Universidade do Estado do Rio de Janeiro
Centro de Tecnologia e Ciências
Instituto de Química

Marcello Montillo Provenza

Previsão de séries temporais para os óbitos no Brasil causados pela COVID-19 no âmbito da pandemia

Rio de Janeiro
2022

Marcello Montillo Provenza

Previsão de séries temporais para os óbitos no Brasil causados pela COVID-19 no âmbito da pandemia

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Engenharia Química, da Universidade do Estado do Rio de Janeiro.
Área de concentração: Bioprocessos e Tecnologia Ambiental.

Orientador: Prof. Dr. Aderval Serevino Luna

Orientador: Prof. Dr. Vinicius Layter Xavier

Rio de Janeiro

2022

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC/Q

P969 Provenza, Marcello Montillo.

Previsão de séries temporais para os óbitos no Brasil causados pela COVID-19 no âmbito da pandemia. – 2022.
52 f.

Orientador(a): Aderval Serevino Luna
Vinicius Layter Xavier

Dissertação (Doutorado) – Universidade do Estado do Rio de Janeiro. Instituto de Química.

1. Engenharia química – Teses. 2. Engenharia química – Processamento de dados – Teses. 3. COVID-19 (Doença) – Teses. 4. Econometria – Teses. I. Luna, Aderval Serevino. II. Xavier, Vinicius Layter III. Universidade do Estado do Rio de Janeiro. Instituto de Química. IV. Título.

CDU 66.0::616-002.6

Marcello Montillo Provenza
Assinatura

18/11/2022
Data

Marcello Montillo Provenza

Previsão de séries temporais para os óbitos no Brasil causados pela COVID-19 no âmbito da pandemia

Tese apresentada, como requisito parcial para obtenção do título de Doutor, ao Programa de Pós-Graduação em Engenharia Química, da Universidade do Estado do Rio de Janeiro.
Área de concentração: Bioprocessos e Tecnologia Ambiental.

Aprovada em 23 de setembro de 2022.

Banca Examinadora:

Prof. Dr. Aderval Severino Luna
Instituto de Química - UERJ

Prof. Dr. Vinicius Layter Xavier
Instituto de Matemática e Estatística - UERJ

Prof. Dr. André Luiz Hemerly Costa
Instituto de Química - UERJ

Prof. Dr. Alexandre Rodrigues Tôrres
Faculdade de Tecnologia - UERJ

Prof. Dr. Eduardo Lima Campos
Escola Nacional de Ciências Estatísticas

Prof. Dr. Marcelo Rubens dos Santos do Amaral
Instituto Brasileiro de Geografia e Estatística

Rio de Janeiro
2022

DEDICATÓRIA

Dedico este trabalho ao meu querido e amado Lion (em memória).

AGRADECIMENTOS

Primeiramente a Deus. Apesar de ser um dos muitos católicos chamados de “não praticante”, acredito na existência de uma força superior que olha por todos nós e garante a existência de tudo.

Ao meu amigo, professor, colega de trabalho e orientador Aderval Severino Luna que abraçou a confecção deste projeto em séries temporais e me atendeu muitas vezes na Universidade do Estado do Rio de Janeiro, por telefone e também online tirando dúvidas, dando conselhos, ensinando todas as técnicas preditivas, de modelagem, de planejamento de experimentos e participando efetivamente desta pesquisa, tornando-o melhor e exequível ao longo desse tempo.

Ao meu amigo, professor, colega de faculdade, de trabalho e orientador Vinicius Layter Xavier que pôde sempre me atender e ajudar na parte de programação do software R. Ninguém melhor do que ele para explicar e resolver os problemas que aparecem nos códigos. Como não poderia ser diferente, o mesmo também participou de modo eficaz na elaboração deste trabalho, tornando-o melhor e executável ao longo desse período.

Ao professor Jefferson Góis, o qual conheci no dia de nossa posse como professores da Universidade do Estado do Rio de Janeiro, e me convidou a conhecer o Programa de Pós-Graduação em Engenharia Química, do Instituto de Química, e me apresentou ao professor Aderval Severino Luna. Sem ele, nada disso estaria acontecendo agora.

Aos meus pais Francisco Provenza (em memória) e Wilma Diva Montillo Provenza pela educação e apoio a mim concedidos, à minha irmã Márcia Phoenix também pelo suporte. Não posso deixar de agradecer ao meu sobrinho e afilhado Gabriel Phoenix que faz parte de nossas vidas desde 2011 e trouxe a alegria de uma criança para a família.

À minha namorada Elizabeth Assis dos Santos pelo apoio incondicional e por conviver ao meu lado durante o tempo de elaboração deste trabalho, muitas vezes ficando sozinha, esperando-me terminar de escrever até tarde da noite e sabendo que no dia seguinte seria a mesma coisa. Ao meu filho Erick Santos Provenza que chegou ao mundo há pouco tempo, e agora é o maior amor de minha vida.

Aos meus amigos, alguns ex-alunos, e agora professores da Universidade do Estado do Rio de Janeiro Paulo Henrique Couto Simões, Jorge Luiz de Jesus Goulart, Igor Campos de Almeida Lima e Julio Cesar Siqueira. As resenhas de quinta-feira e o convívio foram fundamentais para tornar essa passagem mais leve e factível de boas risadas.

Aos meus amigos de longa data Rodrigo Doti Correia, Pedro Paulo Oliveira de Souza Ribeiro, Bruno Araújo Ferreira, Marcos Anjos Martins e Daniel Keidel Bou Haya. Todos são como parte de uma família pra mim.

Aos meus antigos amigos de trabalho do Instituto de Segurança Pública que ajudaram muito na parte de minha formação pessoal e profissional, principalmente em relação as diversas análises e tipos de tratamento de banco de dados.

A todos os professores do Programa de Pós-Graduação em Engenharia Química pelo conhecimento passado a mim, não só nas aulas, mas também no dia a dia.

A todos os meus colegas do Programa de Pós-Graduação em Engenharia Química da Universidade do Estado do Rio de Janeiro, mesmo aqueles que ainda não chegaram nesta parte da jornada, mas o tempo e os estudos em conjunto foram muito proveitosos.

A todos meus antigos e atuais alunos da graduação, orientandos e orientados. O conhecimento não vai só do professor para o aluno, mas também vem do aluno para o professor.

Agradeço também aos que não mencionei, mas que de uma forma ou de outra, contribuíram para a confecção deste trabalho, peço perdão pelo esquecimento.

E claro, como bom aquariano, amante da natureza e dos animais, jamais poderia deixar de mencionar meus lindos, queridos e amados bichinhos de estimação. São quatro gatos chamados Lion (em memória), Lara Croft, Preta (ou Michela ou T'Chala - rainha de Wakanda) e Diguinho; e dois cachorros chamados Pateta e Valentina. É incrível a capacidade de conseguirem atrapalhar quando estou trabalhando, escrevendo, vendo televisão ou fazendo qualquer outra coisa. Todos dão muito trabalho, mas mesmo assim, adoro isto. E também aos pássaros Nemo e Loirinho.

Os humanos são mais numerosos do que qualquer outro grande animal na história da Terra. E isso representa uma forma de desequilíbrio ecológico que não pode continuar para sempre. Em algum momento haverá uma correção natural.

David Quammen

RESUMO

PROVENZA, Marcello Montillo. *Previsão de séries temporais para os óbitos no Brasil causados pela COVID-19 no âmbito da pandemia*. 2022. 52 f. Tese. Doutorado em Engenharia Química – Instituto de Química, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.

As previsões de óbitos por COVID-19 são úteis para a formulação de políticas públicas, permitindo a utilização de estratégias de isolamento social mais eficazes e com menor impacto econômico e social, além de promover indicadores de como a população adere às vacinas. O objetivo deste trabalho é explorar um amplo conjunto de métodos de previsão para identificar os melhores modelos sem cobertura vacinal (Caso 1) e com cobertura vacinal (Caso 2) no Brasil. Foram considerados os métodos de Inteligência Artificial e os métodos clássicos de econometria. A técnica de validação cruzada para séries temporais foi implementada, fornecendo assim uma estimativa precisa para avaliar a capacidade preditiva dos modelos. Cada modelo foi ajustado considerando uma base inicial de treinamento de 30 valores. No Caso 1, foram usadas as mortes diárias e acumuladas da base Oxford COVID-19 *Government Response Tracker*. No Caso 2, o conjunto de dados provém do *Our World in Data*, onde a média móvel de sete dias foi adotada como referência para melhorar a qualidade dos dados. No Caso 1 os modelos foram treinados e testados com 266 amostras considerando um horizonte de previsão de 7 dias. No Caso 2 os modelos foram treinados e testados 494 vezes considerando um horizonte de previsão de sete dias, 486 vezes considerando um horizonte de 15 dias e 471 vezes considerando um horizonte de 30 dias. Foram adotados modelos de diferentes classes: algoritmos ETS, ARIMA, regressão e aprendizado de máquina. A comparação entre as previsões foi feita utilizando os resultados médios das métricas de previsão: R^2, RMSE, MAE e MAPE. No Caso 1, as previsões acumuladas ofereceram melhores resultados do que as diárias, pois os modelos são menos influenciados pelas componentes da série temporal: ciclo e sazonalidade. Os melhores resultados para a predição de óbitos diários foram obtidos pelo método de regressão de Ridge (R^2 = 0,772, RMSE = 136 e MAE = 113). Os melhores resultados para predição de óbitos acumulados foram obtidos pelo método de regressão Cubist (R^2 = 0,993, RMSE = 468 e MAE = 409). No Caso 2, o modelo ARIMA com uma diferenciação apresentou os melhores resultados para um horizonte de sete dias (RMSE = 74 e MAE = 64).

Palavras-chave: Séries temporais. Óbitos. COVID-19. Modelos estatísticos. Aprendizado de máquina.

ABSTRACT

PROVENZA, Marcello Montillo. *Time series forecast for deaths in Brazil caused by COVID-19 in the context of the pandemic.* 2022. 52 f. Tese. Doutorado em Engenharia Química – Instituto de Química, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2022.

The predictions of deaths from COVID-19 are useful for the formulation of public policies, allowing the use of more effective social isolation strategies with less economic and social impact, in addition to promoting indicators of how the population adheres to vaccines. The objective of this work is to explore a broad set of prediction methods to identify the best models without vaccination coverage (Case 1) and with vaccination coverage (Case 2) in Brazil. The methods of Artificial Intelligence and the classical methods of econometrics were considered. The cross-validation technique for time series was implemented, thus providing an accurate estimate to assess the predictive capacity of the models. Each model was adjusted considering an initial training base of 30 values. In Case 1, daily and cumulative deaths from the Oxford COVID-19 Government Response Tracker database were used. In Case 2, the dataset comes from Our World in Data, where the seven-day moving average was adopted as a reference to improve data quality. In Case 1 the models were trained and tested with 266 samples considering a forecast horizon of 7 days. In Case 2 the models were trained and tested 494 times considering a forecast horizon of seven days, 486 times considering a horizon of 15 days, and 471 times assuming a horizon of 30 days. Models of different classes were adopted: ETS algorithms, ARIMA, regression, and machine learning. The forecasts were compared using the average results of the forecast metrics: R^2, RMSE, MAE, and MAPE. In Case 1, the accumulated forecasts offered better results than the daily ones, as the models are less influenced by the components of the time series: cycle and seasonality. The best results for the prediction of daily deaths were obtained by the Ridge regression method (R^2 = 0.772, RMSE = 136, and MAE = 113). The best results for predicting cumulative deaths were obtained by the Cubist regression method (R^2 = 0.993, RMSE = 468, and MAE = 409). In Case 2, the ARIMA model with one differentiation showed the best results for a seven-day horizon (RMSE = 74 and MAE = 64).

Keywords: Time series. Deaths. COVID-19. Statistical models. Machine learning.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

SUMÁRIO

INTRODUÇÃO

Vários casos de pacientes com pneumonia foram associados à doença do novo coronavírus humano (SARS-CoV-2, COVID-19) iniciada em dezembro de 2019 (ZHU et al., 2020). O vírus demonstrou uma capacidade abundante de transmissões inter-humanas, espalhou-se rapidamente pelo mundo e se tornou uma pandemia, causando milhões de mortes. Os pacientes infectados apresentaram sintomas significativamente variados, com casos de indivíduos assintomáticos até mesmo ao óbito. A discriminação de pacientes gravemente enfermos daqueles com sintomas leves pode ajudar a compreender as variações individualizadas do prognóstico da COVID-19. Os estudos, as pesquisas e os conhecimentos relacionados a esta nova doença também podem determinar facilmente o diagnóstico precoce da gravidade do vírus.

O surgimento da SARS-CoV-2 marcou o terceiro coronavírus como altamente patogênico em humanos no século XXI, após a síndrome respiratória aguda grave (SARS) em 2003 e a síndrome respiratória do Oriente Médio (MERS) em 2012 (DROSTEN et al., 2003; ZAKI et al., 2012). Estudos apontam que o SARS-CoV-2 está mais relacionado ao coronavírus de morcego semelhante ao SARS do que ao MERS (LU et al., 2020; WU et al., 2020). O tempo de sobrevivência dos pacientes que chegam a óbito pode ser entre uma e duas semanas após sua admissão numa UTI, ainda que alguns demorem mais tempo (YANG X. et al., 2020).

A pandemia da COVID-19 é um fator novo na sociedade moderna, porém, como ultimamente a tecnologia tem avançado de forma extremamente rápida, é plausível que a ocorrência da mesma pudesse ser prevista, ainda que empiricamente. Fatores como o aumento populacional e o trânsito de pessoas entre diversas partes do mundo favorecem a possibilidade de disseminação de doenças (QUAMMEN, 2020).

Com a própria evolução dos sistemas de dados e bancos de dados, principalmente durante o período da pandemia, onde informações são divulgadas diariamente sobre o número de casos e óbitos da COVID-19, tornou-se um passo natural o uso de técnicas e modelos de previsões, com o intuito de antever o resultado de uma dada região. Neste trabalho, além dos tradicionais métodos estatísticos, foi também explorada uma nova abordagem conceitual de validação

cruzada para base de treino e teste, além dos modelos de regressão e aprendizado de máquina.

Sendo assim, o objetivo deste trabalho consiste em realizar uma avaliação da validação cruzada para séries temporais no contexto de previsões dos óbitos da COVID-19, além de usar uma ampla gama de métodos de previsão para identificar os melhores modelos. É comum que os trabalhos utilizem uma única base de treino e uma única base de teste. Com a ajuda de algoritmos, a proposta é utilizar múltiplos subconjuntos de bases de treinamento e teste. Desta forma, a presente tese está estruturada em cinco capítulos, a saber: introdução, revisão teórica, metodologia, resultados e considerações finais.

O capítulo de revisão teórica trata, de forma geral, dos principais estudos e pesquisas realizadas sobre a COVID-19, iniciando com um breve histórico sobre modelos matemáticos em epidemias, trabalhos de previsão com modelos de aprendizado de máquina e séries temporais, além de apontar levemente o contexto da economia durante a pandemia.

A metodologia aborda as bases de dados escolhidas para o estudo, além de introduzir o conceito de séries temporais e seus elementos como tendência, ciclo, ruído, sazonalidade e estacionariedade. O capítulo trata também dos conceitos dos métodos estatísticos e de aprendizado de máquina, além das métricas de previsão para séries temporais.

Nos resultados foram expostas as projeções para os óbitos por COVID-19 no Brasil usando os métodos estatísticos, de regressão e de aprendizado de máquina. A avaliação e escolha dos melhores modelos foi feita utilizando as médias das métricas de previsão. Ademais, foram utilizadas algumas variáveis exógenas para dar maior exatidão aos resultados, tais como casos, índice de restrição, vacinação e taxa de reprodução do vírus.

Nas considerações finais foi feito um resumo dos resultados desta pesquisa, além de citar os trabalhos desenvolvidos e indicar sugestões para continuação do mesmo.

1 **REVISÃO TEÓRICA**

Nos últimos dois anos trabalhos de pesquisas foram elaborados relacionados a COVID-19. A crise do novo coronavírus criou uma necessidade sem precedentes de rastreamento de contatos em todos os países, exigindo que milhares de pessoas aprendam habilidades essenciais rapidamente. Além disso, medidas de contingenciamento foram tomadas, como distanciamento social, maior e mais cuidados com a higiene pessoal, fechamento de bares, restaurantes, lojas, shoppings, entre outras (SOHRABI, et al., 2020).

A doença atingiu os cinco continentes do planeta e se tornou uma pandemia com milhões de mortes (o primeiro óbito foi registrado em 9 de janeiro de 2020). Vacinas foram desenvolvidas como a AstraZeneca, CoronaVac, Pfizer, Janssen, Moderna e Sputnik V. Ainda assim, a busca por medicamentos eficazes continua, como também continuam os cuidados necessários para evitar o surgimento de variantes do vírus.

Dentro deste cenário, o uso da bioestatística tem sido fundamental para as análises e o tratamento dos dados relacionados a COVID-19. A mesma é essencial à epidemiologia e à medicina baseada em evidência. É utilizada a fim de entender sistemas variáveis, controle de processos, qualidade das informações e sumarização de dados para tomada de decisão (ASSIS; SOUZA; DIAS, 2019).

1.1 **Breve histórico sobre modelos matemáticos em epidemias**

Entre 1927 e 1933, William Kermack e Anderson McKendrick criaram um modelo em que se considera uma população (N) fixa com apenas três classes de indivíduos: Suscetíveis (S(t)), Infectados (I(t)) e Recuperados/Removidos (R(t)) - SIR.

O S(t) é usado para representar o número de indivíduos não infectados com a doença no momento t, ou aqueles suscetíveis à doença. O I(t) representa o número de indivíduos que tenham sido infectadas com a doença no momento t e que são capazes de transmitir a doença aos da categoria suscetível. O R(t) é o número de

infectados e, em seguida, removidos a partir da doença (devido à imunização ou à morte no momento t). Os que estão nesta categoria não são capazes de serem infectados novamente ou transmitir a infecção para outras pessoas (KERMACK; MCKENDRICK, 1927; KERMACK; MCKENDRICK, 1932; KERMACK; MCKENDRICK, 1933). Contudo, isso não vale para COVID-19. O modelo SIR, mesmo tendo sido desenvolvido há muito tempo atrás, é bastante usado até hoje em artigos científicos, pesquisas, trabalhos e estudos em geral sobre vírus e epidemias.

A taxa de reprodução de um vírus, denotado por R_0, é estipulada como o número de infecções secundárias causadas por um único indivíduo infectado incorporado numa população composta integralmente de indivíduos suscetíveis ao longo da infecção (KOEIJER; DIEKMANN; JONG, 2008). Se $R_0 > 1$, a infecção se espalhará exponencialmente. Se $R_0 < 1$, a infecção se espalhará lentamente e acabará desaparecendo. Nos casos em que $R_0 = 1$, a doença se torna endêmica, ou seja, ela permanece na população e se manifesta com frequência em determinadas regiões, geralmente provocada por circunstâncias ou causas locais (a população convive constantemente com a doença). Quanto mais alto o valor de R_0, mais rápido a epidemia irá progredir (ARONSON; BRASSEY; MAHTANI, 2020).

1.2 **A pandemia da COVID-19**

Em dezembro de 2019, um surto local de pneumonia de causa inicialmente desconhecida foi detectado em Wuhan (província de Hubei, China), e foi rapidamente identificado como sendo causado por um novo coronavírus humano, a saber, Coronavírus da Síndrome Respiratória Aguda Grave 2 (SARS-CoV-2). A principal causa da mortalidade deste vírus é a Síndrome do Desconforto Respiratório Agudo (SDRA). Desde então, o surto se espalhou para todas as províncias da China continental, bem como para outros países e regiões (DONG; DU; GARDNER, 2020).

A Organização Mundial da Saúde (OMS), em 11 de março de 2020, declarou o novo surto de coronavírus humano (COVID-19) como uma pandemia global. Em uma entrevista coletiva, o Diretor-Geral da OMS, Dr. Tedros Adhanom Ghebreyesus, disse que a OMS estaria “profundamente preocupada com os níveis alarmantes de disseminação e gravidade e com os níveis alarmantes de inação”, e pediu aos

países que tomem medidas para conter o vírus (SOHRABI, et al., 2020). Há uma hipótese de que o vírus poderia estar adormecido em todo o mundo antes de ser detectado pela primeira vez na China. Fongaro et al. (2021) citam que o RNA do SARS-CoV-2 foi encontrado dentro do esgoto humano no Brasil, no estado de Santa Catarina, em novembro de 2019.

A extensão da pandemia teve impacto significativo nos países. O bloqueio afetou aspectos cruciais da vida diária em todo o mundo, tais como: segurança alimentar, economia global, educação, turismo, hotelaria, violência/abuso doméstico, saúde mental e poluição do ar. O bloqueio global foi iniciado para conter a propagação do vírus e “achatar a curva” da pandemia (ONYEAKA et al., 2021).

A população passou a tomar algumas precauções no dia a dia, como evitar contato próximo com pessoas que estão doentes, assegurar medidas preventivas diárias e lavar as mãos frequentemente com água e sabão ou desinfetante que contenha pelo menos 70% de álcool. Evitar contato em seu rosto, nariz e olhos. Evitar aglomerações, especialmente em espaços mal ventilados. O risco de exposição a vírus respiratórios como o COVID-19 pode aumentar em ambientes lotados e fechados, com pouca circulação de ar, principalmente se houver pessoas doentes na multidão (CUCINOTTA; VANELLI, 2020).

Em 12 de março de 2020, a COVID-19 foi confirmada em 125.048 pessoas em todo o mundo, levando uma mortalidade de aproximadamente 3,7%. Em comparação, a influenza possui uma taxa de mortalidade de menos de 1% (MEHTA et al., 2020). Devido a esta emergência de saúde pública, foram desenvolvidos vários sistemas de coleta e divulgação das informações sobre o vírus. Um deles foi hospedado pelo Centro de Ciência e Engenharia de Sistemas da Universidade Johns Hopkins, Baltimore, Estados Unidos, para visualizar e rastrear casos relatados da COVID-19 em tempo real (DONG; DU; GARDNER, 2020). Outro exemplo foi o banco de dados de Oxford (*Oxford COVID-19 Government Response Tracker* - OxCGRT), que coleta sistematicamente informações sobre várias respostas de políticas que os governos adotaram (HALE, et al., 2020). O site *Our World in Data*, que tem como objetivo divulgar dados para avanço contra os maiores problemas do mundo, se alinhou a OMS e criou uma parte somente para dados sobre a COVID-19 (RITCHIE et al., 2020).

Com a possibilidade da vacinação, passou-se a promover campanhas públicas para a população aderir em massa, buscando assim o controle da doença.

Um questionário online, transversal e auto-administrado foi instrumentalizado para pesquisar participantes adultos sobre a aceitabilidade das vacinas. Aqueles que tomaram a vacina contra influenza sazonal foram mais propensos a aceitar as vacinas contra a COVID-19. Tiveram participantes que acreditavam que havia uma conspiração por trás da COVID-19 e também aqueles que não confiam em nenhuma fonte de informação (EL-ELIMAT et al., 2021; PILTCH-LOEB et al., 2021). A população americana também aceita uma vacina com base na eficácia da mesma, não sendo significativo a probabilidade de efeitos colaterais menores e as chances de uma reação adversa séria (KAPLAN; MILSTEIN, 2021).

Após o início da vacinação, tornou-se constante os estudos sobre efetividade das vacinas. Dagan et al. (2021) revelou que a efetividade estimada da Pfizer foi de 72% para os dias 14 a 20 após a primeira dose. Após sete ou mais dias da segunda dose, a eficácia chegou a 92% para desenvolvimento grave da doença (DAGAN et al., 2021). A efetividade da vacina Moderna contra casos severos, críticos ou fatais foi cerca de 81,6% após a primeira dose e cerca de 95,7% após a segunda dose (CHEMAITELLY et al., 2021). Com a Astrazeneca, a eficácia de duas doses é de 74,5% entre as pessoas com a variante Alfa e 67,0% para a variante Delta (LOPEZ BERNAL et al., 2021). Na imunização completa, a estimativa da eficácia das vacinas foi de 65,9% para a prevenção, 87,5% para a prevenção de hospitalização, 90,3% para a prevenção de UTI, e 86,3% para a prevenção de morte (JARA et al., 2021).

Estudos mostram que indicadores como renda, pobreza e desenvolvimento humano influenciam o acesso às vacinas. Países de extrema pobreza tiveram baixo acesso às vacinas. Já potências econômicas como Reino Unido, Estados Unidos, China e Israel tiveram prioridade (OLIVEIRA et al., 2021). A pandemia intensificou as diferenças globais econômicas existentes entre países ricos e pobres. Entre alguns fatores, destacam-se: a importância do fortalecimento de sistemas de saúde universais, da ciência, tecnologia, inovação e das bases econômicas, em países com diferentes graus de desenvolvimento (LIMA; GADELHA, 2021).

1.3 **Estudos sobre a COVID-19**

O surto da COVID-19 representou um desafio significativo para os pesquisadores, pois os dados disponíveis sobre a trajetória de crescimento inicial foram limitados (MATTA et al., 2021). Além disso, as características epidemiológicas do novo coronavírus ainda não foram totalmente elucidadas, mesmo com mais de dois anos de pandemia. As previsões do surto para diferentes países são úteis para a alocação eficaz de recursos de saúde, e pode atuar como um sistema de alerta precoce para os formuladores de políticas governamentais. Melhorar a vigilância epidemiológica pode ajudar os atores a tomar decisões em tempo hábil, permitindo o uso de estratégias de isolamento social mais eficazes e específicas com menor impacto econômico e social (MINISTÉRIO DA SAÚDE, 2002).

A disseminação da COVID-19 revelou muitos perigos e necessitou de políticas rígidas de contenção, portanto, a previsão de casos e óbitos são fundamentais. Na maioria dos países do mundo, o surto da doença foi grave e o número de casos confirmados da COVID-19 aumentou diariamente (MALEKI et al., 2020). No Japão, 98,8% das escolas de ensino fundamental e médio em todo o país fecharam a partir de 1 de março de 2020. O fechamento das escolas pareceu não reduzir a incidência de infecção por coronavírus. A eficácia da medida começou em 9 de março e os casos reais relatados foram maiores do que o previsto, e com um intervalo de confiança bastante amplo. As análises de sensibilidade usando datas diferentes também não demonstraram eficácia (MELIN et al., 2020). Entretanto, o fechamento de escolas foi apenas para aqueles com idade entre 6 e 18 anos, ou seja, indivíduos vulneráveis provavelmente não foram totalmente protegidos pela medida.

Os algoritmos de aprendizado de máquina surgiram como um paradigma popular nas recentes pesquisas científicas devido à sua flexibilidade para lidar com as especificidades dos dados (PENG; NAGATA, 2020). A versatilidade desta abordagem permite sua aplicação em diversos contextos, desde a previsão de variáveis financeiras até a análise de sentimento em textos e aplicações médicas. A pandemia aumentou a necessidade de decisões clínicas imediatas e o uso eficaz dos recursos de saúde.

O diagnóstico positivo da COVID-19 pode ser previsto com aprendizado de máquina, utilizando como preditores apenas resultados de exames de admissão em pronto-socorro (BATISTA et al., 2020). Durante a urgência global, cientistas, médicos e especialistas em saúde em todo o mundo continuam em busca de novas tecnologias para apoiar no combate à pandemia (MIRANDA et al., 2020). O aprendizado de máquina surgiu como uma ferramenta eficaz para prever o surto, uma alternativa aos modelos SIR e SEIR (ARDABILI et al., 2020). Dentro de um curto período de tempo desde o surto da COVID-19, técnicas avançadas de aprendizado de máquina foram usadas na classificação taxonômica de genomas, ensaio de detecção do vírus e previsão de sobrevivência em pacientes graves (ALIMADADI et al., 2020).

O método do aprendizado de máquina é significativo na área de triagem, predição, previsão, rastreamento de contato e desenvolvimento de medicamentos para epidemias (LALMUANAWMA; HUSSAIN; CHHAKCHHUAK, 2020). No entanto, a maioria dos modelos não são implantados o suficiente para mostrar sua operação no mundo real, mas ainda assim, estão à altura de combater a atual pandemia (ARDABILI et al., 2020). No Brasil, o Hospital Israelita Albert Einstein em São Paulo elaborou um estudo onde foram coletados dados de pacientes adultos dos quais 43% receberam diagnóstico positivo para COVID-19 nos testes de RT-PCR. Cinco algoritmos de aprendizado de máquina foram testados, e o melhor desempenho foi obtido pelas Máquinas de Vetores de Suporte (BATISTA et al., 2020).

Como os grandes surtos virais exigem uma elucidação precoce de classificação taxonômica e da origem da sequência genômica do vírus, Randhawa et al. (2020) elaboraram um estudo que pode auxiliar no planejamento estratégico, contenção e tratamento da COVID-19. O método proposto combina aprendizado de máquina supervisionado com processamento de sinal digital para análises de genoma, melhorado por uma abordagem de árvore de decisão. O modelo forneceu evidências para a classificação taxonômica do vírus, bem como evidências quantitativas que sustentaram a hipótese de origem em morcegos. O método atingiu uma classificação 100% precisa das sequências do vírus e descobriu as relações mais relevantes entre mais de 5.000 genomas virais em poucos minutos, usando apenas dados de sequência de DNA brutos, e sem qualquer conhecimento biológico especializado, treinamento, gene ou anotações do genoma. Isso sugere que esta

abordagem pode fornecer uma opção confiável em tempo real para classificação taxonômica.

Elaziz et al. (2020) propuseram um modelo para classificar as imagens de radiografia de tórax em duas classes: pacientes com ou sem COVID-19. Os recursos foram extraídos das imagens de raios-x de tórax. Na seleção de características, o classificador K-Vizinhos Mais Próximos (KNN) foi usado para decidir se uma dada imagem era de um paciente infectado com o vírus. A abordagem proposta alcançou alto desempenho, bem como baixo consumo de recursos, selecionando as características mais significativas. Já o trabalho de Gao et al. (2020) usou um modelo conjunto com Regressão Logística, Máquina de Vetores de Suporte, Árvores de Decisão e Redes Neurais. A construção deste modelo permitiu calcular a predição de risco de mortalidade da COVID-19 com até 20 dias de antecedência. As características que contribuíram para o resultado foram correlacionadas com idade avançada, sexo masculino e presença de múltiplas comorbidades.

O excesso de mortalidade é uma medida internacional apropriada que evita a subcontagem de mortes por muitos patógenos, incluindo a própria pandemia (BORREGO–MORELL; HUERTAS; TORRADO, 2021). Aproximadamente um milhão de mortes em excesso ocorreram em 2020 em países de alta renda, sendo as taxas de mortalidade maiores em homens do que em mulheres (ISLAM et al., 2021a). As vidas perdidas associadas à pandemia foram mais que cinco vezes superiores as associadas à epidemia de gripe sazonal em 2015 (ISLAM et al., 2021b). Empregando a análise de séries temporais contrafactuais, Modi et al. (2021) estima que, na Itália, país aonde ocorreu o primeiro pico da pandemia, o número de mortes por COVID-19 em 2020 até 9 de setembro foi de 59.000 a 62.000, valores maiores que os divulgados de forma oficial de 36.000.

Usualmente, modelos econométricos simples podem ser úteis para prever a propagação do vírus. A previsão do modelo ARIMA foi realizada nos dados epidemiológicos da Johns Hopkins para a tendência epidemiológica da prevalência e incidência do vírus. Os parâmetros foram estimados pelo gráfico da Função de Autocorrelação (FAC) e Função de Autocorrelação Parcial (FACP). O correlograma mostrou que tanto a prevalência como a incidência da COVID-19 não são influenciadas pela sazonalidade. O modelo ARIMA (1,0,4) foi selecionado para determinar a prevalência enquanto o ARIMA (1,0,3) foi usado para a incidência (BENVENUTO et al., 2020).

Em fevereiro de 2020, o vírus começou a se espalhar no Brasil, e em maio de 2020, São Luís, capital do estado do Maranhão, foi a primeira cidade junto com Fortaleza, capital do estado do Ceará, a entrar em lockdown (PARAGUASSU, 2020). Segundo dados do *Our World in Data* (2022), os picos de casos e óbitos mais altos no Brasil ocorreram em março e abril de 2021, provavelmente uma consequência direta da lenta implementação da vacinação, que começou em janeiro de 2021. Em abril de 2021 o percentual da população brasileira vacinada era de apenas 13,71%, enquanto na mesma data era de 25,15% para a Espanha (BORREGO–MORELL; HUERTAS; TORRADO, 2021).

Na Bahia, houve um aumento da mortalidade materna e sua relação temporal com a incidência da COVID-19 em 2020. Foi analisada a tendência temporal da razão de mortalidade materna por meio de Regressão Polinomial e previstos pelo modelo de Alisamento Exponencial de Holt-Winters Aditivo (CARVALHO-SAUER et al., 2020). Em Tocantins, a segunda onda apresentou as maiores taxas de incidência e letalidade. Contudo, ao final do período das séries temporais analisadas, a incidência e a letalidade apresentaram tendências estáveis, sugerindo um resultado positivo do programa de vacinação (CESAR et al., 2021), o que corrobora com os resultados deste trabalho. Já em Pernambuco, as tendências analisadas nas séries mostraram que a letalidade diminuiu na primeira onda e ficou estacionária na segunda (CAVALCANTI et al., 2022). Esses são alguns exemplos que mostram as diferenças regionais existentes dentro do país, onde se tem uma diferença social grande e medidas como as de contenção e vacinação também sofreram variações (FIOCRUZ, 2022).

1.4 **Lacuna existente e contribuição da tese na literatura**

Os trabalhos publicados utilizam apenas uma base de treinamento e uma base de teste em um período específico da série temporal (SALGOTRA; GANDOMI; GANDOMI, 2020; MELIN et al., 2020; KHAN; GUPTA, 2020). Entretanto, para uma melhor avaliação da capacidade preditiva dos modelos, é necessário utilizar muitos conjuntos de treinamento e teste. Assim, este trabalho tem como contribuição a

utilização de múltiplas bases de treinamento e teste para as séries temporais, que oferecem maior exatidão para previsão dos modelos.

Além disso, os outros trabalhos publicados para a previsão de séries temporais do COVID-19 usam um escopo muito pequeno de métodos preditivos (ARDABILI et al., 2020; MALKI et al., 2020; CHIMMULA; ZHANG, 2020; SALGOTRA; GANDOMI; GANDOMI, 2020; MELIN et al., 2020; KHAN; GUPTA, 2020; MALEKI et al., 2020; BRAGA et al., 2021; DIVINO et al., 2022; MASUM et al., 2022; MOHAN et al., 2022). Este trabalho tem também como diferencial utilizar um amplo escopo de métodos de previsão, possibilitando identificar os modelos que produzem melhores previsões dos óbitos por COVID-19 no Brasil, e que podem ser úteis em futuras epidemias/pandemias.

Ademais, este trabalho também se diferencia dos demais por utilizar, além das séries temporais de óbitos, variáveis exógenas tais como casos, índice de restrição, vacinação, taxa de reprodução do vírus, etc.

2 **METODOLOGIA**

O uso das previsões em séries temporais requer cuidado quando são implementadas. Neste trabalho, o uso da validação cruzada foi aperfeiçoado para as séries temporais para identificar os melhores modelos. O trabalho foi divido em duas partes: na primeira (Caso 1) foram usados dados sem cobertura vacinal e na segunda (Caso 2) foram usados dados com cobertura vacinal no Brasil.

No Caso 1 foi usado o banco de dados *Oxford COVID-19 Government Response Tracker* (OxCGRT) (HALE et al., 2021). Essa base coleta sistematicamente informações diárias sobre várias políticas públicas que os governos adotaram para responder à pandemia, como bloqueio, restrições de viagens, vacinação, taxa de reprodução do vírus, entre outras.

O período analisado no Caso 1 foi de 17 de março de 2020 a 20 de janeiro de 2021 (dados diários). Isso porque a primeira morte por COVID-19 no Brasil foi notificada em 17 de março de 2020, e a vacinação teve início em 17 de janeiro de 2021. Foram nove variáveis utilizadas (uma dependente e oito independentes):

- Variável dependente (Y): óbitos.
- Variáveis independentes: casos (X_1); índice de restrição (X_2); índice de resposta do governo (X_3); índice de saúde (X_4); índice econômico (X_5); lag 7, valores de Y em passos de tempo anteriores ($X_6 = Y(t-7)$); lag 8, valores de Y em passos de tempo anteriores ($X_7 = Y(t-8)$); e lag 9, valores de Y em passos de tempo anteriores ($X_8 = Y(t-9)$).

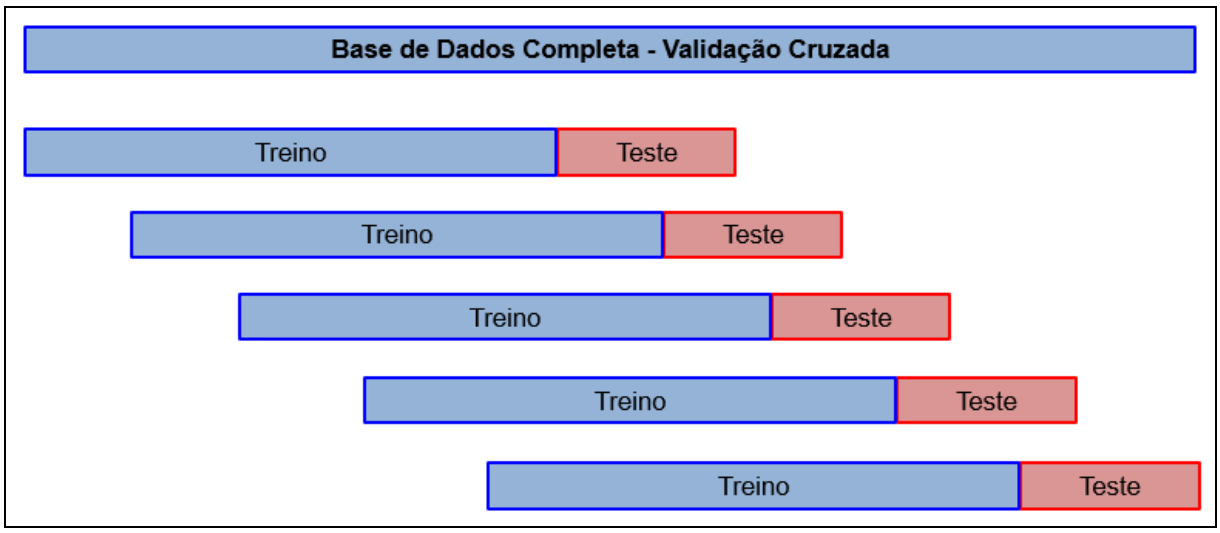
O índice de restrição é a restrição adotada pela região, como fechamento de aeroportos e outros locais. O índice do governo é a resposta do governo à pandemia de COVID-19. O índice de saúde é o investimento feito pelo governo nas políticas de saúde. O índice econômico é o apoio financeiro que o governo fez durante a pandemia, como ajuda financeira a pessoas que perderam o emprego ou não podem trabalhar. Todas as variáveis estão completamente descritas em um painel global de políticas pandêmicas (HALE et al., 2021).

Essas variáveis foram escolhidas por serem medidas utilizadas na Universidade de Oxford. O banco de dados de Oxford analisa países em todo o mundo quanto ao nível de restrição contra a COVID-19. As defasagens das variáveis foram selecionadas de acordo com o horizonte de sete dias. Além disso, em uma

aplicação de estudo de séries temporais, é interessante utilizar inicialmente o período para previsão de sete dias devido ao pequeno número de observações.

A validação cruzada utilizada é adequada para séries temporais, sendo usada a ordem dos valores como base fundamental para previsão. O objetivo é separar as séries temporais em intervalos de tempo fixos. No Caso 1, a base de treino foi fixada em 30 dias, e a base de teste fixada em sete dias a frente, com deslocamento de um dia (Figura 1).

Figura 1 - Validação cruzada para o Caso 1



Fonte: O autor, 2022.

No Caso 2 foi usada a base do *Our World in Data* (OWID) (HASELL et al., 2020; MATHIEU et al., 2021). Esta base de dados é uma publicação científica online, gratuita e acessível a todos. O OWID se concentra em problemas globais significativos como doenças, pobreza, desigualdade, mudança climática, violência, entre outros. É um projeto do *Global Change Data Lab*, uma instituição de caridade registrada na Inglaterra e no País de Gales. O *Global Change Data Lab* é uma organização sem fins lucrativos e uma instituição de caridade registrada em educação. A equipe de pesquisa OWID está sediada na Universidade de Oxford. O banco de dados da COVID-19 no OWID é atualizado sistematicamente com informações como óbitos, casos, taxa de reprodução do vírus, vacinação, bloqueio, entre outras. Essas informações também são divididas por data e país.

O período analisado no Caso 2 foi de 17 de janeiro de 2021 a 31 de julho de 2022. Foram usadas as médias móveis de 7 dias para os óbitos por COVID-19.

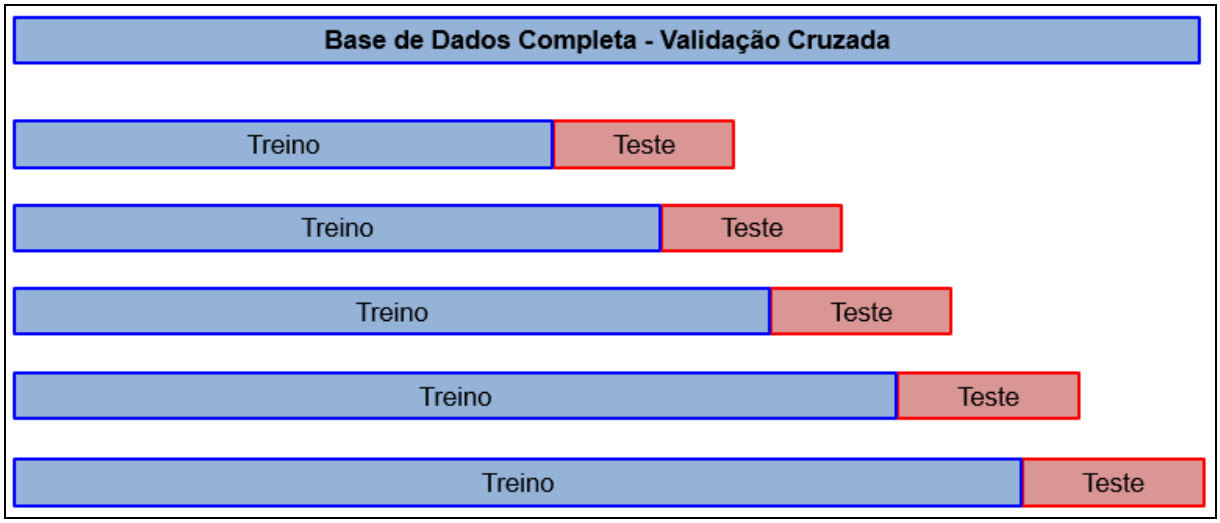
Sendo assim, a vacinação foi incluída nesta segunda parte do estudo. Foram sete variáveis (uma dependente e cinco independentes):

- Variável dependente (Y): óbitos.
- Variáveis independentes: casos (X_1); taxa de reprodução do vírus (R_0) (X_2); indivíduos parcialmente vacinados (uma dose da vacina) (X_3); indivíduos totalmente vacinados (duas doses ou dose única da vacina) (X_4); índice de restrição social (X_5); e lag 30, valores de Y em passos de tempo anteriores (X_6 = Y(t-30)).

Essas variáveis foram selecionadas por serem medidas que podem ser utilizadas posteriormente ao início da vacinação em períodos de pandemia/epidemia. A defasagem de 30 dias foi escolhida de acordo com os horizontes de previsão de sete, 15 e 30 dias.

A validação cruzada foi usada para séries temporais. O objetivo é separar as séries temporais em intervalos de tempo fixos, expandindo o conjunto de treino em cada iteração. No Caso 2, a base de treino foi fixada inicialmente em 30 valores, e a base de teste fixada em sete, 15 e 30 valores (Figura 2).

Figura 2 - Validação cruzada para o Caso 2



Fonte: O autor, 2022.

Dado o uso da validação cruzada, com múltiplos conjuntos de treinamento e teste para o Caso 1 e para o Caso 2, os resultados possuem maior exatidão em relação a outros estudos que usam somente uma única base de treino e uma única base de teste.

Todas as análises de dados, modelagem e programação foram realizadas utilizando o programa R (R CORE TEAM, 2019), com os pacotes: stats (R CORE TEAM, 2019), ggcorrplot (KASSAMBARA et al., 2019), forecast (HYNDMAN et al., 2020) e caret (KUHN et al., 2018).

2.1 **Séries temporais**

O objetivo principal da análise de séries temporais é a previsão. Esta metodologia permite prever valores futuros através dos valores presentes e passados (BOX et al., 2015). Portanto, os modelos são essenciais para fornecer o suporte necessário para a inferência estatística (GUJARATI; PORTER, 2011). Uma série temporal pode ter quatro componentes: tendência, ciclo, sazonalidade e resíduo. A tendência descreve o comportamento da variável ao longo do tempo. O ciclo é uma flutuação periódica em relação à tendência. A sazonalidade é uma mudança que ocorre em períodos específicos de uma série temporal e, diferentemente do ciclo, com recorrência de tamanho fixo. O resíduo é representado por flutuações aleatórias resultantes de fatos inesperados (BOX et al., 2015).

2.2 **Análise exploratória de dados**

A primeira análise de séries temporais é visual, através do gráfico de linhas, com a variável tempo representada no eixo horizontal e os valores da série temporal representados no eixo vertical. Algumas inferências e hipóteses podem ser sugeridas. O boxplot é uma maneira gráfica de exibir a distribuição de dados com base em cinco valores: mínimo, primeiro quartil, mediana, terceiro quartil e máximo. Valores discrepantes podem ser exibidos como pontos individuais. Essa técnica não faz suposições sobre a distribuição estatística envolvida nos dados. Os espaços entre as diferentes partes da caixa indicam o grau de dispersão, a assimetria nos dados e os outliers (ROSS, 2020). Os gráficos de dispersão são usados para

observar as relações entre duas variáveis. A correlação de Pearson mede o grau de associação linear entre as variáveis (CLINE, 2019).

Muitos métodos estatísticos supõem que os dados são de uma população com uma distribuição de probabilidade específica. A característica dessa distribuição pode ser um dos propósitos da análise. Existem testes estatísticos responsáveis por avaliar a distribuição teórica dos dados. Os seguintes testes foram usados neste trabalho: Shapiro-Wilk, Kolmogorov-Smirnov, Anderson-Darling, Cramer-von Mises, Lilliefors, Pearson e Shapiro-Francia para testar se os dados seguem a distribuição gaussiana. Nesses testes, a hipótese nula (H_0) é que os dados seguem uma distribuição normal. A hipótese alternativa (H_1) é que os dados não seguem uma distribuição normal (GHASEMI; ZAHEDIASL, 2012).

2.3 **Modelos ETS e Box-Jenkins**

O modelo ETS (*Error Trend Seasonal*) é uma classe "especial" de suavização exponencial. As previsões pontuais produzidas pelos modelos são idênticas se utilizarem os mesmos valores dos parâmetros de suavização (HYNDMAN; ATHANASOPOULOS, 2018). A caracterização do modelo seguindo a terminologia de Hyndman et al. (2002) e Hyndman et al. (2008) é feita usando uma sequência de caracteres de três caracteres. A primeira letra indica o tipo de erro (A ou M); a segunda letra indica o tipo de tendência (N, A ou M); e a terceira letra indica o tipo de sazonalidade (N, A ou M). Em todos os casos, N = nenhum, A = aditivo e M = multiplicativo.

A metodologia Box-Jenkins consiste em ajustar modelos de Médias Móveis Integradas Autoregressivas (ARIMA). A estratégia de construção do modelo é baseada em um ciclo iterativo. As etapas do ciclo interativo são especificação, identificação, estimativa e diagnóstico. Em geral, os modelos postulados são parcimoniosos, pois contêm um pequeno número de parâmetros, e as previsões obtidas são bastante precisas. Os modelos ARIMA assumem que os valores de uma série temporal possuem uma relação de dependência onde cada valor pode ser explicado por valores anteriores dos dados da série (GUJARATI; PORTER, 2011). O objetivo da metodologia Box-Jenkins é determinar os três componentes que

compõem a estrutura: p (parâmetros autorregressivos), d (processos de diferenciação) e q (parâmetros de média móvel), formando assim o ARIMA (p,d,q) (BOX et al., 2015).

Autocorrelação de grau k é a correlação da variável X(t) consigo mesma no último instante X(t-k), que é chamada de defasagem k. A Função de Autocorrelação (FAC) mede a dinâmica da correlação entre uma variável e suas defasagens. A Função de Autocorrelação Parcial (FACP) é uma medida da correlação entre observações de uma série temporal que são separadas por k unidades de tempo (X(t) e X(t-k)) (GUJARATI; PORTER, 2011; BOX et al., 2015). Ela é um dos coeficientes de um modelo de defasagem distribuída.

2.4 **Modelos de regressão**

Os modelos de regressão procuram identificar alguma relação entre variáveis dependentes e independentes. Essa relação pode ser linear ou não linear. Em modelos de regressão com dados transversais, a ordem das observações é irrelevante para a análise. Em séries temporais, a ordem dos dados é fundamental. Uma característica significativa deste tipo de dados é que as observações vizinhas são geralmente dependentes ao longo do tempo, por isso é interessante analisar e modelar essa dependência (GUJARATI; PORTER, 2011).

Normalmente, os conjuntos de dados possuem muitas variáveis independentes, por isso é necessário saber quais são relevantes para explicar a variável dependente. Nesses casos, são necessários mecanismos para escolher o melhor subconjunto de variáveis independentes para explicar a variável dependente. Para isso, os Métodos de Regularização são recomendados. Esses métodos incorporam uma restrição ao modelo, limitando os coeficientes do modelo e, portanto, selecionando as variáveis independentes mais importantes (JAMES et al., 2013).

Os modelos de regressão dinâmica também são chamados de modelos ARIMA com variáveis exógenas (ARIMAX). Nos modelos de regressão linear, assume-se que o ruído tem média zero, variância constante, distribuição normal e independência, não havendo, portanto, correlação serial (GUJARATI; PORTER,

2011). O modelo ARIMAX combina a dinâmica de séries temporais e o efeito de variáveis explicativas. A variável dependente é explicada por seus valores defasados e valores atuais e passados de variáveis exógenas.

Além do ARIMAX, foram utilizados neste trabalho os seguintes modelos de regressão linear: Múltiplo (MLR), Stepwise, Stepwise com menor valor de Akaike (Stepwise AIC), Lasso, Ridge, Rede Elástica, Boosted, Boosted Tree e Robust. Os modelos não lineares foram: Cubist, Regressão Adaptativa Multivariada Splines (MARS) e MARS com poda de validação cruzada (MARS gCV).

2.5 **Aprendizado de máquina**

Os modelos de aprendizado de máquina exploram o estudo e a construção de algoritmos que podem aprender com os dados e fazer previsões (MUELLER; MASSARON, 2019). Os estudos com Máquinas de Vetores de Suporte (SVM) começaram a ser desenvolvidos na década de 60, na Rússia, por Vapnik, Lemer e Chervonenkis. No entanto, pode-se dizer que o SVM teve seu ponto de partida junto com o desenvolvimento da teoria da aprendizagem estatística por Vapnik em 1979. A forma atual foi desenvolvida por Vapnik no final da década de 1990 e visava encontrar um hiperplano que maximizasse a margem entre classes (NAGUIB & DARWISH, 2012). A Regressão de Vetores de Suporte (SVR) mantém as mesmas características do SVM.

As Florestas Aleatórias (RF) são formadas por diversas árvores de decisão. Cada uma delas fornece uma estimativa. A classificação final é dada pelo resultado mais frequente em todas as árvores (HASTIE et al., 2009; JAMES et al., 2013; MUELLER; MASSARON, 2019).

As Redes Neurais Artificiais (RNA) são técnicas computacionais que apresentam um modelo matemático inspirado no cérebro humano (MUELLER; MASSARON, 2019). A propriedade crucial das RNA é sua capacidade de aprender com seu ambiente e melhorar seu desempenho. Isso é feito através de um processo iterativo de ajuste dos pesos da rede (HASTIE et al., 2009). A técnica de Redes Neurais Autoregressivas combina o modelo estatístico autoregressivo e as redes neurais, resultando em um modelo AR-NN(p). No modelo AR-NNX, as variáveis

exógenas são incluídas, fornecendo novos dados para melhorar o desempenho da previsão. Além dos valores defasados da variável dependente, podem ser adicionadas variáveis independentes que também serão utilizadas (HADDOUN, 2008).

O método Boosting pertence à categoria de aprendizado de máquina chamada *ensemble* (conjunto). As técnicas *ensemble* envolvem grupos de modelos preditivos para obter melhor exatidão e estabilidade do modelo. Boosting refere-se a uma família de algoritmos que convertem aprendizado fraco em aprendizado forte. A previsão de cada aprendizado é combinada para convertê-lo em um aprendizado forte (HASTIE et al., 2009; JAMES et al., 2013). O algoritmo eXtreme Gradient Boosting (XGBoost) combina os modelos Boosting e de árvores de decisão.

2.6 **Métricas de previsão**

No Caso 1, para realizar a avaliação dos modelos, foi implementado o método de reamostragem Jackknife, no qual foi utilizada toda a base de dados. A estratégia remove uma amostra do conjunto total observado, recalculando o estimador a partir dos valores restantes. A utilização desta técnica promove a redução das incertezas, tendo assim uma estimativa precisa para avaliar a capacidade preditiva dos modelos (SHAO, 2012). No Caso 2, a técnica usada foi acrescentar um valor a cada momento t+1 na série na base de treino, buscando assim aumentar o tamanho amostral ao longo do tempo.

As médias das métricas de previsão são calculadas para avaliar os modelos. Neste trabalho, as seguintes métricas foram usadas: Coeficiente de Determinação (R^2), Erro Médio Absoluto (MAE), Erro Percentual Absoluto Médio (MAPE) e Erro Quadrado Médio (RMSE) (MELLO, 2021).

3 **RESULTADOS**

Os resultados encontrados foram divididos em dois estudos distintos. O Caso 1 promove a previsão os óbitos antes do início do período de vacinação (entre 17/03/2020 e 20/01/2021). No Caso 1 foram analisados os óbitos diários e acumulados. O Caso 2 inclui a cobertura vacinal (entre 17/01/2021 e 31/07/2022) e a taxa de reprodução da COVID-19, o R_0. No Caso 2 foram analisadas as médias móveis de sete dias de óbitos.

3.1 **Caso 1**

Entre 17 de março de 2020 e 20 de janeiro de 2021, a média de óbitos por COVID-19 foi de 687 ± 391. A média de casos foi de 27.865 ± 18.461.O primeiro óbito ocorreu quando houve 51 casos notificados (Quadro 1).

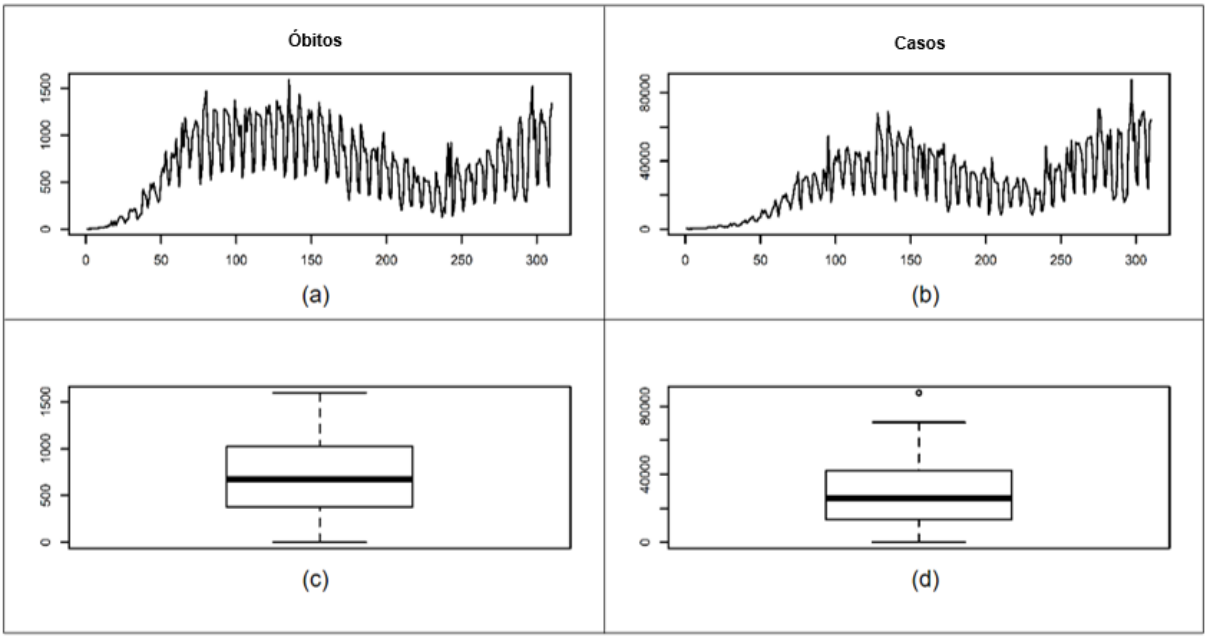
Quadro 1 - Estatísticas descritivas de óbitos e casos diários

Estatística	**Óbitos**	**Casos**
Média	687	27.865
Desvio padrão	391	18.461
Valor Mínimo	1	51
Q_1	376	13.381
Mediana	676	26.017
Q_3	1.018	42.144
Valor Máximo	1.595	87.843

Fonte: O autor, 2022.

Ambas as séries temporais mostram ciclo, com tendências de crescimento e redução. O boxplot não revela nenhum valor discrepante para os óbitos e um valor discrepante para os casos (Figura 3). Todos os p-valores referentes aos testes indicados no Quadro 2 ficaram abaixo de 0,05, informando que os dados não são normalmente distribuídos.

Figura 3 - Séries temporais e boxplots de óbitos e casos



Legenda: (a) e (b) - séries temporais; (c) e (d) boxplots.

Fonte: O autor, 2022.

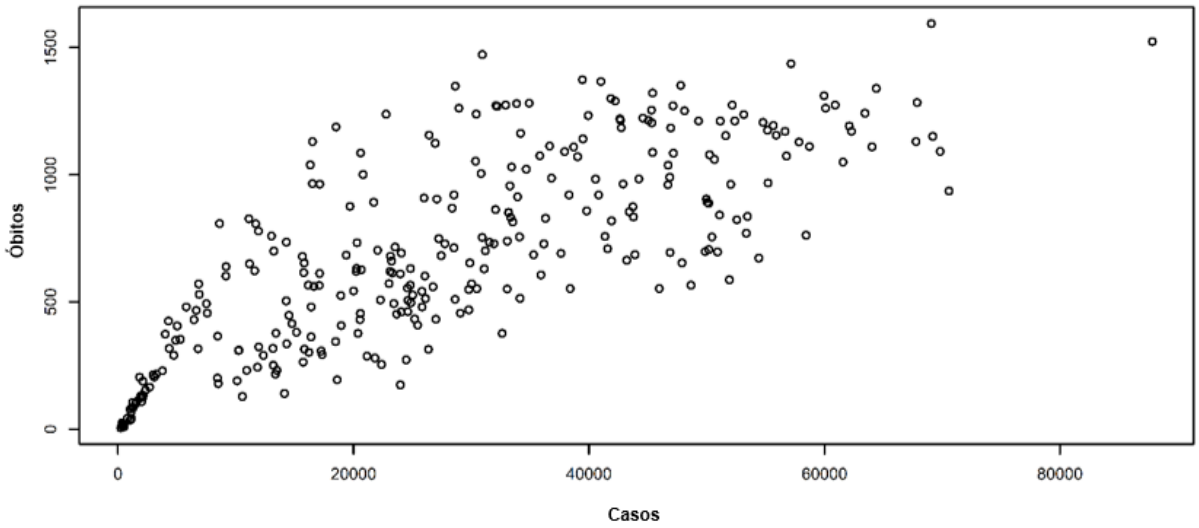
Quadro 2 - p-valor dos testes.

Testes	Óbitos	Casos
Shapiro-Wilk	$6{,}0 \times 10^{-6}$	$2{,}0 \times 10^{-6}$
Kolmogorov-Smirnov	$2{,}2 \times 10^{-16}$	$2{,}2 \times 10^{-16}$
Anderson Darling	$2{,}3 \times 10^{-5}$	$2{,}3 \times 10^{-5}$
Cramer von Mises	0,0009	0,0007
Lilliefors	0,0022	0,0024
Pearson	0,0014	$1{,}8 \times 10^{-7}$
Shapiro-Francês	$4{,}5 \times 10^{-5}$	$1{,}2 \times 10^{-5}$

Fonte: O autor, 2022.

O gráfico de dispersão revela que óbitos e casos têm correlação positiva - quanto maior o número de casos, maior o número de óbitos (Figura 4). Óbitos e casos apresentaram coeficiente de correlação igual a 0,78.

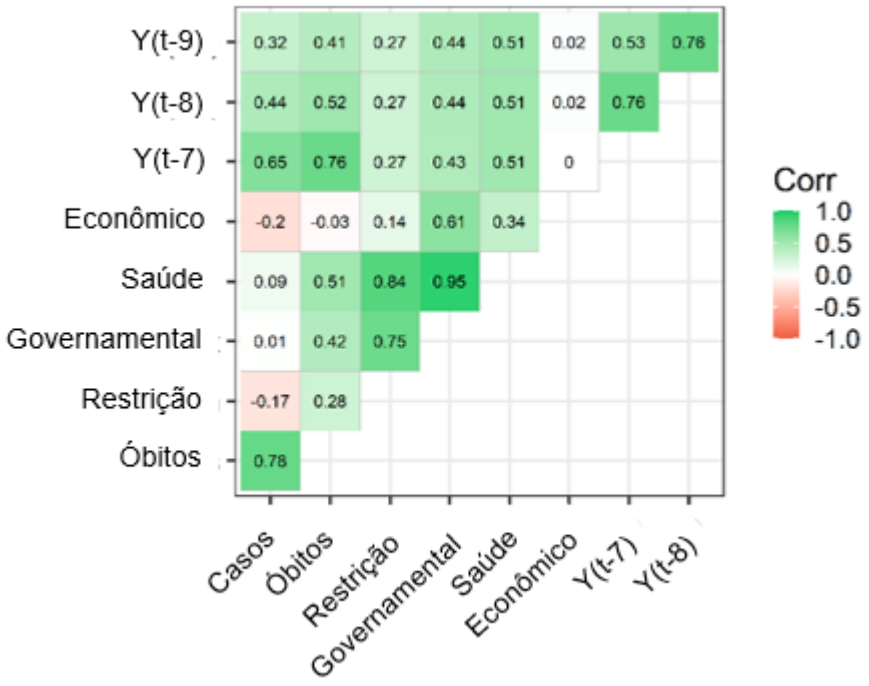
Figura 4 - Gráfico de dispersão entre casos/mortes



Fonte: O autor, 2022.

O correlograma revela a correlação entre todas as variáveis consideradas no Caso 1. Apenas o índice de saúde e o índice governamental possuem coeficiente de correlação superior a 0,90 (Figura 5).

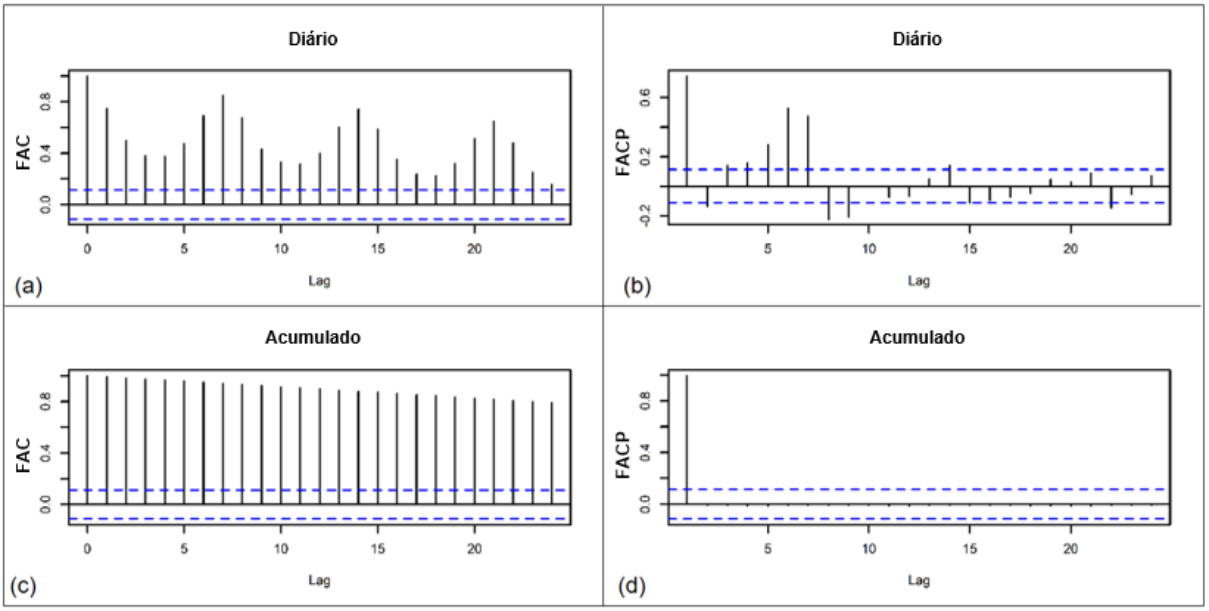
Figura 5 - Correlograma do Caso 1



Fonte: O autor, 2022.

A FAC e a FACP indicam a necessidade de realizar pelo menos uma diferenciação para o modelo ARIMA para a série temporal de óbitos diários e acumulados (Figura 6).

Figura 6 - FAC e FACP de óbitos e casos



Legenda: (a) FAC de óbitos diários; (b) FACP de óbitos diários; (c) FAC de óbitos acumulados; (d) FACP de óbitos acumulados.

Fonte: O autor, 2022.

Após 266 subconjuntos de treinamento e teste para cada modelo, 0066oram calculadas as médias das métricas de previsão para as bases de teste. Os Quadros 3 e 4 mostram as avaliações para os modelos de regressão, Floresta Aleatória, Máquinas de Vetores de regressão, Redes Neurais Artificiais e XGBoost. O R^2 revela que os óbitos acumulados produziram melhores estimativas do que os diários.

Quadro 3 - Resultados dos modelos de regressão, Florestas Aleatórias, Máquinas de Vetores de Suporte, Redes Neurais e XGBoost para previsão de óbitos diários

Modelos - Óbitos Diários	**R^2**	**RMSE**	**MAE**
Regressão Linear Múltipla	0,708	1.034	828
Stepwise	0,750	154	127
Stepwise AIC	0,738	330	263
Lasso	0,790	149	126

Ridge	0,772	136*	113*
Rede Elástica	0,752	149	124
Boosted	0,715	186	157
Boosted Tree	0,689	190	160
Robusta	0,796*	145	120
Cubist	0,745	152	123
Regressão Adaptativa Multivariada Splines	0,731	151	121
Regressão Adaptativa Multivariada Splines gCV	0,724	154	123
Florestas Aleatórias	0,669	158	127
Máquinas de Vetores de Regressão	0,287	306	260
Redes Neurais Artificiais	0,690	804	765
Redes Neurais Artificiais Média	0,580	804	765
XGBoost	0,640	170	137

* Nota: Melhores valores.

Fonte: O autor, 2022.

Quadro 4 - Resultados dos modelos de regressão, Florestas Aleatórias, Máquinas de Vetores de Suporte, Redes Neurais e XGBoost para previsão de óbitos acumulados

Modelos - Óbitos Acumulados	**R^2**	**RMSE**	**MAE**
Regressão Linear Múltipla	0,895	6.464	5.087
Stepwise	0,986	1.230	1.101
Stepwise AIC	0,970	1.918	1.615
Lasso	0,989	717	652
Ridge	0,987	1.104	1.037
Rede Elástica	0,983	898	834
Boosted	0,873	9.811	9.741
Boosted Tree	0,550	6.173	5.968
Robusta	0,986	790	717
Cubist	0,993	468*	409*
Regressão Adaptativa Multivariada Splines	0,995*	524	469
Regressão Adaptativa Multivariada Splines gCV	0,995*	521	466
Florestas Aleatórias	0,575	521	466
Máquinas de Vetores de Regressão	0,735	13.824	13.676
Redes Neurais Artificiais	0,424	114.546	114.522
Redes Neurais Artificiais Média	0,541	114.546	114.522
XGBoost	0,614	3.463	3.082

* Nota: Melhores valores.

Fonte: O autor, 2022.

As Regressões Robust (R^2) e Ridge (RMSE e MAE) tiveram o melhor ajuste para prever os óbitos diários (Quadro 3). A Regressão Robust tem uma função que minimiza os erros absolutos, e a Regressão Ridge tem uma penalidade que reduz a complexidade de um modelo. As Regressões MARS (R^2) e Cubist (RMSE e MAE) tiveram o melhor ajuste para prever os óbitos acumulados (Quadro 4). Esses métodos usam uma abordagem baseada em regras, como as árvores de decisão.

Os Quadros 5 e 6 apresentam as avaliações para os modelos ARIMA, ETS, ARIMAX, AR-NN e AR-NNX. Os modelos ARIMA e ARIMAX foram ajustados com diferentes componentes p e q e variaram de 1 até 5 (ou seja, em cada uma das 266 bases de treinamento, os valores p e q podem ser alterados). As diferenças (d) foram de até 4 para ARIMA e até 5 para ARIMAX. A métrica MAPE revelou que os óbitos acumulados produziram melhores estimativas. As diferenças foram calculadas com o uso da função “ndiffs” no R.

Quadro 5 - Resultados dos modelos ARIMA, ETS, ARIMAX, AR-NN e AR-NNX para previsão de óbitos diários

Modelos - Óbitos Diários	**MAPE (%)**	**RMSE**	**MAE**
ARIMA(p,1,q)	30,9*	218*	182*
ARIMA(p,2,q)	97,4	686	621
ARIMA(p,3,q)	322,5	2.442	2.061
ARIMA(p,4,q)	493,2	4.034	3.169
ETS	42,1	293	247
ARIMAX(p,0,q)	6.266,8	64.338	27.327
ARIMAX(p,1,q)	234,7	1.976	855
ARIMAX(p,2,q)	70,6	459	268
ARIMAX(p,3,q)	687,7	3.709	2.219
ARIMAX(p,4,q)	3.149,4	18.093	10.388
ARIMAX(p,5,q)	104.905,2	480.294	240.537
AR-NN(p)	37,7	286	230
AR-NNX	37,6	286	230

* Nota: Melhores valores.

Fonte: O autor, 2022.

Quadro 6 - Resultados dos modelos de regressão, Florestas Aleatórias, Máquinas de Vetores de Suporte, Redes Neurais e XGBoost para previsão de óbitos acumulados

Modelos - Óbitos Acumulados	MAPE (%)	RMSE	MAE
ARIMA(p,1,q)	1,1	619*	542*
ARIMA(p,2,q)	1,0*	551	486
ARIMA(p,3,q)	3,2	2.733	2.272
ARIMA(p,4,q)	5,9	6.245	4.809
ETS	0,9	630	558
ARIMAX(p,0,q)	23,2	1.788	1.440
ARIMAX(p,1,q)	22,6	1.615	1.247
ARIMAX(p,2,q)	21,8	1.559	1.202
ARIMAX(p,3,q)	22,6	1.637	1.255
ARIMAX(p,4,q)	14,6	1.041	806
ARIMAX(p,5,q)	32,2	2.479	1.807
AR-NN(p)	2,8	1.870	1.588
AR-NNX	2,8	1.872	1.589

* Nota: Melhores valores.

Fonte: O autor, 2022.

O ARIMA(p,1,q) (MAPE, RMSE e MAE) teve o melhor ajuste para prever os óbitos diários (Quadro 5). A modelagem ARIMA contém um pequeno número de parâmetros, e as previsões são bastante precisas. Neste caso, foi necessária uma diferenciação para obter o melhor modelo.

Para os óbitos diários, considerando a métrica MAPE, o modelo ETS obteve o melhor resultado e considerando as métricas RMSE e MAE o modelo ARIMA(p,2,q) (RMSE e MAE) obteve o melhor resultado. (Quadro 6). A modelagem de ETS dá maior peso às observações passadas em relação às recentes. Este modelo também considera elementos como tendência e sazonalidade. No ARIMA(p,2,q), foram necessárias duas diferenças para se obter o melhor modelo.

Na elaboração das previsões, foi observado que as variáveis que mais influenciaram no ajuste foram casos e índice de restrição, sendo casos a dominante. Esta observação foi feita testando os melhores modelos com apenas uma variável exógena, que foi sendo trocada a cada teste e sendo avaliadas conforme as métricas de previsão.

3.2 **Caso 2**

A análise descritiva das variáveis do Caso 2 é apresentada no Quadro 7. Entre 17 de janeiro de 2021 e 31 de julho de 2022, a média de óbitos por COVID-19 foi de 841 ± 799. A média de casos foi de 45.205 ± 34.945. A primeira dose da vacina ocorreu quando havia 209.993 óbitos notificados, a taxa de reprodução do vírus era de 1,06 e a média móvel de sete dias de óbitos era de 962. Todos os p-valores foram inferiores a 0,05, indicando que os dados não são normalmente distribuídos (Shapiro-Wilk).

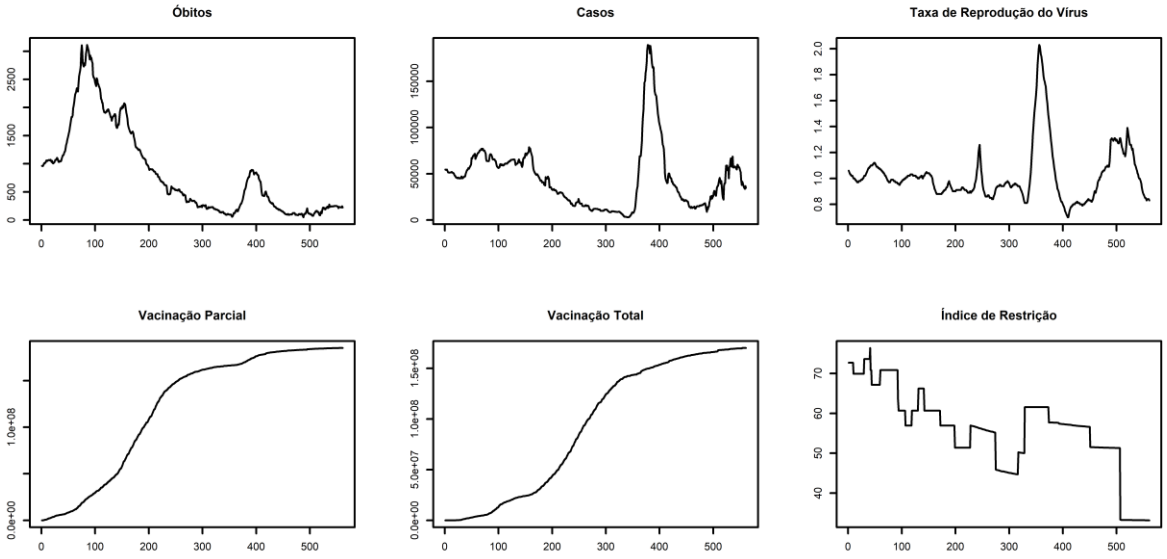
Quadro 7 - Estatísticas para média móvel de 7 dias

Estatísticas	**Casos**	**Óbitos**	**R_0**	**Vacinação Parcial**	**Vacinação Total**	**Índice de Restrição**
Média	45.205	841	1,02	121.296.430	92.768.031	55,87
Desvio-padrão	34.945	799	0,22	66.825.044	65.393.891	10,30
Valor Mínimo	2.984	48	0,70	112	0	33,20
Q_1	17.174	210	0,90	49.685.501	23.126.008	51,39
Mediana	42.284	525	0,98	157.986.323	109.945.760	56,94
Q_3	60.913	1.225	1,06	179.754.579	157.662.860	61,57
Valor Máximo	189.227	3.112	2,03	185.208.286	170.168.354	76,39
p-valor	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001	< 0,001

Fonte: O autor, 2022.

As séries temporais de óbitos, casos e índice de restrição social mostram tendências de redução. Ambas as séries temporais de vacinação revelam tendências de crescimento. A série temporal da taxa de reprodução do vírus apresenta ciclo, com tendências de crescimento e redução (Figura 7).

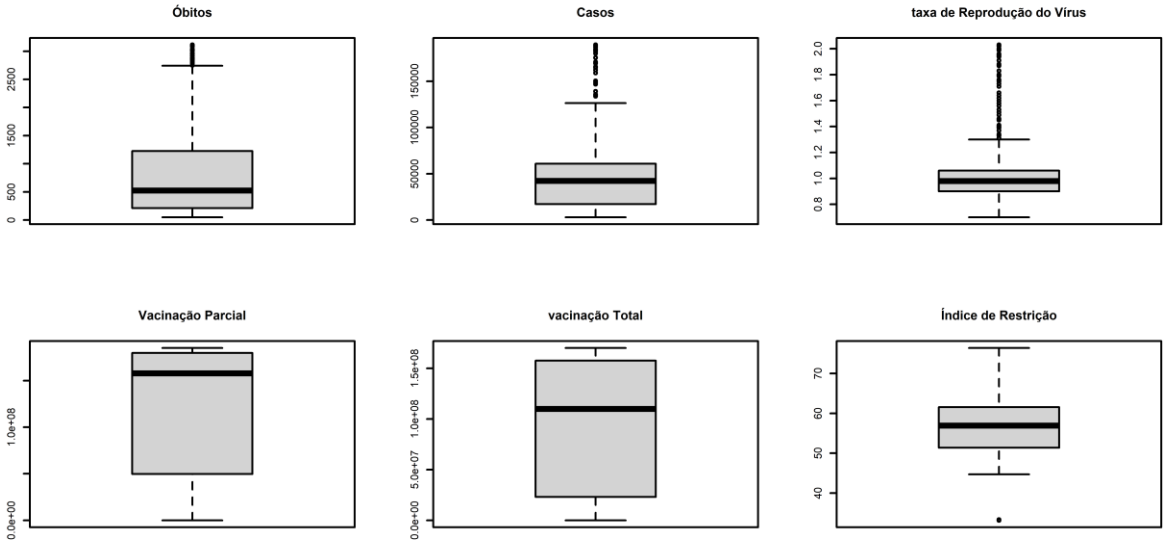
Figura 7 - Séries temporais de óbitos, casos, taxa de reprodução do vírus, vacinação parcial, vacinação total e índice de restrição



Fonte: O autor, 2022.

Os boxplots não revelam valores discrepantes para vacinação (parcial e total). Índice de restrição social apresenta um valor discrepante. Já óbitos, casos e taxa de reprodução do vírus apresentam vários valores discrepantes (Figura 8).

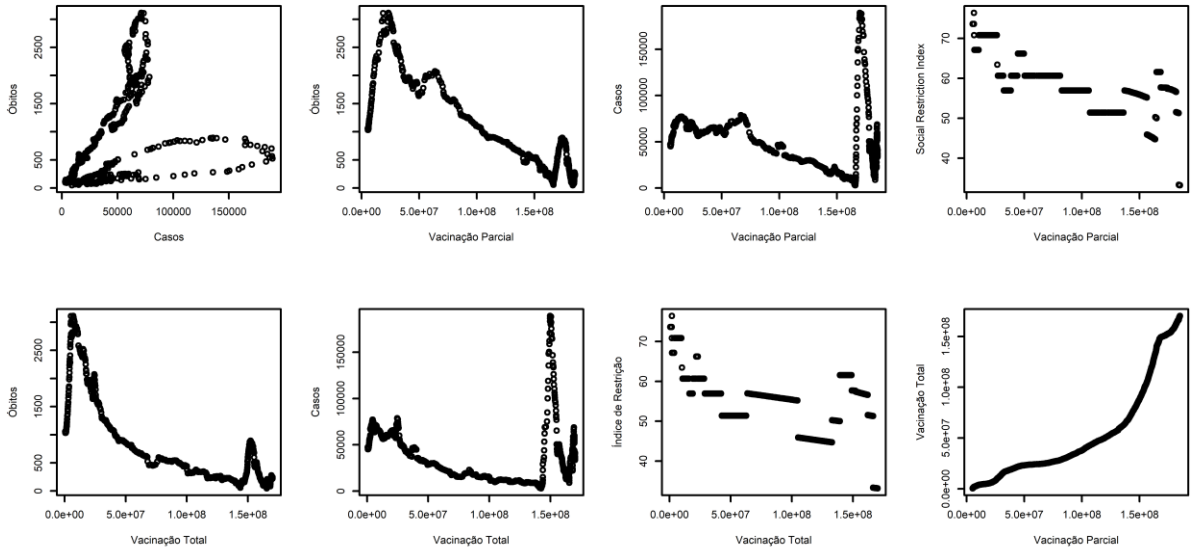
Figura 8 - Boxplots de óbitos, casos, taxa de reprodução do vírus, vacinação parcial, vacinação total e índice de restrição



Fonte: O autor, 2022.

O gráfico de dispersão revela que mortes e casos têm uma correlação positiva - quanto maior o número de casos, maior o número de óbitos. A vacinação (parcial e completa) correlaciona-se negativamente com óbitos, casos e índice de restrição social - quanto maior a quantidade de indivíduos vacinados, menor o número de óbitos, casos e índice de restrição social. A vacinação parcial é correlacionada positivamente com a vacinação total (Figura 9).

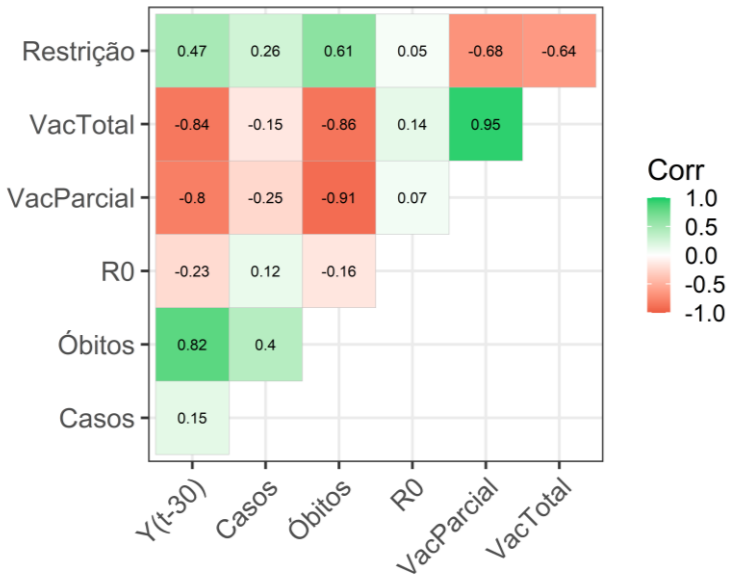
Figura 9 - Diagramas de dispersão



Fonte: O autor, 2022.

O correlograma da Figura 10 mostra a correlação entre todas as variáveis consideradas no Caso 2. Óbitos e casos apresentaram coeficiente de correlação igual a 0,40, valor bem menor do que visto no Caso 1. Nesse sentido, podemos inferir que conforme a vacinação avança, casos e óbitos passam a ser menos correlacionados. Óbitos e vacinação (parcial e completa) apresentaram coeficiente de correlação igual a -0,91 e -0,86, respectivamente. As vacinações parciais e completas têm a maior correlação (0,95).

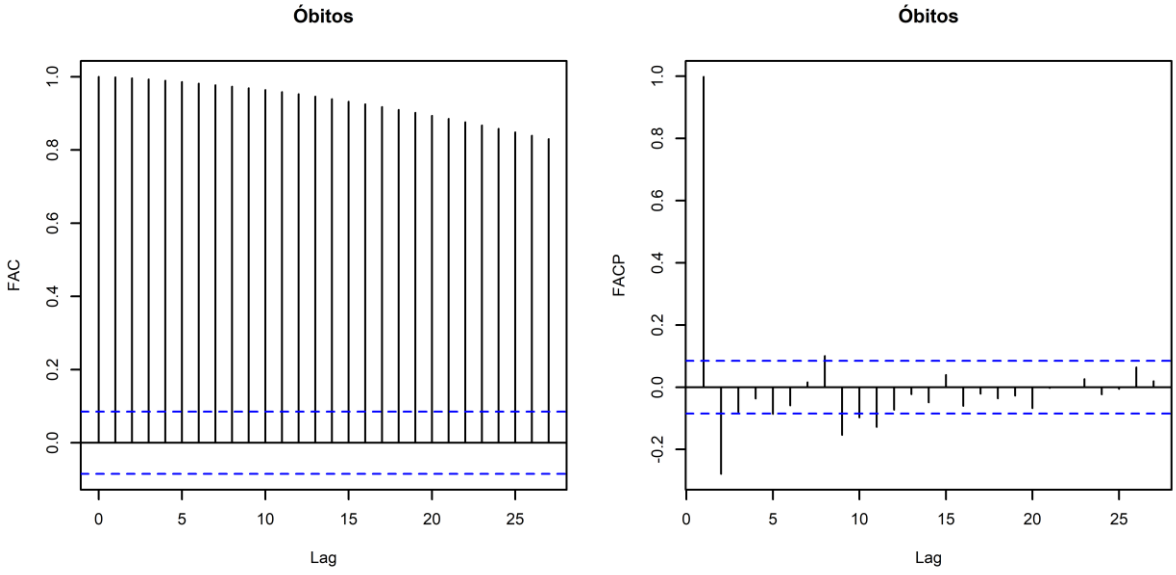
Figura 10 - Correlograma do Caso 2



Fonte: O autor, 2022.

A FAC tem decaimento exponencial lento e a FACP é truncada em 1. Deste modo, a FAC e a FACP indicam a necessidade de realizar pelo menos uma diferenciação para a modelagem ARIMA (Figura 11).

Figura 11 - FAC e FACP de óbitos para o Caso 2



Fonte: O autor, 2022.

O Quadro 8 apresenta os resultados para os modelos clássicos. Os modelos ARIMA e ARIMAX foram gerados com diferentes componentes p e q. As diferenças (d) foram de até 2 para ARIMA e ARIMAX, considerando a FAC e a FACP. As diferenças foram calculadas com o uso da função “ndiffs” no R. Observando o MAPE e considerando o horizonte de previsão de sete e 15 dias, o ARIMAX(p,2,q) teve o melhor ajuste para prever mortes por COVID-19. No ARIMAX(p,2,q), foram necessárias duas diferenças para obter o melhor modelo. Considerando uma previsão de horizonte de 30 dias, o ARIMA(p,1,q) teve o melhor ajuste. Neste caso, foi necessária uma diferenciação para obter o melhor modelo ARIMA. A modelagem ARIMA contém um pequeno número de parâmetros e as previsões são bastante exatas. Observando o RMSE e o MAE, o ARIMA(p,1,q) também teve o melhor resultado.

Quadro 8 - Resultados dos modelos ARIMA, ETS, ARIMAX, AR-NN e AR-NNX para previsão de óbitos em um horizonte de 7, 15 e 30 dias

Modelos	**MAPE**			**RMSE**			**MAE**		
	h = 7	**h = 15**	**h = 30**	**h = 7**	**h = 15**	**h = 30**	**h = 7**	**h = 15**	**h = 30**
ARIMA(p,1,q)	12,7	19,3	31,9*	74*	124*	219*	64*	107*	188*
ARIMA(p,2,q)	13,0	20,6	39,0	75	125	241	66	108	204
ETS	12,8	19,9	35,0	76	130	321	66	112	242
ARIMAX(p,1,q)	14,4	50,3	583,6	432	1.975	57.474	344	1.400	16.628
ARIMAX(p,2,q)	8,0*	17,6*	213,3	232	620	18.301	190	477	6.257
AR-NN	14,4	22,8	38,8	86	149	250	75	128	219
AR-NNX	14,5	22,8	38,7	86	148	250	75	128	218

* Nota: Melhores valores.

Fonte: O autor, 2022.

O Quadro 9 apresenta os resultados para os modelos de regressão e aprendizado de máquina. Observando o R^2 o Boosted Linear teve o melhor ajuste. Os modelos Boosted são métodos de aprendizagem em conjunto e realizam previsões através da combinação dos resultados de árvores individuais. Esses modelos constroem árvores uma de cada vez, onde cada nova árvore ajuda a corrigir erros cometidos por árvores previamente treinadas. A cada árvore adicionada, o modelo torna-se ainda mais expressivo. Observando o RMSE e o MAE

a regressão Cubist teve o melhor ajuste, método que utiliza um procedimento baseado em regras (árvores de decisão).

Quadro 9 - Resultados dos modelos de regressão e aprendizado de máquina para previsão de óbitos em um horizonte de 7, 15 e 30 dias.

Modelos	**R^2**			**RMSE**			**MAE**		
	h = 7	**h = 15**	**h = 30**	**h = 7**	**h = 15**	**h = 30**	**h = 7**	**h = 15**	**h = 30**
RLM	0,53	0,52	0,49	294	391	564	279	357	493
Stepwise	0,59	0,55	0,56	284	388	584	271	352	500
Stepwise AIC	0,54	0,52	0,52	295	422	683	279	381	590
Lasso	0,55	0,53	0,54	283	377	565	271	344	489
Ridge	0,55	0,54	0,53	295	420	609	279	380	542
Rede Elástica	0,55	0,54	0,53	205	262	379	197	242	335
Boosted Linear	0,61*	0,58*	0,59*	267	314	389	262	300	361
Boosted Tree	0,49	0,41	0,37	161	219	300	150	199	269
Boosted Gam	0,61*	0,57	0,54	157	211	319	146	190	274
Robusta	0,56	0,55	0,53	267	393	656	252	353	569
Cubist	0,59	0,56	0,49	91*	142*	223*	81*	124*	192*
MARS	0,60	0,55	0,58	128	213	335	117	187	310
MARS gCV	0,56	0,54	0,53	131	220	426	119	193	361
Florestas Aleatórias	0,45	0,41	0,35	112	171	251	101	152	223
SVR Radial	0,56	0,51	0,47	268	357	482	251	330	443
SVR Polinomial	0,58	0,56	0,54	142	223	307	127	197	281
SVR Linear	0,56	0,55	0,53	252	366	606	238	330	520
RNA	0,55	0,31	0,36	792	786	770	791	782	760
RNA Média	0,48	0,41	0,38	792	786	770	791	782	760
XGBoost	0,41	0,33	0,29	115	192	295	101	166	256

* Nota: Melhores valores.

Fonte: O autor, 2022.

Na elaboração das previsões, foi observado que as variáveis que mais influenciaram no ajuste foram vacinação parcial e vacinação total. Esta observação foi feita testando os melhores modelos com apenas uma variável exógena, que foi sendo trocada a cada teste e sendo avaliadas conforme as métricas de previsão.

CONSIDERAÇÕES FINAIS

A COVID-19 ocasionou os óbitos de milhões de pessoas desde o final de 2019. Com a evolução dos sistemas de dados e o alto contágio, informações passaram a ser publicadas diariamente sobre o número de casos e óbitos causados pelo vírus (HALE et al., 2021). Assim, tornou-se imprescindível o uso de técnicas e modelos de previsão para projetar essa contagem em regiões e países com altos índices (LALMUANAWMA; HUSSAIN; CHHAKCHHUAK, 2020; MALKI et al., 2020; CHIMMULA; ZHANG, 2020; SALGOTRA; GANDOMI; GANDOMI, 2020; MELIN et al., 2020). No Brasil, já foram mais de 650 mil mortes (cerca de 0,3% da população). A vacinação no Brasil começou em 17 de janeiro de 2021. A metodologia de validação cruzada de séries temporais foi aplicada para reduzir possíveis variações temporais, com várias bases de treinamento e teste.

No Caso 1, vários modelos de previsão foram observados para mortes diárias e acumuladas. As séries temporais não possuem distribuição normal. O gráfico de dispersão e o correlograma mostram que casos e óbitos têm correlação positiva. Os índices de saúde e restrição e os índices de governo e restrição também apresentam correlação positiva. No entanto, quando essas variáveis são correlacionadas com outras, apenas os índices de saúde e governamental apresentam forte correlação.

Os números previstos no final da epidemia são altamente dependentes da duração da série temporal usada nos modelos (MARTINEZ; ARAGON; NUNES, 2020). Portanto, foi utilizada uma previsão de sete dias. As variáveis exógenas casos, índice de restrição, índice de resposta do governo, índice de saúde e índice econômico foram utilizadas para obter maior exatidão. A previsão de óbitos acumulados produziu estimativas melhores que as diárias (BRAGA et al., 2021). O uso dos múltiplos subconjuntos de treino e teste mostraram que as regressões não lineares tiveram o melhor ajuste para prever os óbitos acumulados, diferindo de outros estudos publicados anteriormente (DIVINO et al., 2022; MASUM et al., 2022; MOHAN et al., 2022; MALEKI et al., 2020). A variável casos foi a que mais influenciou no ajuste dos modelos no Caso 1.

No Caso 2, vários modelos de previsão foram observados para a média móvel de sete dias das mortes diárias. As séries temporais não possuem distribuição

normal. O gráfico de dispersão e o correlograma mostram que casos e óbitos têm correlação positiva, porém fraca. Ou seja, conforme a vacinação populacional aumenta, casos e óbitos passam a ter menos correlação. As vacinações parciais e completas têm a maior correlação. Óbitos e vacinação apresentaram coeficientes de correlação negativos, o que sugere que quanto maior a vacinação menor o número de óbitos.

Foram utilizadas as previsões para um horizonte de sete, 15 e 30 dias. As variáveis exógenas casos, taxa de reprodução do vírus, vacinação parcial, vacinação completa e índice de restrição foram utilizadas para obter maior exatidão das estimativas. As previsões com horizonte de sete dias produziram estimativas melhores que as demais. O modelo ARIMA(p,1,q) teve o melhor ajuste para prever as médias móveis diárias de sete dias. A cobertura vacinal foi a variável que mais contribuiu para as previsões dos óbitos no Caso 2.

PUBLICAÇÕES

O Caso 1 foi aceito para publicação na revista *Brazilian Archives of Biology and Technology* sob o titulo *Evaluating How the Social Restriction, the Government Response, the Health, and Economic Indices Affected the Prediction of the Number of Deaths Provoked by COVID-19 in Brazil Using Classical Statistical and Machine Learning Models*.

O Caso 2 está sendo produzido para publicação em revista a ser pesquisada conforme os temas: COVID-19, séries temporais, modelos estatísticos e aprendizado de máquina.

TRABALHOS FUTUROS

Como sugestão para trabalhos futuros, pode-se utilizar os métodos que melhores se ajustaram para outros países ou regiões menores como estados e municípios, e também o uso da modelagem SARIMA.

REFERÊNCIAS

ALIMADADI, Ahmad et al. Artificial intelligence and machine learning to fight COVID-19. **Physiological genomics**, v. 52, n. 4, p. 200-202, 2020.

ARDABILI, Sina F. et al. Covid-19 outbreak prediction with machine learning. **Algorithms**, v. 13, n. 10, p. 249, 2020.

BATISTA, Andre Filipe de Moraes et al. COVID-19 diagnosis prediction in emergency care patients: a machine learning approach. **MedRxiv**, 2020.

BENVENUTO, Domenico et al. Application of the ARIMA model on the COVID-2019 epidemic dataset. **Data in brief**, p. 105340, 2020.

BORREGO–MORELL, Jorge A.; HUERTAS, Edmundo J.; TORRADO, Nuria. On the effect of COVID-19 pandemic in the excess of human mortality. The case of Brazil and Spain. **PloS one**, v. 16, n. 9, p. e0255909, 2021.

BOX, George EP et al. **Time series analysis**: forecasting and control. [S.l]: John Wiley & Sons, 2015.

BRAGA, Marcus de Barros et al. Artificial neural networks for short-term forecasting of cases, deaths, and hospital beds occupancy in the COVID-19 pandemic at the Brazilian Amazon. **PLoS One**, v. 16, n. 3, p. e0248161, 2021.

CARVALHO-SAUER, Rita de Cássia Oliveira de et al. Impact of COVID-19 pandemic on time series of maternal mortality ratio in Bahia, Brazil: analysis of period 2011–2020. **BMC Pregnancy and Childbirth**, v. 21, n. 1, p. 1-7, 2021.

CAVALCANTI, Matheus Paiva Emidio et al. Trends in COVID-19 lethality and mortality rates in the State of Pernambuco, Brazil: a time series analysis from april 2020 to june 2021. **Journal of Human Growth and Development**, v. 32, n. 2, p. 327-338, 2022.

CESAR, Andre Evaristo Marcondes et al. Analysis of COVID-19 mortality and case-fatality in a low-income region: an ecological time-series study in Tocantins, Brazilian Amazon. **Journal of Human Growth and Development**, v. 31, n. 3, p. 496-506, 2021.

CHIMMULA, Vinay Kumar Reddy; ZHANG, Lei. Time series forecasting of COVID-19 transmission in Canada using LSTM networks. **Chaos, Solitons & Fractals**, v. 135, p. 109864, 2020.

CHEMAITELLY, Hiam et al. Eficácia da vacina mRNA-1273 COVID-19 contra B. 1.1. 7 e B. 1.351 variantes e doença grave de COVID-19 no Catar. **Medicina da natureza** , v. 27, n. 9, pág. 1614-1621, 2021.

CLINE, Graysen. **Nonparametric Statistical Methods Using R**. [S.l.]: Scientific e-Resources, 2019.

DAGAN, Noa et al. BNT162b2 mRNA Covid-19 vaccine in a nationwide mass vaccination setting. **New England Journal of Medicine**, 2021.

DIVINO, Fabio et al. Unreliable predictions about COVID-19 infections and hospitalizations make people worry: The case of Italy. **Journal of Medical Virology**, v. 94, n. 1, p. 26-28, 2022.

ELAZIZ, Mohamed Abd et al. New machine learning method for image-based diagnosis of COVID-19. **Plos one**, v. 15, n. 6, p. e0235187, 2020.

FIOCRUZ. **MonitoraCovid-19**. Instituto de Comunicação e Informação Científica e Tecnológica em Saúde (ICICT), 2022. Disponível em: <https://bigdata-covid19.icict.fiocruz.br>. Acesso em: 07 jul. 2022.

FONGARO, Gislaine et al. The presence of SARS-CoV-2 RNA in human sewage in Santa Catarina, Brazil, November 2019. **Science of the Total Environment**, v. 778, p. 146198, 2021.

GAO, Yue et al. Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. **Nature communications**, v. 11, n. 1, p. 1-10, 2020.

GHASEMI, Asghar; ZAHEDIASL, Saleh. Normality tests for statistical analysis: a guide for non-statisticians. **International journal of endocrinology and metabolism**, v. 10, n. 2, p. 486, 2012.

GUJARATI, Damodar N.; PORTER, Dawn C. **Econometria básica-5**. Amgh Editora, 2011.

HADDOUN, Abdelhakim et al. Modeling, analysis, and neural network control of an EV electrical differential. **IEEE Transactions on industrial electronics**, v. 55, n. 6, p. 2286-2294, 2008.

HALE, Thomas et al. A global panel database of pandemic policies (Oxford COVID-19 Government Response Tracker). **Nature human behaviour**, v. 5, n. 4, p. 529-538, 2021.

HASELL, Joe et al. A cross-country database of COVID-19 testing. **Scientific data**, v. 7, n. 1, p. 1-7, 2020.

HASTIE, Trevor et al. **The elements of statistical learning**: data mining, inference, and prediction. New York: Springer, 2009.

HYNDMAN, Rob J. et al. A state space framework for automatic forecasting using exponential smoothing methods. **International Journal of forecasting**, v. 18, n. 3, p. 439-454, 2002.

HYNDMAN, Rob et al. **Forecasting with exponential smoothing**: the state space approach. [S.l.]: Springer Science & Business Media, 2008.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. **Forecasting**: principles and practice. [S.l.]: OTexts, 2018.

HYNDMAN, Rob J. et al. Forecasting functions for time series and linear models. **R package version 8.12**. 2020.

ISLAM, Nazrul et al. Excess deaths associated with covid-19 pandemic in 2020: age and sex disaggregated time series analysis in 29 high income countries. **BMJ**, v. 373, 2021a.

ISLAM, Nazrul et al. Effects of covid-19 pandemic on life expectancy and premature mortality in 2020: time series analysis in 37 countries. **BMJ**, v. 375, 2021b.

JAMES, Gareth et al. **An introduction to statistical learning**. New York: Springer, 2013.

JARA, Alejandro et al. Effectiveness of an inactivated SARS-CoV-2 vaccine in Chile. **New England Journal of Medicine**, v. 385, n. 10, p. 875-884, 2021.

KASSAMBARA, Alboukadel. Ggcorrplot: Visualization of a Correlation Matrix using'ggplot2'. **R package version 0.1**, v. 3, 2019.

KHAN, Farhan Mohammad; GUPTA, Rajiv. ARIMA and NAR based prediction model for time series analysis of COVID-19 cases in India. **Journal of Safety Science and Resilience**, v. 1, n. 1, p. 12-18, 2020.

KUHN, M. et al. Classification and Regression Training. **R package version 6.0–81**. 2018.

LALMUANAWMA, Samuel; HUSSAIN, Jamal; CHHAKCHHUAK, Lalrinfela. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. **Chaos, Solitons & Fractals**, p. 110059, 2020.

LIMA, Nísia Trindade; GADELHA, Carlos Grabois. The COVID-19 Pandemic: Global Asymmetries and Challenges for the Future of Health. **China CDC Weekly**, v. 3, n. 7, p. 140, 2021.

LOPEZ BERNAL, Jamie et al. Effectiveness of Covid-19 vaccines against the B. 1.617. 2 (Delta) variant. **New England Journal of Medicine**, 2021.

MALEKI, Mohsen et al. Time series modelling to forecast the confirmed and recovered cases of COVID-19. **Travel Medicine and Infectious Disease**, p. 101742, 2020.

MALKI, Zohair et al. Association between weather data and COVID-19 pandemic predicting mortality rate: Machine learning approaches. **Chaos, Solitons & Fractals**, v. 138, p. 110137, 2020.

MARTINEZ, Edson Zangiacomi; ARAGON, Davi Casale; NUNES, Altacílio Aparecido. Long-term forecasts of the COVID-19 epidemic: a dangerous idea. **Revista da Sociedade Brasileira de Medicina Tropical**, v. 53, 2020.

MASUM, Mohammad et al. Comparative study of a mathematical epidemic model, statistical modeling, and deep learning for COVID-19 forecasting and management. **Socio-Economic Planning Sciences**, v. 80, p. 101249, 2022.

MATHIEU, Edouard et al. A global database of COVID-19 vaccinations. **Nature human behaviour**, v. 5, n. 7, p. 947-953, 2021.

MATTA, Gustavo Corrêa et al. **Os impactos sociais da Covid-19 no Brasil:** populações vulnerabilizadas e respostas à pandemia. 2021.

MELIN, Patricia et al. Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: the case of Mexico. In: **Healthcare**. Multidisciplinary Digital Publishing Institute, 2020. p. 181.

MELLO, Luiz Gustavo de. **Combinação ótima de métodos de previsão segundo o critério Payoff-Jolliffe fatorial: uma abordagem multivariada para a estimação de demanda de gás natural.** 2021. Tese (Doutorado em Engenharia de Produção) - Faculdade de Engenharia, Universidade Federal de Itajubá. Itajubá, 2021.

MINISTÉRIO DA SAÚDE. **Guia de vigilância epidemiológica.** Fundação Nacional de Saúde. 5.ed. Brasília, 2002.

MIRANDA, Beatriz Santos et al. Inteligência artificial como ferramenta de redimensionamento durante a pandemia de covid-19: revisão narrativa. **Revista Brasileira de Saúde Funcional**, v. 10, n. 1, 2022.

MODI, Chirag et al. Estimating COVID-19 mortality in Italy early in the COVID-19 pandemic. **Nature communications**, v. 12, n. 1, p. 1-9, 2021.

MOHAN, Sumit et al. Predicting the impact of the third wave of COVID-19 in India using hybrid statistical machine learning models: A time series forecasting and sentiment analysis approach. **Computers in Biology and Medicine**, v. 144, p. 105354, 2022.

MUELLER, John Paul; MASSARON, Luca. **Aprendizado de Máquina para Leigos**. Alta Books Editora, 2019.

NAGUIB, Ibrahim A.; DARWISH, Hany W. Support vector regression and artificial neural network models for stability indicating analysis of mebeverine hydrochloride and sulpiride mixtures in pharmaceutical preparation: A comparative study. **Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy**, v. 86, p. 515-526, 2012.

OLIVEIRA, Brigitte Renata Bezerra de et al. Determinants of access to the SARS-CoV-2 vaccine: a preliminary approach. **International journal for equity in health**, v. 20, n. 1, p. 1-11, 2021.

PARAGUASSU, L. Major Brazilian cities set lockdowns as virus spreads. **Reuters World News**, 2020.

PENG, Yaohao; NAGATA, Mateus Hiro. An empirical overview of nonlinearity and overfitting in machine learning using COVID-19 data. **Chaos, Solitons & Fractals**, v. 139, p. 110055, 2020.

R CORE TEAM. **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. 2019.

RANDHAWA, Gurjit S. et al. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. **Plos one**, v. 15, n. 4, p. e0232391, 2020.

ROSS, Sheldon M. **Introduction to probability and statistics for engineers and scientists**. [S.l.]: Academic press, 2020.

SALGOTRA, Rohit; GANDOMI, Mostafa; GANDOMI, Amir H. Time series analysis and forecast of the COVID-19 pandemic in India using genetic programming. **Chaos, Solitons & Fractals**, v. 138, p. 109945, 2020.

SHAO, Jun; TU, Dongsheng. **The jackknife and bootstrap**. [S.l.]: Springer Science & Business Media, 2012.