algoritmo K-Means com a métrica de similaridade DTW. Os dados analisados incluem as séries temporais acumuladas do número de mortes por 100 mil habitantes ao longo de 452 dias em todos os estados brasileiros. Os resultados mostraram três padrões de resposta à pandemia, variando de maior a menor controle, com todos os grupos apresentando um aumento significativo no número de mortes nos últimos meses.

Pereira et al. (2020) aplica um algoritmo de agrupamento às regiões do mundo para as quais existem dados epidêmicos disponíveis e a pandemia está em estágio avançado. Em seguida, um conjunto de características representando a resposta dos países à disseminação inicial da pandemia é usado para treinar uma Rede Autoencoder para prever o futuro da pandemia no Brasil. – Este agrupamento foi um elemento chave para evitar treinar o modelo preditivo para um determinado estado em um país que tem uma dinâmica totalmente diferente. Em trabalhos futuros, pretendemos refinar esses clusters periodicamente para considerar a evolução das estratégias das diferentes regiões estudadas.

Kurniawan et al. (2020) este estudo emprega o agrupamento K-means para agrupar 200 países afetados pela pandemia no mundo. Além disso, a análise de risco também necessita de outro método para buscar conhecimento desconhecido nos dados. Uma pesquisa afirma que o conhecimento gerado a partir da matriz de correlação é essencial. A matriz de correlação supera com sucesso os problemas de incerteza entre fatores relacionados. Portanto, tanto as técnicas de agrupamento quanto a matriz de correlação são aplicadas para ver a relação oculta entre componentes nos dados. Conhecimento das relações entre atributos é essencial para determinar métodos apropriados de prevenção e mitigação.

Em Liu et al. (2020) um algoritmo de agrupamento é usado para processar dados de buscas na Internet e alertas de notícias para realizar uma previsão em tempo real do surto.

Para estudar o comportamento epidêmico em diferentes zonas da cidade de Nova York, um algoritmo de agrupamento é proposto em Khmaissia et al. (2020) que modela o surto na cidade.

Em dos Santos Gomes e de Oliveira Serra (2021), o estudo propõem uma nova ferramenta computacional de aprendizado de máquina para rastreamento adaptativo e previsão em tempo real de casos de morte por COVID-19. A abordagem combina

filtros de Kalman e um algoritmo de agrupamento fuzzy tipo-2 com mecanismo de distância de similaridade adaptativa. O conjunto de dados usado para a experimentação, extraído do relatório oficial do Ministério da Saúde do Brasil, consiste em relatórios diários de mortes no período de 29 de fevereiro de 2020 a 18 de maio de 2020. O método foi comparado com LASSO, ARIMA e rede neural recorrente LSTM, Wavelet-Coupled Random Vector Functional Link (WCRVFL).

A abordagem de Rios et al. (2021) emprega um algoritmo de agrupamento hierárquico, um ramo de aprendizado não supervisionado da Inteligência Artificial, junto com a estratégia de ligação média para determinar a pertinência dos países aos grupos ao longo das semanas para analisar como a doença se espalha e afeta diferentes sociedades. As partições de agrupamento foram avaliadas usando a silhueta média para garantir a representatividade da modelagem.

Nicholson et al. (2022) empregam métodos supervisionados e não supervisionados para identificar os fatores demográficos, de mobilidade, clima, capacidade médica e saúde ao nível do condado críticos para estudar a propagação da COVID-19 antes da disponibilidade generalizada de uma vacina. Usamos esse subespaço de características para agregar condados em clusters significativos para apoiar esforços de análise de doenças mais refinados.

Nikolopoulos et al. (2021) propõem o uso de 7 modelos de aprendizado de máquina e um novo método de previsão híbrido baseado em vizinhos mais próximos e agrupamento k-means para prever taxas de crescimento da COVID-19. Eles empregaram LSTM, regressão linear múltipla, regressão ridge, árvores de decisão, floresta aleatória, rede neural e máquinas de vetor de suporte em dados a nível de país (dos EUA, Índia, Reino Unido, Alemanha e Singapura).

Para determinar como avaliar o risco regional relacionado à COVID-19, Lucic et al. (2021) – propõem uma abordagem de modelagem em duas fases enquanto consideram critérios demográficos e econômicos. Primeiro, uma técnica de agrupamento não supervisionado, especificamente k-means, é empregada para agrupar condados dos EUA com base em semelhanças demográficas e econômicas. Em seguida, a previsão de séries temporais de cada cluster de condados é desenvolvida para avaliar o risco de transmissibilidade viral de curto prazo.

Ashouri e Phoa (2022) propõem uma ferramenta interativa baseada na web para agrupar e prever os dados disponíveis dos casos confirmados de infecção por COVID-19 em Taiwan. 4 a árvore baseada em modelo (MOB) e atributos relevantes ao domínio para agrupar o conjunto de dados e exibir previsões.

Jin (2020) apresenta um algoritmo eficaz para o agrupamento de casos confirmados de COVID-19 ao nível do condado nos Estados Unidos. A deformação dinâmica do tempo e a distância euclidiana são examinadas como métricas de distância de agrupamento k-means. A deformação dinâmica do tempo pode comparar séries temporais variando em velocidade, pois os condados frequentemente experimentam tendências de surto semelhantes sem que as linhas do tempo correspondam exatamente. O efeito do pré-processamento de dados no agrupamento foi estudado sistematicamente. Análises adicionais demonstram o valor imediato dos clusters tanto para a interpretação retrospectiva da pandemia quanto como entradas informativas para modelos de previsão de casos.

Usando os dados retrospectivos diários da COVID-19 em 2020 divididos em 24 períodos de meio mês, Sadeghi et al. (2021) aplicaram técnicas de aprendizado de máquina não supervisionado, em particular, análise de agrupamento hierárquico

para agrupar países em cinco grupos dentro de cada período de acordo com sua mortalidade cumulativa diária por COVID-19 ao longo do ano e casos cumulativos diários de COVID-19 por milhão de habitantes ao longo do período de meio mês.

A pesquisa de Soto-Ferrari et al. (2023) - propõe uma métrica de avaliação de risco chamada AGGFORCLUS que integra previsão de séries temporais e agrupamento para transmitir informações conjuntas sobre o crescimento previsto da carga de casos e variabilidade, fornecendo assim uma visão educada e visualmente simples do status de risco.

Caiado e Lúcio (2023) propõem uma nova abordagem de agrupamento para comparar séries temporais financeiras e empregá-la para estudar como a pandemia de COVID-19 afetou o mercado de ações dos EUA.

No geral, os métodos de agrupamento aplicados aos dados de COVID-19 já possuem uma vasta literatura, porém, ainda foram pouco explorados em conjuntos de dados da pandemia no Brasil, especialmente numa escala municipal. Além disso, esses métodos estão frequentemente associados a técnicas de previsão de séries temporais (forecasting), que utilizam os resultados dos agrupamentos para informar estratégias de análise e tomada de decisão. No próximo capítulo, iremos descrever as principais metodologiais utilizadas neste estudo.

3 Metodologia

Nesse Capítulo, inicialmente iremos descrever o conjunto de dados que serviu como base para a análise, bem como as técnicas de aprendizado de máquina aplicadas no estudo e o plano de pesquisa elaborado. Este trabalho busca explorar padrões de evolução da COVID-19 nos municípios de grande porte do Brasil, utilizando a medida de distância DTW e um algoritmo de agrupamento hierárquico com o método de Ward.

3.1 Pré-Processamento dos Dados

Na Figura 3.1, conseguimos ver a curva de novos casos e mortes por COVID-19 no Brasil semanalmente. Para a situação epidemióligica, foram considerados os dados a partir de 25/02/2020 até 31/12/2022. Podemos observar que a magnitude de casos é diferente da magnitude de óbitos, visto que só um percentual pequeno dos casos tem desfecho indesejado.

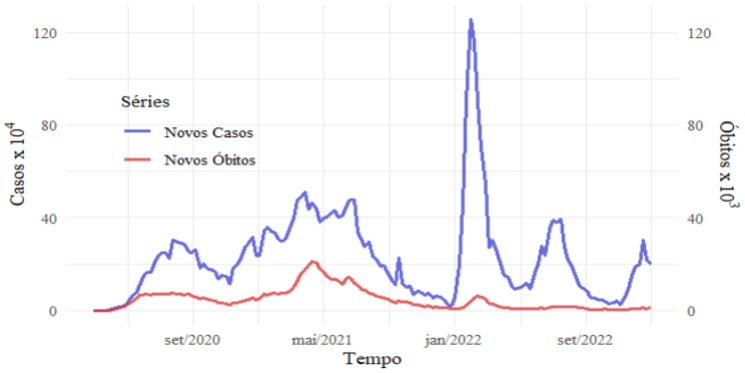

Figura 3.1: Séries temporais semanais de novos casos e de novos óbitos por COVID-19 no Brasil.

Por conta dos municípios possuírem diferentes magnitudes de população, território, PIB e mobilidade urbana, também é esperado diferentes magnitudes de novos casos e óbitos em diferentes municípios. Em outros estudos como Cassão et al. (2022), foram utilizados os dados de novos casos (óbitos) por 100 mil habitantes para mitigar os efeitos de diferentes tamanhos de população.

Neste estudo, utilizamos os dados normalizados por semana epidemiológica para padronizar a comparação das séries temporais. Dessa forma, podemos definir a quantidade de novos casos e óbitos como:

$$x_{it}^{*} = \frac{x_{it} - \bar{x}_i}{s_i}, \qquad (t = 1, 2, \ldots, T; i = 1, 2, \ldots, n).$$

em que x_{it}^{*} representa o número normalizado de novos casos (óbitos) semanais no município i na semana epidemiológica t, x_{it} representa o número de novos casos (óbitos) semanais no município i no dia t, e $\bar{x}_i$ representa a média de novos casos semanais (óbitos) no município i durante todo o período de estudo. s_i é o desvio padrão de novos casos (óbitos) semanais no município i durante todo o período de estudo. (Luo et al., 2023)

3.2 Aprendizado de Máquina Não-Supervisionado

Há uma maneira informativa de visualizar os dados? Podemos descobrir subgrupos entre as variáveis ou entre as observações? O aprendizado não-supervisionado refere-se a um conjunto diversificado de técnicas para responder a perguntas como essas. (James et al., 2013)

Essencialmente, o aprendizado não-supervisionado reside na identificação de propriedades intrínsecas aos dados de entrada, de maneira a construir representações que possam servir a diversos propósitos, como auxílio a tomada de decisões ou descoberta de conhecimento. Essas técnicas são utilizadas, principalmente, quando o objetivo da pesquisa é encontrar padrões ou tendências que auxiliem no entendimento dos dados. Mais precisamente, no aprendizado não supervisionado não existem atributos meta. A partir do conjunto de dados X, um algoritmo de aprendizado de máquina aprende a representar as entradas submetidas segundo algum critério de qualidade. (Faceli, 2021)

Dentre as técnicas disponíveis, iremos utilizar a clusterização, uma ampla classe de métodos para descobrir subgrupos desconhecidos nos dados.

3.2.1 Métodos de Agrupamento de Séries Temporais

A clusterização (agrupamento) refere-se a um conjunto muito amplo de técnicas para encontrar subgrupos, ou clusters, em um conjunto de dados. Quando fazemos a clusterização das observações de um conjunto de dados, procuramos particioná-las em grupos distintos, de modo que as observações dentro de cada grupo sejam bastante semelhantes entre si, enquanto as observações em diferentes grupos sejam bastante diferentes entre si e, para tornar isso concreto, devemos definir o que significa que duas ou mais observações sejam semelhantes ou diferentes. (James et al., 2013)

No contexto de séries temporais, os métodos de agrupamento podem ser definidos em três categorias: shape-based, feature-based ou model-based.(Aghabozorgi et al., 2015) Nesse trabalho, iremos utilizar métodos aplicados aos dados originais de cada série (shape-based), dada a multicausalidade presente nos dados e a complexidade de agrupá-las a partir de uma medida de similaridade. (Luo et al., 2023)

Dado um conjunto de dados de n séries temporais $S = \{x_1, x_2, \ldots, x_n\}$ o processo de partição não supervisionada de S em $G = \{G_1, G_2, \ldots, G_K\}$, de modo que séries temporais homogêneas são clusterizadas (agrupadas) baseado em uma determinada medida de similaridade, é chamado de agrupamento de séries temporais. Então, G_i

é denominado grupo, onde $S = \cup_{i=1}^{K} G_i$ e $G_i \cap G_j = \varnothing$ para $i \neq j$ (Warren Liao, 2005).

3.2.2 Medida de Distância DTW

O processo de agrupar séries temporais em grupos similares depende muito de qual característica dos dados está sendo medida. Para diferentes características, são necessárias métricas distintas. Existem muitas medidas de similaridade entre séries temporais, porém aqui vamos considerar uma em particular que tem por atributo medir a flexibilidade temporal entre séries com durações diferentes, chamada Dynamic Time Warping (DTW).

Seja $x_t = (x_1, \ldots, x_T)$ e $z_{t'} = (x_1, \ldots, x_{T'})$ duas séries de tamanho T e T', respectivamente, e d a distância entre dois elementos de cada uma delas. Em casos assim, a distância Euclidiana é geralmente aceita como a distância mais simples entre duas séries. A distância entre x_t e $z_{t'}$ pode ser definida como:

$$D(x_t, z_{t'}) = \sqrt{d(x_1, z_1)^2 + \ldots + d(x_T, z_{T'})^2}.$$

Infelizmente, essa medida não consegue capturar similaridades flexíveis presentes em séries temporais. Por exemplo, $x = (a, b, a, a)$ e $z = (a, a, b, a)$ são diferentes de acordo com essa distância mesmo que elas representem trajetórias similares.(François Petitjean, 2011)

Dynamic Time Warping (DTW), ou Distorção Dinâmica de Tempo, é um algoritmo que permite que você descubra como alinhar duas séries temporais que podem ter durações diferentes. A técnica nos diz duas coisas: em primeiro lugar, quais pontos de uma das séries correspondem aos pontos de outra série e, em segundo lugar, o quão similares duas ou mais séries são entre si a partir de uma função custo (Ver Figura 3.2)

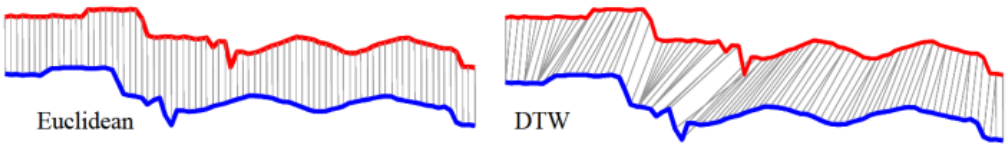

Figura 3.2: Comparação entre distância Euclidiana e Dynamic Time Warping (Keogh, 2002).

Retomando a notação utilizada anteriormente, para alinhar ambas as séries a partir dessa medida de distância, o primeiro passo será estabelecer uma matriz custo local de tamanho $T \times T'$. Cada elemento da matriz é a distância Euclidiana entre cada ponto das séries x_t e $z_{t'}$, em que $t = 1, 2, \ldots, T$ e $t' = 1, 2, \ldots, T'$, e então podemos denotar a função custo como (Luo et al., 2023; Yang et al., 2019)

$$\delta(t, t') = (x_t - z_{t'})^2.$$

No segundo passo, o algoritmo irá definir o caminho de distorção W, um conjunto contíguo (no sentido declarado abaixo) de elementos de matriz que define uma relação entre x_t e $z_{t'}$. O r-ésimo elemento de W é definido como $w_{\mathrm{r}} = (t, t')_{\mathrm{r}}$, então temos:

$$W = w_1, w_2, \ldots, w_{\mathrm{r}}, \ldots, w_{\mathrm{R}} \quad \max(T, T') \leq \mathrm{R} < T + T' - 1.$$

O caminho de deformação que define o alinhamento entre as duas séries temporais está sujeito a várias restrições.

- Limites: $w_1 = (1, 1)$ e $w_r = (T, T')$, isso requer que o caminho comece e termine em cantos diagonalmente opostos da matriz.
- Continuidade: Sendo $w_r = (a, b)$ então $w_{r-1} = (a', b')$ onde $a - a' <= 1$ e $b - b' <= 1$, isso restringe os possíveis passos do caminhamento para células que sejam adjacentes (incluindo diagonalmente adjacentes).
- Monotonicidade: Sendo $w_r = (a, b)$ então $w_{r-1} = (a', b')$ onde $a - a' >= 0$ e $b - b' >= 0$, isso força os pontos de W a serem monotonicamente espaçados no tempo.

Além disso, grande parte da literatura em que se aplica DTW também restringe o caminho de deformação de forma global, limitando o quanto ele pode se desviar da diagonal.(Rakthanmanon et al., 2012; Keogh, 2002) Uma restrição típica é a Banda Sakoe-Chiba, que afirma que o caminho de deformação não pode se desviar mais do que R células da diagonal.(Rakthanmanon et al., 2012; Sardá-Espinosa, 2019)

Existem diversos caminhos de deformação que satisfazem as condições acima, no entanto, estamos interessados apenas no caminho que minimiza o custo de deformação. Conforme (Keogh, 2002; Yang et al., 2019), esse caminho pode ser encontrado utilizando programação dinâmica ao acumular a função custo $\delta(t, t')$ através de toda a matriz, iterando através de $\delta(1, 1)$ até $\delta(T, T')$, usando um princípio de minimização para acumular o custo dos elementos bidimensionais adjacentes sequencialmente, isto é,

$$\sigma(t, t') = \delta(t, t') + \min\{\sigma(t-1, t'), \sigma(t, t'-1), \sigma(t-1, t'-1)\}. \tag{3.1}$$

Onde $\sigma(t, t')$ é o acumulado mínimo das distâncias fragmentárias δ' de $(n, m) = (1, 1)$ até $(t, t') = (T, T')$ para percorrer a matriz bidimensional. De acordo com a Regra de Acumulação 3.1, a distância acumulada mínima $\sigma(T, T')$ pode ser alcançada para obter um caminho encurtado W para medir a distância entre as sequências $\mathbf{x}$ e $\mathbf{z}$. O caminho contíguo W, registrado e distorcido desde a primeira célula $(1, 1)$ até a célula final (T, T'), fornece o custo total de distorção Σ como:

$$\Sigma = \sqrt{\sum_{i \in W} \delta_i(t, t')}$$

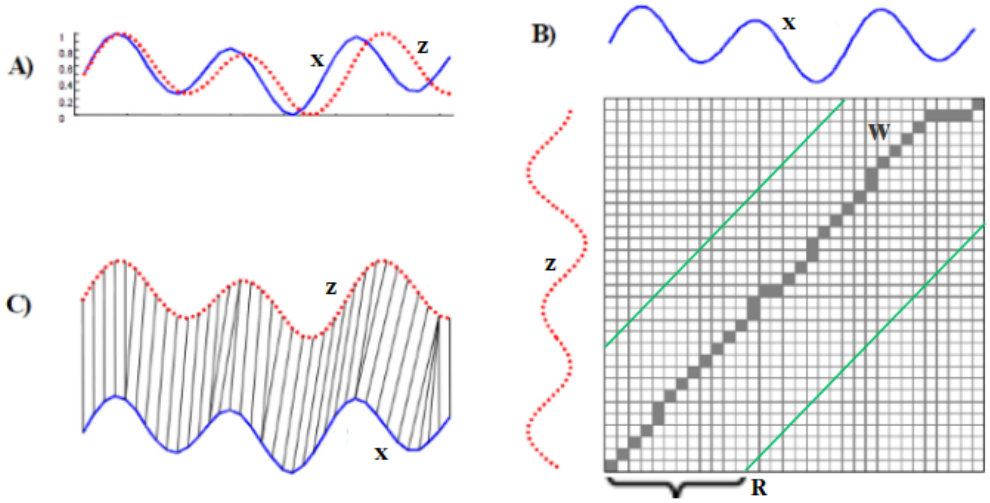

Figura 3.3: **A)** Duas sequências X e Z que são semelhantes, mas estão fora de fase. **B)** Alinhamento de sequências . **C)** O alinhamento resultante. (Keogh, 2002; Rakthanmanon et al., 2012)

Na figura 3.3 temos duas sequências X e Z que são semelhantes (A), mas estão fora de fase. Para alinhar as sequências, construímos uma matriz de deformação e procuramos pelo caminho de deformação ótimo, mostrado com os quadros em cinza. Note que a Banda Sakoe-Chiba com largura R é usada para limitar o caminho de deformação (B). E no gráfico (C) o alinhamento resultante.

3.2.3 Algoritmo de Agrupamento

Conforme descrito em Luo et al. (2023), existem diversos algoritmos de agrupamento, dependendo do critério usado para decidir quais grupos mesclar. Um dos métodos mais conhecidos é o método de Ward, que mescla os dois grupos que têm o menor custo para mesclar, conforme a equação abaixo. Seja G_m o grupo $m, x_i^{(m)}$ a i-ésima amostra em G_m, n_m o número de amostras em G_m e $\tilde{x}^{(m)}$ o protótipo de série temporal (centroide) de G_m. Então, a soma dos quadrados dos desvios dos exemplos em G_m é dada por:

$$W_m = \sum_{i=1}^{n_m} \left(d_i^{(m)}(T,T) \right)^2,$$

onde $d_i^{(m)}(T,T)$ denota a distância DTW entre $x_i^{(m)}$ and $\tilde{x}^{(m)}$. Assim, a soma dos quadrados dos desvios dos K grupos é:

$$W = \sum_{m=1}^{K} W_m.$$

Quando K é determinado, o agrupamento que faz com que W alcance o valor mínimo é escolhida. O método de Ward considera o aumento da soma dos quadrados dos desvios dos dois grupos combinados como a distância quadrada entre grupos. Em outras palavras, a distância quadrada entre o grupo p e o grupo q é definida como:

$$D_{pq}^2 = W_r - W_p - W_q,$$

onde $G_r = G_p \cup G_q, W_r$ é a soma dos quadrados dos desvios de G_r, onde $p, q, r \in \{1, 2, \cdots, K\}$. O método de Ward começa considerando as n amostras como n grupos (no momento,$W = 0$), e então combina dois deles em um grupo de cada vez, reduzindo um grupo. A soma dos desvios quadrados é aumentada para cada combinação. Os dois grupos que causam o menor aumento na soma dos quadrados dos desvios W são selecionados e combinados até que todos as amostras sejam agrupados em um só.

3.3 Medidas de Validação

O agrupamento é comumente considerado um procedimento não supervisionado, portanto, avaliar seu desempenho pode ser bastante subjetivo. No entanto, existem métricas padronizadas de avaliação de agrupamento usando índices de validade de grupo (CVI's) (Sardá-Espinosa, 2019). Dentre as técnicas disponíveis, iremos utilizar o índice da Silhueta e o índice de Dunn que são métodos de validação interna que avaliam apenas os grupos formados.

3.3.1 Índice da Silhueta

O índice da Silhueta pode ser definido como,

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

em que $a(i)$ representa a diferença média entre uma série temporal x_m e todas as demais séries do grupo em que está contida, e $b(i)$ a diferença média mínima entre x_m e todas as demais séries dos grupos em que não está contido (Rousseeuw, 1987). Os valores da silhueta poderão variar entre [-1,1], sendo que quanto mais próximo de 1, maior é a similaridade intra-grupo e menor a similaridade entre-grupos. As medidas podem ser resumidas com o coeficiente da silhueta, que pode ser definido como

$$\mathrm{SC} = \max_k \tilde{s}(k),$$

em que $\tilde{s}(k)$ representa a média $s(i)$ sobre todas as séries do banco de dados para um número específico de grupos K (Kaufman e Rousseeuw, 1990).

3.3.2 Índice de Dunn

Já o índice de Dunn é um índice do tipo razão onde a coesão é estimada pela distância do vizinho mais próximo e a separação pelo diâmetro máximo do grupo. O índice original é definido como (Arbelaitz et al., 2013)

$$\mathrm{D}(C) = \frac{\min_{c_k \in C}\{\min_{c_l \in C c_k}\{\delta(c_k, c_l)\}\}}{\max_{c_k \in C}\{\Delta(c_k)\}},$$

em que

$$\delta(c_k, c_l) = \min_{x_i \in c_k} \min_{x_j \in c_l}\{\ \mathrm{d_e}(x_i, x_j)\},$$
$$\Delta(c_k) = \max_{x_i, x_j \in c_k}\{\ \mathrm{d_e}(x_i, x_j)\}.$$

Um maior índice de Dunn significa que a distância entre amostras de diferentes clusters é grande, enquanto a distância entre amostras do mesmo cluster é pequena, o que resulta em um melhor agrupamento (Luo et al., 2023).

3.4 Comparação de Variáveis Socioeconômicas e Políticas entre os Grupos Formados

Para avaliar os resultados dos agrupamentos, iremos verificar a hipótese de que há diferença entre a mediana dos grupos para as variáveis de Índice de Desenvolvimento Humano Municipal (IDHM) e do partidarismo político em cada município a partir dos resultados das eleições presidenciais de 2018, segundo Xavier et al. (2022) estas duas variáveis tiveram influência nos desfechos de mortalidade por COVID-19 nos municípios brasileiros.

3.4.1 Teste de Kruskal-Wallis

O teste que iremos utilizar será o teste de Kruskal–Wallis, que é uma generalização do modelo de análise de variância (ANOVA), este método não paramétrico testa se as amostras se originam da mesma distribuição. Sejam $F_1, \ldots, F_K$ denotando $K \geqslant 2$ funções de distribuição acumulada (FDAs) populacionais desconhecidas, das quais K grupos de amostras aleatórias independentes de tamanhos $n_1, \ldots, n_K$ são extraídas, respectivamente. As hipóteses nula e alternativa genérica para o teste de Kruskal–Wallis são:

$$H_0 : F_1 = \cdots = F_K = F^0$$
$$H_1 : \text{ ao menos uma } F_i \text{ não é igual a } F^0.$$

Considere amostras aleatórias independentes $\{X_{ij} : i = 1, \ldots, K\,,\ j = 1, \ldots, n_i\}$ coletadas em experimentos envolvendo K grupos de tratamento. Seja R_i a soma das classificações da i-ésima amostra na amostra combinada. A estatística do teste de Kruskal-Wallis para H_0 versus H_1 nas hipóteses acima é

$$S_{KW} = \frac{12}{n(n+1)} \sum_{i=1}^{K} n_i \left(\frac{R_i}{n_i} - \frac{n+1}{2} \right)^2,$$

onde $n = \sum_{i=1}^{K} n_i$ denota o tamanho total da amostra. Sob a hipótese nula H_0, S_{KW} é assintoticamente qui-quadrado distribuído com $K - 1$ graus de liberdade, ou seja, χ^2_{K-1} (Fan et al., 2011).

3.4.2 Teste de Dunn

O teste de Dunn é um procedimento de múltipla comparação par a par não paramétrico, frequentemente utilizado após a rejeição da hipótese nula em análises como o teste de Kruskal-Wallis. Ele permite comparar grupos independentes e identificar quais deles são estatisticamente diferentes em algum nível de significância α. Desse modo, conseguimos identificar quais grupos possuem diferenças significativas ao serem observados a partir da variáveis utilizada na comparação.

4 Resultados e Discussão

O Brasil, com mais de 5 mil municípios, apresenta uma vasta diversidade de características culturais, econômicas e geográficas. Em eventos de grande escala, como a pandemia de COVID-19, é esperado que esses fatores resultem em diferentes padrões de evolução dos novos casos e óbitos ao longo das diversas localidades do país. Para investigar essas variações, aplicamos métodos de agrupamento de séries temporais aos dados dos municípios com mais de 100 mil habitantes, a fim de identificar similaridades e dissimilaridades entre as séries temporais de novos casos e óbitos nessas cidades.

O resultado do agrupamento de novos óbitos também foi analisado utilizando testes de hipótese para avaliar as diferenças em variáveis associadas aos grupos formados, comparando os resultados com o estudo de Xavier et al. (2022) que analisou a mortalidade por COVID-19 nos municípios brasileiros. Especificamente, aplicamos o teste de Kruskal-Wallis para comparar as medianas de variáveis como o Índice de Desenvolvimento Humano Municipal (IDHM), o partidarismo político, e a proporção de indivíduos vulneráveis à pobreza, entre os grupos formados. Essa análise visa compreender como essas variáveis podem ter influenciado a configuração dos grupos e, consequentemente, o padrão de evolução da mortalidade por COVID-19 em diferentes regiões do Brasil .

4.1 Agrupamento de Novos Casos

O primeiro passo foi agrupar as séries temporais de novos casos em todos os 326 municípios utilizando o método de agrupamento descrito acima. A figura 4.1 mostra o dendrograma como resultado visual do método de agrupamento hierárquico, em que todas as séries começam em seu próprio grupo e assim são fundidas sucessivamente até formarem um único grupo. No gráfico, a altura indica a ordem em que os grupos foram unidos e os valores de "altura"refletem a distância ou diferença entre os grupos.

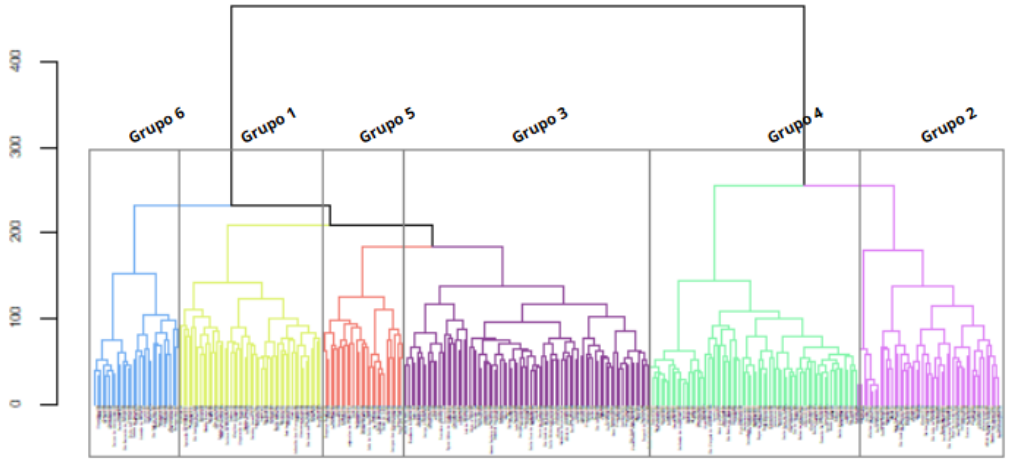

Figura 4.1: Dendrograma feito a partir do agrupamento hierárquico das séries temporais de novos casos dos 326 municípios de grande porte analisados

Podemos visualizar que a partir de seis grupos as diferenças/alturas a cada fusão de dois grupos ficam cada vez maiores, indicando não ser apropriado continuar a fusão dos grupos. Por conta de estarmos realizando uma análise exploratória em um grande volume de dados a partir de métodos não-supervisionados, torna-se complexa a tarefa de definir a quantidade de grupos K somente a partir de resultados visuais. No entanto, os resultados do índice de Dunn e do coeficiente da Silhueta corroboram a decisão de que seis grupos seria o ideal ao indicarem uma maior similaridade intra-grupos para $K = 6$.

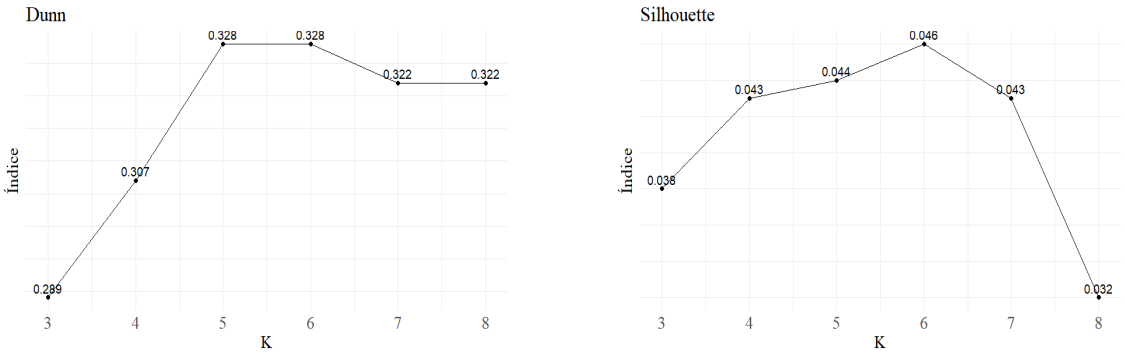

Figura 4.2: Resultado dos métodos de validação de clusters mencionados anteriormente

Na figura 4.3 conseguimos ver as séries temporais de novos casos em cada município de acordo com seu grupo, ficando evidente os diferentes padrões de evolução pandêmicos em nossa amostra de municípios. Os grupos 1,3 e 6 caracterizam-se pelo rápido crescimento na primeira e segunda onda de novos casos ocorrida, respectivamente, em torno de setembro de 2020 e maio de 2021. Com exceção do grupo 6, todos os demais grupos apresentaram uma terceira e quarta onda de novos casos ocorrida no início de 2022. Além disso, os grupos 2,4 e 5 tiveram uma evolução mais lenta e menor de novos casos na primeira e segunda onda que se manteve até o final de 2021.

Na tabela 4.1, podemos ver que aproximadamente 50% dos municípios de grande porte da região Sul estão incluídos no grupo 4, ou seja, tiveram um desenvolvimento da curva de novos casos semelhantes à média representada pelo grupo. Para os municípios do Nordeste tivemos a mesma proporção mas para o grupo 3, com estas cidades possuindo um padrão de evolução mais próximo à média desse grupo. Já os municípios do Sudeste estão mais presentes no grupo 4 e 2, enquanto os do Centro-Oeste e Nordeste estão inclusos majoritariamente no grupo 3. Em outros estudos como Théry (2020) e Xavier et al. (2022), a geografia das cidades também foi um fator de influência na disseminação da COVID-19, os estudos

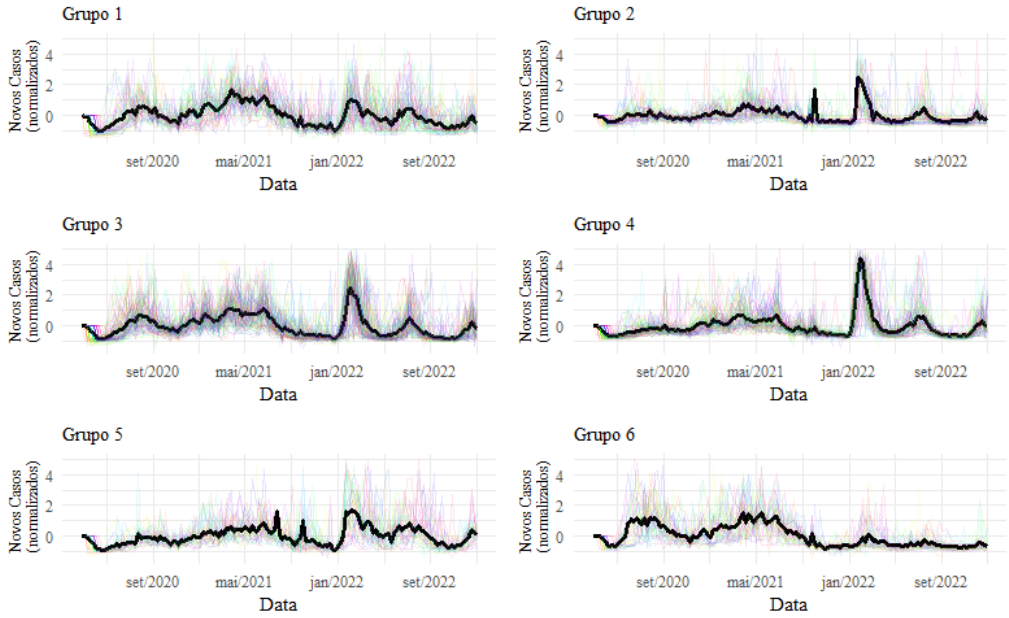

Figura 4.3: O diagrama mostra as séries temporais (normalizadas) de novos casos para cada grupo formado, em que a linha preta representa a série temporal média dos municípios de cada grupo

associam essa distribuição a vários fatores fatores como densidade populacional, condições de saneamento, níveis de educação e renda, entre outros. Suas análises sugerem que áreas com maior pobreza, densidade populacional elevada, e infraestrutura sanitária inadequada têm uma maior disseminação da doença. Além disso, também consideram a influência da religião, observando que práticas religiosas que envolvem aglomerações podem aumentar o risco de contágio. Portanto, a geografia urbana, juntamente com fatores socioeconômicos e culturais, podem ter desempenhado um papel significativo na propagação da doença.

4.2 Agrupamento de Novos Óbitos

Para as séries temporais de novos óbitos dos municípios presentes em nossa amostra foi seguido o mesmo procedimento descrito acima. Abaixo conseguimos ver o dendrograma com o resultado do agrupamento hierárquico desse novo conjunto de dados.

Infelizmente, nesta etapa os métodos de validação interna utilizados não foram tão condizentes com o resultado do dendrograma como ocorrido para os dados de

Tabela 4.1: Proporção do total de municípios de cada região da nossa amostra nos grupos formados a partir dos dados de novos casos

		Regiões				
Grupos	Tamanho do Grupo	CENTRO-OESTE	NORDESTE	NORTE	SUDESTE	SUL
Grupo 1	51	25%	16%	13%	17%	9%
Grupo 2	51	0%	19%	19%	21%	2%
Grupo 3	88	38%	47%	29%	18%	25%
Grupo 4	75	12%	9%	13%	24%	47%
Grupo 5	30	21%	3%	0%	9%	17%
Grupo 6	31	4%	6%	26%	12%	0%
Total	326	100%	100%	100%	100%	100%

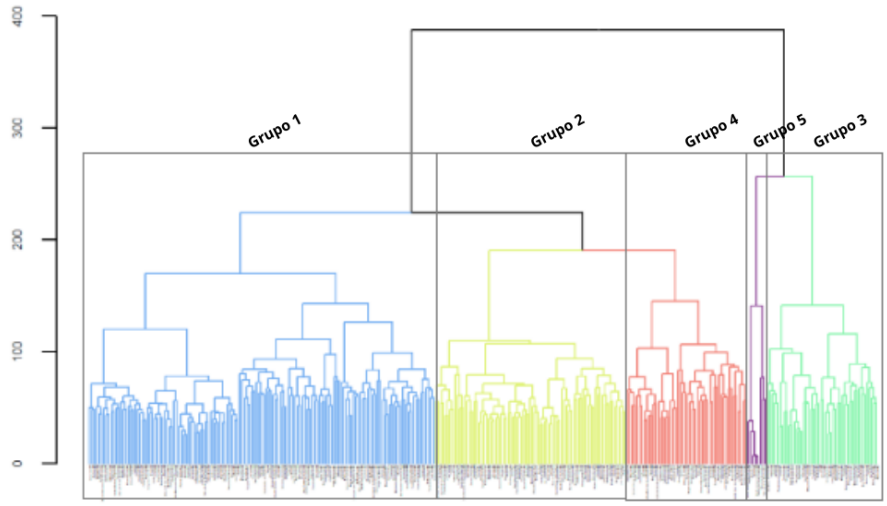

Figura 4.4: Dendrograma feito a partir do agrupamento hierárquico das séries temporais de novos casos dos 326 municípios de grande porte analisados

novos casos. Dessa forma, a escolha final de utilizarmos cinco grupos foi feita principalmente a partir do dendrograma e das alturas das segregações dos grupos.

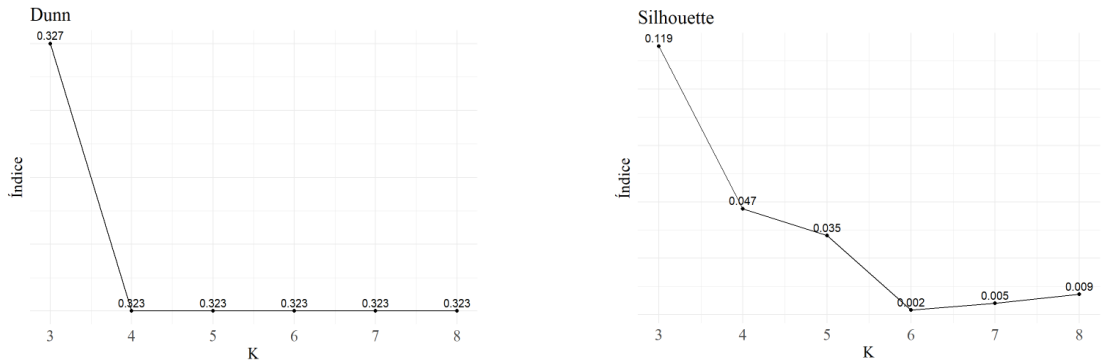

Figura 4.5: Resultado dos métodos de validação de clusters mencionados anteriormente

Na figura 4.6 conseguimos ver as séries temporais de novos óbitos em cada grupo

formado, nos permitindo analisar os diferentes padrões de evolução e intensidade. Os grupos 1 e 2 diferenciam-se principalmente pela intensidade da primeira onda de óbitos ocorrida em 2020, sendo esta significativamente maior na linha média do primeiro grupo. Com exceção do grupo 6, todos os demais grupos apresentaram uma terceira e quarta onda de novos casos ocorrida no início de 2022. Além disso, os grupos 2,4 e 5 tiveram uma evolução mais lenta e menor de novos casos na primeira e segunda onda que se manteve até o final de 2021.

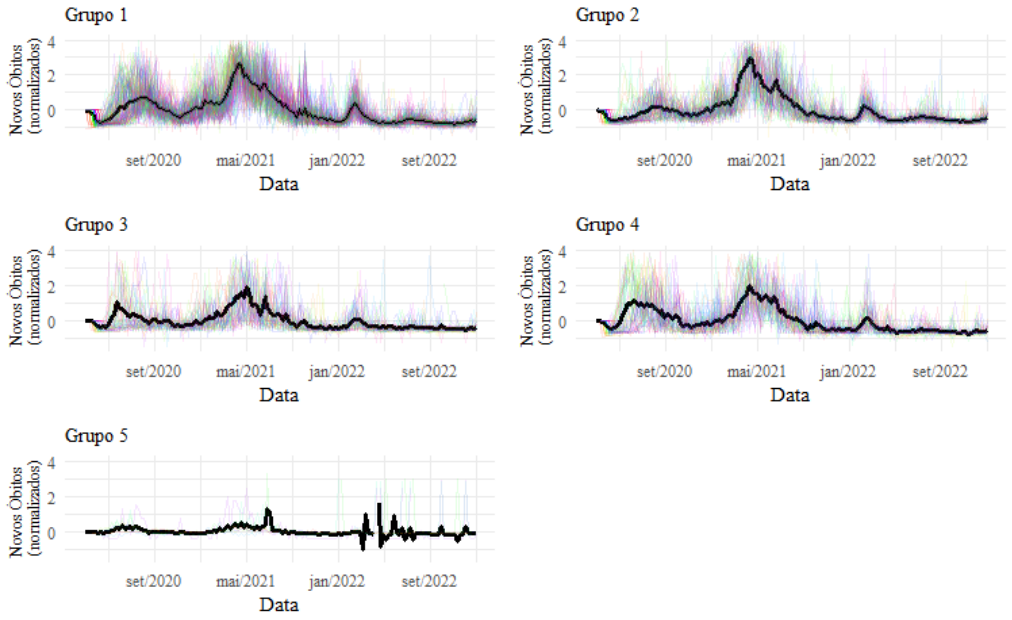

Figura 4.6: O diagrama mostra as séries temporais (normalizadas) de novos óbitos para cada grupo formado, em que a linha preta representa a série temporal média dos municípios de cada grupo

Os resultados da tabela 4.2 apresentam quais grupos possuem uma maior concentração dos municípios por região, e que consequentemente possuem um padrão de evolução semelhante à média de determinado grupo. Podemos notar que, dos municípios de grande porte do Centro-Oeste presentes em nossa amostra, 75% estão inclusos no grupo 1 e 17% no grupo 2, representando 92% das cidades dessa região que foram utilizados na análise. Já os municípios do Nordeste estão presentes principalmente no grupo 4 e 1, representando 47% e 27% do total dos municípios da região, respectivamente. Dos municípios do Norte, 52% foram alocados no grupo 3, enquanto que os municípios do Sudeste e Sul estão inclusos predominantemente no grupo 1 e 2.

Nas próximas subseções, iremos comparar os grupos formados a partir dos dados de novos óbitos utilizando as seguintes variáveis de cada município: IDHM, o partidarismo político nas eleições presidenciais de 2018 e proporção de indivíduos vulneráveis à pobreza

Tabela 4.2: Proporção do total de municípios de cada região da nossa amostra nos grupos formados a partir dos dados de óbitos

		Regiões				
Grupos	Tamanho do Grupo	CENTRO-OESTE	NORDESTE	NORTE	SUDESTE	SUL
Grupo 1	143	75%	27%	13%	53%	42%
Grupo 2	79	17%	8%	10%	27%	47%
Grupo 3	46	4%	9%	52%	12%	8%
Grupo 4	50	4%	47%	23%	6%	4%
Grupo 5	8	0%	9%	3%	1%	0%
Total	326	100%	100%	100%	100%	100%

4.2.1 IDHM

O Índice de Desenvolvimento Municipal (IDHM) foi calculado a partir dos dados municipais sobre educação, renda e longevidade, da mesma forma que o IDH. Os dados sobre o IDH, fornecidos pelo Atlas Brasil (Fundação João Pinheiro, 2022), foram coletados do Censo Demográfico de 2010, que é a fonte mais recente que inclui os índices de todos os municípios brasileiros.

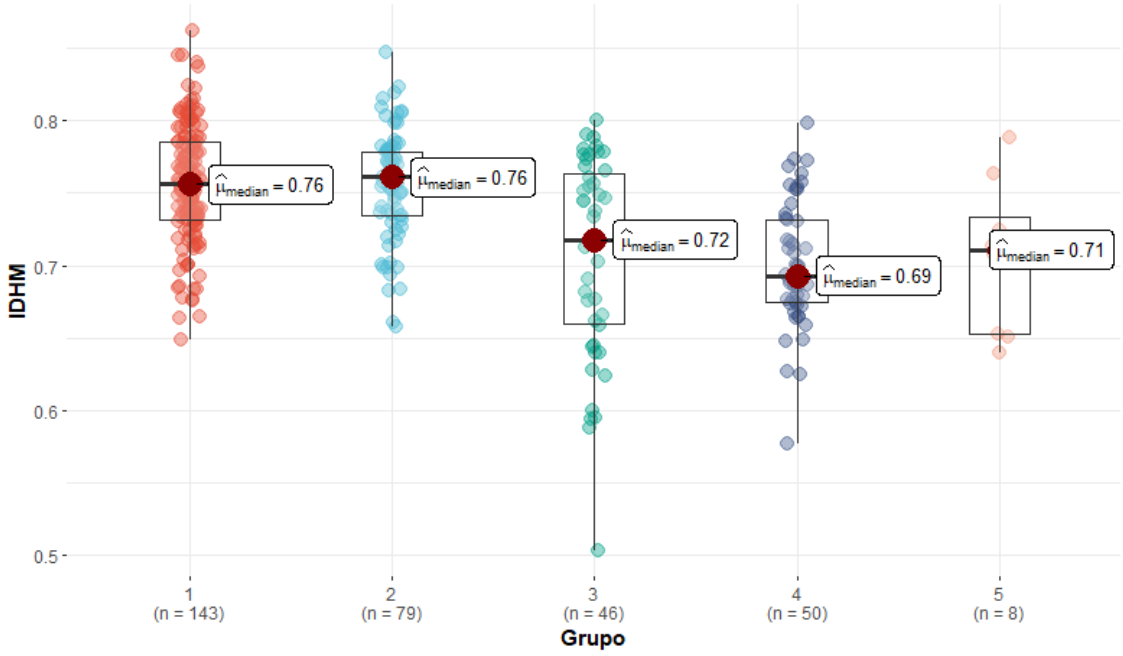

Figura 4.7: Box-plot do IDHM para cada grupo formado, junto da mediana $\hat{\mu}$ e do tamanho n de cada grupo

As análises desses resultados foram compiladas na Seção 4.2.4, além de uma discussão a respeito dos mesmos.

O teste de Kruskal-Wallis utilizado para comparar os grupos obteve um *p-valor* < 0.001, indicando que pelo menos uma das medianas dos grupos comparados na

Tabela 4.3: Resultados do teste de Dunn para comparações múltiplas do IDHM entre os grupos formados a partir dos dados de novos óbitos

Grupo	Grupo Comparado	Estatística do Teste	p-valor
1	2	-3.41	1
	3	-70.76	<0.001*
	4	-107.68	<0.001*
	5	-120.34	0.003*
2	3	-67.35	<0.001*
	4	-104.27	<0.001*
	5	-116.93	0.004*
3	4	-36.92	0.221
	5	-49.58	0.509
4	5	-12.66	1

figura 4.7 difere das demais. Para complementar o resultado, utilizamos o teste de Dunn para comparações múltiplas para verificar quais dos grupos diferem entre si, os resultados das comparações estão na tabela 4.3. As comparações revelam que, a um nível de significância de 5%, os grupos 1 e 2 têm diferenças significativas em relação aos grupos 3, 4 e 5. No entanto, as diferenças entre os grupos 4 e 5, assim como entre os grupos 1 e 2 e do grupo 3 em relação aos grupos 4 e 5, não são estatisticamente significativas, sugerindo que suas medianas são comparáveis.

4.2.2 Partidarismo Político

Os dados das urnas eletrônicas do segundo turno da eleição presidencial em 7 de outubro de 2018 foram obtidos do Repositório de Dados Eleitorais do Tribunal Superior Eleitoral (TSE). Apenas dois candidatos participaram desse segundo turno, a saber, Fernando Haddad (Partido dos Trabalhadores) e Jair Bolsonaro (Partido Social Liberal, na época). Esses dados foram convertidos em percentuais, onde o numerador se referia ao número de votos recebidos pelo candidato presidencial e o denominador era o número de eleitores aptos a participar na eleição no município na época.

O teste de Kruskal-Wallis utilizado para comparar os grupos também obteve um *p-valor* < 0.001, indicando que pelo menos uma das medianas dos grupos comparados na figura 4.8 difere das demais. Utilizamos novamente o teste de Dunn para comparações múltiplas para verificar quais dos grupos diferem entre si, os resultados das comparações estão na tabela 4.4. As comparações revelam que, a um nível de significância de 5%, novamente os grupos 1 e 2 têm diferenças significativas em relação aos grupos 3, 4 e 5. As diferenças entre os grupos 4 e 5, assim como entre os grupos 1 e 2 e do grupo 3 em relação aos grupos 4 e 5, não são estatisticamente significativas, sugerindo que suas medianas são comparáveis.

As análises desses resultados foram compiladas na Seção 4.2.4 a seguir.

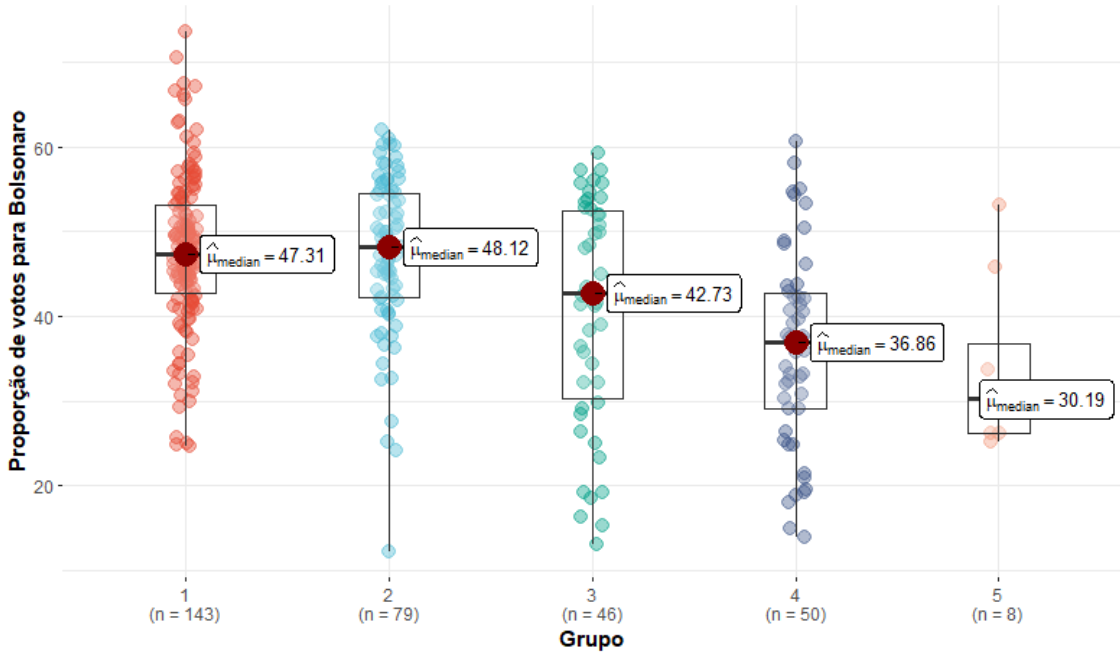

Figura 4.8: Box-plot da proporção de votos para Bolsonaro na eleições presidenciais de 2018 para cada grupo formado, junto da mediana $\hat{\mu}$ e do tamanho n de cada grupo

Tabela 4.4: Resultados do teste de Dunn para comparações múltiplas do partidarismo político entre os grupos formados a partir dos dados de novos óbitos

Grupo	Grupo Comparado	Estatística do Teste	p-valor
1	2	2.99	1
	3	-44.26	0.034*
	4	-87.44	<0.001*
	5	-105.52	0.015*
2	3	-47.25	0.034*
	4	-90.43	<0.001*
	5	-108.51	0.015*
3	4	-43.18	0.099
	5	-61.26	0.269
4	5	-18.08	1

4.2.3 Proporção de Indivíduos Vulneráveis à Pobreza

Proporção dos indivíduos com renda domiciliar per capita igual ou inferior a R$255,00 mensais, em reais de agosto de 2010, equivalente a 1/2 salário mínimo nessa data. O universo de indivíduos é limitado àqueles que vivem em domicílios particulares permanentes

O teste de Kruskal-Wallis utilizado para comparar os grupos também obteve um *p-valor* < 0.001, indicando que pelo menos uma das medianas dos grupos comparados na figura 4.9 difere das demais. Utilizamos novamente o teste de Dunn para comparações múltiplas para verificar quais dos grupos diferem entre si, os resultados

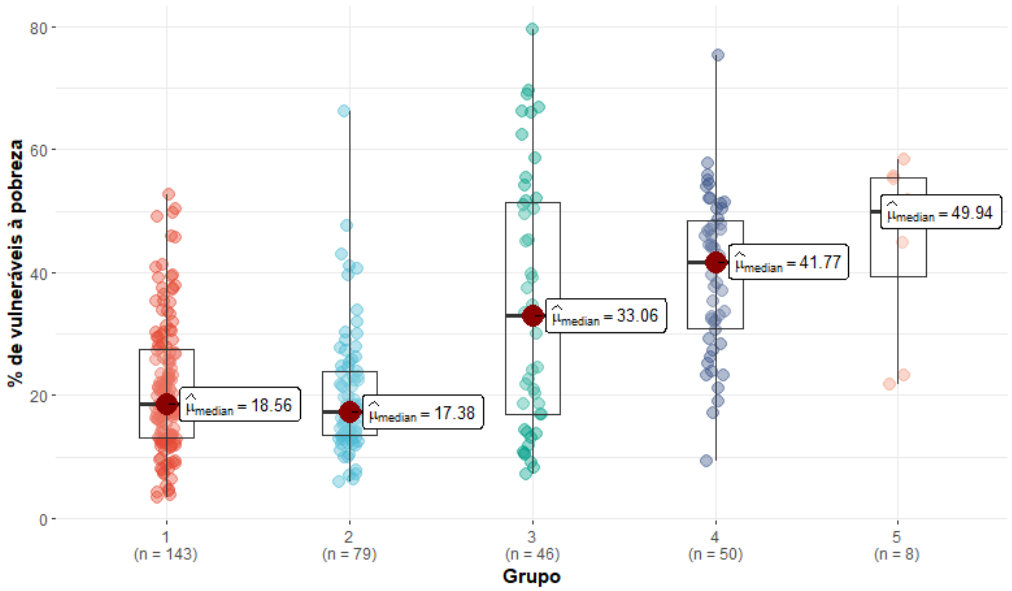

Figura 4.9: Box-plot da proporção de indivíduos vulneráveis à pobreza para cada grupo formado, junto da mediana $\hat{\mu}$ e do tamanho n de cada grupo

Tabela 4.5: Resultados do teste de Dunn para comparações múltiplas da proporção de indivíduos vulneráveis à pobreza entre os grupos formados a partir dos dados de novos óbitos

Grupo	Grupo Comparado	Estatística do Teste	p-valor
1	2	-9.70	0.926
	3	62.81	<0.001*
	4	113.35	<0.001*
	5	131.82	<0.001*
2	3	72.50	<0.001*
	4	123.05	<0.001*
	5	141.52	<0.001*
3	4	50.55	0.035*
	5	69.01	0.168
4	5	18.47	0.926

das comparações estão na tabela 4.5. As comparações revelam que, a um nível de significância de 5%, novamente os grupos 1 e 2 têm diferenças significativas em relação aos grupos 3, 4 e 5, também havendo diferença entre o grupo 3 e 4. As diferenças entre os grupos 4 e 5, assim como entre os grupos 1 e 2 e os grupos 3 e 5, não são estatisticamente significativas, sugerindo que suas medianas são comparáveis.

4.2.4 Discussão

A análise dos agrupamentos de novos casos de COVID-19 revela padrões distintos de evolução da pandemia nos diferentes municípios brasileiros com mais de 100 mil habitantes. O uso de técnicas de agrupamento hierárquico e a validação com o índice de Dunn e o coeficiente da Silhueta indicaram que seis grupos capturam de

forma adequada as similaridades entre as séries temporais. Observa-se que esses grupos refletem não apenas padrões temporais, mas também fatores regionais e socioeconômicos.

O agrupamento de novos óbitos também revelou padrões de evolução distintos em diferentes regiões do Brasil, com as séries temporais sendo organizadas em cinco grupos. Diferentemente dos novos casos, a validação interna com o índice de Dunn e o coeficiente da Silhueta apresentou resultados menos conclusivos, o que levou à escolha final dos grupos com base no dendrograma e nas alturas de segregação, destacando a complexidade da dinâmica de mortalidade em relação à COVID-19.

A partir do resultado do agrupamento dos dados de óbitos, os gráficos dos grupos formados a partir da variável de proporção de indivíduos vulneráveis à pobreza na figura 4.9 demonstraram um comportamento praticamente inverso aos gráficos do partidarismo político e IDHM. Podemos ver que os grupos 1 e 2 na figura 4.9 são os que tiveram as menores medianas de indivíduos vulneráveis à pobreza, enquanto foram os grupos que apresentaram as maiores medianas para partidarismo político (figura 4.8) e de IDHM (figura 4.7). Além disso, nas séries temporais médias dos grupos formados a partir de novos óbtios (figura 4.6), podemos ver que esses dois grupos também tiveram os maiores picos na segunda onda (ocorrida aproximadamente em Maio de 2021) quando comparados aos demais grupos.

Além disso, os resultados do teste Kruskal-Wallis que compara a mediana dos grupos de novos óbitos foi significativo a um nível de 1% de confiança para todas as variáveis utilizadas. Ademais, os grupos 1 e 2 terem diferenças significativas em relação aos grupos 3, 4 e 5 nas comparações múltiplas das três variáveis utilizadas reforçam a eficácia do agrupamento utilizado, em que também podemos observar padrões de evolução distintos entre os grupos na figura 4.6, sugerindo que o padrão de evolução da mortalidade por COVID-19 durante a pandemia pode ter sido influenciado por uma combinação de fatores sociais e políticos preexistentes.

Essas hipóteses podem ser investigadas através de estudos detalhados que analisam dados demográficos, comportamentais e de saúde em diferentes grupos sociais. A complexidade do fenômeno exige uma abordagem multifacetada para entender plenamente as razões por trás das diferenças nos padrões de evolução de mortalidade e novos casos de COVID-19 entre os grupos formados.

Os resultados condizentes com a pesquisa de Xavier et al. (2022) reforçam a viabilidade da metodologia que utilizamos e suas possíveis aplicações em cenários futuros como a COVID-19, tendo sua maior eficácia enquanto cenários pandêmicos ainda estão em desenvolvimento atuando em conjunto com técnicas de previsão de séries temporais (forecasting) como em Nikolopoulos et al. (2021) e dos Santos Gomes e de Oliveira Serra (2021).

5 Conclusão

Nosso estudo teve como objetivo explorar a viabilidade da utilização de métodos de agrupamento de séries temporais aplicados a dados relacionados a pandemia, especificamente para o cenário pandêmico no Brasil. Os resultados encontrados com essa metodologia são similares aos que foram encontrados por outros estudos relacionados ao mesmo tema mas que utilizaram metodologias diferentes, o que nos dá segurança ao analisar os resultados encontrados.

Os métodos de agrupamento em si não permitem tirar conclusões fortes a partir de seus resultados, mas devido ao seu caráter exploratório nos dão outra perspectiva e abrem um leque maior de possibilidades de modelagem a partir de seus resultados. Todos os resultados encontrados nesta pesquisa foram comparados com outros estudos e comprovam como o desenvolvimento de novos casos e óbitos tiveram padrões de evolução distintos pelo Brasil. As variáveis utilizadas que tiveram diferença estatisticamente significativa ao compararmos os grupos se mostram como possíveis candidatas a terem influência direta sobre o processo gerador da série temporal (média) de cada grupo. Entretanto, para tais afirmações é necessário a aplicação de outros métodos estatísticos que comprovem esta nova hipótese, além da colaboração de especialistas de outras áreas de conhecimento.

Em resumo, nosso estudo confirma a eficácia do agrupamento de séries temporais na análise de dados da pandemia no Brasil e esperamos que nossos achados possam servir de referência e suporte para outros pesquisadores, cientistas de dados e formuladores de políticas. Além disso, também esperamos que as ferramentas utilizadas também possam ser eficientes em pesquisas futuras.

Referências Bibliográficas

Aghabozorgi, S., Shirkhorshidi, A. S., e Wah, T. Y. (2015). Time-series clustering – a decade review. *Elsevier*.

Arbelaitz, O., Gurrutxaga, I., Muguerza, J., Pérez, J. M., e Perona, I. (2013). An extensive comparative study of cluster validity indices. *Pattern Recognition*, 46(1):243–256.

Ashouri, M. e Phoa, F. K. H. (2022). Interactive tool for clustering and forecasting patterns of taiwan covid-19 spread. *Plos one*, 17(6):e0265477.

Caiado, J. e Lúcio, F. (2023). Stock market forecasting accuracy of asymmetric garch models during the covid-19 pandemic. *The North American Journal of Economics and Finance*, 68:101971.

Cassão, V., Alves, D., de Andrade Mioto, A. C., Bernardi, F. A., e Brandão Miyoshi, N. S. (2022). Unsupervised analysis of covid-19 pandemic evolution in brazilian states. *Procedia Computer Science*, 196:655–662. International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies 2021.

Cota, W. (2020). Monitoring the number of COVID-19 cases and deaths in brazil at municipal and federative units level. *SciELOPreprints:362*.

dos Santos Gomes, D. C. e de Oliveira Serra, G. L. (2021). Machine learning model for computational tracking and forecasting the covid-19 dynamic propagation. *IEEE Journal of Biomedical and Health Informatics*, 25(3):615–622.

Faceli, K. (2021). *Inteligência Artificial – Uma Abordagem de Aprendizado de Máquina*. Grupo GEN.

Fan, C., Zhang, D., e Zhang, C. (2011). On sample size of the kruskal-wallis test with application to a mouse peritoneal cavity study. *Biometrics*, 67(1):213–224.

Ferrante, L., Duczmal, L., Steinmetz, W. A., Almeida, A. C. L., Leão, J., Vassão, R. C., Tupinambás, U., e Fearnside, P. M. (2021). How brazil's president turned the country into a global epicenter of covid-19. *Journal of Public Health Policy*.

François Petitjean, Alain Ketterlin, P. G. (2011). A global averaging method for dynamic time warping, with applications to clustering. *Pattern Recognition*.

Fundação João Pinheiro (2022). ATLAS Brasil.

James, G., Witten, D., Hastie, T., e Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*, volume 103 of *Springer Texts in Statistics*. Springer, New York.

Jin, Q. (2020). Time warping clustering for the forecast and analysis of covid-19. In *2020 IEEE MIT Undergraduate Research Technology Conference (URTC)*, pages 1–5. IEEE.

Kaufman, L. e Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction To Cluster Analysis*.

Keogh, E. (2002). Chapter 36 - exact indexing of dynamic time warping. In Bernstein, P. A., Ioannidis, Y. E., Ramakrishnan, R., e Papadias, D., editors, *VLDB '02: Proceedings of the 28th International Conference on Very Large Databases*, pages 406–417. Morgan Kaufmann, San Francisco.

Khmaissia, F., Haghighi, P. S., Jayaprakash, A., Wu, Z., Papadopoulos, S., Lai, Y., e Nguyen, F. T. (2020). An unsupervised machine learning approach to assess the zip code level impact of covid-19 in nyc. *arXiv preprint arXiv:2006.08361*.

Kurniawan, R., Sheikh Abdullah, S. N. H., Lestari, F., Nazri, M. Z. A., Mujahidin, A., e Adnan, N. (2020). Clustering and correlation methods for predicting coronavirus covid-19 risk analysis in pandemic countries. In *2020 8th International Conference on Cyber and IT Service Management (CITSM)*, pages 1–5.

Liu, D., Clemente, L., Poirier, C., Ding, X., Chinazzi, M., Davis, J. T., Vespignani, A., e Santillana, M. (2020). A machine learning methodology for real-time forecasting of the 2019-2020 covid-19 outbreak using internet searches, news alerts, and estimates from mechanistic models. *arXiv preprint arXiv:2004.04019*.

Lucic, M. C., Ghazzai, H., Lipizzi, C., e Massoud, Y. (2021). Integrating county-level socioeconomic data for covid-19 forecasting in the united states. *IEEE Open Journal of Engineering in Medicine and Biology*, 2:235–248.

Luo, Z., Zhang, L., Liu, N., e Wu, Y. (2023). Time series clustering of covid-19 pandemic-related data. *Data Science and Management*, 6(2):79–87.

Malta, M., Vettore, M. V., da Silva, C. M. F. P., Silva, A. B., e Strathdee, S. A. (2021). Political neglect of COVID-19 and the public health consequences in Brazil: The high costs of science denial. *The Lancet*.

Nicholson, C., Beattie, L., Beattie, M., Razzaghi, T., e Chen, S. (2022). A machine learning and clustering-based approach for county-level covid-19 analysis. *Plos one*, 17(4):e0267558.

Nikolopoulos, K., Punia, S., Schäfers, A., Tsinopoulos, C., e Vasilakis, C. (2021). Forecasting and planning during a pandemic: Covid-19 growth rates, supply chain disruptions, and governmental decisions. *European journal of operational research*, 290(1):99–115.

Pereira, I. G., Guerin, J. M., Silva Júnior, A. G., Garcia, G. S., Piscitelli, P., Miani, A., Distante, C., e Gonçalves, L. M. G. (2020). Forecasting covid-19 dynamics in brazil: A data driven approach. *International Journal of Environmental Research and Public Health*, 17(14).

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zhu, Q., Zakaria, J., e Keogh, E. (2012). Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 262–270, New York, NY, USA. Association for Computing Machinery.

Rios, R. A., Nogueira, T., Coimbra, D. B., Lopes, T. J., Abraham, A., e Mello, R. F. d. (2021). Country transition index based on hierarchical clustering to predict next covid-19 waves. *Scientific reports*, 11(1):15271.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.

Sadeghi, B., Cheung, R. C., e Hanbury, M. (2021). Using hierarchical clustering analysis to evaluate covid-19 pandemic preparedness and performance in 180 countries in 2020. *BMJ open*, 11(11):e049844.

Sardá-Espinosa, A. (2019). Time-Series Clustering in R Using the dtwclust Package. *The R Journal*, 11(1):22–43.

Soto-Ferrari, M., Carrasco-Pena, A., e Prieto, D. (2023). Aggforclus: A hybrid methodology integrating forecasting with clustering to assess mitigation plans and contagion risk in pandemic outbreaks: the covid-19 case study. *Journal of Business Analytics*, 6(3):217–242.

Théry, H. (2020). Quels sont les facteurs associés à la propagation de l'épidémie de covid-19 au brésil? *Diploweb.com: la revue géopolitique*. Disponível em: https://www.diploweb.com.

Warren Liao, T. (2005). Clustering of time series data—a survey. *Pattern Recognition*, 38(11):1857–1874.

Xavier, D. R., e Silva, E. L., Lara, F. A., e Silva, G. R., Oliveira, M. F., Gurgel, H., e Barcellos, C. (2022). Involvement of political and socio-economic factors in the spatial and temporal dynamics of covid-19 outcomes in brazil: A population-based study. *The Lancet Regional Health-Americas*, page 100221.

Yang, C.-Y., Chen, P.-Y., Wen, T.-J., e Jan, G. E. (2019). Imu consensus exception detection with dynamic time warping—a comparative approach. *Sensors*, 19(10).