UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA
DEPARTAMENTO DE ESTATÍSTICA

Matheus de Azevedo Pessoa

SEGMENTAÇÃO DOS MUNICÍPIOS BRASILEIROS PELO PADRÃO DA DISSEMINAÇÃO INICIAL DA COVID-19

Recife
2023

MATHEUS DE AZEVEDO PESSOA

SEGMENTAÇÃO DOS MUNICÍPIOS BRASILEIROS PELO PADRÃO DA DISSEMINAÇÃO INICIAL DA COVID-19

Trabalho de conclusão de curso apresentado ao Curso de Bacharelado em Estatística da Universidade Federal de Pernambuco como requisito parcial para obtenção do título de Bacharel em Estatística.

Universidade Federal de Pernambuco - UFPE

Orientador: Prof. Dr. Cristiano Ferraz

Coorientador: Prof. Dr. André Leite Wanderley

Recife-PE

Setembro de 2023

Ficha de identificação da obra elaborada pelo autor,
através do programa de geração automática do SIB/UFPE

Pessoa, Matheus de Azevedo .
Segmentação dos municípios brasileiros pelo padrão da disseminação inicial da COVID-19 / Matheus de Azevedo Pessoa. - Recife, 2023.
24 p. : il., tab.

Orientador(a): Cristiano Ferraz
Cooorientador(a): André Leite Wanderley
Trabalho de Conclusão de Curso (Graduação) - Universidade Federal de Pernambuco, Centro de Ciências Exatas e da Natureza, Estatística - Bacharelado, 2023.

1. Modelos de segmentação. 2. Algoritmo K-means. 3. Padrão de crescimento. 4. COVID-19. 5. Munícipios brasileiros. I. Ferraz, Cristiano. (Orientação). II. Wanderley, André Leite. (Coorientação). IV. Título.

310 CDD (22.ed.)

Agradecimentos

Primeiramente, quero expressar minha imensa gratidão à minha mãe, que foi a parte fundamental de minha formação. Seu amor incondicional, sacrifícios e ensinamentos moldaram o indivíduo que sou hoje. Cada gesto, cada palavra de incentivo e cada sacrifício que fez por mim foram a força motriz por trás de cada passo que dei.

Ao meu pai, que mesmo não estando mais entre nós, deixou um legado de valores e ensinamentos que foram fundamentais. Sinto sua falta todos os dias, mas sei que você abençoa e guia meu caminho de onde estiver.

Aos meus irmãos, Gabriel e Luiz, minha eterna gratidão. Seu apoio inabalável e companhia foram a chama que me manteve aceso nos momentos mais sombrios.

Um agradecimento especial ao meu orientador prof. Dr. Cristiano Ferraz e coorientador prof. Dr. André Leite. Sua dedicação, paciência, disponibilidade e sabedoria foram cruciais e os pilares que sustentaram este trabalho desde o início.

A Vinícius, Marcelo, Lucas e Hygor, companheiros de turma e amigos: enfrentamos desafios, compartilhamos conhecimentos e vivemos experiências memoráveis juntos. Cada um de vocês teve uma contribuição essencial em minha trajetória acadêmica, especialmente Vinícius. Sou eternamente grato pelos momentos que compartilhamos e pela constante presença e apoio de todos vocês.

À Gabriel Leite, Matheus Leite, Gabriel Teotônio, Tharso Rossiter e Raíza Oliveira. Vocês foram mais do que colegas, foram mentores e amigos que me guiaram com conselhos valiosos, ensinamentos e apoio durante a minha graduação. Em muitos momentos, quando quis desistir foram as suas palavras e exemplos que me reacenderam a chama da determinação. Vocês são verdadeiras inspirações para mim e me mostraram que, com perseverança e apoio mútuo, podemos superar qualquer desafio.

À Malu Galindo, em cada desafio enfrentado, em cada obstáculo superado, você esteve ao meu lado, oferecendo apoio inabalável. A jornada foi repleta de altos e baixos, e em muitos momentos, a ideia de desistir parecia tentadora. No entanto, com a força e o carinho que recebi de você, encontrei a coragem para seguir em frente. Os momentos compartilhados durante a graduação serão sempre lembrados com carinho. Você e toda sua família são partes disso.

À Stella, Carol, Thaysa, Hélio e tantos outros, obrigado pelos momentos de alegria, pelos sorrisos compartilhados e pelas memórias inesquecíveis que construímos juntos durante a graduação.

A todos os outros familiares e amigos queridos como os do Gera, Talaiba, Virtus e

amigos-de-rods, mesmo sem mencionar cada nome individualmente para evitar o risco de esquecer alguém, minha profunda gratidão. Agradeço por cada conversa, cada gesto de apoio, cada momento de descontração e todas as memórias que construímos juntos. Ter vocês por perto foi essencial para manter meu equilíbrio e bem-estar durante os anos de faculdade e em todos os aspectos da vida. Vocês representam tanto a família que a vida me deu quanto a que escolhi para mim. Vocês são meu porto seguro e meu refúgio!

E, por fim, a todos que, direta ou indiretamente, cruzaram meu caminho e contribuíram para esta conquista, meu mais profundo agradecimento. Cada interação, cada palavra e cada gesto foram essenciais para que eu chegasse até aqui.

Resumo

Este presente trabalho estuda a dinâmica da disseminação inicial da COVID-19 nos primeiros trinta dias da pandemia em municípios brasileiros. Para essa análise, foram ajustados modelos de regressão linear, regressão exponencial e modelo de gompertz, gerando coeficientes de determinação R^2 para cada município. Esses coeficientes indicam a adequação dos padrões observados de disseminação aos modelos. Utilizando os coeficientes de determinação e o coeficiente β_1 como variáveis de entrada, foi possível fazer duas segmentação para categorizar os municípios. Essa abordagem se beneficiou do uso do algoritmo de agrupamento *K-means*, que agrupou municípios com padrões de disseminação semelhantes. Com a finalidade de investigar a interação entre a evolução dos casos acumulados e os contextos demográficos, realizou-se uma análise exploratória para entender as diferenças demográficas entre os agrupamentos. Os resultados proporcionam uma visão inicial para orientar intervenções e a alocação de recursos. Espera-se que trabalhos futuros explorem a abordagem deste trabalho utilizando modelos mais robustos e refinados.

Palavras-chave: COVID-19, Brasil, DATASUS, *K-means*, Modelos lineares, Modelos não lineares, Censo, IBGE, Pandemia

Abstract

This current study examines the dynamics of the initial spread of COVID-19 during the first thirty days of the pandemic in Brazilian municipalities. For this analysis, linear regression, exponential regression, and Gompertz models were adjusted, generating determination coefficients R^2 for each municipality. These coefficients indicate the fit of the observed spread patterns to the models. Using the determination coefficients and the coefficient β_1 as input variables, it was possible to perform two segmentations to categorize the municipalities. This approach benefited from the use of the *K-means* clustering algorithm, which grouped municipalities with similar spread patterns. To investigate the interaction between the evolution of accumulated cases and demographic contexts, an exploratory analysis was conducted to understand the demographic differences between the clusters. The results provide an initial insight to guide interventions and optimize resource allocation. Future work is expected to further explore this study's approach using more robust and refined models.

Keywords: COVID-19, Brazil, DATASUS, *K-means*, Linear models, Non-linear models, Census, IBGE, Pandemic

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

A COVID-19, causada pelo coronavírus SARS-CoV-2, emergiu no final de 2019 na cidade de Wuhan, China, e rapidamente se tornou uma pandemia global, afetando milhões de pessoas em todo o mundo (ZHOU et al., 2020). Esta doença respiratória apresenta uma ampla gama de sintomas, desde casos assintomáticos ou leves até quadros graves que podem levar à morte, especialmente em indivíduos idosos ou com comorbidades.

A rápida disseminação do vírus e seu impacto substancial na saúde pública global levaram a uma mobilização sem precedentes da comunidade científica. Pesquisadores de todo o mundo têm trabalhado incansavelmente para entender a biologia do vírus, os mecanismos de transmissão, desenvolver tratamentos eficazes e vacinas (COHEN, 2020).

Além dos desafios de saúde, a pandemia também trouxe implicações socioeconômicas significativas, com muitos países implementando medidas de bloqueio rigorosas que afetaram a economia global (NICOLA et al., 2020).

Desde que a pandemia emergiu globalmente, diversos países enfrentaram inúmeros desafios em sua gestão e controle sendo considerada uma pandemia mundial. O Brasil, com sua diversidade demográfica e socioeconômica, não foi exceção. Em território brasileiro, a propagação do SARS-CoV-2 não somente evidenciou as limitações dos sistemas de saúde, mas também ressaltou as desigualdades.

No Brasil, a situação foi particularmente grave, com o país sendo um dos mais impactados pela pandemia. Até o início de 2023, o Brasil registrou quase 700.000 mortes devido à doença, uma cifra alarmante que reflete a complexidade e os desafios enfrentados pelo país durante essa crise de saúde (CRIBARI-NETO, 2023)

A variedade na progressão da doença em diferentes municípios brasileiros levantou questões pertinentes sobre como fatores locais, como infraestrutura de saúde e densidade populacional, poderiam estar moldando a dinâmica da doença. O distanciamento social, uma das principais estratégias de mitigação, afetou gravemente os negócios e a economia em geral, levando a consequências significativas na saúde mental da população, com relatos de aumento de estresse, ansiedade e outros problemas relacionados.

Diante desse cenário, pesquisadores têm trabalhado para entender e modelar a progressão da doença no país. (CRIBARI-NETO, 2023) destaca a construção de um modelo de regressão beta para analisar a mortalidade por COVID-19 nas unidades federativas brasileiras, oferecendo visões valiosos sobre os fatores que influenciam as taxas de mortalidade em diferentes estados.

Este trabalho surge com o propósito de analisar os primeiros trinta dias a partir do

primeiro caso em cada município do Brasil, período crucial onde as bases para a gestão do surto foram estabelecidas. Estes primeiros momentos são fundamentais, pois estabelecem o tom da resposta à crise e indicam tendências. Assim, utilizando dos casos acumulados para os primeiros trinta dias foram ajustados modelos estatísticos como a Regressão Linear, Regressão Exponencial e Modelo de Gompertz. O quanto os modelos explicam a variabilidade dos dados, representada pelo coeficiente de determinação R^2 e o coeficiente β_1, foram utilizados como indicadores chaves.

A principal inovação deste estudo é a incorporação dos coeficientes de determinação e dos coeficientes de regressão β_1 como variáveis de entrada em um modelo de classificação. Utilizando o algoritmo *K-means*, realizamos a categorização dos municípios com base na aderência de seus modelos estatísticos aos casos acumulados durante os primeiros trinta dias. Desta forma, identificamos áreas que exibiram padrões exponenciais, lineares ou que demonstraram comportamento conforme previsto pelo modelo Gompertz. Adicionalmente, outra categorização considerou a taxa de crescimento, permitindo a observação dos municípios que mais contribuíram para a disseminação da doença no início do período analisado.

Com a finalidade de observar a interação entre a evolução dos casos acumulados e os contextos demográficos, utilizando dados do Ministério da Saúde e do IBGE realizou-se uma análise exploratória para entender as diferenças demográficas entre os agrupamentos.

2 Metodologia

A metodologia é uma parte crucial de qualquer pesquisa ou estudo, pois fornece um roteiro detalhado das etapas e procedimentos adotados para alcançar os objetivos propostos. Ela permite que outros pesquisadores compreendam e reproduzam o estudo, garantindo sua validade e confiabilidade. Nesta seção, descreveremos minuciosamente os métodos e técnicas utilizados, isso proporcionará uma visão transparente de como os resultados foram obtidos.

2.1 Modelos de Regressão

A análise de Regressão é um método que explora e esclarece as relações entre variáveis. As aplicações são diversas e ocorrem em quase todos os campos, incluindo engenharia, ciências físicas e químicas, economia, ciências biológicas e ciências sociais. (MONTGOMERY et al., 2013)

Neste estudo, foram aplicados modelos de regressão a cada município brasileiro para analisar como cada um descrevia a variabilidades nos casos de COVID-19 ao longo do tempo e com isso, conseguirmos segmentar os municípios baseados nesses padrões de disseminação. É importante destacar que, para diversos municípios cujos primeiros 30 dias de registros ocorreram no início de 2020, a contabilização de casos estava fortemente atrelada ao número de óbitos e aos testes realizados em profissionais da saúde. Isso ocorre porque os testes em massa para a população em geral só se tornaram comuns no final daquele ano.

Em doenças contagiosas, inicialmente os padrões de crescimento tendem a ser não-lineares. Assim, optar também por modelos não-lineares pode oferecer informações valiosas sobre a progressão da doença, especialmente nos estágios iniciais. Portanto, para cada município, foram ajustados os seguintes modelos: Regressão Linear, Regressão Exponencial e Modelo de Gompertz.

Para os estimadores dos mínimos quadrados serem os estimadores lineares não enviesados e eficiente, é preciso que os erros satisfaçam um conjunto de condições. Não ter viés indica que, em média, as estimativas não se desviam sistematicamente do valor verdadeiro, garantindo que não haja erros sistemáticos ou tendências que possam distorcer os resultados. Já ser eficiente indica que as estimativas possuem as menores variâncias possíveis entre todas as estimativas não enviesadas.

- $\mathbb{E}(\epsilon_i = 0)\ \forall i \in \{1, \ldots, n\}$: O valor esperado dos erros precisa ser zero. Isso garante que

o modelo está capturando toda a informação relevante das variáveis independentes.

- $\text{Var}(\epsilon_i) = \sigma^2 \ \forall i \in \{1, \ldots, n\}$: A variância dos erros é constante em todas as observações, indicando homoscedasticidade.
- $\text{Cov}(\epsilon_i, \epsilon_j) = 0 \ \forall i, j \in \{1, \ldots, n\}$ e $i \neq j$: Os erros são não auto correlacionados, ou seja, o erro associado a uma observação não está correlacionado com o erro de qualquer outra observação.

2.1.1 Regressão Linear

A regressão linear é um método estatística que modela e analisa as relações entre duas ou mais variáveis. Em uma regressão linear simples, focamos na relação linear entre uma variável dependente e uma variável independente A ideia central da regressão linear é encontrar a melhor reta que se ajusta a um conjunto de pontos de dados. (DRAPER et al., 1998). Esta reta, é representada por:

$$y_t = \beta_0 + \beta_1 t + \epsilon_t.$$

No qual:

- y_t é a variável dependente, o número acumulado de casos de COVID-19 por dia a partir do primeiro caso.
- t é a variável independente representando o número de dias desde o primeiro caso registrado no município.
- β_0 é o intercepto. Ou seja, o ponto onde a linha de regressão cruza o eixo y.
- β_1 é o coeficiente associado ao número de dias.
- ϵ_t é o erro aleatório.

Para determinar a reta de melhor ajuste, utilizamos o método dos mínimos quadrados seguindo as condições descritas em 2.1.

2.1.2 Regressão Exponencial

A regressão exponencial é similar a regressão linear mas utilizado para analisar relações entre variáveis quando esta relação entre as variáveis pode ser descrita por uma função exponencial. Uma das principais vantagens é sua capacidade de modelar tendências não lineares, o que a torna adequada para muitos fenômenos naturais e sociais.

A forma geral da equação de regressão exponencial é descrita por:

$$y_t = \beta_0 e^{\beta_1 t} \epsilon_t.$$

Entretanto, podemos linearizar para utilizar o método dos mínimos quadrados da seguinte forma:

$$\log(y_t) = \beta_0 + \beta_1 t + \epsilon_t.$$

No qual:

- y_t é a variável dependente, o número acumulado de casos de COVID-19 por dia a partir do primeiro caso.
- t é a variável independente representando o número de dias desde o primeiro caso registrado no município.
- β_0 é o intercepto. Ou seja, o ponto onde a linha de regressão cruza o eixo y.
- β_1 é o coeficiente associado ao número de dias.
- ϵ_t é o erro aleatório.

Para determinar a reta de melhor ajuste nessa linearização, utilizamos o método dos mínimos quadrados seguindo as condições descritas em 2.1.

2.1.3 Modelo de Gompertz

O Modelo de Gompertz é uma função de crescimento sigmoidal frequentemente empregada em epidemiologia, demografia e outras áreas para descrever processos de crescimento que começam rapidamente e depois desaceleram à medida que se aproximam de um limite assintótico. Proposto por Benjamin Gompertz (GOMPERTZ, 1825), este modelo foi inicialmente desenvolvido para estudar a dinâmica populacional mas desde então tem sido aplicado em diversos contextos como por exemplo o crescimento de células (ROSSI et al., 2003).

Caracteristicamente, o modelo apresenta uma fase inicial de crescimento exponencial, seguida por uma desaceleração até que o crescimento se estabilize em um valor assintótico. Como está sendo trabalhado apenas com os primeiros trinta dias da pandemia em cada município, não é possível observar a estabilização no valor assintótico.

A função de Gompertz é definida em (ROSSI et al., 2003) como:

$$y_t = K e^{(-e^{(\beta_0 + \beta_1 t + \epsilon_t)})}.$$

Entretanto, podemos linearizar para utilizar o método dos mínimos quadrados da seguinte forma:

$$\log\left(\log\left(\frac{K}{y_t}\right)\right) = \beta_0 + \beta_1 t + \epsilon_t.$$

No qual:

- y_t é a variável dependente, o número acumulado de casos de COVID-19 por dia a partir do primeiro caso.
- K é o valor assintótico que y pode alcançar, onde nesse estudo utilizamos K assumindo o valor total de casos $+1$. Ou seja, somando 1 ao total de casos confirmados de COVID-19 no município até o dia da extração dos dados, 16 de Julho de 2023.
- t é a variável independente representando o número de dias desde o primeiro caso registrado no município.
- β_0 é o intercepto. Ou seja, o ponto onde a linha de regressão cruza o eixo y.
- β_1 é o coeficiente associado ao número de dias.
- ϵ_t é o erro aleatório.

Para determinar a reta de melhor ajuste nessa linearização, utilizamos o método dos mínimos quadrados seguindo as condições descritas em 2.1.

2.1.4 Coeficiente de determinação R^2

O coeficiente de determinação, denotado por R^2, é uma métrica estatística que quantifica a proporção da variância na variável dependente que é previsível a partir das variáveis independentes. Em outras palavras, R^2 oferece uma medida de quão bem o modelo explica a variabilidade observada na variável independente. (MONTGOMERY et al., 2013).

O coeficiente de determinação é definido como:

$$R^2 = 1 - \frac{\text{Soma dos quadrados dos resíduos}}{\text{Soma total dos quadrados}} = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2}.$$

Em que:

- y_t é o valor observado da i-ésima observação.
- $\hat{y}_t$ é o valor previsto da i-ésima observação.
- $\bar{y}$ é a média dos valores observados.

Um valor de R^2 próximo de 1 indica que uma grande proporção da variabilidade na variável dependente é explicada pelo modelo, enquanto um valor próximo de 0 sugere o contrário. (MONTGOMERY et al., 2013).

2.2 Métodos de Agrupamentos

A análise de agrupamento, é um método estatístico que busca identificar padrões intrínsecos em um conjunto de dados. O objetivo principal é agrupar dados semelhantes em grupos, de modo que os dados em um mesmo grupo sejam mais semelhantes entre si do que com os dados em outros grupos.

Utilizando-se do coeficiente de determinação dos modelos de regressão de cada município, esse estudo propõe utilizar da análise de agrupamentos para segmentar os municípios baseando-se nos padrões de disseminação da doença nos primeiros dias.

A classificação das observações em grupos requer alguns métodos de cálculo de distância ou similaridade entre cada par de observações. O resultado de este cálculo é conhecido como matriz de dissimilaridade ou distância. Existem muitos métodos para calcular essas informações de distância. (KASSAMBARA, 2017)

2.2.1 Distância euclidiana

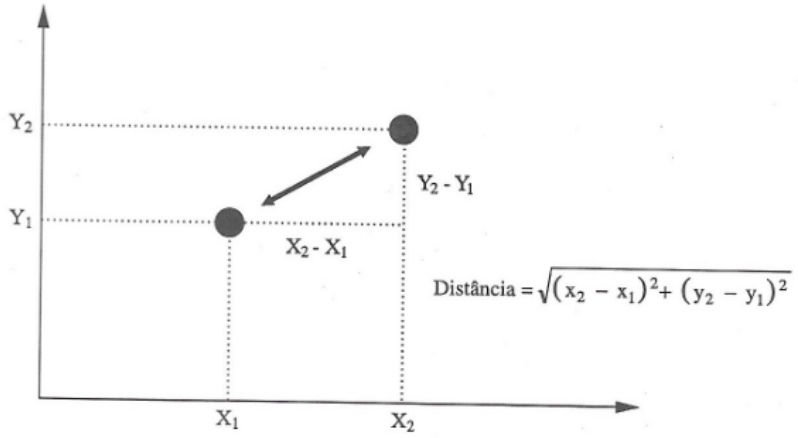
A escolha das medidas de distância é um passo essencial no agrupamento. Ele define como a similaridade de duas unidades é calculada e influenciará a forma dos aglomerados. (KASSAMBARA, 2017)

A distância euclidiana é uma das métricas mais comuns e intuitivas usadas para medir a distância entre dois pontos. Ela é derivada a geometria euclidiana e é frequentemente usada em contextos de agrupamento devido à sua simplicidade e eficácia.

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

Em que x e y são dois vetores de tamanho n.

Figura 1 – Distância Euclidiana



Fonte: (PEREIRA, 2004)

2.2.2 Algoritmo *K-means*

O agrupamento *K-means* (MACQUEEN, 1967) é o método de agrupamento mais utilizado, a sua simplicidade e natureza intuitiva tornam o *K-means* fácil de compreender e implementar. O algoritmo opera atribuindo pontos de dados ao grupo mais próximo com base na distância, onde nesse estudo, será a distância euclidiana.

Este trabalho não tem a intenção de fornecer uma explicação detalhada sobre o algoritmo. Ele é baseado e se alinha com as informações apresentadas em (KASSAMBARA, 2017). Para uma compreensão mais aprofundada, é recomendado consultar diretamente a referência.

No entanto, é essencial reconhecer que o *K-means* tem suas limitações. Uma das principais restrições, o que pode ser desafiador em situações onde a estrutura dos dados não é claramente conhecida. Além disso, o algoritmo é sensível à inicialização dos centroides, o que significa que diferentes inicializações podem levar a resultados variados e por lidar com médias o algoritmo é sensível a *outliers*.

O objetivo principal do algoritmo *K-means* é agrupar os dados de forma que a soma das variações totais dentro dos grupos seja minimizada. Esta medida representa a dispersão acumulada dos pontos de dados em torno dos centroides de todos os grupos.

Em (KASSAMBARA, 2017), dado que k é o total de grupos, foi definido a variação total dentro de um grupo como a soma das distâncias quadradas das distâncias entre os itens e o correspondente centroide:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2. \tag{2.1}$$

Em que:

- x_i: É uma observação dentro do grupo C_k.
- μ_i: é o valor médio dos pontos atribuídos ao grupo C_k.

Cada observação é atribuída a um determinado grupo de modo que a soma dos quadrados da distância da observação aos centros dos grupos atribuídos seja mínima. Definimos a soma das variação totais sendo:

$$\sum_{k=1}^{k} W(C_k) = \sum_{k=1}^{k} \sum_{x_i \in C_k} (x_i - \mu_k)^2. \tag{2.2}$$

O *K-means* busca minimizar a soma total das variações através de um processo iterativo:

- **Inicialização**: Selecione aleatoriamente k pontos de dados aleatórios como centroides iniciais, onde k será o número de grupos.
- **Atribuição**: Atribua cada observação ao centroide mais próximo, com base na distância euclidiana entre a observação e o centroide.
- **Atualização de Centroides**: Calcule a observação média de todas as observações em cada grupo e defina essa observação como o novo centroide.
- **Convergência**: Repita os passos de atribuição e atualização até que os centroides não mude significativamente entre as iterações ou até que um número máximo de iterações seja alcançado.

A cada iteração, a soma das variações é reduzida até que o algoritmo convirja para uma solução em que a variação não pode ser reduzida significativamente por mais reatribuições ou atualizações de centroides.

A escolha inicial dos centroides pode levar a mínimos locais na função objetivo, o que não necessariamente será a melhor solução globalmente. Para mitigar esse problema, foi executado o algoritmo cem vezes com diferentes inicializações de centroides para escolher a solução com a menor soma das variações totais dentro dos grupos.

2.2.3 Biplot

O biplot é uma extensão gráfica da Análise de Componentes Principais (PCA). O biplot permite visualizar simultaneamente as projeções das observações e das variáveis originais no espaço definido pelas componentes principais.

Em um biplot, as observações são representadas como pontos, enquanto as variáveis originais são representadas como vetores. A direção e magnitude dos vetores indicam como as variáveis contribuem para as componentes principais. Assim, o biplot fornece uma visão integrada da estrutura dos dados, permitindo interpretar as relações entre observações e variáveis em termos das componentes principais.

A principal vantagem do biplot é sua capacidade de representar de forma concisa e informativa a estrutura multivariada dos dados. Ele facilita a interpretação das relações entre observações e variáveis, bem como entre as próprias variáveis. Além disso, o biplot pode ser usado para identificar agrupamentos ou padrões nos dados, tornando-se uma ferramenta valiosa para análises exploratórias (GOWER et al., 1996).

2.2.4 Métodos para determinar o número de grupos

Uma das principais dificuldades ao usar o algoritmo *K-means* é determinar o número apropriado de grupos k. Uma escolha inadequada de k pode levar a grupos mal definidos ou a uma representação imprecisa da estrutura subjacente dos dados.

Diversos métodos foram propostos para ajudar a determinar o número ótimo de grupo.

2.2.4.1 Método da Silhueta (*Silhouette Method*)

O Método da Silhueta avalia a qualidade dos grupos com base em quão semelhante cada ponto é em relação aos outros pontos em seu próprio grupo, em comparação com os pontos no grupo mais próximo. (ROUSSEEUW, 1987) Para cada ponto x:

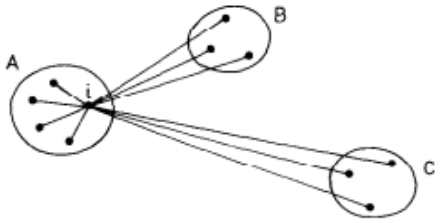
- $a(x)$: É a distância média de x para os outros pontos no mesmo grupo.
- $b(x)$: É a menor distância média de x para os pontos em um grupo diferente.

O coeficiente de silhueta para o ponto x é definido como:

$$s(x) = \frac{b(x) - a(x)}{\max\{a(x), b(x)\}}. \tag{2.3}$$

O coeficiente de silhueta varia entre -1 e 1. Um valor alto indica que o ponto está bem agrupado. Calculando o coeficiente de silhueta médio para todos os pontos, podemos usar como uma métrica para avaliar a qualidade dos grupos para diferentes valores de k onde escolheremos o k que possui a maior média dos coeficientes.

Figura 2 – Ilustração dos elementos envolvidos no cálculo de $s(x)$, onde o objeto x pertence ao grupo A.



Fonte: (ROUSSEEUW, 1987)

2.2.4.2 NbClust

O pacote NbClust (CHARRAD et al., 2014)., disponível para o *software* R, oferece uma solução abrangente para determinar o número ótimo de grupos em um conjunto de dados. Ele calcula 30 índices distintos incluindo o Método da Silhueta, permitindo ao usuário avaliar e comparar diversos esquemas de agrupamento de maneira simultânea.

A Tabela 1 elenca os trinta índices empregados pela função. Uma explanação aprofundada sobre cada índice, bem como as referências bibliográficas pertinentes, podem ser encontradas em (CHARRAD et al., 2014).

Tabela 1 – Índices implementados no pacote NbClust

ch (Calinski and Harabasz 1974)	duda (Duda and Hart 1973)
cindex (Hubert and Levin 1976)	gamma (Baker and Hubert 1975)
ccc (Sarle 1983)	ptbiserial (Milligan 1980, 1981)
db (Davies and Bouldin 1979)	frey (Frey and Van Groenewoud 1972)
tau (Rohlf 1974; Milligan 1981)	ratkowsky (Ratkowsky and Lance 1978)
marriot (Marriot 1971)	ball (Ball and Hall 1965)
tracew (Milligan and Cooper 1985)	friedman (Friedman and Rubin 1967)
rubin (Friedman and Rubin 1967)	kl (Krzanowski and Lai 1988)
gap (Tibshirani et al. 2001)	dindex (Lebart et al. 2000)
hubert (Hubert and Arabie 1985)	sdindex (Halkidi et al. 2000)
pseudot2 (Duda and Hart 1973)	beale (Beale 1969)
gplus (Rohlf 1974; Milligan 1981)	hartigan (Hartigan 1975)
scott (Scott and Symons 1971)	trcovw (Milligan and Cooper 1985)
mcclain (McClain and Rao 1975)	silhouette (Rousseeuw 1987)
dunn (Dunn 1974)	sdbw (Halkidi and Vazirgiannis 2001)

3 Resultados e discussão

Os primeiros dias foram crucial para a evolução da pandemia de COVID-19 nos municípios brasileiros. Esse começo é fundamental para entender a propagação do vírus, pois representam o momento em que as comunidades começam a enfrentar o desafio da doença.

Para a realização desse trabalho, todos os dados utilizados referenciam individualmente cada município. Para ajustar os modelos de regressão foram utilizadas as seguintes informações:

- **Casos acumulados** (y_t): A variável dependente é o total de casos confirmados de COVID-19 acumulados ao longo de trinta dias a partir do primeiro caso do município. Disponíveis no portal do Ministério da Saúde. (DATASUS, 2023)
- **Número de dias**: A variável independente é o número de dias a partir do primeiro caso no município, que vai de um até trinta.

Para a comparação dos agrupamentos de municípios foram utilizadas:

- **Estabelecimentos de Saúde**: Quantidade de Estabelecimentos de Saúde disponível pelo Ministério da Saúde - Cadastro Nacional dos Estabelecimentos de Saúde do Brasil - CNES. (DATASUS, 2023)
- **Profissionais de Saúde**: Quantidade de Profissionais de Saúde disponível pelo Ministério da Saúde - Cadastro Nacional dos Estabelecimentos de Saúde do Brasil - CNES. (DATASUS, 2023)
- **Densidade Populacional por área urbanizada**: É uma métrica que representa o número de pessoas por unidade de área, geralmente expressa em habitantes por quilômetro quadrado (hab/km^2). Nesse caso utilizamos a área urbanizada do município. Ela é uma ferramenta essencial para entender a distribuição da população em uma determinada região ou país. Essa informação pode ser calculada utilizando a área urbanizada disponível no IBGE e a população disponível no IBGE pelo censo de 2022. (IBGE, 2023)
- **Região**: Refere-se a uma das cinco principais divisões territoriais do Brasil, sendo elas: Norte, Nordeste, Centro-Oeste, Sudeste e Sul. Cada região reúne estados que compartilham semelhantes.

Toda parte computacional foi produzida utilizando o software R (R Core Team, 2023).

3.1 Análise Exploratória de Dados

A pandemia da COVID-19 impôs desafios significativos ao Brasil, e a diversidade demográfica do país resultou em uma variação notável na taxa de crescimento da doença entre os municípios.

Os municípios que mais se destacaram, tanto em termos de alta quanto de baixa incidência da doença nos primeiros 30 dias por mil habitantes, são detalhados nas Tabelas 2 e 3. A amplitude das diferenças entre esses municípios é notável, refletindo a heterogeneidade da disseminação do vírus no território nacional.

Tabela 2 – Municípios com a maior incidência de COVID-19 por mil habitantes nos primeiros trinta dias

Município	Estado	Casos per 1000 hab
Porto Alegre do Piauí	Piauí	52,946
Bandeira	Minas Gerais	42,987
Cutias	Amapá	36,235
Igarapé Grande	Maranhão	35,707
Joca Marques	Piauí	35,064

Tabela 3 – Municípios com a menor incidência de COVID-19 por mil habitantes nos primeiros trinta dias

Município	Estado	Casos per 1000 hab
Barreiras	Bahia	0,006
Catalão	Goiais	0,009
Esmeraldas	Minas Gerais	0,009
Ponta Porã	Mato Grosso do Sul	0,011
Juazeiro do Norte	Maranhão	0,011

Ao avaliar a incidência da doença por região, a Tabela 4 revela que a região Norte foi a mais afetada, enquanto a região Sudeste apresentou os menores índices. Estes dados são cruciais para orientar intervenções e otimizar a alocação de recursos em saúde pública.

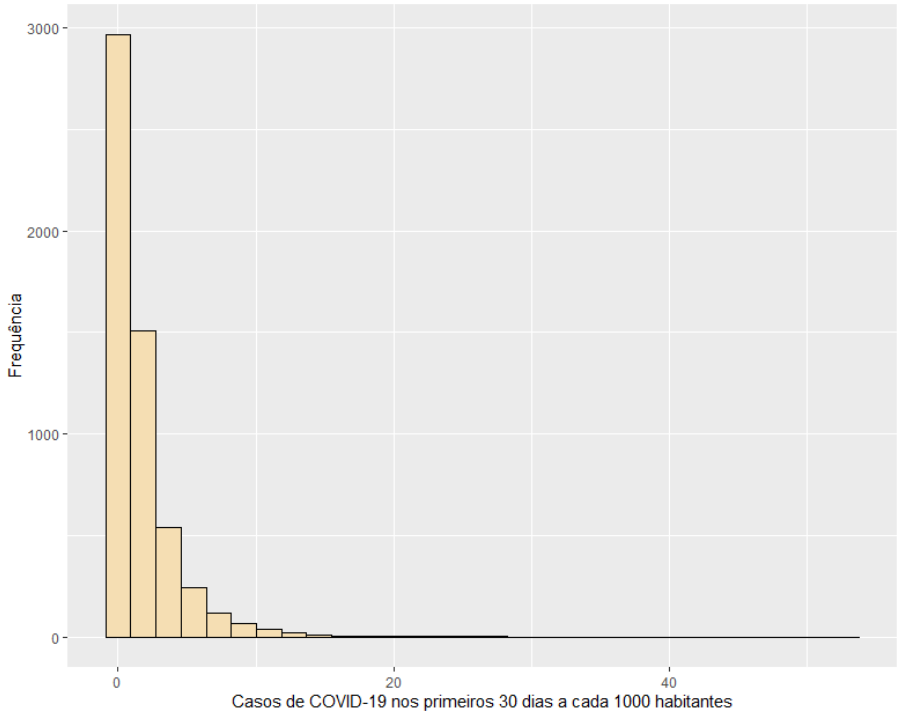
Tabela 4 – Média e mediana de casos de COVID-19 por mil habitantes nos primeiros trinta dias

Região	Média	Mediana
Norte	3,117	1,697
Centro-Oeste	2,193	0,895
Nordeste	1,930	0,922
Sul	1,818	0,831
Sudeste	1,243	0,555

A distribuição dos casos nos municípios é visualmente representada no histograma da Figura 3. A assimetria negativa predominante sugere uma concentração significativa de

municípios com incidências inferiores a 10 casos por mil habitantes no período inicial de trinta dias.

Figura 3 – Distribuição de casos confirmados de COVID-19 nos primeiros trinta dias por mil habitantes



3.2 Análise de Regressão

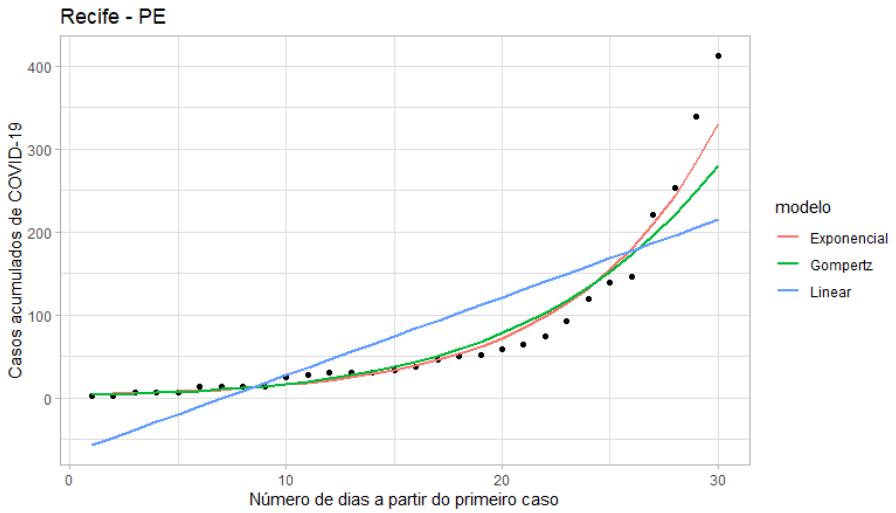
A dinâmica da COVID-19 nos municípios brasileiros foi analisada por meio de diferentes modelos de regressão. Para cada um dos 5.570 municípios brasileiros, foram estimados os modelos de Gompertz, regressão linear e regressão exponencial. Esses modelos foram escolhidos por sua capacidade de capturar diferentes padrões de crescimento e disseminação da doença.

A cidade de Recife serve como um exemplo ilustrativo dessa análise. A Figura 4 mostra o ajuste dos três modelos mencionados para este município. A Tabela 5 destaca que, para Recife, o modelo de Gompertz foi o mais adequado, explicando de forma mais precisa a variabilidade dos casos acumulados nos primeiros trinta dias.

Tabela 5 – Detalhes dos modelos para Recife: coeficientes de determinação

Modelo	R^2
Regressão Linear	0,650
Regressão Exponencial	0,954
Modelo de Gompertz	0,960

Figura 4 – Recife: Ajuste de modelos ao total de casos confirmados de COVID-19 nos primeiros 30 dias a cada mil habitantes



A validade e robustez de um modelo de regressão dependem, em grande parte, da satisfação das condições descritas em 2.1.

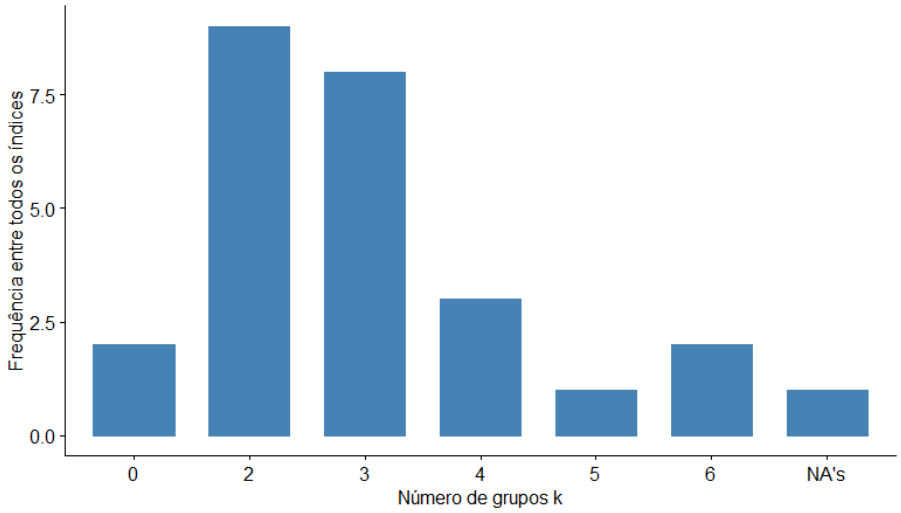
Ao modelar o total acumulado de casos confirmados de COVID-19, é esperada a presença de autocorrelação nos resíduos, devido à natureza cumulativa dos dados. Esta autocorrelação é inerente, pois os totais acumulados são influenciados pelos valores dos dias anteriores.

3.3 Segmentação dos Municípios

A segmentação dos municípios brasileiros, com base nos padrões de crescimento de casos acumulados de COVID-19 nos primeiros trinta dias, é fundamental para entender a dinâmica de propagação da doença e é o principal objetivo desse estudo. Utilizamos o coeficiente de determinação dos modelos para cada município como critério de agrupamento. Ajustes foram realizados nos coeficientes, atribuindo valor zero em situações onde a evolução dos casos acumulados eram constante ou quando os modelos não eram estatisticamente significativos, ou seja quando o p-valor referente ao teste F era maior que 0.05.

A determinação do número ideal de grupos foi realizada através da função NbClust (CHARRAD et al., 2014), a Figura 5 demonstra o resultado de 30 diferentes índices utilizados para selecionar o número ideal de grupos em um particionamento. Dentre eles, estão incluídos o método da silhueta (KASSAMBARA, 2017) e o método do cotovelo (ROUSSEEUW, 1987). Como ilustrado, a maioria dos índices sugeriu a formação de dois ou três grupos, optando-se, neste trabalho, por três por oferecer uma segmentação mais detalhada, permitindo identificar nuances e padrões intermediários que poderiam ser perdidos em uma divisão em dois grupos.

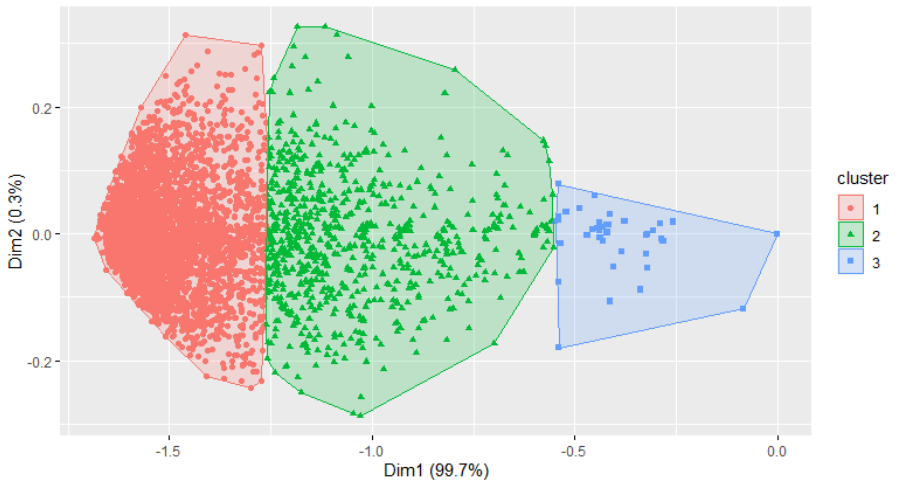
Figura 5 – Avaliação de métodos para identificar o número ótimo de agrupamentos



O método *K-means* (MACQUEEN, 1967) foi empregado para a segmentação, com um critério de até cem iterações e utilizando da distancia euclidiana. Para contornar o desafio dos mínimos locais, o algoritmo foi executado cem vezes com diferentes inicializações de centroides, selecionando-se o modelo com a menor variação total.

A Figura 6 apresenta um biplot, que combina as projeções das observações e das variáveis originais no espaço definido pelas duas primeiras componentes principais (GOWER et al., 1996). O biplot permite uma interpretação conjunta da variabilidade dos dados e da contribuição de cada variável para essa variabilidade. É evidente a distinção clara entre os grupos, sem sobreposições aparentes.

Figura 6 – Visualização da segmentação dos municípios brasileiros utilizando do coeficiente de determinação



A Tabela 6 apresenta as estatísticas descritivas, média e mediana, dos coeficientes de determinação para cada grupo. A segmentação claramente distingue municípios com alta, média e baixa capacidade dos modelos em explicar a variabilidade dos casos.

Tabela 6 – Distribuição dos coeficientes de determinação entre os grupos

Grupo	Medida	R^2 - linear	R^2 - exponencial	R^2 - gompertz
Grupo 1	Média	0,873	0,877	0,892
	Mediana	0,857	0,864	0,879
Grupo 2	Média	0,601	0,606	0,624
	Mediana	0,589	0,587	0,591
Grupo 3	Média	0,000	0,000	0,000
	Mediana	0,044	0,044	0,044

Conforme ilustrado na Tabela 7, a predominância dos municípios se encontra no primeiro grupo onde os modelos explicaram bem a variabilidade dos dados.

Tabela 7 – Contagem de municípios em cada grupo de segmentação

Grupo	Frequência
Grupo 1	3.675
Grupo 2	1.166
Grupo 3	729

A distribuição regional dos municípios em cada grupo é detalhada na Tabela 8. As regiões Norte e Nordeste têm uma presença marcante no primeiro grupo, com 80% e 72% de seus municípios, respectivamente. O que significa que grande parcela dos seus municípios tiveram a variabilidade dos seus dados bem explicado pelos modelos.

Tabela 8 – Distribuição percentual e absoluta de municípios por região e agrupamento

Região	Grupo 1	Grupo 2	Grupo 3
Norte	359 (80%)	27 (06%)	64 (14%)
Nordeste	1.295 (72%)	169 (10%)	330 (18%)
Centro-Oeste	301 (64%)	79 (17%)	87 (19%)
Sudeste	1.043 (63%)	253 (15%)	372 (22%)
Sul	677 (57%)	201 (17%)	313 (26%)

A Tabela 9 destaca as métricas demográficas e de saúde para cada grupo. Embora existam diferenças sutis entre os grupos, é notável que o primeiro grupo, que teve sua variabilidade mais bem capturada pelos modelos, possui a menor proporção de profissionais de saúde por mil habitantes. Além disso, uma densidade populacional urbana mais elevada é observada em municípios onde os modelos puderam explicar a variabilidade de forma mais eficaz.

Tabela 9 – Comparação de indicadores demográficos entre os grupos

Grupo	Medida	Densidade Urb.	Estab. de saúde (mil hab.)	Prof. de saúde (mil hab.)
Grupo 1	Média	4.311	01,48	12,40
	Mediana	3.775	01,27	11,70
Grupo 2	Média	3.982	01,58	12,70
	Mediana	3.527	01,45	11,90
Grupo 3	Média	3.843	01,58	13,00
	Mediana	3.503	01,43	12,10

Com o intuito de discernir quais municípios sofreram maior impacto nos dias iniciais, realizou-se outra segmentação empregando o algoritmo *K-means*. O critério utilizado para essa segmentação foram os coeficientes de regressão β_1, que simbolizam a taxa de crescimento da curva epidemiológica.

O algoritmo *K-means* com a distância euclidiana é sensível à escala das variáveis, portanto, como os coeficientes de uma Regressão Linear, Exponencial e de Gompertz possuem escalas distintas é crucial realizar a padronização dos dados antes de aplicar o algoritmo. A padronização de uma variável envolve a reescala dos dados de forma que a média da variável seja 0 e o desvio padrão seja 1. Isso é feito subtraindo a média de cada observação e, em seguida, dividindo pelo desvio padrão. A padronização de uma variável x pode ser calculada como:

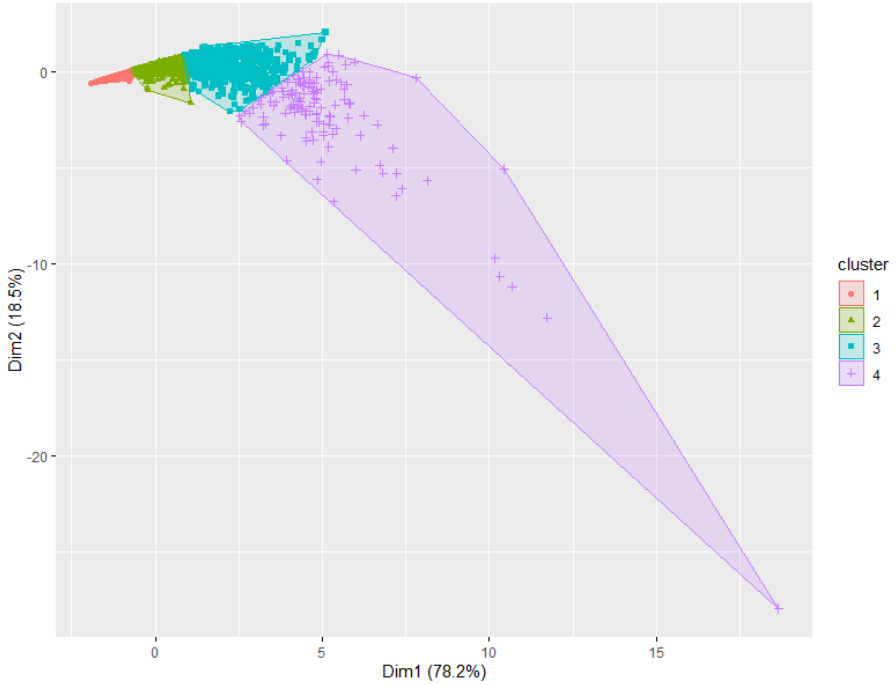
$$z = \frac{x - \bar{x}}{s}.$$

onde z é a variável padronizada, x é o valor original da variável, $\bar{x}$ é a média dos valores da variável, e s é o desvio padrão da variável. A padronização é essencial para garantir que todas as variáveis tenham o mesmo peso no modelo *K-means*, evitando que variáveis com escalas maiores dominem o processo de agrupamento.

Optou-se pela formação de quatro grupos. Esta escolha possibilitou a delimitação de um grupo constituído por municípios que exibiram uma taxa de crescimento elevada. Tal isolamento permite uma análise apurada desses municípios, identificando-os como potenciais áreas de atenção para o enfrentamento de futuras pandemias.

A Figura 7 exibe o biplot resultante dessa segmentação. A distinção entre os grupos é claramente evidente, sem sobreposições visíveis.

Figura 7 – Visualização da segmentação dos municípios brasileiros utilizando do coeficiente de regressão



A Tabela 10 mostra as estatísticas descritivas, média e mediana, dos coeficientes de determinação para cada grupo. A segmentação efetua uma delimitação precisa com um grupo contendo os municípios com as maiores taxas de crescimento que foi o seu principal objetivo.

Tabela 10 – Distribuição dos coeficientes de regressão β_1 entre os grupos

Grupo	Medida	β_1 - linear	β_1 - exponencial	β_1 - gompertz
Grupo 1	Média	0,044	0,024	-0,004
	Mediana	0,059	0,022	-0,004
Grupo 2	Média	0,338	0,076	-0,013
	Mediana	0,428	0,077	-0,013
Grupo 3	Média	1,357	0,126	-0,025
	Mediana	1,697	0,129	-0,026
Grupo 4	Média	7,289	0,173	-0,036
	Mediana	9,092	0,173	-0,037

Como demonstrado na Tabela 11, temos 128 municípios no grupo 4, que corresponde aos grupos delimitados que apresentaram as maiores taxas de crescimento nos primeiros trinta dias após o registro do primeiro caso de COVID-19 no município.

Tabela 11 – Contagem de municípios em cada grupo de segmentação

Grupo	Frequência
Grupo 1	1.955
Grupo 2	2.315
Grupo 3	1.172
Grupo 4	128

A Tabela 12 detalha a distribuição regional dos municípios em cada grupo. O quarto grupo é predominantemente representado pelas regiões Norte e Nordeste, que, juntas, compõem 78% dos municípios deste grupo. Nota-se que 10% dos municípios da região Norte integram este conjunto de municípios caracterizado por altas taxas de crescimento.

Tabela 12 – Distribuição percentual e absoluta de municípios por região e agrupamento

Região	Grupo 1	Grupo 2	Grupo 3	Grupo 4
Norte	82 (18%)	166 (37%)	159 (35%)	43 (10%)
Nordeste	481 (27%)	734 (41%)	522 (29%)	57 (3%)
Centro-Oeste	170 (36%)	195 (42%)	97 (21%)	5 (1%)
Sudeste	670 (40%)	728 (44%)	252 (15%)	18 (1%)
Sul	552 (47%)	492 (41%)	142 (12%)	5 (0%)

A Tabela 13 destaca as métricas demográficas e de saúde associadas a cada grupo. Observa-se que o quarto grupo, que apresentou as taxas de crescimento mais elevadas, possui a menor proporção de estabelecimentos de saúde por mil habitantes e maior densidade populacional urbana.

Tabela 13 – Comparação de indicadores demográficos entre os grupos

Grupo	Medida	Densidade Urb.	Estab. de saúde (mil hab.)	Prof. de saúde (mil hab.)
Grupo 1	Média	3.903	01,60	12,70
	Mediana	3.508	01,46	12,00
Grupo 2	Média	4.113	01,58	12,60
	Mediana	3.668	01,37	11,80
Grupo 3	Média	4.623	01,30	12,20
	Mediana	4.006	01,12	11,20
Grupo 4	Média	5.606	01,03	13,00
	Mediana	5.016	0,79	12,50

Os 128 municípios segmentados no grupo 4 são descritos na tabela 14 identificando-os como potenciais áreas de atenção em futuras pandemias. Observamos que alguns desses municípios são capitais importantes do país como Recife, Salvador e São Paulo.

Tabela 14 – Municípios segmentados com altas taxas de crescimento

Acaraú - CE	Alvarães - AM	Amarante do Maranhão - MA
Amaturá - AM	Anapurus - MA	Araioses - MA
Autazes - AM	Bacuri - MA	Bandeira - MG
Barcelos - AM	Barra do Corda - MA	Barreirinha - AM
Barreiros - PE	Belo Horizonte - MG	Belém - PA
Benjamin Constant - AM	Boa Vista - RR	Bom Jesus das Selvas - MA
Borba - AM	Brasília - DF	Breves - PA
Buriticupu - MA	Caaporã - PB	Camaragibe - PE
Carauari - AM	Careiro - AM	Caucaia - CE
Coari - AM	Codó - MA	Coelho Neto - MA
Concórdia - SC	Coroatá - MA	Curitiba - PR
Darcinópolis - TO	Diadema - SP	Dom Eliseu - PA
Duque de Caxias - RJ	Esperantinópolis - MA	Fonte Boa - AM
Fortaleza - CE	Fátima do Sul - MS	Governador Archer - MA
Grajaú - MA	Guajará-Mirim - RO	Guarabira - PB
Guarulhos - SP	Guia Lopes da Laguna - MS	Igarapé Grande - MA
Igarapé-Miri - PA	Iguaí - BA	Ipixuna - AM
Itambé - PR	Itapiranga - AM	Itinga do Maranhão - MA
Japurá - AM	Joca Marques - PI	Juruaia - MG
Juruena - MT	Lagoa Alegre - PI	Laranjal do Jari - AP
Lima Campos - MA	Macapá - AP	Manacapuru - AM
Manaus - AM	Maracaçumé - MA	Maranhãozinho - MA
Maués - AM	Mesquita - RJ	Moju - PA
Morada Nova - CE	Nossa Senhora do Socorro - SE	Nova Iguaçu - RJ
Olinda - PE	Osasco - SP	Ourilândia do Norte - PA
Ouro Preto - MG	Paragominas - PA	Paulista - PE
Paulo Ramos - MA	Pedreiras - MA	Pinheiro - MA
Pio XII - MA	Pirapemas - MA	Portel - PA
Porto Alegre - RS	Porto Alegre do Piauí - PI	Poção de Pedras - MA
Poções - BA	Presidente Dutra - MA	Recife - PE
Rio Brilhante - MS	Rio Preto da Eva - AM	Rio de Janeiro - RJ
Salinas - MG	Salvador - BA	Santa Helena - MA
Santa Quitéria do Maranhão - MA	Santana - AP	Santo André - SP
Santo Antônio dos Lopes - MA	Santos - SP	Sena Madureira - AC
Senador José Porfírio - PA	Serra - ES	São Bento - MA
São Bernardo - MA	São Bernardo do Campo - SP	São Domingos do Mar. - MA
São Gonçalo do Amarante - CE	São José de Ribamar - MA	São Luís - MA
São Mateus do Maranhão - MA	São Miguel do Guaporé - RO	São Paulo - SP
São Pedro da Água Branca - MA	Tabatinga - AM	Tapauá - AM
Tarauacá - AC	Tefé - AM	Teotônio Vilela - AL
Tutóia - MA	Uarini - AM	Ulianópolis - PA
Vila Velha - ES	Vitória do Xingu - PA	Xambioá - TO
Xanxerê - SC	Águas Belas - PE	

4 Conclusões e Trabalhos Futuros

O estudo realizado buscou compreender a dinâmica da disseminação inicial da COVID-19 nos primeiros trinta dias da pandemia em municípios brasileiros, relacionando a evolução dos casos acumulados com os contextos demográficos. Através da aplicação de modelos de regressão e do algoritmo de agrupamento *K-means*, foi possível categorizar municípios com padrões de disseminação semelhantes e com essa categorização possibilita em trabalhos futuros análises mais detalhadas. No entanto, é importante destacar que os modelos de regressão utilizados não atenderam completamente às suposições, o que pode ter impactado a precisão e a interpretabilidade dos resultados obtidos.

Foram identificados 128 municípios que, no momento inicial da disseminação da doença, foram segmentados como áreas de altas taxas de crescimento. Essa identificação possibilita uma análise mais detalhada desses municípios, destacando-os como potenciais áreas de atenção em face de futuras pandemias.

Esse estudo é desafiador por se utilizar de casos confirmados visto que testes em massa para a população geral só se tornaram comuns após um longo período da pandemia.

Adicionalmente, a adoção de casos confirmados desde o início, em detrimento dos casos acumulados, pode oferecer uma melhora. O modelo gompertz, apesar de sua relevância, apresenta limitações, especialmente ao considerar a necessidade de um parâmetro para determinar sua assíntota. Tal característica pode ser problemática em cenários iniciais de pandemias, onde a magnitude da propagação é incerta. Portanto, é crucial considerar outros modelos alternativos e mais complexos para melhorar a precisão e a aplicabilidade das análises em cenários de pandemia.

Referências

CHARRAD, M. et al. NbClust: An R package for determining the relevant number of clusters in a data set. *Journal of Statistical Software*, v. 61, n. 6, p. 1–36, 2014. Disponível em: <https://www.jstatsoft.org/v61/i06/>. Citado 2 vezes nas páginas 11 e 15.

COHEN, J. The race for coronavirus vaccines: a graphical guide. *Nature*, Nature Publishing Group, v. 580, n. 7805, p. 576–577, 2020. Citado na página 1.

CRIBARI-NETO, F. A beta regression analysis of covid-19 mortality in brazil. *Infectious Disease Modelling*, v. 8, n. 2, p. 309–317, 2023. ISSN 2468-0427. Disponível em: <https://www.sciencedirect.com/science/article/pii/S246804272300012X>. Citado na página 1.

DATASUS. *Dados sobre COVID-19*. 2023. Disponível em: <http://tabnet.datasus.gov.br/>. Citado na página 12.

DRAPER et al. *Applied regression analysis*. 3. ed. New York: Wiley, 1998. ISBN 0-471-17082-8. Citado na página 4.

GOMPERTZ, B. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the royal society of London*, v. 115, p. 513–583, 1825. Citado na página 5.

GOWER, J. et al. *Biplots*. [S.l.]: Chapman & Hall, 1996. Citado 2 vezes nas páginas 10 e 16.

IBGE. *Censo 2022*. 2023. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/populacao/22827-censo-demografico-2022.html?edicao=37225&t=resultados>. Citado na página 12.

KASSAMBARA, A. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*. [S.l.]: STHDA, 2017. v. 1. Citado 3 vezes nas páginas 7, 8 e 15.

MACQUEEN, J. Some methods for classification and analysis of multivariate observations. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. [S.l.: s.n.], 1967. v. 1, n. 14, p. 281–297. Citado 2 vezes nas páginas 8 e 16.

MONTGOMERY, D. C. et al. *Introduction to Linear Regression Analysis*. [S.l.]: Wiley, 2013. v. 5. Citado 3 vezes nas páginas 3, 6 e 7.

NICOLA, M. et al. The socio-economic implications of the coronavirus pandemic (covid-19): A review. *International Journal of Surgery*, Elsevier, v. 78, p. 185–193, 2020. Citado na página 1.

PEREIRA, J. C. R. *Análise de Dados Qualitativos: Estratégias Metodológicas para as Ciências da Saúde, Humanas e Sociais. 3 ed.* [S.l.]: Editora da Universidade de São Paulo., 2004. Citado na página 8.

R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria, 2023. Disponível em: <https://www.R-project.org/>. Citado na página 12.

ROSSI et al. Application of the gompertz equation for the study of xylem cell development. *Dendrochronologia*, v. 21, p. 33–39, 12 2003. Citado na página 5.

ROUSSEEUW, P. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53–65, 1987. Citado 3 vezes nas páginas 10, 11 e 15.

ZHOU, P. et al. A pneumonia outbreak associated with a new coronavirus of probable bat origin. *Nature*, Nature Publishing Group, v. 579, n. 7798, p. 270–273, 2020. Citado na página 1.