



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

Hygor Fernando Alves Menezes

**Análise de regressão do padrão de crescimento de COVID-19 nos
municípios do Brasil**

**Recife
2023**

Hygor Fernando Alves Menezes

Análise de regressão do padrão de crescimento de COVID-19 nos municípios do Brasil

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. Cristiano Ferraz

Coorientador: Prof. 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

Menezes, Hygor Fernando Alves.

Análise de regressão do padrão de crescimento de COVID-19 nos
municípios do Brasil / Hygor Fernando Alves Menezes. - Recife, 2023.
28 p. : il., tab.

Orientador(a): Cristiano Ferraz

Orientador(a): André Leite Wanderley

Coorientador(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 regressão. 2. COVID-19. 3. Padrão de crescimento. 4.
Municípios brasileiros. 5. Análise de resíduos. I. Ferraz, Cristiano. (Orientação).
II. Wanderley, André Leite. (Orientação). III. Wanderley, André Leite.
(Coorientação). IV. Título.

310 CDD (22.ed.)

Agradecimentos

Em primeiro lugar, queria expressar minha gratidão a toda a minha família, meus pais, minha madrasta, meu padrasto, minhas avós, minha irmã, meu tio e meus zêlbrions. É graças a eles que pude enfrentar e até hoje enfrento a vida com tanta leveza e alegria. Sempre pude contar com seu apoio incondicional em todas as minhas decisões, e eles nunca duvidaram da minha capacidade.

Agradeço também aos meus amigos, tanto aos de longa data quanto aos mais recentes. Aos que já não somos tão próximos, pois tomamos rumos diferentes, tenham certeza que vocês deixaram uma marca em minha trajetória e em quem me tornei, e aos que permanecem ao meu lado, agradeço pelo apoio, pela lealdade e por sempre estarem presentes. Minha felicidade também é reflexo da presença de cada um de vocês em minha vida.

É essencial mencionar meus orientadores, André e Cristiano, bem como todos os professores que tive ao longo de minha jornada educacional, em especial durante a graduação. E ao falar de jornada, não posso deixar de lado os mentores que tive em minha trajetória profissional. Recordo-me de cada um de vocês com imenso carinho e sou grato por todas as lições compartilhadas.

A todos vocês, minha sincera gratidão. Esta conquista também é reflexo do apoio, ensinamentos e amor que recebi de cada um.

Resumo

Este trabalho tem como objetivo analisar os primeiros 30 dias da pandemia de COVID-19 em cada município brasileiro. Para tal análise, consideramos o período inicial da pandemia em cada município, a partir do momento do registro do primeiro caso de COVID-19. O primeiro município a registrar um caso foi em 25 de fevereiro de 2020, enquanto o último município atingiu a marca de 30 dias de casos em 12 de janeiro de 2021. Utilizando dados do Ministério da Saúde e do IBGE, a análise se concentra na avaliação do padrão de crescimento dos casos de COVID-19 e na determinação de qual modelo — linear, exponencial ou Gompertz — melhor descreve esse padrão em cada um dos municípios brasileiros. Após a determinação do modelo mais adequado para cada município, foi realizada uma análise adicional, utilizando apenas da regressão linear para entender a relação entre o padrão de crescimento dos casos e variáveis explicativas sociais e demográficas. Além disso, o estudo compara os resultados desses modelos e realiza uma análise de resíduos para avaliar a adequação dos modelos. Este trabalho serve como base para futuras pesquisas na área, visando auxiliar na elaboração de estratégias eficazes para enfrentar futuros surtos e otimizar as medidas de controle e prevenção.

Palavras-chave: COVID-19; Municípios brasileiros; Padrão de crescimento; Modelos de regressão; Análise de resíduos.

Abstract

This study aims to analyze the initial 30 days of the COVID-19 pandemic in each Brazilian municipality. The analysis spans from the first recorded case on February 25, 2020, to January 12, 2021, when the last municipality reached its 30-day case mark. Using data from the Ministry of Health and IBGE, the research focuses on evaluating the growth pattern of COVID-19 cases and determining which model — linear, exponential, or Gompertz — best describes this pattern for each municipality. Upon determining the most suitable model for each municipality, an additional analysis was conducted, using linear regression to understand the relationship between the growth pattern of cases and social and demographic explanatory variables. Furthermore, the research compares the outcomes of these models and conducts a residual analysis to assess the models' adequacy. This work lays the foundation for future research in the field, aiming to assist in devising effective strategies to tackle future outbreaks and optimize control and prevention measures.

Keywords: COVID-19; Brazilian municipalities; Growth pattern; Regression models; Residual analysis.

Lista de ilustrações

Figura 1 – Modelo linear Belo Horizonte - MG	6
Figura 2 – Modelo exponencial Vila Velha - ES	8
Figura 3 – Modelo de Gompertz Belém - PA	10
Figura 4 – Os modelos em Recife – PE	11
Figura 5 – Histograma dos casos acumulados	14
Figura 6 – Modelos por município	14
Figura 7 – Municípios com maiores taxas de crescimento linear	16
Figura 8 – Municípios com maiores taxas de crescimento pelo modelo de Gompertz	16
Figura 9 – Municípios com maiores taxas de crescimento exponencial	17
Figura 10 – Comparação do padrão de crescimento das regiões do Brasil	18
Figura 11 – Melhor modelo por município	19
Figura 12 – Comparação do padrão de crescimento das mesorregiões de Pernambuco	20
Figura 13 – Densidade dos Resíduos por modelo	23
Figura 14 – QQ-Plot dos Resíduos por modelo	24
Figura 15 – Resíduo Studentizado vs Valores Preditos	25

Lista de tabelas

Tabela 1	–	Valores de R^2 ajustado para os modelos em Recife	11
Tabela 2	–	Estatísticas dos casos acumulados	13
Tabela 3	–	Número de municípios por melhor modelo ajustado	15
Tabela 4	–	Capitais e seus modelos	17
Tabela 5	–	Resumo da População em 2022 por Modelo que Melhor explica o município	18
Tabela 6	–	Porcentagem de Modelos por Mesorregião	19
Tabela 7	–	R^2 Ajustado no modelo global	21
Tabela 8	–	R^2 Ajustado no modelo por município	21
Tabela 9	–	Resumo dos Resultados das Regressões	22
Tabela 10	–	P-valores do teste de Shapiro-Wilk para os modelos	24
Tabela 11	–	Valores do teste de Durbin-Watson para os modelos	25

Sumário

1	INTRODUÇÃO	1
2	REGRESSÃO	3
2.1	Estimação dos parâmetros	4
2.2	Regressão linear	4
2.3	Regressão exponencial	6
2.4	Modelo de Gompertz	8
2.5	Comparação entre os modelos	10
3	RESULTADOS E DISCUSSÕES	12
3.1	Dados de COVID-19	12
3.2	Análise exploratória de dados	13
3.3	Modelos de Regressão	20
3.4	Análise de Variáveis	22
3.5	Análise de Resíduos	23
4	CONCLUSÃO	26
	REFERÊNCIAS	28

1 Introdução

A COVID-19, apesar de ter suas origens em Wuhan, China, encontrou terreno fértil no Brasil devido a uma série de características inerentes ao país. Com uma vasta extensão territorial e uma diversidade demográfica e socioeconômica imensa, o Brasil enfrentou desafios particulares na contenção e gestão da pandemia. As respostas ao vírus variaram consideravelmente entre os municípios e regiões, refletindo as disparidades presentes em nosso país. As metrópoles superpopuladas apresentaram desafios significativos em relação à transmissão, enquanto áreas mais remotas e menos desenvolvidas enfrentaram dificuldades com a estrutura de saúde e logística.

Este cenário complexo exige uma análise detalhada para entender as particularidades da pandemia no contexto brasileiro. Um estudo realizado por (SILVA et al., 2020) destaca que as medidas de distanciamento social foram amplamente implementadas no Brasil, principalmente na segunda metade de março de 2020, antes ou na fase inicial da curva de crescimento de casos e mortes de COVID-19 na grande maioria dos estados.

Em cenários de crise de saúde pública, como a pandemia da COVID-19, a análise de dados não é apenas uma ferramenta de suporte, mas sim uma necessidade crítica. Ela permite as autoridades acessarem informações atualizadas, identificar tendências emergentes e fazer ajustes efetivos na resposta à doença. Por exemplo, um estudo realizado por (LIU et al., 2021) apresentou um novo modelo de previsão de pandemia SEAIRD, que utiliza redes neurais para aprender com os dados reais da epidemia e obter parâmetros ótimos.

A análise de dados epidemiológicos e demográficos identifica padrões de comportamento e ajuda a identificar populações ou áreas específicas que podem estar em maior risco. Isso pode significar a diferença entre uma resposta eficaz e uma crise agravada.

O Brasil, com sua extensão territorial e diversidade socioeconômica, é uma nação complexa repleta de desafios e oportunidades na gestão da pandemia. Uma abordagem baseada em dados torna-se ainda mais importante quando consideramos as vastas diferenças entre municípios e regiões.

Além disso, a natureza dinâmica da COVID-19, com suas variantes emergentes e mudanças no padrão de transmissão, torna a análise contínua uma parte essencial da resposta à pandemia. A compreensão das nuances locais e regionais da disseminação do vírus pode oferecer percepções sobre áreas de vulnerabilidade e resiliência. Ao focar nos municípios brasileiros, este estudo também destaca a importância de abordagens descentralizadas na gestão de crises de saúde pública. Enquanto políticas nacionais e diretrizes gerais são cruciais, a capacidade de adaptar estratégias com base em dados locais

pode ser a chave para uma intervenção mais eficaz. Através deste trabalho, buscamos contribuir fornecendo uma análise que pode ser usada por formuladores de políticas, profissionais de saúde e pesquisadores para ajudar a moldar uma resposta à COVID-19 e a outras ameaças de saúde pública.

Neste trabalho, utilizando dados do Ministério da Saúde (SAUDE, 2022) e do Censo do IBGE de 2022 ((IBGE), 2022), buscamos entender o padrão de crescimento dos casos de COVID-19 nos municípios brasileiros. Procuramos identificar qual modelo de crescimento — linear, exponencial ou Gompertz — melhor representa a evolução da doença em diferentes contextos municipais. E após isso entender quais variáveis sociais e demográficas melhor se relacionam com esse padrão. Esta análise visa oferecer uma perspectiva clara sobre a trajetória da pandemia em território brasileiro e estabelecer um alicerce para futuros estudos na área. Almejamos que os resultados deste estudo contribuam para a formulação de estratégias mais eficazes no combate a futuros surtos e na otimização das medidas de controle e prevenção.

Foi considerado os 30 primeiros dias da pandemia em cada município a partir do registro do primeiro caso de COVID-19, com o primeiro município registrando um caso de COVID-19 em 25 de fevereiro de 2020 e o último município alcançou a marca de 30 dias de casos de COVID-19 no dia 12 de janeiro de 2021. Todos os municípios brasileiros foram incluídos no estudo. A variável resposta escolhida foi os casos acumulados de COVID-19, representando o total de casos confirmados até a data de referência em cada município. Como variável explicativa, utilizamos o número de dias, que indica o número de dias desde o início da observação, ou seja, desde o registro do primeiro caso em cada município.

Após a determinação do melhor modelo para cada município, foi feita uma regressão linear tendo a taxa de crescimento encontrada no modelo como variável resposta, e com as variáveis explicativas sendo densidade populacional do município por área urbanizada, leitos de UTI por habitante e enfermeiros por habitante. Essa análise visa um melhor entendimento de fatores que possam explicar ou não o padrão de crescimento de cada município.

2 Regressão

A análise de regressão é um método estatístico que busca estudar a relação entre uma variável dependente e uma ou mais variáveis independentes. No contexto de estudos epidemiológicos, é comum adotar modelos de regressão não-lineares, pois dependendo das variáveis a relação entre elas pode não ser linear, especialmente em situações como a propagação de doenças infecciosas. Em situações epidemiológicas, como a propagação da COVID-19, é comum observar padrões de crescimento que não são lineares, especialmente nas fases iniciais de um surto. Portanto, a escolha de modelos não-lineares pode fornecer percepções relevantes sobre a dinâmica da doença.

Neste estudo, a unidade de análise escolhida é o município. Cada município brasileiro é tratado como uma entidade única, com suas características socioeconômicas e demográficas. A variável resposta, y_t , é um valor acumulado, refletindo o total de casos até o momento, que aumenta com o passar dos dias. Vale ressaltar que, para muitos municípios que tiveram seus primeiros 30 dias no começo de 2020, a coleta do número de casos passa muito pelo número de óbitos e pelos testes feitos em profissionais da saúde, já que os testes em massa da população geral só vieram a ser uma realidade em setembro do mesmo ano.

Três modelos de regressão foram empregados: Regressão Linear, Regressão Exponencial e Gompertz. O objetivo desta análise é determinar qual desses modelos melhor se adequa ao padrão de crescimento dos casos de COVID-19 nos municípios brasileiros. Levando em consideração os primeiros 30 dias do início da pandemia em cada município, a data de registro do primeiro caso de COVID-19 variou, sendo São Paulo o primeiro a registrar um caso em 25 de fevereiro de 2020, e o último município a atingir 30 dias de casos de COVID-19 foi Cedro do Abaeté, em Minas Gerais, em 12 de janeiro de 2021.

Outro modelo que embora não seja usado neste trabalho, porém tem que ser citado, devido a sua importância na tomada de decisões na pandemia de COVID-19 e foi fundamental na análise da propagação da COVID-19, especialmente no contexto brasileiro, é o modelo SIR (Susceptível - Infectado - Removido). Um estudo focado no Estado da Paraíba e na cidade de Campina Grande utilizou este modelo para entender a evolução da doença na região. Os resultados indicaram uma redução no número de reprodução efetivo, destacando a importância de compreender a dinâmica da transmissão da infecção e avaliar a eficácia das medidas de controle para prevenir a transmissão sustentada em novas áreas (FRANCO; DUTRA, 2021).

A regressão linear também será usada em uma análise posterior para entender a relação das variáveis densidade populacional por área urbanizada, leitos de UTI por habitante e enfermeiros por habitante com a variável resposta que será a taxa de crescimento

de casos obtido através de um dos modelos citados acima que melhor se ajustar ao município.

2.1 Estimação dos parâmetros

O Método dos Mínimos Quadrados (MMQ) é uma técnica estatística fundamental para ajustar modelos a conjuntos de dados. No contexto da pandemia de COVID-19, essa técnica torna-se essencial para entender a evolução e propagação do vírus em diferentes municípios brasileiros.

Ao ajustar uma linha de regressão aos dados de COVID-19, assumimos que existe uma relação, possivelmente linear, entre os casos acumulados (variável dependente y_t) e o número de dias desde o primeiro caso registrado (variável independente t). A linha de regressão ajustada nos permite prever a resposta, y_t , para um dado valor de t . (DISCOVERY, n.d.)

$$\mu_{Y|t} = \beta_0 + \beta_1 t$$

No entanto, para qualquer observação específica, o valor real de y_t pode desviar do valor previsto. Estes desvios são chamados de erros ou resíduos. O MMQ busca minimizar a soma dos quadrados destes resíduos, proporcionando o melhor ajuste possível ao modelo.

$$\sum e_i^2 = \sum (y_t - \bar{y}_t)^2$$

As vantagens de utilizar o MMQ no contexto da COVID-19 incluem:

- **Precisão:** O MMQ fornece estimativas não viesadas dos parâmetros, garantindo um ajuste confiável aos dados.
- **Flexibilidade:** Pode ser aplicado a diversos modelos, incluindo linear, exponencial e Gompertz, permitindo análises diversificadas da propagação do vírus.
- **Interpretabilidade:** Os coeficientes estimados têm interpretações claras, como a taxa de crescimento do vírus.

2.2 Regressão linear

A Regressão Linear é uma técnica estatística que busca modelar e analisar as relações entre uma variável dependente e uma ou mais variáveis independentes. Segundo (MAULUD; ABDULAZEEZ, 2022), a regressão linear é um dos algoritmos estatísticos e de aprendizado de máquina mais comuns e abrangentes, sendo usada para encontrar uma relação linear entre um ou mais preditores. A principal característica da regressão linear é

a sua capacidade de fornecer uma linha reta que melhor se ajusta aos dados observados, minimizando a soma dos quadrados das diferenças (resíduos) entre os valores observados e os valores previstos pelo modelo.

No contexto deste trabalho, o modelo de regressão linear é especificamente expresso como:

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

No qual:

- y_t é a variável dependente, representando o número acumulado de casos no dia t .
- t é a variável independente representando o número de dias desde o primeiro caso registrado.
- β_0 é o intercepto.
- β_1 é o coeficiente associado ao número de dias.
- ϵ_t é o erro aleatório.

As condições dos erros aleatórios, ϵ_t , são fundamentais para a regressão linear:

- **Valor esperado dos erros:** $E(\epsilon_t) = 0$. Isso implica que os erros têm uma média zero, ou seja, eles não são nem sistematicamente positivos nem negativos.
- **Homoscedasticidade:** $Var(\epsilon_t) = \sigma^2$. A variância dos erros é constante ao longo do tempo.
- **Independência dos erros:** $Cov(\epsilon_t, \epsilon_s) = 0$ para $t \neq s$. Isso significa que os erros em diferentes pontos no tempo são não correlacionados.
- **Normalidade dos resíduos:** Os resíduos seguem uma distribuição aproximadamente normal. Facilitando testes de hipóteses e inferências estatísticas.

Estas condições garantem que as estimativas dos coeficientes β_0 e β_1 sejam não viesadas e eficientes.

Para ajustar o modelo de regressão linear aos dados, é comum utilizar o método dos mínimos quadrados, que busca minimizar a soma dos quadrados dos resíduos. O coeficiente de determinação ajustado, R^2 ajustado, é frequentemente usado para avaliar a qualidade do ajuste do modelo, indicando a proporção da variância na variável dependente que é previsível a partir das variáveis independentes.

Neste trabalho, a Regressão Linear será empregada para analisar a taxa de crescimento dos casos de COVID-19 nos municípios brasileiros, com base nos dados obtidos do

Ministério da Saúde e do Censo do IBGE de 2022 e a relação de variáveis demográficas e sociais com essa taxa. A variável dependente são os casos acumulados, enquanto a variável independente é número de dias, que representa o número de dias desde o início da observação na regressão linear inicial.

Na regressão secundária, a variável resposta é a taxa de crescimento, enquanto as variáveis explicativas incluem densidade populacional por área urbanizada, leitos de UTI por habitante e enfermeiros por habitante. A escolha dessas variáveis se baseia na suposição de que a quantidade e a densidade de pessoas por área urbanizada são fatores críticos, dada a alta transmissibilidade da COVID-19. Presume-se que, quanto maior o número de pessoas por metro quadrado, maior será o número de casos. Por outro lado, a quantidade de leitos de UTI e de enfermeiros por habitante são indicadores de saúde selecionados para compreender quais recursos estão mais correlacionados com os padrões de crescimento da doença.

O ajuste do modelo será avaliado pelo coeficiente de determinação ajustado R^2 ajustado. Os resultados deste estudo proporcionarão percepções sobre a trajetória da pandemia em diferentes municípios brasileiros, contribuindo para uma compreensão mais aprofundada da evolução da doença no país.

Pegando um exemplo, foi escolhido um município com um alto R^2 ajustado no modelo linear, Belo Horizonte – MG.

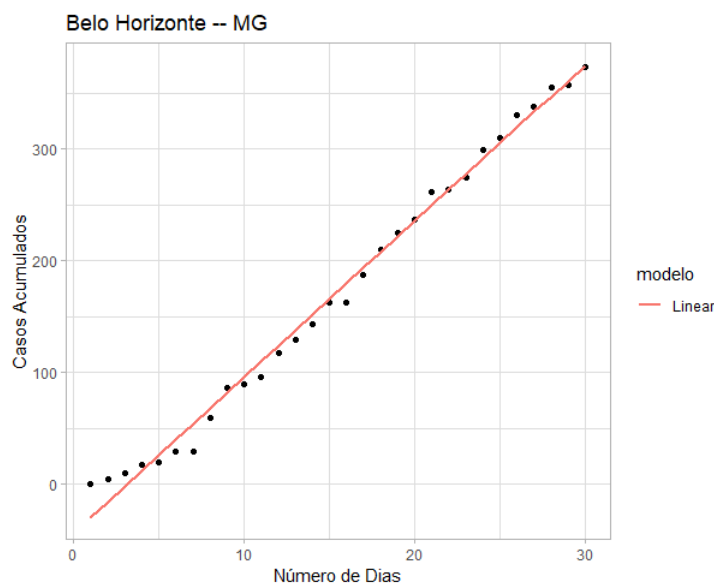


Figura 1 – Modelo linear Belo Horizonte - MG

2.3 Regressão exponencial

A Regressão Exponencial é empregada quando se deseja capturar padrões de crescimento que não são lineares, mas sim exponenciais. Em muitos contextos biológicos,

como o crescimento bacteriano citado por (ROHIM et al., 2020), observa-se que a taxa de aumento de uma variável dependente é proporcional ao seu valor atual. Diferentemente da regressão linear, que estabelece uma relação direta entre as variáveis, a regressão exponencial é adequada para situações em que as mudanças na variável dependente aceleram ou desaceleram com o tempo. Esta técnica é essencial para modelar fenômenos que exibem crescimento exponencial, proporcionando uma compreensão mais profunda das dinâmicas subjacentes.

A forma geral da equação de regressão exponencial é:

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

No contexto deste trabalho, o modelo de regressão exponencial é linearizado e é expresso da seguinte forma:

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

No qual:

- y_t é a variável dependente, representando o número acumulado de casos no dia t .
- t é a variável independente representando o número de dias desde o primeiro caso registrado.
- β_0 é o intercepto.
- β_1 é o coeficiente associado ao número de dias.
- ϵ_t é o erro aleatório.

As condições para os resíduos (erros) em modelos de regressão, seja ela linear ou exponencial, são geralmente as mesmas, o que muda é que, no modelo exponencial, você está linearizando o modelo ao tomar o logaritmo da variável resposta. Isso pode ter implicações na maneira como você verifica e interpreta as condições dos resíduos. Por exemplo, a homoscedasticidade no modelo exponencial refere-se à variância constante dos resíduos em relação aos valores ajustados no espaço log-transformado.

Para ajustar o modelo de regressão exponencial aos dados, é comum transformar a equação exponencial em uma equação linear através da aplicação do logaritmo natural. Uma vez transformado, o método dos mínimos quadrados pode ser aplicado para estimar os coeficientes. O coeficiente de determinação ajustado, R^2 ajustado, é usado para avaliar a qualidade do ajuste do modelo, indicando a proporção da variância na variável dependente que é previsível a partir das variáveis independentes.

Neste trabalho, a Regressão Exponencial será empregada para modelar a taxa de crescimento dos casos de COVID-19 em municípios brasileiros. A escolha deste modelo é

motivada pela natureza exponencial da propagação de doenças infecciosas, especialmente nas fases iniciais de um surto. A adequação do modelo será avaliada pelo coeficiente de determinação ajustado R^2 ajustado, e os resultados ajudarão a identificar os principais determinantes da propagação da doença em diferentes municípios.

Pegando um exemplo, foi escolhido um município com um alto R^2 ajustado no modelo exponencial, Vila Velha – ES.

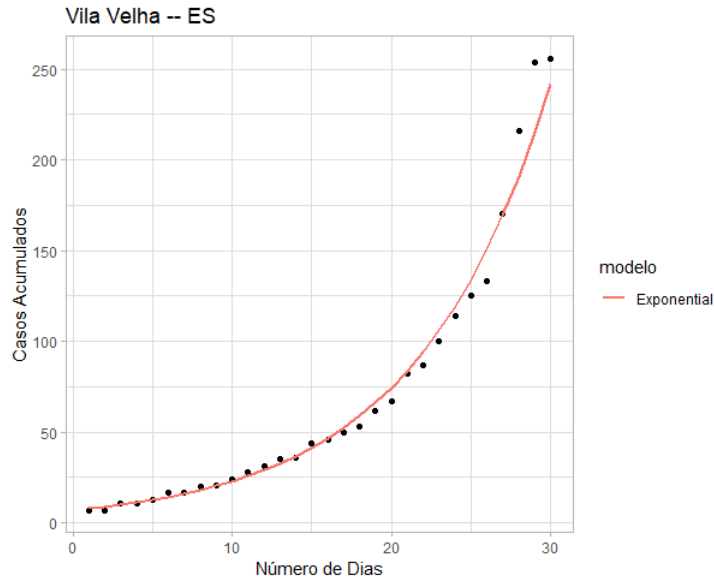


Figura 2 – Modelo exponencial Vila Velha - ES

2.4 Modelo de Gompertz

O modelo Gompertz é uma função sigmoideal frequentemente utilizada para descrever processos de crescimento biológico e populacional. Este modelo é caracterizado por um crescimento inicial exponencial que desacelera e se aproxima assintoticamente de um limite superior. Em contextos epidemiológicos, o modelo Gompertz pode ser particularmente útil para descrever a dinâmica de crescimento de doenças infecciosas, especialmente quando se espera que a taxa de infecção desacelere à medida que se aproxima da saturação, logo consegue lidar bem com um fluxo alto de casos, logo se ajusta melhor em municípios mais populosos.

A função de Gompertz é definida como:

$$y(t) = Ke^{(-e^{(\beta_0 + \beta_1 t + \epsilon_t)})}$$

No contexto deste trabalho, o modelo de Gompertz é linearizado e é expresso da seguinte forma:

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

Em que:

- y_t é a variável dependente, representando o número acumulado de casos no dia t .
- K é o limite assintótico $+ 1$, ou seja, o valor máximo que y_t pode atingir $+ 1$.
- t é a variável independente representando o número de dias desde o primeiro caso registrado.
- β_0 é o intercepto.
- β_1 é o coeficiente associado ao número de dias.
- ϵ_t é o erro aleatório.

O ajuste do modelo Gompertz aos dados é geralmente realizado por meio de técnicas de otimização não-linear, buscando estimar os parâmetros a , b e c que minimizam a diferença entre os valores observados e os valores previstos pelo modelo, mas como linearizamos o modelo, podemos usar o MMQ para estimar os parâmetros. O coeficiente de determinação ajustado, R^2 ajustado, é frequentemente usado para avaliar a qualidade do ajuste do modelo.

Neste trabalho, o modelo Gompertz será empregado para descrever a trajetória de crescimento dos casos de COVID-19 em municípios brasileiros. A escolha deste modelo é motivada pela sua capacidade de capturar a dinâmica de crescimento inicial acelerado seguido de uma desaceleração à medida que a propagação da doença se aproxima de um limite. A adequação do modelo será avaliada pelo coeficiente de determinação ajustado, R^2 ajustado, fornecendo percepções sobre a natureza da propagação da COVID-19 em diferentes contextos municipais.

Em particular, o modelo Gompertz tem sido empregado em estudos recentes para modelar a propagação da COVID-19 em diferentes contextos, demonstrando sua relevância e aplicabilidade em cenários epidemiológicos. (CATALA et al., 2020)

Pegando um exemplo, foi escolhido um município com um alto R^2 ajustado no modelo de Gompertz, Belém – PA.

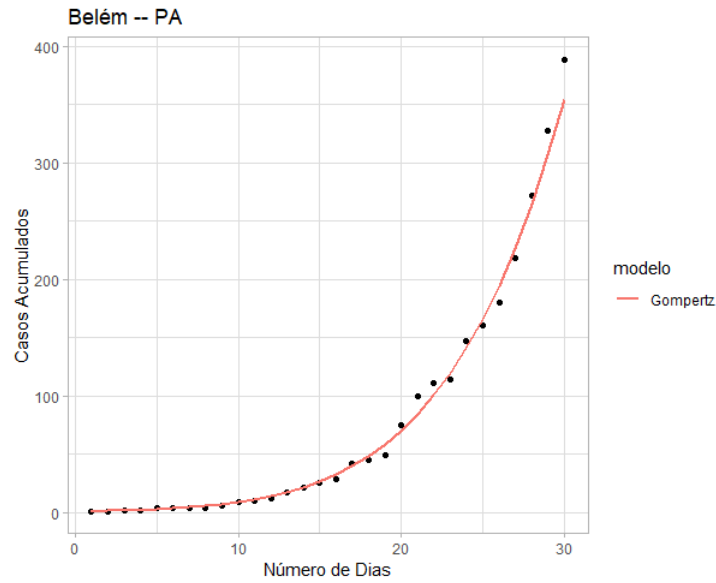


Figura 3 – Modelo de Gompertz Belém - PA

2.5 Comparação entre os modelos

Depois de mostrar todos os modelos estatísticos, vale lembrar que para cada um deles, os parâmetros têm interpretações específicas:

- No modelo linear, β_0 é o valor da variável dependente y_t quando t é zero (intercepto), e β_1 representa a mudança esperada em y_t para uma unidade de mudança em t .
- No modelo exponencial, após a transformação logarítmica, β_0 é o logaritmo do valor da variável dependente y_t quando t é zero, e β_1 indica a taxa de crescimento logarítmica.
- No modelo de Gompertz, β_0 é uma constante relacionada ao limite assintótico e à forma da curva de crescimento, enquanto β_1 está associado à taxa de crescimento da curva.

Um município específico foi escolhido com o intuito de olhar os 3 modelos numa curva, o município foi escolhido Recife – PE. Apesar de visualmente o modelo exponencial parecer lidar melhor com a curva da capital pernambucana, podemos ver pelo R^2 ajustado que na verdade é o modelo de Gompertz que melhor explica essa curva:

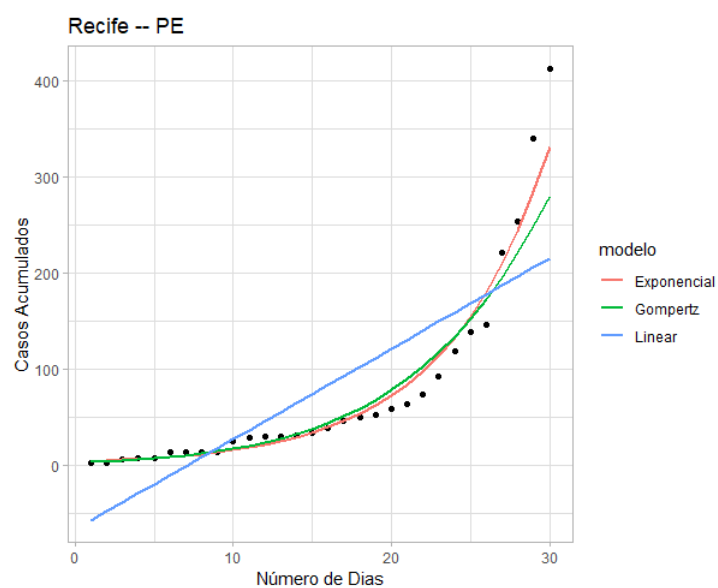


Figura 4 – Os modelos em Recife – PE

Tabela 1 – Valores de R^2 ajustado para os modelos em Recife

Modelo	R^2 ajustado
Linear	0.638
Exponencial	0.952
Gompertz	0.959

3 Resultados e Discussões

3.1 Dados de COVID-19

Cada município experimentou seus próprios 30 primeiros dias da pandemia, começando no dia em que o primeiro caso foi registrado localmente. Estes dias iniciais são fundamentais para entender a propagação do vírus, pois representam o momento em que as comunidades começam a enfrentar o desafio da doença. Este capítulo detalha os dados coletados durante esses períodos iniciais em diferentes municípios, fornecendo uma visão detalhada da situação epidemiológica em diversas partes do Brasil.

Os 3 modelos de regressão a serem feitos (Linear, Exponencial e Gompertz) utilizam as seguintes variáveis na primeira análise:

- **Casos acumulados (y_t):** A variável resposta é o total de casos confirmados de COVID-19 até a data de referência.
- **Número de dias:** A variável explicativa é o número de dias desde o início da observação.

No segundo momento, será feito um modelo de regressão linear para tentar explicar a taxa de crescimento obtida por uma das regressões anteriores e vão possuir as seguintes variáveis

- **Taxa de crescimento (y_t):** A variável resposta é a taxa de crescimento obtida pelo modelo que melhor se encaixou no município.
- **Densidade populacional por área urbanizada:** Representa o número de habitantes por unidade de área urbanizada.
- **Leitos de UTI por habitante:** Indica a quantidade de leitos de UTI existentes por pessoa em um município.
- **Enfermeiros por habitante:** Reflete a proporção de enfermeiros em relação à população total.

Para as regressões, utilizou-se apenas essas variáveis. No entanto, foram coletados outros dados que serão utilizados em análises exploratórias. Todos os dados foram obtidos a partir de fontes oficiais, especificamente do Ministério da Saúde e do Censo do IBGE de 2022. As variáveis coletadas para análises exploratórias e suas respectivas descrições são:

- **Código do município (IBGE):** Código numérico único atribuído pelo Instituto Brasileiro de Geografia e Estatística (IBGE) a cada município brasileiro.
- **Mesorregião:** Divisão regional que agrupa diversos municípios com características geográficas, sociais e econômicas semelhantes.
- **Estado:** Unidade federativa brasileira onde o município está localizado.
- **Região:** Uma das cinco grandes áreas em que o Brasil é dividido (Norte, Nordeste, Centro-Oeste, Sudeste e Sul) e que agrupa estados com características geográficas e culturais semelhantes.

Antes de prosseguir com as análises, foi necessário realizar um processo rigoroso de limpeza e preparação dos dados. Isso envolveu a remoção de registros duplicados, tratamento de valores ausentes e correção de possíveis inconsistências. Além disso, devido à natureza das fontes de dados, houve a necessidade de unir diferentes tabelas para obter um conjunto de dados completo e coeso. Esta união foi realizada com base no código do município, garantindo que as informações de diferentes fontes fossem corretamente alinhadas.

Para a análise dos dados, foi utilizado o software R. As visualizações geográficas dos casos, por sua vez, foram elaboradas com o auxílio da ferramenta QGIS, permitindo uma representação espacial clara da disseminação do vírus nos municípios brasileiros.

Para tratar o banco de dados e poder realizar as regressões diversos pacotes do R foram utilizados para facilitar e otimizar o processo. O pacote `tidyverse` foi fundamental, pois oferece uma coleção de pacotes relacionados à manipulação e visualização de dados. O `ggplot2` permitiu a criação de visualizações gráficas avançadas, enquanto o `dplyr` foi essencial para a manipulação de dados. Além disso, o `lubridate` ajudou na gestão de datas e o `stringr` na manipulação de strings. Juntos, esses pacotes proporcionaram uma análise de dados robusta e eficiente.

3.2 Análise exploratória de dados

Para começar a análise exploratória, podemos fazer uma análise descritiva e olhar o histograma dos casos acumulados:

Tabela 2 – Estatísticas dos casos acumulados

Estatística	Valor
Mediana	3
Média	9.59
Máximo	1794

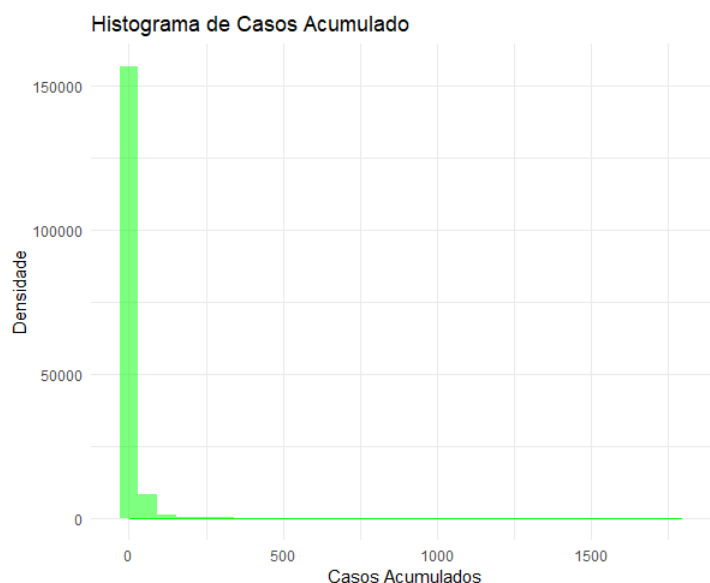


Figura 5 – Histograma dos casos acumulados

Isso já nos indica que existem valores elevados influenciando o aumento da média, o que pode sugerir a presença de outliers, que podem ser confirmados pelo histograma. Embora seja intuitivo pensar que o maior número de casos ocorreria em locais com maior população, como São Paulo, o número máximo de casos acumulados foi registrado na capital cearense, Fortaleza.

Podemos examinar por meio de um mapa coroplético a distribuição espacial de quais modelos funcionam melhor em cada município.

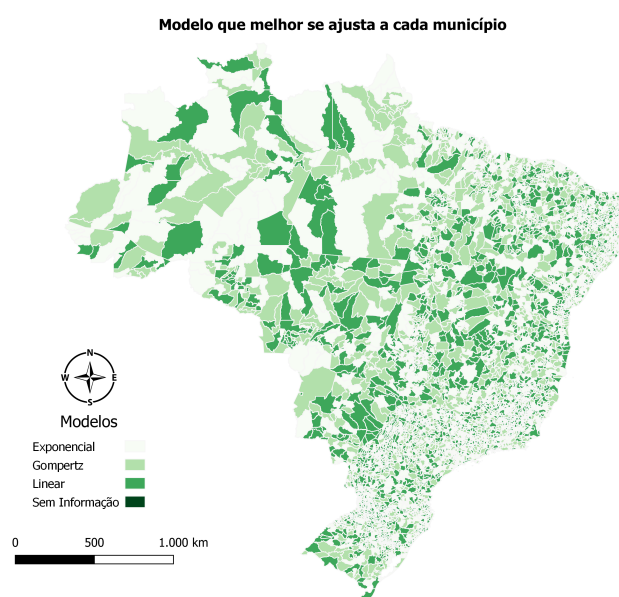


Figura 6 – Modelos por município

Visualmente, não podemos detectar um padrão por área, o que sempre é uma dificuldade ao trabalhar com uma área tão extensa como todo o território brasileiro, porém um fato interessante é que o modelo de Gompertz é o que melhor explica a maioria das capitais. Logo abaixo podemos ver a quantidade de municípios que se ajustaram melhor por modelo.

Tabela 3 – Número de municípios por melhor modelo ajustado

Modelo	Municípios
Exponencial	1485
Gompertz	2032
Linear	2053

Temos uma distribuição equilibrada de municípios entre os modelos. O modelo de Gompertz tende a se adequar melhor a municípios mais populosos. Por isso, é compreensível que o modelo linear explique um número maior de municípios, uma vez que existem mais municípios com populações menores do que aqueles com grandes populações. Por outro lado, o modelo exponencial parece ser o mais adequado para os municípios que não se ajustam bem nem ao Gompertz nem ao linear.

Podemos também analisar os municípios que apresentam as maiores taxas de crescimento em cada modelo, a fim de identificar possíveis padrões e compreender as diferenças entre os modelos. A taxa de crescimento foi determinada utilizando o coeficiente β_1 da variável explicativa no modelo linear. No entanto, para os modelos exponencial e Gompertz, é necessário reverter os dados para suas formas originais, uma vez que haviam sido linearizados. No modelo exponencial, a taxa de crescimento é representada por $\exp(\beta_1)$. Já no modelo de Gompertz, a taxa é dada por $K/\exp(\exp(\beta_1))$.

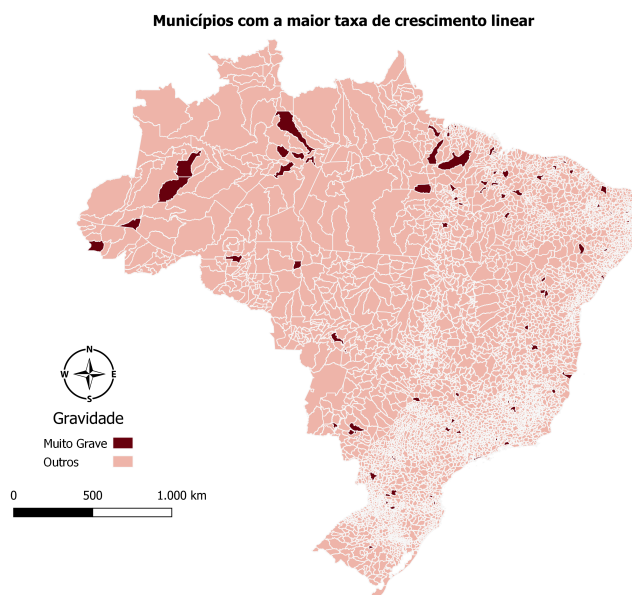


Figura 7 – Municípios com maiores taxas de crescimento linear

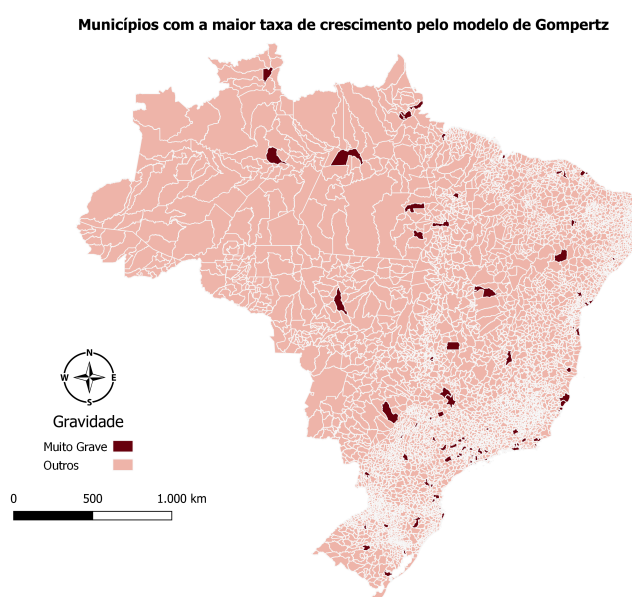


Figura 8 – Municípios com maiores taxas de crescimento pelo modelo de Gompertz

Sobre o mapa do modelo linear, ele entrega muito pouca informação sobre padrões visualmente, mas um fato interessante é que 2 capitais (Belo Horizonte e Cuiabá) estão nos 5% municípios com maior taxa de crescimento linear e as outras 5 capitais que são melhores explicadas por esse modelo são: Maceió, Campo Grande, Aracaju, Palmas e Rio Branco, todos essas capitais, com exceção de BH, aparentam ter uma população menor.

Quando verificamos o modelo de Gompertz não temos um padrão espacial bem definido, porém temos um padrão de capitais. O modelo explica melhor 17 capitais, das quais todos estão nessa lista das maiores taxas de crescimento, as 17 que ele explica melhor são as capitais mais populosas do país, com exceção da maior metrópole, São Paulo.

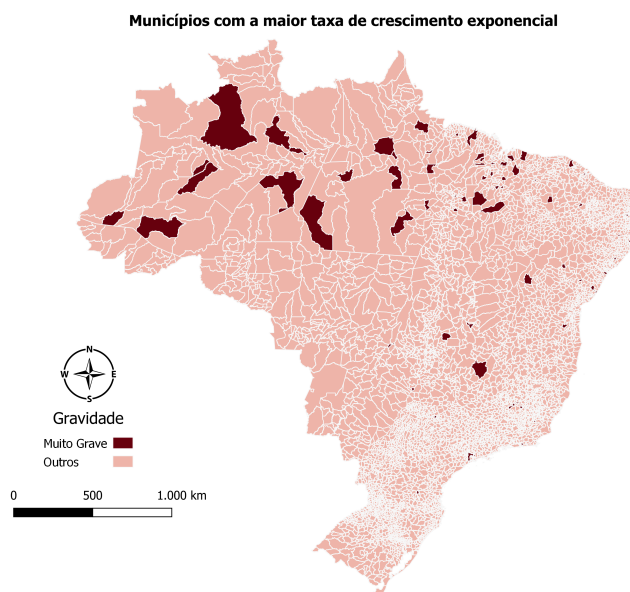


Figura 9 – Municípios com maiores taxas de crescimento exponencial

Ao olharmos para o modelo exponencial percebemos um leve padrão indicando uma possível maior presença na região Norte, porém como o Brasil é muito grande, fica difícil de confirmar essa teoria. Esse modelo explica melhor 3 capitais, Porto Velho, São Luís e São Paulo. Tudo isso pode ser visto na tabela abaixo.

Tabela 4 – Capitais e seus modelos

Modelo linear	Modelo exponencial	Modelo de Gompertz
Belo Horizonte - MG	Porto Velho - RO	Belém - PA
Cuiabá - MT	São Luís - MA	Boa Vista - RR
Maceió - AL	São Paulo - SP	Brasília - DF
Campo Grande - MS		Curitiba - PR
Aracaju - SE		Florianópolis - SC
Palmas - TO		Fortaleza - CE
Rio Branco - AC		Goiânia - GO
		João Pessoa - PB
		Natal - RN
		Recife - PE
		Rio de Janeiro - RJ
		Salvador - BA
		Teresina - PI
		Vitória - ES
		Porto Alegre - RS
		Manaus - AM

Em conclusão, a primeira impressão é que os modelos linear e exponencial tendem a se focar em municípios com uma menor população, porém existem exceções como é o caso da capital mineira e da capital paulista. Para tentarmos entender melhor isso podemos fazer uma análise descritiva das populações dos municípios por modelo.

Podemos enxergar pela média e 3º quartil que de fato o Gompertz lida melhor com

Tabela 5 – Resumo da População em 2022 por Modelo que Melhor explica o município

Modelo	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Gompertz	877	5,576	12,120	48,180	27,661	6,625,849
Exponencial	998	5,874	12,245	36,046	24,951	12,200,180
Linear	836	4,807	9,943	27,434	21,365	2,392,678

municípios mais populosos, enquanto o exponencial parece lidar melhor com municípios com uma população mais mediana e o linear com municípios de baixa população. Para compreender ainda melhor o padrão de crescimento, é útil observar o comportamento dos casos acumulados nos primeiros 30 dias da pandemia em cada região brasileira. Esta análise, semelhante à realizada em (FERRAZ et al., 2020), pode nos oferecer percepções sobre as diferenças no padrão de crescimento entre as regiões

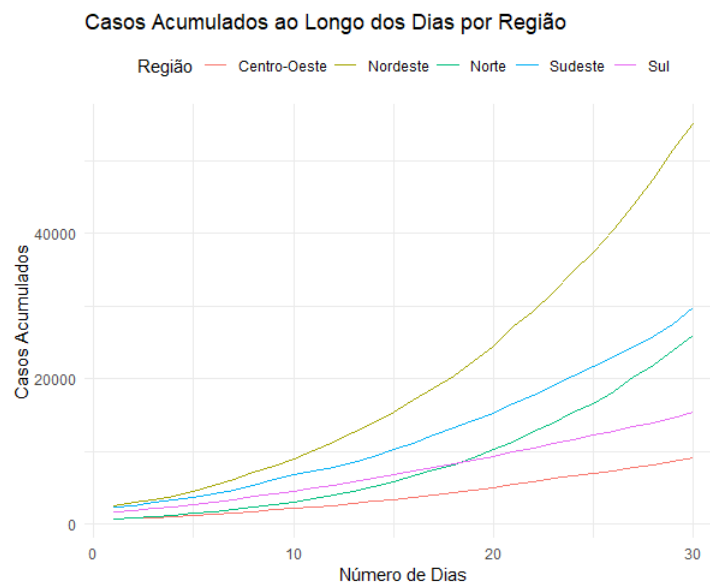


Figura 10 – Comparação do padrão de crescimento das regiões do Brasil

Observa-se que cada região apresenta um padrão de crescimento distinto, o que reforça a importância de explorar diversos modelos para determinar qual se ajusta melhor a cada município.

Conforme evidenciado no mapa coroplético, há uma dificuldade de visualização ao considerar todos os municípios brasileiros simultaneamente. Ao focar em um único estado, é possível analisar a distribuição dos modelos em relação as mesorregiões e a capital. Nesse contexto, optou-se por examinar o estado de Pernambuco, terra natal do autor.

Ao examinar o mapa, notamos que a região metropolitana, que envolve a capital pernambucana, apresenta a maior proporção de municípios cujo modelo mais adequado foi o de Gompertz. Por outro lado, na mesorregião do São Francisco, a menos populosa de Pernambuco, os modelos linear e exponencial foram os que melhor explicaram a distribuição dos municípios. Essa observação sugere que o número de habitantes pode influenciar o

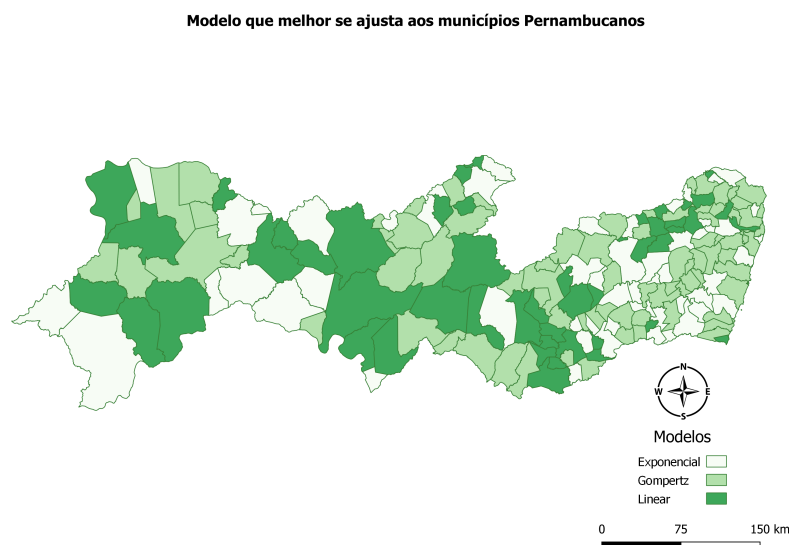


Figura 11 – Melhor modelo por município

Tabela 6 – Porcentagem de Modelos por Mesorregião

Mesorregião	Linear	Exponencial	Gompertz
Agreste	33.8%	31.0%	35.2%
Região Metropolitana	13.3%	20.0%	66.7%
Sertão	31.7%	22.0%	46.3%
São Francisco	40.0%	46.7%	13.3%
Zona da Mata	11.6%	34.9%	53.5%

número de casos. Contudo, é intrigante que o município com o maior número de casos não seja a capital e cidade mais populosa, Recife, mas sim Paulista, localizada na região metropolitana. Esse fato coloca uma dúvida sobre a significância da correlação entre população e número de casos. Seriam necessárias análises mais profundas e classificações adicionais para entender completamente esse padrão.

Uma análise relevante envolve examinar o padrão de crescimento das mesorregiões de Pernambuco, permitindo avaliar se, em um escopo mais restrito, as curvas apresentam maior similaridade.

Observa-se uma sutil semelhança nas curvas das mesorregiões do Agreste e da Zona da Mata. No entanto, as demais regiões demonstram padrões distintos. Isso evidencia que, mesmo com uma análise mais detalhada, os municípios mantêm características singulares. Isso reforça a necessidade de desenvolver um modelo específico para cada município e considerar uma futura clusterização para identificar variáveis que influenciam na similaridade dos modelos.

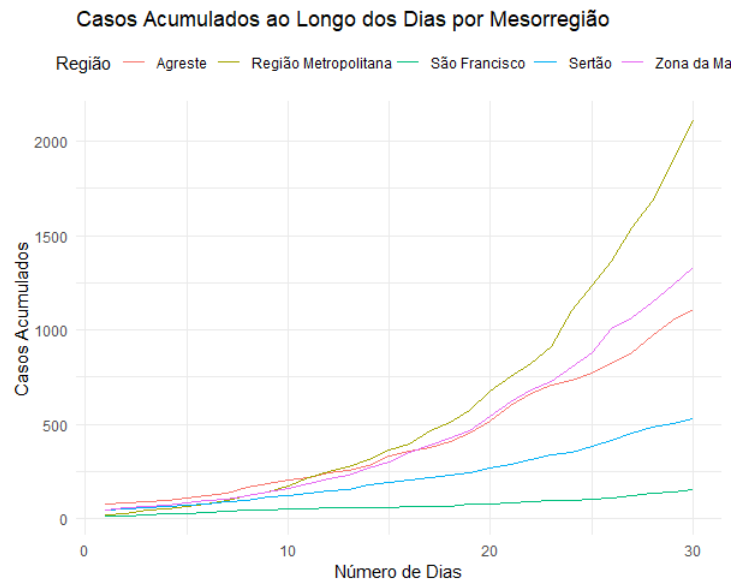


Figura 12 – Comparação do padrão de crescimento das mesorregiões de Pernambuco

3.3 Modelos de Regressão

Como dito anteriormente, foram feitos 3 modelos de regressão (Linear, Exponencial e Gompertz) nos quais a variável dependente são os casos acumulados, enquanto a variável independente é o número de dias. Uma vez que os modelos foram ajustados, métricas de ajuste foram extraídas para avaliar o desempenho de cada modelo. A métrica principal considerada foi o R^2 ajustado, que indica a proporção da variância nos dados que é explicada pelo modelo.

Porém, como estamos falando de um país com uma vasta extensão territorial com muitas diferenças espaciais, podemos comparar abordagens para ver qual funciona melhor. Abordagem Global vs. Abordagem por Município: Ao usar a função do R `lm` diretamente em todo o conjunto de dados, estarei buscando uma abordagem global. Isso significa que tentarei encontrar uma única relação linear que melhor se ajuste a todos os meus dados. Por outro lado, ao usar a função do R `map` para ajustar modelos lineares a subconjuntos de dados, estarei adotando uma abordagem por município. Isso pode revelar variações nas relações lineares entre diferentes municípios (KAN; DOE; SMITH, 2019).

Ao usar o modelo global temos uma simplicidade computacional, porém ao usar o modelo por município ganhamos em 2 aspectos, Flexibilidade e Personalização. Por exemplo, ao descobrir que um modelo linear não é adequado para um município específico, podemos optar por ajustar um modelo diferente apenas para esse municípios. Ou também ao ajustar modelos separados para subconjuntos de dados, podemos descobrir nuances ou padrões específicos do município que podem ser obscurecidos em uma análise global.

Em resumo, enquanto o `lm` fornece uma visão global dos seus dados, o uso do `map` para ajustar modelos a subconjuntos de dados permite uma análise mais granular e

específica. Ao comparar os resultados de ambas as abordagens, obtemos uma compreensão mais completa e matizada das relações em seus dados. Devido a grande extensão territorial e a vasta diferença entre os municípios é muito provável que o modelo global não se ajuste bem, aqui está uma tabela com os R^2 ajustado nos modelos 1m

Tabela 7 – R^2 Ajustado no modelo global

Modelo	R^2 Ajustado
Linear	0.066
Exponencial	0.253
Gompertz	0.142

O modelo exponencial é levemente melhor, mas não obtém um resultado significativo, como era esperado, logo o ideal é usar a abordagem por município. Para determinar qual modelo se ajusta melhor aos dados em geral, foram calculadas a média e a mediana do R^2 ajustado para cada modelo obtido pela função `map`. A discrepância entre média e mediana pode ser atribuída à presença de outliers ou a uma distribuição assimétrica dos valores de R^2 ajustado. Enquanto a média é sensível a outliers, a mediana oferece uma visão mais centrada da distribuição dos dados.

A escolha entre média e mediana, ao avaliar o desempenho dos modelos, pode ter implicações práticas significativas. Por exemplo, se estivermos interessados em um modelo que funcione bem para a maioria dos municípios (e não necessariamente em média), a mediana pode ser uma métrica mais relevante. Por outro lado, se estivermos interessados em um desempenho geral, a média pode ser mais informativa.

Tabela 8 – R^2 Ajustado no modelo por município

Modelo	Média R^2 ajustado	Mediana R^2 ajustado
Linear	0.743	0.801
Exponencial	0.770	0.833
Gompertz	0.760	0.831

O objetivo principal da análise é determinar qual modelo se ajusta de forma mais adequada aos diferentes municípios, logo o resultado da mediana é mais interessante. Os resultados da mediana do R^2 ajustado revelam que, para ao menos 50% dos municípios, o R^2 ajustado é superior a 0.8. Isso é um indicativo positivo, sugerindo que esses três modelos podem efetivamente representar a maioria dos padrões de crescimento dos municípios. Isso é interessante dado que o modelo exponencial foi o modelo com o menor número de estados na qual ele explicava melhor. Uma análise de agrupamento, que será abordada em um trabalho futuro, se apresenta como o próximo passo lógico.

3.4 Análise de Variáveis

Agora que foi possível analisar os modelos iniciais, o próximo passo é usar a regressão linear para tentar explicar a taxa de crescimento pelas variáveis densidade populacional por área urbanizada, leitos de UTI por habitante e enfermeiros por habitante. A taxa de crescimento pode ser descrita em cada modelo como:

- **Modelo linear:** A taxa de crescimento foi calculada utilizando o coeficiente β_1 da variável explicativa, representando diretamente a taxa de crescimento no modelo linear.
- **Modelo exponencial:** Dado que o modelo original estava linearizado, para obter uma noção da gravidade na escala original, é necessário converter os dados de volta para o modelo exponencial. A taxa de crescimento, nesse caso, é dada por $\exp(\beta_1)$.
- **Modelo de Gompertz:** Similarmente, para o modelo de Gompertz, a taxa de crescimento é obtida revertendo a linearização, resultando na fórmula $K/\exp(\exp(\beta_1))$.

Agora podemos aplicar a regressão linear que desejávamos a eles, e esses foram os p-valores e R^2 de cada modelo.

Tabela 9 – Resumo dos Resultados das Regressões

Modelo	Variável	R^2 Ajustado	p-valor
Linear	(Intercepto)	0.03545	$6.89e - 05$
	Leitos UTI/hab.		$1.91e - 10$
	Enfermeiros/hab.		0.903
	Densidade por área urbanizada		$3.26e - 06$
Exponencial	(Intercepto)	0.02854	$< 2e - 16$
	Leitos UTI/hab.		$5.99e - 05$
	Enfermeiros/hab.		0.0479
	Densidade por área urbanizada		$3.27e - 07$
Gompertz	(Intercepto)	0.1729	$1.64e - 09$
	Leitos UTI/hab.		$< 2e - 16$
	Enfermeiros/hab.		$5.36e - 09$
	Densidade por área urbanizada		$1.21e - 09$

No modelo linear, a variável enfermeiros por habitante não é estatisticamente significativa, enquanto as demais variáveis demonstram significância. É importante destacar que este modelo é mais adequado para municípios com populações menores. No modelo exponencial, que se adapta melhor a municípios de porte médio em termos de população, a variável enfermeiros por habitante torna-se significativa ao nível de 0.05. Já no modelo de Gompertz, que é mais indicado para municípios com grandes populações, todas as

variáveis são estatisticamente significativas. Ao considerarmos o R^2 ajustado, observamos valores baixos, com o modelo de Gompertz apresentando um valor um pouco mais elevado.

As suposições feitas por meio da análise descritiva da população precisa ser aprofundada em um modelo de classificação, para que possamos dizer com certeza que dependendo do porte populacional do município a significância das variáveis muda. Estas conclusões ressaltam a necessidade de estratégias de intervenção e políticas públicas adaptadas às características específicas de cada município para combater eficientemente a pandemia.

3.5 Análise de Resíduos

A análise de resíduos é uma etapa crucial na avaliação de modelos estatísticos, pois permite verificar se as suposições do modelo são atendidas. Uma das principais suposições é a normalidade dos resíduos. Para verificar essa suposição, faremos o histograma da densidade dos resíduos, o QQ-Plot e empregaremos o Teste de Shapiro-Wilk, que é amplamente reconhecido por sua eficácia na detecção de desvios da normalidade.

A não normalidade dos resíduos pode ter implicações significativas para a validade e interpretação do modelo. Se os resíduos não seguem uma distribuição normal, isso pode indicar a presença de não linearidades não capturadas pelo modelo, heteroscedasticidade ou outliers influentes. Em tais situações, as inferências padrão sobre os coeficientes do modelo podem não ser válidas. Portanto, é crucial abordar e remediar quaisquer violações das suposições do modelo para garantir conclusões robustas e confiáveis.

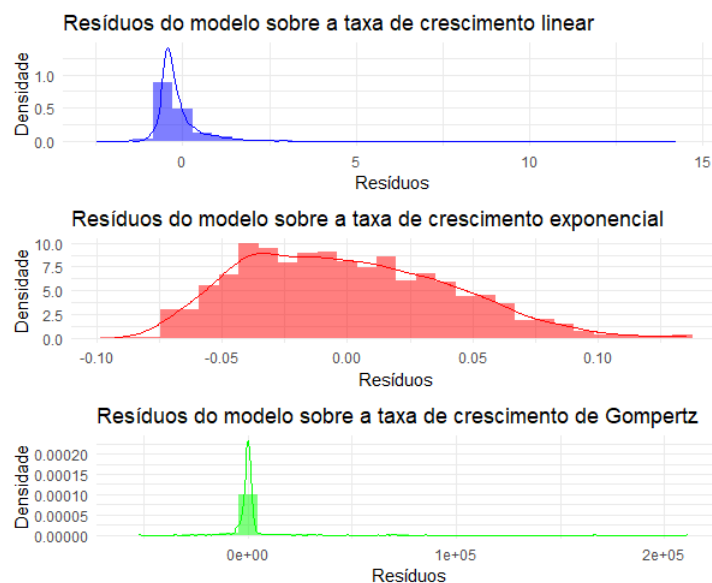


Figura 13 – Densidade dos Resíduos por modelo

Observando o gráfico, notamos que, em nenhum deles podemos perceber uma certa normalidade dos dados, sendo mais próximo o modelo exponencial, mas ainda muito

distante de uma normalidade. Para reforçar essa análise, podemos também avaliar o QQ-plot.

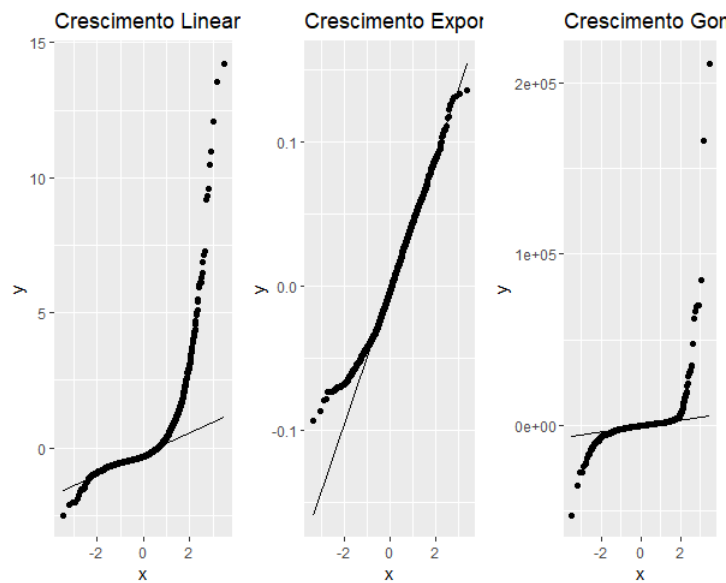


Figura 14 – QQ-Plot dos Resíduos por modelo

Mais uma vez podemos ver que nos modelos não podemos identificar normalidade de jeito nenhum e existe a presença de um outlier no modelo linear e de Gompertz, mais uma vez o modelo exponencial se apresenta um pouco melhor que os demais, mas nada que indique normalidade, podemos tirar a prova usando o teste de Shapiro-Wilk.

Tabela 10 – P-valores do teste de Shapiro-Wilk para os modelos

Taxa de Crescimento	P-valor
Linear - Segunda Regressão	1.06×10^{-58}
Exponencial - Segunda Regressão	7.77×10^{-14}
Gompertz - Segunda Regressão	1.30×10^{-66}

Isso confirma a não normalidade dos resíduos. Uma estratégia comum para abordar a não normalidade é a transformação de Box-Cox. Embora, neste caso específico, ela possa não ser eficaz, em muitos contextos, essa transformação é uma alternativa valiosa para lidar com a não normalidade.

Contudo, a análise de resíduos não se restringe apenas à verificação da normalidade. A homoscedasticidade, que se refere à constância da variância dos resíduos em todos os níveis de uma variável independente, também é crucial. A heteroscedasticidade, que é o oposto da homoscedasticidade, pode ser problemática. Ela viola uma das suposições fundamentais da regressão linear, o que pode resultar em estimativas de coeficientes imprecisas e testes de hipóteses inválidos.

Para avaliar formalmente a homoscedasticidade, recorreremos ao gráfico de resíduo studentizado vs valores preditos. Este gráfico é especialmente útil para detectar padrões específicos de heteroscedasticidade. Aqui podemos observar os gráficos para cada modelo.

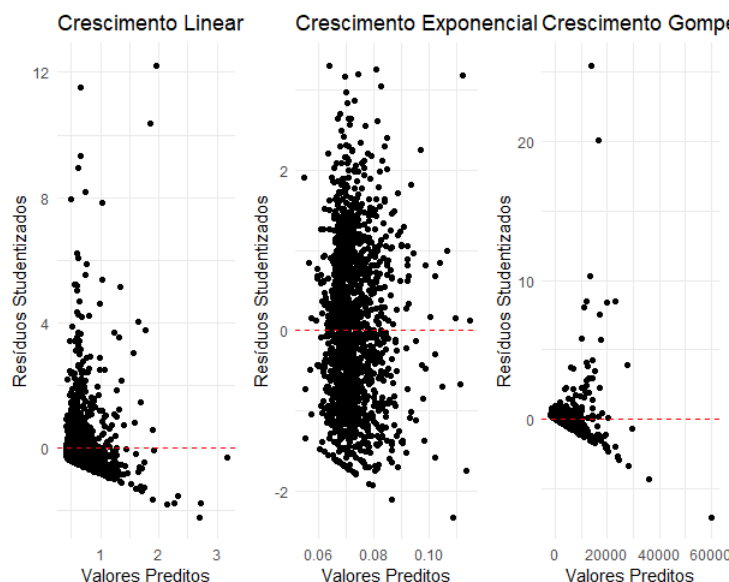


Figura 15 – Resíduo Studentizado vs Valores Preditos

Os resíduos não estão distribuídos de forma aleatória ao longo do eixo x, o que indica heteroscedasticidade, uma forma de tentar resolver essa falha na suposição é método dos mínimos quadrados generalizados (GLS). Podemos ver também que todos os modelos apresentam traços de heteroscedasticidade claro e outliers assim como os gráficos anteriores tinham apontado, além de ser possível apontar alguns padrões, o que pode indicar a presença de autocorrelação.

Contudo ainda devemos checar se de fato existe autocorrelação nos dados. A presença de autocorrelação nos resíduos de um modelo de regressão é problemática, pois viola a suposição de que os resíduos são independentes entre si. Esta suposição é crucial para a validade das inferências estatísticas feitas a partir do modelo. Que pode ser atestada pelo gráfico de resíduo studentizado pelo tempo. Aqui podemos ver os resultados do teste de Durbin-Watson.

Tabela 11 – Valores do teste de Durbin-Watson para os modelos

Taxa de Crescimento	Valor do teste
Linear	0.367
Exponencial	0.202
Gompertz	0.910

O teste produz um valor de estatística entre 0 e 4, na qual entre 0 e 2 sugere que há autocorrelação positiva em todos os modelos.

4 Conclusão

O estudo mostra que a análise do número acumulado de casos de COVID-19 nos primeiros meses de 2020 é desafiadora devido à dependência da contabilização em óbitos e testes em profissionais da saúde, visto que testes em massa para a população geral só se tornaram comuns mais tarde no ano. Devido a esse fato, muitos municípios tiveram menos casos registrados, o que atrapalha a identificação do melhor modelo. Além disso, mostra como é essencial reconhecer a relevância do modelo SIR (Susceptível - Infectado - Removido) na tomada de decisões durante a pandemia, dada sua capacidade de capturar a dinâmica de transmissão da doença.

É evidenciado que a análise de dados em contextos pandêmicos, especialmente nos estágios iniciais, é de suma importância. Ela fornece percepções valiosas sobre a propagação de doenças e pode informar estratégias de intervenção eficazes. No entanto, ao tentar generalizar a situação em um país tão vasto e diversificado como o Brasil, encontramos desafios significativos. A abordagem global compara com a abordagem por município para a modelagem de regressão e a análise espacial, revelou que as nuances locais podem ser ofuscadas quando se tenta analisar o país como um todo, o que pode exemplificado pelo mapa coroplético.

Este estudo também reforçou a ideia de que a modelagem deve ser realizada a nível municipal. Mesmo ao analisar os dados no nível mais granular, como a mesorregião de um estado, observou-se que as curvas de crescimento dos casos variavam amplamente. Esta diversidade intrínseca em padrões de crescimento reforça a necessidade de abordagens de modelagem adaptadas a cada município.

Ao avaliar os modelos propostos, os valores de R^2 ajustado e as análises de resíduos sugerem que os três modelos podem ser adequados para descrever a maioria dos municípios, especialmente aqueles com um fluxo considerável de casos. Isso é promissor, pois indica que, apesar das variações locais, certos padrões podem ser identificados e modelados com precisão.

Ao tentar explicar a taxa de crescimento com base em variáveis sociais e demográficas, os modelos não atenderam a todas as suposições. No entanto, foi revelado que a densidade populacional por área urbanizada e a quantidade de leitos são variáveis significativas para a taxa de crescimento. Além disso, há indícios de que a variável "enfermeiros por habitante" tem influência particularmente em municípios mais populosos. Esta suposição poderia ser explorada e validada por meio de um processo de classificação. Considerando as limitações do modelo linear empregado neste estudo para explicar a taxa de crescimento, pesquisas subsequentes poderiam explorar modelos de regressão mais robustos e adequados.

Por fim, é importante salientar que a motivação principal deste trabalho não é apenas fornecer uma análise retrospectiva da pandemia de COVID-19, mas também deixar um legado para pesquisas futuras. A esperança é que apesar das limitações, este estudo possa servir como base para classificações subsequentes, permitindo a identificação de grupos entre os municípios. Ao fazer isso, podemos começar a explicar algumas descobertas e suposições feitas nesse artigo sobre como variáveis socioeconômicas e demográficas influenciam o padrão inicial de crescimento dos casos de COVID-19, fornecendo percepções valiosos para enfrentar futuros desafios de saúde pública.

Referências

CATALA, M. et al. Empirical model for short-time prediction of covid-19 spreading. *PLOS Computational Biology*, 2020. Disponível em: <<https://dx.doi.org/10.1371/journal.pcbi.1008431>>. Citado na página 9.

DISCOVERY, J. S. *The Method of Least Squares*. n.d. Disponível em: <<https://www.jmp.com/en-fi/statistics-knowledge-portal/what-is-regression/the-method-of-least-squares.html>>. Citado na página 4.

FERRAZ, C. et al. Covid-19: Monitoramento por gráficos de shewhart. *Revista Brasileira de Estatística*, Universidade Federal de Pernambuco, Universidade Estadual de Campinas, Universidade Federal do Rio de Janeiro, Macaé, v. 78, n. 245, p. 23–41, julho-dezembro 2020. Computational Agricultural Statistics Laboratory. Disponível em: <<https://a.storyblok.com/f/134103/ff26467371/3027364.pdf>>. Citado na página 18.

FRANCO, C.; DUTRA, R. Sir model for propagation of covid-19 in the paraíba's state (brazil). *Intermaths*, 2021. Disponível em: <<https://periodicos2.uesb.br/index.php/intermaths/article/download/9696/6464>>. Citado na página 3.

(IBGE), I. B. de Geografia e E. *Censo 2022*. 2022. Disponível em: <<https://www.ibge.gov.br/estatisticas/sociais/populacao/22827-censo-demografico-2022.html?edicao=37225&t=resultados>>. Citado na página 2.

KAN, A.; DOE, B.; SMITH, C. Analysis of linear and non-linear regression models in epidemiological studies. *Journal of Epidemiological Research*, v. 45, n. 2, p. 123–135, 2019. Citado na página 20.

LIU, X.-X. et al. A new seaird pandemic prediction model with clinical and epidemiological data analysis on covid-19 outbreak. *Applied Intelligence*, v. 51, p. 1–20, 2021. Disponível em: <<https://link.springer.com/content/pdf/10.1007/s10489-020-01938-3.pdf>>. Citado na página 1.

MAULUD, D. H.; ABDULAZEEZ, A. A review on linear regression comprehensive in machine learning. *Journal Name*, 2022. Disponível em: <<https://jastt.org/index.php/jasttpath/article/download/57/20>>. Citado na página 4.

ROHIM, R. A. et al. Modeling the growth of bacteria streptococcus sobrinus using exponential regression. *Pesquisa Brasileira em Odontopediatria e Clínica Integrada*, 2020. Disponível em: <<https://www.scielo.br/j/pboci/a/PFvygzNPhNzdrBhgNvjRtBD/?format=pdf&lang=en>>. Citado na página 7.

SAUDE, M. da. *Dados sobre COVID-19*. 2022. Disponível em: <https://infoms.saude.gov.br/extensions/covid-19_html/covid-19_html.html>. Citado na página 2.

SILVA, L. L. S. d. et al. Social distancing measures in the fight against covid-19 in brazil: description and epidemiological analysis by state. *Cadernos de Saúde Pública*, v. 36, n. 9, p. 1–11, 2020. Disponível em: <<https://www.scielo.br/j/csp/a/gR6mkQmSqBHqvZb5YMNYjxD/?lang=pt&format=pdf>>. Citado na página 1.