UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO DE JOINVILLE
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE SISTEMAS ELETRÔNICOS

Eduard Hermes Anschau

Modelagem e Identificação de Dados Epidemiológicos Associados à Pandemia de COVID-19 no Estado de Santa Catarina

Joinville
2022

Eduard Hermes Anschau

Modelagem e Identificação de Dados Epidemiológicos Associados à Pandemia de COVID-19 no Estado de Santa Catarina

Dissertação submetida ao Programa de Pós-Graduação em Engenharia de Sistemas Eletrônicos da Universidade Federal de Santa Catarina como requisito parcial para a obtenção do título de mestre em Engenharia de Sistemas Eletrônicos.
Orientador: Prof. Alexandro Garro Brito, Dr.
Coorientador: Prof. Pablo Andretta Jaskowiak, Dr.

Joinville
2022

Ficha de identificação da obra elaborada pelo autor,
através do Programa de Geração Automática da Biblioteca Universitária da UFSC.

Anschau, Eduard
Modelagem e Identificação de Dados Epidemiológicos Associados à Pandemia de COVID-19 no Estado de Santa Catarina / Eduard Anschau ; orientador, Alexandro Garro Brito, coorientador, Pablo Andretta Jaskowiak, 2022.
152 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Campus Joinville, Programa de Pós-Graduação em Engenharia de Sistemas Eletrônicos, Joinville, 2022.

Inclui referências.

1. Engenharia de Sistemas Eletrônicos. 2. Epidemiologia. 3. COVID-19. 4. ARIMA. 5. Nonlinear Autorregressive Moving Average With Exogenous Input. I. Garro Brito, Alexandro . II. Andretta Jaskowiak, Pablo. III. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Engenharia de Sistemas Eletrônicos. IV. Título.

Eduard Hermes Anschau

Modelagem e Identificação de Dados Epidemiológicos Associados à Pandemia de COVID-19 no Estado de Santa Catarina

O presente trabalho em nível de mestrado foi avaliado e aprovado por banca examinadora composta pelos seguintes membros:

Prof. Lucas Weihmann, Dr.
Universidade Federal de Santa Catarina

Prof. Benjamin Grando Moreira, Dr.
Universidade Federal de Santa Catarina

Certificamos que esta é a **versão original e final** do trabalho de conclusão que foi julgado adequado para obtenção do título de mestre em Engenharia de Sistemas Eletrônicos.

ICP-Edu
Documento assinado digitalmente
Moises Ferber de Vieira Lessa
Data: 13/10/2022 08:53:37-0300
CPF: ***.157.286-**
Verifique as assinaturas em https://v.ufsc.br

Coordenação do Programa de Pós-Graduacão

ICP-Edu
Documento assinado digitalmente
Alexandro Garro Brito
Data: 11/10/2022 17:12:22-0300
CPF: ***.180.626-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Alexandro Garro Brito, Dr.
Orientador

ICP-Edu
Documento assinado digitalmente
Pablo Andretta Jaskowiak
Data: 11/10/2022 17:21:53-0300
CPF: ***.115.189-**
Verifique as assinaturas em https://v.ufsc.br

Prof. Pablo Andretta Jaskowiak, Dr.
Coorientador

Joinville, 2022.

Este trabalho é dedicado ao meu amado vô
Mário Hermes, que já deixa enorme saudade e
um vazio intransponível em minha vida.

AGRADECIMENTOS

Primeiramente, agradeço à minha família, sem a qual nada faria sentido. Presto honras aos meus pais, Genésio e Jaqueline, ao meu irmão Martin, aos meus avós Mário e Isolde, à tia Carine e aos tios Gilberto e Gilmar, e à namorada Débora, a qual acompanhou com carinho a minha trajetória nos últimos anos e tem sido uma companheira ímpar.

Sou enormemente grato aos professores Alexandro e Pablo, por terem feito um excelente papel na orientação deste trabalho.

Por fim, agradeço à FAPESC pelo apoio financeiro concedido à concretização do presente projeto de dissertação.

RESUMO

O coronavírus (COVID-19) difundiu-se por todo o globo e tornou-se uma das grandes mazelas da contemporaneidade, impactando profundamente o Brasil, o qual configura como uma das nações mais afetadas pela doença. Desse modo, a necessidade por sistemas tecnológicos de combate à crise sanitária tornou-se ainda mais urgente nesse país. À vista disso, o presente trabalho apresenta um estudo comparativo entre três técnicas de modelagem e previsão de dados epidemiológicos associados à pandemia de COVID-19 no Brasil, especificamente no estado de Santa Catarina. Foram considerados modelos do tipo ARIMA, *Non-Linear Autoregressive model with eXogenous input* (NARX) e *Non-Linear Autoregressive model Moving Average with eXogenous input* (NARMAX) polinomiais para importantes séries de dados associadas à doença. O paradigma de validação dos modelos preditivos consistiu em janelas deslizantes de avanço não ancorado, sendo que o desempenho preditivo, avaliado por meio da aplicação de métricas de desempenho tradicionais, mostrou que, num âmbito geral, a técnica ARIMA gerou predições superiores em comparação com as modelagens NARX e NARMAX.

Palavras-chave: Epidemiologia. COVID-19. ARIMA. Nonlinear Autorregressive Moving Average With Exogenous Input.

ABSTRACT

The coronavirus (COVID-19) has spread all over the globe and has become one of the great ailments of contemporaneity, deeply impacting Brazil, which is one of the nations most affected by the disease. Thus, the need for technological systems to combat the health crisis has become even more urgent in this country. In view of this, the present work presents a comparative study between three modeling and forecasting techniques for epidemiological data associated with the pandemic of COVID-19 in Brazil, specifically in the state of Santa Catarina. We considered polynomial ARIMA-type models, *Non-Linear Autoregressive model with eXogenous input* (NARX) and *Non-Linear Autoregressive model Moving Average with eXogenous input* (NARMAX) for important data series associated with the disease. The validation paradigm for the predictive models consisted of unanchored forward sliding windows, and the predictive performance, assessed by applying traditional performance metrics, showed that, overall, the ARIMA technique generated superior predictions compared to the NARX and NARMAX models.

Keywords: Epidemiology. COVID-19. Nonlinear Autorregressive Moving Average With Exogenous Input.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ADF	*Augmented Dickey-Fuller*
AIC	*Akaike Information Criteria*
AM	Aprendizado de Máquina
ANFIS	*Adaptative Network-based Fuzzy Inference System*
AP	Aprendizado Profundo
ARIMA	*AutoRregressive Integrated Moving Average*
BIC	*Bayes Information Criteria*
EQMM	Erro Quadrático Médio Mínimo
ERR	*Error Reduction Ratio*
FIR	*Finite Impulse Response*
GNN	*Graph Neural Network*
IA	Inteligência Artificial
LSTM	Long Short-Term Memory
MAE	*Mean Absolute Error*
MAPE	*Mean Absolute Percentage Error*
MAR	*Missing At Random*
MBI	Modelo Baseado em Indivíduos
MCAR	*Missing Completely At Random*
MLP	*MultiLayer Perceptron*
MNAR	*Missing Not At Random*
MQ	Mínimos Quadrados
MQO	Mínimos Quadrados Ortogonais
MSE	*Medium Squared Error*
NAR	*Nonlinear AutoRregressive*
NARMAX	*Nonlinear AutoRregressive Moving Average with eXogenous input*
NARX	*Nonlinear AutoRregressive with eXogenous input*
OMS	Organização Mundial da Saúde
PRBS	*PseudoRandom Binary Sequence*
PSO	*Particle Swarm Optimizer*
RF	*Random Forest*
RMSE	*Root Mean Squared Error*
RRSE	Root Relative Squared Error
SG	Savitsky-Golay
SIR	*Susceptible - Infected - Recovered*
SVR	*Support Vector Regression*
SVM	*Support Vector Machine*
SUS	Sistema Único de Saúde
VMD	Variational Mode Decomposition

SUMÁRIO

1 INTRODUÇÃO

A modelagem e a simulação de sistemas são ferramentas de decisão importantes, as quais podem ser úteis na análise e controle de doenças humanas (ANDERSON; MAY, 1979; THIEME, 2003; IVORRA *et al.*, 2014). No entanto, como cada comorbidade apresenta características biológicas particulares, os modelos de previsão precisam ser adaptados a cada caso específico, a fim de se tornarem aptos a enfrentar situações reais (BRAUER *et al.*, 2012; YAN; CAO, 2019).

Na contemporaneidade, destaca-se o COVID-19 – doença infecciosa, cujos primeiros casos de infecção são datados de dezembro de 2019. Em 30 de janeiro de 2020, a Organização Mundial da Saúde, (2020), declarou que tal fenômeno epidêmico se tratava de uma Emergência de Saúde Pública de Importância Internacional. A doença se difundiu rapidamente por todo o mundo e, em 11 de fevereiro de 2020, a mesma instituição a renomeou como SARS-CoV-2 (GORBALENYA *et al.*, 2020). Em 11 de março de 2020, a doença já havia sido confirmada em cerca de 118 000 casos, relatados em 114 países, com mais de 90 por cento dos casos concentrados em apenas quatro deles. Dessa forma, a OMS declarou se tratar de uma pandemia. No Brasil, o primeiro caso confirmado da doença ocorreu em 2 de fevereiro de 2020. Considerando que, até o final do mês de janeiro de 2022, foram contabilizados aproximadamente 25 milhões de casos e cerca de 627 mil óbitos, o que faz com que esse país seja uma das nações mais afetadas pela doença. Especificamente, no estado de Santa Catarina, foram confirmados em torno de 1,35 milhões de casos e pouco mais de 20 mil óbitos pela enfermidade, até o dia 21/01/2022. À vista disso, constata-se a severidade da situação sanitária a que o Brasil, bem como o estado catarinense, estiveram submetidos. Tendo em conta o impacto causado pela pandemia de COVID-19, faz-se fundamental que as autoridades sanitárias e governamentais estejam dotadas de recursos tecnocientíficos para análise e descrição de dados associados à dinâmica das doenças.

A epidemiologia avançou significativamente nos últimos dois séculos, graças ao progresso intelectual nas esferas da biologia celular, biologia molecular e imunologia. O resultado direto desse desenvolvimento pode ser notado na

redução das taxas de mortalidade e no crescimento de índices associados à expectativa de vida (ANDERSON; MAY, 1992). No que se refere a esse domínio científico, a epidemiologia matemática tem exercido papel de crescente destaque na análise de fenômenos epidêmicos, sendo que o interesse em modelar doenças infecciosas tem sido objeto de inúmeros trabalhos em todo o mundo (HETHCOTE, 2000; YANG, H. M., 2001). À vista disso, Yang (2001) afirma que "modelos matemáticos têm auxiliado os sanitaristas na escolha do melhor mecanismo de controle de doenças infecciosas por meio de vacinações". Clancy, (1999), ressalta que o principal motivador para o estudo de modelos matemáticos de análise de proliferação de doenças está na possibilidade de que uma melhor compreensão dos mecanismos de transmissão possa proporcionar estratégias de combate mais efetivas. Nesse sentido, a epidemiologia matemática tem se beneficiado de técnicas e metodologias típicas de engenharia. Assim, procura-se modelar, predizer e projetar ações de controle, a fim de minimizar os danosos efeitos de epidemias. Complementarmente, esses modelos servem para estudar e avaliar situações antes que ela ocorram.

Dentre as ferramentas matemáticas pioneiras em matéria de análise epidemiológica, destaca-se o modelo compartimental SIR, fundamentado na tríade "suscetíveis, infectados e recuperados" (HETHCOTE, 2000). Esse modelo descreve a epidemia como um sistema de equações diferenciais que relaciona as parcelas da população contidas nos compartimentos S, I e R, de forma análoga à modelagem da interação entre partículas segundo o princípio da ação de massas (KEELING; ROHANI, 2002). Entretanto, o SIR não é capaz de explicar a persistência ou erradicação de doenças infecciosas (KEELING; GRENFELL, 2000; LLOYD, 2001). Defende-se que a principal razão para isso é que esse modelo considera uma distribuição de indivíduos espacial e temporalmente homogênea (KEELING; ROHANI, 2002). Dessa forma, uma abordagem para lidar com a questão de populações heterogêneas, estudada em ecologia, são os chamados Modelos Baseados em Indivíduos, MBI (ou IBM, do inglês *Individual Based Model*) (GRIMM, 1999; KEELING; GRENFELL, 2000; NEPOMUCENO *et al.*, 2006). Conforme Grimm (1999), "cada indivíduo é tratado como uma entidade única e discreta que possui idade e ao menos mais uma propriedade que muda ao longo do ciclo da vida, tal como peso, posição social, entre outras".

O estado da arte dos modelos descritivos e preditivos é em grande parte representado por tecnologias mais avançadas, a exemplo da inteligência artificial (IA), a qual engloba os conceitos de aprendizado de máquina (AM) e aprendizagem profunda (AP) - as quais podem ser empregadas para identificar e prever distintos aspectos de dados epidêmicos, por meio da aplicação, a exemplo, de redes neurais artificiais e algoritmos genéticos. As principais áreas onde essas técnicas podem ser aplicadas são no diagnóstico precoce de doenças, no rastreamento de contato, no desenvolvimento de medicamentos e vacinas, bem como na previsão de casos de contração de doenças (ZHU *et al.*, 2019; HEWAMALAGE *et al.*, 2021). No âmbito da identificação de séries temporais associadas à epidemias, tais como a difundida globalmente pelo novo coronavírus, destaca-se a aplicação de Redes Neurais Recorrentes, capazes de lidar com não linearidades, assim como com interdependências inerentes aos dados (HEWAMALAGE *et al.*, 2021; HOCHREITER; SCHMIDHUBER, 1997). Outra classe de técnicas bem sucedida na identificação de séries de dados corresponde ao modelo NARX (*Nonlinear Autoregressive with Exogenous Inputs*), o qual mostra bom desempenho na identificação de séries temporais não-lineares, bem como agrega uma importante característica: a facilidade com que certos tipos de conhecimentos podem ser extraídos e incorporados na modelagem (AGUIRRE, 2004).

O presente trabalho propõe a aplicação três de métodos de modelagem para análise e previsão de dados epidemiológicos acerca da pandemia de COVID-19. Para tanto, propõe-se uma análise comparativa entre duas três técnicas de previsão, aplicadas à importantes séries temporais associadas à pandemia no estado de Santa Catarina, a saber: os modelos ARIMA, NARX e NARMAX.

Acredita-se que esta pesquisa trará impactos tanto do ponto de vista científico como social. De posse de tais ferramentas, cientistas e autoridades serão capazes de prever comportamentos e tomar decisões para frear o avanço da epidemia e prover ações de manutenção da atividade econômica em nível local e estadual.

1.1 OBJETIVOS

1.1.1 Objetivo Geral

O presente trabalho tem como objetivo principal a concepção de um estudo comparativo entre três técnicas de modelagem e previsão de dados epidemiológicos associados à pandemia de COVID-19 no Brasil, especificamente no estado de Santa Catarina. Serão considerados modelos do tipo *Non-Linear Autoregressive with eXogenous input* (NARX) e *Non-Linear Autorregressive Moving Average with eXogenous input* (NARMAX) polinomiais, como contraponto ao modelo ARIMA, utilizado aqui como referência comparativa - ou *baseline*. Para tanto, as técnicas supramencionadas serão aplicadas na predição de importantes séries de dados associadas à doença no estado catarinense, como a quantidade de casos confirmados no estado e óbitos, bem como as quantidades de pacientes recuperados da doença e leitos SUS ocupados.

Dessa forma, pretende-se desenvolver análises epidemiológica preditivas úteis na previsão de dados atinentes pandemia em Santa Catarina.

1.1.2 Objetivos Específicos

- Elaboração e descrição de base de dados associada à pandemia de COVID-19 no Estado de Santa Catarina, visando as quantidades diárias de casos confirmados, número de óbitos, quantidade de pacientes recuperados e quantidade de leitos SUS ocupados por enfermos da doença.
- Aplicação do modelo ARIMA na previsão das séries históricas, o qual será utilizado como base comparativa com os outros modelos.
- Aplicação dos algoritmos de identificação NARX e NARMAX polinomiais para modelagem matemática das séries de dados supracitadas, assim como algoritmos de cálculo de estimação de parâmetros, critério de informação e para determinação do *Error Reduction Ration* (ERR) associado aos termos dos modelos NARX.
- Análise e comparação entre as técnicas de modelagem empregadas.

1.1.3 Justificativa

Para uma eficiente mitigação e prevenção das mazelas associadas à pandemia de COVID-19, faz-se necessário o tratamento dos dados atinentes à doença a partir de ferramentas consagradas na literatura científica - possibilitando a análise epidemiológica e posterior compartilhamento da informação. Dessa forma, em conformidade com Lopes (2020), os sistemas de vigilância em saúde podem ser aprimorados a fim de permitir o acompanhamento em tempo real da doença.

Nesse sentido, o estudo da dinâmica de transmissão de doenças infecciosas, utilizando modelos matemáticos preditivos, viabiliza subsídio ao planejamento adequado das estratégias de intervenção sanitária. De acordo com Bistrian et al., (2019), previsões realizadas por meio de modelagem matemática proporcionam melhorias às campanhas de vacinação, bem como permitem a alocação adequada de recursos, a qual evita a redundância na distribuição de profissionais de saúde. A partir dessa linha de pensamento, Bontempi et al., (2013), sugere que o tratamento dos dados, efetuado mediante algoritmos computacionais, permite a predição da ocorrência através da análise de dados pregressos. Dessa forma, segundo o autor, possibilita-se a compreensão de eventos em diferentes áreas do conhecimento.

Tendo em vista as séries históricas tratadas pelo presente trabalho, tem-se que os conjuntos de dados relativos às quantidades de casos de infecção, óbitos e recuperações estão entre os mais abordados em trabalhos envolvendo a predição numérica no âmbito da pandemia de COVID-19. A quantidade de casos de infecção pela doença é, a exemplo, a única variável tratada nos trabalhos de Singh et al. (2020) e Moftakhar e Seif (2020). Em suma, o primeiro trabalho aplica a modelagem ARIMA para a previsão de casos diários na Malásia e, a partir da experimentação de diversos parâmetros da técnica, o modelo mais eficiente em termos da métrica de desempenho e critério de informação adotados, foi efetivamente utilizado para as previsões. Ainda, esse trabalho buscou relatar o comportamento da tendência dos dados previstos, o que enriqueceu a análise. Já o segundo trabalho supracitado utiliza-se da mesma técnica preditiva para tratar os dados do Irã. Nesse trabalho, o horizonte de previsão é de um mês,

sendo que o conjunto de dados de treinamento também possui comprimento mensal. Nesse caso, pode-se dizer que a base de dados é bastante limitada para a efetuação de previsões eficazes.

Ainda, há os trabalhos que efetuam previsões acerca das três séries históricas (casos de infecção, óbitos e recuperações), tais como os artigos de Swaraj et al. (2020) e Fang et al. (2020). Em Swaraj et al. (2020), uniu-se as técnicas ARIMA e NAR (*nonlinear autorregressive*) em um modelo híbrido, de tal modo que a parcela ARIMA do modelo foi responsável por identificar a relação linear contida nos dados. Por outro lado, a parcela NAR é incumbida de detectar componentes autorregressivas de natureza não linear das séries históricas. O horizonte preditivo adotado é de um mês, sendo que foi possível notar importante tendência nos dados previstos para dados de até um mês à frente. No entanto, esse trabalho se utiliza de um conjunto de dados de treinamento de apenas um mês - o que sugere uma base de dados consideravelmente reduzida quando o objetivo é o estudo pormenorizado dos resultados das predições.

Em Fang et al., (2020), modelos ARIMA foram desenvolvidos para prever os casos cumulativos confirmados, mortos e recuperados, na Rússia. Esse trabalho indica que a capacidade do sistema de saúde da Rússia, à época em que o artigo foi submetido, pode responder com eficácia à pandemia COVID-19, considerando que os três modelos concebidos têm um bom efeito de ajuste e podem ser usados para previsões de curto prazo.

Como alternativa aos trabalhos que utilizam ARIMA como técnica preditiva, o trabalho de Wei e Billings (2021), efetua análises preditivas por meio de modelagem NARMAX. Nesse trabalho, são tratadas as variáveis de casos de infecção e óbitos diários, o que reforça a ênfase da literatura preditiva, concernente à pandemia de COVID-19, em relação às séries históricas a serem tratadas na presente dissertação. No trabalho supramencionado, houve a adição do fator de reprodução (número R) epidêmico, associado à disseminação dos casos de COVID-19, como variável exógena do modelo NARMAX, a fim de enriquecer o processo de identificação das quantidades diárias de casos e mortes pela doença.

Os trabalhos supracitados revelam a preocupação científica quanto à análise preditiva das séries históricas associadas aos casos, óbitos e recupe-

rações de pacientes no âmbito da pandemia de COVID-19. Já a necessidade pela eficiente previsão da ocupação de leitos faz alusão à pressão que os sistemas de saúde sofrem com a alocação de recursos clínicos para pacientes em estado crítico. De fato, Noronha et al. (2020) aponta a importância desse fenômeno, bem como menciona uma variedade de trabalho nacionais e internacionais acerca desse tema. Ainda, o autor constata que, sem nenhuma ação de ampliação da oferta de leitos, há alta probabilidade de saturação dos sistemas de saúde em um espaço de tempo relativamente curto.

De fato, a preocupação com o estudo e previsão da ocupação de departamentos de emergência não é novidade na literatura médica. Os trabalhos de Hoot et al., (2007), e Schweigler et al., (2008), por exemplo, já abordavam a necessidade de se prever, no curto prazo, indicadores técnicos ambulatoriais e a taxa de ocupação dos setores de emergência, a fim de melhor gerenciar os recursos disponíveis. Um estudo mais atual, proposto por Tuominen et al., (2022), reforça a importância dada a esse tipo de temática, ao estudar com mais afinco a seleção de preditores exógenos para prever de forma mais aprimorada a ocupação de departamentos de emergência.

Como exemplo de trabalhos que consideram a influência da ocupação de leitos para o enfrentamento da pandemia de COVID-19 no Brasil, Rache et al. (2020) apontaram as regiões de saúde mais vulneráveis à pandemia da COVID-19, assim como a necessidade de leitos de UTI. Nesse trabalho, as estimativas de oferta foram realizadas considerando-se apenas leitos do Sistema Único de Saúde (SUS) e não levaram em consideração fatores demográficos associados à variação regional dos casos.

Tendo em vista as informações supracitadas, faz-se evidente a pertinência não apenas da temática de pesquisa abordada por esta dissertação, como também do uso das quatro séries históricas abordadas ao longo do trabalho.

1.1.4 Organização do Trabalho

Este trabalho se encontra dividido em cinco capítulos. Os assuntos abordados por cada um deles são descritos a seguir. O Capítulo 2 se refere à fundamentação teórica do presente trabalho, bem como apresenta o estado

da arte atinente à modelagem preditiva de dados associados à pandemia de COVID-19. No Capítulo 3, são apresentados os resultados obtidos. Já o método proposto, assim como informações acerca da continuidade do estudo, são relatados no Capítulo 4. O Capítulo 5 se destina à conclusão do trabalho.

2 REVISÃO BIBLIOGRÁFICA

No decurso desta seção, são expostas as bases teóricas que fundamentam os modelos de identificação não linear preconizados pelo presente trabalho. Adicionalmente, ao final da seção, é descrito o estado da arte concernente à literatura de previsão de dados epidemiológicos, sobretudo àqueles associados à pandemia de COVID-19.

2.1 SÉRIES TEMPORAIS

Conforme Downing e Clark (2006), “As séries temporais (ou históricas) são conjuntos de medidas de uma mesma grandeza, relativas a vários períodos consecutivos”. De fato, a série temporal configura-se como uma sucessão de valores de determinada variável, observada em intervalos regulares de tempo. A variável independente inerente a essa categoria é o tempo e as séries temporais são ordenadas cronologicamente (SILVA, Elio Medeiros da; SILVA, Ermes Medeiros da, 1999), de tal modo que variar a ordem das observações modifica a informação contida na série. Se denotarmos a série temporal como sendo $Z(t)$, o valor da série no momento t pode ser escrito como $Z(t = 1, 2, ..., n)$, em que t é o parâmetro de tempo. A curva obtida no gráfico da série temporal define sua trajetória e o conjunto de todas as possíveis trajetórias fundamenta um processo estocástico. Dessa forma, uma série temporal pode ser definida como uma realização - ou amostra - desse processo estocástico.

O conjunto de observações ordenadas no tempo pode ser discreto como o número de atendimentos diários em um Pronto Socorro ou a quantidade diária de casos notificados de uma doença específica – como no caso deste trabalho. Por outro lado, o conjunto de dados pode ter natureza contínua, como o registro de um eletrocardiograma ou a mensuração dos valores de temperatura e umidade ao longo do dia. Pode-se obter uma série temporal discreta a partir de uma amostra de pontos de uma série contínua ou por meio de um parâmetro como, por exemplo, a média de períodos fixos de tempo.

Ainda, conforme Weigend (2018), pode-se fazer a decomposição de uma série temporal no seguintes itens.

- Tendência: elementos de longo prazo associados à série.
- Ciclos: longas ondas, com certo grau de regularidade, situadas em torno de uma linha de tendência.
- Sazonalidade: padrões regulares da série temporal.
- Aleatório: todos os efeitos não contabilizados pelas componentes supracitadas, ou seja, trata-se do resíduo.

Esses componentes podem ser estocásticos ou determinísticos, a depender do conjunto de dados que compõe a série.

Na análise de uma série temporal, primeiramente deseja-se efetuar uma análise estatística e descritiva acerca do fenômeno estudado para, a partir daí, descrever o comportamento da série a partir de modelagem e, por último, realizar estimativas futuras. Tarefa crucial atinente a essa temática também cabe à avaliação dos fatores que influenciaram o comportamento da série histórica, buscando definir ou mensurar relações de causa e efeito entre conjuntos distintos de dados. Para cada um desses propósitos, há uma variedade de critérios e ferramentas - dessas quais, algumas serão descritas e, após, aplicadas nesta pesquisa.

2.2 ANÁLISE E TRATAMENTO DE DADOS

Esta seção abordará os conceitos de análise exploratória de dados utilizados para o desenvolvimento do trabalho. A investigação dos dados está dividida em três etapas - a coleta de dados, a análise e o pré-processamento, que configuram etapa crucial, anterior à aplicação de modelos preditivos. De fato, para que a identificação de séries temporais seja aperfeiçoada e os modelos preditivos estejam aptos a desempenhar adequadamente, é necessário que se tenha compreensão acerca da base de dados, bem como que ela tenha boa qualidade.

2.2.1 Coleta e Análise de Dados

A coleta de dados é a primeira etapa para a identificação do sistema. Para sistemas dinâmicos de natureza não autônoma, os sinais de saída devem ser observados por meio introdução de sinais de excitação. Dessa forma, o comportamento do sistema é observado através de medições diretas das variáveis de saída (quando disponíveis) ou por meio da observação das variáveis de estado. Nesse contexto, os sinais de entrada são aplicados para estimular a dinâmica do sistema e, para isso, o sinal de entrada necessita constituir características adequadas no âmbito de seu espectro de potência. No caso de modelos não lineares, enquanto a amplitude deve permanecer pequena, a fim de evitar a condução do sistema fora da faixa de comportamento linear escolhido, os dados de entrada devem ser capazes de explorar as características não lineares do sistema e, muitas vezes, isso requer excitar o sistema com sinais de maior amplitude. Certas condições devem ser atendidas para escolher sinal de excitação apropriado. Na prática, ruído branco gaussiano e, especialmente no caso de sistemas lineares, a sequência binária pseudo-aleatória (PRBS) é comumente utilizada como sinal de entrada. Finalmente, a amostragem de dados deve ser realizada considerando o contexto espectral dos sinais.

No caso do presente trabalho, o sistema epidêmico que se pretende modelar associa-se a um tipo especial de dado, denominado como série temporal. Pragmática e operacionalmente, uma série temporal corresponde a uma sucessão de mensurações de pelo menos um determinado objeto de investigação no tempo. No âmbito da Epidemiologia, as mensurações são obtidas discreta e agregadamente no tempo (*clustering in time*) (FORATTINI, 1996b). Ou seja, a sua unidade de informação e consequente unidade de análise é o resultado de uma medida agregada no tempo, espaço ou tempo-espaço. Esses dados têm como característica principal a dependência entre observações vizinhas, ou seja, o caráter cronológico de disposição das informações (BROCKWELL, Peter J; DAVIS, Richard A, 2002). Especificamente, no caso de fenômenos epidêmicos, a exemplo da pandemia de COVID-19, séries temporais são comumente dispostas em bancos de dados de propriedade de órgãos de saúde públicos ou privados, bem como de hospitais ou departamentos acadêmicos voltados à te-

mática da saúde. Sendo assim, no aspecto técnico, a coleta desse tipo de dado dá-se pela aquisição ou transferência desse banco de dados a um *software* que reconheça sua tipologia.

2.2.2 Análise Exploratória

A análise exploratória dos dados consiste no estudo e investigação da base de dados. Com isso, é realizada a inspeção qualitativa e estatística que, a fim de caracterizar a base de dados e como suas observações estão distribuídas. Do ponto de vista estatístico, a análise exploratória de dados é uma abordagem de análise de conjuntos de dados para resumir suas principais características, geralmente usando gráficos estatísticos e outros métodos de visualização de dados, tais quais: o histograma; gráficos de densidade probabilística e gráficos *box-plot*. Modelos estatísticos de identificação do dados podem, ou não, ser usados, mas essa abordagem serve para investigar a essência dos dados além da modelagem formal e, portanto, contrasta o teste de hipóteses tradicional. A análise exploratória é diferente da análise de dados inicial, que se concentra mais estritamente na verificação de suposições necessárias para o ajuste do modelo e teste de hipóteses, além de lidar com valores ausentes e fazer transformações de variáveis conforme necessário (CHATFIELD, 1995).

Segundo Tukey et al. (1977), os objetivos da Análise Exploratória de Dados são:

- Propiciar descobertas inesperadas acerca dos dados.
- Sugerir hipóteses sobre as causas do fenômeno observado.
- Acessar suposições para as quais foram fundamentadas inferências estatísticas.
- Auxiliar na seleção de técnicas e ferramentas estatísticas adequadas ao conjunto de dados.

2.2.3 Identificação de *Outliers* e Imputação de Valores Faltantes

O pré-processamento dos dados refere-se à efetuação de procedimentos para conversão dos dados originais coletados em formatos úteis e eficientes. Em síntese, o pré-processamento compreende a preparação, organização e estruturação de dados, constituindo atividade essencial que precede a realização de análises e modelagem. Nesse sentido, a identificação de anomalias na estrutura dos dados é parte preponderante da análise de séries temporais. *Outliers* são, talvez, o tipo mais comum de anomalia presente em cenários de pesquisa que envolvem a coleta e análise de dados (PEARSON, 2005). A conceituação clássica de *outlier*, ou anomalia, é dada por Hawkins (1980), que define o fenômeno como sendo a observação que se desvia tanto das outras observações presentes no conjunto de dados, ao ponto de levantar suspeita acerca do processo que a gerou. Dessa forma, diz-se que *outliers* são variações substanciais da norma (MEHROTRA *et al.*, 2017) e, segundo Inoue et al. (2017), identificá-las consiste na tarefa de encontrar padrões nos dados que não estejam em conformidade com o comportamento esperado. Assume-se, por exemplo, que os valores esperados estejam situados em regiões de alta densidade probabilística do modelo estocástico, enquanto que as anomalias são associadas a regiões de baixa probabilidade (DAVIS, N. *et al.*, 2020).

De acordo com Chen e Liu (1993), dados de séries temporais são frequentemente contaminados com valores discrepantes devido à influência de eventos incomuns e não repetitivos. Conforme Aggarwal (2017), no que se refere a dados temporais, a detecção de *outliers* é feita tendo-se em conta dois propósitos distintos. Em alguns procedimentos de limpeza de dados, é necessário identificar dados anômalos para, então, suprimi-los. Em outros casos, busca-se identificar os *outliers* para analisar o eventual fenômeno incomum que os tenha originado. Nesse caso, a finalidade do procedimento de detecção visa o próprio dado destoante, a fim de conhecer melhor suas características.

Ainda, em conformidade com Blázquez-Garica et al. (2021), as técnicas de identificação de outliers em séries históricas dependem das seguintes particularidades:

- O tipo de dados das séries de entrada: em resumo, esse ponto refere-se

à faceta univariada ou multivariada das séries temporais de entrada.

- O tipo de *outlier*: esse aspecto visa a natureza da informação anômala que o método busca detectar. Nesse sentido, a anomalia pode estar associada a um ou mais pontos isolados na série, associados a instantes específicos no tempo. Ainda, outro tipo de *outlier* pode referir-se a uma sequência de pontos, mesmo que cada observação avulsa do conjunto não se configure como *outlier* pontual. Complementarmente, uma série inteira pode ser considerada um *outlier*, a qual poderá ser identificada somente para um caráter multivariados das séries históricas de entrada do sistema (BLÁZQUEZ-GARCIA *et al.*, 2021).

- A natureza dos métodos de detecção de *outliers*: assim como no caso do tipo de dado de entrada, o caráter da técnica de identificação de anomalias pode ser univariado ou multivariado. Em síntese, o método univariado aplica-se a uma série temporal enquanto que um método multivariado operará, simultaneamente, em mais de uma série temporal. No entanto, um método univariado pode ser aplicado a um conjunto de dados multivariado, atuando em uma série histórica por vez.

De forma complementar à formulação supracitada, tem-se a designação elaborada por Gupta et al. (2014), a qual descreve novas especificidades de análise de *outliers* em séries temporais. Sinteticamente, os aspectos abordados pelos autores são os seguintes.

- Supervisão de dados: existem exemplos de dados anômalos previamente identificados na série temporal em estudo, os quais podem servir como base para detecção de novos valores discrepantes?

- A natureza contínua ou discreta: as séries temporais são contínuas ou discretas?

- Detecção pontual ou sequencial: busca-se detectar *outliers* pontuais ou janelas anômalas de dados?

O impacto de *outliers* nas estimativas dos parâmetros de modelos de previsão para séries temporais tem sido extensivamente estudado desde que Fox (1972) propôs os conceitos de *outliers* aditivos e *outliers* inovativos. Ao estudarem o efeito de *outliers* aditivos nas previsões, Hillmer (1984) e Ledolter (1989) descobriram que os intervalos de previsão são bastante sensíveis a *outliers* aditivos, mas que as previsões pontuais não são consideravelmente afetadas, a menos que o valor discrepante ocorra nas proximidades da origem da previsão. Posteriormente, Chen e Liu (1993), enfatizaram a importância do tratamento de *outliers*, visto que a presença de dados anômalos pode acarretar a redução na precisão de modelos preditivos devido a (1) um efeito de transferência do valor discrepante na previsão pontual e (2) uma polarização nas estimativas dos parâmetros do modelo.

Tendo em vista as informações e conceitos discutidos, pode-se justificar a aplicação de procedimento de detecção de dados discrepantes como etapa do tratamento dos dados atinentes a este trabalho.

2.2.3.1 Filtro de Hampel

A partir da descrição de *outliers* concebida, bem como da sua importância no âmbito da análise de séries temporais, percebe-se que existe a necessidade de um procedimento de detecção de *outliers* versátil e, ao mesmo tempo, robusto. Nesse sentido, para esta pesquisa, optou-se pelo filtro de Hampel como ferramenta de identificação de dados discrepantes, o qual é um detector estatístico e opera independentemente de suposições acerca da distribuição dos dados, assim como é capaz de detectar com destreza vários *outliers*, sem a necessidade de supervisionamento de dados discrepantes previamente rotulados (BATISTA JÚNIOR; PIRES, 2014). Esse filtro, considerado um dos identificadores de *outliers* mais robustos e eficientes (LIU, H. *et al.*, 2004), satisfaz os requisitos supracitados, baseando-se no paradigma de janelas móveis de dados, o que é conveniente quando se trabalha com dados temporais, visto que o filtro se ajusta de maneira móvel no decurso cronológico da série histórica. Em suma, o filtro de Hampel é uma adaptação robusta da regra probabilística 3σ, por operar utilizando-se da mediana dos dados - e não da média aritmética. A

regra 3σ expressa uma heurística convencional de que quase todos os valores associados a uma distribuição estão contidos entre os limites de três desvios padrão da média.

A operação do filtro de Hampel é descrita da seguinte forma. Concebe-se uma janela móvel, a qual percorre os dados da série temporal, denotada por,

$$W_k^N = \{x_{k-N}, \ldots, x_k, \ldots, x_{k+N}\}.$$

onde N denota a metade da janela móvel e é determinado por um número inteiro positivo. Por outro lado, k denota o índice da janela de dados, ou seja, faz referência a uma janela específica de observações. Utiliza-se o filtro de mediana padrão $\mathcal{M}_K$, introduzido por J. W. Tukey em 1974 (TUKEY, J., 1974), o qual calcula a mediana da janela móvel W_k^K e é dada pela Equação (1).

$$\mathcal{M}_k = \text{mediana}\left\{W_k^N\right\}. \tag{1}$$

A versão do filtro de Hampel aqui considerada representa uma implementação de janela móvel do identificador Hampel descrito por Davies e Gather (1993). Especificamente, a resposta desse filtro é dada pela Equação (2).

$$y_K = \begin{cases} x_k, & |x_k - m_k| \leq t.S_k, \\ m_k, & |x_k - m_k| > t.S_k. \end{cases} \tag{2}$$

onde $t = 3$ - em conformidade com a regra 3σ; m_k é a mediana da janela de dados móvel e S_k é o desvio absoluto médio (DAM) dos dados, definido pela Equação (3).

$$S_k = k \times \text{mediana}_{j\in[-M,M]}\left\{|x_{k-j} - m_k|\right\}. \tag{3}$$

k é um fator multiplicativo que depende da distribuição dos dados de toda a série temporal e é expressa pela razão entre o desvio padrão e o desvio médio absoluto da amostra.

Em conformidade com as conceituações descritas por Blásquez-García et al. (2021) e Gupta et al. (2014), o filtro de Hampel pode ser descrito como um método de detecção de outliers pontuais para séries temporais discretas e univariáveis, que não se utiliza do paradigma de supervisão.

Em síntese, e reforçando alguns dos argumentos supracitados, a pertinência do uso do filtro de Hampel é dada pelas seguintes justificativas.

- O método utiliza o paradigma de janelas deslizantes, o qual é mais apropriado, em comparação a um método estatístico aplicado a toda série histórica, visto que o método ajusta à sequência cronológica dos dados.
- Unifica características de robustez à simplicidade de implementação (ROUSSEEUW; CROUX, 1993).
- É um método de detecção versátil, o qual é útil, a priori, a qualquer tipo de série temporal, e que se utiliza de cálculos estatísticos robustos. De fato, o filtro de Hampel já foi utilizado em trabalhos que visam séries históricas de diversos sistemas, com aplicação na engenharia química (BELISÁRIO *et al.*, 2020); eletromiografia superficial (ALLEN, s.d.); sensoriamento de redes *wireless* (CHEN, Y. C. *et al.*, 2010) e análise trajetória de radares em aviação (JÚNIOR; PIRES, 2014).
- Utilizado no âmbito de epidemias, com trabalhos relacionados à previsão de séries epidêmicas, inclusive associados à pandemia de COVID-19 (JIN, 2020).

2.3 IMPUTAÇÃO DE DADOS AUSENTES

Em estudos que envolvem a análise de dados, um dos desafios mais comuns se dá pela ausência de algumas observações no conjunto de dados. Desse fenômeno, surge a necessidade de considerar algum critério de estimação dessas observações faltantes.

Em concordância com Donders et al. (2006), os principais tipos de observações ausentes são descritos da seguinte forma:

- *Missing At Random* (MAR): referem-se aos dados que estão ausentes na base de dados de maneira aleatória, cuja falta resulta de algum fator externo ou característica já conhecida. Um exemplo desse tipo de observação refere-se ao fato de que o homem é menos propenso a preencher questionários acerca de pesquisas sobre depressão. No entanto, tal fato não se relaciona com o nível de depressão desse homem, depois de conhecido o seu sexo;

- *Missing Completely At Random* (MCAR): se o conjunto de observações munido de valores ausentes é subconjunto da coleção completa, diz-se que os dados ausentes são do tipo MCAR, i.e., são totalmente aleatórios. Dessa forma, a ausência da observação não está relacionada a nenhuma outra característica da base de dados. A perda acidental de amostras de um estudo é exemplo desse caso;
- *Missing Not At Random* (MNAR): refere-se aos valores que existem mas foram omitidos da base dados. Nesse cenário, há perda de informação valiosa da base de dados. Uma pessoa que não declara seu imposto de renda é exemplo desse caso.

Considerando-se que a ausência de dados epidêmicos de uma base de dados, a exemplo da falta de algumas observações diárias acerca da quantidade de casos de infecção por COVID-19, referem-se ao caso MCAR, o presente trabalho se utilizou da técnica de interpolação linear como método de imputação de valores ausentes.

2.3.1 Interpolação Linear

Este método assume que um ponto estimado se situará no vetor que conecta os pontos adjacentes à observação que se pretende estimar. Em suma, esse método utiliza uma função linear para aproximar a função que representa bem os dados e calcular o valor ausente (RANTOU, 2017).

2.4 SUAVIZAÇÃO DE RUÍDO

A dinâmica de propagação das pandemias é semelhante ao comportamento de outros sistemas não lineares, a exemplo de mapas caóticos e fluxos turbulentos (BONASERA; ZHANG, 2020). Nesses sistemas, em geral, o comportamento caótico indica que o sistema possui aspecto ruidoso, sendo que algumas características particulares da dinâmica caótica podem ser usadas para a análise de pandemias humanas. A investigação da dinâmica caótica de pandemias não é novidade na literatura científica, a exemplo dos estudos concernentes à febre do Vale do Rift (PEDRO *et al.*, 2014); à doenças infantis

(BILLINGS, L.; SCHWARTZ, 2002); e à epidemias em ecossistemas (EILERSEN *et al.*, 2020).

No âmbito da pandemia de COVID-19, os trabalhos de Debbouche et al. (2021) e Sapkota et al. (2021) mostram, a partir da aplicação de ferramentas de análise espectral, de atratores e da estabilidade de pontos de equilíbrio, que as séries temporais epidemiológicas atinentes à doença pertencem à categoria de sistemas caóticos.

O ruído presente nos dados limita a capacidade de extrair informações quantitativas de sinais que variam no tempo. Em muitos casos científicos, há o interesse em uma série temporal de dados produzidos por um sistema cujo comportamento introduz comportamento caótico, como o caos de baixa dimensão. Dessa forma, faz-se pertinente efetuar a distinção entre a componente ruidosa e o sinal real de séries temporais, a fim de que se torne possível análise mais clara acerca dos dados. Tal objetivo pode ser alcançado por meio da minimização do ruído presente nos dados. Para esse propósito, filtros simples de média móvel são os mais tradicionais; contudo, estudos recentes têm aplicado dispositivos mais complexos, como o ajuste móvel de polinômios interpoladores (ARABZADEH *et al.*, 2021; RASJID *et al.*, 2021).

2.4.1 Filtro de Savitsky-Golay

Com o objetivo de suavizar os ruídos presentes nas séries temporais tratadas no presente trabalho, optou-se pela aplicação do filtro Savitzky-Golay, o qual se configura como filtro digital de resposta finita ao impulso (ou *finite impulse response* - FIR), utilizado a fim de aumentar a precisão de um conjunto de dados discretos, sem distorcer a tendência do sinal. Tal ferramenta angariou popularidade em 1964, a partir da publicação do trabalho de Savitzky e Golay (1964), que visava a construção de um sistema de tratamento de ruídos e análise espectral de processos químicos. O método, fundamentado em procedimentos matemáticos bem estabelecidos, pode ser pensado como uma generalização do filtro de médias móveis e é desenvolvido a partir da combinação de duas características presentes nos filtros FIR: a equivalência entre a filtragem digital passa-baixa e a suavização de ruído por meio de ajuste po-

linomial. Esta abordagem consiste na interpolação de um conjunto de dados amostrados pelo método dos mínimos quadrados. O ajuste polinomial por mínimos quadrados é, em suma, efetuado pela convolução dos dados numéricos de entrada em uma janela de tamanho $2m+1$, $m \in R$, determinando o valor suavizado do ponto central do conjunto através de uma regressão polinomial de grau n, deslocando a janela ponto a ponto até que todo o conjunto de dados seja suavizado. A pormenorização matemática desse procedimento é fornecida em Savitsky e Golay (1964), bem como em Sadeghi, Behnia e Amiri (2020). A expressão do filtro pode ser descrita pela Equação 4.

$$y_j^* = \frac{\sum_{i=-m}^{i=m} c_i \times y_{j+i}}{2m+1}, \tag{4}$$

onde y refere-se à observação original; y_j^* é a observação filtrada; c_i é o coeficiente para a i-ésimo valor do filtro (no interior de uma janela de suavização); e $N = 2m+1$ refere-se ao tamanho da janela de suavização. Ainda, o índice móvel j é associado à suavização ordenada no interior de uma janela de suavização.

No entanto, a utilização do método de filtragem de Savitsky e Golay implica no uso de algum critério de determinação dos seus dois parâmetros de implementação, os quais são o tamanho da janela de dados e a ordem do polinômio de ajuste - ou ordem do filtro. Nesse sentido, a presente pesquisa apoiou-se no trabalho *On the Selection of Optimum Savitzky-Golay Filters*, dos autores Sunder Ram Krishnan e Chandra Sekhar Seelamantul (KRISHNAN; SEELAMANTULA, 2013). Em suma, o problema abordado pelo referido artigo é a escolha de um comprimento ou ordem de filtro SG que minimiza o erro quadrático médio mínimo (EQMM), bem como a preservação da estrutura temporal de um sinal variável no tempo. Para tanto, os autores elaboraram um método de compensação do problema de polarização-variância associado ao EQMM usando o estimador de risco imparcial de Stein, apresentado na forma de um algoritmo em (KRISHNAN; SEELAMANTULA, 2013).

De forma complementar, pode-se ressaltar a aplicação do filtro de Savitsky-Golay em trabalhos de previsão epidêmica concernente à pandemia de COVID-19, como em (RASJID *et al.*, 2021) e (ZHAN *et al.*, 2021).

2.4.2 Seleção de Preditores Exógenos

A análise exploratória dos dados consiste no estudo e investigação da base de dados. Por meio dessa análise, é realizada a inspeção qualitativa e estatística, a qual visa caracterizar a base de dados. Em se tratando de epidemiologia preditiva, as variáveis contidas na base de dados, acerca de uma doença, são a principal ferramenta de concepção de modelos de previsão ou identificação matemática das dinâmicas associadas a uma doença. Nesse sentido, podem ser utilizadas observações de outras séries temporais na construção do modelo preditivo de uma série histórica específica. Desse modo, diz-se que o modelo se utiliza de preditores auxiliares - ou preditores exógenos.

A etapa de análise de dados envolve também a seleção dos atributos, ou seja, a escolha criteriosa dos preditores exógenos da série temporal a ser identificada. Nesse sentido, deve-se utilizar alguma ferramenta que indique quais as variáveis que melhor explicam, do ponto de vista preditivo, a dinâmica de determinada variável dependente. Neste trabalho, a seleção das variáveis exógenas foi feita a partir do cálculo da correlação de Spearman.

2.4.2.1 Correlação de Spearman

Em estatística, o coeficiente de correlação de postos de Spearman, denotado como r_S, é um medida não paramétrica de correlação de postos (dependência estatística entre os postos de duas variáveis). Considere-se que os postos de uma variável corresponde ao conjunto ordenado, em ordem crescente, das observações da referida variável. Em síntese, a correlação de Spearman avalia o quão bem a relação entre duas variáveis pode ser descrita por meio de uma função monotônica (SPEARMAN, 1961).

A correlação de Spearman entre duas variáveis é obtida a partir do cálculo da correlação de Pearson entre os valores de postos dessas duas variáveis. Portanto, assim como na correlação de Pearson, o valor do coeficiente de correlação de Spearman pertence ao intervalo real $[-1, +1]$, de tal modo que a correlação de Spearman entre duas variáveis adquire:

- valor próximo da unidade, quando as observações tiverem uma classificação semelhante para ambas as variáveis;

- valores muitos próximos de -1, os quais indicam indicam correlação inversa entre os postos;
- valores muito próximos de +1, quando há relação direta entre os postos.

Ainda, uma correlação de Spearman perfeita (+1 ou -1) ocorre quando cada uma das variáveis é uma função monótona perfeita da outra. Ressalte-se que enquanto a correlação de Pearson avalia as relações lineares entre pares de variáveis, a correlação de Spearman avalia as relações monotônicas, as quais podem ser lineares ou não. Para variáveis não correlacionadas monotonicamente, o coeficiente de Spearman adquire valor nulo. Em síntese, a gradação do valores para coeficiente de Spearman é descrita na Tabela 1.

Tabela 1 – Classificação dos valores para o coeficiente de Spearman (SPEARMAN, 1961).

Valor absoluto	**Grau de correlação**
$r_S = 0$	ausência de correlação
$0 < r_S \leq 0,19$	muito fraca
$0,20 < r_S \leq 0,39$	fraca
$0,40 < r_S \leq 0,59$	moderada
$0,60 < r_S \leq 0,79$	forte
$0,80 < r_S \leq 1,00$	muito forte
$r_S = 1,00$	correlação monotônica

Supondo uma amostra de tamanho n, os n dados brutos X_i, Y_i, associados, respectivamente às séries temporais $X(t = 1, 2, \ldots, n)$ e $Y(t = 1, 2, \ldots, n)$. As séries históricas convertidas em postos podem ser denotadas como $\mathbf{R}(X)$, $\mathbf{R}(Y)$, onde $\mathbf{R}$ é o operador de conversão em postos, i.e., é o vetor ordenado das observações das referidas séries históricas. Tendo em vista que o coeficiente de correlação de Spearman é definido como o coeficiente de correlação de Pearson entre variáveis classificadas em postos, sua expressão é dada pela Equação (5).

$$r_S = \rho_{\mathbf{R}(X),\mathbf{R}(Y)} = \frac{\mathrm{cov}(\mathbf{R}(X), \mathbf{R}(Y))}{\rho_{\mathbf{R}(X)} \rho_{\mathbf{R}(Y)}}. \tag{5}$$

onde

- ρ denota a correlação de Pearson usual;

- cov é o operador de covariância;
- $\rho_{\mathbf{R}(X)}$ e $\rho_{\mathbf{R}(Y)}$ correspondem aos desvios padrão das séries convertidas em postos.

A utilização da correlação de Spearman para seleção de preditores exógenos, no presente trabalho, pode ser justificada a partir dos seguintes argumentos:

- É um método consagrado na literatura concernente à previsão de dados.
- O coeficiente de Spearman permite a quantificação de dependências monotônicas entre variáveis, diferentemente da correlação de Pearson, a qual lida apenas com dependências estatísticas lineares.
- É estatisticamente menos sensível aos efeitos de *outliers*, em comparação com o coeficiente de correlação de Pearson (PERNET, 2012).

2.5 MODELOS PARAMÉTRICOS PARA PREVISÃO DE SÉRIES TEMPORAIS

Para um modelo estocástico, quando o valor das observações futuras de dados ordenados temporalmente é regida por um modelo matemático, diz-se que a modelagem é paramétrica. Uma classe importante desse tipo de modelagem para descrição de séries temporais é a dos modelos estacionários, que são baseados na hipótese de que o processo permanece em equilíbrio em torno de um nível médio constante. Em outras palavras, o processo evolui no tempo de modo que a escolha de uma origem dos tempos não é importante, ou seja, as características de $Z(t+k)$, para todo k, são as mesmas de $Z(t)$, onde k é um deslocamento arbitrário ao longo do eixo do tempo (MORETTIN; TOLOI, 2004). Dessa forma, a média $\mu(t)$ e a variância $\sigma^2(t)$ de Z são constantes para todo $t \in T$, ou seja,

$$\mu(t) = \mathrm{E}[Z(t)] = \mu$$

e

$$\sigma^2(t) = \mathit{Var}[Z(t)] = \mathrm{E}[(Z(t)-\mu)^2] = \sigma^2.$$

Note-se que E denota o operador de esperança matemática. Mesmo quando uma série é não-estacionária, pode-se transformar os dados originais, a fim de tentar estacionarizá-la. O procedimento mais utilizado para esse fim consiste em diferenciar sucessivamente a série original, até se obter uma série estacionária. Diferenciar, aqui, significa considerar diferenças sucessivas da série original. Denotando-se z^{-1} como operador de atraso em tempo discreto, a primeira diferença de $Z(t)$ é definida por

$$\nabla Z(t) = [1 - z^{-1}]Z(t) = Z(t) - z^{-1}Z(t) = Z(t) - Z(t-1),$$

Já a segunda diferença é dada por

$$\nabla^2 Z(t) = [1-z^{-1}]^2 Z(t) = \nabla[\nabla Z(t)] = \nabla[Z(t)-Z(t-1)] = Z(t)-2Z(t-1)+Z(t-2).$$

Dessa forma, pode-se generalizar que *n*-ésima diferenciação de um série histórica pode ser expressa como

$$\nabla^n Z(t) = [1 - z^{-1}]^n Z(t) = \nabla[\nabla^{n-1} Z(t)].$$

Neste trabalho, serão abordadas três modelagens paramétricas, a mencionar:

- modelo autorregressivo integrado de médias móveis (ARIMA);
- modelo não linear autorregressivo de médias móveis (NARX)
- modelo não linear autorregressivo de médias móveis com entradas exógenas (NARMAX).

Modelos ARIMA são lineares e têm consagrada aplicação na literatura associada à análise de séries temporais, incluindo diversos trabalhos atinentes à pandemia de COVID-19 (SINGH, S. *et al.*, 2020), (SWARAJ *et al.*, 2021), (MOFTAKHAR; SEIF, 2020), (FANG *et al.*, 2020).

Já os modelos NARMAX utilizam mapeamentos não lineares e são, do ponto de vista científico, mais recentes, com aplicação bem sucedida em análise matemática de séries históricas. Até o momento presente, nenhum trabalho

brasileiro aplicou essa categoria de modelo no âmbito de séries epidemiológicas concernentes à pandemia de COVID-19.

Para a construção dos modelos ARIMA, George Box e Gwilym Jenkins sugeriram as seguintes etapas iterativas (PANKRATZ, 2009).

1. Coleta e pré-processamento de dados.
2. Escolha da representação do modelo.
3. Seleção de estrutura.
4. Estimação de parâmetros.
5. Validação do modelo.

Ressalte-se que esse procedimento tornou-se tradicional na identificação da dinâmica de sistemas e sua aplicação generaliza-se ao processo de concepção de modelos paramétricos e não paramétricos.

Na presente seção, serão descritas as principais bases teóricas dos dois modelos de identificação aplicados neste trabalho, para a previsão de dados epidêmicos em Santa Catarina. As seções Seção 2.5.1 e Seção 2.5.2 expõem, respectivamente, o detalhamento concernente aos modelos ARIMA e NARMAX.

2.5.1 Modelo Autorregressivo Integrado de Médias Móveis

Os modelos estocásticos ARIMA foram popularizados por George Box e Gwilym Jenkins no início dos anos 70, a partir da sistematização da informação necessária para entender e usar esses modelos para séries temporais monovariáveis (PANKRATZ, 2009). Esses modelos são robustos do ponto de vista conceitual e estatístico, proporcionam previsões probabilísticas e são de fácil implementação. De fato, como esse modelo representa uma generalização de diversos métodos de análise de séries temporais, bem como é consagrado na literatura científica concernente a essa temática, justifica-se a sua aplicação como técnica de base no presente trabalho.

Ressalte-se que o modelo ARIMA faz duas suposições acerca do comportamento da série temporal, sendo elas:

1. A série temporal corresponde a um processo estacionário.
2. A série temporal é um sistema dinâmico monovariável - ou seja, é apenas função do tempo.

Diversos autores optam por explicar, de maneira separada, cada componente que constitui o modelo ARIMA, da seguinte maneira.

(a) AR (p): componentes autorregressivas, utilizam a relação de dependência entre a observação corrente e as p observações pregressas da série;

(b) I (d): componente de integração que indica a quantidade d de diferenças a tempo discreto utilizadas para tornar estacionária a série temporal;

(c) MA (q): componentes de médias móveis, às quais utilizam a dependência entre uma observação e um erro residual de um modelo de média móvel centrado em q e aplicado às observações defasadas.

Tendo em vista as características supracitadas, modelos ARIMA fazem-se apropriados para descrever séries temporais não estacionárias homogêneas, ou seja, séries que, apesar de não progredirem em torno de uma média constante ao longo do tempo, quando diferenciadas d vezes, tornam-se estacionárias.

Sumariamente, a representação matemática de um modelo ARIMA é dada pela Equação (6).

$$y(t) = \varphi_1 y(t-1) + \cdots + \varphi_{p+d} y(t-p-d) + \zeta(t) - \theta_1 \zeta(t-1) - \cdots - \theta_q \zeta(t-q), \quad (6)$$

onde

- y é a d-ésima deferença da variável de interesse;
- φ_1, φ_1, $\ldots$, φ_p são os parâmetros autorregressivos;
- θ_1, θ_2, $\ldots$, θ_q são os parâmetros de média móvel;
- ζ corresponde ao ruído branco ou erro aleatório.

Ressalte-se que, no caso em houver característica sazonal associada à série histórica que se pretende modelar, há a necessidade de se considerar uma sazonalidade estocástica ajustada à série. Para tanto, o modelo a ser implementado denomina-se ARIMA com sazonalidade, ou SARIMA (*seasonal autoregressive integrated moving average*). Neste caso, o processo é referido como SARIMA $(p,d,q)(P,D,Q)_S$ pode ser escrito conforme a Equação (7) (PANKRATZ, 1983).

$$\varphi_p(z)\Phi_P(z^{-S})\nabla^d\nabla_s^D y(t) = \theta_q(z)\Theta_Q(z^{-S})\tau(t), \tag{7}$$

em que

- p, q representam a ordem da componente não sazonal do modelo, sendo d o número de diferenças para torná-la estacionária;
- P, Q representam a ordem da componente sazonal do modelo, sendo D o número de diferenças para tornar estacionária essa componente;
- S é o período da sazonalidade;
- $\varphi_p(z) = (1-\varphi_1 z^{-1}-\varphi_2 z^{-2}-\cdots-\varphi_p z^{-p})$ e $\theta_p(z) = (1-\theta_1 z^{-1}-\theta_2 z^{-2}-\cdots-\theta_p z^{-p})$ são, respectivamente, os operadores autorregressivo e de média móvel não sazonais;
- $\Phi_P(z) = (1-\Phi_S z^{-S}-\Phi_{2S} z^{-2S}-\cdots-\Phi_{PS} z^{-PS})$ e $\Theta_Q(z) = (1-\Theta_S z^{-S}-\Theta_{2S} z^{-2S}-\cdots-\Theta_{QS} z^{-QS})$ são, respectivamente, os operadores autorregressivo e de média móvel sazonais;
- ∇^d e ∇^D são, respectivamente, os operadores de diferença não sazonal e sazonal;
- $\tau(t)$ é a decomposição sazonal da série temporal.

Um detalhamento minucioso do modelo SARIMA é fornecido em Pankratz, 1983.

2.5.2 Modelo Não Linear Autorregressivo com Variáveis Exógenas

Haja vista o comportamento complexo das variáveis associadas a uma epidemia, faz-se pertinente identificá-las a partir de modelos cujas estruturações matemáticas são não lineares. Isso se deve, sobretudo, à habilidade com que representações dessa natureza identificam certos regimes dinâmicos que modelos lineares não são aptos a caracterizar. Nesse sentido, justifica-se a utilização do modelo não linear autorregressivo com variáveis exógenas (NARMAX), cujas bases de formulação são devidamente esmiuçadas em (AGUIRRE, 2004). O NARMAX, de forma geral, é um modelo autorregressivo munido de média móvel, cuja variante MISO (*multiple input, single output*) é expressa matematicamente pela Equação (8).

$$\begin{aligned} y(t) = F^l[&y(t-1), y(t-2), ..., y(t-n_y), u_1(t-1), \\ &u_1(t-2), ..., u_1(t-n_{u_1}), u_2(t-1), u_2(t-2), \\ &..., u_2(t-n_{u_2}), ..., u_n(t-1), u_n(t-2), ..., \\ &u_n(t-n_{u_n}), e(t-1), e(t-2), ..., e(t-n_e)] + e(t), \end{aligned} \tag{8}$$

em que F^l representa uma função não-linear geral munida de grau de não-linearidade l. As variáveis $y(t)$ e $e(t)$ são, respectivamente, a saída e o ruído aditivo do sistema, cujos atrasos máximos são representados por n_y e n_e. Similarmente, as n entradas exógenas utilizadas para conformação do modelo são denotadas por $u_i(t)$, com atrasos máximos denotados por n_{u_i}, de forma que i=1,...,n. No caso do presente estudo, F^l corresponde a uma expansão polinomial com grau de não-linearidade l. Adicionalmente, este trabalho também considera casos em que as séries temporais epidemiológicas não possuem atraso puro de tempo e que nenhum dos parâmetros a ser estimado depende de $e(t)$. A esse tipo de modelo dá-se a denominação de NARX - caso especial da modelagem NARMAX que não se utiliza de fatores associados a médias móveis. Uma vez que tanto o NARMAX quanto o NARX de mapeamento polinomial produzem uma estrutura linear nos parâmetros, a construção computacional desses modelos é facilitada. A estimação de parâmetros é concebida por meio de aplicação de algoritmo baseado em mínimos quadrados (MQ), o qual confere minimização dos erros de previsão. Adicionalmente, o nível de importância dos

termos regressivos utilizado para conformar a expressão matemática descrita pela Equação (8) é determinado por meio do parâmetro ERR (*error reduction ratio*), o qual representa quantitativamente a capacidade de cada termo regressivo explicar a variância do sinal de saída identificado. Já a quantidade de termos polinomiais utilizados para compor o modelo é determinada pelo AIC.

2.5.2.1 Estimação de Parâmetros

2.5.2.1.1 Polarização e Variância

O problema crucial de identificação é definir uma função f que represente uma boa aproximação do sistema, a partir de sua aplicação sobre uma massa de dados de treinamento. Supondo que o verdadeiro sistema pode ser representado pela função de aproximação $\hat{f}$, de modo que

$$y \sim \hat{f}(\mathbf{x}, \hat{\theta}),$$

onde $\mathbf{x}$ é denotado como o conjunto de termos regressivos e $\hat{\theta}$ corresponde ao conjunto de parâmetros. A eficácia do modelo proposto, $\hat{f}(\mathbf{x}, \hat{\theta})$, pode ser definida como a sua capacidade de generalização. Para quantificar essa generalização, pode-se utilizar a métrica do erro quadrático médio (MSE) (Equação (9)), calculado sobre um conjunto de dados de validação.

$$\text{MSE} = \text{E}[(y - \hat{f}(\mathbf{x}, \hat{\theta})^2], \tag{9}$$

em que E[$\cdot$] é o operador correspondente à esperança matemática. Como mostrado em Geman et al. (1992), o MSE pode ser representado a partir da decomposição da polarização e da variância, segundo a Equação (10).

$$\text{MSE} = \text{Var}(\hat{f}) + [\text{Bias}(\hat{f})]^2 + \text{Var}(\varepsilon), \tag{10}$$

onde Var(ε) é a variância do ruído e Bias é o operador de polarização. Se, em média, $\hat{f}$ é diferente de E[y], então $\hat{f}$ é considerado um estimador polarizado de E[y]. Para minimizar o MSE, é requerida a minimização dos termos de polarização e variância, o que é, na maioria das situações práticas, uma tarefa conflitante de otimização. A variância do ruído não pode ser reduzida, haja vista

que é independente do modelo e de seus parâmetros. Digamos, por exemplo, que a previsão de saída não depende das variáveis de entrada. Ao negligenciar os dados de entrada, a saída pode ser definida para um valor constante, independentemente dos valores de entrada. Nesse caso, a variância do estimador será definitivamente minimizada, mas a saída será, provavelmente, muito diferente da função real, ou seja, a polarização será excessiva. Por outro lado, se o modelo projetado, $\hat{f}$, é muito flexível, capaz de interpolar perfeitamente o conjunto de dados de treinamento, a polarização do estimador será reduzida e, no entanto, o modelo sofrerá com um grande aumento na variância.

A polarização descreve o quão longe o modelo está, em média, do verdadeiro processo. Já a variância representa o quanto a previsão do modelo varia entre as realizações. O ajuste equilibrado de polarização-variância é, em síntese, a negociação entre superajuste e subajuste. Alto valor de polarização pode fazer com que o modelo não siga a relação de entrada-saída do sistema original, enquanto a considerável variação presente nos dados causa superajuste, onde o modelo estimado tende a modelar o ruído em vez da saída.

De fato, definir a complexidade do modelo pode ser uma tarefa desafiadora, visto que não há uma maneira analítica para determinar o tamanho ideal do modelo. Uma solução possível para estimar a complexidade do modelo dá-se pelo emprego de técnicas de validação cruzada, onde o tamanho do modelo é escolhido a fim de minimizar o erro de previsão em dados ainda não vistos pelo modelo de previsão.

2.5.2.1.2 Método dos Mínimos Quadrados Ortogonais

Uma vez que a estrutura do modelo é definida, faz-se necessário estimar os parâmetros associados aos seus termos. O estimador de parâmetros mais utilizado para modelos de identificação de sistemas lineares nos parâmetros é o Métodos dos Mínimos Quadrados Ordinário (MQO). Uma base teórica aprofundada concernente a tal modelo pode ser, assim como uma implementação algorítmica dessa ferramenta, podem ser encontradas na obra de Aguirre (2004).

2.5.2.2 *Error Reduction Ratio*

Seja $\mathcal{R} = \{\psi_1, \psi_2, \ldots, \psi_m\}$ o conjunto de m termos regressivos da estrutura de um modelo NARMAX. Em suma, o problema de seleção da estrutura do modelo é escolher, a partir dos elementos de $\mathcal{R}$, um subconjunto de termos regressivos para compor o modelo final. Para isso, um critério amplamente utilizado é o *error reduction ratio* (ERR) (BILLINGS; CHEN, a. K., 1989), o qual quantifica a redução na variância dos resíduos, a partir da adição de um novo termo regressivo na estrutura do modelo, normalizado em relação à variação de saída. O ERR devido à inclusão do i-ésimo regressor no modelo pode ser escrito segundo a Equação (11).

$$[\text{ERR}_1]_i = \frac{\text{MS1PE}(\mathcal{M}_{i-1}) - \text{MS1PE}(\mathcal{M}_i)}{\langle \mathbf{y}, \mathbf{y} \rangle}, \tag{11}$$

para $i = 1, 2, \ldots, m$, onde $\text{MS1PE}(\mathcal{M}_1)$ refere-se ao erro quadrático médio de predição de um passo à frente para o modelo com i termos regressivos; m é o número de regressores candidatos; e $\mathcal{M}$ representa uma família de modelos munida de estruturas aninhadas e, portanto, $\mathcal{M}_{i-1} \subset \mathcal{M}_i$. Na Equação (11), o numerador fornece o valor numérico da redução da variância dos resíduos devido à inclusão do i-ésimo regressor. Já o denominador corresponde aos dados da variância. Como vantagem, o ERR_1 pode ser representado de forma compacta, conforme a Equação (12) (BILLINGS; CHEN, a. K., 1989).

$$[\text{ERR}_1]_i = \frac{\hat{g}_i \langle \mathbf{w}_i, \mathbf{w}_i \rangle}{\langle \mathbf{y}, \mathbf{y} \rangle}, \tag{12}$$

onde $\mathbf{w}_i$ é o i-ésimo termo regressivo ortogonal e $\hat{g}_i$ é o parâmetro estimado correspondente. Assim, a cada etapa, o termo com o maior ERR_1 é adicionado ao modelo.

O objetivo de usar o critério ERR é tornar o modelo apto a organizar um conjunto de regressores candidatos em ordem decrescente de relevância. A quantidade de termos utilizada para compor o modelo, ou seja, o ponto de corte em que o ERR deixará de incluir termos no modelo, pode ser determinado por meio da aplicação de outros critérios complementares, a exemplo dos critérios de informação (AGUIRRE, 2004). Para esse fim, o presente trabalho utilizou-se

do critério de informação de Akaike e Rissanen (AIC), o qual será detalhado, de forma sucinta, a seguir.

2.5.2.3 Critério de Informação de Akaike

No âmbito da estruturação matemática do modelo, é desejável encontrar uma representação otimizada em termos da polarização e da variância contida nos dados. Determinar um modelo que atenda de forma adequada esses requisitos é, na maioria das vezes, uma tarefa difícil, visto que a função de erro do modelo é, comumente, desconhecida. O modelo deve ser rico o suficiente para capturar e simular o comportamento do sistemas, mas não flexível ao ponto de incorrer em sobreparametrização. Se o modelo for muito simples, munido de quantidade pequena de parâmetros, provavelmente não generalizará bem os dados de treinamento. Por outro lado, se houver um grande conjunto de parâmetros, a representação não se adequará a novos dados. Equilibrar o modelo correto entre polarização e variância, ou seja, escolher a complexidade do modelo, deve ser realizada de alguma forma (JAMES *et al.*, 2013).

Uma forma de avaliar o desempenho do modelo dá-se pela análise da função de verossimilhança. O princípio da máxima verossimilhança seleciona os parâmetros do modelo que tornam os dados mais condizentes com o sistema real. Se a complexidade do modelo for aumentada, a função de probabilidade também será aumentada, pois os dados serão mais prováveis de constarem no modelo a partir dos parâmetros definidos. A Estimativa de Máxima Verossimilhança é uma técnica que lida com o aspecto de polarização da função do erro médio quadrático e, portanto, o efeito de superajuste não é contabilizado. Para lidar com as consequências referentes ao aumento da variância advinda da flexibilização do modelo, pode-se penalizar os modelos que se mostram mais complexos. Em (AKAIKE, 1974), o autor propõe um critério para comparar e selecionar modelos, a partir de um grupo de candidatos. A ideia geral é apresentada a seguir.

Considere duas funções de verossimilhança, sendo f a verdadeira função de verossimilhança, e f^* a função de verossimilhança estimada, obtida a partir dos dados. Para comparar essas funções de distribuição de probabilidade, uma

das maneiras mais comuns é usar a Divergência de Kullback-Leibler, expressa pela Equação (13).

$$D(p, q) = \sum_x p \log \frac{p}{q}, \tag{13}$$

onde p e q são distribuições sobre os dados x. Seja $f_{j\theta^*}$ a distribuição estimada para um modelo particular j. Considerando-se a distribuição verdadeira e a distribuição estimada, a função de perda é dada pela Equação (14),

$$\text{loss} = D(f, f_{j\theta^*}). \tag{14}$$

Considerando que $f_{j\theta^*}$ é dependente dos dados, para que ela se aproxime de f, a esperança da função de perda, expressa por $E_f[D(f, j\theta^*)]$, deve ser minimizada. Dessa forma, se o modelo superajustar os dados, $f_{j\theta^*}$ não se aproximará de f e o modelo é penalizado. Da mesma forma, se o modelo tiver um desempenho inferior nos dados de treinamento Y, as duas funções de distribuição serão diferentes e o modelo também será penalizado. Portanto, esse procedimento leva em conta a polarização e a variância presente nos dados.

Pode-se comprovar que a divergência entre f e $f_{j\theta^*}$ é expressa pela Equação (15).

$$D(f, f_{j\theta^*}) = c - A(f, f_{j\theta^*}), \tag{15}$$

onde

$$c = \sum_y f(Y) \log(f(Y))$$

é a entropia e

$$A(f, f_{j\theta^*}) = \sum_y f(Y) \log(f_{j\theta^*}(Y)).$$

A fim de minimizar a divergência, a esperança $\mathrm{E}_f[D(f, j\theta^*)]$ deve otimizar a expressão $A(f, f_{j\theta^*})$. O desafio concernente à minimização de $A(f, f_{j\theta^*})$ é que a verdadeira distribuição, f, é desconhecida. Para solucionar tal problema, é necessário estimar f a partir de $f_{j\theta^*}(Y)$, o que leva a uma estimativa enviesada. A magnitude da polarização é proporcional à quantidade k de parâmetros do modelo. Em resumo, o Critério de Informação de Akaike (AIC) é representado pela probabilidade $\mathcal{L}$, corrigida pela polarização k, conforme a Equação (16).

$$\mathrm{E}_f[D(f, j\theta^*)] \approx \log(\mathcal{L}(\theta^*|Y)) - k. \tag{16}$$

Considerando o caso de máxima verossimilhança, a definição do critério de informação de Akaike é estabelecida pela Equação (17).

$$\text{AIC} = -2\log(\mathcal{L}(\theta^*|x)) + 2k. \tag{17}$$

Quando o tamanho da amostra é pequeno, eleva-se a probabilidade de o AIC selecionar modelos com grande conjunto de parâmetros, o que leva ao superajuste do modelo. Desse modo, uma versão corrigida do critério, levando em consideração o tamanho da amostra dos dados, denotado como AICc, foi desenvolvido por (CAVANAUGH, 1997), e é expresso pela Equação (18).

$$\text{AICc} = \text{AIC} + \frac{2k^2 + 2k}{n - k - 1}. \tag{18}$$

Por meio da aplicação de AIC (e AICc), os modelos podem ser comparados e classificados considerando a otimização das características de polarização e variância. Portanto, modelos com AIC reduzido devem ser selecionados, haja vista que atendem melhor tal critério.

2.5.3 Validação do modelo

2.5.3.1 Métricas de Desempenho

A fim de avaliar o desempenho de dado modelo de de identificação, é necessário aplicar ferramentas matemáticas que determinam o quão bem as suas predições se aproximam dos dados observados. A habilidade preditiva dos modelos concebidos por este trabalho foi avaliada com base no cálculo de métricas tradicionais de avaliação de desempenho preditivo (FORATTINI, 1996a), são elas: Erro Médio Absoluto (MAE), Erro Percentual Médio Absoluto (MAPE), Raiz do Erro Quadrático Médio (RMSE) e Raiz do Erro Quadrático Relativo (RRSE).

O MAE mede a magnitude média dos erros em um conjunto de previsões, sem considerar sua direção. É a média sobre a amostra de teste das diferenças absolutas entre a previsão e a observação real, onde todas as diferenças individuais possuem igual peso.

O MAPE indica percentualmente quanto, em média, o preditor erra, sem compensar erros negativos com erros positivos. A partir dessa métrica, as medições são tomadas sobre o desvio absolutos das previsões em relações aos

dados reais observados e, usualmente, expressa a eficiência do preditor como uma porcentagem.

Já a raiz do erro quadrático médio (RMSE) de um estimador, calcula a raiz quadrada da média dos quadrados dos erros - ou seja, a raiz quadrada da diferença quadrática média entre os valores estimados e o que é estimado. Essa métrica consiste em uma função de risco, correspondente ao valor esperado da raiz da perda quadrática do erro. O fato de o RMSE ser quase sempre estritamente positivo (e não zero) é devido à aleatoriedade ou ao fato de o estimador não levar em conta as informações que poderiam produzir uma estimativa mais precisa.

O RRSE é uma métrica estatística relativa à eficiência de predição caso um preditor simples tivesse sido aplicado sobre os dados. Esse preditor simples, mais especificamente, é apenas a média dos valores reais dos dados. Assim, o erro quadrático relativo, calculado pelo RRSE, calcula o erro quadrático total e o normaliza por meio da divisão pelo erro quadrático total do preditor simples. A variação numérica dessa métrica está contida no intervalo real positivo.

As expressões matemáticas para cada uma das métricas são descritas pelas Equações (19) – (22).

$$\text{MAE} = \sum_{t=1}^{n} \frac{|y(t)-\hat{y}(t)|}{n}, \tag{19}$$

$$\text{MAPE} = \frac{1}{n} \sum_{t=1}^{n} \left|\frac{y(t)-\hat{y}(t)}{y(t)}\right|, \tag{20}$$

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^{n} (y(t) - \hat{y}(t))^2}, \tag{21}$$

$$\text{RRSE} = \sqrt{\sum_{t=1}^{n} \frac{y(t)-\hat{y}(t)}{\overline{y}-\hat{y}(t)}}. \tag{22}$$

Onde

- $y(t)$, $\hat{y}(t)$, $\overline{y}(t)$ referem-se, respectivamente, aos sinais de saída real, previsto pelo modelo de identificação não-linear e previsto por um preditor simples baseado na média aritmética;

- n indica a quantidade de observações utilizadas para o cálculo das métricas.

Para modelos de identificação ideais, as métricas supracitadas devem adquirir valor nulo. Logo, em cenários práticos de identificação de sistemas, modelos preditivos eficientes possuem valores bastante reduzidos para tais parâmetros.

2.5.3.2 Conjunto de Dados de Validação

Numa atividade de previsão de séries temporais ou identificação de sistemas, é preponderante estimar o erro de teste associado ao modelo de regressão para um conjunto de dados observados. A técnica do conjunto de dados de validação é uma estratégia simples para abordar essa problemática. Esse mecanismo consiste em dividir os dados observados em duas partes, um conjunto de treinamento e um conjunto de validação. O modelo é efetivamente concebido a partir de sua aplicação sobre o conjunto de treinamento. O modelo construído é, então, utilizado para efetuar previsões no conjunto de dados de validação, os quais são as informações ainda não vistas pelo modelo. Uma métrica de desempenho aplicada ao conjunto de validação fornece uma estimativa da taxa de erro esperada em um conjunto de teste (novos dados), a fim de fornecer informações sobre a qualidade do modelo. Uma possível desvantagem dessa técnica refere-se ao fato de que a estimativa de erro de validação pode ser consideravelmente variável, dependendo de como os dados são divididos e quais observações são incluídas

Para análises mais rigorosas quanto à qualidade dos modelos, é pertinente que um procedimento de validação mais refinado seja utilizado, a exemplo da validação cruzada.

2.5.3.3 Janela Deslizante

A técnica de validação por janelas deslizantes segue o mesmo princípio do mecanismo de validação apresentado na Seção 2.5.3.2. Para os dados de séries temporais, onde a dimensão temporal das observações tem crucial relevância, não se pode dividi-las aleatoriamente em grupos. Em vez disso,

deve-se particionar os dados de forma a respeitar a ordem temporal em que os valores foram observados. Na previsão de séries temporais, essa avaliação de modelos em dados históricos pode ser realizada por meio da técnica de janelas deslizantes.

A técnica das janelas deslizantes se utiliza de um padrão de divisão sincronizado dos conjuntos de dados de treino e validação, por meio de subconjuntos adequados da amostra total. Em suma, neste procedimento se utiliza um intervalo de dados de tamanho pré-determinado - a janela propriamente dita - a qual se constitui de uma parcela de dados para treinamento, onde será identificado o modelo preditivo; e um parcela de dados destinada à validação, a qual configura o conjunto em que o modelo criado realizará predições. Essa janela se movimentará, respeitando a disposição cronológica dos dados, a um passo de tamanho fixo. No âmbito da validação de modelos preditivos por meio do avanço de janelas deslizantes, há duas abordagens de interesse:

- Janela deslizante de avanço não ancorado: quando um ciclo de predição é efetuado, à medida em que a janela move-se de acordo com o tamanho do passo, faz-se a exclusão de um intervalo de dados do tamanho do passo movimentação no início da janela e é adicionado um conjunto de dados do tamanho do passo ao final da janela. Nesse cenário, o conjunto de treinamento possui tamanho fixo e tem-se a vantagem de o modelo adaptar-se melhor à ordem temporal dos dedos, refletindo seus aspectos sazonais (KIRKPATRICK II; DAHLQUIST, 2010).

- Janela deslizante de avanço ancorado: quando uma etapa de predição é finalizada, adiciona-se ao final da janela um conjunto de dados de mesma magnitude do passo de movimentação, não havendo exclusão dos dados do início da janela. Desse modo, há sempre utilização de dados antigos na conformação dos modelos preditivos.

No presente estudo, foi utilizada a janela deslizante de avanço não ancorado, haja vista que são utilizada séries temporais, cujo caráter cronológicos dos dados é preponderante. Desse modo, tal abordagem ajusta-se à evolução temporal das séries. Adicionalmente, a janela sem ancoramento é mais interes-

sante para as análises, visto que não sofre influência de dados muito antigos associados às informações epidêmicas.

2.6 ESTADO DA ARTE DOS MODELOS DE PREVISÃO EPIDEMIOLÓGICA

A previsão do número de casos de infecção associados a uma doença é crucial ao planejamento apropriado de políticas públicas de prevenção e alocação de recursos. Nesse contexto, técnicas de previsão tradicionais de séries temporais, a exemplo do ARIMA, bem como modelos baseados em aprendizado de máquina, têm sido consideravelmente exploradas no contexto da pandemia de COVID-19. Por conseguinte, tais redes neurais têm sido utilizadas em várias aplicações associadas a projeções futuras de dados dispostos cronologicamente (KEELING; GRENFELL, 2000; KEELING; ROHANI, 2002). Adicionalmente, diversos países têm empregado, como ferramentas auxiliares de combate a epidemias, *softwares* de aprendizado de máquina na estimação de quantidades futuras de infecção, bem como da trajetória espacial dos indivíduos acometidos pela doença. Como contraponto às abordagens preditivas supracitadas, há métodos fundamentados na identificação matemática de sistemas não-lineares, sendo que o presente trabalho concentrou-se no estudo de técnicas auto-regressivas. No que se refere aos modelos estatísticos autorregressivos lineares, a técnica ARIMA (*autoregressive integrated moving average*) é referência na literatura relativa à previsão de séries temporais, inclusive no âmbito de análise epidemiológica. Complementarmente, será enfatizada a modelagem de caráter não linear fundamentada na técnica NARMAX e sua variante NARX, implementadas no presente estudo. Tais modelos já foram aplicados com sucesso em matéria de epidemiologia matemática, mas, no entanto, têm sido pouco explorado no âmbito da pandemia de COVID-19, no cenário brasileiro.

2.6.1 Modelos Autorregressivos de Identificação Linear

Em Singh et al., (2020), desenvolveu-se um modelo ARIMA de previsão para casos diários de COVID-19, na Malásia, com base em casos observados no período que compreende o dia 22 de janeiro a 31 de março de 2020. A partir da análise de diversos parâmetros epidêmicos da doença, o modelo de

predição mais eficiente foi selecionado, fundamentando-se num subconjunto desses parâmetros. O modelo ARIMA (0,1,0) produziu o melhor ajuste aos dados observados, com um valor de erro médio absoluto (MAPE) de 16,01 e um valor de *Bayes Information Criteria* (BIC) de 4,170. Os valores previstos mostraram uma tendência decrescente de casos COVID-19 até 1 de maio de 2020. Os casos observados durante o período de previsão foram previstos com precisão e foram colocados nos intervalos de previsão gerados pelo modelo ajustado.

Um modelo híbrido, o qual integra o modelo ARIMA e uma rede neural autorregressiva não linear (NAR), é proposto em Swaraj et al., (2020). Nesse trabalho, os modelos ARIMA são usados para extrair as correlações lineares, ao passo que a operação da rede neural objetiva modelar os resíduos gerados pelo modelo ARIMA, adicionando componentes não lineares dos dados. Assim, são efetuadas comparações com base em parâmetros de avaliação de desempenho entre a forma pura do modelo ARIMA, o modelo ARIMA-NAR e algumas outras técnicas preditivas existentes, os quais foram aplicados nos dados associados à pandemia de COVID-19 em diferentes países. A combinação híbrida apresentou redução significativa nos valores de RMSE (16,23 %), MAE (37,89 %) e MAPE (39,53 %) quando comparada com o modelo ARIMA para casos diários de infecção. Resultados semelhantes, com porcentagens de erro reduzidas, foram encontrados para mortes relatadas diariamente, bem como casos de recuperação. Os resultados sugeriram a eficácia do novo modelo híbrido em relação ao modelo ARIMA puro na captura dos padrões lineares e não lineares dos dados COVID-19.

Em Moftakhar e Seif, (2020), utilizou-se o ARIMA na previsão dos dados pandêmicos referentes ao Irã. Os dados usados nesse estudo foram obtidos a partir de relatórios diários do Ministério da Saúde iraniano, assim como conjuntos de dados fornecidos pela Universidade Johns Hopkins, incluindo o número de novos casos infectados desde 19 de fevereiro de 2020 até o dia 21 de março de 2020. A partir de um horizonte de previsão de 30 dias, o modelo ARIMA previu um aumento exponencial no número de novos pacientes detectados. O resultado deste estudo também mostra que se o padrão de disseminação continuar o mesmo de antes, o número de novos casos diários seria de 3574 até a

data final de previsão.

Em Fang et al., (2020), modelos ARIMA foram desenvolvidos para prever os casos cumulativos confirmados, mortos e recuperados, na Rússia. Os modelos ARIMA (2,2,1), ARIMA (3,2,0) e ARIMA (0,2,1) foram obtidos para previsão de casos confirmados, óbitos e casos de recuperação, respectivamente. Os testes computacionais resultaram em percentuais de erro médio absoluto (MAPE) de 0,6, 3,9 e 2,4, respectivamente. Esse trabalho indica que a capacidade do sistema de saúde da Rússia pode responder com eficácia à pandemia COVID-19, considerando que os três modelos concebidos têm um bom efeito de ajuste e podem ser usados para previsões de curto prazo.

2.6.2 Modelos Autorregressivos de Identificação Não-Linear

Em Nepomuceno et al., (2006), é proposta a obtenção de um modelo NARMAX polinomial para previsão dos parâmetros de uma epidemia, de forma on-line, a fim de se aplicar um método de controle sobre os parâmetros identificados. Para tanto, a técnica de controle implementada nesse trabalho é chamada de Proporcional-Narmax (ou P-NARMAX), uma vez que atua como um controle proporcional em malha fechada. Em síntese, tal modelo foi utilizado para o cálculo da probabilidade de um indivíduo ser infectado, de acordo com o número de infectados e pessoas suscetíveis à infecção num cenário epidêmico. Baseado no valor de probabilidade previsto, foi possível aplicar a ação de controle. Uma vantagem do método proposto é que é necessário conhecer apenas os dados estatísticos sobre a epidemia (número de indivíduos suscetíveis, infectados e número total de indivíduos) não sendo obrigatório conhecer os parâmetros inerentes da epidemia.

Em Billings e Wei, (2019), é apresentado o uso do modelo NARMAX no âmbito da gripe ocasionada pelo vírus influenza. Em primeira análise, um simples modelo foi estabelecido para representar relações entre a taxa de mortalidade semanal e a doença em países como Inglaterra e País de Gales. Adicionalmente, outros nove modelos foram construídos a partir de séries temporais associadas à doença, os quais revelam como os sintomas gerais da gripe, em regiões dos Estados Unidos, estão relacionados temporalmente com os sinto-

mas vistos em outras partes do mundo. Os resultados obtidos pela pesquisa sugerem que a metodologia empregada com o uso do NARMAX representa uma ferramenta de análise de fenômenos clínicos, seja na previsão epidêmica, seja na interpretação e compreensão da relação entre as diversas variáveis utilizada na conformação do modelo.

É descrito, em Rahim et al., (2007), a aplicação de modelos NARMAX no diagnóstico de infecção por dengue. O sistema desenvolvido baseia sua previsão exclusivamente nos parâmetros de bioimpedância e dados fisiológicos dos pacientes, tais quais sexo, peso, presença de sintomas como o vômito e o dia de origem da febre. São considerados, na construção dos modelos, três critérios de seleção de estrutura diferentes, nomeadamente FPE, AIC e Lipschitz. A pesquisa foi divida em duas abordagens, as quais são a abordagem não regularizada e a abordagem regularizada. Os resultados mostram que o uso do critério de Lipschitz com abordagem regularizada corresponde aos melhores desempenhos referentes ao diagnóstico da doença, com índices de desempenho de 88,40%. Além disso, a análise mostra que o modelo NARMAX produz melhor precisão em comparação com o modelo de média móvel autorregressiva com entrada exógena (ARMAX) no sistema de diagnóstico.

A utilização de métodos preditivos não lineares, em comparação com abordagens lineares, na previsão de hipóxia ocasionada em passageiros em voos de avião, é avaliada em Billings et al., (2013). Entre os métodos não lineares utilizados na pesquisa, menciona-se o NARMAX. Os autores concluem que esse modelo opera com melhor desempenho, sobretudo quando se há acesso a maiores conjuntos de dados. Por fim, o trabalho reitera que estudos maiores acerca dos modelos autorregressivos de identificação permitiria o desenvolvimento de equações preditivas ainda mais eficientes.

Um sistema de modelagem de doenças cardíacas, implementado com base em sinais sonoros provenientes do coração, é proposto por Shamsuddin e Taib (2011). O sistema preditivio utiliza o modelo ARX como vetor de regressão e redes neurais para estruturação de modelos não lineares gerando, assim, um modelo ARX não linear. O modelo resultante, fundamentado na união entre o paradigma de aprendizado de máquina e modelagem autorregressiva, ajusta-se com excelência no mapeamento dos sinais cardíacos originais com R-quadrado

médio acima de 99,9%. Por fim, os parâmetros de peso dos algoritmos neurais foram estimados e analisados com o propósito de classificação de doenças cardíacas.

Lauraitis e Maskeliūnas (2017), apresentam um modelo para prever o nível de capacidade funcional para pessoas com disfunções motoras, intimamente relacionadas com as síndromes de Huntington ou Parkinson. O modelo proposto é concebido aplicando-se técnicas de aprendizagem supervisionada e modelos de previsão estatística - incluindo o NARX. Em síntese, os resultados atinentes à modelagem NARX sugerem um ótimo desempenho de predição, com valores não inferiores a 0,99 para a métrica de desempenho R-quadrado, bem como valores de erros médios quadráticos não superiores a 0,11.

2.6.3 Modelos Baseados em Aprendizado de Máquina

Um estudo comparativo entre algoritmos de aprendizado de máquina, no que concerne à previsão de dados de COVID-19, foi realizado em Ardabili et al., (2020). Os autores desse trabalho utilizaram-se de implementações de computação evolucionária, a exemplo do algoritmo genético, *Particle Swarm Optimization* (PSO) e *Gray Wolf Optimizer*, bem como modelos mais populares, como o *Multilayer Perceptron* (MLP) e *adaptative network-based fuzzy inference system* (ANFIS). O desempenho dessas técnicas foi avaliado no âmbito da capacidade previsão de dados sob distintos horizontes de tempo, atinentes a cinco países, sendo que as técnicas MLP e ANFIS produziram os melhores resultados, alcançando níveis de correlação muito próximos da unidade.

Kapoor et al., (2020), uma nova abordagem que se utiliza da ferramenta *Google Research*, a qual combina dados temporais e espaciais, é proposta. Nesse estudo, redes neurais para grafos, ou *Graph Neural Networks* (GNN), foram usados em conjunto com dados de mobilidade espacial fornecidos pelo *Google*, a fim de desvendar as diversas interações espaçotemporais presentes na disseminação da pandemia. Os resultados provenientes de experimentos numéricos indicam que a união entre tais recursos mostrou-se poderosa na previsão de relações entre tempo e espaço atinentes à pandemia de COVID-19, sendo que a técnica GNN alcançou correlações de valor 0,998.

Melin et al., (2020), empregou-se uma rede neural empilhada, constituída por três módulos, para efetuar previsões de dados relativos à pandemia de COVID-19 no México, num período de 110 dias. As previsões realizadas pelos módulos desse modelo de rede são combinadas por meio de integradores do tipo *fuzzy*, designados para lidar com incertezas, a fim de gerar uma única saída de rede. Os experimentos, que consistiram na previsão de casos de infecção e óbitos visando um horizonte de dez dias, indicam que a técnica supera modelos simples de redes neurais munidas de uma saída. O índice RMSE obtido para a previsão de casos de infecção confirmados no México é de 0,0808, considerando que tal parâmetro varia de 0,0322 a 0,2157, para os estados desse país. Já o índice RMSE resultante da predição do número de óbitos no país é de 0,0914, variando, nos estados mexicanos, de 0,0175 a 0,2094.

Arora et al., (2020), foram utilizadas três variantes da rede neural LSTM para produzir as predições epidêmicas. Nesse sentido, o trabalho implementa os modelos bidirecional, convolucional e empilhado, com o objetivo de prever casos de infecção em horizontes de um dia e uma semana. A modalidade bidirecional da rede apresentou os melhores desempenhos preditivos quanto à métrica de desempenho adotada (MAPE), sendo que os resultados menos satisfatórios referem-se ao modelo empilhado de rede. Os valores resultantes para a métrica de desempenho são de 5,05%, 4,81%, e 3,22%, respectivamente para os modelos empilhado, convolucional e bidirecional das redes construídas.

Diversos algoritmos baseados em AM são comparados em da Silva et al., (2020), para previsão de casos futuros de infecção por COVID-19, no Brasil e nos Estados Unidos. Inclui-se, nesse estudo, o uso de Redes Neurais Bayesianas, Regressão cúbica, *k-nearest neighbors*, *Random Forest* e *Support Vector Machine*, aplicadas com o uso de pré-processamento de dados do tipo *variational mode decomposition* (VMD), assim como entradas exógenas associadas às variáveis de temperatura e precipitação. Os melhores resultados apresentados não indicaram com clareza o modelo mais eficaz. Os experimentos mais promissores resultaram em erros de predição de aproximadamente 3%, bem como sugerem que a utilização do VMD, na etapa de pré-processamento, aprimora o desempenho das redes neurais.

Uma comparação entre modelos baseados em AM e abordagens estatísticas foi realizada em Shahid et al., (2020). Especificamente, utilizou-se o modelo estatístico ARIMA como contraponto às redes neurais do tipo LSTM e SVR, a fim de prever dados acerca da quantidade de casos de infecção, número de óbitos e quantidade de pacientes recuperados, referentes ao cenário pandêmico de dez países distintos. Obteve-se os modelos finais por meio do treinamento sobre os dados dos últimos 110 dias, para previsão das quantidades atinentes aos próximos 48 dias, num período de testes que compreende o mês janeiro ao mês junho do ano de 2020. A pesquisa indicou superioridade da rede neural LSTM, por meio da obtenção de valores de erro médio absoluto de 2,0463 e 0,0095, no âmbito da previsão das quantidades de casos de infecção confirmados e óbitos, respectivamente.

Baseando-se nos dados de um período de 56 dias, Rustam et al., (2020), realizou a previsão das taxas de contágio, óbitos e casos de recuperação dos próximos 10 dias, no âmbito global. Para tanto, utilizou-se como modelos de previsão os métodos estatísticos de regressão linear, a regressão LASSO e o amaciamento exponencial. Tais modelos desempenharam de forma mais precisa que a técnica SVR, representante do paradigma de aprendizado de máquina, sendo que o modelo de suavização exponencial garantiu os melhores resultados na predição da taxa de óbitos e a regressão LASSO configurou-se como a mais eficiente na previsão da taxa de infecção.

Direkoglu et al., (2020), aplicou-se redes neurais recorrentes da variante LSTM para prever a quantidade de infecções e mortes no leste europeu, na China, assim como as quantidades referentes a todo planeta. A partir do uso das informações de três dias anteriores, efetuou-se a previsão epidemiológica dos 10 dias posteriores. A rede neural concebida resultou em valores médios de 1,5% para a raiz do erro quadrático médio - métrica de desempenho utilizada no trabalho.

Ribeiro et al., (2020), realizou a previsão dos casos confirmados de COVID-19 no Brasil por meio de seis modelos baseados em AP. Neste artigo, ARIMA, regressão cúbica (CUBIST), o algoritmo *random forest* (RF), bem como os métodos de regressão RIDGE e *support vector regression* (SVR) são os modelos preditivos empregados para prever os casos cumulativos de COVID-19

em dez estados brasileiros, a partir de horizontes de predição de um, três e seis dias. Nesse trabalho, os modelos de regressão CUBIST, RF, RIDGE e SVR também foram unidos, a fim de gerar um modelo de aprendizado de máquina mais abrangente, e adotados como os aprendizes básicos da rede híbrida, sendo que uma técnica de processamento Gaussiano dos dados foi utilizada como meta-aprendiz. A eficácia dos modelos foi avaliada com base em métricas de desempenho, tais quais o erro absoluto médio e critérios de erro percentual absoluto médio simétrico. Na maioria dos casos, o SVR e o modelo híbrido alcançaram os melhores no que se refere aos parâmetros de desempenho aplicados. Em geral, os modelos desenvolvidos resultaram em predições precisas, alcançando erros na faixa de 0,87% – 3,51%, 1,02% – 5,63% e 0,95% – 6,90% para um, três e seis dias de antecedência, respectivamente. A classificação dos modelos, em ordem decrescente de desempenho preditivo, em todos os cenários, é SVR, modelo híbrido, ARIMA, CUBIST, RIDGE e modelos RF.

3 MATERIAIS E MÉTODOS

3.1 COLETA DE DADOS

Para a conformação do conjunto de dados epidêmicos atinentes à pandemia de COVID-19 em Santa Catarina, utilizou-se como fonte os boletins de ocorrência fornecidos pelo site oficial da Secretaria de Saúde do Estado de Santa Catarina. Por meio desses boletins foram captadas as seguintes séries de dados:

- quantidade acumulada de casos confirmados;
- quantidade acumulada de óbitos;
- quantidade acumulada de casos de pacientes recuperados da doença;
- número de leitos SUS ocupados por enfermos COVID-19.

Tem-se que o primeiro ponto de dados informado pela Secretaria de Saúde é datado do dia 13 de março de 2020, quando foram computados três casos confirmados da doença no estado catarinense. O último ponto de dados refere-se ao dia 21 de janeiro de 2022. O conjunto de dados foi compilado em formato .csv, o que permite sua manipulação e análise por meio da biblioteca *Pandas* em linguagem de programação *Python*.

3.2 ANÁLISE PRELIMINAR DAS SÉRIES DE DADOS

A presente seção tem como objetivo um breve esclarecimento quanto às características dos dados epidemiológicos da pandemia de COVID-19 no estado de Santa Catarina. Para tanto, serão descritos aspectos concernentes à representação gráfica das séries temporais empregadas no presente trabalho, bem como algumas características estatísticas presentes nos dados coletados.

Na Figura 1, são expostos graficamente os dados atinentes às quantidades cumulativas de casos de infecção, óbitos e recuperações, obtidas por meio de inspeção dos boletins epidêmicos.

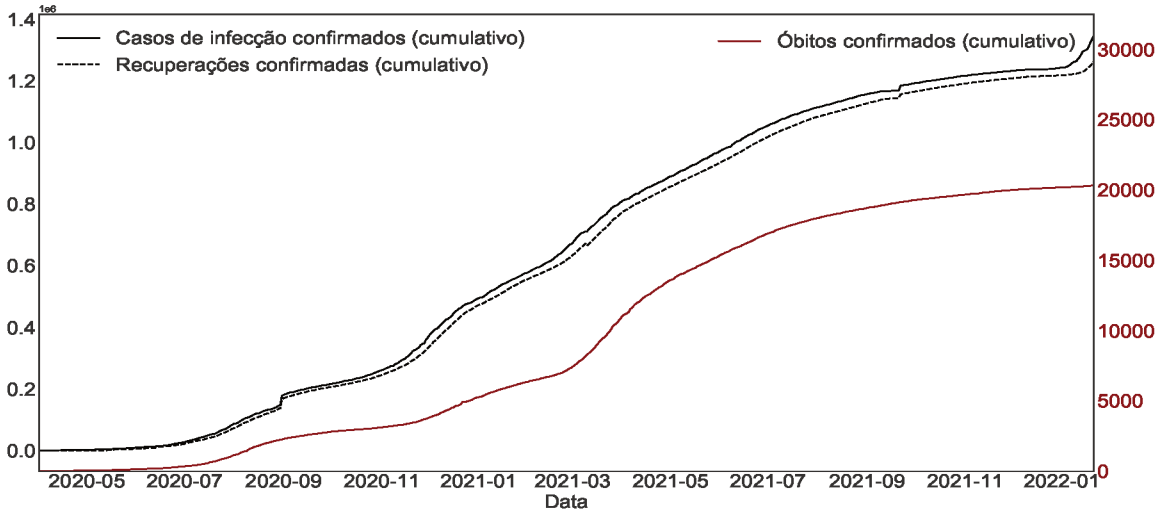


Figura 1 – Gráfico das séries temporais cumulativas, obtidas por meio de boletins lançados pela Secretaria de Saúde do Estado de SC.

Já na Figura 2, há a representação gráfica da série temporal relativa à quantidade de leitos SUS ocupados em Santa Catarina.

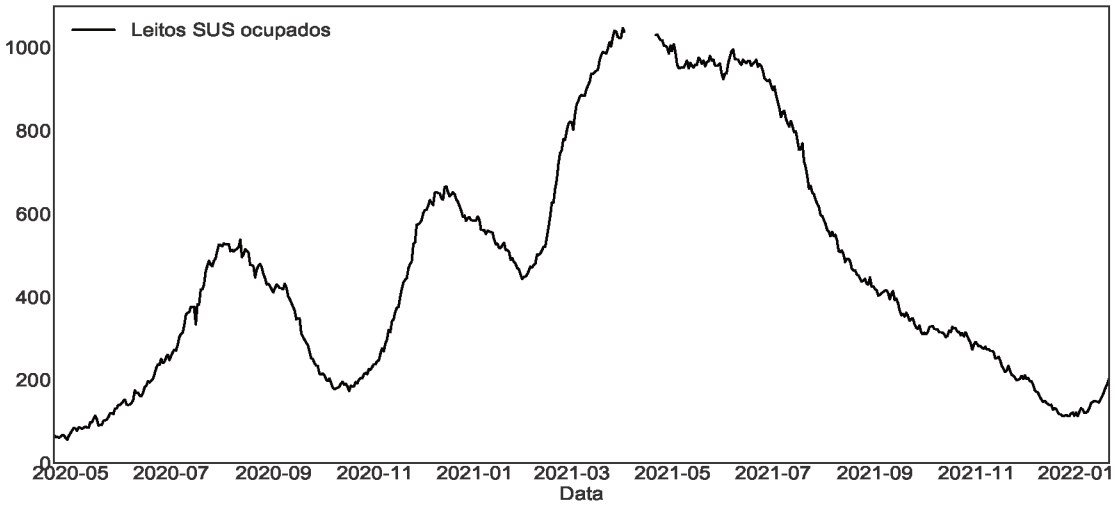
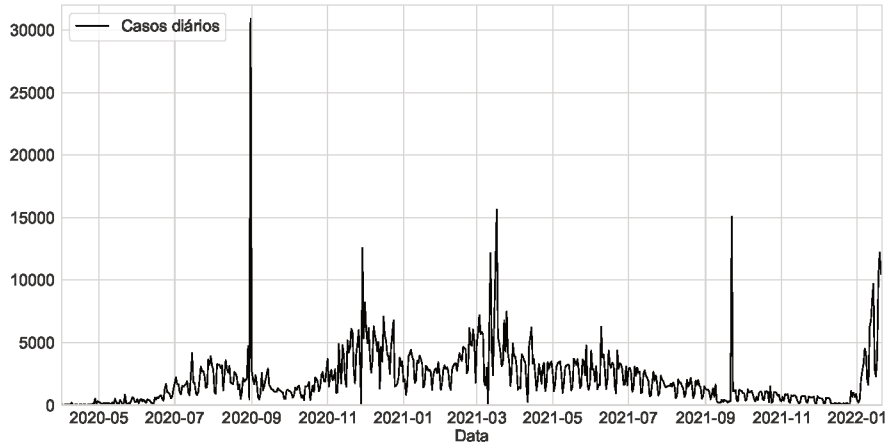
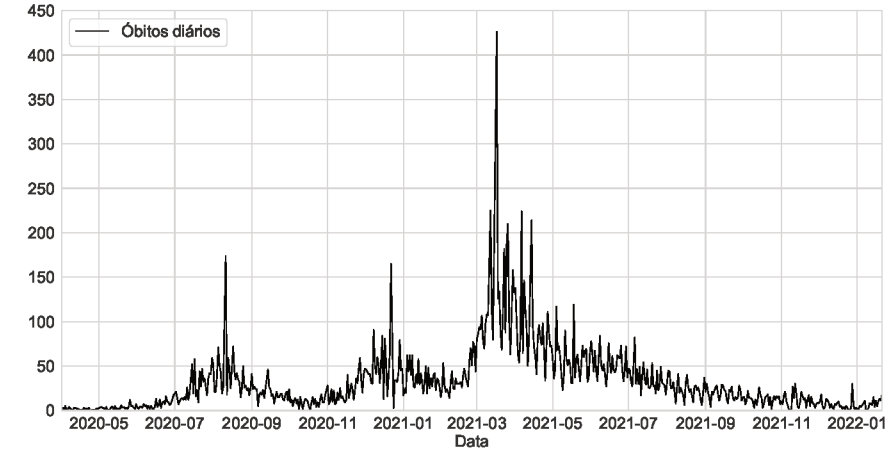


Figura 2 – Gráfico representativo da curva associada à série temporal leitos SUS de Santa Catarina ocupados por pacientes com COVID-19.

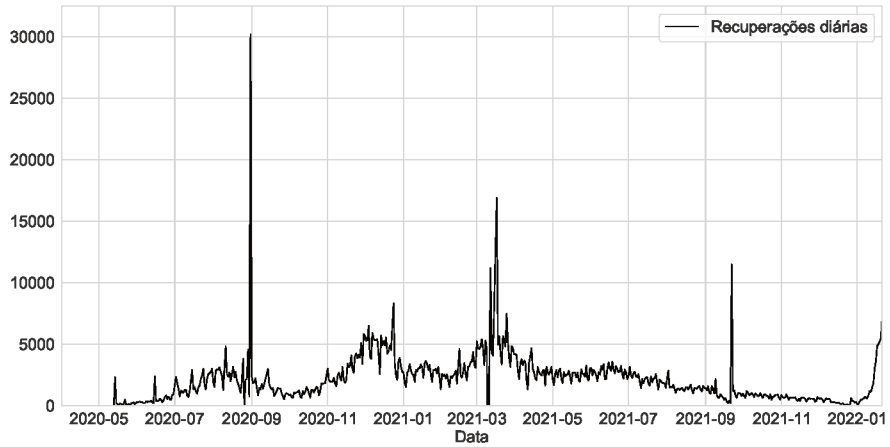
Por meio de simples inspeção visual das Figuras 1 e 2, chamam atenção algumas características: há uma descontinuidade associada ao dia 14 de setembro de 2020 para as séries de dados atinentes à quantidade de casos e óbitos notificados; a lacuna de dados em parte do mês de maio de 2021, relativa à série histórica de ocupação de leitos. De fato, os problemas mencionados devem-se a um equívoco de atualização do banco de dados, à época, por parte da Secretaria de Saúde. Devido à presença de lacuna, ou seja, ausência de sequência de observações referentes à ocupação de leitos, essa porção da série histórica foi submetida ao procedimento de interpolação linear prévio, para imputação das observações faltantes. Adicionalmente, observam-se incrementos abruptos nas quantidades acumuladas de casos de infecção e recuperação informados pela secretaria de saúde do estado, para o mês de janeiro de 2022, o que concorda com a fato de que nesse período, o Brasil, incluindo Santa Catarina, passavam por uma alta de casos diários da doença. A Figura 3 mostra as curvas associadas aos valores diários obtidos pela manipulação das séries das quantidades acumuladas para as variáveis referidas às séries temporais cumulativas de casos, óbitos e recuperações diárias.



(a)



(b)



(c)

Figura 3 – Gráficos para as curvas concernentes às séries temporais diárias de: (a) casos de infecção; (b) óbitos e; (c) recuperações.

Juntamente com a série histórica de ocupação de leitos, as séries históricas associadas às quantidades diárias de casos, óbitos e recuperações, serão efetivamente operadas. Tais conjuntos de dados serão submetidos a etapas de pré-processamento para, então, fornecerem dados históricos adequados à fase de previsão. Em aspecto visual, sobretudo as séries históricas atinentes aos casos de infecção, de óbitos e recuperações, suas respectivas curvas exibem aspecto ruidoso. Comparativamente, a série temporal associada à ocupação de leitos apresenta menor caráter ruidoso, munida de ruídos de menor amplitude.

Note-se que a série referente à quantidade de infecções diárias, assim como à quantidade de recuperações diárias, têm um valor negativo cada, -110 e -6439, respectivamente. Essas são as observações de valor mínimo para ambas as séries históricas e, obviamente, não são possíveis na realidade. Tal fenômeno também pode ser explicado por algum equívoco na atualização dos dados por parte da Secretaria de Saúde catarinense ou na alteração dos critérios de notificação adotados pela instituição. Ressalte-se que valores negativos como esses são excluídos e, após, em seus lugares, são atribuídos novos valores a partir um procedimento de interpolação polinomial, na fase de pré-processamento de dados.

Outro aspecto a ser levado em conta, quanto à disposição dos dados, é o fato de que foram capturadas 661 observações para cada série temporal, com exceção da série histórica de leitos ocupados, a qual possui 607 observações. Caso a secretaria de saúde efetuasse o lançamento de boletins informativos a uma taxa perfeitamente diária, as séries teriam 680 observações cada. Dessa forma, esses dados faltantes também são especialmente tratados em etapa de pré-processamento.

A Figura 4 exibe os gráficos *box-plot* das séries diárias. Já a Tabela 2 sumariza a caracterização estatística concernente às séries, bem como complementa as informações contidas nos gráficos mostrados na Figura 4. Por fim, supondo que as quatro séries históricas de trabalho possam ser descritas a partir do modelo de decomposição aditiva de séries temporais, tendo em vista que a amplitude de seus períodos sazonais não se altera com o passar do tempo, a Figura 5 mostra as componentes de tendência e sazonalidade para cada conjunto de dados.

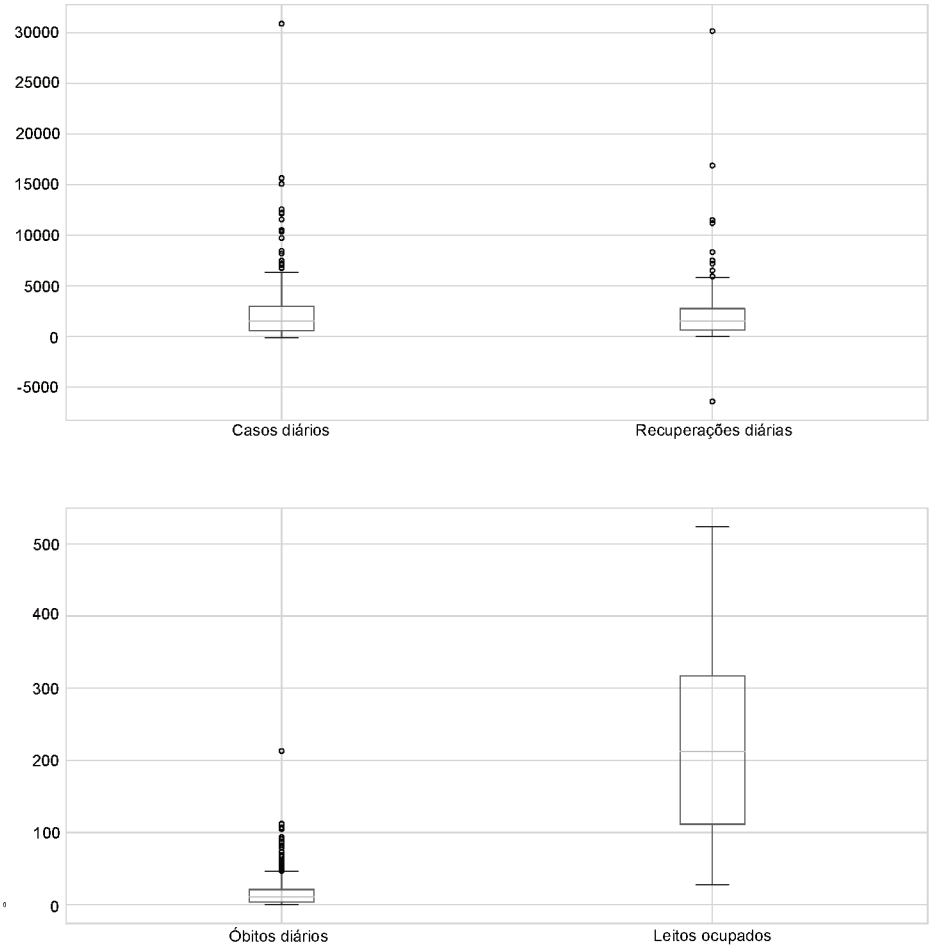


Figura 4 – Gráficos *box-plot* relativos às séries históricas diárias de casos de infecção, óbitos, recuperações e ocupação de leitos.

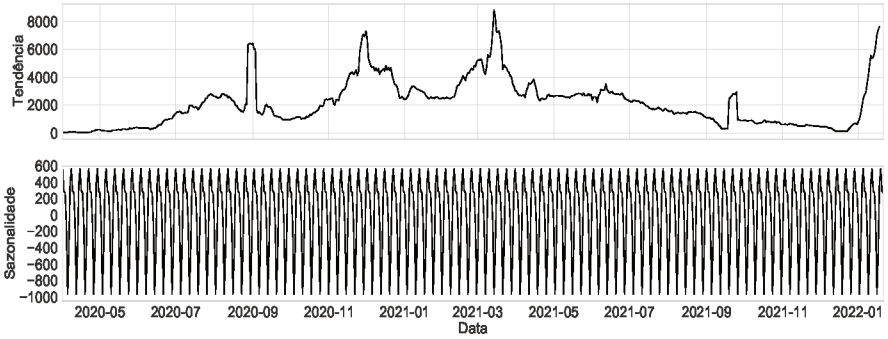
Os gráficos *box-plot* enfatizam, por meio do patamar das medianas, que os conjuntos de dados associados aos casos confirmados de infecção e de recuperação têm ordens de grandeza relativamente próximas. O mesmo não ocorre com as séries históricas de óbitos diários e ocupação de leitos, sendo que esta possui uma distribuição de dados mais abrangente, comparada aos outros conjuntos de dados. Adicionalmente, observa-se maior simetria no conjunto de dados atinente à ocupação de leitos, haja vista que, proporcionalmente, os tamanhos das caudas superior e inferior dos outros conjuntos de dados possuem

diferença considerável. Ainda, por se tratarem de séries temporais, os *outliers* - definidos apropriadamente na Seção 2.2.3, não serão tratados mediante análise de *box-plot*, visto que métodos desse tipo não levam em conta a ordem temporal da distribuição dos dados. Nesse sentido, verificam-se observações bastante discrepantes, a exemplo dos valores máximos de 30193 casos confirmados e 30186 recuperações, enfatizados na Tabela 2, as quais não condizem com a realidade epidêmica e se devem a algum equívoco na atualização de dados.

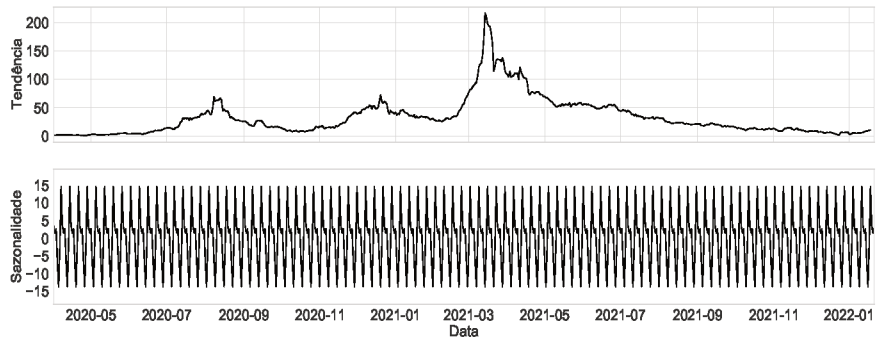
Já os gráficos exibidos na Figura 5 fornecem outras informações preliminares acerca da caracterização dos dados.

Tabela 2 – Descrição estatística das séries de dados correspondentes às quantidades diárias das principais variáveis de análise da pandemia de COVID-19 em SC.

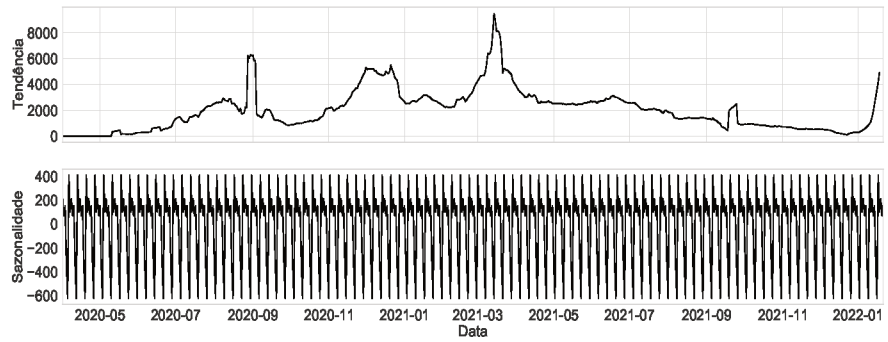
Característica	**Casos**	**Óbitos**	**Recuperados**	**Leitos ocupados**
Média aritmética	2041,7	30,8	1912,8	468,4
Desvio padrão	2314,9	36,5	2026,7	285,5
Valor mínimo	-110,0	0,0	-6439,0	56,0
Valor máximo	30193,0	426,0	30186,0	1047,0
Primeiro quartil	546,0	8,0	633,0	223,0
Segundo quartil	1502,0	21,0	1529,0	425,0
Terceiro quartil	2948,0	42,0	2749,0	633,5



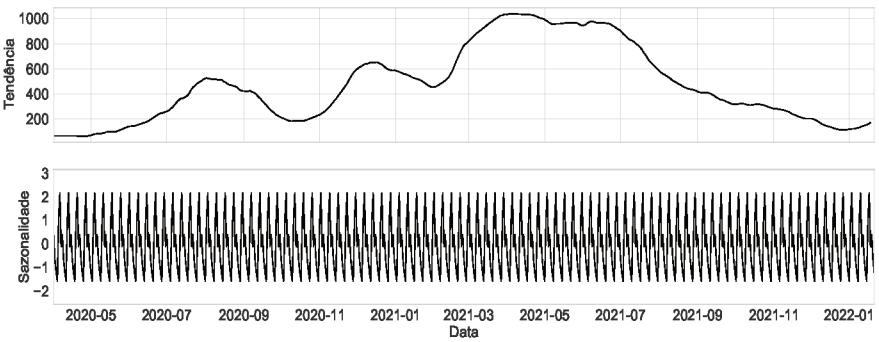
(a)



(b)



(c)



(d)

Figura 5 – Gráficos das componentes aditivas de tendência e sazonalidades das séries temporais diárias de: (a) casos de infecção; (b) óbitos; (c) recuperações e; (d) leitos ocupados.

Da análise visual das componentes de tendência, mostradas na Figura 5, verifica-se facilmente o caráter não estacionário das séries temporais estudadas neste trabalho, visto que cada uma delas apresenta momentos de crescimento, decrescimento, bem como variância inconstante nos dados. Ainda, inspecionando-se as componentes aditivas de sazonalidade, confirma-se sazonalidade semanal para todos os conjuntos de dados. De fato, essa característica se relaciona com o modo em que a Secretaria de Saúde do Estado de Santa Catarina atualizou a base de dados concernente à pandemia de COVID-19.

Para atividades de organização, manipulação e análise descritiva dos conjuntos de dados foi utilizada precipuamente a biblioteca *pandas*[1], especificamente em sua versão 1.2.3, ferramenta de código aberto criada para a linguagem Python, para fins manipulação e análise de dados.

3.3 PRÉ-PROCESSAMENTO DOS DADOS

Posteriormente à coleta, análise exploratória e obtenção das séries temporais referidas aos dados epidemiológicos pontuais diários, estas foram submetidas às etapas de pré-processamento, as quais respeitam a seguinte ordem de execução.

1. Identificação de *outliers* por meio do filtro de Hampel.
2. Imputação de valores ausentes por meio de interpolação linear.
3. Suavização de ruído com uso do filtro de Savitsky-Golay.

A aplicação desses procedimentos será descrita a seguir.

3.3.1 Remoção de *Outliers* e Imputação de Valores Ausentes

A remoção de dados discrepantes foi efetuada aplicando-se o filtro de Hampel - cuja formulação é dada na Seção 2.2.3.1 - a cada uma das séries históricas, separadamente, e sua implementação foi feita com a utilização da

[1] Informações relativas à versão 1.2.3 da biblioteca *pandas* podem ser consultadas em https://pypi.org/project/pandas/1.2.3/.

biblioteca *Numpy*[2] - versão 1.19.5 - com o uso de métodos para o cálculo da mediana e do desvio absoluto da mediana, para concepção das Equações 2 e 3. Dessa forma, foram excluídas as observações externas aos limites superior e inferior estabelecidos na Equação (2), os quais são identificados como discrepantes por esse processo de filtragem. Ressalte-se que foram utilizadas janelas móveis de dados de tamanho mensal para identificação de *outliers* e, por fim, foram identificados:

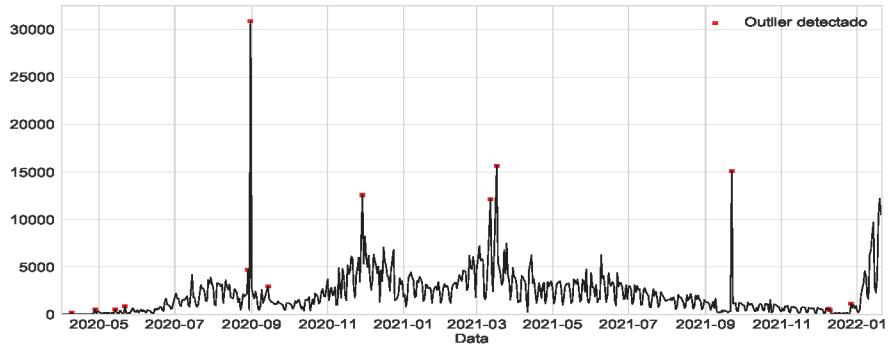
- 14 valores discrepantes para a série de casos diários de infecção;
- 12 para a série de pacientes recuperados diariamente;
- 15 para a série de óbitos diários;
- 9 para a série temporal de leitos SUS ocupados.

A Figura 6 fornece esclarecimento visual concernente aos resultados da aplicação desse procedimento. Note-se que as observações discrepantes estão destacadas com marcações em cor vermelha nos gráficos mostrados nessa figura.

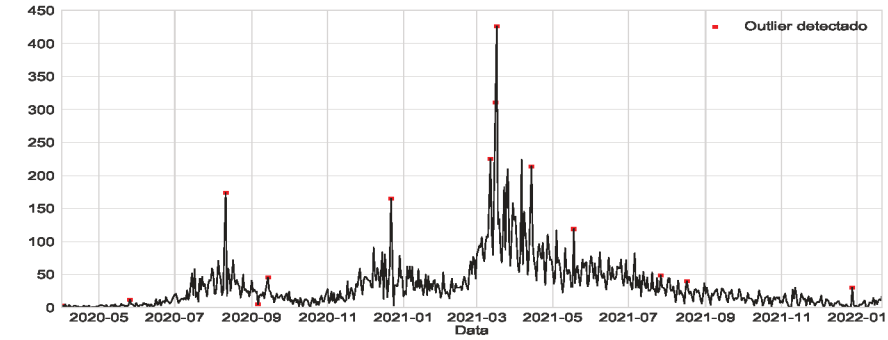
Considerando que os conjuntos de dados deste trabalho possuem quantidade relativamente baixa de dados, faz-se adequado substituir as observações excluídas por valores mais condizentes com os dados. Para cumprir tal propósito, utilizou-se como método de imputação de valores ausentes a interpolação polinomial de primeira ordem (ou interpolação linear). Reforça-se que tal método também é aplicado às observações atinentes à base de dados crua - cujas séries temporais ainda não foram submetidas aos procedimentos de pré-processamento de dados. No caso deste trabalho, as observações ausentes contidas na base de dados crua não constituem por si só uma perda de dados, visto que, comumente, o órgão de saúde responsável pela atualização desses dados divulgam essas observações cumulativamente em data posterior.

Adicionalmente, esse procedimento é pertinente, haja vista que a etapa de suavização de ruído, posterior a esta, utiliza-se de uma técnica que requer

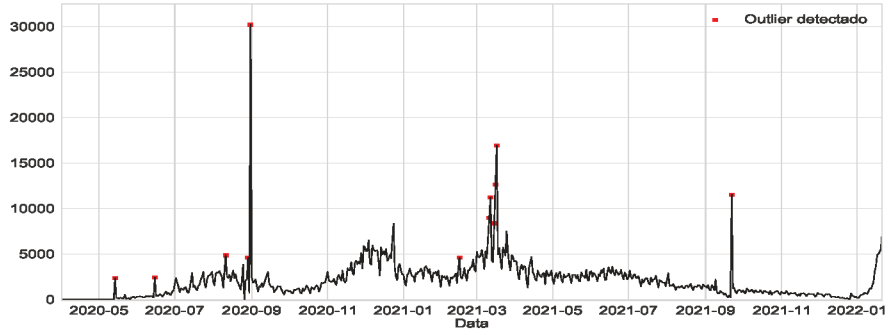
[2] O projeto da biblioteca *Numpy*, de versão 1.19.5, pode ser acessado no endereço eletrônico https://pypi.org/project/numpy/1.19.5/.



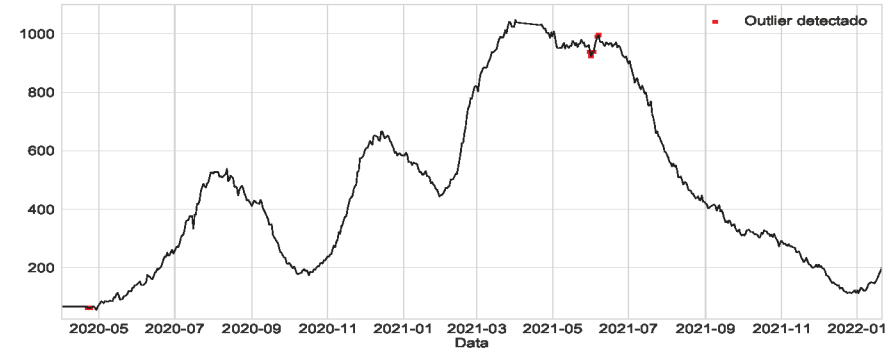
(a)



(b)



(c)



(d)

Figura 6 – *Outliers* identificados pelo filtro de Hampel para as séries temporais diárias de: (a) casos; (b) óbitos; (c) recuperações e; (d) leitos ocupados.

que as observações sejam igualmente espaçadas entre si. No entanto, vale ressaltar os efeitos oriundos da imputação de valores ausentes por meio de métodos como a interpolação. Nesse sentido, Çokluk e Kayri (2011), após efetuarem comparações entre os efeitos ocasionados a partir da imputação de valores ausentes entre diversos métodos (entre eles, a interpolação linear), concluíram que a imputação de valores ausentes diminui a variância explicada pelos dados.

3.3.2 Suavização de Ruído

Após a construção de um conjunto de dados sem *outliers* e lacunas, com observações igualmente espaçadas a um intervalo diário, aplica-se o filtro de Savitsky-Golay (SAVITZKY; GOLAY, 1964), como ferramenta para redução de ruído presente nas séries históricas. Fundamentando-se no trabalho de (KRISHNAN; SEELAMANTULA, 2013), mencionado na Seção 2.4.1, utilizou-se como abordagem sugerida pelo referido trabalho a fixação de um valor para a ordem polinomial do filtro e a subsequente escolha ótima do tamanho de sua janela móvel, em termos da minimização do erro quadrático médio mínimo. Dessa forma, utilizou-se de filtros de Savitsky e Golay de ordem fixa de valor 4, sendo que os tamanhos ótimos das janelas móveis dos filtros foram calculados conforme o modelo proposto por a partir de implementação algorítmica fornecida pelo referido trabalho, de modo a otimizar aspectos de viés e variância presentes nos dados. Com o uso de tal método, foram obtidas janelas móveis de:

- 25 observações para a série de casos diários;
- 25 observações para a série de mortes diárias;
- 23 observações para a série de recuperações diárias;
- 12 observações para a série de leitos ocupados.

Na Figura 7, são mostrados os gráficos comparativos das séries cruas, ou seja, aquelas coletadas diretamente da plataforma da Secretaria de Saúde do Estado de Santa Catarina, e das séries que foram submetidas aos tratamentos supracitados.

De fato, após o procedimento de suavização, são obtidas as séries históricas que servirão de base para a etapa de predição dos dados, ou seja, essas são as séries temporais de trabalho. A Tabela 3 sumariza a descrição estatística das séries temporais submetidas às etapas de pré-processamento supracitadas.

Tabela 3 – Descrição estatística das séries temporais submetidas às etapas de pré-processamento.

Característica	**Casos**	**Óbitos**	**Recuperados**	**Leitos ocupados**
Média aritmética	1993,3	30,4	1848,7	460,7
Desvio padrão	1656,5	29,5	1430,6	303,6
Valor mínimo	0,2	0,1	0,0	56,0
Valor máximo	12594,6	143,6	6943,6	1037,8
Primeiro quartil	659,3	10,0	618,8	202,8
Segundo quartil	1709,2	21,7	1705,7	413,8
Terceiro quartil	2748,6	43,7	2632,9	639,6

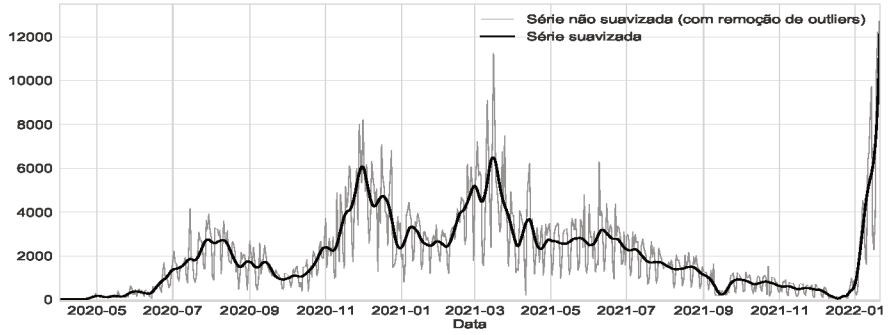
Sabe-se que o filtro de Savitsky-Golay é largamente utilizado em aplicações de temática biomédica. Como exemplos de trabalhos concernentes à eletrocardiografia, podem ser citados os artigos de Nishida et al., (2017), e Hargittai, (2005). Os trabalhos de Haider et al., (2018), e Singh et al., (2020), utilizam o referido filtro com a finalidade de suavizar sinais provenientes da atividade pulmonar em seres humanos. Por sua vez, no estudo de Sadikoglu e Kavalcioglu, (2017), aplica-se o método para filtragem de sinais de monitoramento de glicose.

A utilização do filtro de Savistky-Golay, em detrimento de filtros de média móvel tradicionais, dá-se por conta, e em conformidade com Abusam (2022), de que a referida técnica preserva características importantes das séries temporais, a exemplo da altura e largura dos picos e vales, bem como retém as tendências do sinal a ser tratado. Nesse sentido, Isnanto, (2011), ao comparar o filtro de Savitsky-Golay ao filtro clássico de média móvel, aponta que o método mais tradicional comumente retira maiores porções de alta frequência dos sinais, o que limita a capacidade desse filtro em preservar as regiões de pico e regiões de vale.

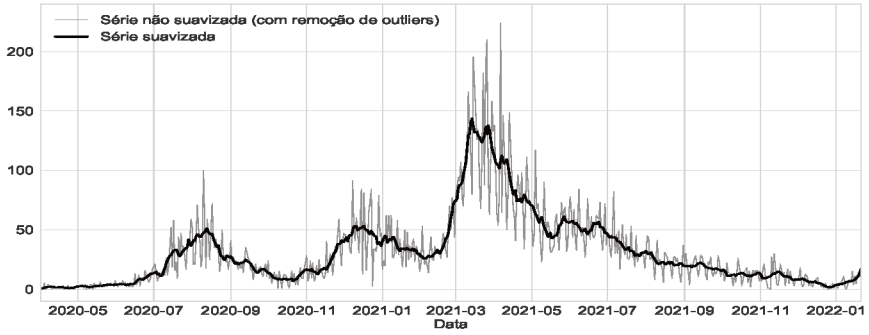
Ainda, Haider et al., (2018), os quais aplicaram o filtro de SG para suavização de sinais pulmonares, proferem o seguinte: "*O filtro de Savitsky-Golay tem várias vantagens sobre os métodos mais comuns de filtragem. Ele supera o filtro de média móvel na preservação de sinal válido*". De fato, os aspectos técnicos supracitados já foram mencionados por Abusam, (2022), e Zhan et al., (2021) - os quais, assim como o presente trabalho, abordam a análise preditiva de dados epidemiológicos. Zhan et al. (2021) optaram por tratar os dados epidemiológicos concernentes à pandemia em 184 países a partir do filtro de Savitsky-Golay, visando os momentos de pico das séries temporais associadas aos casos ativos de infecção pela doença. Já Abusam (2022) optou pelo uso do referido filtro, em vez de aplicar o filtro clássico de média móvel, tendo em vista as vantagens técnicas já mencionadas.

Por fim, pode-se citar outros trabalhos associados à pandemia de COVID-19 que se utilizaram do filtro de Savitsky-Golay. Parolini et al. (2021) criaram um *dashboard* para efetuação de análises estatísticas à respeito da pandemia na Itália, sendo que a filtragem via método de Savitsky-Golay foi aplicada com o intuito de suavização das séries históricas de casos de infecção. Fontal et al. (2021) se utiliza do filtro para suavizar sinais de temperatura, parâmetro utilizado como atributo associado aos dados epidemiológicos. Já Cui e Kertész (2021) aplicam o filtro de Savitsky-Golay na etapa de pré-processamento das séries temporais atinentes à infecção por COVID-19 em regiões da China.

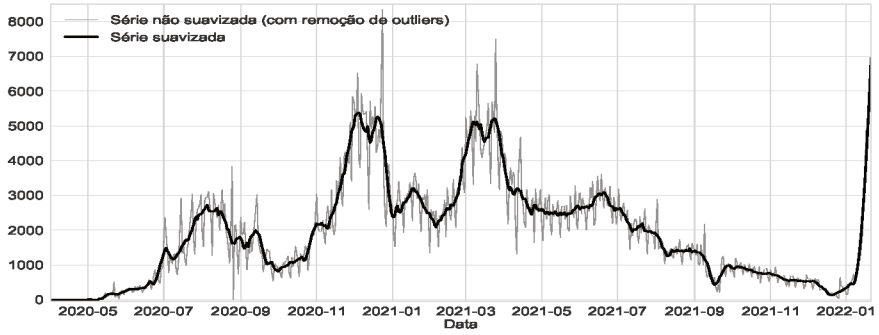
Ressalte-se que, antes da etapa de previsão propriamente dita, tem-se a aplicação do último procedimento concernente à análise dos dados, a qual corresponde ao critério e correspondente seleção de preditores exógenos.



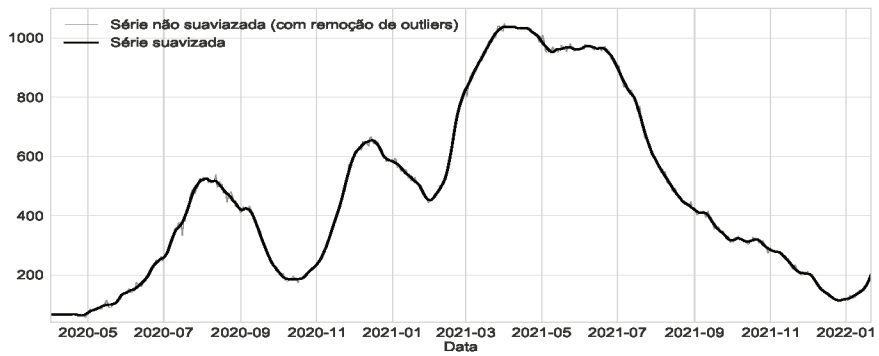
(a)



(b)



(c)



(d)

Figura 7 – Séries temporais suavizadas por meio de filtro de Savitsky-Golay: (a) casos; (b) óbitos; (c) recuperações e; (d) leitos ocupados.

3.3.3 Seleção de Preditores Exógenos

Como o presente trabalho prevê a utilização de modelos preditivos que agregam preditores exógenos, tais quais os modelos NARX e NARMAX polinomiais, faz-se crucial a adoção de critério para a escolha desses preditores externos.

Inicialmente, decidiu-se pela listagem de conjuntos de séries históricas candidatas a serem preditores exógenos das variáveis a serem identificadas. Sejam $c(t)$, $m(t)$, $r(t)$ e $l(t)$ as denotações para séries históricas associadas aos óbitos, recuperações e ocupação de leitos, respectivamente, descreve-se os preditores exógenos candidatos da seguinte maneira:

- Constituem o conjunto candidato de preditores exógenos da série temporal de casos diários de infecção as séries históricas de óbitos diários, recuperações diárias e leitos ocupados munidas de atrasos discretos diários de 1 até 7 dias. Logo, são candidatos a preditores exógenos dos casos diários as séries: $m(t-1)$, $m(t-2)$, ..., $m(t-7)$, $r(t-1)$, $r(t-2)$, ..., $r(t-7)$, $l(t-1)$, $l(t-2)$, ..., $l(t-7)$.

- Constituem o conjunto candidato de preditores exógenos da série temporal de recuperações diárias as séries históricas de casos diários, mortes diárias e leitos ocupados, munidas atrasos discretos diários de 1 até 7 dias. São candidatas a preditores exógenos da série de óbitos as séries: $c(t-1)$, $c(t-2)$, ..., $c(t-7)$, $r(t-1)$, $r(t-2)$, ..., $r(t-7)$, $l(t-1)$, $l(t-2)$, ..., $l(t-7)$.

- Constituem o conjunto candidato de preditores exógenos da série temporal de óbitos diários as séries históricas de casos diários, recuperações diárias e leitos ocupados, munidas de atrasos discretos diários de 1 a 7 dias. Dessa forma, constituem candidatos a preditores exógenos para as recuperações diárias, as séries: $c(t-1)$, $c(t-2)$, ..., $c(t-7)$, $m(t-1)$, $m(t-2)$, ..., $m(t-7)$, $l(t-1)$, $l(t-2)$, ..., $l(t-7)$.

- Constituem o conjunto candidato de preditores exógenos da série temporal de casos diários de infecção as séries históricas de óbitos diários,

recuperações diárias e leitos ocupados, munidas atrasos discretos diários de 1 a 7 dias. Ou seja, candidatam-se como preditoras exógenas da série asssociada à ocupação de leitos, as séries: $c(t-1)$, $c(t-2)$, ..., $c(t-7)$, $m(t-1)$, $m(t-2)$, ..., $m(t-7)$, $r(t-1)$, $r(t-2)$, ..., $r(t-7)$.

Determinou-se, como critério de aprovação dos preditores candidatos, que esses tivessem uma correlação de Spearman (SPEARMAN, 1961), cuja descrição é feita na Seção 2.4.2.1. Os preditores exógenos escolhidos devem possuir correlação de Spearman no mínimo forte, em relação à variável de interesse. Conforme descrito na Tabela 1, o valor de correlação que respeita esse critério possui um valor de no mínimo $r_s = 0,7$, em que r_s é o coeficiente de correlação de Spearman. Ou seja, com uma correlação de Spearman dessa ordem, garante-se uma relação monotônica evidente entre o preditor exógeno e a série temporal que se pretende prever.

Optou-se por omitir os valores dos coeficientes de correlação de Spearman, visto que verificou-se que todos os preditores exógenos candidatos referentes às séries temporais defasadas estão aptos a serem utilizados efetivamente nas atividades de previsão, haja vista que a elas estão associadas correlações de Spearman no mínimo fortes. Por outro lado, os preditores atinentes às séries munidas de uma diferenciação possuem coeficientes de correlação baixos e, portanto, não serão utilizadas em procedimentos de previsão.

3.4 MODELAGEM DAS SÉRIES TEMPORAIS

Após a efetuação do pré-processamento dos dados e a seleção de preditores exógenos, efetua-se a sistematização dos experimentos de previsão, visto que, a essa altura, já se dispõe de uma massa de dados tratada e apta a realização de predições. No decurso desta seção são descritos como as técnicas de predição utilizadas neste trabalho foram experimentadas sobre os conjuntos de dados.

3.4.1 Identificação e Teste

O processo de identificação dos modelos preditivos foi realizado por meio do método de avanço não ancorado de uma janela deslizante (Figura 8), no qual foi estabelecida uma janela fixa de 150 dias de observações para treinamento dos modelos. Em outras palavras, essa janela corresponde ao conjunto móvel de observações em que há a identificação das séries temporais.

As predições para cada variável foram efetuadas para o dia seguinte à janela de treinamento, constituindo, portanto, um paradigma de predição de um passo à frente (ou um dia à frente). Ou seja, a janela de teste, na qual se efetua as predições, tem tamanho constante de uma observação. A janela de identificação de tamanho constante progride (ou desliza) um dia a cada ciclo de identificação e teste, sendo que esse processo se repete até o final da base de dados, munida de 680 observações para cada variável a ser predita. Portanto, são realizadas 530 predições pontuais diárias.

Na prática, essa abordagem efetua a identificação de um modelo a cada etapa de identificação e teste e prioriza informações mais recentes à medida que a janela deslizante avança sobre a série temporal, utilizando-se, para a concepção dos modelos, de informações relativas aos cinco últimos meses.

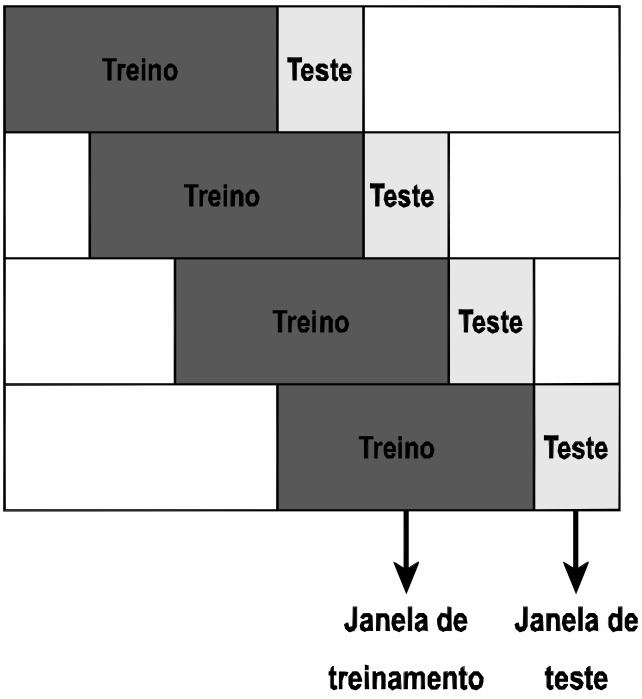


Figura 8 – Método de validação por meio de janela deslizante com avanço não ancorado.

3.4.1.1 Considerações Sobre o Horizonte de Previsão

Na literatura relacionada à pandemia de COVID-19, a abordagem preditiva de um dia à frente, do ponto de vista de gestão de saúde, é consideravelmente menos utilizada se comparada à aplicações com preditores munidos horizontes de predição mais elevados, a exemplo da previsão de uma semana à frente.

Nesse sentido, Ribeiro et al. (2020), visando ajudar o sistema público de saúde a partir de um planejamento estratégico, utilizaram cinco técnicas de modelagem distintas a fim de efetuarem previsões de um dia, três dias e uma semana à frente, da quantidade cumulativa de casos de infecção por COVID-19 em dez estados brasileiros. Os autores do referido trabalho argumentam que a contribuição primária do estudo é a criação de um sistema de predição, cuja precisão dos modelos auxilia governantes na tomada de decisão para conter a pandemia e estratégias sobre o sistema de saúde brasileiro.

Outra contribuição refere-se à aplicação de preditores considerando-se

mais de um horizonte de previsão, fator que colabora para uma análise mais realista do ponto de vista de gestão de recursos médicos.

Em Deschepper et al., (2021), desenvolveu-se um modelo preditivo, que abrange a previsão de até 10 dias à frente, para estimação da quantidade de leitos do hospital universitário Ghent, na Bélgica. Nesse trabalho, a previsão de um passo à frente foi utilizada como etapa de validação do modelo, em termos de exatidão. Ou seja, na pesquisa a variante de preditores de um dia à frente foi utilizada não apenas com o objetivo da predição propriamente dita, mas como parte da análise de eficiência preditiva dos modelos.

Já o trabalho de Goic et al., (2021), relacionado à estimação de leitos para pacientes enfermos por COVID-19 em regiões do Chile, utiliza-se apenas de preditores de uma e duas semanas à frente. Segundo os autores, o sistema de previsão desenvolvido estava sendo usado ativamente, à época da submissão do artigo, para planejamento de capacidade de hospitais chilenos. De fato, isso sugere que a abordagem preditiva proposta tornou-se útil do ponto de vista prático. Outra contribuição mencionada pelos autores do referido estudo, é de que a abordagem pode ser facilmente replicada em outros países que enfrentam restrições agudas de capacidade em relação aos leito de unidade de saúde emergenciais, haja vista de que se trata de uma pesquisa de código aberto.

No entanto, há trabalhos que visam a previsão do estado hospitalar em intervalos reduzidos de tempo. A exemplo, Tuominen et al., (2022). Nesse trabalho, os autores investigam a seleção de preditores exógenos adequados para a previsão da quantidade de pacientes do dia seguinte do departamento de emergência do hospital universitário de Tampere, na Finlândia.

Ainda, vale mencionar estudos que indicaram a importância em fazer previsões em intervalos de tempo muito menores do que um dia ou uma semana, a exemplo de Schweigler et al., (2009), e Hoot et al., (2008). Em Schweigler et al., (2009), investigou-se a taxa de ocupação de departamentos de emergência de três hospitais, a partir de horizontes de previsão de 4 horas e 12 horas. Nessa pesquisa, objetivou-se verificar a viabilidade desse tipo de previsão, a partir de técnicas de previsão autorregressivas, como o filtro de média móvel simples e o modelo ARIMA. Os modelos desenvolvidos no estudo foram concebidos para serem ferramentas auxiliares aos hospitais, no intuito de mitigar situações

de superocupação do departamento de emergência e de redes ambulatoriais regionais.

Em conformidade com as informações descritas, atinentes ao horizonte de previsão, pode-se deduzir que o tipo de preditor a ser utilizado depende precipuamente do cenário de aplicação. Para contextos mais emergenciais, em que as condições hospitalares se alteram consideravelmente de um dia para o outro, previsões diárias podem se tornar adequadas, a depender da capacidade de resposta das unidades de gestão de saúde. No entanto, preditores de uma semana à frente certamente proporcionam maior abrangência aos gestores de saúde, fornecendo maior gama de dados a fim de gerenciar os recursos hospitalares dentro de uma semana.

3.4.2 ARIMA

Para a concepção dos modelos ARIMA, utilizou-se como ferramenta para determinação dos parâmetros p, d e q, os correlogramas de autocorrelação e autocorrelação parcial das variáveis de interesse, cada qual com suas particularidades. Os parâmetros p, d, q - associados, respectivamente, às ordens de autorregressão, integratividade e média móvel do modelo - são descritos na Seção 2.5.1. Os correlogramas de autocorrelação e autocorrelação parcial são, respectivamente, os diagramas das funções de autocorrelação e autocorrelação parcial para uma determinada distribuição de dados. Seja X_t uma série estacionária, a função de autocovariância para o atraso h da referida série é

$$\gamma_X(h) = \text{cov}(X_{t+h}, X), \tag{23}$$

em que cov é o operador de covariância. A função de autocorrelação de X_t, para um atraso h, é definida como:

$$\rho_X \equiv \frac{\gamma_X(h)}{\gamma_X(0)}. \tag{24}$$

Ou seja, a autocorrelação para o atraso h da série X_t nada mais é do que a correlação linear entre a série original X_t e a série defasada de h passos. Já a autocorrelação parcial, para determinado atraso h de uma série temporal, é uma medida da correlação entre série defasada de h atrasos e a série temporal

original, excluindo-se a influência dos outros atrasos para o cálculo. A exemplo, a autocorrelação parcial para um atraso de 5 unidades é apenas a correlação que não contabilizada por atrasos de 1 a 4 unidades. Um maior esclarecimento acerca da autocorrelação parcial, bem como sua formulação, é fornecida em (DURBIN, 1960).

Sublinha-se que os modelos ARIMA foram criados por meio da biblioteca ARIMA, a qual implementada para programação em linguagem Python. Essa biblioteca constitui parte do pacote de análise de séries temporais *statsmodels.tsa*, sendo que a versão 0.13.2 foi utilizada.

Em síntese, respeitou-se as seguintes etapas de modelagem.

1. Primeiramente, verifica-se o caráter estacionário da série temporal original completa, a partir da aplicação do teste de Dickey-Fuller Aumentado (ADF), desenvolvido por Dickey e Fuller (1981). Esse teste visa a verificar a existência de raiz unitária em séries temporais a partir de testes de hipóteses, sendo que hipótese nula é de que existe raiz unitária (série não estacionária) e a hipótese alternativa determina que a série é estacionária. A estatística ADF, usada no teste, é um número negativo, e quanto mais negativo, mais indicativo o teste se torna de rejeitar a hipótese nula de que existe raiz unitária na série. Adicionalmente, o *p*-valor resultante da teste de hipóteses deve ser menor que o valor fixado de 0,05 (GAUVREAU; PAGANO, 1994). Para um maior detalhamento teórico acerca do método, o leitor pode se referir ao trabalho (DICKEY; FULLER, 1981). A partir do uso do teste supracitado, checou-se a estacionariedade das duas primeira diferenças das séries temporais, quantidade suficiente de diferenciação para que todas as variáveis de interesse tornem-se estacionárias, segundo o referido teste.

2. Para determinação da ordem de autorregressão *p* e de média móvel *q* dos modelos ARIMA, utilizou-se a a abordagem descrita na Tabela 5, cuja formulação teórica é detalhada em (MONTGOMERY *et al.*, 2008).

Tendo em vista que o paradigma de validação dos modelos preditivos se utilizará de janelas deslizantes, nem todas essas janelas de observações são

explicadas pelo conjunto de parâmetros p, d e q resultantes dessa abordagem. Dessa forma, a fim de considerar mais cenários de previsão, utilizou-se modelos ARIMA com até sete níveis de autorregressão, de forma a respeitar aos critérios adotados para escolha dos preditores exógenos, exposto na Seção 3.3.3. De forma complementar, também foram incluídas três ordens de média móvel aos modelos implementados, de tal forma que os valores de q levados em conta são 0, 1 ou 2.

Como cada série temporal de interesse apresenta particularidades acerca de aspectos como estacionariedade e aspectos de correlação, faz-se pertinente exibir as análises para cada uma delas, as quais serão discutidas a seguir.

Os resultados para os coefientes estatísticos do teste de Dickey-Fuller Aumentado, denotado por ADF, e seus respectivos p-valores, são mostrados na Tabela 4. Os correlogramas das séries temporais (incluindo suas primeira e segunda diferenças) são apresentados nas Figuras 20-23, situadas no Apêndice A.

Tendo em vista que valor crítico do teste estatístico ADF, para séries temporais munidas de 680 observações, é de -2,866, referido a um nível de confiança de 95%, pode-se chegar às seguintes conclusões - em conformidade com os dados mostrados na Tabela 4:

- **Para a série temporal de casos diários**: observa-se que essa série temporal é estacionarizada apenas com a imposição de uma diferença, quando o teste estatístico ADF resulta em um valor mais negativo que o valor crítico referido a um intervalo de confiança de 95%. Tal resultado é corroborado a partir de um p-valor = 0,044, por meio do qual se pode rejeitar a hipótese nula. Dessa forma, faz-se pertinente identificar toda a série a partir de modelos ARIMA munidos de d = 1. Ainda, torna-se viável identificar tal conjunto de dados a partir de modelos com duas ordens de diferenciação, visto que, nesse caso, a rejeição da hipótese nula é ainda mais contundente.

- **Para a série temporal de mortes diárias**: verifica-se que a série adquire caráter estacionário a partir da aplicação de uma diferença. Dessa forma, um modelo ARIMA adequado para teste tem parâmetro de diferenciação

unitário. Por meio de inspeção do valor do teste estatístico, com duas diferenças, esse parâmetro possui um valor consideravelmente menor (ADF = -7,334). Dessa forma, faz-se pertinente o teste com modelos ARIMA cujo $d = 2$.

- **Para a série temporal de recuperações diárias**: a estacionariedade é alcançada apenas com duas diferenças, visto que o teste estatístico retorna um valor de magnitude superior ao valor crítico para um critério de 5%.
- **Para a série temporal de ocupação de leitos**: depreende-se análise similar à elaborada para a série temporal de recuperações diárias.

Considerando-se mais cenários de previsão, bem como observando-se a faixa de valores do parâmetro *d* para garantia de estacionariedade das séries históricas, experimentou-se diferenças de primeira e segunda ordem aos cenários preditivos.

Tabela 4 – Resultados do teste estatístico de Dickey-Fuller Aumentado para as séries temporais.

Série temporal	**ADF**	***p*-valor**
Casos		
Série original	-0,636	0,863
1ª diferença	-3,086	0,044
2ª diferença	-5,941	0,000
Mortes		
Série original	-2,561	0,101
1ª diferença	-4,595	0,000
2ª diferença	-7,334	0,000
Recuperações		
Série original	-1,836	0,363
1ª diferença	-2,986	0,048
2ª diferença	-6,527	0,000
Ocupação de leitos		
Série original	-2,203	0,205
1ª diferença	-3,342	0,013
2ª diferença	-5,792	0,000

Adicionalmente, tendo em vista as variantes estacionárias das séries de referência, dos correlogramas apresentados no Apêndice A, depreendem-se as seguintes conclusões acerca dos níveis de autorregressão e média móvel dos modelos ARIMA, as quais são fundamentadas a partir da Tabela 5 (MONTGOMERY *et al.*, 2008), para identificação de modelos ARMA - ou modelos ARIMA (p, q, d) cuja ordem de diferenciação já garanta estacionariedade da série.

Tabela 5 – Tabela para identificação de modelos ARMA (p, q).

	ACF (ρ_k)	**PACF (ϕ_{kk})**
AR (p)	Função exponencial e/ou senoidal amortecida	$\phi_{kk} = 0$, $k > p$
MA (q)	$\rho_k = 0$, $k > q$	Função exponencial e/ou senoidal amortecida
ARMA (p, q)	Função exponencial e/ou senoidal amortecida a partir do atraso máx$(0, q-p)$	Função exponencial e/ou senoidal amortecida a partir de um atraso máx$(0, p-q)$

- **Para a série temporal de casos diários (Figura 20)**:
  - Série com uma diferença: para a série com uma diferença, verifica-se que a função de autocorrelação pode ser modelada a partir de um envoltória senoidal amortecida. Por outro lado, o gráfico de autocorrelação parcial mostra que há uma queda abrupta dessa magnitude a partir da observação com um atraso.
  - Série com duas diferenças: a partir da análise dos correlogramas, um modelo pertinente para teste corresponde ao ARIMA $(0, 2, 0)$, haja vista que o ponto de corte de ambos os gráficos, de autocorrelação e autocorrelação parcial, apresentam ponto de corte para a observação atual. No entanto, nesse cenário ocorre sobrediferenciação, o qual corresponde a um padrão muitas vezes nocivo à identificação (HOSSAIN *et al.*, 2019), em que os sinais dos valores de autocorrelação se alternam para atrasos consecutivos. Ou seja, essa forma de modelo não será aplicada nos cenários de previsão.
  - Conclusões: para esta série, faz-se adequado a experimentação de um modelo ARIMA $(1, 1, 0)$ para a série original.
- **Para a série temporal de óbitos diários (Figura 21)**:

  - Série com uma diferença: por meio de análise gráfica da função de autocorrelação parcial, o ponto de corte para essa variável ocorre no primeiro atraso. No entanto, não há clareza, a partir de inspeção visual, se, a partir do primeiro atraso, a função de autocorrelação assume aspecto de função senoidal amortecida.
  - Série com duas diferenças: com duas diferenças, a série apresenta pontos de corte nas observações atuais para ambos os correlogramas. Dessa forma, depreende-se conclusão análoga à descrita para a série temporal de casos diários munida de duas diferenças.
  - Conclusões: sustenta-se a hipótese de que a série pode ser adequadamente identificada a partir de um modelo ARIMA $(1, 1, 0)$.

- **Para a série temporal de recuperações diárias (Figura 22)**: para a série de recuperações diárias, depreendem-se análises perfeitamente análogas àquelas feitas para série histórica de casos diários. Dessa forma, sustenta-se que tal conjunto de dados pode ser pertinentemente associado a um modelo ARIMA $(1, 1, 0)$.

- **Para a série temporal de ocupação de leitos:**

  - Série com uma diferença: o gráfico de autocorrelação parcial mostra que há uma queda abrupta dessa magnitude a partir do primeiro atraso. Adicionalmente, verifica-se que a função de autocorrelação assume a forma de uma função exponencial decrescente, também a partir do primeiro atraso.
  - Série com duas diferenças: há algumas possibilidades de interpretação dos correlogramas concernentes à série histórica de ocupação de leitos com duas diferenças. No entanto, para todas as interpretações, valores máximos para p e q seriam de duas unidades. A exemplo, pode-se inferir, por meio de análise gráfica da função de autocorrelação parcial, que o ponto de corte da referida função ocorre a partir do segundo atraso. Ainda, a função de correlação adquire aspecto forma similar à forma da função de autocorrelação parcial, indicando que termos de média móvel não são cruciais para a explicação do conjunto de dados.
  - Conclusões: supõe-se que modelos ARIMA $(1, 1, 0)$ e ARIMA $(2, 2, 0)$ são sugestões adequadas à identificação da série histórica.

Ressalte-se que como os modelos sugeridos anteriormente referem-se às séries temporais completas de cada variável, faz-se adequado a experimentação de outros modelos num cenário em que é utilizada a abordagem de janelas deslizantes, cada qual referida a uma série temporal única munida de 150 observações. Tendo em vista que, para todas as variáveis, foram aplicados até dois níveis de diferenças; foi utilizado um padrão de uso de até sete atrasos para as variáveis preditoras; bem como há indícios de que os termos de média móvel não explicam adequadamente as séries observadas; o sistema de experimentação do modelo ARIMA no presente trabalho terá a seguinte configuração.

- Serão testados modelos ARIMA $(p, 1, 0)$, em que p varia de 1 a 7.
- Modelos ARIMA $(p, 2, 0)$, com p variando de 1 a 7, também serão experimentados.
- Considerando a inclusão de teste com termo não nulo de média móvel, foram testados modelos do tipo ARIMA $(p, 1, 1)$, com variação de 1 a 7 para o parâmetro p.

Por meio da sistematização supracitada, a experimentação do modelo ARIMA consistirá de 21 tipos de ajuste do modelo para cada variável a ser predita.

3.4.3 NARX e NARMAX Polinomiais

A construção dos modelos NARX e NARMAX de mapeamento polinomial, discutidos na 2.5.2 e aprofundado em (AGUIRRE, 2004), fundamentou-se nas seguintes etapas.

1. A partir de um grau de não-linearidade l, bem como o número de atrasos considerados para cada variável constituinte do modelo, gera-se o conjunto de todos os possíveis regressores.
2. Os termos candidatos são classificados conforme o algoritmo dos Mínimos Quadrados Ortogonais - MQO, apresentado em (AGUIRRE, 2004), o qual faz o cálculo dos parâmetros associados aos regressores.
3. Calcula-se o *error reduction ratio* (ERR), retratado na Seção 2.5.2.2, associado a cada regressor, com base em algoritmo apresentado em (AGUIRRE, 2004).

4. Fixa-se uma quantidade de termos, já elencados de maneira decrescente em termos de importância pelo ERR, a serem avaliados pelo critério de informação de Akaike (AIC), descrito na Seção 2.5.2.3.

5. O número de regressores incorporados ao modelo relaciona-se com o ponto de corte relativo ao critério de informação adotado.

Dessa forma, os experimentos de predição são sistematizados da forma seguintes.

- Utiliza-se, como preditores, as séries temporais atrasadas em até sete dias.
- Graus de não linearidade 2, 3 e 4 são considerados.
- Para o modelo NARMAX, os termos de ruído seguem as condições supracitadas. Ou seja, são usados até sete atrasos dos termos do ruído, com graus de não linearidade 2, 3 e 4.
- Um valor máximo possível de 15 termos é utilizado para cálculo do AIC.

Dessa forma, serão efetuados 21 cenários de previsão referentes à cada tipo de modelo (NARX e NARMAX), para as quatro variáveis a serem preditas.

No âmbito da expressão matemática dos modelos NARX polinomiais obtidos, será utilizado o seguinte padrão de notação: $m(k)$, $r(k)$ e $l(k)$ referem-se, respectivamente, às quantidades à tempo discreto associadas às quantidades de mortes, pacientes recuperados e aos leitos SUS ocupados. Já $y(k)$ denota a quantidade em tempo discreto da variável de saída identificada.

4 RESULTADOS E DISCUSSÃO

Visando uma discussão mais abrangente dos resultados obtidos, serão elaboradas duas formas de análise concernentes aos resultados obtidos da aplicação da metodologia de predição descrita no Capítulo 3.

1. Uma mais geral, onde se busca discutir os principais aspectos de todos os modelos preditivos concebidos.
2. Outro debate enfatizará os melhores cenários de predição a partir de ARIMA, NARX e NARMAX, criados para cada variável de interesse. Ressalte-se que os melhores casos de previsão relacionam-se com aqueles que resultaram em melhores valores relativos às métricas de desempenho.

4.1 ANÁLISE GERAL DOS MODELOS

A Tabela 23, situada no Apêndice B, sumariza o desempenho métrico, quanto às métricas de desempenho MAE e RMSE, resultante da aplicação dos modelos preditivos para as séries temporais diárias atinentes aos casos, óbitos, recuperações e ocupação de leitos.

Em conformidade com os resultados descritos na Seção 4.1, a Tabela 6 evidencia os melhores casos preditivos abordados por este trabalho.

Tabela 6 – Resultados métricos (MAE e RMSE) para os melhores cenários preditivos.

Série	Modelo	Descrição	MAE	RMSE
Casos	ARIMA	$p = 1$, $d = 1$, $q = 0$	88,679	129,538
	NARX	$l = 2$, qtd. de atrasos = 2	109,903	201,958
	NARMAX	$l = 2$, qtd. de atrasos = 2	122,703	200,583
Óbitos	ARIMA	$p = 5$, $d = 1$, $q = 0$	1,359	2,061
	NARX	$l = 2$, qtd. de atrasos = 2	1,550	2,455
	NARMAX	$l = 4$, qtd. de atrasos = 1	1,703	2,566
Recuperações	ARIMA	$p = 5$, $d = 1$, $q = 0$	48,016	70,728
	NARX	$l = 2$, qtd. de atrasos = 2	51,376	75,544
	NARMAX	$l = 2$, qtd. de atrasos = 2	51,844	76,559
Oc. de leitos	ARIMA	$p = 2$, $d = 2$, $q = 0$	0,986	1,265
	NARX	$l = 2$, qtd. de atrasos = 2	1,127	1,456
	NARMAX	$l = 2$, qtd. de atrasos = 4	1,437	2,084

Visando os resultados exibidos nessa tabela, faz-se uma análise pormenorizada para cada variável, a seguir.

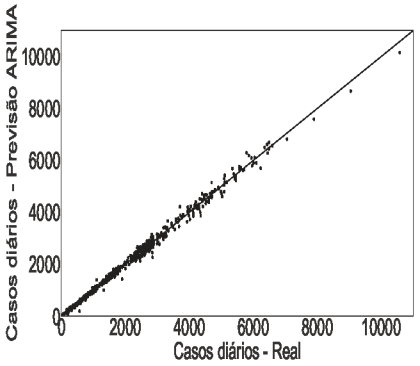
4.1.1 Casos Diários

Tendo em vista as métricas de desempenho utilizadas, tem-se que o modelo ARIMA (1,1,0) foi o que melhor se ajustou sobre o conjunto de dados concernente aos casos diários de infecção. Observa-se que, para os modelos ARIMA $(p, 1, 0)$, os desempenhos preditivos são bastante similares e relativamente superiores aos desempenhos observados para os modelos munidos de duas ordens de diferença ou modelos que incorporam parâmetro não nulo de média móvel. De fato, os modelos que incorporam termos de média móvel são aqueles com pior desempenho, no que se refere às métricas aplicadas. Dessa forma, infere-se que esse conjunto de dados não é bem explicado por termos de média móvel e que a suposição inicial, de que um modelo ARIMA (1,1,0) se ajusta adequadamente a essa série histórica, estava correta. Para o modelos NARX e NARMAX experimentados para essa série, verificou-se similaridade quanto aos ajustes que acarretaram melhores resultados métricos. De fato, para ambos, os modelos mais eficientes nesse aspecto se utilizam de um grau de não linearidade l=2 e incorporam até dois atrasos nas variáveis empregadas no modelo. De modo geral, para ambos os modelos, verificou-se as seguintes características:

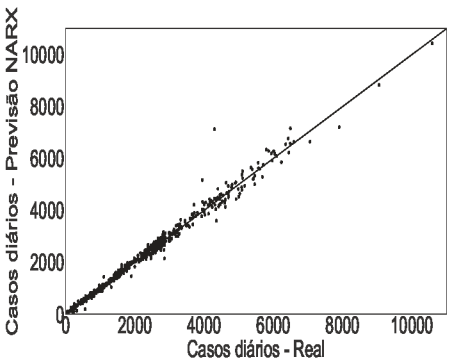
- para modelos, com mesmo grau de não linearidade aplicado, o desempenho piora com o aumento de atrasos incorporados pelo modelo;
- à medida que se aumenta o grau de não linearidade dos modelos, reduz-se sua eficiência preditiva.

Complementarmente, observa-se que, para modelos NARX e NARMAX análogos, i.e., munidos de parâmetros similares de linearidade e atraso contido nas variáveis, tem-se que os modelos NARX são ligeiramente superiores em termos métricos.

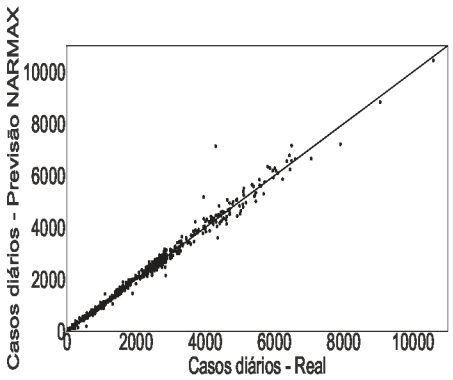
A Figura 9 mostra os diagramas de dispersão referentes aos melhores cenários preditivos no âmbito da previsão dos casos diários de COVID-19.



(a) Previsão ARIMA.



(b) Previsão NARX.



(c) Previsão NARMAX.

Figura 9 – Diagramas de dispersão para previsões de casos diários.

A partir de análise dos diagramas de dispersão, verifica-se melhor ajuste para o modelo preditivo ARIMA, em comparação com os modelos de previsão não lineares, visto que esses, nitidamente, possuem pontos distantes da curva associada a uma função linear ideal. De fato, os coeficientes de correlação linear obtidos para as previsões ARIMA, NARX e NARMAX, em relação série temporal real de casos diários de infecção são, respectivamente: 0,9969; 0,9926 e 0,9927. Ou seja, mostra-se, numericamente, que as previsões pontuais ARIMA são mais precisas, sendo que as previsões NARX e NARMAX obtiveram desempenhos muito próximos.

4.1.2 Óbitos Diários

Para a série temporal concernente às mortes diárias por COVID-19, o modelo ARIMA que melhor se ajustou, conforme as métricas de desempenho adotadas, tem

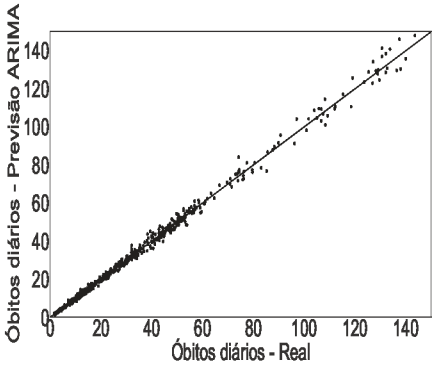
parâmetros $p = 5$, $d = 1$ e $q = 0$. Ou seja, supõe-se que, para esse conjunto de dados, em média, a utilização de cinco atrasos da própria variável predita, é eficaz para o processo de previsão. Nota-se, ainda, quanto aos modelos ARIMA, que a inserção de duas ordens de diferença acarreta em perda de capacidade preditiva. Ainda, tem-se que, em média, não há considerável diferença entre os modelos do tipo ARIMA $(p, 1, 0)$ e ARIMA $(p, 1, 1)$. Ou seja, a inserção de um termo de média móvel no modelo não gera alteração substancial no que diz respeito às magnitudes das métricas de desempenho.

Quanto aos modelos preditivos do tipo NARX e NARMAX, segue-se que:

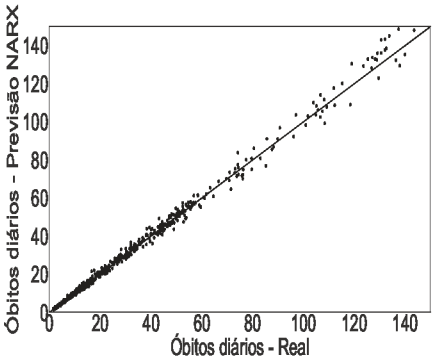
- o melhor cenário de previsão para o modelo NARX, quanto às métricas de desempenho, refere-se ao modelo com grau de não linearidade de dois níveis e o uso de até dois atrasos nas variáveis incorporadas pelo modelo;
- também alcançam valores relativamente bons para as métricas de desempenho os modelos que se utilizam de grau de não linearidade de quatro níveis para o NARMAX, com uma ou duas defasagens para cada variável utilizada;
- de maneira geral, considerando-se os cenários abordados, para modelos com mesmo grau de não linearidade, o desempenho desse modelos piora com a incorporação de atrasos maiores para as variáveis.

Em suma e de modo geral, tendo em vista o desempenho métrico dos modelos, a modelagem ARIMA acarretou em melhores valores para as métricas MAE e RMSE, comparada aos cenários de modelagem não linear implementadas. Em média, a modelagem autorregressiva de caráter não linear seguiu, em termos métricos, os principais aspectos vistos no âmbito da previsão dos casos diários de infecção, sendo que os cenários de previsão NARX podem ser ditos superiores aos casos de implementação do modelo NARMAX.

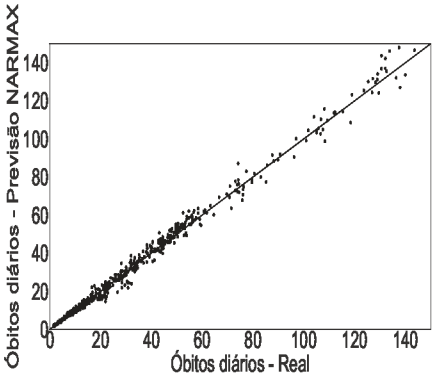
Considerando os melhores modelos obtidos via análise de desempenho métricos, tem-se os diagramas de dispersão mostrados na Figura 10.



(a) Previsão ARIMA.



(b) Previsão NARX.



(c) Previsão NARMAX.

Figura 10 – Diagramas de dispersão para previsões de mortes diárias.

Em suma, os valores de correlação linear para esses cenários são 0,9977; 0,9970 e; 0,9965. Os valores são referidos às previsões ARIMA, NARX e NARMAX, respectivamente. De fato, esses resultados numéricos mostram ligeira superioridade do método linear de identificação. Complementarmente, o modelo NARX moustra-se mais preciso que o modelo NARMAX.

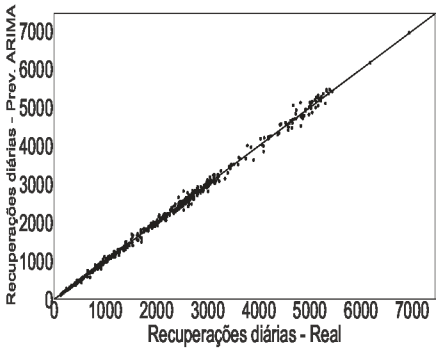
4.1.3 Recuperações Diárias

Dentre os modelos ARIMA experimentados, o cenário que acarretou melhor previsão, de acordo com as métricas de desempenho analisadas, refere-se àquele munido de parâmetros $p = 5$ e $d = 1$, desprovido de termo de média móvel. No entanto, nota-se que, entre os modelos que não aplicam termos de média móvel, a diferença dos resultados métricos é pouco considerável, não havendo, de fato, um modelo de

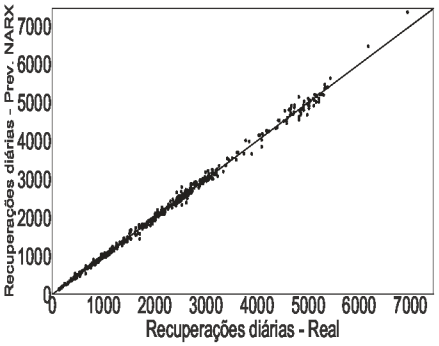
grande destaque entre os testados. Ainda, verifica-se que os modelos caracterizados por dois níveis de diferenciação geram piores resultados que os que se utilizam de apenas um nível.

Quanto aos modelos de identificação não linear, NARX e NARMAX, aqueles que implementam um grau de não linearidade $l = 2$ e duas defasagens para as variáveis de modelo, são os mais destacados no âmbito dos resultados métricos. De forma semelhante ao que ocorre na previsão dos casos diários de infecção, tem-se que, para modelos NARX e NARMAX com determinado grau de não linearidade, a capacidade preditiva é deteriorada ao se incorporar mais defasagens nas variáveis. Ainda, com o aumento do grau de não linearidade, também há redução da aptidão preditiva. Por fim, de forma geral ambos os modelos não lineares forneceram resultados semelhantes no âmbitos das métricas de desempenho abordadas. No entanto, os casos de previsão a partir da técnica ARIMA se mostraram um pouco mais precisos.

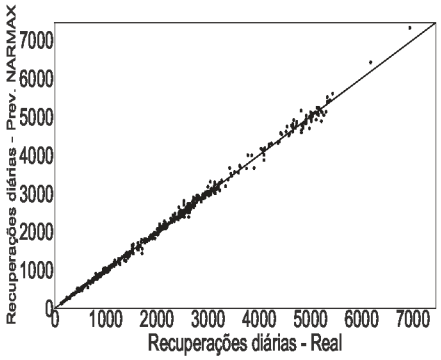
Tendo em vista os melhores modelos obtidos, quanto à análise do MAE e RMSE, tem-se em conta os diagramas de dispersão mostrados na Figura 11.



(a) Previsão ARIMA.



(b) Previsão NARX.



(c) Previsão NARMAX.

Figura 11 – Diagramas de dispersão para previsões de recuperações diárias.

Sendo que os coeficientes de correlação para os cenários preditivos ARIMA, NARX e NARMAX são, respectivamente, 0,9987; 0,9985 e 0,9984. De fato, os modelos mostram correlações lineares bastante próximas, notando-se ligeira superioridade para a identificação ARIMA.

4.1.4 Ocupação de Leitos

O modelo ARIMA que apresenta menores magnitudes relativas às métricas MAE e RMSE, ao longo de todo o processo preditivo, é constituído de duas ordens de diferenciação, bem como duas ordens de autorregressão. Ainda, esse modelo não incorpora termos de média móvel. Em aspecto geral, os modelos do tipo ARIMA $(p, 2, 0)$ possuem maior grau de eficiência preditiva, quanto às métricas abordadas, em comparação com modelos ARIMA $(p, 1, 0)$. Já os modelos que incluem um termo de média

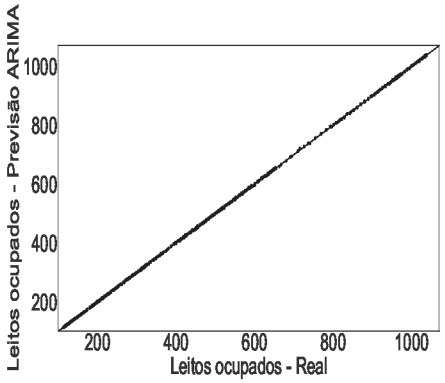
móvel, representados pela estrutura ARIMA $(p, 1, 1)$ obtiveram valores piores para as métricas de desempenho, em comparação com as outras vertentes supracitadas. Posto isso, supõe-se que a série temporal concernente à ocupação de leitos não é, ao menos para os cenários preditivos aqui estudados, bem representado por termos de média móvel, assim ocorre com a outras variáveis preditas. Adicionalmente, pode-se dizer que essa série temporal necessita de dois graus de diferenciação para gerar preditores ARIMA mais eficientes.

O modelo NARX, por sua vez, apresentou melhores características preditivas ao longo do procedimento de previsão, para um cenário em que há incorporação de um grau de não linearidade $l = 2$, com aplicação de até dois atrasos para as variáveis implementadas nos modelos.

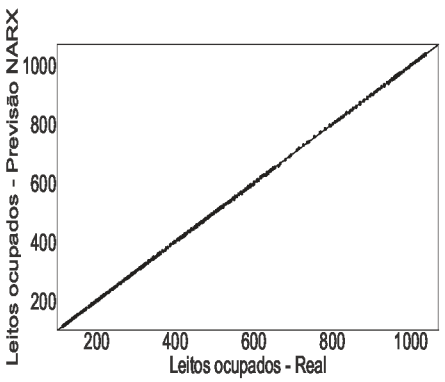
Já em relação à modelagem NARMAX, verifica-se que dois dos melhores modelos referem-se àqueles que se utilizam de grau de não linearidade $l = 2$, com incorporação de 4 e 6 atrasos para as variáveis preditoras endógenas e exógenas.

Em termos mais abrangentes, a modelagem linear acarretou, em média, previsões mais precisas comparada às técnicas de identificação não lineares. Adicionalmente e de modo geral, verifica-se superioridade preditiva da modelagem NARX em comparação à modelagem NARMAX.

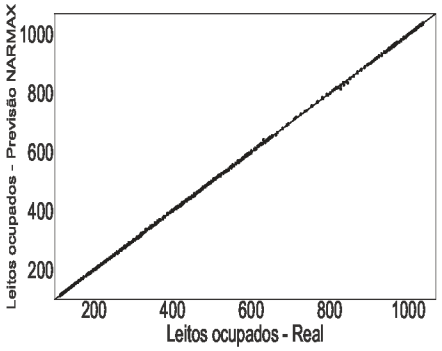
Para os melhores casos de previsão, obtiveram-se os diagramas de dispersão apresentados na Figura 12.



(a) Previsão ARIMA.



(b) Previsão NARX.



(c) Previsão NARMAX.

Figura 12 – Diagramas de dispersão para previsões de ocupação de leitos.

Por meio de análise dos diagramas de dispersão, verifica-se uma correlação linear aproximadamente perfeita para todas as técnicas de previsão utilizadas, indicando considerável aptidão preditiva em relação à série temporal de ocupação de leitos. De fato, para os três casos a correlação linear alcançou coeficientes de 0,9999.

4.1.5 Considerações Gerais

Levando-se em conta todos os modelos aplicados, infere-se que, no escopo experimental deste trabalho, o cenários preditivos mais eficientes estão associados aos modelos ARIMA, visto que, ao longo de todo processo de previsão, essa técnica apresentou melhores valores para as métricas MAE e RMSE, se comparada com aqueles resultantes da aplicação dos modelos NARX e NARMAX de mapeamento polinomial.

Tendo em vista os melhores cenários de previsão para as três técnicas aplicadas no presente estudo, descritas na Tabela 6, verifica-se as seguintes características.

- Série de casos diários: o menor MAE obtido é de 88,68 casos por dia, referente ao modelo ARIMA $(1, 1, 0)$, seguido por MAEs de valores 109,90 e 122,70; atinentes aos melhores preditores NARX e NARMAX, respectivamente. Tendo em vista as características dessa série temporal, dadas pela Tabela 3, o erro médio absoluto obtido pelo modelo ARIMA representa cerca de 4,45% da magnitude média dos casos diários reais, em comparação com 5,51% e 6,15%, para os preditores NARX e NARMAX, respectivamente.
- Série de óbitos diários: o menor erro médio absoluto obtido para essa série temporal refere-se à aplicação do modelo ARIMA $(5, 1, 0)$ - MAE = 1,359 mortes diárias, seguidas por valores de 1,550 e 1,703 obtidos para os melhores cenários de predição NARX e NARMAX, respectivamente. Esse valores representam, respectivamente, 4,47%, 5,10% e 5,60% do valor médio de óbitos diários.
- Série de recuperações diárias: tendo em vista que o valor médio de recuperações diárias da série pré-processada é de 1848,70, tem-se que os erros médios absolutos para os melhores cenários de previsão representam 2,60%, 2,78% e 2,80% dessa quantidade, respectivamente para os melhores modelos ARIMA, NARX e NARMAX obtidos via análise de desempenho métrico.
- Série de ocupação de leitos: para essa série observou-se MAEs de valor 0,986; 1,127 e 1,437 para os melhores cenários de previsão obtidos. De fato, essas são magnitudes de erro médio ínfimas comparadas a magnitude média da série temporal de trabalho, a qual indica um valor médio de 460,70 leitos ocupados por dia.

Complementarmente, sustenta-se que a técnica NARMAX demonstra-se ligeiramente inferior ao NARX, utilizando-se como critério apenas os resultados métricos abordados. Aliando essa informação ao fato de que os modelos cenários ARIMA mais eficientes não se utilizaram de termos de média móvel, deduz-se que as séries temporais analisadas tratados no presente trabalho não são efetivamente explicadas ou formadas a partir de processos de média móvel. Supõe-se que isso decorra do fato de que as séries temporais cruas foram submetidas a um processo de tratamento de ruído a partir do filtro de Savitsky-Golay, o qual pode ser interpretado como a generalização de um filtro de média móvel.

4.2 ANÁLISE DOS MELHORES CENÁRIOS DE PREDIÇÃO

Os modelos preditivos associados aos melhores resultados métricos, devidamente descritos na Tabela 6, são o cerne das análises seguintes, as quais pretendem investigar intervalos nos quais os modelos demonstraram maior dificuldade para efetuar as predições. Na sequência desta seção, as previsões de cada variáveis serão analisadas de forma isolada, de modo que alguns intervalos preditivos serão investigados acerca das características dos modelos preditivos concebidos ao longo do processo de previsão.

No decurso da presente seção, os modelos preditivos se utilizarão da seguinte denotação:

- $y(t)$ refere-se à variável de interesse (aquela que se pretende identificar);
- $c(t)$, $m(t)$, $r(t)$ e $l(t)$ referem-se, respectivamente, às variáveis relativas às séries temporais de casos, óbitos, recuperações e ocupação de leitos.

4.2.1 Casos Diários

Para a série histórica de casos de infecção por COVID-19, considerou-se cinco regiões de interesse, descritas a seguir:

1. Região 1, que compreende o intervalo 10/08/2020 - 17/08/2020.
2. Região 2, que compreende o intervalo 06/02/2021 - 11/02/2021.
3. Região 3, que compreende o intervalo 14/01/2022 - 19/01/2022.

Essas regiões estão destacadas na Figura 24, situada no Apêndice C.

4.2.1.1 Região 1 (11/08/2020 - 17/08/2020)

O gráfico das previsões para a Região 1 é mostrado na Figura 13.

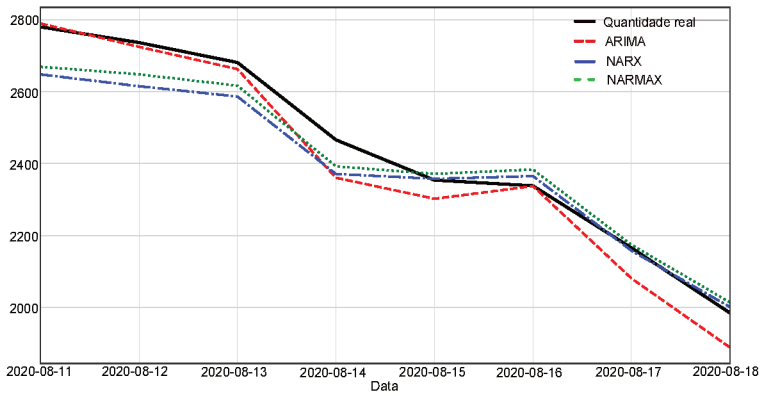


Figura 13 – Previsões para o período de tempo compreendido no intervalo 11/08/2020 - 18/08/2020.

Esse intervalo indica erro relevante associado ao dia 14/08/2020, para todos os modelos preditivos considerados, bem como desvios mais acentuados, em relação à curva real, para o modelo ARIMA.

Tendo em vista que o modelo ARIMA $(1,1,0)$ ajustado para esse intervalo é escrito conforme a Equação (25).

$$\Delta y(t) = \alpha \Delta y(t-1) + \beta. \tag{25}$$

onde $\Delta y(t) = y(t) - y(t-1)$, a evolução dos parâmetros α e β dos modelos é exibida na Tabela 7.

Tabela 7 – Evolução do modelo ARIMA para a previsão associada à Região 1 (casos diários).

		Data de previsão pontal						
Modelo	**Par.**	**11/08/20**	**12/08/20**	**13/08/20**	**14/08/20**	**15/08/20**	**16/08/20**	**17/08/20**
ARIMA								
	α	0,48	0,48	0,49	0,52	0,52	0,53	0,52
	β	18,73	18,42	17,38	14,52	14,71	14,69	15,3
	Erro Abs.	9,52	11,76	18,24	105,16	51,71	0,70	84,61

Em termos gerais, o modelo ARIMA não apresenta variação substancial para o parâmetro α. No entanto, há variação substancial do coeficiente β, associado ao termo constante, a partir do dia 14/08/2020, quando esse termo assume um valor de 14,52. É justamente nesse ponto em que há previsão concernente ao modelo ARIMA

adquire erro absoluto máximo, de aproximadamente 105 casos. Por outro lado, para esse intervalo de predição, os modelos NARX e NARMAX podem ser escritos de acordo com a Equação (26).

$$\begin{aligned} y(t) = {} & \theta_1 y(t-1) \\ & + \theta_2 y(t-2) \\ & + \theta_3 r(t-2) \\ & + \theta_4 y(t-1)y(t-2) \\ & + \theta_5 y(t-2)l(t-1) \\ & + \theta_6 y(t-2)^2 \\ & + \theta_7 e(t-1). \end{aligned} \tag{26}$$

Onde θ_n, $n \in [1, 2, \ldots, 7]$ denota os parâmetros associados aos regressores. Ressalte-se que apenas o modelo NARMAX contém termos de ruído, representado pela variável *e*. Desse modo, a evolução dos parâmetros concernentes aos modelos NARX e NARMAX são mostrados na Tabela 8.

Tabela 8 – Evolução dos modelos NARX e NARMAX para a previsão associada à Região 1 (casos diários).

		Data de previsão pontual						
Modelo	**Par.**	**11/08/20**	**12/08/20**	**13/08/20**	**14/08/20**	**15/08/20**	**16/08/20**	**17/08/20**
NARX								
	θ_1	1,34	1,36	1,39	1,40	1,41	1,41	1,41
	θ_2	-0,41	-0,43	-0,45	-0,46	-0,46	-0,46	-0,46
	θ_3	-0,00021	-0,00020	-0,00019	0,00020	-0,00019	-0,00019	-0,00018
	θ_4	0,0014	0,0013	0,0013	0,0014	0,0013	0,0013	0,0012
	θ_5	-0,16	-0,14	-0,12	—-	—-	—-	-0,11
	θ_6	——	——	——	-0,015	-0,014	-0,013	——
	Erro abs.	131,97	121,28	94,56	94,78	3,65	26,91	7,66
NARMAX								
	θ_1	1,46	1,45	1,40	1,38	1,39	1,39	1,39
	θ_2	-0,51	-0,49	-0,44	-0,41	-0,43	-0,43	-0,42
	θ_3	-0,00017	-0,00017	-0,00017	0,00018	-0,00018	-0,00018	-0,00017
	θ_4	0,0012	0,0011	0,0011	0,0013	0,0013	0,0012	0,0011
	θ_5	-0,14	-0,13	-0,11	—-	—-	—-	-0,12
	θ_6	——	——	——	-0,016	-0,015	-0,015	——
	θ_7	-0,0015	-0,0015	-0,0017	-0,0017	-0,0017	-0,0017	-0,0016
	Erro abs.	110,83	88,33	64,45	73,66	17,41	45,42	8,95

Por meio de análise da Tabela 8, também é observada uma mudança de padrão na estruturação dos modelos NARX e NARMAX, os quais dependiam até então dos termos associados aos cinco primeiros coeficientes. Nesse ponto, há independência do

termo $\theta_5 y(t-2)I(t-1)$ e incorporação do termo $y(t-2)^2$ aos modelos, o que pode explicar a variação brusca desses modelos no referido ponto de predição. Sugere-se que algum evento associado à pandemia possa ter alterado o perfil dinâmico de identificação a partir desses modelos. Ainda, a mudança de perfil paramétrico pode ter sido ocasionada devido à alteração da curva atinente à quantidade real (série pré-processada) de casos diários que, a partir do dia 14/08/2020 ao dia 16/08/2020 apresentou decremento em sua inclinação.

4.2.1.2 Região 2 (06/02/2021 - 11/02/2021)

O gráfico das previsões para a Região 2 é mostrado na Figura 14.

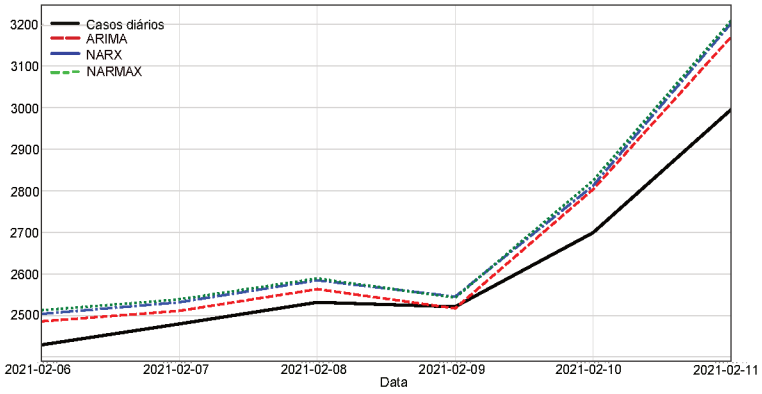


Figura 14 – Previsões para o período de tempo compreendido no intervalo 06/08/2021 - 11/02/2021.

Na Região 2, destaca-se o desvio, ao final do período, dos modelos em relação ao conjunto real de dados. Ainda, verifica-se comportamento similar para todos os modelos, sendo que a previsões obtidas via modelagem ARIMA são as que mais se aproximam das observações reais.

A evolução dos modelos ARIMA ajustados para efetuação das previsões pontuais desse intervalo é mostrada na Tabela 9.

Tabela 9 – Evolução do modelo ARIMA para a previsão associada à Região 2 (casos diários).

		Data de previsão pontual					
Modelo	**Par.**	**06/02/21**	**07/02/21**	**08/02/21**	**09/02/21**	**10/02/21**	**11/02/21**
ARIMA							
	α	0,57	0,57	0,57	0,57	0,57	0,58
	β	7,45	6,48	5,34	4,63	7,93	10,49
	Erro Abs.	11,88	33,62	52,94	43,64	196,59	6,56

Nota-se, visando os dados da Tabela 9, que o parâmetro α, associado ao primeiro atraso da série diferenciada, é quase constante ao longo do intervalo preditivo, sendo que o termos constante, i.e., o β sofre variações nítidas com o passar dos dias. Além de o termo constante apresentar valores positivos relativamente altos para os dois últimos pontos do intervalo, o aspecto constante do parâmetro α também pode ter influenciado na característica da curva. Sugere-se que, para um cenário mais adequado de previsão, esse parâmetro teria seu valor reduzido, sobretudo nos últimos pontos do intervalo, de forma a aproximar as previsões dos dados reais.
Já os modelos NARX e NARMAX concebidos para a previsão dessa região podem ser escritos pela Equação (27).

$$\begin{aligned} y(t) = {} & \theta_1 y(t-1) \\ & + \theta_2 y(t-2) \\ & + \theta_3 y(t-2) m(t-1) \\ & + \theta_4 y(t-2) m(t-2) \\ & + \theta_5 m(t-1) \\ & + \theta_6 e(t-1). \end{aligned} \tag{27}$$

A variação desses parâmetros no decorrer da efetuação das previsões para os pontos da Região 2 é apresentada na Tabela 10.

Tabela 10 – Evolução dos modelos NARX e NARMAX para a previsão associada à Região 2 (casos diários).

		Data da previsão pontual					
Modelo	**Par.**	**06/02/21**	**07/02/21**	**08/02/21**	**09/02/21**	**10/02/21**	**11/02/21**
NARX							
	θ_1	1,44	1,48	1,48	1,48	1,47	1,47
	θ_2	-4,00	-4,49	-4,56	-4,47	-4,46	-4,45
	θ_3	-0,000036	-0,0000078	-0,0000075	-0,0000079	-0,0000079	-0,0000079
	θ_4	0,00029	—	—	—	—	—
	θ_5	-0,69	—	—	—	—	—
	Erro abs.	74,48	52,04	53,02	24,27	112,99	207,13
NARMAX							
	θ_1	1,64	1,52	1,53	1,53	1,55	1,56
	θ_2	-6,30	-4,88	-4,97	-5,07	-5,26	-5,30
	θ_3	-0,000033	-0,0000080	-0,0000077	-0,0000077	-0,0000074	-0,0000073
	θ_4	0,00035	—	—	—	—	—
	θ_5	-0,89	—	—	—	—-	—-
	θ_6	0,0056	0,0063	0,0065	0,11	0,12	0,20
	Erro abs.	83,18	59,05	57,56	21,64	125,23	213,43

Nota-se, para os modelos NARX e NARMAX que, a partir do dia 07/02/2021, ambos os modelos adquirem um padrão estrutural que permanece até o final do intervalo em questão. Sugere-se que manutenção dos valores dos parâmetros pode ser o fator responsável pelo incremento sucessivo nos valores de predição. Daí, as curvas atinentes a esses modelos se desviam consideravelmente da curva associada aos dados reais.

4.2.1.3 Região 3 (14/01/2022 - 19/01/2022)

Graficamente, as predições realizadas para a Região 3 são mostradas na Figura 15.

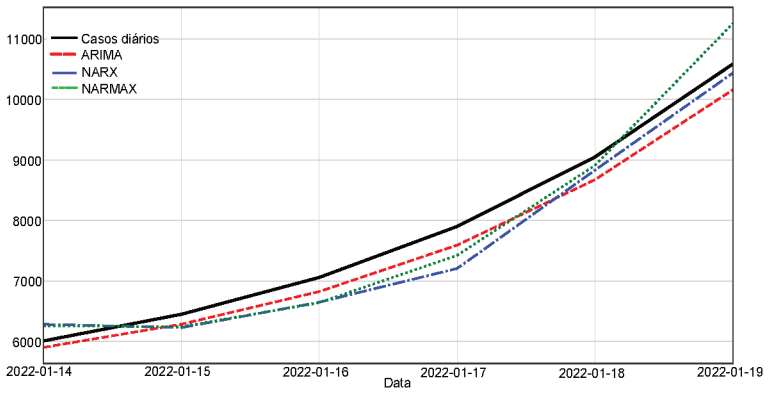


Figura 15 – Previsões para o período de tempo compreendido no intervalo 11/08/2020 - 18/08/2020.

A Região 3 mostra a porção final de todo o conjunto de previsão, quando a série temporal de casos diários adquire um crescimento considerável em seus valores. Nesse intervalo, observa-se um considerável desvio das predições via modelo NARMAX, sobretudo ao final do intervalo.

A variação dos modelos ARIMA na Região 3 é exibida na Tabela 11.

Tabela 11 – Evolução do modelo ARIMA para a previsão associada à Região 3 (casos diários).

		Data de previsão pontual					
Modelo	**Par.**	**14/01/22**	**15/01/22**	**16/01/22**	**17/01/22**	**18/01/22**	**19/01/22**
ARIMA							
	α	0,81	0,82	0,83	0,86	0,91	0,96
	β	33,24	39,58	47,96	57,36	84,38	173,39
	Erro Abs.	107,20	164,23	233,56	501,24	375,77	426,15

Em conformidade com as informações contidas na Tabela 11, nota-se aumento gradual de ambos os coeficientes do modelo, o que acarreta, consequentemente, no crescimento dos valores das predições, em concordância com a dinâmica real da série temporal.

Já os modelos NARX e NARMAX concebidos para a previsão dessa região podem ser

escritos conforme a Equação (28).

$$\begin{aligned} y(t) = {} & \theta_1 y(t-1) \\ & + \theta_2 y(t-2) \\ & + \theta_3 y(t-1) y(t-2) \\ & + \theta_4 y(t-1)^2 \\ & + \theta_5 y(t-1) m(t-1) \\ & + \theta_6 y(t-1) r(t-1) \\ & + \theta_7 r(t-1) \\ & + \theta_8 I(t-1) \\ & + \theta_9 e(t-1). \end{aligned} \tag{28}$$

A variação desses parâmetros no decorrer da efetuação das previsões para os pontos da Região 2 é apresentada na Tabela 12.

Tabela 12 – Evolução dos modelo NARX e NARMAX para a previsão associada à Região 3 (casos diários).

		Data da previsão pontual					
Modelo	**Par.**	**06/02/21**	**07/02/21**	**08/02/21**	**09/02/21**	**10/02/21**	**11/02/21**
NARX							
	θ_1	1,83	1,99	1,98	2,01	1,83	1,82
	θ_2	-0,68	-0,69	-0,89	-0,89	-0,90	-0,71
	θ_3	-0,00051	—	—	-0,00033	-0,00047	-0,00056
	θ_4	0,000076	—	—	—	—	—
	θ_5	—	0,000069	0,000071	0,000075	—	0,000053
	θ_6	—	—	—	—	0,000040	—
	θ_7	-0,000022	-0,000072	-0,000075	-0,000080	-0,000021	-0,000012
	θ_8	—	-0,0051	-0,0058	—	—	—
	Erro abs.	274,00	220,97	409,52	694,03	217,64	148,38
NARMAX							
	θ_1	2,33	2,31	1,98	2,28	2,26	2,00
	θ_2	-1,34	-1,28	-1,25	-1,28	-0,88	-1,59
	θ_3	-0,00019	—	—	-0,0000050	-0,00034	-0,00027
	θ_4	0,000078	—	—	—	—	—
	θ_5	—	0,00012	0,00017	0,000095	—	0,000015
	θ_6	—	—	—	—	0,000039	—
	θ_7	-0,000014	-0,000012	-0,000015	-0,000085	-0,000021	-0,000012
	θ_8	—	-0,0021	-0,0028	—	—	—
	θ_9	0,0062	0,0064	0,0064	0,021	0,035	0,092
	Erro abs.	246,60	205,32	415,84	477,42	138,71	669,40

De acordo com a Tabela 12, observa-se que, para o dia final do intervalo de predição, o modelo NARMAX incorpora o termo de ruído com coeficiente mais elevado

de todo o período. Possivelmente, isso explica o comportamento final das previsões associadas a esse modelo.

4.2.2 Mortes Diárias

As regiões de análise para a série temporal de óbitos diários estão destacadas em vermelho no gráfico mostrado na Figura 25, exposta no Apêndice C. Como se pode notar, são duas as regiões de análise escolhidas para a investigação da série temporal de óbitos diários, as quais são descritas a seguir.

1. Região 1, que compreende o intervalo 12/11/2020 - 18/11/2020.
2. Região 2, que compreende o intervalo 08/03/2021 - 13/03/2021.

Para a investigação da série histórica de óbitos diários associados ao COVID-19, um bom cenário preditivo referente à modelagem ARIMA, no que tange às métricas de desempenho, corresponde ao modelo com uma ordem de diferença, cinco ordens de autorregressão, bem como é desprovido de termos de média móvel. Nesse sentido, esse modelo será representado pela estrutura representada pela Equação (29).

$$\begin{aligned} \Delta y(t) = {} & \alpha_1 \Delta y(t-1) \\ & + \alpha_2 \Delta y(t-2) \\ & + \alpha_3 \Delta y(t-3) \\ & + \alpha_4 \Delta y(t-4) \\ & + \alpha_5 \Delta y(t-5) \\ & + \beta. \end{aligned} \tag{29}$$

4.2.2.1 Região 1 (12/11/2020 - 18/11/2020)

Em termos visuais, as previsões realizadas no decurso da região podem ser visualizadas na Figura 16.

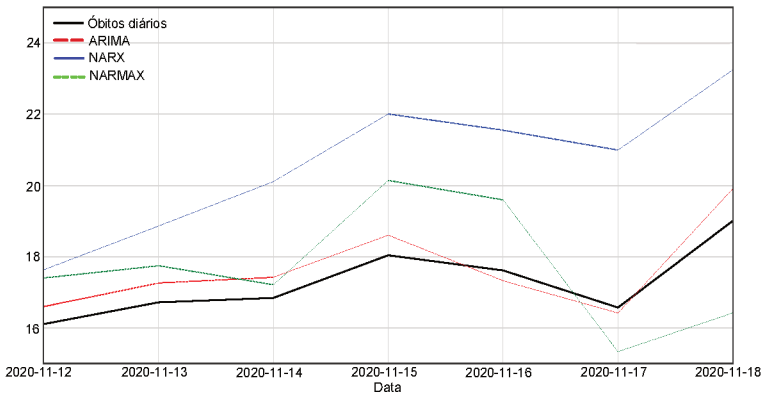


Figura 16 – Previsões para o período de tempo compreendido no intervalo 12/11/2020 - 18/11/2020.

A figura nos mostra um comportamento relativamente errôneo angariado via modelagem NARX e NARMAX. Enquanto os modelos NARX fazem predições de valores consideravelmente mais altos do que o ideal, os modelos NARMAX efetuam previsões com variações mais abruptas que o adequado. A previsão via modelagem ARIMA, por outro lado, mostrou-se a mais precisa, em termos gerais, para esse intervalo.

A evolução dos parâmetros ARIMA, ao longo do processo de previsão dessa região, é mostrada na Tabela 13.

Tabela 13 – Evolução do modelo ARIMA para a previsão associada à Região 1 (óbitos diários).

		Data de previsão pontual						
Modelo	**Par.**	**12/11/20**	**13/11/20**	**14/11/20**	**15/11/20**	**16/11/20**	**17/11/20**	**18/11/20**
ARIMA								
	α_1	0,036	0,036	0,035	0,035	0,035	0,035	0,034
	α_2	-0,031	-0,030	-0,034	-0,032	-0,029	-0,032	-0,038
	α_3	0,19	0,19	0,19	0,19	0,19	0,19	0,20
	α_4	0,067	0,067	0,069	0,072	0,067	0,067	0,064
	α_5	-0,099	-0,099	-0,098	-0,10	-0,10	-0,11	-0,10
	β	0,095	0,090	0,070	0,071	0,052	0,049	0,086
	Erro Abs.	0,49	0,54	0,58	0,56	0,29	0,15	0,90

Os modelos NARX identificados para a previsão da Região 1 podem ser expres-

sos pela Equação (30).

$$\begin{aligned} y(t) = {} & \theta_1 y(t-1) \\ & + \theta_2 y(t-2) \\ & + \theta_3 y(t-1) y(t-2) \\ & + \theta_4 y(t-2)^2 \\ & + \theta_5 y(t-1) c(t-1) \\ & + \theta_6 c(t-1). \end{aligned} \tag{30}$$

Note-se que a variação dos modelos NARX, ao longo do processo preditivo para esse intervalo, é mostrado na Tabela 14.

Tabela 14 – Evolução do modelo NARX para a previsão associada à Região 1 (óbitos diários).

		Data de previsão pontual						
Modelo	**Par.**	**12/11/20**	**13/11/20**	**14/11/20**	**15/11/20**	**16/11/20**	**17/11/20**	**18/11/20**
NARX								
	θ_1	1,14	1,14	1,14	1,08	1,08	1,10	1,11
	θ_2	-0,23	-0,23	-0,23	-0,22	-0,22	-0,23	-0,23
	θ_3	-0,00013	-0,00013	-0,00013	—	—	—	—
	θ_4	0,0000087	0,0000087	0,0000088	0,0000074	0,0000072	0,0000069	0,0000059
	θ_5	-0,00000074	-0,00000074	-0,00000073	-0,00000082	-0,00000083	-0,00000081	-0,00000076
	θ_6	—	—	—	0,00057	0,00055	0,00052	0,00046
	Erro Abs.	1,52	2,14	3,27	3,97	3,94	4,42	4,24

Por outro lado, os modelos NARMAX concebidos possuem a seguinte estruturação definida pela Equação (31).

$$\begin{aligned} y(t) = {} & \theta_1 y(t-1) \\ & + \theta_2 l(t-1) c(t-1) \\ & + \theta_3 l(t-1) r(t-1) \\ & + \theta_4 c(t-1) r(t-1) \\ & + \theta_5 r(t-1)^2 \\ & + \theta_6 c(t-1)^3 \\ & + \theta_7 c(t-1)^4 \\ & + \theta_8 r(t-1)^4 \\ & + \theta_9 c(t-1)^3 r(t-1) \\ & + \theta_{10} l(t-1) r(t-1)^3 \\ & + \theta_{11} e(t-1). \end{aligned} \tag{31}$$

Os modelos NARMAX concebidos para a predição da Região 2 são mostrados na Tabela 15.

Tabela 15 – Evolução do modelo NARMAX para a previsão associada à Região 1 (óbitos diários).

		Data de previsão pontual						
Modelo	**Par.**	**12/11/20**	**13/11/20**	**14/11/20**	**15/11/20**	**16/11/20**	**17/11/20**	**18/11/20**
NARMAX								
	θ_1	0,85	0,85	0,85	0,86	0,86	0,87	0,88
	θ_2	0,000017	0,000018	0,000018	0,000024	0,000025	0,000013	0,000015
	θ_3	-0,000011	-0,000011	-0,000011	-0,000018	-0,000019	-0,0000062	-0,000011
	θ_4	—	—	—	-0,0000032	-0,0000035	—	—
	θ_5	—	—	—	0,0000028	0,0000031	-0,0000023	—
	θ_6	$-2{,}65 \times 10^{-10}$	$-2{,}63 \times 10^{-10}$	$-2{,}63 \times 10^{-10}$	—	—	—	—
	θ_7	—	—	—	—	—	$-4{,}2 \times 10^{-14}$	—
	θ_8	$4{,}80 \times 10^{-14}$	$4{,}75 \times 10^{-14}$	$5{,}17 \times 10^{-14}$	—	—	—	—
	θ_9	—	—	—	—	—	—	$-8{,}3 \times 10^{-14}$
	θ_{10}	—	—	—	—	—	—	$4{,}1 \times 10^{-13}$
	θ_{11}	$5{,}21 \times 10^{-8}$	$4{,}90 \times 10^{-8}$	$4{,}91 \times 10^{-8}$	$6{,}22 \times 10^{-8}$	$6{,}26 \times 10^{-8}$	$4{,}31 \times 10^{-8}$	$5{,}21 \times 10^{-8}$
	Erro Abs.	1,29	1,03	0,37	2,10	1,98	1,24	2,56

O comportamento preditivo obtido via modelagem NARMAX, apesar de mais preciso em comparação com as previsões NARX, incorpora termos de ordem mais elevada - o que pode explicar as variações mais notáveis nas predições pontuais resultantes dos modelos NARMAX.

4.2.2.2 Região 2 (08/03/2021 - 13/03/2021)

Graficamente, as previsões da Região 2 para a série temporal podem ser vistas na Figura 17.

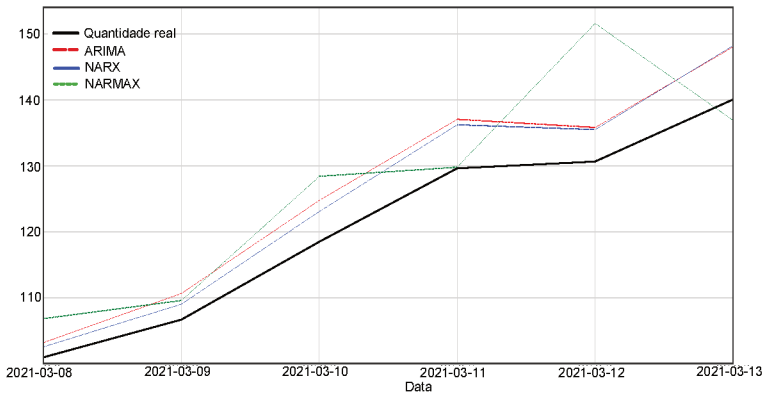


Figura 17 – Previsões para o período de tempo compreendido no intervalo 08/03/2021 - 13/03/2021.

Para esse intervalo, é destacado o desvio cometido pela modelagem NARMAX no penúltimo ponto de predição, correspondente ao dia 11/03/2021. Adicionalmente, é interessante notar a similaridade entre as curvas obtidas via modelos NARX e ARIMA, sendo que a previsão que mais se aproxima da curva real durante quase a totalidade do intervalo relaciona-se com a modelagem NARX.

O modelo ARIMA, para essa região, se comporta em conformidade com dados mostrados na Tabela 16.

Tabela 16 – Evolução do modelo ARIMA para a previsão associada à Região 2 (óbitos diários).

		Data de previsão pontual					
Modelo	**Par.**	**08/03/21**	**09/03/21**	**10/03/21**	**11/03/21**	**12/03/21**	**13/03/21**
ARIMA							
	α_1	0,25	0,27	0,32	0,37	0,033	0,29
	α_2	0,03	0,06	0,08	0,09	0,03	0,07
	α_3	0,29	0,28	0,34	0,34	0,34	0,38
	α_4	0,07	0,06	0,03	0,05	0,06	0,07
	α_5	0,03	0,02	-0,01	-0,03	-0,04	-0,04
	β	0,70	0,79	1,03	1,26	0,98	1,14
	Erro Abs.	2,23	3,96	6,27	7,43	5,15	7,98

Os modelos NARX identificados para a previsão da Região 2 podem ser expres-

sos pela Equação (32).

$$\begin{aligned} y(t) = {} & \theta_1 y(t-1) \\ & + \theta_2 y(t-2) \\ & + \theta_3 y(t-1) y(t-2) \\ & + \theta_4 y(t-2) r(t-1) \\ & + \theta_5 y(t-1) r(t-2) \\ & + \theta_6 I(t-1) \\ & + \theta_7 r(t-1) \\ & + \theta_8 r(t-2) \\ & + \theta_9 u(t). \end{aligned} \tag{32}$$

Onde $u(t)$ denota o degrau unitário em tempo discreto. Note-se que a variação dos modelos NARX, ao longo do processo preditivo para esse intervalo, é mostrado na Tabela 17.

Tabela 17 – Evolução do modelo NARX para a previsão associada à Região 2 (óbitos diários).

		Data de previsão pontual					
Modelo	**Par.**	**08/03/21**	**09/03/21**	**10/03/21**	**11/03/21**	**12/03/21**	**13/03/21**
NARX							
	θ_1	0,97	0,96	0,91	0,84	0,73	0,54
	θ_2	-0,33	-0,37	-0,26	-0,19	-0,20	0,000013
	θ_3	—-	—-	—-	—-	0,00046	0,00083
	θ_4	0,000021	0,000019	0,000027	0,000027	—-	—-
	θ_5	0,00019	0,00024	—-	—-	—-	-0,0000023
	θ_6	—-	—-	0,0014	—-	—-	—-
	θ_7	—-	—-	—-	—-	0,011	0,0090
	θ_8	—-	—-	—-	0,0015	—-	—-
	θ_9	2,78	3,04	2,84	2,93	2,50	2,70
	Erro Abs.	1,58	2,34	4,57	6,60	4,82	8,17

Já os modelos NARMAX podem ser descritos a partir da Equação (33).

$$\begin{aligned} y(t) = {} & \theta_1 y(t-1) \\ & + \theta_2 c(t-1) \\ & + \theta_3 r(t-1) \\ & + \theta_4 l(t-1) \\ & + \theta_5 l(t-1)c(t-1) \\ & + \theta_6 y(t-1)^2 \\ & + \theta_7 l(t-1)^2 \\ & + \theta_8 l(t-1)^3 \\ & + \theta_9 y(t-1)^4 \\ & + \theta_{10} l(t-1)^4 \\ & + \theta_{11} l(t-1)^3 c(t-1) \\ & + \theta_{12} u(t) \\ & + \theta_{13} e(t-1). \end{aligned} \tag{33}$$

A variação dos modelos NARMAX ao longo das previsões pontuais da Região 2 é mostrada na Tabela 18.

Tabela 18 – Evolução do modelo NARMAX para a previsão associada à Região 2 (óbitos diários).

		Data de previsão pontual					
Modelo	**Par.**	**08/03/21**	**09/03/21**	**10/03/21**	**11/03/21**	**12/03/21**	**13/03/21**
NARMAX							
	θ_1	0,73	0,79	0,89	0,81	1,04	0,82
	θ_2	—-	0,0024	—-	0,0015	-0,20	0,0015
	θ_3	0,00045	—-	—-	—-	—-	—-
	θ_4	0,033	—-	—-	0,000027	—-	—-
	θ_5	—-	-0,0000039	—-	—-	—-	—-
	θ_6	—-	—-	-0,0023	—-	-0,0041	—-
	θ_7	-0,000039	—-	—-	—-	—-	—-
	θ_8	—-	$4{,}17 \times 10^{-8}$	$3{,}11 \times 10^{-8}$	—-	—-	—-
	θ_9	—-	—-	$1{,}60 \times 10^{-7}$	—-	$2{,}05 \times 10^{-7}$	—-
	θ_{10}	$5{,}20 \times 10^{-11}$	—-	—-	$3{,}10 \times 10^{-11}$	$3{,}54 \times 10^{-11}$	$6{,}52 \times 10^{-11}$
	θ_{11}	—-	—-	—-	$-7{,}15 \times 10^{-12}$	—-	$-7{,}14 \times 10^{-12}$
	θ_{12}	—-	-0,26	0,71	-0,20	0,18	-0,21
	θ_{13}	$5{,}21 \times 10^{-10}$	$5{,}44 \times 10^{-10}$	$3{,}22 \times 10^{-10}$	$5{,}81 \times 10^{-10}$	$5{,}74 \times 10^{-10}$	$3{,}48 \times 10^{-10}$
	Erro Abs.	5,87	2,90	9,92	0,19	20,95	3,09

Em suma, além de incorporar termos com ordens mais elevadas (terceira e quarta potências), o desvio para a previsão efetuada pelo modelo NARMAX, pode ser explicado pelo alto valor para o primeiro parâmetro ($\theta_1 = 1,04$), no dia 12/03/2021.

4.2.3 Recuperações Diárias

As regiões investigadas para a série histórica de recuperações diárias são destacadas na Figura 26, do Apêndice C. Essas regiões são alusivas aos seguintes intervalos.

- Região 1, que compreende o intervalo 19/03/2021 - 23/03/2021.
- Região 2, que compreende o intervalo 06/08/2021 - 11/08/2021.

Ressalte-se que o modelo ARIMA (5, 1, 0) será representado pela Equação (34).

$$\begin{aligned} \Delta y(t) = {} & \alpha_1 \Delta y(t-1) \\ & + \alpha_2 \Delta y(t-2) \\ & + \alpha_3 \Delta y(t-3) \\ & + \alpha_4 \Delta y(t-4) \\ & + \alpha_5 \Delta y(t-5) \\ & + \beta. \end{aligned} \tag{34}$$

4.2.3.1 Região 1 (19/03/2021 - 23/03/2021)

O gráfico mostrado na Figura 18, exibe as previsões atinentes à Região 1, para a série histórica de recuperações diárias.

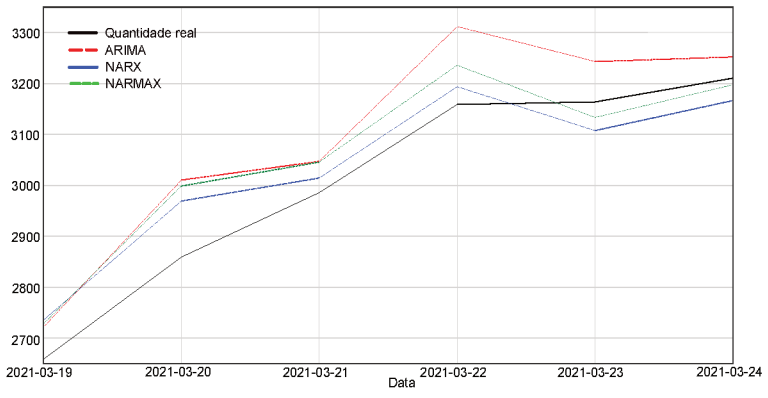


Figura 18 – Previsões para o período de tempo compreendido no intervalo 19/03/2021 - 24/03/2021.

Esse intervalo mostra previsões pontuais do modelo ARIMA munidas de magnitude superior em comparação às observadas para os modelos NARX e NARMAX. De fato, as previsão pontuais do modelo ARIMA para os dias 20/03/2021 e 22/03/2021 representam as predições que acarretaram maior erro absoluto para essa região de interesse. Adicionalmente, observa-se que as curvas das predições associadas aos modelos não lineares têm dinâmica parecida de crescimento e decrescimento, sendo que as predições NARMAX angariam valores maiores em comparação com as predições dos modelos NARX.

Para a Região 1, o modelo ARIMA tem seus parâmetros variados de acordo com os dados provenientes da Tabela 19.

Tabela 19 – Evolução do modelo ARIMA para a previsão associada à Região 1 (recuperações diárias).

		Data de previsão pontual					
Modelo	**Par.**	**19/03/21**	**20/03/21**	**21/03/21**	**22/03/21**	**23/03/21**	**24/03/21**
ARIMA							
	α_1	0,53	0,53	0,53	0,53	0,52	0,52
	α_2	0,05	0,05	0,05	0,06	0,07	0,06
	α_3	0,22	0,24	0,24	0,23	0,21	0,21
	α_4	0,15	0,14	0,14	0,14	0,15	0,14
	α_5	-0,23	-0,24	-0,24	-0,23	-0,23	-0,23
	β	22,91	28,06	28,87	31,59	28,18	26,95
	Erro Abs.	63,15	151,33	61,45	151,13	79,32	42,05

Conforme os dados vistos na Tabela 19, nota-se dois fatos que propiciaram magnitudes maiores para os dias associados aos maiores erros absolutos:

- no dia 20/03/2021, o modelo ARIMA incorporou um termo constante (β = 28,06) de valor substancialmente maior, em comparação ao incorporado pelo modelo identificado no dia anterior (β = 22,91), sendo esse o fator que mais tenha contribuído para o distanciamento da predição em relação à curva dos dados reais;
- de modo parecido com o evento supracitado, o modelo agregou maior valor ao termo constante no dia 22/03/2021, o que condiz com o crescimento positivo de casos apresentado pela curva dos dados reais. No entanto, essa elevação de valor do termo constante pode ter sido maior que a necessária e isso se alia ao fato de que o restante dos coeficientes se manteve basicamente inalterado, com valores similares aos do dia 21/03/2021. Um ajuste alternativo desses valo-

res poderia compensar o crescimento proporcionado pela magnitude do termo constante.

Os modelos NARX e NARMAX identificados para a Região 1 podem ser expressos pela Equação (35).

$$\begin{aligned} y(t) = {} & \theta_1 y(t-1) \\ & + \theta_2 y(t-2) \\ & + \theta_3 c(t-1) \\ & + \theta_5 y(t-2)c(t-1) \\ & + \theta_5 y(t-2)c(t-2) \\ & + \theta_6 e(t-1). \end{aligned} \tag{35}$$

A variação desses modelos, bem como os erros absolutos pontuais, são mostrados na Tabela 20.

Tabela 20 – Evolução dos modelos NARX e NARMAX para a previsão associada à Região 1 (recuperações diárias).

		Data da previsão pontual					
Modelo	**Par.**	**19/03/21**	**20/03/21**	**21/03/21**	**22/03/21**	**23/03/21**	**24/03/21**
NARX							
	θ_1	1,42	1,42	1,42	1,41	1,41	1,41
	θ_2	-0,51	-0,51	-0,51	-0,50	-0,50	-0,50
	θ_3	0,083	0,081	0,079	0,081	0,081	0,081
	θ_4	-0,020	-0,021	-0,021	-0,022	-0,023	-0,023
	θ_5	0,0025	0,0027	0,0027	0,0028	0,0029	0,0029
	Erro abs.	76,87	110,13	28,84	34,99	56,29	43,89
NARMAX							
	θ_1	1,52	1,57	1,56	1,56	1,54	1,55
	θ_2	-0,59	-0,63	-0,63	-0,63	-0,61	-0,62
	θ_3	0,068	0,061	0,060	0,063	0,065	0,063
	θ_4	-0,016	-0,016	-0,017	-0,017	-0,018	-0,018
	θ_5	0,0022	0,0021	0,0021	0,0022	0,0023	0,0024
	θ_6	0,00021	0,00022	0,00021	0,00021	0,00021	0,00021
	Erro abs.	69,20	139,56	60,13	77,13	30,48	12,62

Tendo em vista as informações contidas na Tabela 20, observa-se que ambos os modelos, NARX e NARMAX, conservam a estrutura do modelo à medida que as previsões são feitas, para toda a Região 1.

Infere-se, ainda, que as predições associadas ao NARMAX possuem maior valor que as previsões dos modelos NARX por conta de uma maior magnitude, em termos

gerais, dos seus parâmetros, incluindo um termo positivo associado ao ruído em todos os dias concernentes ao intervalo de interesse.

4.2.3.2 Região 2 (06/08/2021 - 11/08/2021)

A Figura 19 mostra o gráfico atinente às previsões efetuadas para a Região 2.

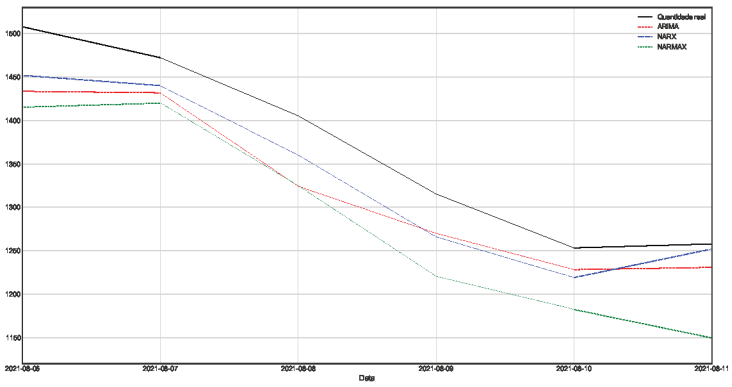


Figura 19 – Previsões para o período de tempo compreendido no intervalo 06/08/2021 - 11/08/2021.

O fenômeno mais interessante desse intervalo refere-se ao repentino distanciamento da curva relativa às previsões NARMAX em relação às outras curvas, sendo que ao final do período, no dia 11/08/2021, há uma previsão pontual bastante errônea desse modelo.

De modo geral, os modelos ARIMA efetuam predições de valores menores que o ideal, fazendo com que a curva atinente a esses modelos tenha patamar ligeiramente inferior à curva de dados reais. Comportamento parecido ocorre nas previsões fundamentadas por modelos NARX.

Os modelos ARIMA variam conforme a descrição presente na Tabela 21.

Tabela 21 – Evolução do modelo ARIMA para a previsão associada à Região 2 (recuperações diárias).

		Data de previsão pontual					
Modelo	**Par.**	**06/08/21**	**07/08/21**	**08/08/21**	**09/08/21**	**10/08/21**	**11/08/21**
ARIMA							
	α_1	0,49	0,49	0,52	0,58	0,55	0,54
	α_2	-0,04	-0,05	-0,08	-0,07	-0,08	-0,07
	α_3	0,21	0,22	0,22	0,26	0,29	0,29
	α_4	0,23	0,23	0,21	0,13	0,12	0,12
	α_5	-0,34	-0,34	-0,32	-0,32	-0,30	-0,28
	β	-24,32	-24,05	-23,47	-27,39	-25,44	-23,83
	Erro Abs.	74,26	40,48	81,14	45,24	24,97	26,83

Já os modelos NARX e NARMAX seguem a seguinte estrutura definida pela Equação (36).

$$\begin{aligned} y(t) = {} & \theta_1 y(t-1) \\ & + \theta_2 y(t-2) \\ & + \theta_3 c(t-1) \\ & + \theta_5 m(t-1) \\ & + \theta_6 l(t-2) \\ & + \theta_7 y(t-1) l(t-1) \\ & + \theta_8 e(t-1). \end{aligned} \tag{36}$$

Esses modelos têm seus parâmetros variados, bem como erros absolutos de predição, mostrados na Tabela 22.

Tabela 22 – Evolução dos modelo NARX e NARMAX para a previsão associada à Região 2 (recuperações diárias).

		Data da previsão pontual					
Modelo	**Par.**	**06/08/21**	**07/08/21**	**08/08/21**	**09/08/21**	**10/08/21**	**11/08/21**
NARX							
	θ_1	1,36	1,38	1,43	1,40	1,38	1,35
	θ_2	-0,43	-0,45	-0,49	-0,47	-0,45	-0,46
	θ_3	0,051	-0,037	-0,024	-0,021	-0,027	0,0038
	θ_4	0,057	0,22	0,18	0,17	0,19	0,21
	θ_5	—-	—-	—-	—-	—-	0,000012
	θ_6	—-	0,0000097	0,0000083	0,0000086	0,0000096	—-
	Erro abs.	55,89	32,29	45,19	49,41	34,16	5,74
NARMAX							
	θ_1	1,64	1,31	1,69	1,65	1,64	1,19
	θ_2	-0,68	-0,39	-0,73	-0,70	-0,69	-0,39
	θ_3	0,037	-0,031	-0,028	-0,034	-0,036	-0,075
	θ_4	-0,0094	0,19	0,093	0,11	0,12	0,35
	θ_5	—-	—-	—-	—-	—-	0,000034
	θ_6	—-	0,0000091	0,0000071	0,0000083	0,0000087	—-
	θ_7	-0,000009	-0,000013	-0,000013	-0,000012	-0,000012	-0,000010
	Erro abs.	98,48	51,94	80,38	94,80	70,92	108,40

Enfatizando os modelos não lineares e tendo em vista sobretudo os últimos três pontos de predição associados à Região 2, nota-se que o modelo NARMAX, além de incorporar termos de ruído munidos de coeficiente negativo, implementa coeficientes mais negativos para θ_2 e θ_3 nas predições realizadas, respectivamente nos dias 09/08/2021 e 10/08/2021, em comparação com a modelagem NARX. Já na última predição efetuada no intervalo, o modelo NARMAX possui um coeficiente θ_1 = 1,19, consideravelmente menor que aquele incorporado pelo modelo NARX (θ_1 = 1,35); bem como agrega valor negativo ao parâmetro θ_3, ao contrário do que é visto para o modelo NARX. Dessa forma, esses aspectos caracterizam o comportamento final da curva atinente às predições efetuadas via modelagem NARMAX.

4.2.4 Considerações Gerais

De acordo com os resultados expostos, sugere-se as seguintes hipóteses:

- Os dados da pandemia são meramente lineares, já que são melhor explicados por um modelo linear (ARIMA) que por modelos não-lineares (NARX e NARMAX). Tal hipótese contraria o perfil apresentado pela pandemia, o qual não corresponde a um comportamento dinâmico regular (linear, em instância).

- Aparentemente, as técnicas de modelagem utilizadas atuam em condições e hipóteses muito distintas para se questionar a qualidade de um modelo em detrimento do outro. Mais estudos da estatística dos resíduos seriam necessários, os quais serviriam como guia à efetuação de possíveis mudanças nos modelos NARX/NARMAX para se elaborar uma comparação mais rica deste com a modelagem ARIMA.
- No âmbito da média móvel, observou-se que a filtragem efetuada na etapa de pré-processamento das séries temporais reduziu drasticamente a ação do ruído sobre os dados - ao ponto de que parte correspondente aos termos de média móvel concernentes aos modelos ARIMA e NARMAX ser desprezível.

Em suma, ao se observar os cenários de predição utilizados no presente trabalho e expostos na nesta seção, pode-se dizer que o modelo ARIMA foi o melhor entre os testados, sendo que a modelagem NARX foi superior à NARMAX em quase todos os cenários. De forma complementar, aparentemente, os termos de média móvel não contribuem para a melhoria dos modelos. Por fim, as técnicas não lineares são menos informativas em comparação com a técnica linear.

Na sequência do texto serão apresentados alguns fundamentos acerca da aptidão da modelagem ARIMA de prever dados epidemiológicos, a fim de melhor de melhor esclarecer a primeira hipótese supracitada.

4.2.4.1 Considerações Sobre a Aptidão Preditiva da Técnica ARIMA com Relação a Dados Epidemiológicos

De fato, quando se trata de modelagem ou previsão de sistemas eminentemente não lineares, a exemplo das séries históricas de uma epidemia, espera-se que técnicas de modelagem não lineares se ajustem melhor aos dados, em comparação com técnicas lineares. No caso deste trabalho, nota-se o oposto, já que as previsões obtidas por meio de modelagem ARIMA resultaram, no geral, em melhor desempenho métrico para todas as séries históricas em estudo. No entanto, a literatura epidemiológica pode nos fornecer algumas informações acerca do sucesso da referida técnica em modelar ou prever dados epidêmicos.

A primeira aplicação de modelagem ARIMA para previsão de dados epidemiológicos remonta ao trabalho de Choi e Thacker, (1981), os quais utilizaram a referida técnica para predizer a quantidade de mortes por pneumonia e influenza em 121 cidades americanas, fundamentando-se por meio de um banco de dados que abrangia

o intervalo do ano de 1962 a 1979. De fato, a modelagem ARIMA foi proposta como contraponto às técnicas preditivas vigentes utilizadas pelos Centros de Controle e Prevenção de Doenças (CDC) dos Estados Unidos - baseadas em análise regressiva.

De fato, a análise comparativa efetuada pelos autores demonstra que o método proposto fornece melhores previsão de mortalidade por pneumonia e influenza em comparação com outros métodos que existiam à época.

Posteriormente ao trabalho supracitado, Helfenstein (1986) utilizou-se da vertente sazonal da modelagem ARIMA para identificar séries temporais associadas à incidência de catapora e caxumba. A partir da aplicação da referida técnicas, obteve-se resultados interessantes, os quais não haviam sido descritos em estudos sobre as mesmas doenças.

Por meio desse trabalho, a exemplo, verificou-se que, apesar de ambas as doenças serem clinicamente distintas, elas apresentam essencialmente a mesma estrutura estatística gerada pela identificação ARIMA. Adicionalmente, para ambas as doenças, depois de modelada a sazonalidade, os resíduos das séries temporais são identificados por um processo AR (1) simples. Assim, além do efeito sazonal, o número de novos casos em um mês é determinado pelo número de casos no mês anterior com a adição de um choque aleatório.

Já no começo dos anos 2000, Reis e Mandl (2003) desenvolveram modelos robustos de utilização de departamentos de emergência dos Estados Unidos, com o objetivo de definir as taxas de visitas esperadas. Para tanto, os autores ajustaram modelos ARIMA para séries históricas fundamentadas em observações de um período de aproximadamente uma década. Como resultados, uma previsão precisa da utilização do departamento de emergência foi alcançada, com um erro médio absoluto percentual de 9,37%, ressaltando-se que a média diária de visitações ao departamento de emergência é de mais de 120 visitas.

Tendo em vista o histórico bem sucedido da aplicação da técnica ARIMA em contextos de previsão epidemiológica, pode-se inferir que há pertinência ao uso dessa técnica ao caso da pandemia de COVID-19. De fato, alguns trabalhos indicam que a modelagem ARIMA pode efetuar previsões superiores aos obtidos por modelos mais complexos, com aqueles baseados em redes neurais - os quais desempenham características não lineares. Como exemplo, em Wang et al., (2021), comparou-se a técnica ARIMA às previsões via amaciamento exponencial e modelagem por rede neura de regressão generalizada (GRNN). Essas técnicas foram utilizadas para prever a incidência diária de casos de COVID-19 na Índia e Estados Unidos, a fim de auxiliar as estratégias

de gestão pública. Em suma, para o cenário indiano, o modelo ARIMA utilizado gerou as melhores previsões, superando a técnica não linear.

Já em Wang et al., (2022), propôs-se a aplicação dos modelos ARIMA, SARIMA e Prophet - técnica não linear de aprendizagem de máquina - para prever novos casos diários e cumulativos casos confirmados nos EUA, Brasil e Índia num horizonte preditivo de 30 dias, com base em dados fornecidos pela OMS. Revelou-se que a modelagem Prophet tem mais vantagens na previsão do COVID-19 dos EUA, haja vista que tal técnica captura características periódicas quando os dados se alteram de forma significativa. Já os modelos ARIMA e SARIMA produziram melhores desempenhos aos casos preditivos no Brasil e na Índia. Ainda, ressalta-se o fato de que o modelo SARIMA capturou um período de sete dias implícito nas séries temporais de casos diários de COVID-19 para os três países. Na previsão de novos casos cumulativos, o modelo ARIMA tem melhor capacidade de ajustar e prever os dados com tendência de crescimento positivo em diferentes países no Brasil e na Índia.

Outro exemplo de estudo comparativo entre a modelagem ARIMA e outra técnica não linear é apresentado por Rahman et al., (2022), os quais apresentaram a referida modelagem como contraponto à previsão via *Extreme Gradient Boosting* (XGBoost), o qual é um algoritmo de aprendizagem de máquina fundamentado em árvores de decisão. Para tanto, os autores utilizaram tais modelos para previsão da incidência de casos e mortes de COVID-19, a uma taxa diária, em Bangladesh. A partir da avaliação da precisão preditiva dos modelos por meio de métricas de desempenho como o RMSE, o MAE e o MAPE, verificou-se vantagem preditiva dos modelos ARIMA.

Tendo em vista as informações supramencionadas, pode-se deduzir que a modelagem ARIMA desempenhou papel importante no âmbito da literatura de previsão de dados epidemiológicos, com aplicações bem sucedidas desde o início dos anos 1980. Dessa forma, não se faz surpreendente a larga utilização dessa técnica no atual cenário pandêmico. Tendo isso em conta, apesar de se tratar de uma técnica linear, foi visto que, por vezes, modelagem ARIMA apresentou previsões mais precisas quando comparada a técnicas sabidamente não lineares.

5 CONCLUSÃO

O presente estudo teve como objetivo a obtenção de modelos de identificação de séries temporais de dados associados à pandemia de COVID-19, no estado brasileiro de Santa Catarina. Para tanto, foram utilizados três métodos de previsão para séries temporais: o modelo ARIMA e os modelos NARX e NARMAX polinomiais. A pesquisa se fundamentou nas quantidades de casos, óbitos, pacientes recuperados e leitos SUS ocupados por pacientes com COVID-19 e a base de dados foi concebida a partir da captação de dados provenientes dos boletins epidemiológicos lançados pela Secretaria de Saúde do Estado de Santa Catarina.

Em primeira instância, foi realizada uma análise exploratória da base de dados histórica disponível. Por meio desse estudo prévio e da aplicação de pré-processamento dos dados a partir de métodos consagrados em literatura atinente à epidemiologia preditiva, o qual consistiu em:

- Identificação de *outliers* por meio do filtro de Hampel.
- Imputação de valores ausentes por meio de interpolação linear.
- Suavização de ruído por meio de aplicação do filtro de Savitsky-Golay.

Posteriormente à etapa de pré-processamento, conduziu-se os experimentos preditivos propriamente ditos. Os modelos de predição foram concebidos considerando-se janelas deslizantes de dados munidas de avanço não ancorado, avaliados para a previsão do dia seguinte a partir dos dados dos últimos 150 dias. Optou-se pelo paradigma de janelas deslizantes não ancoradas devido ao fato de que essa abordagem faz-se adequada à aplicações temporais, em que a ordem cronológica dos dados deve ser respeitada. Desse modo, uma abordagem estática não se faria pertinente.

De acordo com os resultados obtidos, sugere-se as seguintes hipóteses:

- Os dados da pandemia são meramente lineares, já que são melhor explicados por um modelo linear (ARIMA) que por modelos não-lineares (NARX e NARMAX). Tal hipótese contraria o perfil apresentado pela pandemia, o qual não corresponde a um comportamento dinâmico regular (linear, em instância).
- Aparentemente, as técnicas de modelagem utilizadas atuam em condições e hipóteses muito distintas para se questionar a qualidade de um modelo em detrimento do outro. Mais estudos da estatística dos resíduos seriam necessários,

os quais serviriam como guia à efetuação de possíveis mudanças nos modelos NARX/NARMAX para se elaborar uma comparação mais rica deste com a modelagem ARIMA.

- No âmbito da média móvel, observou-se que a filtragem efetuada na etapa de pré-processamento das séries temporais reduziu drasticamente a ação do ruído sobre os dados - ao ponto de que parte correspondente aos termos de média móvel concernentes aos modelos ARIMA e NARMAX ser desprezível.

Em síntese, ao se analisar os cenários de predição utilizados no presente trabalho e expostos na nesta seção, pode-se dizer que o modelo ARIMA foi o melhor entre os testados, sendo que a modelagem NARX foi superior à NARMAX em quase todos os cenários. De forma complementar, aparentemente, os termos de média móvel não contribuem para a melhoria dos modelos. Por fim, as técnicas não lineares são menos informativas em comparação com a técnica linear.

Em aspecto social, o presente estudo demonstra a viabilidade de análise preditiva em situações de pandemia. Em cenários futuros, técnicas de predição como as utilizadas neste trabalho, podem ser empregadas em sistemas de gestão de saúde, a exemplo do que é feito em Hasan et al. (2022), o qual desenvolve um sistema de gerenciamento de dados concernente à pandemia de COVID-19 em Nova Delhi, incluindo-se um algoritmo preditivo baseado em ARIMA, para fins de gestão de saúde.

De fato, as vantagens dos trabalhos preditivos concernentes à pandemia de COVID-19 são bem descritas por Zhao et al., (2021), cujo trabalho refere-se à previsão de curto prazo de casos de infecção por COVID-19, por meio da aplicação do modelo SEIR. Segundo os autores, a previsão da dinâmica de novas infecções pelo novo coronavírus é fundamental para o planejamento de saúde pública, no sentido de facilitar a alocação eficiente de insumos hospitalares, bem como o monitoramento dos efeitos das intervenções políticas.

Adicionalmente, em conformidade com a OMS ([OMS], 2020b), trabalhos que visam a previsão de dados epidemiológicos são cruciais, haja vista que a análise preditiva nos permite estimar o comportamento da pandemia dentro de um grau de incerteza, estabelecendo quando e em que condições os países podem esperar aumentos de casos, picos e reduções de novos casos e mortalidade. Por meio de tais informações, pode-se calcular a demanda por serviços médicos agudos, determinar prazos para levantamento parcial ou total das medidas de contenção, e até mesmo prever necessidades para ondas subsequentes da pandemia. Além disso, as previsões constituem

ferramentas auxiliares à estimativa da demanda por serviços e materiais hospitalares, a fim de permitir o planejamento da aplicação de tecnologias necessárias (como o uso de ventiladores), do controle da cadeia de suprimentos, e da gestão de recursos humanos para um resposta adequada e oportuna no combate à pandemia.

Para trabalhos futuros, tem-se a intenção de aprofundar-se acerca do uso das técnicas preditivas aplicadas para o âmbito epidêmico. Um estudo futuro deve implementar ferramentas estatísticas mais robustas do ponto de vista da análise dos resíduos, de forma a possibilizar a concepção de modelos mais informativos acerca da dinâmica dos dados pandêmicos. Além disso, faz-se pertinente a aplicação de cenários preditivos mais abrangentes quanto ao espaço amostral definido pelo conjunto de parâmetros relativos às técnicas de identificação utilizadas.

Outro tópico a ser abordado em estudos posteriores faz alusão à análise de tendência dos dados epidemiológicos, a qual se também se trata de um segmento de pesquisa bastante explorado no âmbito da pandemia de COVID-19. A análise de tendências, diferentemente da efetuação de previsões pontuais, refere-se à previsão do movimento futuro das séries históricas.

Como exemplo dessa abordagem, em Paiva et al., (2021), desenvolveu-se um sistema computacional para a análise de tendência dos dados pandêmicos a partir de formulações matemáticas, munidas de calibração automática de parâmetros. No referido trabalho, utiliza-se uma função sigmoidal assimétrica para descrição dos dados epidêmicos, considerando que o ajuste automático das variáveis dessa função é efetuado a partir de um algoritmo de otimização numérico, para melhor prever a tendência de curto prazo dos dados.

Em suma, a análise de tendência, em combinação com a abordagem preditiva a partir de previsões pontuais, agregaria utilidade ao presente estudo, haja vista que fornece informações diferentes a um processo decisório em termos de gerenciamento de saúde. Essa distinção decorre do fato de que, diferentemente da previsão pontual, a análise de tendência dos dados provê maior noção acerca do regime de crescimento (ou decrescimento) das séries históricas epidemiológicas.

Por fim, cabe afirmar que a pesquisa descrita nesta dissertação apresenta considerações e resultados úteis acerca do uso de técnicas de análise de séries temporais no estudo preditivo da dinâmica dos dados associados à pandemia de COVID-19 no cenário catarinense. Adicionalmente, acredita-se que o presente trabalho possa colaborar na prevenção de mazelas de ordem clínica e social, oriundas da pandemia de COVID-19 no estado de Santa Catarina. De fato, foram produzidas predições com bom

grau de exatidão e entende-se que os modelos aqui utilizados podem muito bem ser futuramente aplicados ou incorporados em *softwares* de gestão de saúde, a fim de auxiliar atividades de análise epidemiológica.

REFERÊNCIAS

[OMS]. **WHO Director-General's opening remarks at the media briefing on COVID-19 - 11 March 2020**. 2020a. Disponível em: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020. Acesso em: 20 fev. 2021.

[OMS]. **Why Predictive Modeling is Critical in the Fight against COVID-19**. [*S.l.*: *s.n.*], 2020b. https://iris.paho.org/bitstream/handle/10665.2/52276/PAHOEIHISCOVID-19200007_eng.pdf?sequence=8. Accessed: 2022-02-09.

ABUSAM, Abdallah. Dynamics of COVID-19 in the Gulf Cooperation Council (GCC) countries. **Journal of Taibah University Medical Sciences**, Elsevier, 2022.

AGGARWAL, Charu C. An introduction to outlier analysis. *In*: OUTLIER analysis. [*S.l.*]: Springer, 2017. P. 1–34.

AGUIRRE, Luis Antonio. **Introdução à identificação de sistemas–Técnicas lineares e não-lineares aplicadas a sistemas reais**. [*S.l.*]: Editora UFMG, 2004.

AKAIKE, Hirotugu. A new look at the statistical model identification. **IEEE transactions on automatic control**, Ieee, v. 19, n. 6, p. 716–723, 1974.

ALLEN, DP. A frequency domain Hampel filter for blind rejection of sinusoidal interference from electromyograms (Translation Journals style). **Journal Neuroscience Methods**, v. 155, n. 177, p. 2.

ANDERSON, Roy M; MAY, Robert M. **Infectious diseases of humans: dynamics and control**. [*S.l.*]: Oxford university press, 1992.

ANDERSON, Roy M; MAY, Robert M. Population biology of infectious diseases: Part I. **Nature**, Nature Publishing Group, v. 280, n. 5721, p. 361–367, 1979.

ARABZADEH, Rezgar; GRÜNBACHER, Daniel Martin; INSAM, Heribert; KREUZINGER, Norbert; MARKT, Rudolf; RAUCH, Wolfgang. Data filtering methods for SARS-CoV-2 wastewater surveillance. **Water Science and Technology**, IWA Publishing, v. 84, n. 6, p. 1324–1339, 2021.

ARDABILI, Sina F; MOSAVI, Amir; GHAMISI, Pedram; FERDINAND, Filip; VARKONYI-KOCZY, Annamaria R; REUTER, Uwe; RABCZUK, Timon; ATKINSON, Peter M. Covid-19 outbreak prediction with machine learning. **Algorithms**, Multidisciplinary Digital Publishing Institute, v. 13, n. 10, p. 249, 2020.

ARORA, Parul; KUMAR, Himanshu; PANIGRAHI, Bijaya Ketan. Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. **Chaos, Solitons & Fractals**, Elsevier, v. 139, p. 110017, 2020.

BATISTA JÚNIOR, Aguinaldo Bezerra; PIRES, Paulo Sérgio da Motta. An approach to outlier detection and smoothing applied to a trajectography radar data. **Journal of Aerospace Technology and Management**, SciELO Brasil, v. 6, p. 237–248, 2014.

BELISÁRIO, Ana Brandão *et al.* Análise de emissões em caldeiras de recuperação quımica de fábricas de celulose Kraft: predição e análise de sensibilidade com redes neurais artificiais. Universidade Federal de Minas Gerais, 2020.

BILLINGS; CHEN, and Korenberg. Identification of MIMO non-linear systems using a forward-regression orthogonal estimator. **International journal of control**, Taylor & Francis, v. 49, n. 6, p. 2157–2189, 1989.

BILLINGS, Lora; SCHWARTZ, Ira B. Exciting chaos with noise: unexpected dynamics in epidemic outbreaks. **Journal of Mathematical Biology**, v. 44, p. 31–48, 2002.

BILLINGS; WEI. NARMAX model as a sparse, interpretable and transparent machine learning approach for big medical and healthcare data analysis. *In*: IEEE. 2019 IEEE 21st International Conference on High Performance Computing and Communications; IEEE 17th International Conference on Smart City; IEEE 5th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). [*S.l.*: *s.n.*], 2019. P. 2743–2750.

BILLINGS; WEI, Hua-Liang; THOMAS, Patrick; LINNANE, Seamus J; HOPE-GILL, Ben DM. The prediction of in-flight hypoxaemia using non-linear equations. **Respiratory medicine**, Elsevier, v. 107, n. 6, p. 841–847, 2013.

BISTRIAN, DA; DIMITRIU, G; NAVON, IM. Processing epidemiological data using dynamic mode decomposition method. *In*: AIP PUBLISHING LLC, 1. AIP Conference Proceedings. [*S.l.*: *s.n.*], 2019. P. 080002.

BLÁZQUEZ-GARCIA, Ane; CONDE, Angel; MORI, Usue; LOZANO, Jose A. A review on outlier/anomaly detection in time series data. **ACM Computing Surveys (CSUR)**, ACM New York, NY, USA, v. 54, n. 3, p. 1–33, 2021.

BONTEMPI, G; TAIEB, SB. Yann-Aël Le Borgne. **Machine learning strategies for time series forecasting**, European Business Intelligence Summer School. ULB—Universite Libre de Bruxelles, v. 8, 2013.

BRAUER, Fred; CASTILLO-CHAVEZ, Carlos; CASTILLO-CHAVEZ, Carlos. **Mathematical models in population biology and epidemiology**. [*S.l.*]: Springer, 2012. v. 2.

BROCKWELL, Peter J; DAVIS, Richard A. **Introduction to time series and forecasting**. [*S.l.*]: Springer, 2002.

BROCKWELL, Peter J.; DAVIS, Richard A. **Introduction to Time Series and Forecasting**. 3. ed. [*S.l.*]: Springer, 2016.

CAVANAUGH, Joseph E. Unifying the derivations for the Akaike and corrected Akaike information criteria. **Statistics & Probability Letters**, Elsevier, v. 33, n. 2, p. 201–208, 1997.

CHATFIELD, Chris. **Problem solving: a statistician's guide**. [*S.l.*]: CRC Press, 1995.

CHEN, Chung; LIU, Lon-Mu. Joint estimation of model parameters and outlier effects in time series. **Journal of the American Statistical Association**, Taylor & Francis, v. 88, n. 421, p. 284–297, 1993.

CHEN, Yu Chi; SUN, Wei Chih; JUANG, Jyh Ching. Outlier detection technique for RSS-based localization problems in wireless sensor networks. *In*: IEEE. PROCEEDINGS of SICE Annual Conference 2010. [*S.l.*: *s.n.*], 2010. P. 657–662.

CLANCY, Damian. Optimal intervention for epidemic models with general infection and removal rate functions. **Journal of mathematical biology**, Springer, v. 39, n. 4, p. 309–331, 1999.

COKLUK, Omay; KAYRI, Murat. The Effects of Methods of Imputation for Missing Values on the Validity and Reliability of Scales. **Educational Sciences: Theory and Practice**, ERIC, v. 11, n. 1, p. 303–309, 2011.

CUI, Hao; KERTÉSZ, János. Attention dynamics on the Chinese social media Sina Weibo during the COVID-19 pandemic. **EPJ data science**, Springer Berlin Heidelberg, v. 10, n. 1, p. 8, 2021.

DAVIES, Laurie; GATHER, Ursula. The identification of multiple outliers. **Journal of the American Statistical Association**, Taylor & Francis, v. 88, n. 423, p. 782–792, 1993.

DAVIS, Neema; RAINA, Gaurav; JAGANNATHAN, Krishna. A framework for end-to-end deep learning-based anomaly detection in transportation networks. **Transportation research interdisciplinary perspectives**, Elsevier, v. 5, p. 100112, 2020.

DEBBOUCHE, Nadjette; OUANNAS, Adel; BATIHA, Iqbal M; GRASSI, Giuseppe. Chaotic dynamics in a novel COVID-19 pandemic model described by commensurate and incommensurate fractional-order derivatives. **Nonlinear Dynamics**, Springer, p. 1–13, 2021.

DESCHEPPER, Mieke; EECKLOO, Kristof; MALFAIT, Simon; BENOIT, Dominique; CALLENS, Steven; VANSTEELANDT, Stijn. Prediction of hospital bed capacity during the COVID- 19 pandemic. **BMC health services research**, BioMed Central, v. 21, n. 1, p. 1–10, 2021.

DICKEY, David A; FULLER, Wayne A. Likelihood ratio statistics for autoregressive time series with a unit root. **Econometrica: journal of the Econometric Society**, JSTOR, p. 1057–1072, 1981.

DIREKOGLU, Cem; SAH, Melike. Worldwide and regional forecasting of coronavirus (covid-19) spread using a deep learning model. **medRxiv**, Cold Spring Harbor Laboratory Press, 2020.

DONDERS, A Rogier T; VAN DER HEIJDEN, Geert JMG; STIJNEN, Theo; MOONS, Karel GM. A gentle introduction to imputation of missing values. **Journal of clinical epidemiology**, Elsevier, v. 59, n. 10, p. 1087–1091, 2006.

DOWNING, Douglas; CLARK, Jeffrey. Applied Statistics. **Saraiva, Sao Paulo, Brazil**, v. 2, 2006.

DURBIN, James. The fitting of time-series models. **Revue de l'Institut International de Statistique**, JSTOR, p. 233–244, 1960.

EILERSEN, Andreas; JENSEN, Mogens H; SNEPPEN, Kim. Chaos in disease outbreaks among prey. **Scientific reports**, Nature Publishing Group, v. 10, n. 1, p. 1–7, 2020.

FANG, Lanlan; WANG, Dingjian; PAN, Guixia. Analysis and estimation of COVID-19 spreading in Russia based on ARIMA model. **SN comprehensive clinical medicine**, Springer, v. 2, n. 12, p. 2521–2527, 2020.

FONTAL, Alejandro; BOUMA, Menno J; SAN-JOSÉ, Adrià; LÓPEZ, Leonardo; PASCUAL, Mercedes; RODÓ, Xavier. Climatic signatures in the different COVID-19 pandemic waves across both hemispheres. **Nature Computational Science**, Nature Publishing Group, v. 1, n. 10, p. 655–665, 2021.

FORATTINI, Oswaldo Paulo. **Epidemiologia Geral**. 2. ed. São Paulo: Springer, 1996a. v. 2.

FORATTINI, Oswaldo Paulo. Epidemiologia geral, 1996b.

FOX, Anthony J. Outliers in time series. **Journal of the Royal Statistical Society: Series B (Methodological)**, Wiley Online Library, v. 34, n. 3, p. 350–363, 1972.

GAUVREAU, K; PAGANO, M. **WHY 5-PERCENT**. v. 10. [*S.l.*]: ELSEVIER SCIENCE INC 655 AVENUE OF THE AMERICAS, NEW YORK, NY 10010, 1994. P. 93–94.

GOIC FIGUEROA, Marcel; BOZANIC LEAL, Mirko Slovan; BADAL, Magdalena; BASSO SOTZ, Leonardo. COVID-19: short-term forecast of ICU beds in times of crisis. Public Library Science, 2021.

GORBALENYA, Alexander E *et al.* Severe acute respiratory syndrome-related coronavirus: The species and its viruses–a statement of the Coronavirus Study Group. BioRxiv, 2020.

GRIMM, Volker. Ten years of individual-based modelling in ecology: what have we learned and what could we learn in the future? **Ecological modelling**, Elsevier, v. 115, n. 2-3, p. 129–148, 1999.

GUPTA, Manish; GAO, Jing; AGGARWAL, Charu; HAN, Jiawei. Outlier detection for temporal data. **Synthesis Lectures on Data Mining and Knowledge Discovery**, Morgan & Claypool Publishers, v. 5, n. 1, p. 1–129, 2014.

HAIDER, Nishi Shahnaj; PERIYASAMY, R; JOSHI, Deepak; SINGH, BK. Savitzky-Golay filter for denoising lung sound. **Brazilian Archives of Biology and Technology**, SciELO Brasil, v. 61, 2018.

HARGITTAI, S. Savitzky-Golay least-squares polynomial filters in ECG signal processing. *In*: IEEE. COMPUTERS in Cardiology, 2005. [*S.l.*: *s.n.*], 2005. P. 763–766.

HASAN, Iqbal; DHAWAN, Prince; RIZVI, SAM; DHIR, Sanjay. Data analytics and knowledge management approach for COVID-19 prediction and control. **International Journal of Information Technology**, Springer, p. 1–18, 2022.

HAWKINS, Douglas M. Multivariate outlier detection. *In*: IDENTIFICATION of outliers. [*S.l.*]: Springer, 1980. P. 104–114.

HELFENSTEIN, Ulrich. Box-Jenkins modelling of some viral infectious diseases. **Statistics in medicine**, Wiley Online Library, v. 5, n. 1, p. 37–47, 1986.

HETHCOTE, Herbert W. The mathematics of infectious diseases. **SIAM review**, SIAM, v. 42, n. 4, p. 599–653, 2000.

HEWAMALAGE, Hansika; BERGMEIR, Christoph; BANDARA, Kasun. Recurrent neural networks for time series forecasting: Current status and future directions. **International Journal of Forecasting**, Elsevier, v. 37, n. 1, p. 388–427, 2021.

HILLMER, Steven. Monitoring and adjusting forecasts in the presence of additive outliers. **Journal of Forecasting**, Wiley Online Library, v. 3, n. 2, p. 205–215, 1984.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. **Neural computation**, MIT Press, v. 9, n. 8, p. 1735–1780, 1997.

HOOT, Nathan R; ZHOU, Chuan; JONES, Ian; ARONSKY, Dominik. Measuring and forecasting emergency department crowding in real time. **Annals of emergency medicine**, Elsevier, v. 49, n. 6, p. 747–755, 2007.

HOSSAIN, Zakir; RAHMAN, Atikur; HOSSAIN, Moyazzem; KARAMI, Jamil Hasan. Over-differencing and forecasting with non-stationary time series data. **Dhaka University Journal of Science**, v. 67, n. 1, p. 21–26, 2019.

INOUE, Jun; YAMAGATA, Yoriyuki; CHEN, Yuqi; POSKITT, Christopher M; SUN, Jun. Anomaly detection for a water treatment system using unsupervised machine learning. *In*: IEEE. 2017 IEEE international conference on data mining workshops (ICDMW). [*S.l.*: *s.n.*], 2017. P. 1058–1065.

IVORRA, Benjamin; MARTINEZ-LÓPEZ, Beatriz; SÁNCHEZ-VIZCAINO, José M; RAMOS, Ángel M. Mathematical formulation and validation of the Be-FAST model for classical swine fever virus spread between and within farms. **Annals of operations research**, Springer, v. 219, n. 1, p. 25–47, 2014.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. **An introduction to statistical learning**. [*S.l.*]: Springer, 2013. v. 112.

JIN, Qixuan. Time Warping clustering for the forecast and analysis of COVID-19. *In*: IEEE. 2020 IEEE MIT Undergraduate Research Technology Conference (URTC). [*S.l.*: *s.n.*], 2020. P. 1–5.

KAPOOR, Amol; BEN, Xue; LIU, Luyang; PEROZZI, Bryan; BARNES, Matt; BLAIS, Martin; O'BANION, Shawn. Examining covid-19 forecasting using spatio-temporal graph neural networks. **arXiv preprint arXiv:2007.03113**, 2020.

KEELING, Matt J; GRENFELL, Bryan T. Individual-based perspectives on R0. **Journal of theoretical biology**, Elsevier, v. 203, n. 1, p. 51–61, 2000.

KEELING, Matt J; ROHANI, Pejman. Estimating spatial coupling in epidemiological systems: a mechanistic approach. **Ecology Letters**, Wiley Online Library, v. 5, n. 1, p. 20–29, 2002.

KIRKPATRICK II, Charles D; DAHLQUIST, Julie A. **Technical analysis: the complete resource for financial market technicians**. [*S.l.*]: FT press, 2010.

KRISHNAN, Sunder Ram; SEELAMANTULA, Chandra Sekhar. On the Selection of Optimum Savitzky-Golay Filters. **IEEE Transactions on Signal Processing**, v. 61, p. 380–391, 2013.

LAURAITIS, Andrius; MASKELIŪNAS, Rytis. Investigation of predicting functional capacity level for Huntington disease patients. *In*: SPRINGER. INTERNATIONAL Conference on Information and Software Technologies. [*S.l.*: *s.n.*], 2017. P. 142–149.

LEDOLTER, Johannes. The effect of additive outliers on the forecasts from ARIMA models. **International Journal of Forecasting**, Elsevier, v. 5, n. 2, p. 231–240, 1989.

LIU, Hancong; SHAH, Sirish; JIANG, Wei. On-line outlier detection and data cleaning. **Computers & chemical engineering**, Elsevier, v. 28, n. 9, p. 1635–1647, 2004.

LLOYD, Alun L. Realistic distributions of infectious periods in epidemic models: changing patterns of persistence and dynamics. **Theoretical population biology**, Elsevier, v. 60, n. 1, p. 59–71, 2001.

LOPES-JÚNIOR, LC. A Saúde Coletiva no epicentro da pandemia de COVID-19 no Sistema Único de Saúde. **Saúde Colet.(Barueri)**, v. 10, n. 56, p. 3080–9, 2020.

MEHROTRA, Kishan G; MOHAN, Chilukuri K; HUANG, HuaMing. **Anomaly detection principles and algorithms**. [*S.l.*]: Springer, 2017. v. 1.

MELIN, Patricia; MONICA, Julio Cesar; SANCHEZ, Daniela; CASTILLO, Oscar. Multiple ensemble neural network models with fuzzy response aggregation for predicting COVID-19 time series: the case of Mexico. *In*: MULTIDISCIPLINARY DIGITAL PUBLISHING INSTITUTE, 2. HEALTHCARE. [*S.l.*: *s.n.*], 2020. P. 181.

MOFTAKHAR, Leila; SEIF, Mozhgan. The exponentially increasing rate of patients infected with COVID-19 in Iran. **Archives of Iranian medicine**, Academy of Medical Sciences of the IR Iran, v. 23, n. 4, p. 235–238, 2020.

MONTGOMERY, Douglas C; JENNINGS, Cheryl L; KULAHCI, M. Introduction to Time Series Analysis and Forecasting, A John Wiley & Sons. **Inc., New Jersey**, 2008.

MORETTIN, Pedro Alberto; TOLOI, Clélia Maria de Castro. Análise de séries temporais, 2004.

NEPOMUCENO, EG; AGUIRRE, LA; TAKAHASHI, RHC; LAMPERTI, RD; ALVARENGA, LR; KURCBART, SM. Modelagem de sistemas epidemiológicos por meio de modelos baseados em indivıduos. *In*: ANAIS do XVI Congresso Brasileiro de Automática. [*S.l.*: *s.n.*], 2006. P. 2399–2404.

NISHIDA, Erica N; DUTRA, Odilon O; FERREIRA, Luıs HC; COLLETTA, Gustavo D. Application of Savitzky-Golay digital differentiator for QRS complex detection in an electrocardiographic monitoring system. *In*: IEEE. 2017 IEEE International Symposium on Medical Measurements and Applications (MeMeA). [*S.l.*: *s.n.*], 2017. P. 233–238.

NORONHA, Kenya Valeria Micaela de Souza *et al.* Pandemia por COVID-19 no Brasil: análise da demanda e da oferta de leitos hospitalares e equipamentos de ventilação assistida segundo diferentes cenários. **Cadernos de Saúde Pública**, SciELO Brasil, v. 36, 2020.

PAIVA, Henrique Mohallem; AFONSO, Rubens Junqueira Magalhães; LIMA ALVARENGA, Fabiana Mara Scarpelli de; ANDRADE VELASQUEZ, Ester de *et al.* A computational tool for trend analysis and forecast of the COVID-19 pandemic. **Applied Soft Computing**, Elsevier, v. 105, p. 107289, 2021.

PANKRATZ, Alan. **Forecasting with univariate Box-Jenkins models: Concepts and cases**. [*S.l.*]: John Wiley & Sons, 2009.

PAROLINI, Nicola; ARDENGHI, Giovanni; DEDE', Luca; QUARTERONI, Alfio. A mathematical dashboard for the analysis of Italian COVID-19 epidemic data. **International Journal for Numerical Methods in Biomedical Engineering**, Wiley Online Library, v. 37, n. 9, e3513, 2021.

PEARSON, Ronald K. **Mining imperfect data: Dealing with contamination and incomplete records**. [*S.l.*]: SIAM, 2005.

PEDRO, Sansao A; ABELMAN, Shirley; NDJOMATCHOUA, Frank T; SANG, Rosemary; TONNANG, Henri EZ. Stability, bifurcation and chaos analysis of vector-borne disease model with application to rift valley fever. **PloS one**, Public Library of Science San Francisco, USA, v. 9, n. 10, e108172, 2014.

PERNET, Cyril. Improving standards in brain-behavior correlation analyses. Citeseer, 2012.

RACHE, Beatriz; NUNES, Letıcia; ROCHA, Rudi; LAGO, Miguel; FRAGA, Arminio. Como Conter a Curva no Brasil? Onde a Epidemiologia e a Economia se Encontram. **Nota Técnica**, v. 4, 2020.

RAHIM, H Abdul; IBRAHIM, F; TAIB, MN. A novel prediction system in dengue fever using NARMAX model. *In*: IEEE. 2007 International Conference on Control, Automation and Systems. [*S.l.*: *s.n.*], 2007. P. 305–309.

RAHMAN, Md Siddikur; CHOWDHURY, Arman Hossain; AMRIN, Miftahuzzannat. Accuracy comparison of ARIMA and XGBoost forecasting models in predicting the incidence of COVID-19 in Bangladesh. **PLOS Global Public Health**, Public Library of Science San Francisco, CA USA, v. 2, n. 5, e0000495, 2022.

RANTOU, K. Missing data in time series and imputation methods. **University of the Aegean, Samos**, 2017.

RASJID, Zulfany Erlisa; SETIAWAN, Reina; EFFENDI, Andy. A comparison: prediction of death and infected COVID-19 cases in Indonesia using time series smoothing and LSTM neural network. **Procedia computer science**, Elsevier, v. 179, p. 982–988, 2021.

REIS, Ben Y; MANDL, Kenneth D. Time series modeling for syndromic surveillance. **BMC medical informatics and decision making**, BioMed Central, v. 3, n. 1, p. 1–11, 2003.

RIBEIRO, Matheus Henrique Dal Molin; SILVA, Ramon Gomes da; MARIANI, Viviana Cocco; SANTOS COELHO, Leandro dos. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for Brazil. **Chaos, Solitons & Fractals**, Elsevier, v. 135, p. 109853, 2020.

ROUSSEEUW, Peter J; CROUX, Christophe. Alternatives to the median absolute deviation. **Journal of the American Statistical association**, Taylor & Francis, v. 88, n. 424, p. 1273–1283, 1993.

RUSTAM, Furqan; RESHI, Aijaz Ahmad; MEHMOOD, Arif; ULLAH, Saleem; ON, Byung-Won; ASLAM, Waqar; CHOI, Gyu Sang. COVID-19 future forecasting using supervised machine learning models. **IEEE access**, IEEE, v. 8, p. 101489–101499, 2020.

SADEGHI, Mohammad; BEHNIA, Fereidoon; AMIRI, Rouhollah. Window selection of the Savitzky–Golay filters for signal recovery from noisy measurements. **IEEE Transactions on Instrumentation and Measurement**, IEEE, v. 69, n. 8, p. 5418–5427, 2020.

SADIKOGLU, Fahreddin; KAVALCIOĞLU, Cemal. Filtering continuous glucose monitoring signal using Savitzky-Golay filter and simple multivariate thresholding. **Procedia Computer Science**, Elsevier, v. 102, p. 342–350, 2016.

SAPKOTA, Nabin; KARWOWSKI, Waldemar; DAVAHLI, Mohammad Reza; AL-JUAID, Awad; TAIAR, Redha; MURATA, Atsuo; WRÓBEL, Grzegorz; MAREK, Tadeusz. The chaotic behavior of the spread of infection during the COVID-19 pandemic in the United States and globally. **Ieee Access**, IEEE, v. 9, p. 80692–80702, 2021.

SAVITZKY, Abraham; GOLAY, Marcel JE. Smoothing and differentiation of data by simplified least squares procedures. **Analytical chemistry**, ACS Publications, v. 36, n. 8, p. 1627–1639, 1964.

SCHWEIGLER, Lisa M; DESMOND, Jeffrey S; MCCARTHY, Melissa L; BUKOWSKI, Kyle J; IONIDES, Edward L; YOUNGER, John G. Forecasting models of emergency department crowding. **Academic Emergency Medicine**, Wiley Online Library, v. 16, n. 4, p. 301–308, 2009.

SHAHID, Farah; ZAMEER, Aneela; MUNEEB, Muhammad. Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. **Chaos, Solitons & Fractals**, Elsevier, v. 140, p. 110212, 2020.

SHAMSUDDIN, Noraishah; TAIB, Mohd. Nasir. Nonlinear ARX modeling of heart diseases based on heart sounds. *In*: 2011 IEEE 7th International Colloquium on Signal Processing and its Applications. [*S.l.*: *s.n.*], 2011. P. 382–387. DOI: 10.1109/CSPA.2011.5759907.

SILVA, Elio Medeiros da; SILVA, Ermes Medeiros da. **Matemática e estatıstica aplicada**. [*S.l.*]: Atlas, 1999.

SILVA, Ramon Gomes da; RIBEIRO, Matheus Henrique Dal Molin; MARIANI, Viviana Cocco; SANTOS COELHO, Leandro dos. Forecasting Brazilian and American COVID-19 cases based on artificial intelligence coupled with climatic exogenous variables. **Chaos, Solitons & Fractals**, Elsevier, v. 139, p. 110027, 2020.

SINGH, Divya; SINGH, Bikesh Kumar; BEHERA, Ajoy Kumar. Comparative analysis of Lung sound denoising technique. *In*: IEEE. 2020 First International Conference on Power, Control and Computing Technologies (ICPC2T). [*S.l.*: *s.n.*], 2020. P. 406–410.

SINGH, Sarbhan; SUNDRAM, Bala Murali; RAJENDRAN, Kamesh; LAW, Kian Boon; ARIS, Tahir; IBRAHIM, Hishamshah; DASS, Sarat Chandra; GILL, Balvinder Singh. Forecasting daily confirmed COVID-19 cases in Malaysia using ARIMA models. **The Journal of Infection in Developing Countries**, v. 14, n. 09, p. 971–976, 2020.

SPEARMAN, Charles. The proof and measurement of association between two things. Appleton-Century-Crofts, 1961.

SWARAJ, Aman; VERMA, Karan; KAUR, Arshpreet; SINGH, Ghanshyam; KUMAR, Ashok; SALES, Leandro Melo de. Implementation of stacking based ARIMA model for prediction of Covid-19 cases in India. **Journal of Biomedical Informatics**, Elsevier, v. 121, p. 103887, 2021.

THIEME, Horst R. Princeton series in theoretical and computational biology. *In*: MATHEMATICS in Population Biology. [*S.l.*]: Princeton University Press, 2003.

TUKEY, John W *et al.* **Exploratory data analysis**. [*S.l.*]: Reading, MA, 1977. v. 2.

TUKEY, JW. Nonlinear (nonsuperposable) methods for smoothing data. **Proc. Cong. Rec. EASCOM'74**, p. 673–681, 1974.

WANG, Gang; WU, Tiantian; WEI, Wudi; JIANG, Junjun; AN, Sanqi; LIANG, Bingyu; YE, Li; LIANG, Hao. Comparison of ARIMA, ES, GRNN and ARIMA–GRNN hybrid models to forecast the second wave of COVID-19 in India and the United States. **Epidemiology & Infection**, Cambridge University Press, v. 149, 2021.

WEI, Hua-Liang; BILLINGS, Stephen A. Modelling COVID-19 Pandemic Dynamics Using Transparent, Interpretable, Parsimonious and Simulatable (TIPS) Machine Learning Models: A Case Study from Systems Thinking and

System Identification Perspectives. **medRxiv**, Cold Spring Harbor Laboratory Press, 2021.

WEIGEND, Andreas S. **Time series prediction: forecasting the future and understanding the past**. [*S.l.*]: Routledge, 2018.

YAN, Dongxue; CAO, Hui. The global dynamics for an age-structured tuberculosis transmission model with the exponential progression rate. **Applied Mathematical Modelling**, Elsevier, v. 75, p. 769–786, 2019.

YANG, Hyun Mo. Epidemiologia matemática: estudos dos efeitos da vacinação em doenças de trasmissão direta. *In*: EPIDEMIOLOGIA matemática: estudos dos efeitos da vacinação em doenças de trasmissão direta. [*S.l.*: *s.n.*], 2001. P. 239–239.

ZHAN, Choujun; WU, Zhengdong; WEN, Quansi; GAO, Ying; ZHANG, Haijun. Optimizing broad learning system hyper-parameters through particle swarm optimization for predicting COVID-19 in 184 Countries. *In*: IEEE. 2020 IEEE International Conference on E-health Networking, Application & Services (HEALTHCOM). [*S.l.*: *s.n.*], 2021. P. 1–6.

ZHAO, Hongwei; MERCHANT, Naveed N; MCNULTY, Alyssa; RADCLIFF, Tiffany A; COTE, Murray J; FISCHER, Rebecca SB; SANG, Huiyan; ORY, Marcia G. COVID-19: Short term prediction model using daily incidence data. **PloS one**, Public Library of Science San Francisco, CA USA, v. 16, n. 4, e0250110, 2021.

ZHU *et al.* Attention-based recurrent neural network for influenza epidemic prediction. **BMC bioinformatics**, BioMed Central, v. 20, n. 18, p. 1–10, 2019.

APÊNDICE A – CORRELOGRAMAS DAS SÉRIES TEMPORAIS

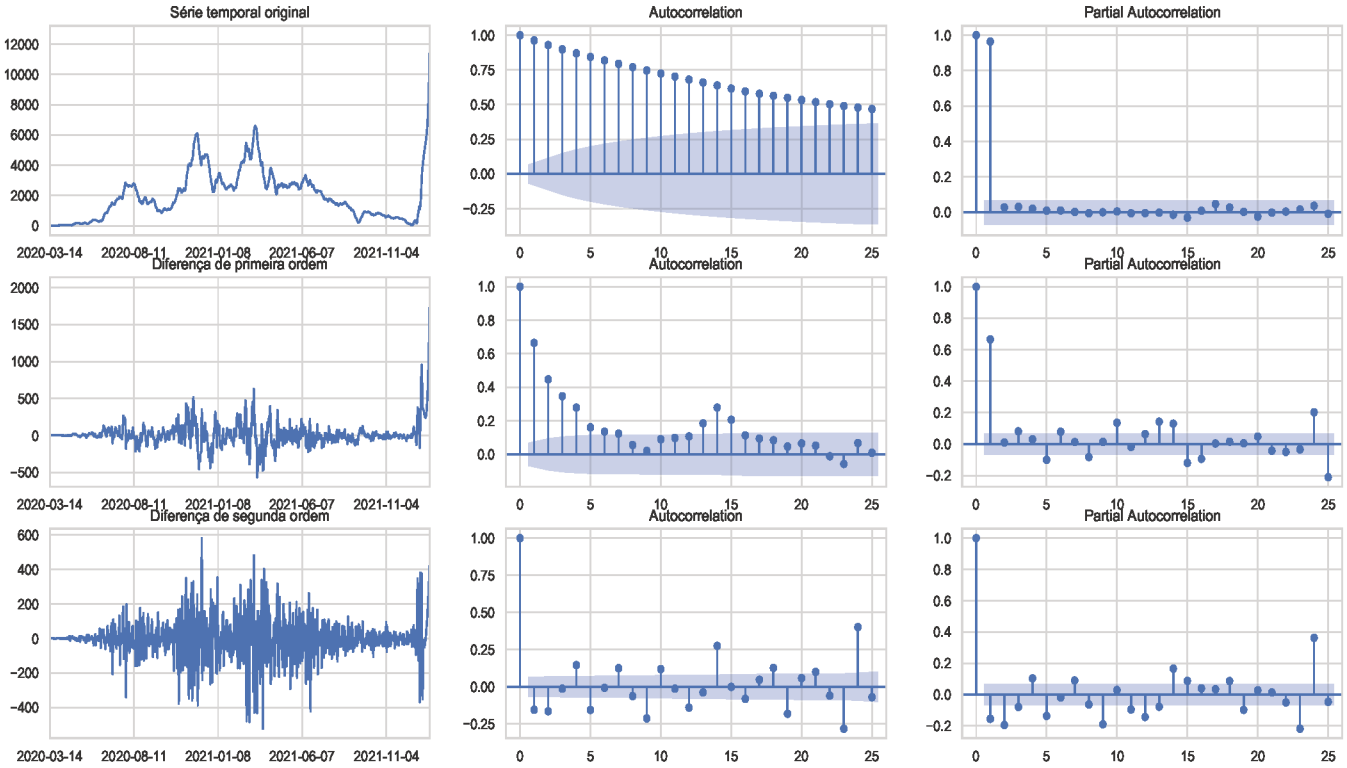


Figura 20 – Correlogramas para a série temporal de casos diários de infecção.

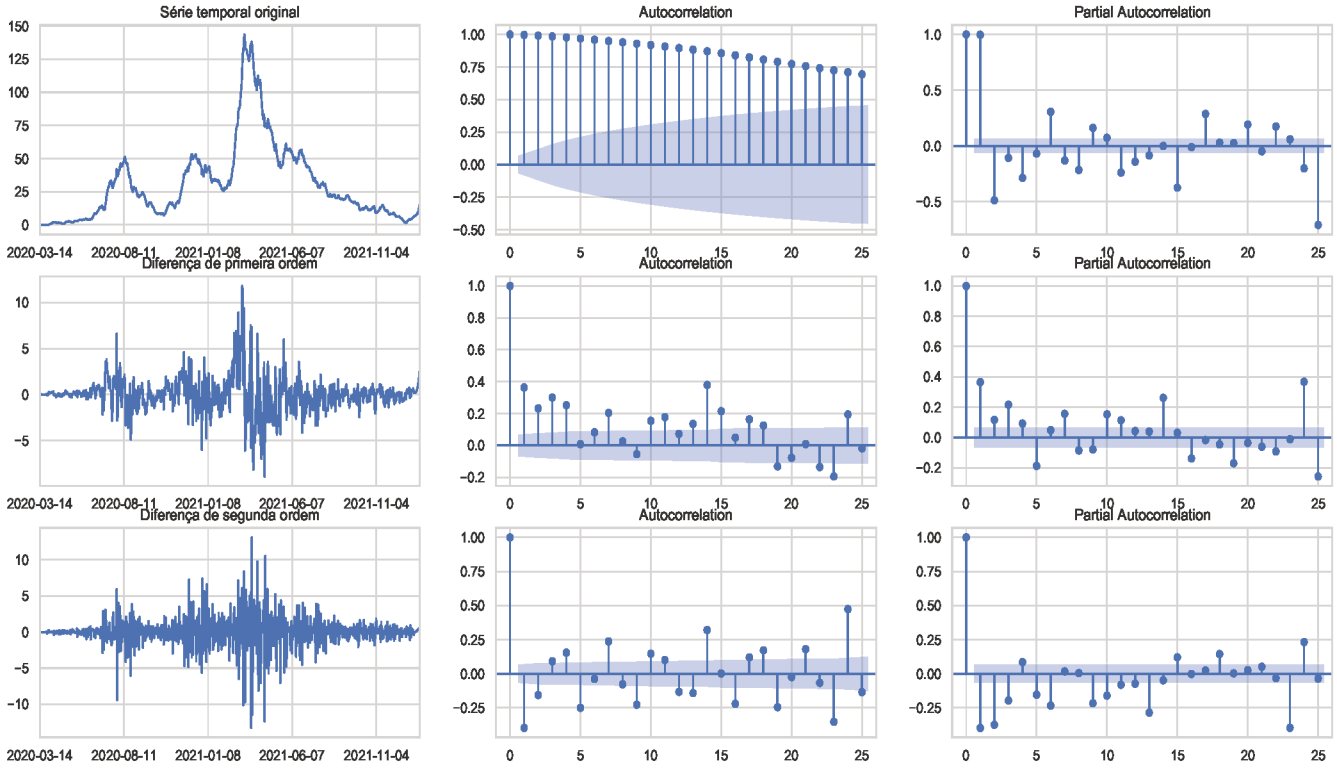


Figura 21 – Correlogramas para a série temporal de quantidade diária de óbitos.

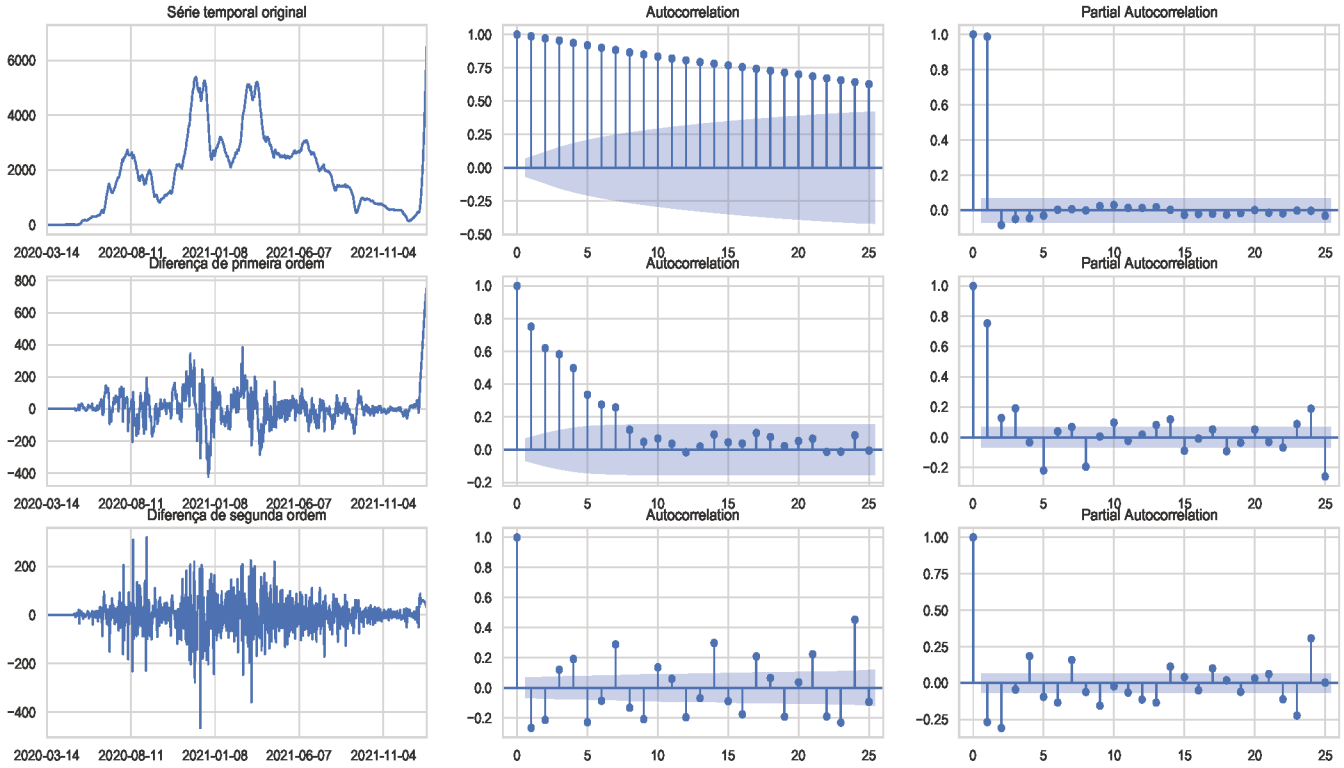


Figura 22 – Correlogramas para a série temporal de quantidade diária de recuperações.

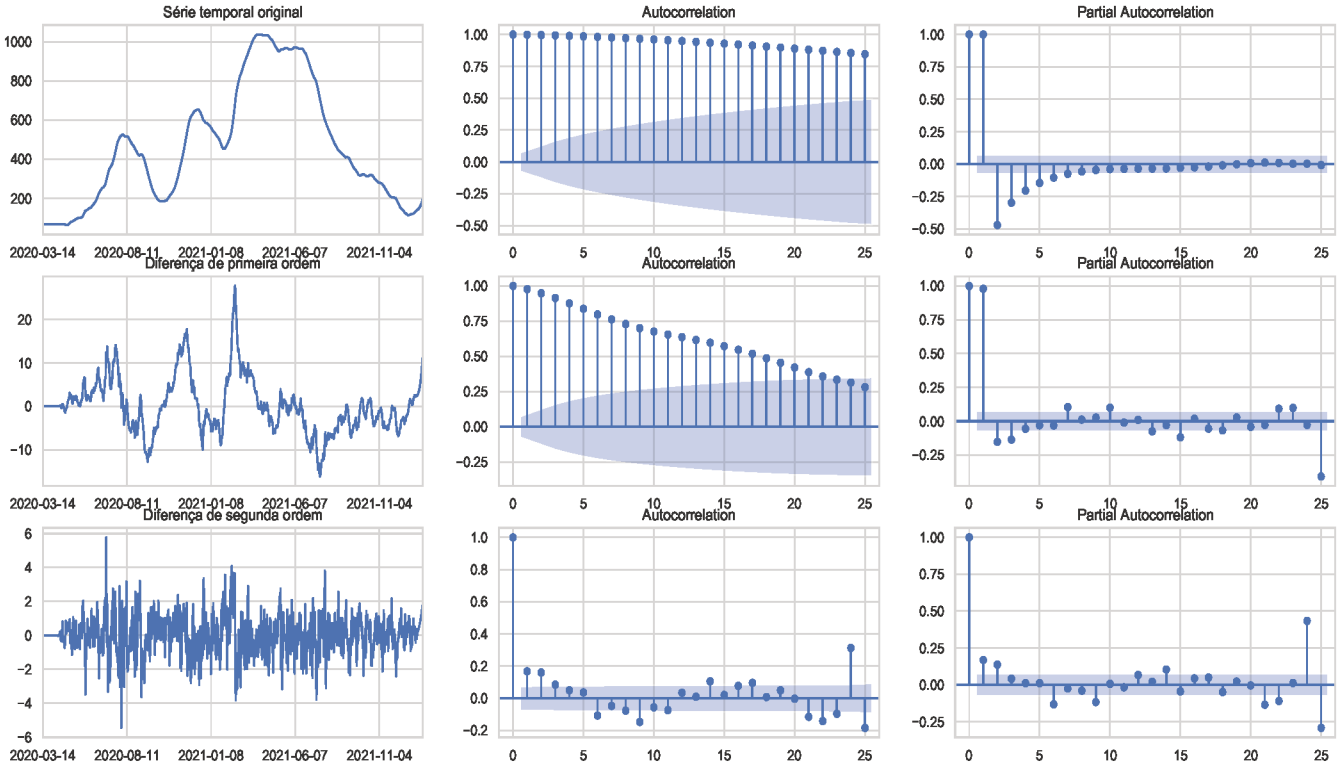


Figura 23 – Correlogramas para a série temporal de ocupação de leitos.

APÊNDICE B – RESULTADOS MÉTRICOS DOS MODELOS DE PREVISÃO

Tabela 23 – Desempenho, em relação às métricas MAE RMSE, dos modelos experimentados para as variáveis preditas.

Modelo	Casos		Óbitos		Recuperações		Oc. de leitos	
ARIMA	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
$p = 1$, $d = 1$, $q = 0$	88,679	129,538	1,379	2,104	51,230	73,860	1,021	1,293
$p = 2$, $d = 1$, $q = 0$	89,216	129,651	1,381	2,106	51,563	73,943	1,008	1,278
$p = 3$, $d = 1$, $q = 0$	88,703	130,22	1,375	2,074	49,483	72,077	0,996	1,271
$p = 4$, $d = 1$, $q = 0$	89,163	131,808	1,387	2,087	49,758	72,446	0,996	1,275
$p = 5$, $d = 1$, $q = 0$	87,979	130,158	1,359	2,061	48,016	70,728	1,004	1,282
$p = 6$, $d = 1$, $q = 0$	88,273	131,307	1,367	2,075	48,287	71,048	0,995	1,277
$p = 7$, $d = 1$, $q = 0$	88,319	132,149	1,376	2,083	48,312	71,450	0,997	1,883
$p = 1$, $d = 2$, $q = 0$	97,171	137,920	1,451	2,140	54,118	77,879	0,998	1,273
$p = 2$, $d = 2$, $q = 0$	93,929	135,656	1,453	2,144	50,173	74,013	0,986	1,265
$p = 3$, $d = 2$, $q = 0$	93,657	135,148	1,448	2,138	50,449	74,194	0,989	1,272
$p = 4$, $d = 2$, $q = 0$	94,069	135,842	1,454	2,149	49,911	73,607	0,995	1,276
$p = 5$, $d = 2$, $q = 0$	92,595	136,168	1,445	2,135	49,917	73,567	0,991	1,276
$p = 6$, $d = 2$, $q = 0$	92,334	137,262	1,450	2,141	48,856	73,073	0,990	1,278
$p = 7$, $d = 2$, $q = 0$	95,563	137,545	1,452	2,144	49,011	73,125	0,995	1,351
$p = 1$, $d = 1$, $q = 1$	99,462	152,181	1,380	2,082	53,211	76,412	1,031	1,393
$p = 2$, $d = 1$, $q = 1$	109,733	160,562	1,384	2,116	53,321	76,632	1,020	1,351
$p = 3$, $d = 1$, $q = 1$	107,075	162,986	1,377	2,081	54,986	77,655	1,012	1,342
$p = 4$, $d = 1$, $q = 1$	111,214	165,452	1,389	2,091	55,056	77,943	1,011	1,339
$p = 5$, $d = 1$, $q = 1$	112,312	166,731	1,365	2,068	55,612	78,129	1,011	1,337
$p = 6$, $d = 1$, $q = 1$	113,411	169,387	1,371	2,082	56,367	78,547	1,011	1,327
$p = 7$, $d = 1$, $q = 1$	113,632	170,002	1,381	2,090	57,098	79,121	1,015	1,360
NARX	Casos		Óbitos		Recuperações		Oc. de leitos	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
$l = 2$, atrasos = 1	129,437	224,751	1,606	2,529	70,616	110,957	4,480	5,680
$l = 2$, atrasos = 2	109,93	201,958	1,550	2,455	51,376	75,544	1,127	1,456
$l = 2$, atrasos = 3	117,881	248,023	1,622	2,693	53,042	79,486	1,166	1,475
$l = 2$, atrasos = 4	126,843	315,837	1,683	2,997	55,905	80,000	1,145	1,469
$l = 2$, atrasos = 5	150,649	464,931	1,783	3,185	56,738	81,346	1,150	1,480
$l = 2$, atrasos = 6	150,034	430,455	1,678	2,910	54,193	77,430	1,147	1,483
$l = 2$, atrasos = 7	134,688	323,812	1,777	3,037	54,181	77,195	1,271	1,729
$l = 3$, atrasos = 1	145,546	351,787	1,668	2,670	74,968	119,633	4,650	6,096

Tabela 23 continuada a partir da página anterior

l = 3, atrasos = 2	123,464	332,937	1,558	2,546	53,165	78,830	1,156	1,494
l = 3, atrasos = 3	129,374	327,491	1,700	2,957	55,375	85,537	1,179	1,498
l = 3, atrasos = 4	154,262	657,984	1,802	3,660	58,165	85,145	1,163	1,488
l = 3, atrasos = 5	185,861	961,302	2,041	4,723	58,501	85,145	1,186	1,579
l = 3, atrasos = 6	173,514	574,445	2,025	4,496	58,080	87,877	1,237	1,668
l = 3, atrasos = 7	159,287	458,442	2,134	4,534	58,062	87,591	1,317	1,821
l = 4, atrasos = 1	161,894	547,862	1,598	2,462	72,676	123,979	4,599	6,597
l = 4, atrasos = 2	144,140	567,959	1,566	2,607	56,442	87,621	1,168	1,517
l = 4, atrasos = 3	154,027	587,319	1,935	4,829	59,153	93,400	1,192	1,519
l = 4, atrasos = 4	215.327	1482,690	1,917	4,674	60,904	91,745	1,184	1,526
l = 4, atrasos = 5	273,263	2202,578	2,168	6,146	64,134	130,179	1,201	1,564
l = 4, atrasos = 6	227,730	778,389	2,396	6,407	62,639	101,244	1,225	1,616
l = 4, atrasos = 7	198,345	641,440	2,585	7,190	61,987	100,878	1,329	1,859
NARMAX	Casos		Óbitos		Recuperações		Oc. de leitos	
	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
l = 2, atrasos = 1	164,422	265,610	2,211	3,278	59,160	83,903	11,468	16,684
l = 2, atrasos = 2	122,703	200,583	1,963	2,897	51,844	76,559	1,752	2,828
l = 2, atrasos = 3	138,004	234,843	2,054	3,163	56,950	86,122	1,639	2,617
l = 2, atrasos = 4	143,387	306,071	2,323	3,635	84,562	123,698	1,437	2,084
l = 2, atrasos = 5	168,253	500,393	2,174	3,508	71,996	104,393	1,508	2,342
l = 2, atrasos = 6	156,146	410,293	1,815	3,067	71,139	99,273	1,431	2,133
l = 2, atrasos = 7	147,044	286,168	1,960	3,304	76,983	112,183	1,541	2,191
l = 3, atrasos = 1	170,764	331,945	1,950	2,957	79,854	122,961	8,814	12,943
l = 3, atrasos = 2	127,320	362,662	1.861	2,833	54,415	79,120	4,026	9,586
l = 3, atrasos = 3	145,116	355,589	2,134	3,464	55,891	89,013	2,114	6,874
l = 3, atrasos = 4	184,812	622,104	2,334	3,909	74,622	110,135	1,499	2,194
l = 3, atrasos = 5	248,416	1605,585	2,338	4,771	74,789	111,056	2,157	0,979
l = 3, atrasos = 6	205,146	619,624	2,827	6,761	75,126	112,322	—-	—-
l = 3, atrasos = 7	421,667	2184,288	8,021	55,336	75,655	112,796	—-	—-
l = 4, atrasos = 1	154,725	478,745	1,702	2,566	77,170	128,075	4,698	6,472
l = 4, atrasos = 2	145,601	678,733	1,737	2,737	58,619	92,716	3,452	4,982
l = 4, atrasos = 3	139,928	305,525	2,417	5,741	63,021	104,610	1,807	4,982
l = 4, atrasos = 4	227,021	1622,276	2,239	3,712	74,440	113,674	1,693	2,324
l = 4, atrasos = 5	302,932	2080,871	2,512	6,898	76,445	116,321	—-	—-
l = 4, atrasos = 6	234,411	823,936	2,724	8,278	76,586	116,455	1.527	2,368
l = 4, atrasos = 7	199,788	661,514	2,647	7,160	77,088	117,632	—-	—-

APÊNDICE C – GRÁFICOS PARA ANÁLISE DOS MODELOS DE PREVISÃO

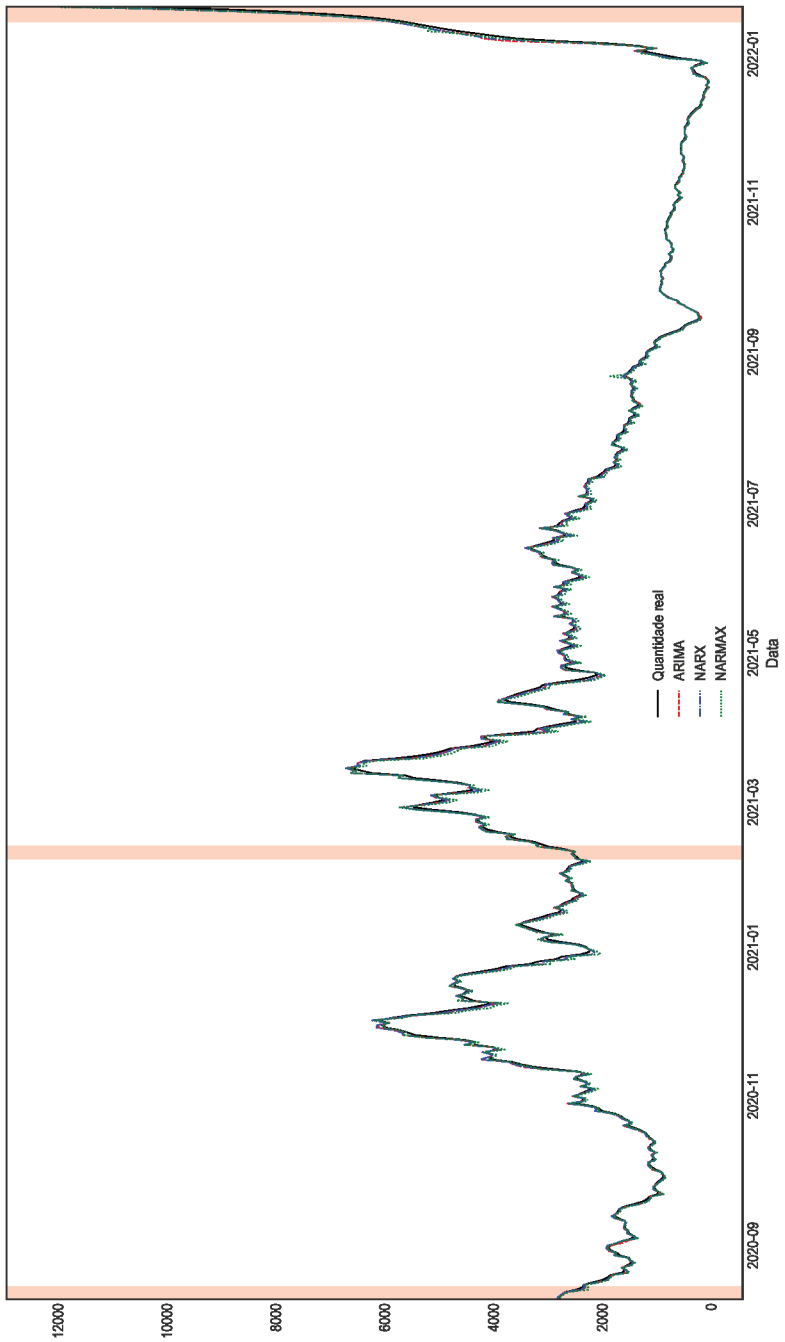


Figura 24 – Regiões de interesse para análise das previsões concernentes à série temporal de casos diários.

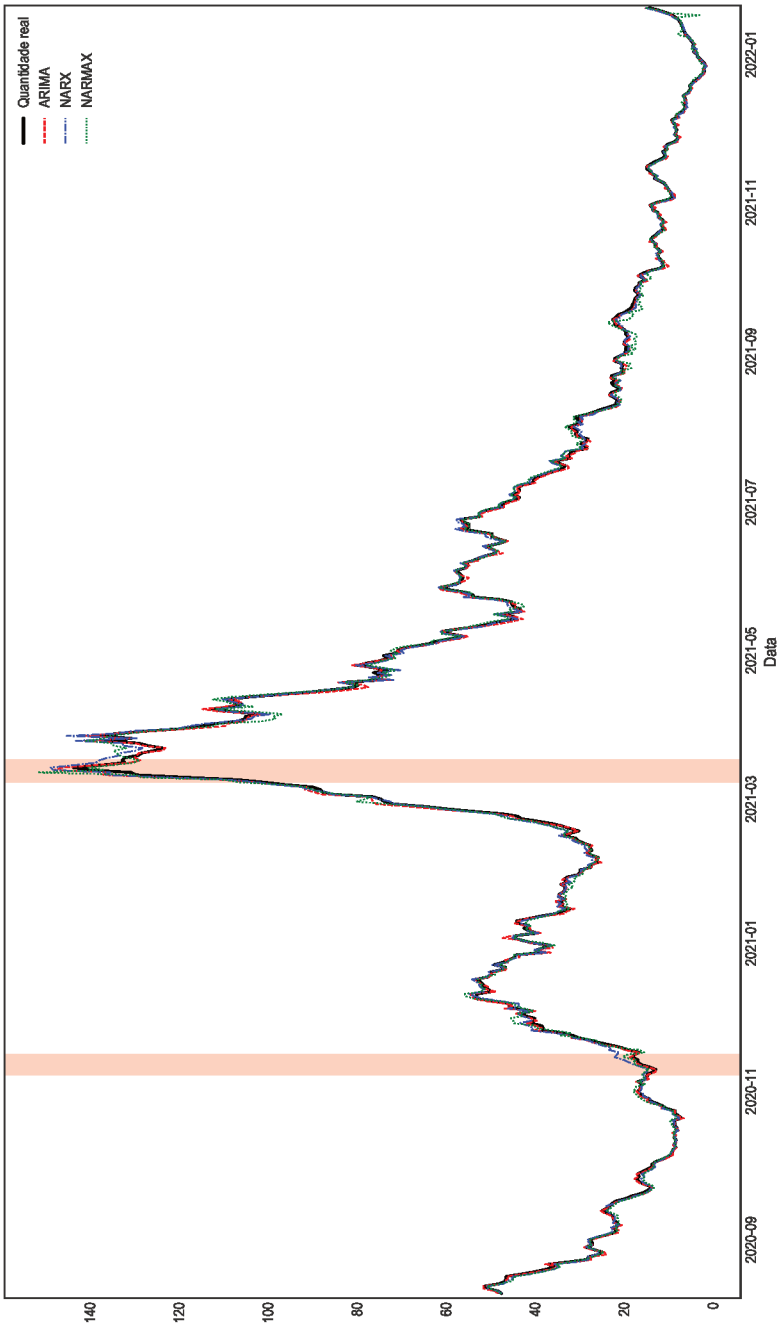


Figura 25 – Regiões de interesse para análise das previsões concernentes à série temporal de óbitos diários.

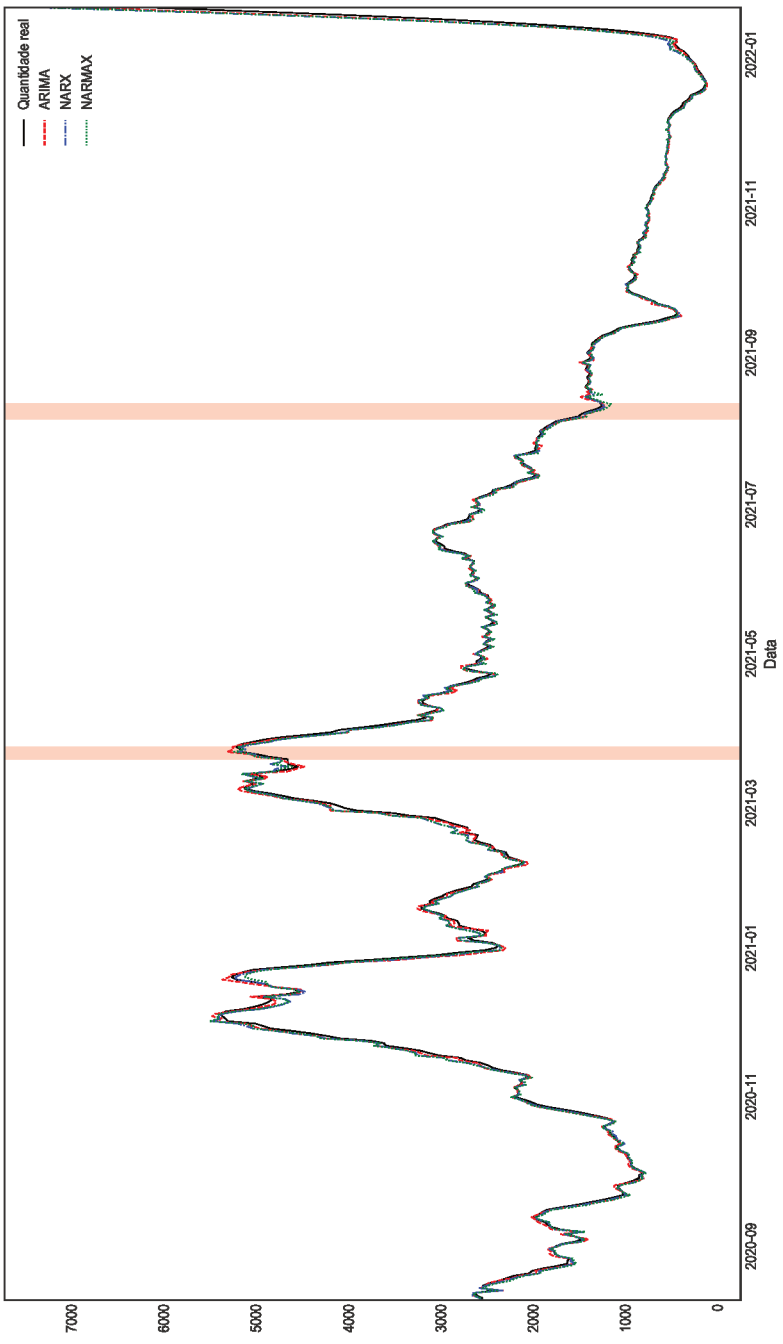


Figura 26 – Regiões de interesse para análise das previsões concernentes à série temporal de recuperações diárias.