UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

IÚRI BATISTA TELES

MAPDI - MODELO AUTOAJUSTÁVEL PARA PREDIÇÃO DO AUMENTO DO NÚMERO DE CASOS DE DOENÇAS INFECTOCONTAGIOSAS

Recife

2021

IÚRI BATISTA TELES

MAPDI - MODELO AUTOAJUSTÁVEL PARA PREDIÇÃO DO AUMENTO DO NÚMERO DE CASOS DE DOENÇAS INFECTOCONTAGIOSAS

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Inteligência Computacional

Orientador (a): Patrícia Cabral de Azevedo Restelli Tedesco

Recife

2021

Catalogação na fonte
Bibliotecária: Mônica Uchôa, CRB4-1010

T269m Teles, Iúri Batista.
MAPDI - Modelo Autoajustável para Predição do Aumento do Número de Casos de Doenças Infectocontagiosas / Iúri Batista Teles. – 2021.
113 f.: il., fig., tab; quad.

Orientadora: Patrícia Cabral de Azevedo Restelli Tedesco.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn. Programa de Pós-graduação em Ciência da Computação. Recife, 2021.
Inclui referências.

1. Doenças infectocontagiosas. 2. ARIMA. 3. LSTM. 4. Prophet. 5. CID. I. Tedesco, Patrícia Cabral de Azevedo Restelli (Orientadora). II. Título.

681.3 CDD (23. ed.) UFPE- CCEN 2021 - 192

Iúri Batista Teles

“MAPDI - Modelo Autoajustável para Predição do Aumento do Número de Casos de Doenças Infectocontagiosas”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 13/09/2021.

BANCA EXAMINADORA

Profa. Dra. Renata Maria Cardoso Rodrigues de Souza
Centro de Informática / UFPE

Prof. Dr. Sergio Crespo Coelho da Silva Pinto
Departamento de Ciência da Computação / UFF

Profa. Dra. Patrícia Cabral de Azevedo Restelli Tedesco
Centro de Informática/ UFPE
(Orientador)

Dedico esta conquista a minha mãe Irene pelo incentivo e apoio em todas as minhas escolhas e decisões.

AGRADECIMENTOS

Foi um processo intenso e com vários desafios, mas, em nenhum momento, estive sozinho na minha trajetória. A presente dissertação teve o apoio de várias pessoas, que deram a sua contribuição de forma direta e indireta. Com isso quero deixar os meus agradecimentos para a minha família e, em especial, à minha mãe, **Irene Batista Teles**, meu irmão, **Igor Batista Teles**, e minha tia **Maria José dos Santos**, por terem me apoiado e acompanhado os meus passos, sempre que necessário.

Também gostaria de agradecer à minha orientadora, Profª. **Patrícia Cabral de Azevedo Restelli Tedesco**, por participar e me apoiar nesta jornada, incentivando, orientando e permanecendo sempre terna. Uma grande profissional, admirável mulher. E, agradecer também a **Ilda Tedesco** pela excelente ajuda nas correções desde trabalho, meu muito obrigado a vocês duas.

Agradeço em especial a **Cloves Alberto Chaves Lima** pela experiência, apoio, motivação, persistência, incentivo, análise e orientação no decorrer de toda esta trajetória. Cloves, foi a essência fundamental para a finalização e prospecção da evolução gradual deste trabalho.

Quero também demonstrar o meu carinho por todos aqueles que, indiretamente ou diretamente, possibilitaram a conclusão desta dissertação, seja pela compreensão do meu esforço e dedicação ou por toda a motivação que me passaram no decorrer destes 2 anos.

RESUMO

Esta dissertação propõe-se a apresentar a construção de um processo nomeado de MAPDI (Modelo autoajustável para previsão doenças infectocontagiosas), que visa ajudar gestores em saúde na tomada de decisões, de forma proativa, em relação ao crescimento desenfreado do número de casos de doenças infectocontagiosas. Com isso, fizemos o uso de três algoritmos de séries temporais (ARIMA, LSTM e *Prophet*) nos dados oriundos de consultas clínicas, realizadas nas unidades especializadas em SRAG (Síndrome Respiratória Aguda Grave), do município do Recife, localizado no estado de Pernambuco. Os dados utilizados são referentes aos diagnósticos aplicados nos prontuários médicos de pacientes, por meio da Classificação Internacional de Doenças (CID) ou por meio da Classificação Internacional de Assistência Primária (CIAP). Com isso, os CIDs/CIAPs que apresentarem comportamentos anômalos (maior quantidade de ocorrência) serão encaminhados de forma automatizada para os algoritmos de séries temporais, que auto ajustarão seus parâmetros visando entregar os melhores resultados para previsão da semana seguinte. O nosso estudo não tem como premissa definir quais são os melhores algoritmos, mas sim validar se não há diferença estatística entre os resultados obtidos e os dados observados. Assim, disponibilizamos para os gestores três dos possíveis cenários para o crescimento ou declínio do número de casos para doenças com alto grau de transmissão, facilitando a tomada de decisão de forma proativa e não reativa, como vem sendo realizado na saúde pública do país. Os resultados encontrados em relação aos modelos produzidos a partir da automação, em comparação aos dados observados, apresentaram equivalência. Com isso, observamos que o uso de mais de um algoritmo pode complementar a percepção dos gestores na tomada de decisão, sem divergir das ocorrências reais.

Palavras-chave: doenças infectocontagiosas; ARIMA; LSTM; *Prophet*; CID.

ABSTRACT

This dissertation aims to demonstrate the construction of a process named MAPDI (Self-adjusting Model for Predicting the Increase in the Number of Cases of Infectious Diseases, or "Modelo Autoajustável para Predição do Aumento do Número de Casos de Doenças Infectocontagiosas", in Portuguese), which aims to help health managers in making proactive decisions with regards to the unrestrained growth in the number of cases of infectious diseases. Thus, we used three time series algorithms (ARIMA, LSTM and Prophet) in data collected from clinical appointments, carried out in units specialized in SRAS (Severe Acute Respiratory Syndrome), in the municipality of Recife, capital of the state of Pernambuco. The data used refer to diagnoses applied to the medical records of patients through the International Classification of Diseases (ICD) or through the International Classification of Primary Care (ICPC). Thus , the ICDs/ICPCs that present anomalous behavior (higher number of occurrences) will be forwarded in an automated way to the time series algorithms that will self-adjust their parameters in order to deliver the best results. Our study does not have the premise of defining the best algorithms, but of rather validating that there is no statistical difference between them and the observed data. Thus, we provide managers with three scenarios of possible increases or decreases in the number of cases for diseases with a high degree of transmission, facilitating decision-making in a proactive and non-reactive way, as has been the norm in the Brazilian Public Health System. The results found in relation to the models generated in comparison with the observed data were quite satisfactory and, therefore, consolidated our initial hypothesis that the usage of more than one algorithm may complement the initial perception of the decision-makers without presenting divergentes in comparison to the real data.

Keywords: infectious diseases; ARIMA; LSTM; *Prophet*; ICD.

LISTA DE FIGURAS

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Aprendizagem de Máquina
ARIMA	*Autoregressive Integrated moving Average*
CID	Classificação internacional de doenças
CIAP	Classificação Internacional de Assistência Primária
IA	Inteligência Artificial
LSTM	*Long Short-Term Memory*
MAPDI	Modelo Auto-Ajustável para Predição de Doenças Infectocontagiosas
SARG	Síndrome Respiratória Aguda Grave
SARS	*Severe Acute Respiratory Syndrome*

SUMÁRIO

1 INTRODUÇÃO

Embora as Síndromes Respiratórias Agudas Graves (SARS, do inglês *Severe Acute Respiratory Syndrome*, ou SRAG, em português) possam ter um potencial devastador como infecção, modelos computacionais conhecidos na literatura podem prover contribuições significativas no combate à disseminação destas doenças. Modelos dinâmicos, aqueles que toleram variações nos dados de entrada, são muito úteis em gerar *insights* para os gestores de órgãos de saúde pública, tornando-se ainda mais relevantes quando as informações sobre uma possível emergente ou re-emergente epidemia ainda são limitadas (MASSAD et al., 2005).

A SARS tornou-se uma ameaça à saúde global, devido ao seu alto grau de transmissibilidade. Para doenças infecciosas como esta, alguns protocolos (práticas de controle e avanço de doenças infectocontagiosas), já consolidados na academia, são utilizados para conter ou atenuar o seu avanço, como as medidas de isolamento horizontal e vertical, e o acompanhamento dos infectados (LEE et al., 2003). De acordo com Souza et al. (2020) em seu estudo, abrangendo os anos de 2010 a 2017, identificou-se que os municípios brasileiros localizados sobretudo na região Norte, parte do Nordeste e Centro-Oeste necessitam de uma maior atenção em relação às doenças infecciosas. Pobreza e saneamento básico são indicadores que influenciam o grau de desafios que a região irá enfrentar para lidar com as doenças, visto que as medidas de controle, citadas anteriormente, podem ser difíceis de serem implementadas na prática.

No estado de Pernambuco, o espaço demográfico e socioeconômico do município do Recife revela grandes contrastes que refletem a complexidade de pensar nas políticas e nos programas sociais aplicados à saúde pública. No âmbito da atenção primária à saúde [1] existem diversas estratégias e metodologias que visam a prevenção, controle, diagnóstico e tratamento de doenças, evitando assim o agravamento e a sua própria evolução. Entretanto, devido à situação delicada da saúde em todo o país, e com o seu agravamento por meio da nova pandemia do SARS-COV2, algumas dessas ações podem estar sendo negligenciadas.

Recife possui um território diversificado e desafiador para o gerenciamento em saúde, formado geograficamente por: morros — 67,43%, planícies — 23,26%, áreas aquáticas — 9,31%, Zonas Especiais de Preservação Ambiental (ZEPA) — 5,58% (MUNICIPAL et al., 2014) de sua região.O município está dividido em 94 bairros, 6 Regiões Político-Administrativas

[1] É o primeiro nível da atenção em saúde e se caracteriza por um conjunto de ações de saúde, no âmbito individual e coletivo, que abrange a promoção e a proteção da saúde, a prevenção de agravos, o diagnóstico, o tratamento, a reabilitação, a redução de danos e a manutenção da saúde com o objetivo de desenvolver um cuidado integral que impacte positivamente na situação de saúde das coletividades.

(RPA) e 8 Distritos Sanitários (DS) (MUNICIPAL et al., 2014).

Atualmente, devido às características heterogêneas da cidade, os distritos sanitários comportam bairros com diferentes níveis sociais e econômicos, o que dificulta ainda mais o acompanhamento dos infectados, quando consideradas as características homogêneas dos habitantes, que faz parte de uma premissa do controle e prevenção (MENDES; TEIXEIRA, 1993).

Devido às dificuldades apresentadas anteriormente, que não são apenas uma particularidade da cidade do Recife, é possível observar um crescente movimento no compartilhamento de dados e estudos sobre doenças infecciosas, disponíveis na 'internet' (GRASSLY; FRASER, 2008). Este movimento, possibilita a construção de aplicações que utilizam modelos matemáticos e/ou computacionais para análise científica desses dados, bem como o desenvolvimento de estratégias práticas de controle de doenças que foram e estão sendo utilizadas por gestores de saúde em todo o mundo. Esses dados também podem ser utilizados para o aprimoramento e modelagem de estruturas inteligentes, que tratem com eficiência e rapidez as informações e, com isso, agregar ainda mais valor para a tomada de decisão das políticas de saúde pública. Todavia, além desses dados públicos, o acesso diário às informações advindas das unidades de saúde, por meio de modelagens inteligentes, podem auxiliar na tomada de decisão proativa dos gestores. Ou seja, antes de uma dada doença infecciosa se tornar um grave problema de saúde pública, os gestores poderiam atuar de maneira preditiva, prática e rápida evitando uma propagação desenfreada do patogênico. Para tal, as aplicações de aprendizagem de máquina (AM) (ou do inglês: *Machine Learning*-ML) e aprendizagem profunda (ou do inglês: *Deep Learning*-DL) na sociedade vêm ganhando mais notoriedade, devido ao aprimoramento do potencial computacional e obtenção de insumos (dados e informações) para treinamento, avaliação e aperfeiçoamento por parte dos modelos desenvolvidos (CHEN; LIU; PENG, 2019). Além disso, o desenvolvimento das tecnologias da informação e comunicação vêm proporcionando uma melhor coleta, armazenamento e distribuição dos dados necessários para os estudos e as análises que estão sendo realizados em todo o globo.

Atualmente, existem diversos estudos literários relacionados à predição do aumento dos números de casos das doenças infecciosas (MASSAD et al., 2005; OLSAVSZKY et al., 2020; BATTINENI; CHINTALAPUDI; AMENTA, 2020; ABBOTT et al., 2020). Muitos desses estudos visam à aplicação de seus modelos em ambientes reais, para ajudar na antecipação dos eventos e condições que possam fugir do controle. Modelos matemáticos e estatísticos vêm sendo aprimorados com o passar dos anos, com o intuito de se adaptarem aos novos modus operandi das mutações dos vírus com altos índices de contaminação, por exemplo. Com isso, o planejamento a

curto, médio e longo prazo pode apresentar desafios na elaboração de novas metodologias que possam auxiliar uma tomada de decisão que evite o crescimento desenfreado de uma dada doença. Estes desafios podem ser as incertezas que uma nova mutação traz consigo como taxa de mortalidade, progressão de casos, contaminação, hospitalização, etc. Entretanto, diante de tais desafios, dados coletados continuamente, de forma acurada e segura, podem ajudar significativamente na antecipação da resposta a esses agentes infecciosos (LEMOS, 2006).

Neste trabalho, são utilizados alguns dos principais algoritmos de predição de séries temporais, como o ARIMA, LSTM e *Prophet*, com o intuito de buscar uma predição de melhor acurácia nas projeções de ocorrências de casos futuros para uma dada doença. Esta iniciativa visou analisar preditivamente CIDs/CIAPs (Código internacional de doenças/Classificação internacional de assistência primária) que possam requerer atenção por parte dos gestores de saúde. Utilizamos os três algoritmos citados como o mesmo conjunto de dados. Com o pré-processamento dos dados, obtivemos diversos resultados considerando os CIDs/CIAPs com maior ocorrência em unidades especializadas no tratamento da SRAG no Recife, capital do estado de Pernambuco. A partir daí, construímos um processo chamado MAPDI (Modelo Autoajustável para Predição do Aumento do Número de Casos de Doenças Infectocontagiosas), o qual consiste em efetuar, de forma automática, a otimização dos modelos e propor a integração, por meio da indexação dos resultados, e a visualização no sistema *Kibana*[2], atualmente utilizado para análise dos dados das teleorientações da cidade do Recife.

Para a construção do MAPDI utilizamos o método da previsão quantitativa, devido à necessidade de realizar as previsões sobre a demanda. Essa necessidade ocorre devido ao comportamento dos dados visto que, diariamente, há uma quantidade significativa de atendimento clínico, gerando assim vários diagnósticos que possuem um ou mais CIDs/CIAPs associados. Logo, para identificarmos os CIDs/CIAPs anômalos (i.e., aqueles que apresentam maior ocorrência), precisamos coletar diariamente esses dados e, com isso, criarmos projeções para a semana seguinte.

Este estudo não visa apresentar um estudo comparativo dos algoritmos anteriormente citados, mas sim usá-los em conjunto para possibilitar várias visões do comportamento de uma dada doença na sociedade, usando informações que não são muito utilizadas devido à dificuldade de sua coleta, os CIDs/CIAPs. Com isso, o MAPDI não foi proposto apenas para a necessidade atual da Covid-19, mas para dar subsídio para outras doenças infectocontagiosas que assolam o país anualmente, elevando os custos com a saúde e afetando diretamente aquela

[2] *Kibana*: <https://www.elastic.co/pt/what-is/kibana>

parcela da população menos favorecida.

Neste Capítulo, buscamos mostrar, em linhas gerais, o contexto no qual o presente trabalho está inserido e um resumo das contribuições feitas por sua elaboração. No Capítulo 2 está a fundamentação teórica, onde abordamos os conceitos relacionados às doenças infecciosas que serão tema do nosso estudo. Também apresentamos como as séries temporais são utilizadas no contexto epidemiológico e quais os critérios para a escolha dos algoritmos utilizados na presente pesquisa. No Capítulo 3, encontram-se os trabalhos relacionados, onde trazemos uma breve discussão sobre as pesquisas similares. O Capítulo 4 visa apresentar as metodologias utilizadas, descrevendo desde a origem dos dados utilizados até os conceitos associados às métricas de análises dos algoritmos. No Capítulo 5 apresentamos com detalhes a construção do processo MAPDI, onde nas seções também foram abordados alguns conceitos que consolidam as fases do processo proposto. O Capítulo 6 apresenta a construção do modelo para cada algoritmo, incluindo a escolha dos melhores parâmetros e seus respectivos resultados. A conclusão é apresentada no Capítulo 7, com as limitações encontradas e sugestões de continuidade a partir desta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Atualmente as aplicações de inteligência artificial (IA), tais como, aprendizagem de máquina (ML, do inglês *Machine Learning* ou AM, em português), e sua subárea à aprendizagem profunda (DL, do inglês *Deep Learning*), estão conquistando cada vez mais espaço em todas as áreas do conhecimento, tendo sua aplicação na economia, ciências naturais, finanças entre outros (BOX et al., 2015). A versatilidade e adaptabilidade nos problemas do cotidiano faz com que as aplicações que utilizam IA possam ser utilizadas em diversas esferas da sociedade, sendo capazes de ajudar na compreensão do comportamento de novas doenças e, assim, contribuir positivamente para a gerenciamento da saúde (OLSAVSZKY et al., 2020).

Tendo em vista que a aprendizagem de máquina (AM) consiste na representação dos dados de entrada e a generalização de padrões contidos nos dados (NAJAFABADI et al., 2015), IA tem o potencial de identificar, classificar e prever um determinado acontecimento ou característica. Aplicações utilizando IA podem ser utilizadas em diversas esferas da sociedade, dentre elas a AM e a aprendizagem profunda pode ajudar a lidar com doenças e melhorar os sistemas de saúde (OLSAVSZKY et al., 2020). Considerando as doenças infecciosas, o controle ocorre pelas intervenções ecológicas, terapêuticas, saneamento e, frequentemente, medidas de controle específicas para cada doença, como medidas socioeducativas e a conscientização da população (SOUZA et al., 2020).

Chen, Liu e Peng (2019) afirmam que, para as aplicações AM que analisam situações reais, diversos fatores podem influenciar na construção de um modelo de IA, como: tempo da inferência — no qual o resultado precisa ser obtido; calibrações — onde são considerados os ajustes necessários para uma melhor precisão dos resultados; interpretabilidade — o quão os resultados precisam ser compreendidos; "*overfitting*" — o modelo é pouco genérico para novos resultados; e "*underfitting*" — onde os resultados são pouco satisfatórios mesmo nas execuções de teste e treino do modelo. Fatores devem ser analisados em conjunto com as características dos dados e resultados esperados.

O entendimento de doenças infecciosas e sua abrangência na sociedade se faz necessário para um melhor entendimento dos padrões e desafios que os modelos de predição temporal podem enfrentar. A utilização de AM na previsão das ocorrências em uma determinada série temporal é objeto de estudo de diversas pesquisas. Séries temporais relacionadas às doenças com potencial de aumento no número de casos, como: as doenças infectocontagiosas

(MARTINEZ; SILVA; FABBRO, 2011; TUITE et al., 2011; BRAS et al., 2014; AZEEZ et al., 2016; OLSAVSZKY et al., 2020) e mais recentemente relacionado à Síndrome Respiratória Aguda Grave (SRAG) (RIBEIRO et al., 2020; SHASTRI et al., 2020), possuem considerável relevância de pesquisas.

Uma série temporal nada mais é do que uma variável descrita no decorrer do tempo. No entanto, outros aspectos também devem ser considerados tais como: a estacionariedade, sazonalidade, se as variáveis alvo são automaticamente correlacionadas, entre outros. É composta por 4 elementos base: tendência - o sentido do deslocamento ao longo do tempo; ciclo - movimento ondulatório temporalmente; sazonalidade - movimento ondulatório em período curto de tempo em geral inferior a um ano e o erro - variável intrínseca dos dados que não pode ser modelada (MORETTIN; TOLOI, 1985).

2.1 DOENÇAS INFECTOCONTAGIOSAS

Na literatura encontramos estudos sobre diversas doenças infecciosas (IVORRA et al., 2020; MUELLNER et al., 2018; BETTENCOURT; RIBEIRO, 2008; SMIESZEK; FIEBIG; SCHOLZ, 2009; CHOISY; GUGAN; ROHANI, 2007). Alguns desses estudos tentam relatar o impacto de tais doenças na sociedade, considerando o seu período histórico e quais avanços foram necessários para mitigá-las (ANDERSON; ANDERSON; MAY, 1992).

Muitas doenças infecciosas, que ainda geram preocupações às autoridades de saúde, possuem séculos de história. A tuberculose, por exemplo, teve a sua primeira aparição na Índia há 3.300 anos e depois sua segunda aparição na China há 2.300 anos (LEVNTAL, 1957). A hanseníase, outra doença preocupante, possui relatos de ocorrência no antigo Egito (THIN, 1891). Doenças infecciosas como a sífilis tornaram-se conhecidas na Europa no final do Século XV, disseminando-se rapidamente pelo continente, e posteriormente assolando a população em escala mundial (AVELLEIRA; BOTTINO, 2006). Apesar do longo período em que a humanidade lida com as doenças infectocontagiosas, ainda é muito difícil erradicá-las por completo, visto que diversos fatores interferem na extinção total dos casos, desde fatores sociais até os naturais.

O hospedeiro de uma doença infecciosa pode ser o vetor de sua transmissão. A infecção pode ou não culminar em lesões derivadas do próprio hospedeiro em resposta manifestada às alterações fisiológicas, bioquímicas e histopatológicas dos agentes intrusos (BEZERRA, 2016). De acordo com (MORENS; FAUCI, 2013), o diagnóstico de doenças infecciosas aconteciam

mesmo sem o conhecimento da existência dos agentes infecciosos. Segundo Bezerra (2016) os agentes infecciosos são microrganismos como tais vírus, bactérias, protozoários ou fungos, que podem estar presentes nos organismos sem causar danos aparentes. No entanto, quando há alguma alteração no sistema imunológico ou alguma outra condição clínica, esses microrganismos podem causar danos que facilitam a entrada de outros microrganismos. Ainda conforme o autor, as doenças infecciosas podem tornar-se preocupantes para uma comunidade, devido à sua associação com o número expressivo de transmissores da doença.

Em virtude da dinâmica da infecção ser não linear, o comportamento dos agentes infecciosos possuem complexidade adicional relacionada ao meio de interação (direta ou indireta), entre os indivíduos suscetíveis à infecção, variando assim, a taxa de novas infecções e a evolução temporal do número de infectados (BELLAN et al., 2012). Diversos fatores podem contribuir para o aparecimento/surgimento destas doenças, como: fatores socioeconômicos, genéticos e biológicos (MORENS; FOLKERS; FAUCI, 2004; WEISS; McMichael, 2004; ZHU; GILLINGS; PENUELAS, 2020; WOOLHOUSE et al., 2001). Deste modo, encontrar medidas de prevenção eficientes é um desafio, tanto para os gestores públicos quanto para os pesquisadores.

As doenças infecciosas aumentam seus riscos à sociedade quando estão associadas a um expressivo nível de transmissão. O atual cenário mundial deixa clara a complexidade que há no controle de diversas doenças. A pandemia, do até então recente vírus Covid-19, mostra que não só apenas os avanços que aconteceram na interligação das fronteiras mundiais auxiliaram na fácil comunicação e coordenação de ações em escala global, como foram fatores-chave para a disseminação da doença. A tecnologia digital atual, dentre outras atribuições, serviu como ferramenta de rastreamento da evolução da disseminação do vírus, possibilitando um gerenciamento mais reativo de agir, ou seja, as ações serão tomadas a partir do comportamento relativo à disseminação do vírus. Entretanto, metodologias proativas de identificação e mensuração de número de casos de agentes infecciosos trazem consigo resistências que dificultam sua aceitação em âmbito geral (TAYLOR, 2019), uma vez que a proatividade pode estar lidando com especulações. Doenças infectocontagiosas, a exemplo do HIV/AIDS, ou casos como H1N1, e a Síndrome Respiratória Aguda Grave (SRAG), possuem impactos globais e a sua identificação, tratamento e/ou cura dependem inteiramente de políticas públicas. Tais políticas podem divergir de região para região, ocasionando surtos e mortes anualmente, principalmente em esferas sociais de vulnerabilidade (MORENS; FOLKERS; FAUCI, 2004; JONES et al., 2008).

Apesar do enorme avanço tecnológico, a globalização e o acesso mais facilitado aos meios

de locomoção, trazem consigo um aumento significativo na complexidade de contenção e estudos relacionados ao avanço das doenças infecciosas (MORENS; FAUCI, 2013). Quanto mais interação entre as pessoas e regiões, maiores são as possibilidades de novas surtos, endemias, epidemias e pandemias ocorrerem.

Doenças infecciosas podem causar diversos transtornos à sociedade, e muitas delas não possuem alternativas práticas que possibilitem sua efetiva erradicação. A sociedade convive com doenças infecciosas, e a iminência do aumento dos casos é uma constante (TAYLOR, 2019). Assim, os estudos e análises de tais doenças podem possibilitar a aplicação de ações práticas, visando mitigar maiores danos causados à população. Diversos estudos, através de meios computacionais, estão relacionados à análise da dinâmica de tais doenças (ROOSA; CHOWELL, 2019; VOLKOVA et al., 2017). Da mesma forma, diversas doenças são objeto de estudo, e SRAG obteve recentemente um grau de relevância global considerável. A disseminação do vírus Covid-19, identificado no final de 2019 (ZHI, 2020), resultou nessa alta relevância de SRAG, devido à pandemia ocasionada.

2.2 SÍNDROME RESPIRATÓRIA AGUDA GRAVE

A Síndrome Respiratória Aguda Grave (SRAG) é uma doença viral emergente do século XXI, com alta capacidade de disseminação entre a população. Esta doença possui características de propagação similar ao resfriado e à gripe, e sua transmissão pode ocorrer de diferentes formas, desde superfícies contaminadas ao simples contato com um hospedeiro infectado (ORGANIZATION, 2021). Na sua grande maioria, os novos casos desta infecção são identificados a partir do aumento das internações, por meio das unidades locais de atenção básica e hospitais que reportam as ocorrências durante a propagação da doença. A SRAG tornou-se uma ameaça à saúde global, e com alto grau de transmissão. As chaves para combatê-la são medidas preventivas proativas, isolamento social e tratamento eficaz para os infectados (LEE et al., 2003).

Os primeiros relatos da SRAG foram alertados pela *World Health Organization* (WHO) com a assistência do *Global Outbreak Alert and Response Network* (GOARN). A identificação do surto, no final de fevereiro de 2003, foi iniciada por um caso no Vietnã de um paciente de Hanói, vindo de Xangai e Hong Kong, China. Através de um paciente infectado por um vírus até então desconhecido, o hospital teve mais 20 casos com sintomas semelhantes (ORGANIZATION, 2003; XU et al., 2004).

A SRAG possui uma alta taxa de contágio, podendo ocasionar graves problemas à saúde pública: os gestores públicos precisam tomar medidas de controle e contenção do aumento das ocorrências de novos casos, além de prover suporte à parcela da população infectada (BITOUN et al., 2020). Um dos vírus causadores de infecções classificadas por SRAG, é o SARS-CoV-2 que possui uma taxa de hospitalização, a depender da carga viral, de 29,1% em carga viral alta, 20,8% para carga viral média e 14,9% para a baixo (MAGLEBY et al., 2020). A incubação é um fator relevante ao considerar a dinâmica da doença, uma vez que os pacientes tendem a ocupar um leito hospitalar por diversos dias.

Considerando as probabilidades, quanto maior o número de ocorrências, maior será a possibilidade de ocorrer superlotação nas unidades de saúde, considerando que os leitos também são usados para outras enfermidades. Tais taxas podem ser vistas como umas das métricas para análise da necessidade de criação de novos leitos, antes que ocorra a superlotação das unidades, por exemplo. Ou seja, considerando as taxas de ocorrência e a situação dos pacientes, os gestores poderão melhor avaliar a necessidade de aumento ou redução de leitos hospitalares, além de gerenciar os recursos humanos necessários para atender um possível iminente aumento de demanda, diminuindo óbitos que talvez pudessem ser evitáveis.

Segundo o estudo da Health (2020), apesar da substancial queda entre os anos 2000 a 2019, período em que foram registradas globalmente 2,6 milhões de mortes , infecções do trato respiratório inferior são classificadas como a quarta causa de morte global, sendo que as mortes causadas pelo SRAG, devido à pandemia do Sars-Cov-2, ainda estão em curso.

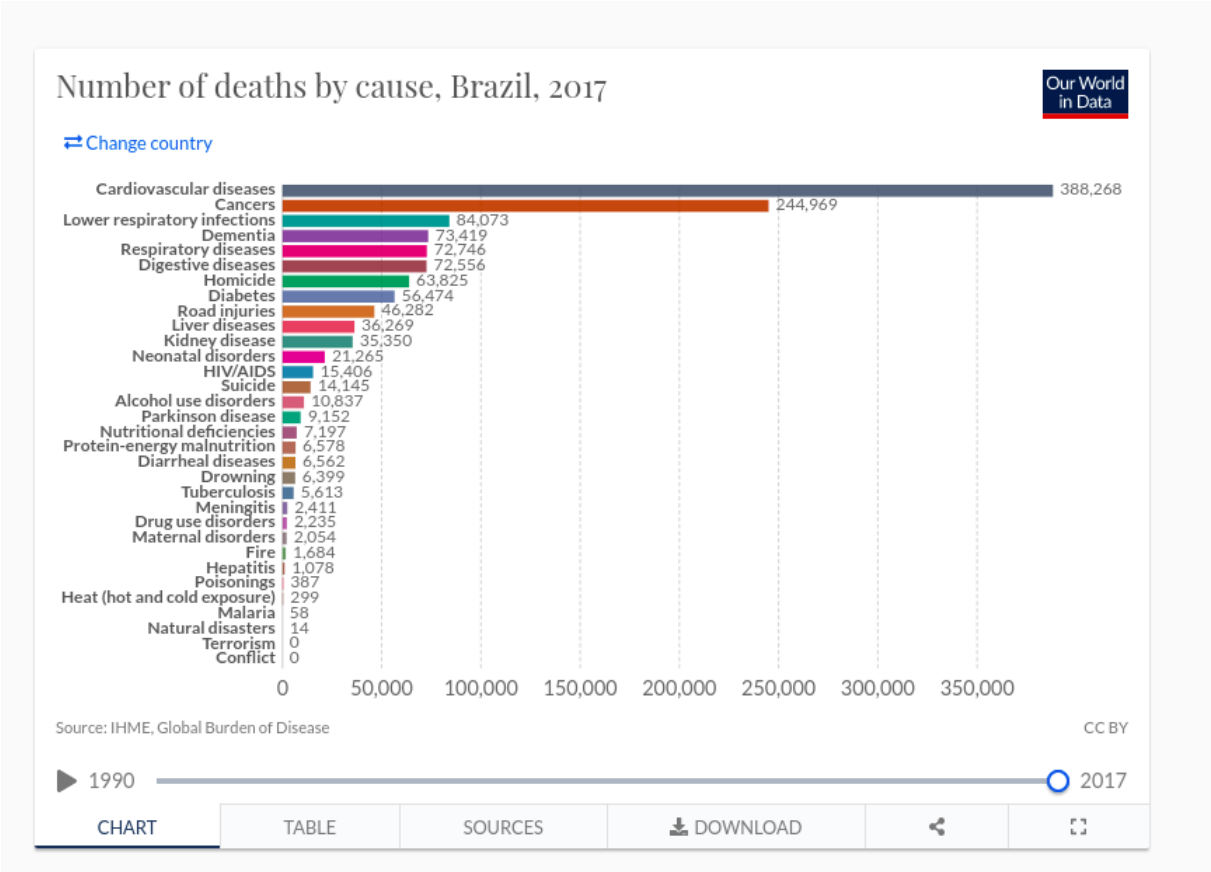
Devido à pandemia provocada pelo Sars-Cov-2, um dos causadores da SRAG, um vasto arsenal de informações sobre o comportamento da doença foi disponibilizado por diversas empresas que lidam com um número massivo de dados de mobilidade da população em geral, como: Google[1], Facebook[2] e Apple[3]. Estas empresas disponibilizam análises e notícias para orientação e ajuda nas pesquisas, tendo o Google[4] reportado, até 20 de janeiro de 2021, 2.054.218 mortes globais com uma incidência de 96.097.101 infectados. Ou seja, SRAG se tornou de ampla relevância no contexto global no ano de 2020.

No Brasil, em períodos anteriores, é possível ver esse crescimento analisando brevemente os dados históricos. Entre os anos de 2008 e 2017, as infecções do trato respiratório infe-

[1] Google: <https://about.google/>
[2] Facebook: <https://about.fb.com/>
[3] Apple Inc.: <https://www.apple.com/business/>
[4] Coronavírus (COVID-19) - Google Notícias: <https://news.google.com/covid19/map?hl=pt-BR&gl=BR&ceid=BR:pt-419>

rior (regiões dos brônquios, bronquíolos, traqueia, pulmões, e os alvéolos pulmonares) são classificadas como a terceira causa de morte no Brasil, como apresentado na Figura 1.

Figura 1 – Principais causas de mortes no Brasil no ano de 2017



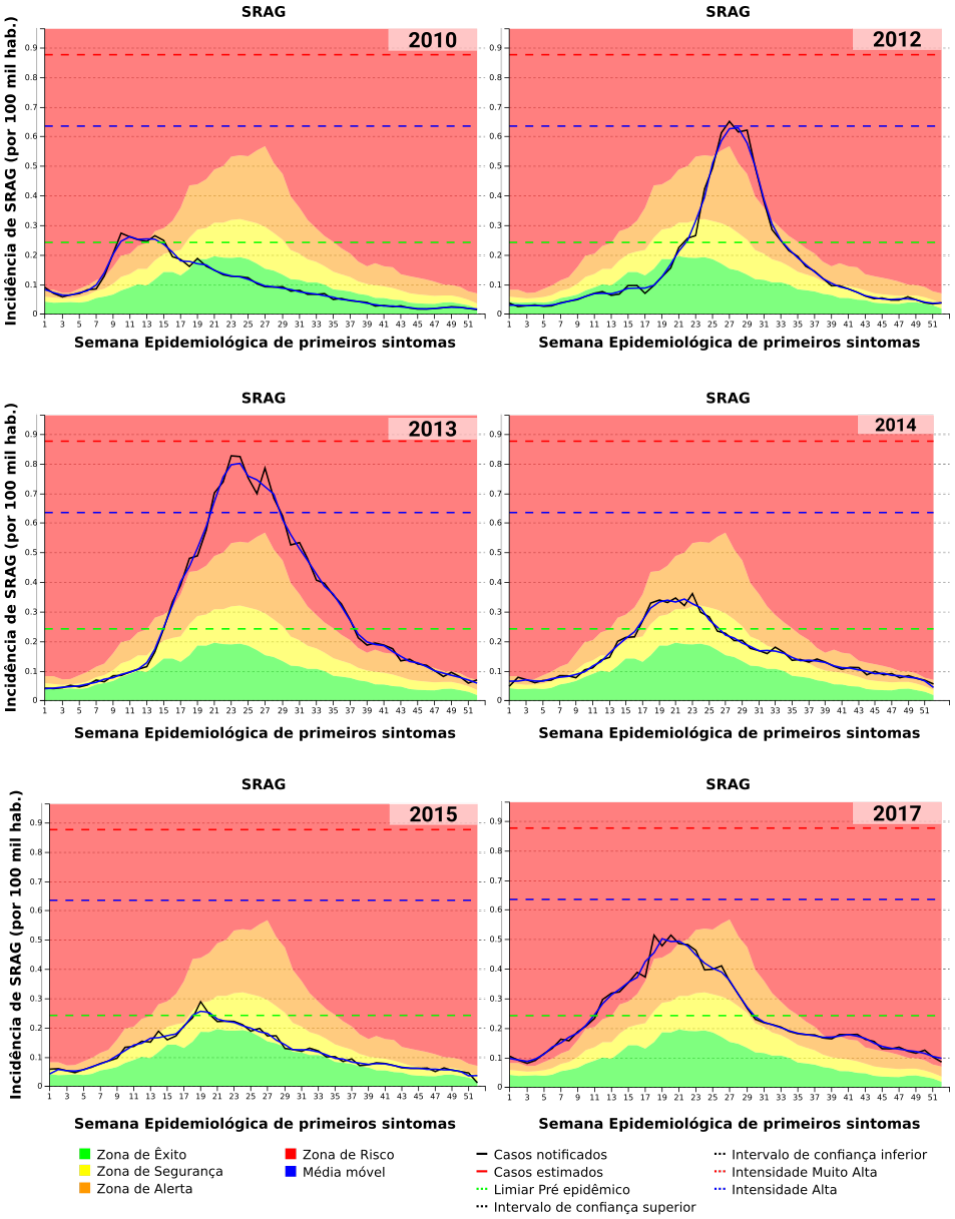
Fonte: RITCHIE; ROSER (2018)

Conforme os dados do InfroGripe (FIOCRUZ, 2020) coletados e reportados no SINAN[5](Sistema de Informação de Agravos de Notificação) em 2020, o Brasil obteve uma incidência de Síndrome Respiratória Aguda Grave (SRAG) de 202,36 a cada 100 mil habitantes. A maioria das notificações de primeiros sintomas ocorreu na 21ª semana epidemiológica.

Considerando os dados do InfoGripe[6], nenhuma das zonas delimitadoras de alerta entre os anos de 2010 a 2017 apresentaram números acima da zona de risco para caso de incidência de SRAG, como se evidencia na Figura 2. Entretanto, no ano de 2020, a partir da 10ª semana epidemiológica, a incidência de SRAG ultrapassou a zona de risco delimitada pelo tracejado em vermelho, reflexo da pandemia do Sars-Cov-2 (Covid-19), como apresentado na Figura 3.

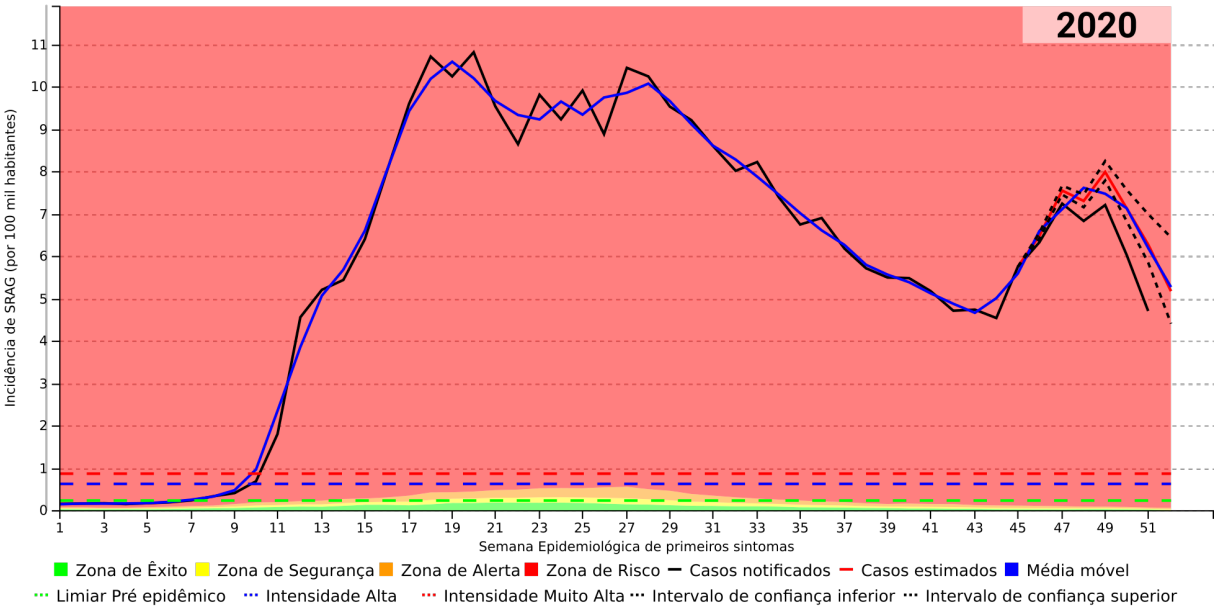
[5] SINAN: <www.saude.gov.br/sinan>=
[6] InfoGripe: <http://info.gripe.fiocruz.br>

Figura 2 – Séries temporais de síndrome respiratória aguda grave (SRAG) no país independente de apresentar febre, por semana de primeiros sintomas. Os anos de 2010, 2012, 2013, 2014, 2015 e 2017 são temporadas regulares segundo o Info Gripe.



Fonte: FIOCRUZ (2020)

Figura 3 – Séries temporais de síndrome respiratória aguda grave (SRAG) no país independente de apresentar febre, por semana de primeiros sintomas do ano de 2020.

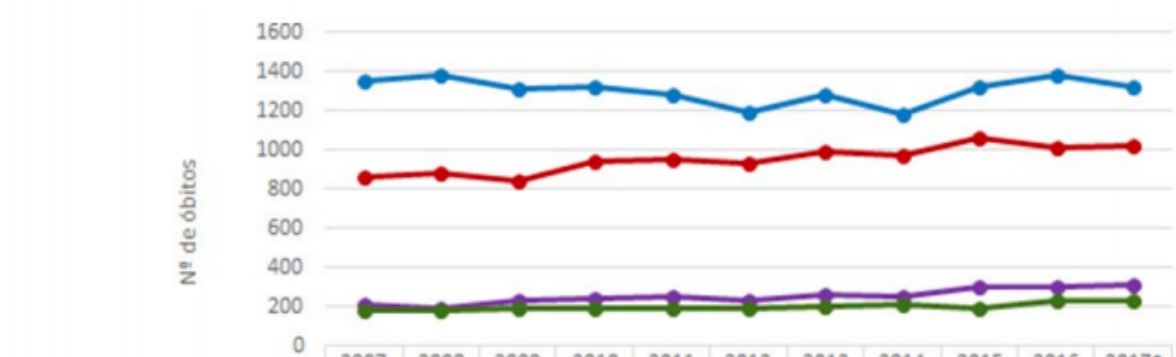


Fonte: FIOCRUZ (2020)

No nordeste brasileiro, especificamente no Recife, capital de Pernambuco, as doenças do aparelho respiratório ocasionaram 1.281.167 internações, nos anos de 2001 a 2013 (MUNICIPAL et al., 2014). Esse número é composto não apenas de infecções no aparelho respiratório, mas também de doenças crônicas associadas, como: asma, rinite alérgica e doenças pulmonares obstrutivas (MUNICIPAL; RECIFE; GERAL, 2018). Diante de tais afirmativas, podemos considerar que as doenças do aparelho respiratório, em 2009, foram consideradas as que mais causaram internações na cidade do Recife (MUNICIPAL et al., 2014).

Na Figura 4 são apresentadas as principais causas de óbitos, entre 30 e 69 anos. Podemos observar que doenças respiratórias crônicas estão relacionadas à terceira maior causa de morte, considerando doenças crônicas não transmissíveis.

Figura 4 – Distribuição dos óbitos prematuros (30 a 69 anos) pelas principais doenças crônicas não transmissíveis. Recife, 2007 a 2017



	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017*
D. do aparelho circulatório	1347	1380	1309	1318	1282	1193	1278	1184	1319	1384	1321
Câncer	866	878	840	945	953	928	992	976	1062	1013	1026
D. respiratórias crônicas	213	193	231	244	257	235	267	249	300	300	313
Diabetes	181	184	193	194	196	194	206	214	195	233	228

Fonte: MUNICIPAL; RECIFE; GERAL (2018)

Doenças do trato respiratório têm um grau considerável de relevância no âmbito da saúde pública. Os estudos de tais doenças podem ajudar na execução de medidas que possibilitem a criação de políticas públicas a fim de direcionar mais eficientemente os esforços para redução das consequências do aumento de novos infectados. SRAG ganhou um alto grau de relevância ultimamente, direcionando diversos estudos na análise e predição do comportamento da disseminação da doença (MELLO, 2020; ABBOTT et al., 2020; CMMID COVID-19 working group et al., 2020; EVOY et al., 2020). Entretanto, apesar de SRAG ganhar uma atenção maior ultimamente, ela não é a única que chama a atenção das autoridades. Diversas outras doenças, como a sífilis, hanseníase e tuberculose preocupam constantemente os órgãos e os atores dos sistemas de saúde.

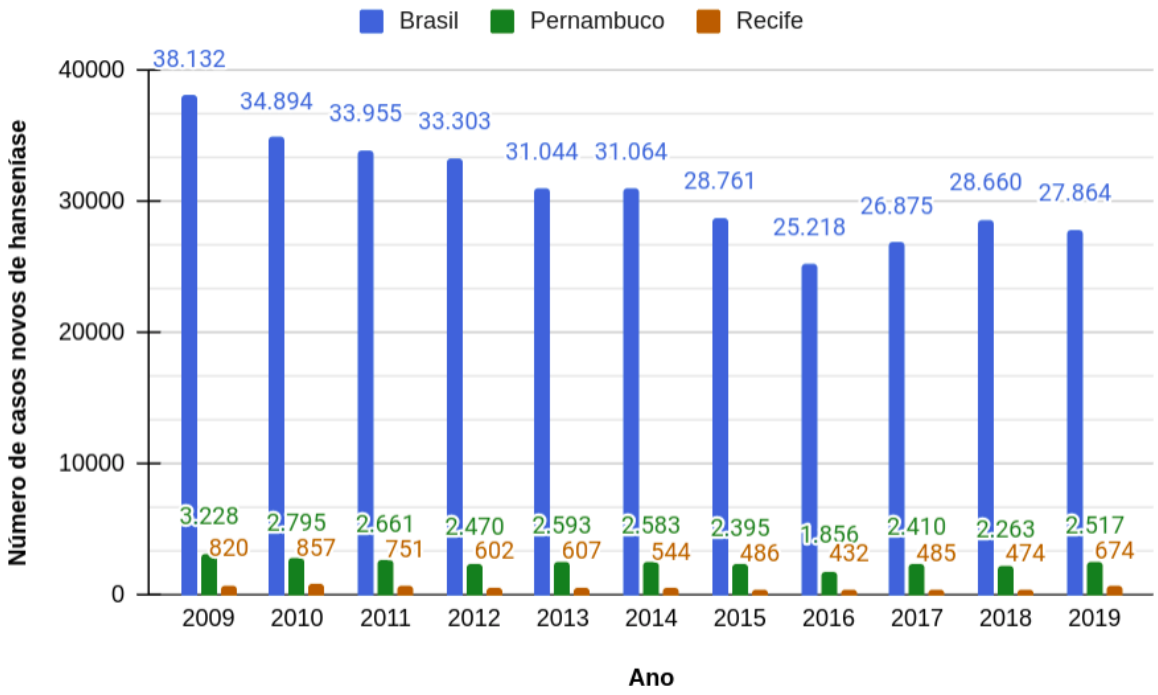
2.3 TUBERCULOSE, HANSENÍASE E SÍFILIS

Todos os anos, a secretaria de saúde de Recife lida com a ocorrência de surtos de doenças infectocontagiosas, como, por exemplo, sífilis, hanseníase e tuberculose, identificados de forma reativa, gerando custos significativos para a saúde pública (ABATH, 2013; MUNICIPAL et al., 2014). Gestores da saúde devem ficar atentos para os períodos que podem apresentar maiores ocorrências, muitas das vezes analisando, empiricamente, os dados e relatos dos gestores das unidades de saúde.

A hanseníase é um problema bem comum no Brasil, sendo um desafio para as autoridades sanitárias. No Recife, a hanseníase concentra cerca de 40% de todas as ocorrências do estado de Pernambuco. Com isso, Recife vem sendo a 2ª no ranque das capitais do Brasil que apresentam os maiores números de casos da doença em crianças de até 15 anos, segundo o levantamento de Municipal et al. (2014). Apesar da redução do número de novos casos de hanseníase na população geral entre 2007 e 2017, esta é considerada uma doença de alto risco no Recife, quando consideramos a existência das subnotificações (MUNICIPAL et al., 2014). Apesar de ter apresentado tendência de redução até 2018, nos anos de 2018 e 2019, foram diagnosticados 474 e 674 casos respectivamente.

Em 2007 a taxa de detecção foi de 58,7 casos por 100 mil habitantes e passou de 29,2 casos em 2017, sendo que em menores de 15 anos passou de 4,8 casos para 9,3 respectivamente. O percentual de cura do total de casos notificados, nesse período, foi de 79% (MUNICIPAL; RECIFE; GERAL, 2018) Figura 5.

Figura 5 – Número de novos casos de hanseníase na população geral. Brasil, Pernambuco e Recife, 2010-2019. Dados preliminares até 05/06/2020. Departamento de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis - DCCI.



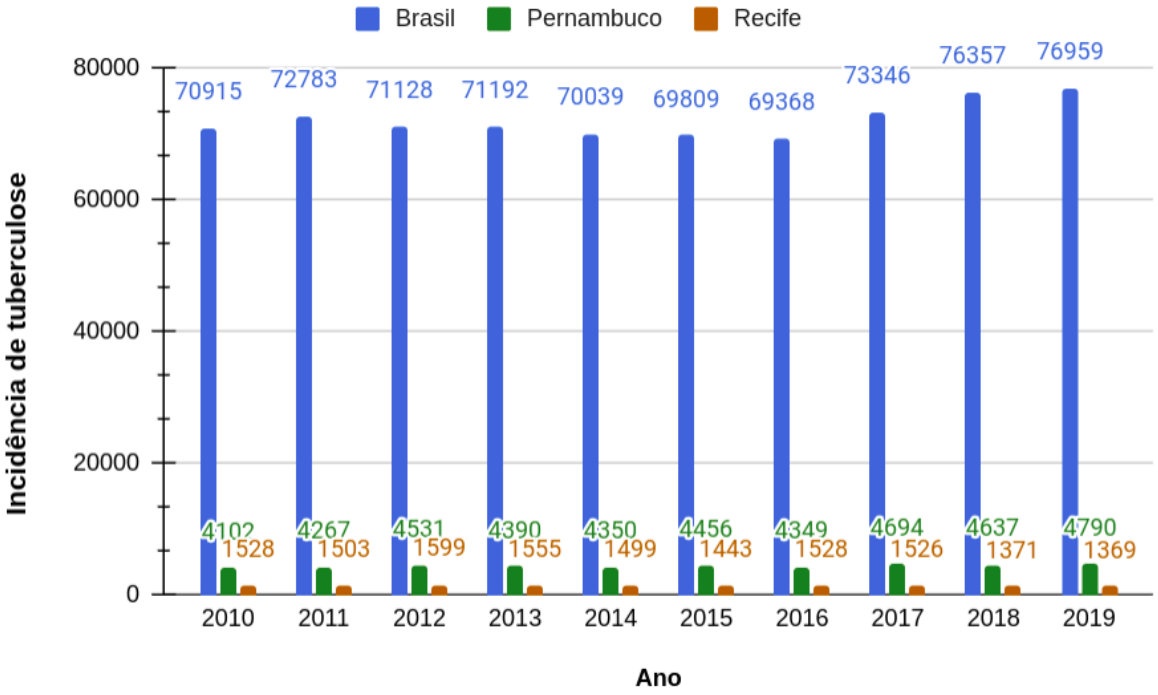
Fonte: SES/MS/SINAN/IBGE. Dados coletados em 05/2020

A tuberculose, outra doença que assola os recifenses, constitui-se também em um grave problema mundial que está intimamente ligada ao aumento da pobreza, à má distribuição de

renda e à urbanização acelerada não planejada. Segundo os dados do Municipal et al. (2014), Recife possui metas acima da média na detecção da tuberculose, sendo de 80/

Os dados fornecidos pelo Departamento das Doenças de Condições Crônicas (DCCI), através dos painéis de Indicadores Epidemiológicos[7] apresentou um aumento na incidência de todas as formas de tuberculose a partir do ano de 2016 no Brasil, ocorrendo um aumento no estado de Pernambuco e sutil declínio, quando analisada a cidade de Recife, como apresentado na Figura 6.

Figura 6 – Incidência de tuberculose por todas as formas. Brasil, Pernambuco e Recife, 2010-2019. Dados preliminares até 05/06/2020. Departamento de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis - DCCI.



Fonte: SES/MS/SINAN/IBGE. Dados coletados em 05/2020

A Sífilis é outro exemplo de doença infecciosa e contagiosa que requer atenção no país. Os dados de sífilis apresentaram um avanço alarmante entre 2010 e 2018, com um aumento de 4.050%, como apresentado na Figura 7. Conforme apresentado por Saúde (2020), os dados de 2019 podem não refletir a realidade do ano:

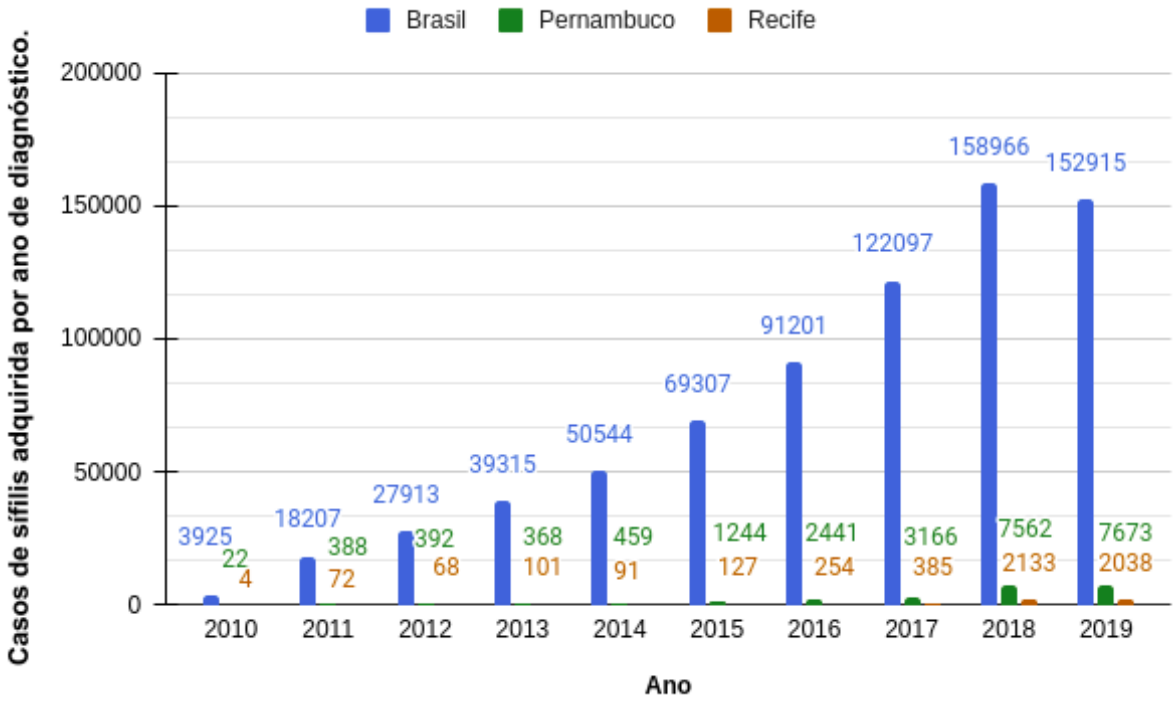
> "Parte dessa redução pode estar relacionada à demora na notificação e na alimentação das bases de dados devido à mobilização dos profissionais de saúde para

[7] Painel de Indicadores Epidemiológicos: <http://www.aids.gov.br/pt-br/gestores/painel-de-indicadores-epidemiologicos>

ações voltadas ao controle da pandemia da Covid-19 Saúde (2020)."

No estado de Pernambuco o aumento da sífilis segue a tendência da esfera nacional. De 2017 a 2018, os casos mais que dobraram no estado, conforme mostra a Figura 7. É alarmante, sendo em Recife o crescimento anualmente atingido com mais severidade. Entre 2001 e 2013, foram notificados, na cidade, 3.114 casos de sífilis congênita (MUNICIPAL et al., 2014). Os coeficientes de detecção (CD) por 1.000 nascidos vivos (NV) apresentaram níveis elevados no mesmo período, predizendo um aumento do agravamento da doença (MUNICIPAL et al., 2014).De 2017 para 2018, a cidade de Recife obteve um aumento de 554% nos casos de sífilis, ver Figura 7.

Figura 7 – Casos de sífilis adquirida por ano em diagnósticos. Brasil, Pernambuco e Recife, 2010-2019. Dados preliminares até 05/06/2020. Departamento de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis - DCCI.



Fonte: MS/SVS/Departamento de Doenças de Condições Crônicas e Infecções Sexualmente Transmissíveis. Dados até 30/06/2020

Como a Síndrome Respiratória Aguda Grave (SRAG), existem outras doenças que necessitam de constante atenção. Muitas doenças podem ser descritas como uma série temporal no decorrer dos anos, meses e dias, e essa constância ajuda na análise de seu comportamento. A análise de tais doenças, além de possibilitar a rastreabilidade e gerenciamento de recursos necessários para lidar com elas, ajuda na tomada de medidas de prevenção e controle.

2.4 PREVENÇÃO E CONTROLE

Muitas doenças infecciosas podem ser transmitidas pela ingestão de água ou alimentos contaminados, secreções respiratórias, ato sexual, sangue, exposição cutânea ao o patogênico, entre outros. As doenças infectocontagiosas são aquelas transmitidas de pessoa para pessoa (SAÚDE, 2010). O fator da prevenção de muitas doenças infecciosas se baseia no controle da transmissão. Quando a contenção inicial do grupo infectado é possível, a utilização de recursos, que até podem ser limitados no início, será mais eficiente (RILEY, 2007). Por exemplo, a sífilis pode ser controlada por tratamento medicamentoso, sendo que a dosagem e o tempo do tratamento podem variar conforme a fase da doença (AVELLEIRA; BOTTINO, 2006). O governo brasileiro anualmente efetua investimento em campanhas para conscientização sobre Tuberculose [8, 9,10], visando reduzir as ocorrências. Por exemplo, a tuberculose, segundo Arruda (2014), possui um custo de tratamento para o sistema público de cerca de R$ 293,91 em 2014 e de R$ 3.119,40 por caso tratado na família, chegando a comprometer 48% da renda familiar. Com relação aos casos de hanseníase, o Brasil se dispôs ao compromisso de reduzir para 10 casos por 100 mil habitantes, com a mortalidade de 1 a cada 100 mil (MUNICIPAL; RECIFE; GERAL, 2018). Bem como, a meta de 1 caso por 1.000 bebês nascidos vivos, atualmente distante de alcançar (MUNICIPAL; RECIFE; GERAL, 2018). Considerando que países e regiões pobres possuem mais dificuldades em lidar com tais doenças, o índice de mortalidade e complicações aumentam.

Países com Índice de Desenvolvimento Humano (IDH) mais baixo possuem maiores índices de mortalidade por doenças transmissíveis do que as que não são transmissíveis. Mesmo com a queda entre os anos 2000 e 2019, seis doenças contaminantes estavam entre as 10 causas de mortes mais comuns (HEALTH, 2020). Os esforços do Sistema de Informação de Agravos de Notificação (SINAN), baseados nos modelos descritos no trabalho de Bastos et al. (2019), visam analisar os dados coletados através de modelos matemáticos, para auxiliar no entendimento e predição da evolução da incidência de SRAG no Brasil.

O painel de indicadores também visa analisar dados de outras doenças, tais como: Aids, Sífilis, Hepatites, monitoramento clínico de HIV, Tuberculose e Hanseníase. Além da análise

[8] Campanha Nacional de Luta Contra a Tuberculose - 2018 - <http://www.aids.gov.br/pt-br/campanha/campanha-nacional-de-luta-contra-tuberculose-2018>

[9] Campanha Nacional de Luta Contra a Tuberculose - 2019 - <http://www.aids.gov.br/pt-br/campanha/campanha-nacional-de-luta-contra-tuberculose-2019>

[10] Campanha Nacional de Luta Contra a Tuberculose - 2020 - <http://www.aids.gov.br/pt-br/campanha/campanha-nacional-de-luta-contra-tuberculose>

da utilização de insumos, tais como distribuição de medicamentos a Gestantes Vivendo com HIV (GVHIV), entre outros. As análises visam auxiliar gestores, profissionais da saúde e até a população em geral (i.e, pela divulgação em escolas, universidades e disseminação da informação através de meios digitais), na tomada de decisão em diferentes níveis do gerenciamento e regiões do país. Por exemplo, pais e universidades podem orientar os jovens sobre como está a dinâmica de sífilis atualmente, visando uma maior atenção para métodos de proteção e prevenção.

Entretanto, a qualidade de cada painel apresenta variação, uma vez que são coletados de múltiplas fontes disponíveis e posteriormente analisados e apresentados de forma visual. Tais fatores podem implicar em: 1 - imprecisão na coleta dos dados para análise; 2 - resposta reativa a partir da ocorrência dos casos em cada região em um passado distante.

O objetivo da vigilância epidemiológica é reduzir a morbimortalidade, provendo uma implantação de medidas de prevenção e cuidados adequados à população (SAÚDE, 2010). Muitas das análises e identificações de doenças ocorrem por meio do quadro clínico. Tais quadros clínicos podem ser classificados conforme os sintomas e condições do paciente. A Classificação Internacional de Doenças (CID) é uma das formas de normatizar e internacionalizar as condições do paciente, bem como a Classificação Internacional de Atenção Primária (CIAP) (WONCA, 2005). Devido ao advento da pandemia de Covid-19 no início de 2020, muitas das classificações tinham as atenções voltadas àqueles pacientes classificados com sintomas e condições que poderiam estar relacionadas à pandemia em curso, possibilitando uma maior eficácia na identificação dos novos casos.

Na cidade do Recife, por exemplo, o aplicativo para autotriagem, nomeado Atende em Casa[11], utiliza classificação de doenças baseado no CID-10, a décima revisão da classificação. O Atende em Casa permite que a população efetue uma autoavaliação sobre seus sintomas de Covid-19 ao responder algumas perguntas objetivas, evitando assim o deslocamento prematuro a hospitais, ajudando na atenuação e prevenção de novos casos, através da orientação, bem como um melhor controle da pandemia da Covid-19. Outra aplicação é o Atende APS[12] que apresenta características similares ao Atende em Casa, porém com uma triagem em três etapas executadas por profissionais e técnicos da saúde. Com a coleta e análise dos dados provenientes de tais aplicações, os gestores podem entender a situação e a demanda atual de cada doença

[11] Atende em Casa: <http://www2.recife.pe.gov.br/noticias/26/03/2020/pcr-e-governo-lancam-aplicativo-web-para-populacao-ser-orientada-distancia-por>

[12] Atende APS manual: <https://drive.google.com/file/d/1eImS6DRgEcAsfCwuJsh-_lDzakIWrPMV/view>

para desenvolver medidas de prevenção e controle mais eficientes.

Entretanto, embora uma análise das condições atuais seja bastante relevante para entender a dinâmica das doenças, temos que considerar que as medidas por parte dos órgãos de saúde se limitam, muitas vezes, a ações reativas ao identificá-las. Os dados coletados podem ser transcritos em séries temporais, particularmente onde o tempo é correlacionado às ocorrências de tais patologias. Dessa forma, se obtém os subsídios para aplicação de técnicas que permitam trabalhar com os dados classificados.

2.5 SÉRIES TEMPORAIS

Uma série temporal nada mais é do que uma sequência de observações relacionadas a uma variável ao longo de uma faixa temporal (BOX et al., 2015). Devido ao grande número de fontes e dados disponibilizados em forma de séries temporais, em conjunto com atuais avanços sensores e poder computacional, as séries possibilitam uma nova abordagem de como sistemas complexos são monitorados e controlados (CHENG et al., 2015). Sistemas e programas de grande porte têm a possibilidade de analisar melhor os custos, desempenho, integridade e, entre outras vantagens, seus dados. Esses dados podem, por exemplo, ser representados de diversas maneiras, tais como: as ocorrências de acidentes em uma determinada rodovia, o número de voos sendo feitos de uma região para outra, quantidade diária de dias ensolarados, o fluxo diário de mercadoria, o desgaste de um catalisador no decorrer do tempo, dentre outros. Dessa forma, a apresentação dos dados em séries temporais podem proporcionar a visualização de padrões e relações dos dados com fatores que possam ser relevantes para a análise dessas informações.

De acordo com Box et al. (2015) existem cinco áreas de aplicação para séries temporais:

1. Predição para o futuro, tendo como base o presente e o passado.
2. Determinação da função de transferência de um sistema propício à inércia.
3. O uso da entrada das variáveis da função de transferência para identificar eventos anômalos na série.
4. A análise das interrelações entre as variáveis de séries temporais correlacionadas e a identificação dos modelos multivariados apropriados.
5. Esquematização dos potenciais desvios da saída do sistema alvo visado.

A aprendizagem de máquina e a aprendizagem profunda, dentre outras aplicações, podem utilizar séries temporais para tentar predizer o seu comportamento futuro, considerando os dados disponíveis até então. Essas aplicações podem auxiliar na economia, engenharia, ciências naturais, negócios, ciências sociais entre outras (BOX et al., 2015).

2.5.1 Estudos Epidemiológicos Com Uso De Séries Temporais

Nos últimos 50 anos, os estudos da dinâmica das doenças têm se tornado um campo interdisciplinar, abrangendo diversas áreas do conhecimento relacionadas à matemática, biologia, epidemiologia, saúde pública, entre outras (HEESTERBEEK et al., 2015). O desafio prático recai na definição da melhor forma de estabelecer uma apropriada coleta de dados e a definição dos modelos matemáticos a serem seguidos (HEESTERBEEK et al., 2015). O comportamento da sociedade ao lidar com as doenças, especialmente no início do surto, podem aumentar a complexidade da definição do modelo, visto que, diversos acontecimentos podem não ser mapeados ou completamente entendidos. Outro fator é a influência das mudanças comportamentais do grupo populacional e as respostas do sistema de saúde, tornando-se extremamente complexo fazer predições a longo prazo (HEESTERBEEK et al., 2015). Diversos estudos vêm considerando abordagens de inteligência artificial para analisar e predizer vários tipos de surtos: dengue, por exemplo, no trabalho de Xu et al. (2020), Martinez, Silva e Fabbro (2011); tuberculose (WANG et al., 2018; LIU et al., 2019; MOOSAZADEH; KHANJANI; BAHRAMPOUR, 2013; WILLIS et al., 2012; AZEEZ et al., 2016); influenza (XU et al., 2017; VOLKOVA et al., 2017; LAMPOS et al., 2015; HE et al., 2017; BROOKS et al., 2018) e doenças relacionadas à mão, pé e boca (SONG et al., 2015; ZHAN et al., 2019; TIAN; WANG; LUO, 2019) de acordo com Olsavszky et al. (2020).

O Quadro 1 a seguir contém alguns trabalhos que utilizaram modelos de aprendizagem de máquina visando analisar o comportamento de algumas doenças e dos próprios algoritmos/-técnicas utilizadas. Muitos dos modelos nos artigos são calibrados para atender ao problema específico, visto que os modelos são sensíveis à qualidade dos dados. No decorrer das pesquisas os modelos podem sofrer variações na sua estrutura padrão, visto que é comum a personalização dos mesmos e até mesmo uma abordagem híbrida com diferentes técnicas. Assim, a predição de algumas doenças podem ter desempenhos melhor ou pior a depender dos ajustes feitos.

Atualmente a Classificação Internacional de Doenças (CID) e a Classificação Internacional de Assistência Primária (CIAP) podem ser utilizadas para padronizar e facilitar a análise

Quadro 1 – Trabalhos visando analisar algumas das principais doenças estudadas pela comunidade científica e respectivas técnicas utilizadas.

Referência	Doenças	Técnicas/Algoritmo
Xu et al. (2020)	Dengue	LSTM
Martinez, Silva e Fabbro (2011)	Dengue	SARIMA
extWang et al. (2018)	Tuberculose	SARIMA e SARIMA-GRNN
Singh et al. (2018)	Tuberculose	LSTM
Liu et al. (2019)	Tuberculose	ARIMA e BPNN
Moosazadeh, Khanjani e Bahrampour (2013)	Tuberculose	SARIMA
Azeez et al. (2016)	Tuberculose	SARIMA e NNAR
Xu et al. (2017)	Gripe	ARIMA, GLM, LASSO e aprendizagem profunda
Volkova et al. (2017)	Gripe	LSTM
He et al. (2017)	Gripe	ARDL e GRNN
He e Tao (2018)	Gripe	ARIMA
Tian, Wang e Luo (2019)	Mão, pé e boca	SARIMA
Song et al. (2015)	Mão, pé e boca	SARIMA
Gu et al. (2019)	Mão, pé e boca	LSTM

Fonte: Elaborada pelo autor (2020)

global das doenças. No Recife, capital do estado de Pernambuco, o sistema de saúde utiliza classificação CID, versão 10, como base. Uma vez padronizado, é possível rastrear, com a ajuda dos dados fornecidos pelo próprio sistema de saúde, a temporalidade da quantidade de ocorrências de determinado CID/CIAP em determinada localidade. Aplicativos como Atende APS[13], que está presente nas unidades de Atenção Primária à Saúde (APS). tem o objetivo de identificar e gerenciar as ocorrências de usuários sintomáticos respiratórios (síndrome gripal), possibilitando que esses dados possam ser coletados diretamente das unidades de saúde. A eficiência e a utilização de tais sistemas, bem como a disponibilização de seus dados para estudo e análise, proporcionam grandes contribuições para os estudos do círculo de incidência das doenças, além de prover vantagens operacionais nestas unidades onde estão implantados.

Técnicas de IA vêm sendo utilizadas para análise de dados da saúde, com o intuito de ajuda no combate e melhorar os sistemas de saúde através de uma maior eficiência (OLSAVSZKY et al., 2020), visto que modelos de predição de séries temporais (dados observados ao longo do tempo) podem auxiliar, caso possuam um nível de precisão alto ou aceitável, a quantidade de ocorrências de determinada patologia no futuro.

[13] Atende APS manual: <https://drive.google.com/file/d/1eImS6DRgEcAsfCwuJsh-_IDzakIWrPMV>

2.6 CONSIDERAÇÕES DO CAPÍTULO

Neste trabalho, foram selecionados três algoritmos de predição de séries temporais: dois desses algoritmos (ARIMA e LSTM), estão fortemente representados nas pesquisas relacionadas a doenças infecciosas (ver Seção 2.5.1). Além desses algoritmos, selecionamos o *Prophet* devido à sua capacidade de considerar os impactos significativos na evolução/redução do número de casos em feriados/finais de semana. Entre os algoritmos, *Prophet* é o mais recentemente desenvolvido.

Nesta seção, foram discutidas algumas doenças e seus impactos nas políticas públicas. Considerando sua incidência e grau de contaminação, os estudos de séries temporais relacionados à epidemiologia vêm ganhando cada vez mais notoriedade. Diversos estudos vêm contribuindo para a aplicação da prevenção e controle epidemiológico, o que definimos como de fundamental importância além do estudo e a análise da aplicação e utilização de técnicas desenvolvidas em pesquisa, em problemas reais e de grande impacto na sociedade.

3 TRABALHOS RELACIONADOS

Modelos estatísticos são usados, em saúde pública, para fornecer respostas a questões vitais, tais como: "Qual será a dimensão do surto", "Como se desenvolverá ao longo do tempo" e talvez o mais importante, "Como podemos controlar isso?". Outras questões também devem ser respondidas, como: "Qual será o novo surto?", "Será uma nova epidemia ou uma endemia".

Alguns estudos tentam responder a essas questões, no entanto, alguns resultados e a falta de conhecimento dos profissionais de saúde, podem levar a percepções negativas dos modelos preditivos criados, o que faz com que, às vezes, sejam esquecidos em relação a outros métodos mais tradicionais. No trabalho de Olsavszky et al. (2020) foi analisado o CID, versão 10, da base de dados nacional da Romênia, de pacientes hospitalizados por diversas doenças, entre 2008 a 2018. Foi utilizado o AutoTS[1], para um rápido experimento de aprendizagem de máquina, o qual obteve dados que ajudam a compreender os benefícios da utilização de tais métodos, para ajudar na tomada de decisão. O AutoTS utilizado simplifica o trabalho, uma vez que executa, em paralelo, buscas heurísticas para identificar o melhor ou melhores modelos, baseados nos dados e objetivos da predição. Segundo os autores, o AutoTS desempenhou bem as predições, mesmo considerando o viés da coleta dos dados. Consideram que a dinâmica prevista da contagem dos casos pode ajudar os agentes a alocar recursos ou conscientizar a operacionalização de medidas de suporte para lidar com as doenças.

Martinez, Silva e Fabbro (2011) efetuaram a análise da evolução dos casos de dengue na cidade de Campinas, no estado de São Paulo. No artigo são apresentados os resultados utilizando o modelo de Média Móvel Sazonal Autoregressiva Integrada (SARIMA, do inglês *Autoregressive Integrated Moving Average*), utilizando as ocorrências de casos de 1998 a 2008. Os autores afirmam que os resultados das predições das ocorrências para 1, 2 e até 12 meses do ano subsequente, tiveram resultados muito próximos dos observados. Assim, o SARIMA produziu boas predições das ocorrências de casos de dengue. Os autores citam que o modelo pode não ser confiável para predizer um ano epidêmico inteiro, devido à falta de imunização da população, uma vez que muitas pessoas podem entrar em contato pela primeira vez com a doença. Entretanto, vemos que modelos estatísticos de séries temporais podem direcionar um melhor entendimento da dinâmica de doenças.

Com o advento da pandemia da COVID-19, diversos trabalhos foram desenvolvidos com a

[1] AutoTS - https://github.com/winedarksea/AutoTS

finalidade de analisar e entender o comportamento de novos casos (SHASTRI et al., 2020; BATTINENI; CHINTALAPUDI; AMENTA, 2020; HU et al., 2020; XIE, 2020), como no caso de (SHAHID; ZAMEER; MUNEEB, 2020), que compararam o desempenho de técnicas de inteligência artificial, tais como: ARIMA, GRU, SVR, LSTM e o bidimensional LSTM, ao predizer a quantidade de ocorrências futuras com dados de diferentes países. Neste trabalho foi mostrado que os resultados apresentados pelo LSTM (do inglês: *Long Short Term Memory*), GRU (do inglês: *Gated Recurrent Network*) e Bi-LSTM (do inglês: *Bidirectional Long Short Term Memory*), apresentaram maior robustez. Como resultado, o autor cita que os modelos podem ser utilizados na saúde para um melhor planejamento e gerenciamento.

Outros modelos também vêm sendo estudados em decorrência da pandemia do COVID19. É o caso do trabalho de Battineni, Chintalapudi e Amenta (2020) o algoritmo Prophet e Tandon et al. (2020) o ARIMA. Nesses trabalhos são analisados os desempenhos dos modelos de aprendizagem de máquina na predição de novos casos. Como afirmado nos artigos, a utilização dos modelos visa melhorar significativamente as estimativas de novos casos de pessoas infectadas, possibilitando melhores formas de planejar e desenvolver políticas na saúde para conter, de forma proativa, o avanço da pandemia.

Com isso, é possível visualizar que a utilização de inteligência artificial, para um melhor gerenciamento e entendimento da dinâmica das doenças, não é um assunto novo. Diversos trabalhos apresentam propostas de utilização de modelos computacionais para prover suporte no planejamento e gerenciamento de tais doenças, por parte dos órgãos e gestores da saúde. De fato, é fundamental que tais análises possam ser utilizadas na prática, para possibilitarem um melhor gerenciamento do sistema de saúde

4 MÉTODO

Com o intuito de melhorar as medidas preventivas e as ações futuras, em decorrência de doenças com tendência de propagação ao nível de um surto ou epidemia, foram utilizadas técnicas de predição de séries temporais, tais como: ARIMA (BOX et al., 2015), LSTM (HOCHREITER; SCHMIDHUBER, 1997) e *Prophet* (TAYLOR; LETHAM, 2017), para analisar a evolução dos números de casos de algumas doenças infectocontagiosas. Tais técnicas podem ser utilizadas para ajudar no acompanhamento periódico das doenças, por meio da análise dos CIDs/CIAPs (MARSDEN-HAUG et al., 2007) e calibração automática dos modelos citados, além de prover suporte para futuros estudos com foco em outras doenças de características similares.

4.1 DELINEAMENTO DO ESTUDO

No âmbito deste trabalho foram utilizados os registros diários de ocorrência de 11 CIDs/-CIAPs, fornecidos pela secretaria municipal de saúde da cidade do Recife, coletados em 24 unidades de saúde distribuídas em 96 bairros, utilizando o sistema Atende APS (Atenção Primária de Saúde). O Atende APS é um sistema desenvolvido para triagem e atendimento da população que apresentem sintomas relacionados à síndrome respiratória aguda grave (SRAG). Dessa forma, os CIDs/ CIAPs categorizados na ferramenta estão associados a doenças respiratórias contagiosas, totalizando 50.657 observações totais, entre 26 de abril de 2020 a 7 de março de 2021.

É fundamental citar que os CIDs/CIAPs são atribuídos aos pacientes por meio das análises técnicas e clínicas. O protocolo de análise foi criado pelo departamento de saúde local, com base no *Manchester Triage System* (MTS) (MACKWAY-JONES; MARSDEN; WINDLE, 2013), sendo amplamente utilizado por unidades médicas, com recursos oriundos do *Royal College of Physicians*[1], e protocolo de combate à Covid-19, devido ao atual cenário pandêmico. As análises técnicas realizadas pelos profissionais da saúde nos pacientes, consistem na aferição dos seguintes sinais vitais: (1) pressão arterial, (2) frequência respiratória (ipm), (3) frequência cardíaca (bpm), (4) saturação de oxigênio (%), (5) temperatura corporal (ºC), (6) glicemia capilar, em caso de diabéticos (mg/dL). Além das análises técnicas, enfermeiros e médicos seguem um protocolo específico para padronizar as avaliações e resultados clínicos. Após reali-

[1] *Royal College of Physicians*: https://www.rcplondon.ac.uk/projects/outputs/national-early-warning-score-news-2

zados todos os procedimentos, o sistema classifica os pacientes indicando o nível de gravidade dos sintomas, apresentados pelas cores azul, verde, amarelo e vermelho. Com isso, após a consulta médica, o paciente também recebe em seu prontuário médico o diagnóstico clínico, com a indicação de um CIDs/CIAPs apresentado no Quadro 2.

Quadro 2 – CIDS e CAPS classificados através do Atende APS - Considerando as ocorrências da cidade do Recife

CID/CIAP	Ocorrências	Descrição
CID U 07.2	16.247	Diagnóstico de Covid-19 confirmado por exames Laboratoriais
CID J11	6.214	Influenza [Gripe] devida a vírus não identificado
CIAP R80	6.031	Gripe
CIAP R74	4.893	Infecção aguda do aparelho respiratório superior (IVAS)
CIAP R83	3.694	Outra infecção respiratória
CID U 07.1	1.540	Diagnóstico de covid-19 confirmado por exames Laboratoriais
CID B 34.2	1.164	Infecção por coronavírus de localização não especificada
CID J11.0	529	Influenza [gripe] com pneumonia, devida a vírus não identificado
CID J 80	93	Síndrome do desconforto respiratório do adulto
Outros	63	Não identificado
CID U 04	53	Síndrome respiratória aguda grave (SRAG)
CID U 04.9	49	Síndrome respiratória aguda grave (SRAG), não especificada

Fonte: Elaborada pelo autor (2021)

4.1.1 Casuística

Os dois primeiros casos confirmados de Sars-CoV-2 causados pelo Covid-19, no estado de Pernambuco, ocorreram na zona sul do Recife no início de março de 2020 (MEIRELES, 2020). Com isso, as autoridades em saúde tiveram que agir rápido para tentar conter a sua propagação, a fim de resguardar o sistema de saúde. Para tal, os órgãos da saúde ampliaram a assistência à população criando as unidades especializadas e sistemas de suporte a essas unidades.

A análise da quantidade de ocorrências diárias de CIDs/CIAPs coletadas por meio do sistema Atende APS, nas unidades especializadas, provê suporte à criação de um sistema capaz de identificar o crescimento de números de casos de doenças infectocontagiosas e, assim, prever o seu impacto nos bairros da cidade do Recife. Com isso, a coleta diária desses dados, nas unidades de saúde, pode influenciar positivamente na criação de políticas públicas proativas.

Para termos uma maior compreensão do cenário atual, o município do Recife possui uma área territorial de 218,843 km^2 (censo 2019), com uma população estimada de 1.653.461 pessoas (censo 2020), com densidade demográfica de 7.039,64 hab/km^2 (censo 2010) e o IDH de 0,777, tendo este tido um aumento gradual desde 1991, quando se encontrava no patamar de 0,550[2]. Segundo a classificação climática de Köppen (1936), o clima do município é Am, ou seja, tropical úmido com chuvas de outono-inverno, apresentando suscetibilidade à SRAG, como acontece em muitas cidades brasileiras. A cidade possui um aeroporto internacional e diversos polos universitários que recebem estudantes de todas as regiões do país, assistida pelo SUS com 274 unidades de saúde registradas em 2009[2]. Com esses dados sociais e demográficos diversificados, fica claro que o gerenciamento público necessita de mecanismos automatizados que reduzam os esforços e recursos extras para contenção de novos surtos, epidemias ou, até mesmo, reflexos de uma nova pandemia. Para tal objetivo, tanto modelos estatísticos quanto computacionais podem ser utilizados, como será mostrado na próxima seção.

4.2 MODELAGEM ESTATÍSTICA

A modelagem estatística possibilita a interpretação dos dados, considerando fatores ainda não observados na análise inicial, período este, onde muitas das informações não são reportadas ou são reportadas equivocadamente (BASTOS et al., 2019). Os dados epidemiológicos possuem essas características, visto que tanto a qualidade quanto a quantidade das informações podem variar significativamente.

Doenças infecciosas estão intrinsecamente relacionadas à média dos indivíduos infectados, ou seja, a capacidade de infectar outros indivíduos é o que pode torná-las relevantes no ambiente. Apesar da imprevisibilidade em detectar se um indivíduo está ou não infectado, a dinâmica populacional segue as expectativas matemáticas, visto que são estudadas por longos períodos. Os ruídos (falhas nos dados coletados) entre as detecções individuais, tendem a se

[2] IBGE - Panorama de Recife: <https://cidades.ibge.gov.br/brasil/pe/recife/panorama>

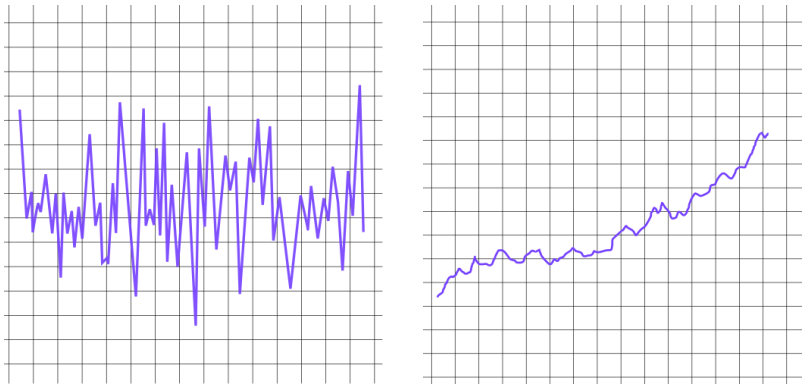
anular à medida que os dados aumentam, tendo em vista a lei dos grandes números (GRASSLY; FRASER, 2008). Em outras palavras, quanto maior a identificação correta dos infectados, mais aumenta a irrelevância dos erros de detecção e melhora a análise estatística.

Na epidemiologia, os modelos probabilísticos são estudados (HELD; MEYER; BRACHER, 2017; HE; TAO, 2018) com o intuito de verificar a influência de doenças e a sua provável tendência de projeção futura. Tais estudos permitem a melhoria e o desenvolvimento de novos modelos, provendo contribuições práticas à sociedade que possibilitam a compreensão do comportamento das doenças infecciosas Farrington et al. (1996), Held, Höhle e Hofmann (2005), Stoner, Economou e Silva (2019). Vale ressaltar que, os modelos estatísticos possibilitam a compreensão, resultados e métodos utilizados na aprendizagem da dinâmica dos dados. Entretanto, a estatística infere dados da população a partir de amostras, mas por outro lado, a aprendizagem de máquina generaliza um padrão de comportamento dos dados. Ambas as abordagens podem utilizar as ocorrências de uma determinada doença com o intuito de fazer projeções relacionadas ao crescimento do número de casos. A seguir, são apresentados alguns conceitos que levam à compreensão dos modelos que serão utilizados neste estudo.

4.2.1 Processos estocásticos

Encontramos na literatura que uma análise estocástica, tendo uma distribuição de probabilidades aleatórias, é mais realística do que modelos determinísticos, considerando o estudo de modelos epidemiológicos (LIN; JIANG; LIU, 2015; ZHANG; ZHOU, 2019; LIU; JIANG, 2017). Um processo estocástico é uma família de variáveis aleatórias inseridas no mesmo espaço de probabilidade. O modelo Autorregressivo Integrado de Médias Móveis (ARIMA) é um dos mais populares modelos de séries temporais estocásticas, onde se supõe que a série é considerada linear e segue uma distribuição estatística conhecida, como a distribuição normal (ADHIKARI; AGRAWAL, 2013). É provável que nenhum fenômeno seja totalmente determinístico, dado que variáveis externas podem influenciar o fenômeno estudado. Ao invés de calcular um modelo preciso, é possível derivar um modelo considerando um limiar específico de confiança (BOX et al., 2015). Considerando séries temporais compostas de dados epidemiológicos, a projeção da série é geralmente não determinística, devido aos fatores do ambiente que, sejam eles sociais, ambientais e econômicos, influenciam a quantidade de ocorrências de novos casos em um determinado espaço de tempo.

Figura 8 – Exemplos de série estacionário e série não estacionária



(a) Série estacionária (b) Série não estacionária

Fonte: Elaborada pelo autor (2021)

4.2.2 Estacionariedade

Quando as características estatísticas de uma série possuem valores constantes temporalmente como média, variância, autocorrelação, entre outras, é correto afirmar que a série temporal é estacionária, ou seja, todas as características probabilísticas do comportamento não são alteradas no tempo (BOX et al., 2015). Por outro lado, quando há mudanças na distribuição estatística em decorrência do tempo, a série é denominada como não estacionária (CAO; GU, 2002). Em alguns casos, a não estacionariedade de uma série pode ser observada por meio de representações gráficas que sugerem o comportamento de "*trend*" (tendência) como mostrado na Figura 8.

Entretanto, nem sempre apenas com a análise visual do comportamento das séries é possível identificar a sua distribuição, como na Figura 8. Nesses casos, alguns testes não paramétricos podem ser utilizados, como o teste de *Dikey-Fuller* (DF) ou o de *Kwiatkowski*, *Phillips*, *Schmidt and Shin test* (KPSS) (SYCZEWSKA, 1997; KOOP; DIJK, 2000). Logo, os testes possuem metodologias diferentes para analisar qual o tipo de distribuição da série. O teste de Dickey-Fuller, por exemplo, pode ser representado pela seguinte equação:

$$y_t = py_{t-1} + e_t, \tag{4.1}$$

onde e_t é o termo que representa o erro da estacionariedade e a hipótese nula é p=1. Já no teste KPSS, a estacionariedade é verificada por meio da representação do espaço:

$$y_t = \tau_t + e_t, \tag{4.2}$$

$$\tau_t = \tau_{t-1} + u_t, \tag{4.3}$$

onde u_t representa o ruído branco com a variância σ_u^e e u_t, e e_s são independentes. Sendo o valor nulo da hipótese, considera a série estacionária por $\sigma_u^e = 0$.

A análise da estacionariedade visa identificar se uma projeção é invariante sobre a translação. Isso significa que serve para identificar se a intensidade da série não varia na região do espaço d-dimensional na primeira ordem e se os eventos estão associados somente em decorrência da distância entre eles (FERREIRA, 2017). Métodos de previsão estacionarizam os dados da série temporal para efetuar as previsões e, em seguida, retornam para o estado original e exibem os resultados (STOCK; WATSON, 2018). No mundo real, é muito difícil encontrar uma série temporal que não sofra efeito de fatores externos que influenciem os valores da série por meio do tempo. A análise e existência da estacionariedade é importante para muitas das técnicas que dependem da condição estacionária para execução dos algoritmos.

4.2.3 Autocorrelação

Na utilização de modelos para análise de séries temporais, é necessário considerar, mesmo com métodos estatísticos ou com redes neurais artificiais, a relação entre a observação atual e as observações anteriores. A função de autocorrelação (FAC, ou do inglês ACF - *Auto-Correlation Function*) é utilizada para identificar a relação dos valores das séries em relação aos seus antecedentes (MAKRIDAKIS; WHEELWRIGHT; HYNDMAN, 1997). A FAC ajuda a identificar o quanto de informação de um valor da série contém sobre o próximo valor ou, se existe pouca relação do valor observado com o seu sucesso (MAKRIDAKIS; WHEELWRIGHT; HYNDMAN, 1997). A autocorreção também ajuda a identificar, no caso do ARIMA, o valor da diferenciação para que a série se torne estacionária, condição necessária para a execução do modelo. Segundo Anderson, Grenfell e May (1984), a autocorrelação é definida pelo coeficiente r_k(onde K=0,1,...N-1, sendo a representação do tamanho da série), que verifica a correlação de diferentes observações a diferentes distâncias. Onde, r_k $(-1 \leq r_k \leq 1)$ mede a correlação entre a série temporal original e o deslocamento da mesma série em k observações.

Para identificar a ordem de séries auto-regressivas, utilizamos a Função de Autocorrelação Parcial (FACP, ou do inglês: PACF - *Partial Auto-Correlation Function*). A FACP verifica a correlação entre duas observações da série, essa medida corresponde a medida de X_h e X_0,

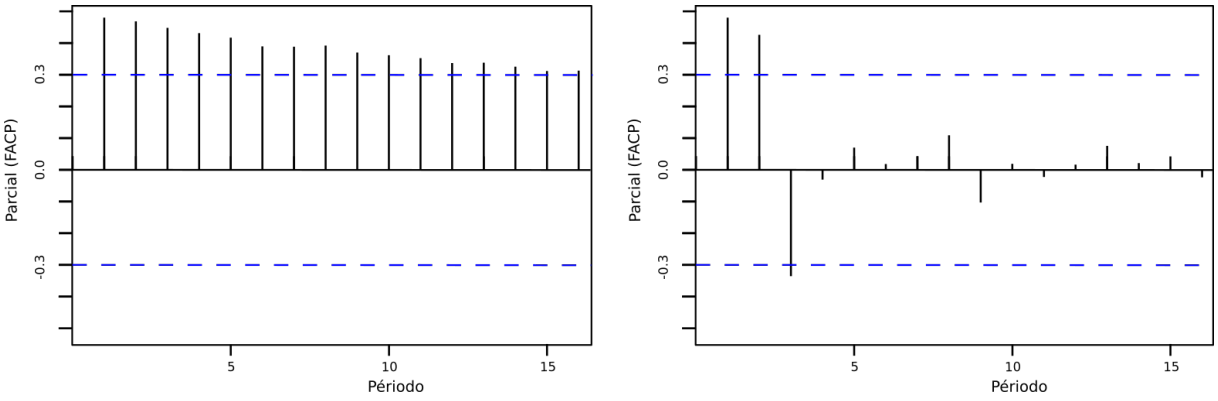
sendo:

$$\alpha(h) = Corr\{X_h - P(X_h|X_1, ..., X_{h-1}), X_0 - P(X_0|X_1, ..., X_{h-1})\}, \tag{4.4}$$

onde $P(X_t|X_1, ..., X_{t-1})$ é a melhor projeção linear de X_t em relação a $X_1, ..., X_{t-1}$.

Na Figura 9 temos um exemplo de função de autocorreção sem diferenciação à esquerda e com diferenciação à direita. A FAC mede a correlação entre todos os períodos da série, enquanto a FACP mede a correlação entre dois períodos da série. Dessa forma, caso os valores dos parâmetros não estejam muito próximos dos limites não estacionários, a equação de Yule-Walker poderá ser utilizada para estimativas aproximadas dos modelos autorregressivos sucessivos (BOX et al., 2015). Em resumo, a análise da correção ajuda a identificar qual a ordem da diferenciação necessária para tornar a série estacionária.

Figura 9 – Autocorreção parcial (FACP). À esquerda representação sem diferenciação e à direita com diferenciação.



Fonte: Elaborada pelo autor (2021)

Com isso, as duas metodologias de análise da correção apresentadas ajudam a identificar qual ordem de diferenciação é necessária para transformar uma série não estacionária em estacionária, de acordo com o que é necessário para a execução dos modelos.

4.2.4 Modelos Lineares e Não Lineares

Como destacado por Souza (2008), existem diversos tipos de modelos lineares como: análise espectral de métodos não paramétricos, modelos de espaço de estados, modelos autoregressivos (AR, do inglês: *Auto Regressive*), modelos de média móvel (MA, do inglês: *Moving Average*), modelos combinados ARMA (*Auto Regressive Moving Average*) e ARIMA (*Auto*

Regressive Integrated Moving Average), modelos autoregressivos com heterocedasticidade condicional GARCH (*Generalized Auto Regressive Conditional Heteroskedasticity*) e ARCH (*Auto Regressive Conditional Heteroskedasticity*), entre outros. O ARIMA, sendo um modelo probabilístico, tem um forte destaque na literatura (SOUZA, 2008). Sabemos que, em geral, modelos lineares produzem melhores resultados quando a parte linear da série é maior do que a parte não linear (YOLCU; EGRIOGLU; ALADAG, 2013). Os Modelos lineares são basicamente descritos por uma reta, definida pela interceptação e sua inclinação, podendo ser descrita pela seguinte equação 4.5 (EDWARDS, 1976):

$$y = \alpha + \beta x + \epsilon \tag{4.5}$$

onde y é a variável dependente (representando o eixo y), x é a variável independente, β é a inclinação da linha e α é a interceptação de y (EDWARDS, 1976).

Podemos também adicionar um elemento de resíduo ou erro, representado pela equação ϵ na equação 4.5, podendo ser uma variável aleatória com distribuição normal de média 0 e desvio padrão (σ). Qualquer série temporal, que não puder ser escrita na forma da equação acima, é considerada uma série temporal não linear.

O estudo de séries temporais é repleto de modelos não lineares. Dentre eles, destacam-se os modelos de redes neurais artificiais, como o LSTM que será descrito nas próximas seções.

4.3 LSTM

O LSTM (Redes de Memória de Curto Prazo Longo, do inglês: *Long short-term memory*), descrito por Hochreiter e Schmidhuber (1997), é uma variação da arquitetura RNN (Rede neural recorrente, do inglês: *recurrent neural network*), sendo uma abordagem de rede neural recursiva comumente aplicada a modelagem de dados sequenciais (SHERSTINSKY, 2020). O LSTM possui como principal característica descrever o desempenho dinâmico dos sistemas. A RNN pode manter um vetor de ativações para cada intervalo de tempo, tornando-a uma rede neural profunda (TEALAB, 2018; SAGHEER; KOTB, 2019; SHERSTINSKY, 2020). No entanto, pode ser difícil treinar uma RNN a ponto de aprender as dependências de longo prazo em dados de séries temporais, devido aos problemas de explosão ou desaparecimento do gradiente (FAKHFAKH et al., 2020). Ainda segundo o autor, ambos os problemas são causados devido à sua propriedade iterativa, cujo gradiente é igual à matriz de peso recorrente, mas elevado a

uma alta potência. Essas potências de matriz fazem com que o gradiente aumente ou diminua a uma taxa que é proporcional ao número de etapas de tempo no conjunto de dados (SINGH et al., 2019).

Entretanto, o problema de gradientes explosivos é considerado um problema relativamente simples, podendo ser resolvido por meio de um recorte, que é simplesmente encolher os gradientes cujas normas excedem um limite predefinido. Dessa forma, é possível, manter o gradiente pequeno na maior parte do tempo de aprendizado, sem reduzi-lo, para que o aprendizado ou a convergência não sejam prejudicados. Em contraste, o problema do gradiente de desaparecimento é extremamente difícil de resolver, uma vez que a característica do gradiente nas tendências, que correspondem às dependências de longo prazo são menores, o que pode prevenir que os pesos da rede mudem de valor, fazendo com que o modelo pare de aprender, enquanto faz com que o componente do gradiente nas tendências que correspondem ao curto prazo possua dependências de termo maiores/grandes. Por este motivo, a RNN sofre com as dependências de longo prazo, mas pode aprender as dependências de curto prazo, com certa facilidade (BENGIO; SIMARD; FRASCONI, 1994).

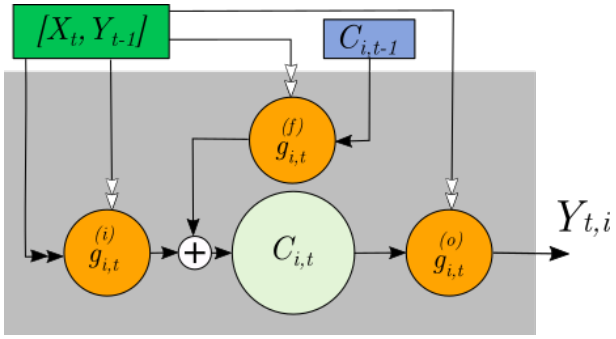
O LSTM surgiu como uma solução eficaz para tratar os problemas de gradiente da RNN (KOLEN; KREMER, 2009; HOCHREITER; SCHMIDHUBER, 1997; TEALAB, 2018; SAGHEER; KOTB, 2019). Ele usa uma célula de memória capaz de representar as dependências de longo prazo em dados sequenciais. As células de memória do LSTM são compostas por três gates/unidades que controlam as interações entre diferentes unidades de memória (GU et al., 2019):

- *input gate* - controla se o sinal de entrada pode modificar o estado da célula de memória ou não;
- *output gate* - controla se pode modificar o estado de outras células de memória ou não;
- *forget gate* - possui a capacidade de escolher entre esquecer ou lembrar seu status anterior.

A Figura 10 ilustra o modelo LSTM, onde o g^o e o g^i representam o *input* e *output gates* respectivamente, os quais protegem a célula de memória. A fórmula do LSTM inclui: um *input gate* (do português, portão de entrada), onde a adição dos conhecimentos úteis no estado da célula é feito; um *output gate* (do português, portão de saída), no qual a extração das informações úteis da célula atual são apresentadas; e um *forget gate* (do português, portão de esquecimento), onde as informações não relevantes são removidas. O *forget gate* pode

ser utilizado para limpar dados não necessários na célula de memória, não compartilhando as mesmas propriedades matemáticas do projeto original do LSTM. Devido à memória finita, as informações mais recentes são frequentemente consideradas as mais relevantes.

Figura 10 – LSTM com *forget gates*. Os círculos laranja são unidades de células multiplicativas. As setas duplas estão representando a saída de um neurônio com múltiplas entradas, enquanto setas únicas representam o caminho de um único valor. Ou seja, a entrada é múltipla enquanto a saída é unitária.



Fonte: PULVER; LYU (2017)

As portas, saídas ocultas e os estados das células, podem ser representados a partir das seguintes equações:

$$f_t = \sigma(X_t U_f + S_{t-1} W^f + b_f) \tag{4.6}$$

$$i_t = \sigma(X_t U^i + S_{t-1} W^i + b_f) \tag{4.7}$$

$$o_t = \sigma(X_t U^o + S_{t-1} W^o + b_o) \tag{4.8}$$

$$\tilde{C}_t = tanh(X_t U^c + S_{t-1} W^c + b_c) \tag{4.9}$$

$$C_t = C_{t-1} \otimes f_t \oplus i_t \otimes \tilde{C}_t \tag{4.10}$$

$$S_t = o_t \otimes tanh(C_t)) \tag{4.11}$$

onde, (Wf, Wi, Wo, Wc), (Uf, Ui, Uo, Uc) e (bf, bi, bo, bc) são pesos recorrentes, pesos de entrada e bias, respectivamente. O estado oculto e o estado da célula no passo de tempo t - 1 são representados por St - 1 e Ct - 1, respectivamente. As atividades sigmoides, adição

pontual, e as multiplicações pontuais, são representados pelos símbolos: σ, $\otimes$ e $\oplus$ (PULVER; LYU, 2017).

Sabemos que existem outras tentativas de superar o problema do gradiente de desaparecimento da RNN, entretanto, segundo (SAGHEER; KOTB, 2019), o LSTM é a tentativa mais eficiente. Desta forma, o LSTM é uma categoria de rede neural recorrente que pode aprender a dependência de ordem entre os itens em uma sequência, corrigindo os problemas de gradiente da RNN e tendo a promessa de conseguir aprender o contexto necessário para efetuar previsões em séries temporais.

4.4 ARIMA

O termo ARIMA é a abreviação de '*Auto Regressive Integrated Moving Average*' que é, na verdade, uma classe de modelos que 'explica' uma determinada série temporal com base em seus próprios valores anteriores, ou seja, seus próprios atrasos e erros de previsão defasados, para que a equação possa ser usada para prever valores futuros (NAU, 2020). Ainda segundo o autor, qualquer série temporal "não sazonal" que exibe padrões, e não é um ruído aleatório, pode ser modelada com modelos ARIMA.

Uma série temporal não estacionária pode ser tornar "estacionária" por diferenciação, se necessário. Sendo que, uma variável aleatória é estacionária se suas propriedades estatísticas forem constantes ao longo do tempo. Portanto, uma série estacionária não tem tendência de variações em torno de sua média, possui amplitude constante e, também, oscila de forma consistente, ou seja, os padrões de aleatoriedade no curto prazo sempre parecem os mesmos. Outro ponto importante é que as correlações com seus próprios desvios anteriores da média são sempre constantes. Assim, um modelo ARIMA pode ser interpretado como um "filtro" que tenta separar o sinal do ruído e, em seguida, o sinal é extrapolado no futuro para obter previsões (NAU, 2020).

Para que o modelo ARIMA possa ser estabelecido, primeiramente é necessário entender a combinação ARMA. O ARMA é formado pelo modelo autoregressivo (AR) com o modelo de média móvel (MA). Esses modelos podem ser aplicados diretamente a dados estacionários. Caso contrário, os dados não estacionários devem ser tratados por meio de processamento diferencial. Após a diferenciação, o modelo ARMA torna-se ARIMA (p,d,q) (P,D,Q), onde (p,d,q) é a sua parte não sazonal do modelo e (P,D,Q) é a parte sazonal, caso necessário. Assim, o modelo ARIMA é caracterizado por 3 partes (NAU, 2020):

- p - é a ordem do termo AR;
- q - é a ordem do termo MA;
- d - é o número de diferenciação(ões) necessário(s) para tornar a série estacionária.

O modelo ARIMA é obtido por meio da seguinte equação:

$$\phi_{\mathbf{p}}(\mathbf{B})\nabla^{\mathbf{d}}\mathbf{x_t} = \theta_{\mathbf{q}}(\mathbf{B})\mathbf{a_t} \tag{4.12}$$

As expressões na Eq. 4.12 são definidas a partir das seguintes equações (DJERBOUAI; SOUAG-GAMANE, 2016):

$$\phi_\rho\left(B\right) = (1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_\rho B^\rho) \tag{4.13}$$

$$\theta_q(B) = (1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q) \tag{4.14}$$

$$B^k x_t = x_{t-k} \tag{4.15}$$

$$\nabla^d = (1 - B)^d \tag{4.16}$$

onde, x_t é a parte não estacionária da série; o ruído branco usual do processo Gaussiano representado por a_t; ϕ_ρ e $\theta_q(B)$ são os AR e MA, representados respectivamente pelas ordens p e q sendo que p é a ordem da auto regressão e q é a ordem da média móvel; ∇^d é o operador de diferença não sazonal d; e B é chamado operador de retrocesso.

Para a construção de um modelo, haverá a necessidade da escolha dos melhores parâmetros. Os hiperparâmetros, p, q, P, Q e suas estimativas, são baseados nas funções de autocorrelação (ACF) ,i.e., a correlação entre as observações no momento atual e as observações em todos os momentos anteriores (BOX; PIERCE, 1970), função de autocorrelação parcial (PACF) que expressa a correlação entre as observações feitas em dois pontos no tempo, enquanto contabiliza qualquer influência de outros pontos de dados) (BOX; PIERCE, 1970); e o critério de informação de Akaike (AIC), originalmente descrito por Akaike (1974), que possui a seguinte formulação matemática (HYNDMAN; ATHANASOPOULOS, 2018; ANDO; KURAKAMI, 2016):

$$AIC = -2log(L) + 2(p + q + k + 1) \tag{4.17}$$

onde, o L representa a função de verossimilhança, visando medir a qualidade do ajuste do modelo estatístico de uma amostra de dados, a fim de determinar os valores dos parâmetros não conhecidos: k = 1 se c $\neq$ 0 e k = 0 se c = 0 (HYNDMAN; ATHANASOPOULOS, 2018; ANDO; KURAKAMI, 2016).

O critério de informação Bayesiano (BIC), descrito por Schwarz (1978) também pode ser utilizado para avaliar preditores da regressão, podendo ser descrito pela fórmula:

$$BIC = AIC + [log(T) - 2](p + q + k + 1) \tag{4.18}$$

Considerando o modelo ARIMA (p,d,q), tais critérios de informação podem ser utilizados para identificar os valores de p e q, não sendo bons para identificar o valor da diferenciação (d), uma vez que a diferenciação altera os dados da probabilidade calculada (HYNDMAN; ATHANASOPOULOS, 2018).

Com isso, o modelo ARIMA pode ser considerado simples devido a ser essencialmente um modelo autorregressivo linear geral, que solicita apenas variáveis endógenas sem expressar a necessidade de obter outras variáveis exógenas. No entanto, o modelo ARIMA captura essencialmente apenas relacionamentos lineares. Em outras palavras, os dados devem ser estacionários (LIU; TIAN; LI, 2012). Para dados não estacionários, a diferenciação é necessária, como falado anteriormente, mas nesse processo, algumas informações numéricas podem ser perdidas.

4.5 *PROPHET*

De acordo com Taylor e Letham (2017), o *Prophet* é um algoritmo de predição desenvolvido pelo Facebook, com o objetivo de criar previsões de negócios de alta qualidade. Ainda conforme o autor, a necessidade de seu desenvolvimento aconteceu após se observar que as técnicas de previsão que são totalmente automáticas, podem ser frágeis e muitas vezes inflexíveis para incorporar suposições úteis, ou heurísticas. Além disso, facilita a análise dos dados por cientistas de dados que possuem pouco conhecimento específico em séries temporais. Dessa forma, além da facilidade de uso, o *Prophet* lida com tendências e diferentes sazonalidades, discrepâncias, observações ausentes e mudanças de padrões.

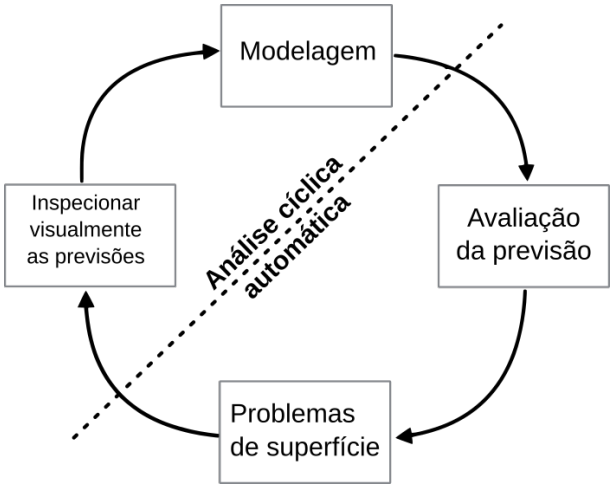
Prophet é um modelo regressivo aditivo, tendo um crescimento logístico linear ou para curva da tendência. Ele utiliza a série de Fourier sazonal anual e comportamento semanal

utilizando variáveis fictícias (SERVICES INC., 2020). Como descrito por J. e Letham (2017), *Prophet* é ideal para dados que possuem pelo menos uma das seguintes características:

- Observações periódicas de pelo menos alguns meses, seja por horas, dias ou semanas;
- Forte sazonalidade de semana e época do ano;
- Importantes feriados ocorrendo em intervalos irregulares durante o ano;
- Consideráveis observações ausentes, ou grandes *outliers* (acontecimentos fora da curva normal);
- Mudanças históricas devido a acontecimentos sociais ou econômicos; e/ou,
- Possuem tendências que são curvas de crescimento não linear, onde a tendência satura ou atinge o limite natural de crescimento.

Prophet, já com os valores definidos por padrão, possui boa precisão nos resultados das previsões, sendo um algoritmo de fácil manipulação (J.; LETHAM, 2017). Com uma análise cíclica da previsão e seus resultados é possível obter bons resultados para diversos casos (J.; LETHAM, 2017). Como apresentado na Figura 11, ele possui internamente um processo automático de ajuste dos seus parâmetros.

Figura 11 – *Prophet*: Esquema da análise cíclica para previsão, utilizada pelo algoritmo



Fonte: TAYLOR; LETHAM (2018)

O *Prophet* usa um modelo de série temporal composto por três componentes principais: tendência, sazonalidade e feriados. Estes componentes são combinados por meio da seguinte equação (TAYLOR; LETHAM, 2017):

$$y(t) = g(t) + s(t) + h(t) + \epsilon_t \tag{4.19}$$

onde, $g(t)$ é a função de tendência que modela as mudanças não periódicas no valor da série, $s(t)$ representa as mudanças periódicas (ex., sazonalidade semanal, mensal e anual) e $h(t)$ representa os efeitos dos feriados/aniversários/eventos que ocorreram em tempos potencialmente irregulares durante um ou mais dias. O termo t é a taxa de erro que representa quaisquer mudanças idiossincráticas que não são acomodadas pelo modelo.

Com o *Prophet* é possível implementar dois modelos de tendência para $g(t)$. O primeiro é chamado crescimento não linear e saturado, representado como um modelo de crescimento logístico (THIYAGARAJAN et al., 2020):

$$g(t) = \frac{C}{1 + exp{-k(t - m))}} \tag{4.20}$$

onde, C é a capacidade de carga (sendo o valor máximo da curva), k é a taxa de crescimento (que representa "a inclinação" da curva) e m é um parâmetro de deslocamento.

Sendo o segundo, um modelo linear por partes, uma modificação do modelo linear convencional. Entretanto, por padrão, o *Prophet* utiliza o modelo de crescimento logístico.

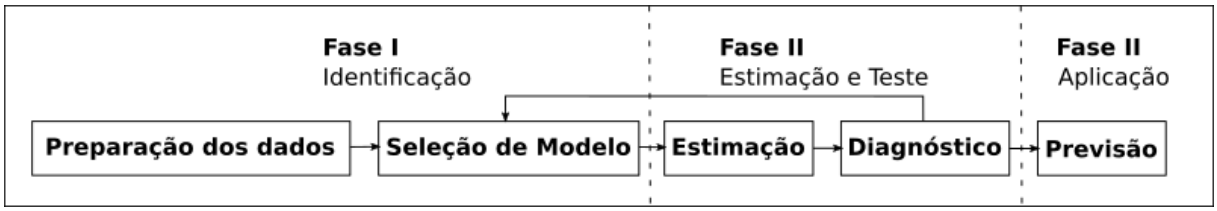
Essa equação logística permite modelar o crescimento não linear com saturação, ou seja, quando a taxa de crescimento de um valor diminui com o seu crescimento. Pode ser bem aplicado aos CIDs/CIAPs de séries não estacionárias.

4.6 AVALIAÇÃO DOS MODELOS

Os modelos foram avaliados por meio das fases descritas no estudo de Box et al. (2015), sendo comumente utilizadas na literatura (MARTINEZ; SILVA; FABBRO, 2011) para construção de modelos paramétricos de séries temporais univariadas. O modelo é definido resumidamente por: preparação dos dados, seleção do modelo, estimação dos parâmetros, diagnóstico e previsão, como mostrado resumidamente na Figura 12.

A definição de cada etapa é apresentada a seguir:

Figura 12 – Metodologia Box-Jenkins para o processo



Fonte: MAKRIDAKIS; WHEELWRIGHT; HYNDMAN (1997)

- **Preparação dos dados**: Análise e transformação dos dados visando estabilizar a variação, onde é realizada a diferenciação para converter os dados em uma série estacionária, quando necessário;
- **Seleção do modelo**: Verificação dos resultados da função de autocorrelação (FAC) e autocorrelação parcial (FACP);
- **Estimação**: Estima os parâmetros no modelo em potencial, e realiza a seleção dos melhores hiperparâmetros (valores de ajuste do modelo) usando o critério adequado;
- **Diagnóstico**: Análise dos resíduos (valores reais e valores previstos). Caso exista inconsistência nessa etapa, a etapa de seleção do modelo é reiniciada;
- **Previsão**: Usa o modelo para previsão.

Essas cinco etapas são executadas sequencialmente até a etapa do diagnóstico, onde é analisada a existência de ruídos brancos, ou seja, se os resíduos não estão correlacionados. A existência de ruído branco indica que o modelo está ajustado, caso contrário, será necessário executar novamente a etapa da seleção do modelo. O teste estatístico comumente utilizado na análise de ruído branco é o Ljung-Box (LB) (LJUNG; BOX, 1978), onde a regra de decisão para H0 é *p-value* $>$ 0,05, definida pelas seguintes hipóteses:

H0: define a existência de ruído branco e dessa forma, não existe falha no modelo.

H1: inexistência de ruído branco, o modelo apresenta falha de ajuste.

O ruído branco nada mais é do que uma série estacionária, de média zero e variância igual a um, além de suas funções de autocorrelação sempre serem nulas, dependendo apenas do número de defasagens (BOX et al., 2015). Um ruído branco é o mais fundamental exemplo de

um processo estacionário quando a interdependência e aleatoriedade das variáveis σ_a^2 pode ser assumida pela média zero e a variância σ_a^2 (BOX et al., 2015), dada por:

$$y_k = E[a_1 a_{1+k}] = \begin{cases} \sigma_a^2 & k = 0 \\ 0 & k \neq 0 \end{cases} \tag{4.21}$$

Caso se concentre na propriedade da segunda ordem, então uma sequência a_1 não correlacionada, tendo média zero, e variância comum de σ_a^2 tem a mesma covariância da funcão y_k, e é estacionária fraca de segunda ordem. Apesar de possuir propriedade simples, o ruído branco também é associado a mais complexas propriedades (BOX et al., 2015).

4.7 CONSIDERAÇÕES DO CAPÍTULO

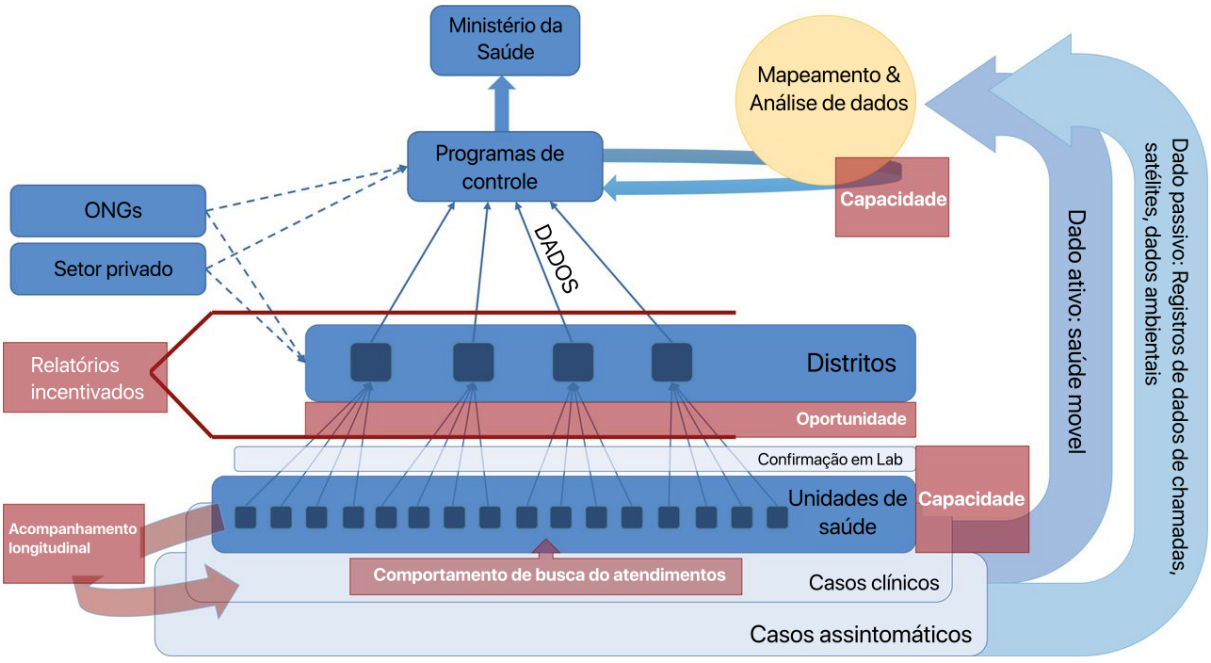
Nesta seção, foram apresentados os conceitos que embasaram o processo de desenvolvimento do trabalho, bem como, o delineamento do estudo, a casuística, processos estatísticos, a apresentação dos modelos utilizados na pesquisa e os processos de avaliação dos modelos.

5 MAPDI: MODELO AUTOAJUSTÁVEL PARA PREDIÇÃO DO AUMENTO DO NÚMERO DE CASOS DE DOENÇAS INFECTOCONTAGIOSAS

Surtos de novas doenças infecciosas com potencial pandêmico representam uma ameaça considerável à vida e ao desenvolvimento humano (MURRAY et al., 2006). A resposta e a contenção ideal a um surto de doença infecciosa podem ser bastantes melhoradas se as medidas de resposta à saúde e controle de surtos puderem ser focadas, de forma proativa, em áreas previstas com maior risco de ocorrência de novos surtos (BUCKEE et al., 2018).

A metodologia utilizada na saúde pública do Brasil para identificar um possível crescimento no número de casos de doenças infecciosas ocorre, em sua grande maioria, por meio dos dados coletados passivamente e disponibilizados via sistemas de vigilância. Nesses sistemas, os profissionais de saúde recebem a incumbência de inserir as informações dos pacientes em um banco de dados central, usado para determinar tendências ao longo do tempo e mapear a distribuição geográfica da carga das doenças. Esses dados regionais servem como uma base importante para as decisões de alocação de recursos. A Figura 13 ilustra o fluxo de dados e possíveis obstáculos enfrentados pelos programas de controle e as maneiras pelas quais novas abordagens podem ser usadas paralelamente aos sistemas tradicionais (BUCKEE et al., 2018).

Figura 13 – Fases da coleta de dados sobre pandemias



Fonte: BUCKEE et al. (2018)

A Figura 13 mostra a fluidez dos dados pelos sistemas de saúde (azul) e os principais desa-

fios enfrentados pelos sistemas de controle (vermelho). Um subconjunto de casos clínicos, que geralmente representam apenas uma parcela das infecções totais, tanto assintomáticas quanto clínicas, na sua grande maioria, são detectados inicialmente por profissionais dos municípios, mais comumente em unidades de saúde e hospitais. Os profissionais também são responsáveis pelo acompanhamento dos indivíduos que apresentam quadros clínicos de infecções crônicas e que requerem tratamentos por períodos extensos. Algumas frações desses casos são confirmados em laboratórios e relatados aos centros regionais ou distritais que, no que lhe concerne, se reportam aos programas nacionais de controle. As ONGs e o setor privado também podem produzir uma quantidade significativa de dados epidemiológicos. Os programas nacionais de controle agregam e analisam dados para mapear a distribuição da carga de doenças, a eficácia da intervenção e assim por diante. Novas abordagens diretas, como, por exemplo, vigilância participativa e dados coletados passivamente (ex.: de telefones celulares via registro de chamados (CDRs, do inglês: *Call Data Records*), etc.) podem ser usados diretamente por programas de controle para mapear riscos subjacentes e distribuições da população. Em todos os níveis, a capacidade de coleta de informação e o que fazer com esses dados, continua sendo um enorme problema para a vigilância de rotina: a capacitação dos profissionais envolvidos para essas novas abordagens são um desafio para a maioria dos programas de controle.

Diante de tais fatos, vimos a necessidade de desenvolver um sistema que coletasse os dados de forma ativa, diretamente nas unidades de saúde, e identificasse anomalias (crescimento exagerado) de algumas doenças infecciosas para, desta forma, analisar o seu crescimento preditivamente por algoritmos de Inteligência Artificial (IA), com autoajuste dos modelos automatizados. Com isso, será possível gerar análises inteligentes para que os gestores da saúde pública, mesmo sem conhecimento de IA, possam agir proativamente no âmbito das doenças analisadas.

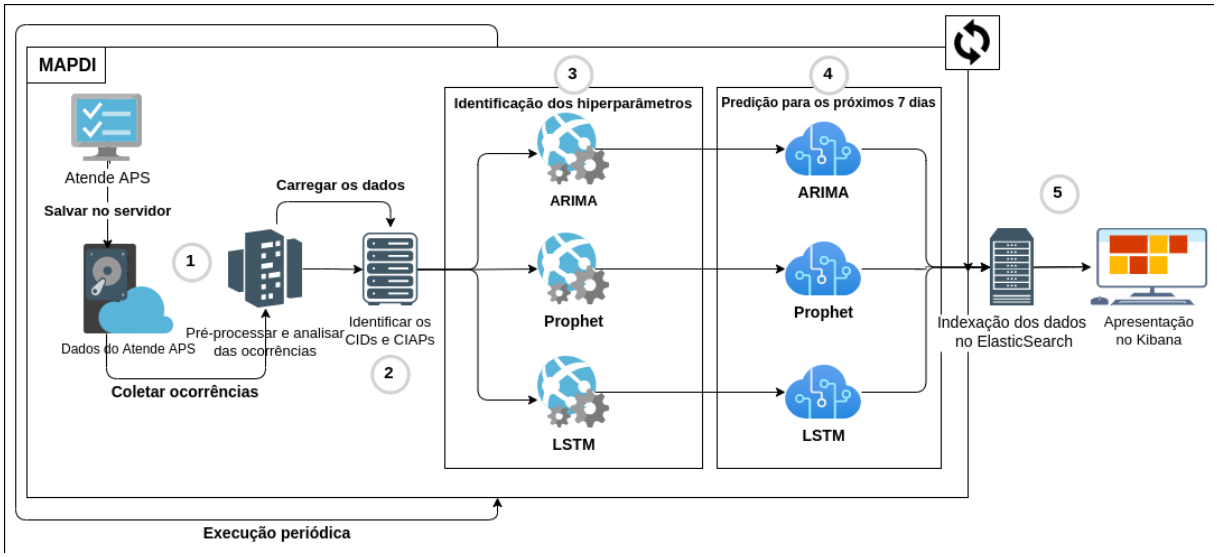
Com base nesta premissa, o processo MAPDI (Modelo Auto Ajustável para Predição do Aumento do Número de Casos de Doenças Infectocontagiosas) foi desenvolvido e um protótipo de painel de análise de dados foi criado para auxiliar e alertar os gestores sobre os possíveis aumentos desenfreados de um dado CID/CIAP.

5.1 VISÃO GERAL

O MAPDI (Modelo Autoajustável para Predição de Aumento de Número de Casos de Doenças Infectocontagiosas) visa identificar de forma inteligente a possível progressão das

ocorrências de uma determinada doença, visto que o processo de coleta e otimização dos modelos de previsão se dá de forma automática. Dessa forma, tem potencial de auxiliar a atuação de políticas públicas de forma proativa com a previsão resultante do processo, ajudando a evitar que o aumento de casos de determinadas doenças venham a se tornar epidêmicas ou mais complexas de lidar. Para tal, o MAPDI foi construído em 5 fases, conforme mostrado na Figura 14.

Figura 14 – As 5 etapas do processo MAPDI



A seguir temos uma breve descrição das 5 fases do processo, que serão detalhadas nas seções seguintes.

1. **Coleta e pré-processamento dos dados** - Os dados são coletados das unidades de saúde especializadas e então pré-processados, possibilitando a identificação dos CIDs/-CIAPs e os bairros de suas ocorrências.

2. **Identificação CIDs/CIAPs anômalos** - É responsável por identificar anomalias (ou seja, a ocorrência significativa de casos de um CID/CIAP de um determinado distrito sanitário). O algoritmo desenvolvido para esta fase agrupa todos os CIDs/CIAPs, a respectiva comunidade do evento e observa o crescimento de novos casos durante uma semana, onde calculará e selecionará aqueles mais alarmantes.

3. **Ajuste automático dos algoritmos de séries temporais** - Nesta fase, um conjunto de rotinas será executado para identificar o tipo de distribuição e de série temporal dos

dados e, com isso, identificar os melhores hiperparâmetros (parâmetros ajustáveis do modelo) para cada um dos algoritmos.

4. **Execução das previsões** - Com os melhores parâmetros para execução dos modelos, são feitas as previsões das ocorrências de cada CID/CIAP em um futuro próximo e seus resultados serão armazenados.

5. **Indexação e apresentação dos resultados** - Os resultados obtidos serão indexados em uma ferramenta de recuperação de informação e, por meio de um painel interativo, os dados serão exibidos em um dashboard para os gestores.

5.2 COLETA E PRÉ-PROCESSAMENTO DOS DADOS

Como mencionado no Capítulo 4, Seção 4.1, os dados foram obtidos diretamente de 24 unidades de saúde distribuídas em 96 bairros da cidade do Recife por meio de uma API[1] (do ingês: *Application Programming Interface*) do aplicativo Atende APS. Esses dados são compostos por todos os atendimentos realizados no período de um ano, logo, sabemos que a qualidade dos dados tem um impacto direto no desempenho dos algoritmos. Dessa forma, uma representação ruim pode tornar o modelo ineficiente, mesmo os mais complexos e avançados (NAJAFABADI et al., 2015). Diante disso, inicialmente reduzimos a dimensionalidade dos dados brutos, visto que a estrutura dos dados foi construída de forma heterogênea para dar suporte a diferentes situações e a própria evolução do sistema. Todas as unidades de saúde que possuem esse sistema, executam 3 etapas principais, no atendimento aos pacientes:

1. Triagem por meio de perguntas básicas relacionadas ao seu estado atual de saúde.

2. Caso na triagem o paciente apresente sintomas mais urgentes, uma avaliação técnica pode ser feita para obter dados relacionados aos seus sinais vitais.

3. Um profissional da saúde irá fazer uma avaliação das informações obtidas nas etapas anteriores e novas perguntas serão feitas ao paciente, de modo a classificá-lo por meio de um ou mais CIDs/CIAPs, consolidando-se em uma consulta médica.

[1] O que é API?: <https://www.redhat.com/pt-br/topics/api/what-are-application-programming-interfaces>

Tabela 1 – Estatísticas do conjunto de dados coletados do sistema Atende APS

Descrição	Valor
Número de variáveis	**109**
Número de observações	28.613
Entradas ausentes	933.269
Entradas ausentes (%)	29,9%
Tamanho total na memória	23,8 MB

Fonte: Elaborada pelo autor (2021)

O processo de triagem e avaliação médica gera uma quantidade significativa de informações do paciente, que para o nosso estudo não são relevantes, uma vez que consideramos apenas o resultado da classificação final.

Identificamos 109 atributos, como mostra a Tabela 1,e muitos desses atributos não foram considerados relevantes. Por isso se fez necessário uma segunda fase dedicada à remoção de dados não relevantes, devido à quantidade significativa de valores ausentes ou dados fora de contexto, como nos atributos: data de atualização, descrição do resultado, estado de saúde do ciap2, dados do apoio emocional, latitude e longitude da unidade de saúde, entre outros.

Também foi identificada a existência de registros de outras regiões próximas à cidade de Recife, que também foram descartados visando diminuir possíveis ruídos na informação. Por Recife ser a capital pernambucana, é comum que pessoas das cidades vizinhas procurem as unidades de saúde em busca de melhores atendimentos.

Os dados originais contêm 1.176 rótulos de bairros, que apresentam variações com erros ortográficos, visto que foram obtidos a partir do CEP informado pelos pacientes e coletados os bairros a partir deste CEP de diferentes fontes. Como resultado, decidimos usar uma normalização semiautomática com precisão de associação de 80%, utilizando o difflib[2], visando uniformizar a nomenclatura para os 96 bairros do Recife. Também, foi necessário ajustar, para alguns dos modelos, os finais de semana e feriados da série temporal, visto que muitas das unidades de saúde não atendem nesses dias.

5.3 IDENTIFICAÇÃO CIDS/CIAPS ANÔMALOS

Para encontrar as Classificações Internacionais de Doenças (CIDs) e as Classificações Internacionais de Atenção Primária (CIAPs) que apresentam comportamentos anômalos, foi

[2] difflib: https://docs.python.org/3/library/difflib.html

necessário o desenvolvimento de um filtro simples, apresentado no pseudocódigo Algoritmo 1. Foi necessário agrupar os identificadores de cada classificação utilizados em uma consulta médica, verificando se doenças infecciosas e contagiosas estavam contidas em uma matriz de valores. Além disso, as informações foram agrupadas por meio dos Distritos Sanitários (DSs) (conjunto de bairros) da cidade, com o intuito de possibilitar futuramente a análise pela dimensão técnica da cidade. Após essa análise, foi possível identificar e organizar os dados para futuras operações.

Algoritmo 1: Coleta de dados por distritos de saúde

```
Result: Obtenção de CIDs/CIAPs infecciosos e contagiosos de cada registro
        relacionado aos distritos sanitários
Coletar os dados de consultas médicas;
Identificar qual bairro está incluído em cada Distrito Sanitário;
Identificar os CIDs e CIAPs;
while Não verificado todos das consultas do
    for CIDs/CIAPs identificados por consulta do
        if Contém CIDs/CIAPs infecciosos e contagiosos then
            Identificar o bairro da consulta;
            for Distritos de saúde do
                if Pertence a um Distrito Sanitário then
                    Atualiza o correspondente distrito sanitário com a ocorrência;
                end
            end
        end
    end
end
```

Fonte: Elaborado pelo autor (2021)

Previamente são identificados os bairros de cada DS e os CIDs/CIAPs relacionados a doenças infecciosas. Uma vez que, o resultado da consulta médica consiste na identificação de uma lista de CIDs/CIAPs, é necessário extrair tais ocorrências individualmente de cada paciente. A primeira rotina condicional interage na classificação e posteriormente identifica a ocorrência da classificação de doenças infectocontagiosas. No próximo passo, o CIDs/CIAPS é agrupado no DS correspondente. Dessa forma, podemos posteriormente, caso necessário, agrupar os dados e contabilizar a quantidade de ocorrências pelos distritos e possibilitar uma

melhor acompanhamento relacionado às zonas distritais. Atualmente, devido a características heterogêneas da cidade, os distritos sanitários comportam bairros de diferentes níveis sociais e econômicos. Tal informação é importante, dado que áreas distritais maiores podem ofuscar inícios de surto localizados em bairros/regiões menores.

5.4 CONJUNTO DE DADOS

Considerando os Distritos Sanitários individualmente, a Tabela 2 apresenta a quantidade de ocorrência dos cinco CIDs/CIAPs que tiveram maiores ocorrências em cada DS. É possível observar que o CID U07.2, correspondente ao "Diagnóstico Clínico ou Epidemiológico de Covid-1", possui a maior quantidade de ocorrências. o CID é justificado pelo objetivo primário do sistema Atende APS, o qual lida com a pandemia do vírus Covid-19 em curso e define a estreita relação com a situação brasileira atual.

Tabela 2 – Quantidade de ocorrências dos cinco CIDs ou CIAPs que tiveram maiores ocorrências em cada Distrito Sanitário

Distrito	CID U07.2	CID J11	CIAP R80	CIAP R74	CIAP R83	Total de ocorrências
DS I	1.166	389	313	291	503	2.662
DS II	2.256	659	1.509	377	587	5.388
DS III	1.696	447	289	267	216	2.915
DS IV	2.073	1.440	1942	1.265	1.026	7.746
DS V	1.861	1307	395	388	337	4.288
DS VI	2.335	138	226	1.178	391	4.268
DS VII	2.471	963	764	625	276	5.099
DS VIII	2.079	731	481	449	284	4.024

Fonte: Elaborado pelo autor (2021)

A Tabela 2, possui os dados utilizados para o treinamento de cada modelo, dividido por distrito sanitário, para melhor elucidar o cenário das predições. Com isso, cada CID/CIAP foi dividido em 70% para compor a base treinamento e 30% para compor a base teste, deixando os 7 últimos dias de dados observados para compor os dados de validação.

5.5 AJUSTE AUTOMÁTICO DOS ALGORITMOS DE SÉRIES TEMPORAIS

Cada um dos algoritmos utilizados possui suas particularidades para a escolha dos melhores parâmetros. Com isso, para explicar com maior autonomia, as seções 6.3, 6.4 e 6.5 elucidaram

todo o processo necessário para escolher os melhores hiperparâmetros individualmente por algoritmo. Hiperparâmetros nada mais são do que os valores ajustáveis na entrada de cada algoritmo de IA. A qualidade do resultado de um modelo de IA está intrinsecamente ligada a uma boa definição dos valores dos hiperparâmetros. Eles possibilitam que o modelo se ajuste a cada contexto, tornando o modelo genérico o suficiente para lidar com diferentes situações.

No ARIMA, foi utilizado o AutoArima para identificar automaticamente os melhores hiperparâmetros (HYNDMAN; KHANDAKAR, 2008) desabilitando o parâmetro "*stepwise*" (passo a passo) e o "*approximation*", com o objetivo de identificar os melhores hiperparâmetros[3] a fim de recalibrar o modelo uma vez por semana, mesmo que, de acordo com a necessidade, o algoritmo pudesse ser executado diariamente. Descrito por Hyndman e Khandakar (2008), o algoritmo *stepwise* é utilizado, salvo quando, o método padrão para selecionar diferenças sazonais, é baseado em uma estimativa da força sazonal (WANG; SMITH; HYNDMAN, 2006) ao invés do teste Canova-Hansen (CANOVA; HANSEN, 1995). Mesmo que, analisar o modelo sem o *stepwise* possa, principalmente com dados não sazonais, ser mais lento.

No *Prophet* foi utilizada a capacidade do próprio modelo para definir automaticamente os melhores hiperparâmetros, com um pequeno ajuste no algoritmo original para ligar valores negativos. Dessa forma, foi avaliada a função linear por partes, para manter a tendência positiva e evitar previsões negativas, as quais não fazem sentido no escopo dessa pesquisa. Como citado anteriormente o *Prophet*, por padrão, possui boa precisão (J.; LETHAM, 2017), como analisaremos no Capítulo 6.

Para o LSTM, a identificação automática dos hiperparâmetros foi feita pela implementação do *Grid Search*. O *Grid Search* é uma técnica mais exploratória, onde podemos definir um limiar de valores possíveis dos hiperparâmetros, como descrito na Seção 6.4, para serem testados e obter o melhor resultado dentre os valores obtidos. *Grid Search* é uma busca exaustiva que foi realizada sobre os possíveis parâmetros do modelo. Devido à sua característica exploratória, com os dados desta pesquisa e a faixa de análise definida, foi o que requisitou mais tempo para a obtenção dos melhores parâmetros de ajuste do modelo. Enquanto o *Prophet* e o ARIMA obtêm os melhores hiperparâmetros em questão de minutos, o LSTM leva algumas horas.

Na fase de execução das previsões, com os melhores hiperparâmetros devidamente definidos, são geradas previsões para os próximos 7 dias. Uma vez utilizados três algoritmos nesta fase, damos a capacidade do MAPDI de fornecer aos gestores visões de diferentes algorit-

[3] Ajuste do melhor modelo ARIMA em séries temporais univariadas: https://www.rdocumentation.org/packages/forecast/versions/8.13/topics/auto.arima

mos com grau de confiança acima de 95%. Desta forma, permitimos a análise com diferentes perspectivas.

5.6 INDEXAÇÃO E APRESENTAÇÃO DOS RESULTADOS

A indexação é a organização do dado em um formato que permite a partição dos dados em uma determinada maneira (ELASTIC, 2021). Segundo Bindá, Brandt e Piedade (2013) a indexação é um processo em que o conteúdo dos documentos é convertido em um formato que permite a recuperação rápida dos elementos nele contido. Esse processo gera um índice como saída, que na sua grande maioria pode ser uma estrutura de dados que possibilitam o acesso aos termos de cada documento, tendo a mesma funcionalidade de um índice de um livro (CORREIA, 2018). A qualidade do processo de indexação está inteiramente relacionada ao desempenho que as consultas irão desempenhar (JONES; WILLETT, 1997).

Os índices, produtos do processo de indexação, são estruturas de dados antigas e bastante difundidas (BAEZA-YATES; RIBEIRO-NETO, 2011). Um índice equivale a uma coleção de palavras associadas aos documentos. A construção de forma eficiente de tais índices contribui para um alto desempenho nas consultas e, consequentemente, em uma maior qualidade no conjunto de respostas (GONZALEZ; LIMA, 2002).

Para tal, o MAPDI utiliza, para o processo de indexação dos resultados a ferramenta *Elasticsearch*[4], que agrupa os documentos/dados nos índices com os quais podem relacionar-se entre si. Cada documento é indexado em formato JSON[5], com chaves e seus respectivos valores (*strings*, números, booleanos, datas, matrizes de valores, geolocalizações ou outras categorias de dados).

Com os dados devidamente indexados na plataforma, utilizamos uma ferramenta auxiliar denominada *Kibana*[6], que fornece recursos de busca e visualização para os dados indexados. Abaixo, seguem as Figuras 15, 16, 17, 18 e 19 que apresentam as visualizações geradas na validação dos modelos no processo do MAPDI por meio da ferramenta *Kibana*.

As Figuras apresentam os resultados das previsões das ocorrências de cada CID/CIAP analisado do mês de fevereiro de 2021 até 07 de março de 2021, onde:

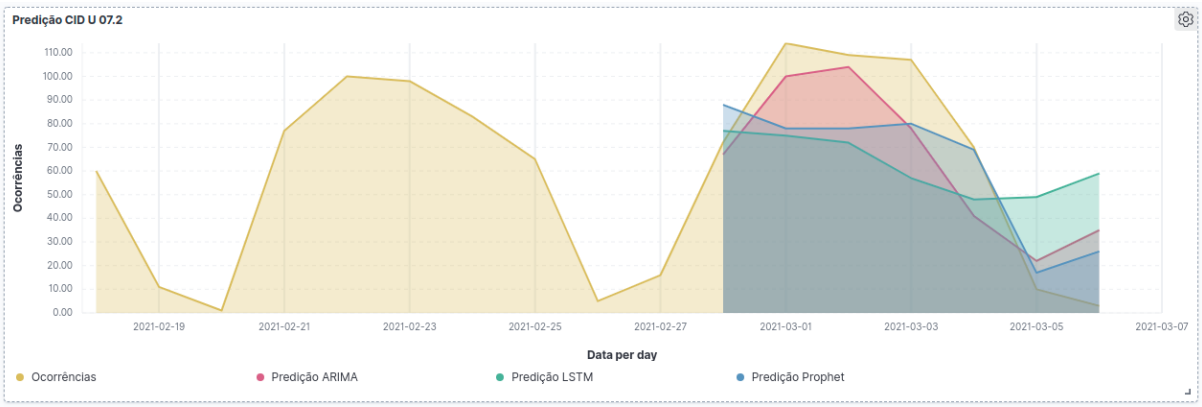
- A série de cor **amarela** representa a ocorrência atual dos casos de ocorrências diárias

[4] Elasticsearch:<https://www.elastic.co/pt/>
[5] Introducing JSON: <https://www.json.org/json-en.html>
[6] O que é o *Kibana*?: <https://www.elastic.co/pt/what-is/kibana>

reais.

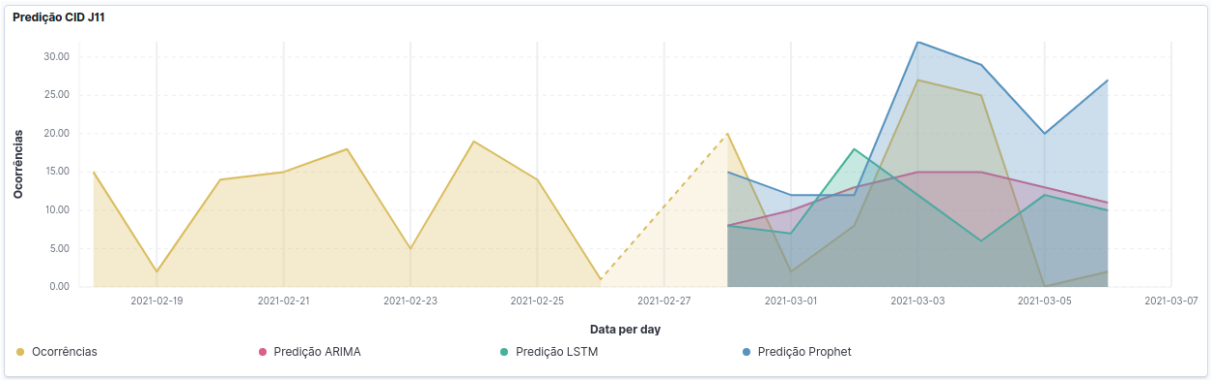
- A série de cor **vermelha** representa os casos diários preditos pelo modelo ARIMA.
- A série de cor **verde** representa os casos diários preditos pelo modelo LSTM.
- A série de cor **azul** representa os casos diários preditos pelo modelo *Prophet*.

Figura 15 – Apresentação dos resultados das predições do CID U 07.2 em comparação com as ocorrências reais no ambiente*Kibana*



Fonte: Elaborada pelo autor (2021)

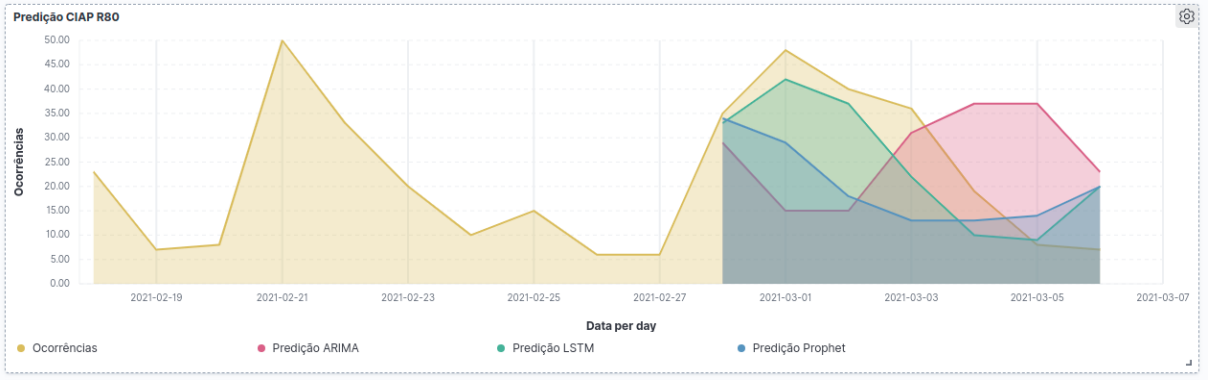
Figura 16 – Apresentação dos resultados das predições do CID J11 em comparação com as ocorrências reais no ambiente *Kibana*



Fonte: Elaborada pelo autor (2021)

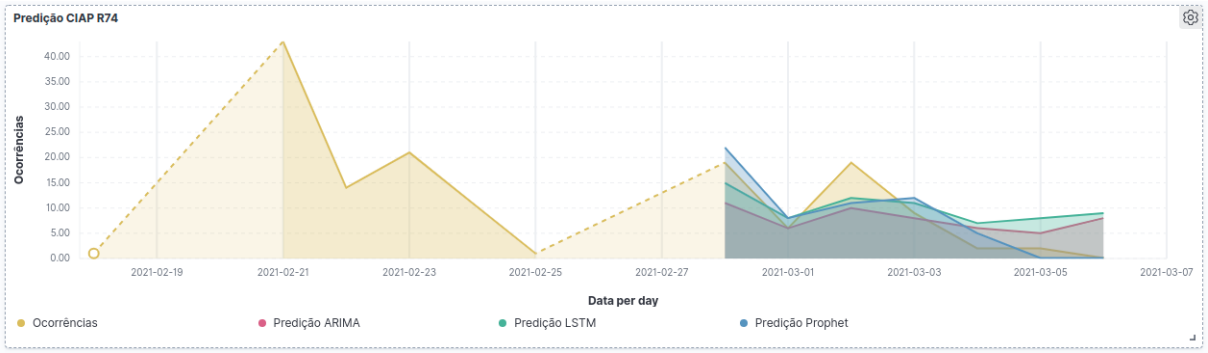
As Figuras citadas demonstram como é possível apresentar os resultados das previsões na *interface* do *Kibana*. Além disso, o *Kibana* fornece filtros por data e pelos campos existentes nos dados que foram indexados no *Elasticsearch* e, possibilita a produção de novos gráficos pela própria ferramenta. É possível, considerando os dados de uma série temporal, apresentar

Figura 17 – Apresentação dos resultados das predições do CIAP R80 em comparação com as ocorrências reais no ambiente *Kibana*



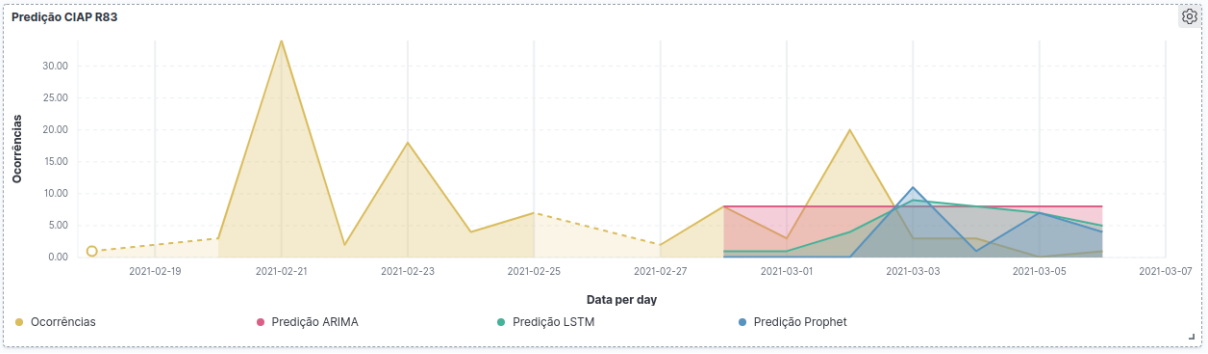
Fonte: Elaborada pelo autor (2021)

Figura 18 – Apresentação dos resultados das predições do CIAP R74 em comparação com as ocorrências reais no ambiente *Kibana*



Fonte: Elaborada pelo autor (2021)

Figura 19 – Apresentação dos resultados das predições do CIAP R83 em comparação com as ocorrências reais no ambiente *Kibana*



Fonte: Elaborada pelo autor (2021)

e personalizar elementos gráficos da visualização, e até mesmo ajustar as séries em função da soma acumulada, média móvel, percentil, entre outras funcionalidades. Dessa forma, os gestores da prefeitura do Recife podem ter um vasto arcabouço de possibilidades, sendo a execução e envio das análises realizadas de forma periódica e automática.

6 ANÁLISE DOS RESULTADOS

A presente seção visa demonstrar os resultados obtidos e as métricas utilizadas para alcançá-los.

6.1 MÉTRICAS DE ANÁLISE DOS DESEMPENHOS

Para a análise do desempenho do modelo foram utilizadas as seguintes métricas: Média Percentual Absoluta do Erro (MAPE, do inglês: *Mean Absolute Percentage Error*), Erro Médio Absoluto (MAE, do inglês: *Mean Absolute Error*), Raiz Quadrada do Erro-Médio (RMSE, do inglês: Root Mean Squared Error) e a Precisão Direcional Média (MDA, do inglês: *Mean Directional Accuracy*). Abaixo segue breve explanação sobre cada uma das métricas de erro: MAPE - define erro da acurácia em porcentagem, ou seja, o valor do MAPE representa a porcentagem em que a predição está errando com relação aos dados reais. Sendo este calculado a partir do erro absoluto em cada período, dividido pelos valores observados naquele período (KHAIR et al., 2017):

1. **RMSE** - resume a discrepância entre os valores observados e os valores esperados no modelo (OLSAVSZKY et al.,). Em resumo, a raiz quadrada da média do quadrado da diferença entre a previsão e as observações reais, onde é mais útil quando grandes erros são particularmente não desejados (SCHULLER et al., 2020).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i^{obs} - y_i^{pred})^2} \tag{6.1}$$

   onde, y_i^{obs} é a observação atual e y_i^{pred} é a observação predita.

2. **MAE** - mede a magnitude média dos erros na predição, sendo a média sobre a amostra de teste das diferenças absolutas entre a previsão e a observação real, com pesos iguais (SCHULLER et al., 2020). O MAE é representada pela seguinte fórmula:

$$MAE = \frac{\sum_{i=1}^{n}\left|y_i^{obs} - y_i^{pred}\right|}{n} \tag{6.2}$$

   onde, y_i^{obs} é a observação atual e y_i^{pred} é a observação predita.

   MAE e RMSE são ambos valores orientados negativamente onde a medida de erro do RMSE é preferível (WILLMOTT; MATSUURA, 2005).

3. **MDA** - é utilizada para mensurar a precisão de previsão de um método estatístico. O MDA considera a direção da previsão com a direção real. Dessa forma, característica que se adequa bem a modelos de economia financeira, onde a oscilação da direção da série pode ser importante para a tomada de decisão (SCHULLER et al., 2020). A fórmula a seguir define a equação:

$$MDA = \frac{1}{N}\sum^{t} 1sign(X_t - X_{t-1}) == sign(F_t - X_{t-1}) \tag{6.3}$$

onde, X_i é a observação atual, F_i a estimação, N é pontos não faltantes, $sign()$ é função sinal onde o retorno é compreendido pelos retornos:

$$sing(x) = \begin{cases} -1 & se \quad x < 0 \\ 0 & se \quad x = 0 \\ 1 & se \quad x > 0 \end{cases} \tag{6.4}$$

As métricas transcritas acima, serão utilizadas para analisarmos os resultados dos processos de escolha dos melhores hiperparâmetros. Com isso, os modelos que apresentarem um menor resultado nestas métricas, serão os hiperparâmetros selecionados. As seções seguintes apresentarão com mais detalhes todo o processo.

6.2 ANÁLISE UNIVARIADA CIDS/CIAPS ANÔMALOS

Devido à peculiaridade de cada algoritmo de série temporal utilizada neste trabalho, fez-se necessário analisar o comportamento de cada conjunto de dados, para identificar se a sua série temporal apresenta características estacionárias. Diante disso, foi necessária a utilização de duas metodologias para determinar a estacionariedade. A primeira foi a estatística móvel — que consiste em analisar a média móvel e o desvio padrão móvel. Para que uma série possa ser considerada estacionária, se faz necessário observar no gráfico se o desvio padrão e a média permanecem constantes em relação ao tempo no gráfico. Para tal observação, as linhas deverão estar retas e paralelas ao eixo X. A outra forma de verificar a estacionariedade é usando o teste de Dickey-Fuller aumentado (ADF test, do inglês: *Augmented Dickey-Fuller test*. Para melhor compreensão precisamos observar as seguintes hipóteses:

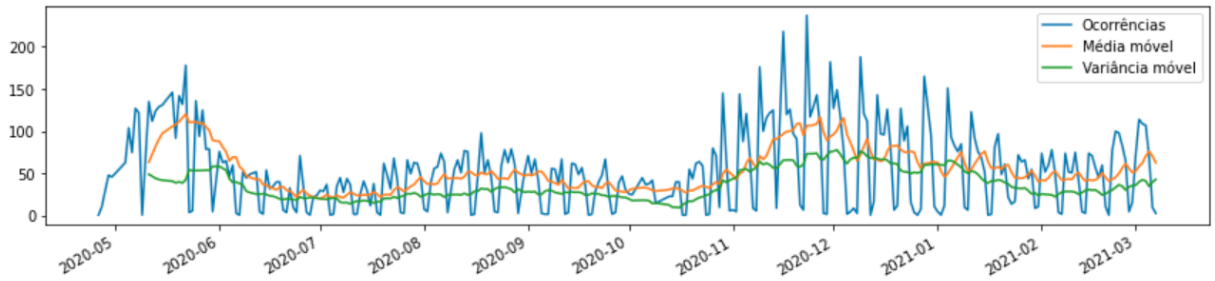
H0: A série temporal não é estacionária.

H1: A série temporal é estacionária.

Portanto, se o valor do p-value do teste for menor que a significância de 0,05, deve se rejeitar a hipótese nula e inferir que a série temporal é de fato estacionária (*i.e.*, outros intervalos de confiança podem ser inseridos e analisados). Nessa análise, foi utilizada uma janela de 12 meses tanto para o desvio padrão, quanto para a média.

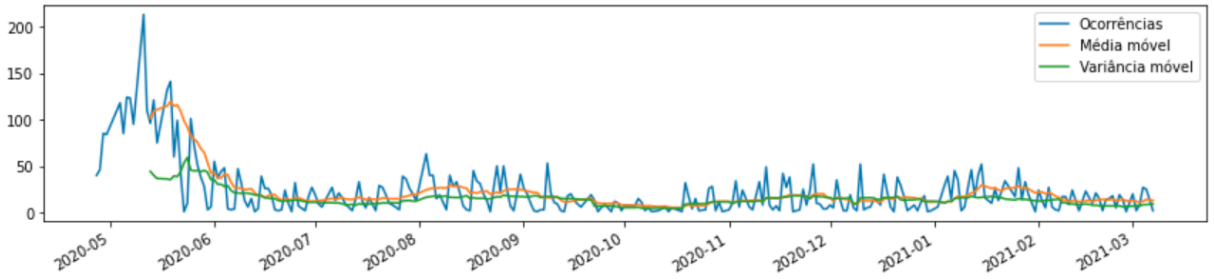
Abaixo seguem os resultados para todos os CIDs/CIAPs analisados (apenas entraram na análise os CIDs/CIAPs que apresentaram comportamento anômalo):

Figura 20 – Apresentação da média e variância móvel do CID U07.2



Fonte: Elaborada pelo autor (2021)

Figura 21 – Apresentação da média e variância móvel do CID J11

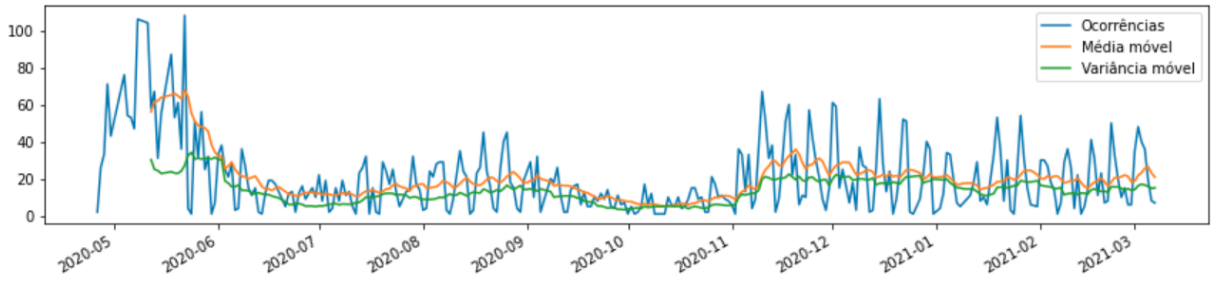


Fonte: Elaborada pelo autor (2021)

As imagens dos gráficos gerados para a identificação do comportamento da série não trouxeram muita clareza para a tomada de decisão para os CIDs J11, R74 e R83 mas, utilizando o teste ADF, obtivemos resultados muito interessantes.

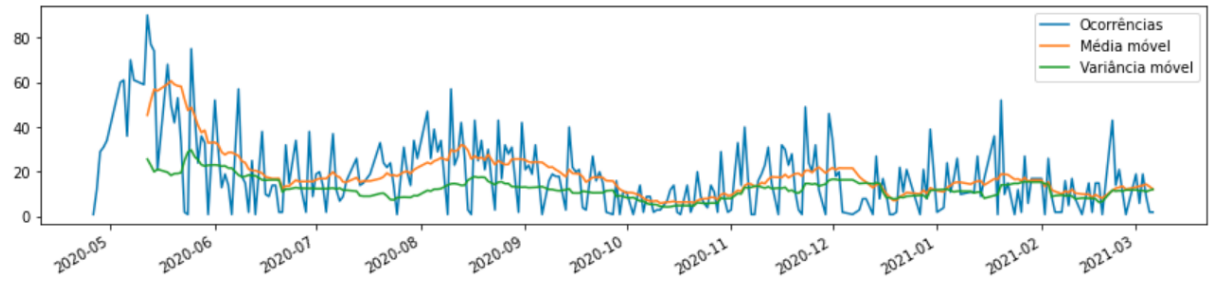
Conforme os resultados do teste ADF apresentados na Tabela 3, foi possível observar que os CIDs U 07.2, J11 e o CID CIAP R80 possuem p-value superior a 0,05 logo, aceitando a hipótese nula. Com isso consideramos que a série para esses CIDs não são estacionárias. Para

Figura 22 – Apresentação da média e variância móvel do CIAP R80



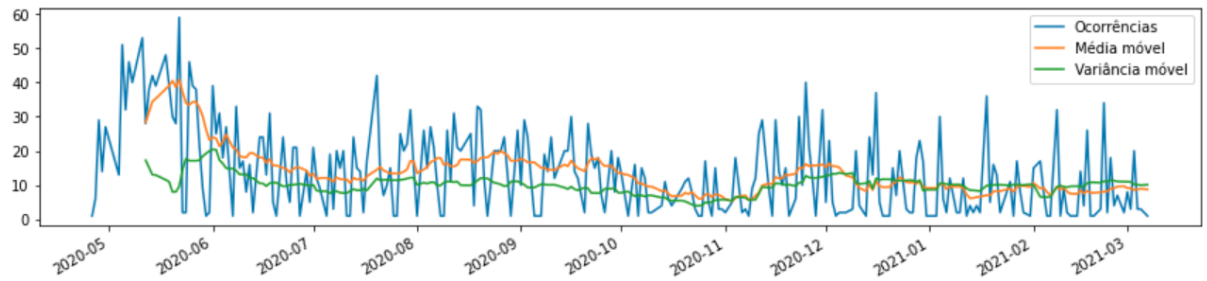
Fonte: Elaborada pelo autor (2021)

Figura 23 – Apresentação da média e variância móvel do CIAP R74



Fonte: Elaborada pelo autor (2021)

Figura 24 – Apresentação da média e variância móvel do CIAP R83



Fonte: Elaborada pelo autor (2021)

Tabela 3 – Teste de estacionariedade utilizando Dickey-Fuller Aumentado (ADF) nos 5 CIDs/CIAPs analisados

CID/CIAP	ADF estatístico	p-value
CID U 07.2	-2,14	0,22
CID J11	-7,45	5,39
CIAP R80	-2,45	0,12
CIAP R74	-3,45	0,009
CIAP R83	-2,96	0,038

Fonte: Elaborada pelo autor (2021)

os CIDs CIAPs R74 e R83 ambos apresentaram o p-value inferior a 0,05, assim, a hipótese nula foi rejeitada e as séries foram consideradas estacionárias.

Para identificarmos se os dados tendem a uma distribuição normal para fins de possível remoção de *outliers*, se fez necessário o uso do teste *Shapiro-Wilk* publicado em 1965. O teste de *Shapiro-Wilk* é baseado na correlação entre os dados e os escores normais correspondentes (PEAT; BARTON, 2008), e fornece melhor poder do que outros testes (STEINSKOG; TJøSTHEIM; KVAMSTø, 2007), sendo recomendado como uma escolha assertiva para testar a normalidade dos dados (THODE, 2002).

O teste de *Shapiro-Wilk* é usado para calcular uma estatística W que testa se uma amostra aleatória, x1, x2,. . . , xn vem (especificamente) de uma distribuição normal. Pequenos valores de W são evidências de desvio da normalidade e pontos percentuais para a estatística W, são obtidos por meio de simulações de Monte Carlo (PEAT; BARTON, 2008). Para considerarmos se a amostra é ou não uma distribuição normal, as seguintes hipóteses foram consideradas:

H0: Os dados seguem a distribuição normal.

H1: Os dados não seguem a distribuição normal.

Logo, o p-value abaixo de 0,05 faz com que rejeitemos a hipotética nula, assim a amostra não segue uma distribuição normal.

Abaixo seguem os resultados para cada CID/CIAP:

Como é possível observar na Tabela 4, todos os CIDs/CIAPs tiveram p-values abaixo de 0,05, logo é certo afirmar que nenhum deles possui a distribuição normal. Com isso, técnicas de remoção de ruído, que usem a distribuição normal, são descartadas para a nossa amostra de dados. Desta forma, utilizaremos os dados em sua completude para a construção dos modelos.

Tabela 4 – Resultado do teste de normalidade *Shapiro-Wilk*

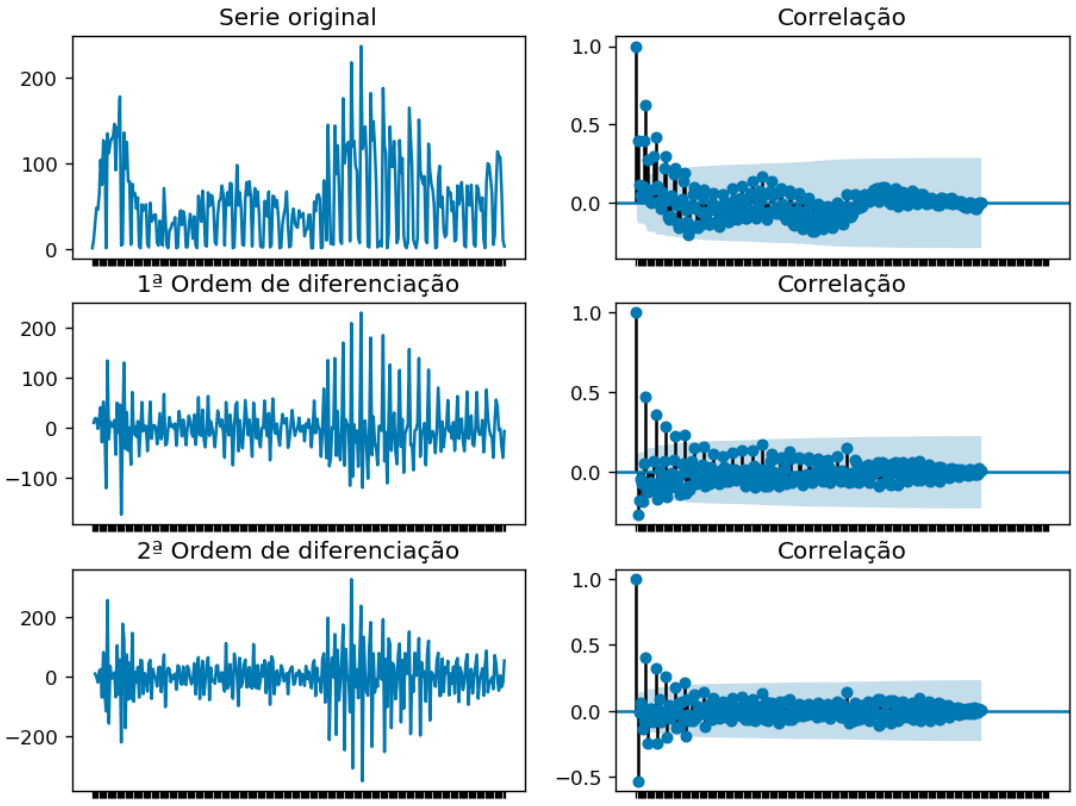
CID/CIAP	P-value
CIAP R80	0,000
CIAP R83	0,000
CID J11	0,000
CID U 07.2	0,000
CIAP R74	0,000

Fonte: Elaborada pelo autor (2021)

6.3 ARIMA: EXECUÇÃO E RESULTADOS

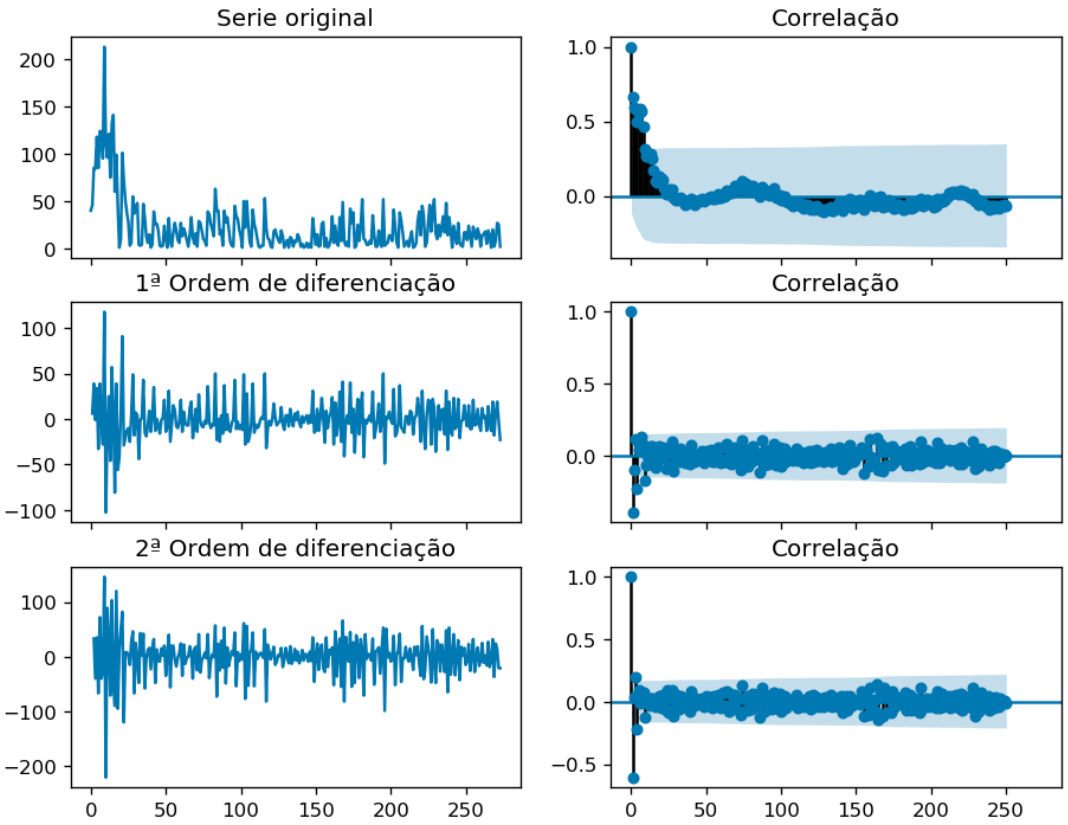
Devido aos resultados apresentados na Seção 6.2, foi preciso calcular o número mínimo de diferenciação necessário para tornar a série temporal estacionária, ver Seção 4.2.2, para os CIDs U07.2, J11 e CIAP R80. Seguem os resultados da autocorrelação:

Figura 25 – CID U07.2: Diferenciação e Correlação



As Figuras 25, 26, 27,28 e 29 mostram que a série temporal atinge a estacionariedade com duas ordens de diferenciação. Mas, ao olhar para o gráfico de autocorrelação para a 2ª diferenciação, o *lag* vai para a zona negativa distante bem rapidamente, o que indica que a

Figura 26 – CID J11: Diferenciação e Correlação



Fonte: Elaborada pelo autor (2021)

série pode ter sido diferenciada demais. Portanto, assumimos a ordem de diferenciação como 1, embora a série não seja perfeitamente estacionária.

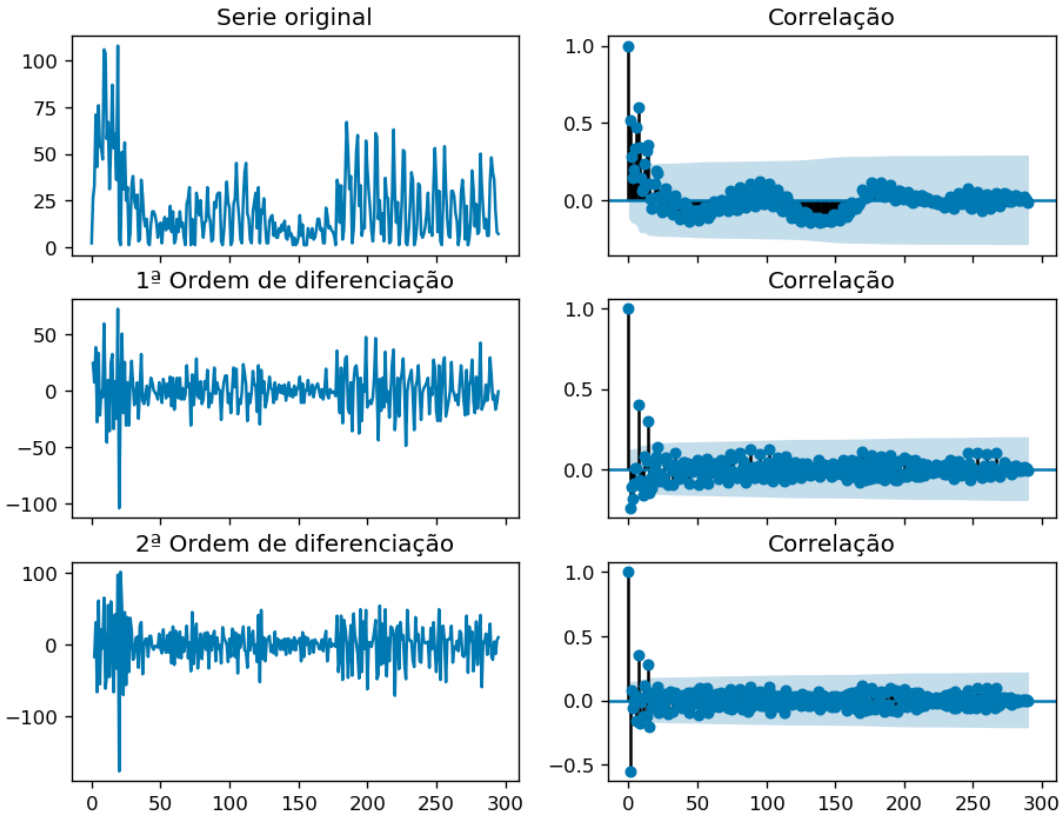
Para a validação deste resultado, fizemos uma análise combinatória nos valores dos parâmetros do ARIMA para cada CID/CIAP. Além da validação do parâmetro d, nosso intuito também foi de encontrar os melhores parâmetros para AR e MA, sem a necessidade de gerar os gráficos de autocorrelação parcial e autocorrelação. Para tal, usamos uma análise combinatória de parâmetros utilizando o teste ADF para identificarmos as melhores combinações que apresentassem os menores valores de AIC. A Tabela 5 apresenta os melhores resultados.

Tabela 5 – Resultados dos melhores parâmetros para o ARIMA

CID/CIAP	Parâmetros	AIC
CID U 7.2	ARIMA(2,1,3)	3027,035
CIAP RJ11	ARIMA(2,1,3)	2374,685
CIAP R80	ARIMA(2,1,3)	2400,519

Fonte: Elaborada pelo autor (2021)

Figura 27 – CIAP R80: Diferenciação e Correlação



Fonte: Elaborada pelo autor (2021)

Conforme a Tabela 5 acima, todas as séries apresentam os mesmo valores para p, d e q dos parâmetros do ARIMA.

Para validar a qualidade dos modelos gerados pelos parâmetros, segue tabela dos valores de erro percentual médio absoluto (MAPE), precisão direcional média (MDA) e desvio médio Quadrado(MSD) para cada CID/CIAP.

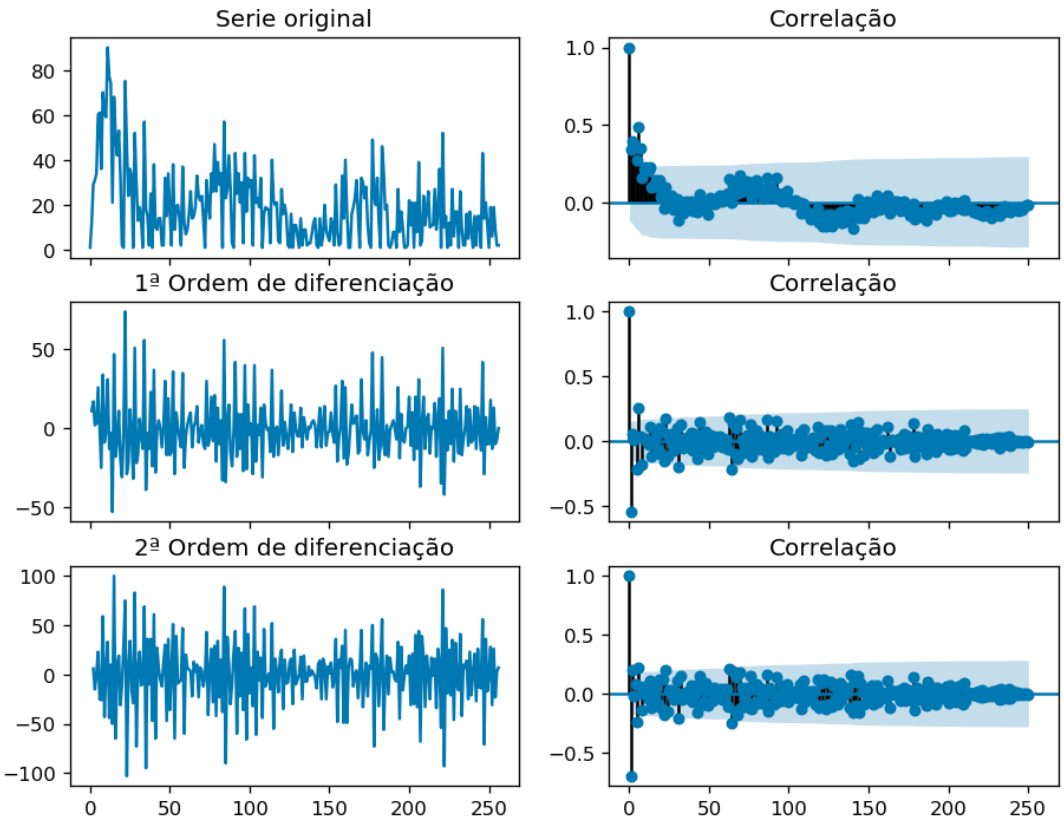
Tabela 6 – Acurácia do modelo

CID/CIAP	MAPE	MDA	MSD
CID U 7.2	1,94	0,66	1 271,71
CID J11	3,41	0,66	116,71
CIAP R80	0,55	0,83	121,06

Fonte: Elaborada pelo autor (2021)

A Tabela 6 mostra que, para os ajustes no modelo do CID U7.2, erram em menos de 2% dos dados com desvio padrão da média de 0,66 dias, e apresentam uma quantidade significativa de *outliers*. Para o CID J11, temos um erro de menos de 4% dos dados com um desvio padrão

Figura 28 – CID R74: Diferenciação e Correlação



Fonte: Elaborada pelo autor (2021)

da média de 0,66 dias, apresentando poucos *outliers*. Finalizando, temos o CIAP R80 que apresentou erro em menos 1% dos dados, com um desvio padrão da média em apenas 0,83.

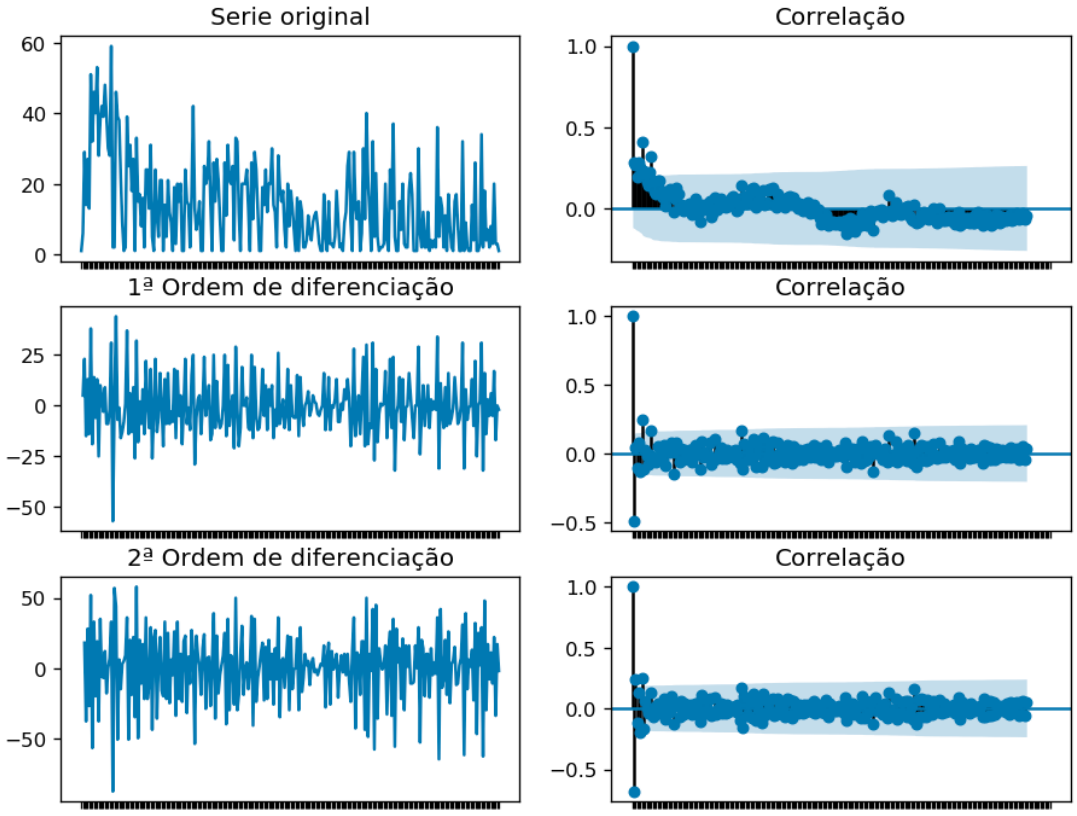
Para finalizar a nossa análise nos modelos gerados, para os CIDs/CIAPs com séries não estacionárias, observamos que o *p-value* para cada parâmetro de AR e MA usados estão abaixo de 0,05, o que implica que os parâmetros realmente são significativos nos modelos, como mostram as Tabelas 7, 8 e 9.

Tabela 7 – Estimativa dos parâmetros AR e MA CID para o CID U 07.2

Tipo	Coef	Std err	z	P
AR(1)	1,21	0.022	56,127	0.000
AR(2)	-0,976	0.020	-48,925	0.000
MA(1)	-1,997	0.067	-30,013	0.000
MA(2)	1,828	0.118	15,475	0.000
MA(3)	-0,708	0.061	-11,553	0.000

Fonte: Elaborada pelo autor (2021)

Figura 29 – CIAP R83: Diferenciação e Correlação



Fonte: Elaborada pelo autor (2021)

Tabela 8 – Estimativa dos parâmetros AR e MA para o CID J11

Tipo	Coef	Std err	z	P
AR(1)	1,213	0,079	15,400	0,000
AR(2)	-0,773	0,107	-7,207	0,000
MA(1)	-1,901	0,106	-18,024	0,000
MA(2)	1,474	0,151	9,791	0,000
MA(3)	-0,374	0,104	-3,604	0,000

Fonte: Elaborada pelo autor (2021)

As Figuras 30, 31, 32, 33 e 34 exibem os gráficos residuais para os CIDs/CIAPs estudados. Um ligeiro desvio dos resíduos da linha reta pode ser observado nos gráficos. Isso indica que os erros estão um pouco próximos do normal, com alguns valores discrepantes. Portanto, a suposição de normalidade é seguida. O histograma residual confirma essa suposição. O gráfico entre os resíduos e os valores ajustados exibe uma pequena dispersão. Isso implica que a suposição de variância constante também é satisfeita pelo modelo.

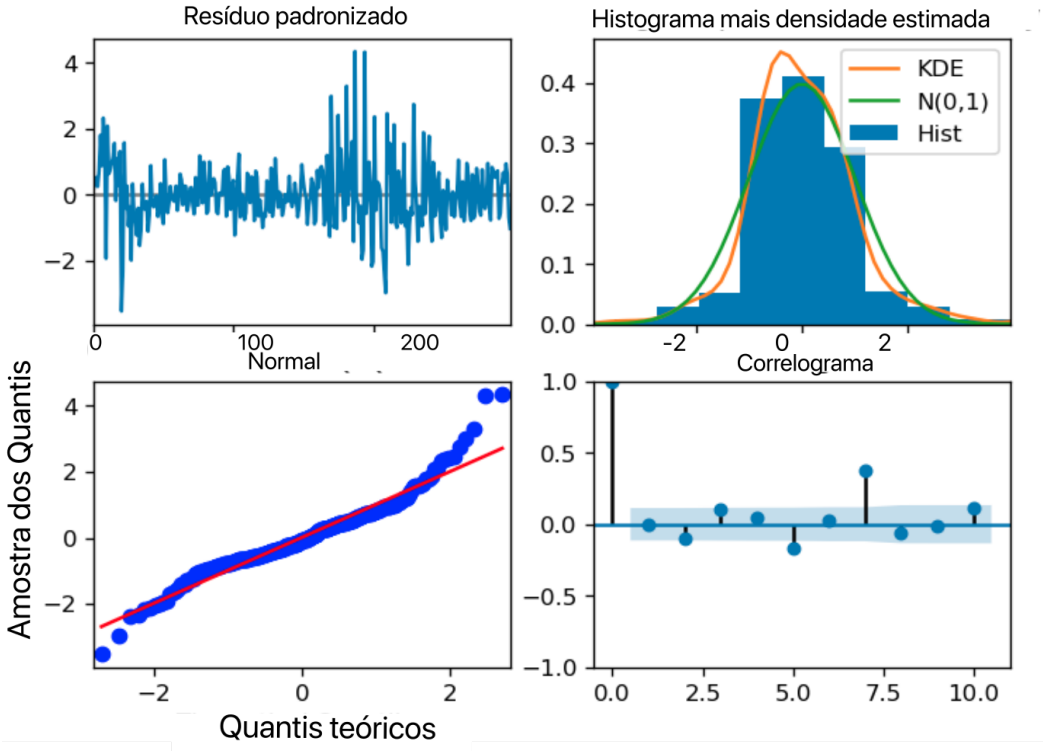
Os gráficos residuais apresentaram um bom comportamento, consolidando que os parâme-

Tabela 9 – Estimativa dos parâmetros AR e MA para o CIAP R80

Tipo	Coef	Std err	z	P
AR(1)	1,180	0,029	41,085	0,000
AR(2)	-0,937	0,030	-31,643	0,000
MA(1)	-1,907	0,060	-31,618	0,000
MA(2)	1,615	0,097	16,695	0,000
MA(3)	-0,541	0,063	-8,606	0,000

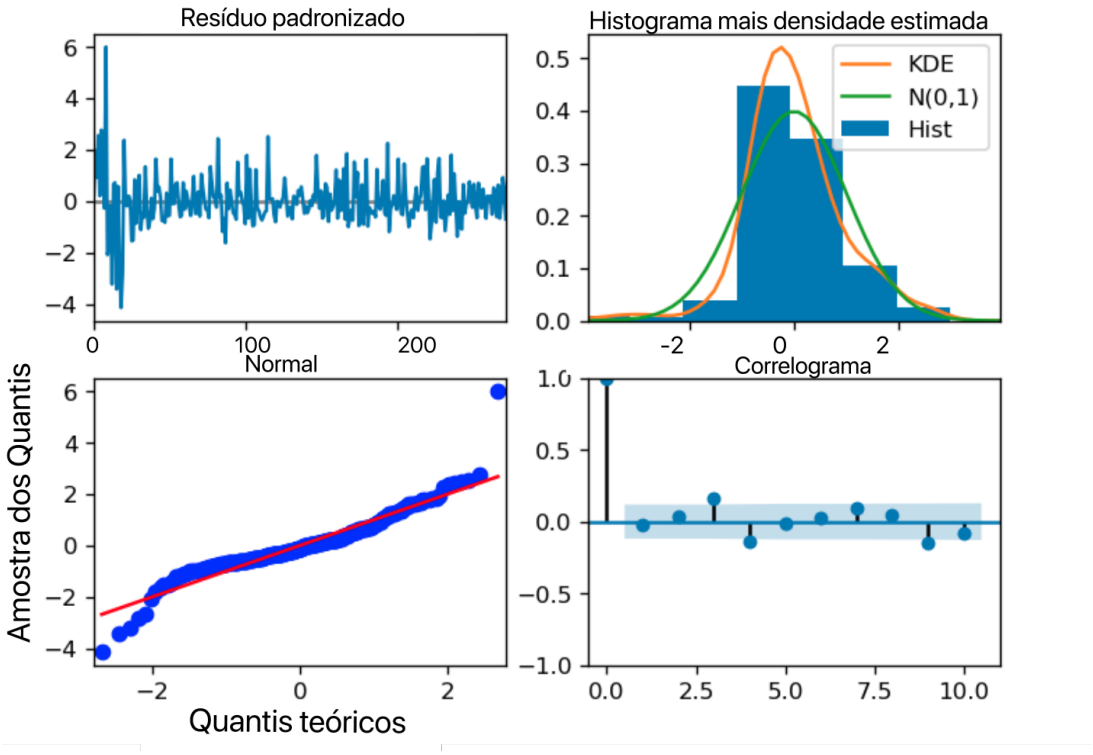
Fonte: Elaborada pelo autor (2021)

Figura 30 – CID U07.2: Diagnóstico para avaliação dos parâmetros



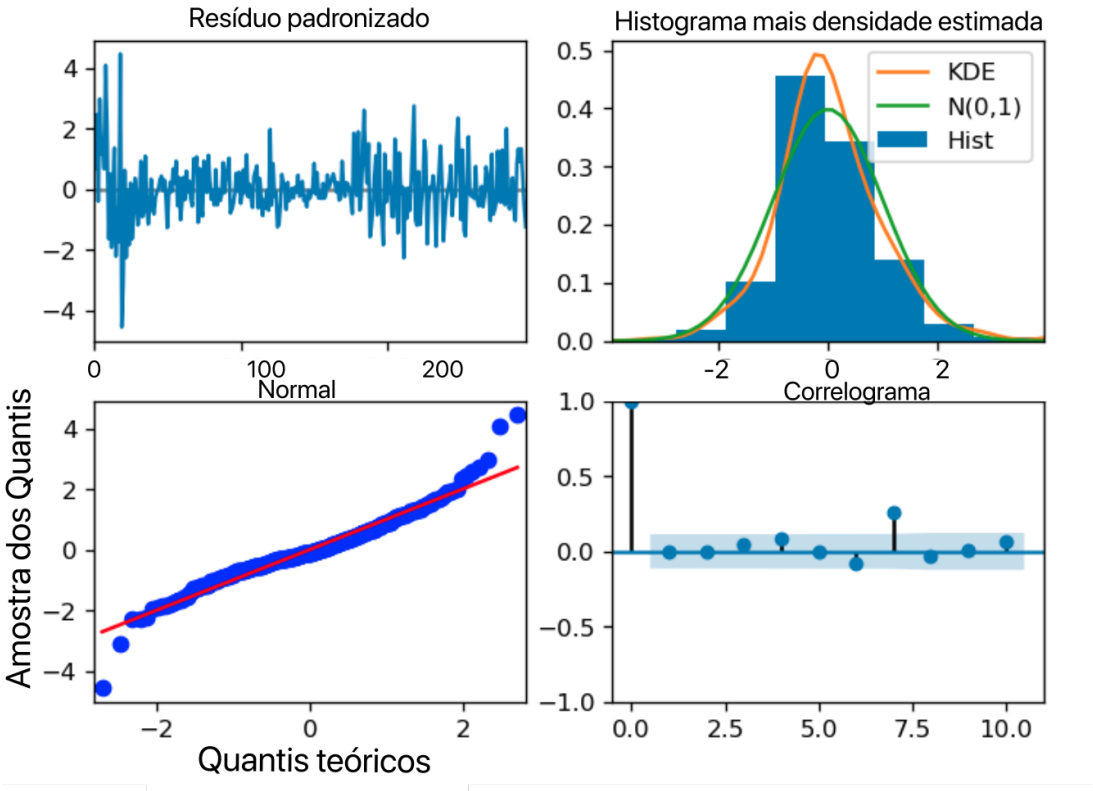
Fonte: Elaborada pelo autor (2021)

Figura 31 – CID J11: Diagnóstico de resíduos avaliação dos parâmetros para o ARIMA



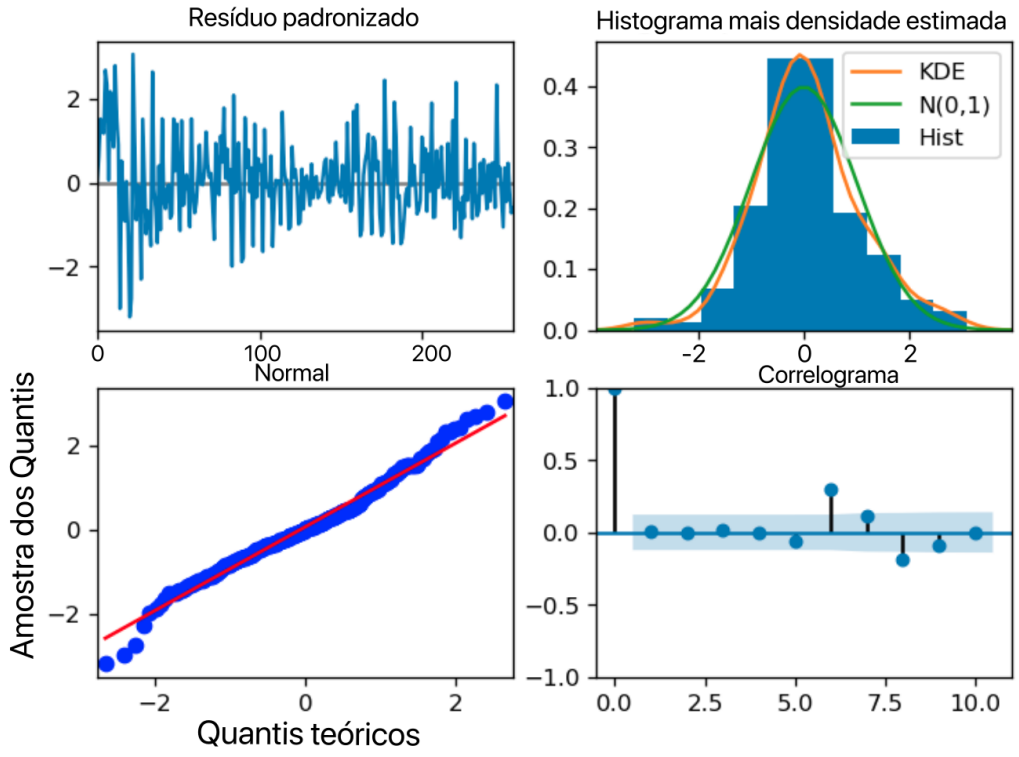
Fonte: Elaborada pelo autor (2021)

Figura 32 – CIAP R80: Diagnóstico de resíduos avaliação dos parâmetros para o ARIMA



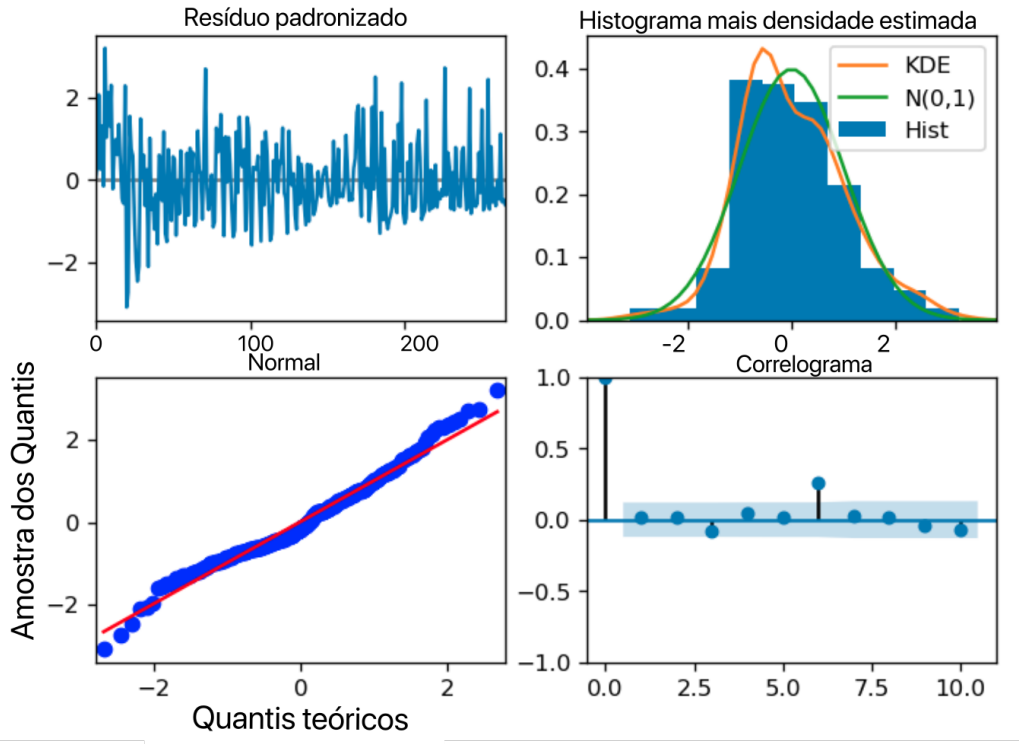
Fonte: Elaborada pelo autor (2021)

Figura 33 – CIAP R74: Diagnóstico de resíduos avaliação dos parâmetros para o ARIMA



Fonte: Elaborada pelo autor (2021)

Figura 34 – CIAP R83: Diagnóstico de resíduos avaliação dos parâmetros para o ARIMA



Fonte: Elaborada pelo autor (2021)

tros para os modelos são significativos.

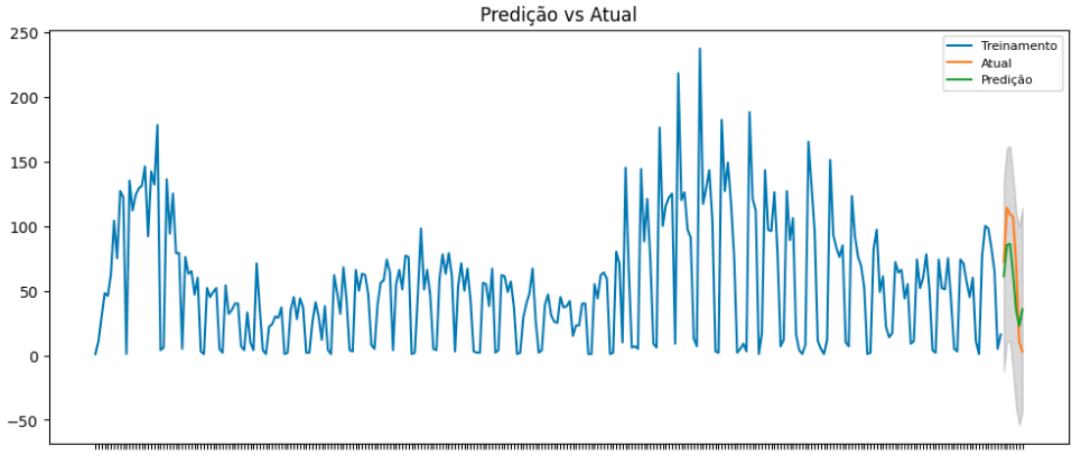
Para finalizar, as Figuras abaixo apresentam os gráficos gerados para validar o modelo, apresentando o fluxo de casos atuais em relação aos preditos para cada CID/CIAP anômalo. Além das Figuras, a Tabela 10 apresenta o acumulado de casos previstos para sete dias.

Tabela 10 – ARIMA - Número de casos acumulados para sete dias

CID/CIAP	Ocorrências
CID U 07.2	446
CID J11	85
CIAP R80	172
CIAP R74	54
CIAP R83	49

Fonte: Elaborada pelo autor (2021)

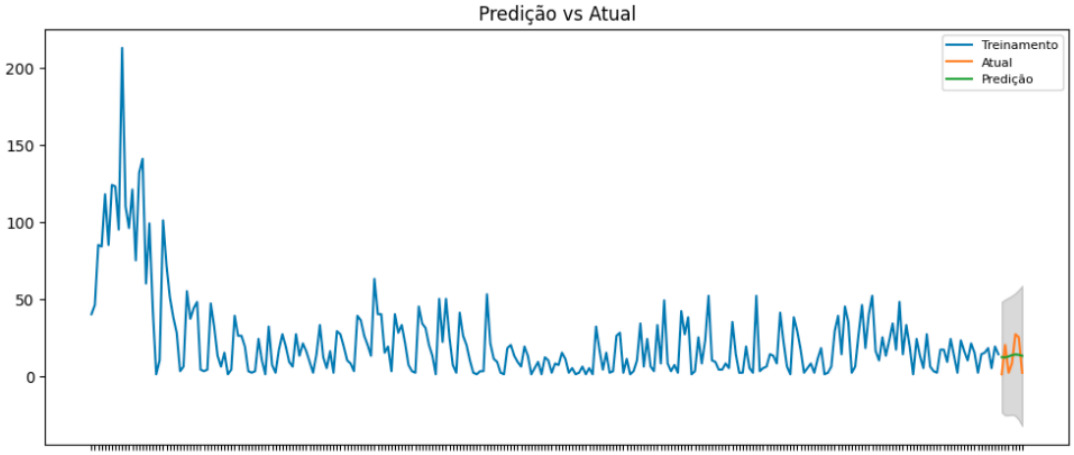
Figura 35 – ARIMA - CID U07.2: Análise do desempenho das ocorrências da predição vs atual



Fonte: Elaborada pelo autor (2021)

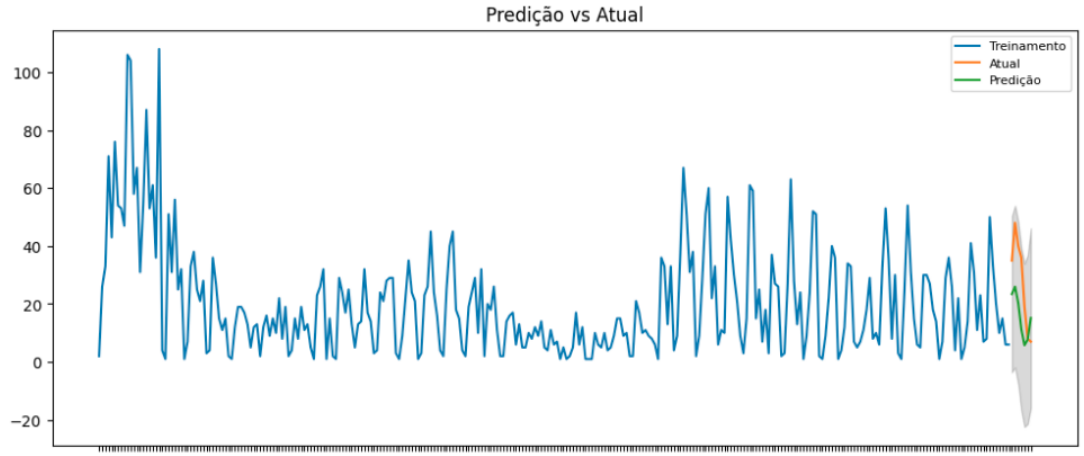
Como é possível observar nas Figuras 35, 36, 37, 38 e 39, todos os resultados de predição estão muito próximos dos dados reais, mostrando que o estudo realizado e os modelos gerados se apresentam satisfatórios para a predição dos CIDs/CIAPs.

Figura 36 – ARIMA - CID J11: Análise do desempenho das ocorrências da predição vs atual



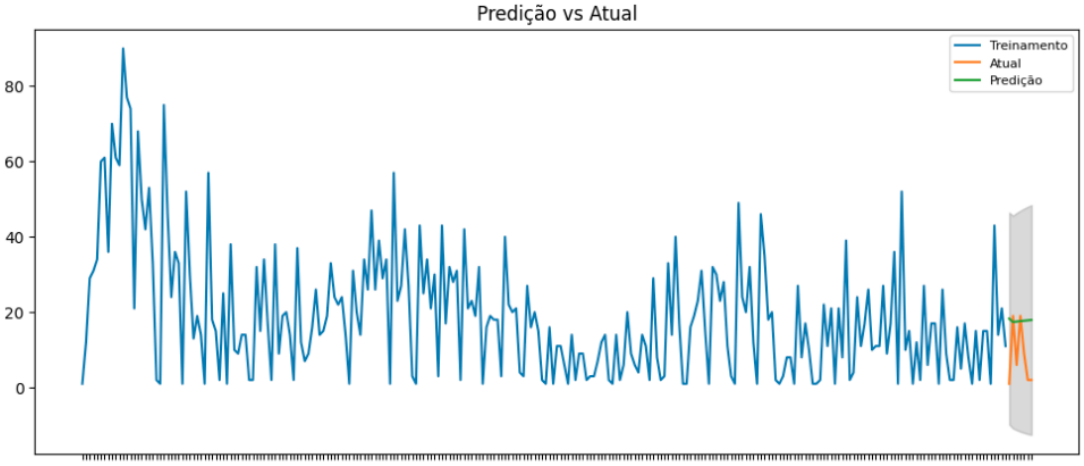
Fonte: Elaborada pelo autor (2021)

Figura 37 – ARIMA - CIAP R80: Análise do desempenho das ocorrências da predição vs atual



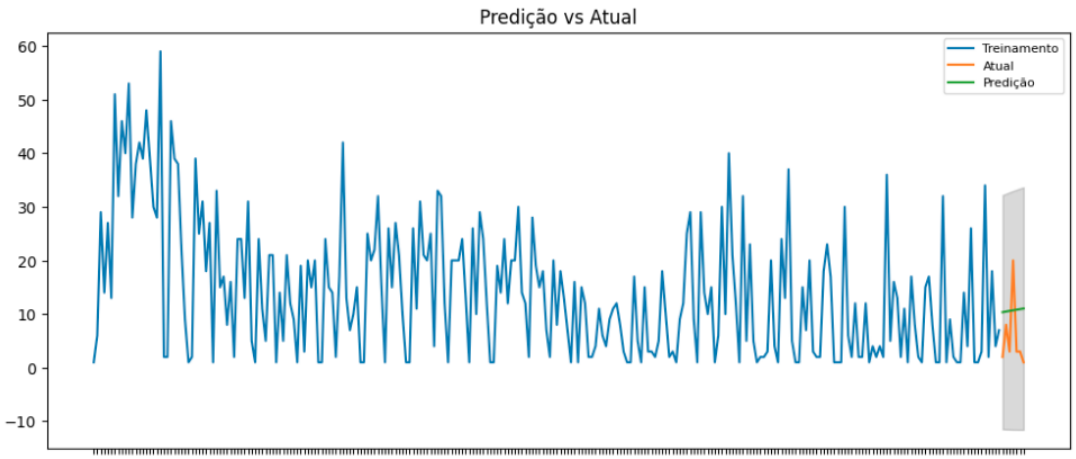
Fonte: Elaborada pelo autor (2021)

Figura 38 – ARIMA - CIAP R74: Análise do desempenho das ocorrências da predição vs atual



Fonte: Elaborada pelo autor (2021)

Figura 39 – ARIMA - CIAP R83: Análise do desempenho das ocorrências da predição vs atual



Fonte: Elaborada pelo autor (2021)

6.4 LSTM: EXECUÇÃO E RESULTADOS

Para a execução do treinamento da série usando o LSTM, faz-se necessário encontrar os melhores hiperparâmetros que consigam, após o treinamento, obter a menor taxa de erro. A metodologia de *gradient search* geralmente não é uma operação indicada para algoritmos de *deep learning*, devido à quantidade de dados utilizados para o seu treinamento. Para os casos em que os conjuntos de dados são menores, como séries temporais univariadas, é possível usar tal técnica para ajustar os hiperparâmetros de um modelo. Como já conhecemos as séries de cada CIDs, ver Seção 4.1, também se fez necessário analisar o número de observações, para assim, definirmos a quantidade ou período que iria compor os conjuntos de treinamento e teste. Na tabela 11, seguem as observações:

Tabela 11 – Número de observações para cada CID/CIAP

CID/CIAP	Ocorrências
U 07.2	300
R 80	296
J11	274
R 83	266
R 74	257

Fonte: Elaborada pelo autor (2021)

Sabemos que o intervalo dos dados é de quase 12 meses, ver Seção 4.1. Logo a ocorrência de tais CIDs/CIAPS são baseadas nas consultas realizadas pelos profissionais de saúde, e isso

não implica que o mesmo CID/CIAP deva ocorrer diariamente. Devido a tal fato, há uma certa diferença entre as observações. Outro fator importante é que esses são os CIDs/CIAPs anômalos, fruto de nossa pesquisa, e como explicado anteriormente, ver Capítulo 4, esses são os que tiveram o maior número de ocorrências. Com isso, o nosso conjunto de testes foi criado com 30 observações para cada CID/CIAP, e as observações restantes foram utilizadas para o conjunto de treinamento.

Para ser possível encontrarmos os melhores parâmetros, utilizando a técnica de *gread search* foi necessário definirmos, como método de avaliação, a abordagem *walk-forward*. O *walk-forward* é uma abordagem em que o modelo faz uma previsão para cada observação no conjunto de dados de teste. Depois de cada previsão, uma observação é adicionada ao conjunto de dados de teste e disponibilizada para o modelo. Os modelos mais simples podem ser reajustados com a observação, antes de se efetuar a previsão subsequente. Já os modelos mais complexos, como redes neurais, não são reajustados devido ao custo computacional. No entanto, as observações adicionadas para cada etapa de tempo, podem então ser usadas como parte da entrada para as previsões na próxima etapa de tempo. Com isso, avaliamos as configurações do modelo várias vezes, por meio da *walk-forward*.

Outro fator importante para a melhoria do modelo, foi calcular a diferenciação para os CIDs/CIAPs não estacionários. Assim, construímos um *Grid Search Multilayer Perceptron* ajustando apenas 5 parâmetros:

- *n_input*: O número de entradas anteriores para usar como entrada para o modelo;
- *n_nodes*: O número de nós a serem usados na camada oculta;
- *n_epochs*: O número de períodos de treinamento;
- *n_batch*: O número de amostras a serem incluídas em cada mini lote;
- *n_diff*: A ordem da diferença.

Como resultado, na escolha dos melhores hiperparâmetros, foi calculada a média para os valores de RMSE. Abaixo, nas Tabelas 12, 13, 14, 15 e 16, segue o resultado dos melhores hiperparâmetros e seus respectivos RMSE.

Com os hiperparâmetros devidamente definidos, obtivemos os resultados de previsão para sete dias. As Figuras abaixo mostram os resultados obtidos pelos modelos, e a Figura 43 contém os resultados acumulados das predições para cada CID e CIAP.

Tabela 12 – LSTM: Hiperparâmetros para o CID U 07.2

Parâmetro	Valor
n_input	7
n_nodes	50
n_epochs	7
n_batch	60
n_diff	7

Fonte: Elaborada pelo autor (2021)

Tabela 13 – LSTM: Hiperparâmetros para o CID R74

Parâmetro	Valor
n_input	12
n_nodes	150
n_epochs	30
n_batch	90
n_diff	0

Fonte: Elaborada pelo autor (2021)

Tabela 14 – LSTM: Hiperparâmetros para o CIAP R80

Parâmetro	Valor
n_input	12
n_nodes	30
n_epochs	30
n_batch	60
n_diff	0

Fonte: Elaborada pelo autor (2021)

Tabela 15 – LSTM: Hiperparâmetros para o CIAP R83

Parâmetro	Valor
n_input	7
n_nodes	50
n_epochs	100
n_batch	90
n_diff	0

Fonte: Elaborada pelo autor (2021)

Tabela 16 – LSTM: Hiperparâmetros para o CID J11

Parâmetro	Valor
n_input	7
n_nodes	50
n_epochs	100
n_batch	90
n_diff	0

Fonte: Elaborada pelo autor (2021)

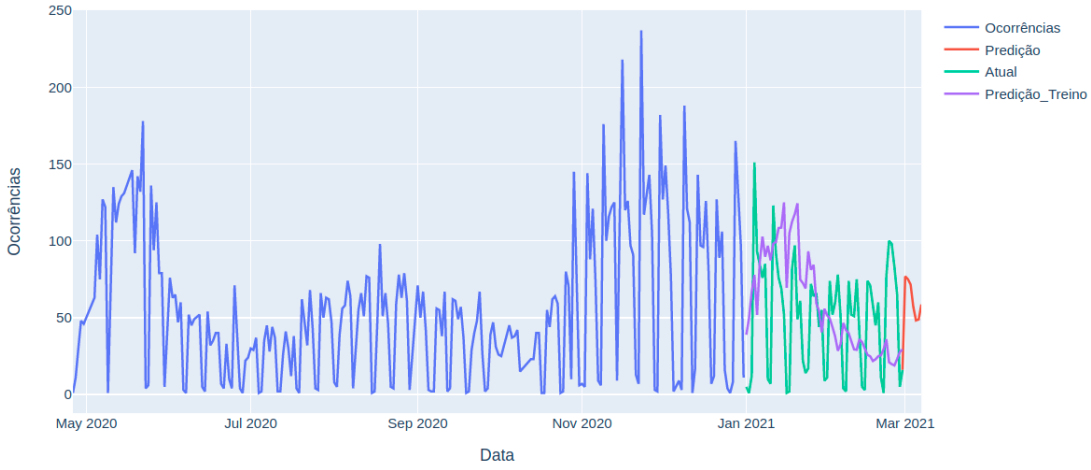
Tabela 17 – LSTM - Número de casos acumulados para sete dias

CID/CIAP	Ocorrências
CID U 07.2	570
CID J11	90
CIAP R80	214
CIAP R74	54
CIAP R83	72

Fonte: Elaborada pelo autor (2021)

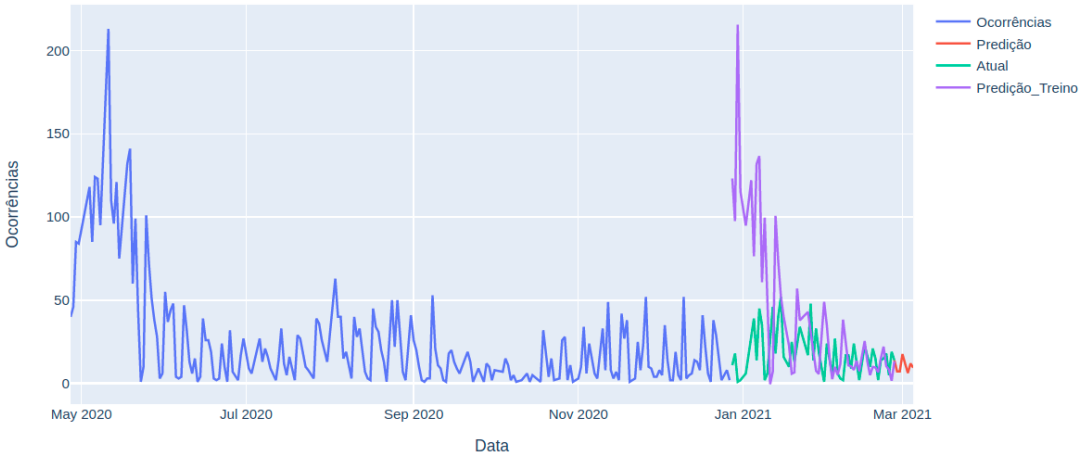
Conforme as Figuras 40, 41, 42, 43 e 44, as linhas de previsão para todos os CIDs e CIAPs apresentam direções similares aos sete dias anteriores. Na Seção 6.6, iremos apresentar os valores da base de validação, que utilizaremos como referência para os sete dias previstos da tabela 17 de casos acumulados.

Figura 40 – LSTM - CID U07.2: Análise do desempenho das ocorrências da predição vs atual



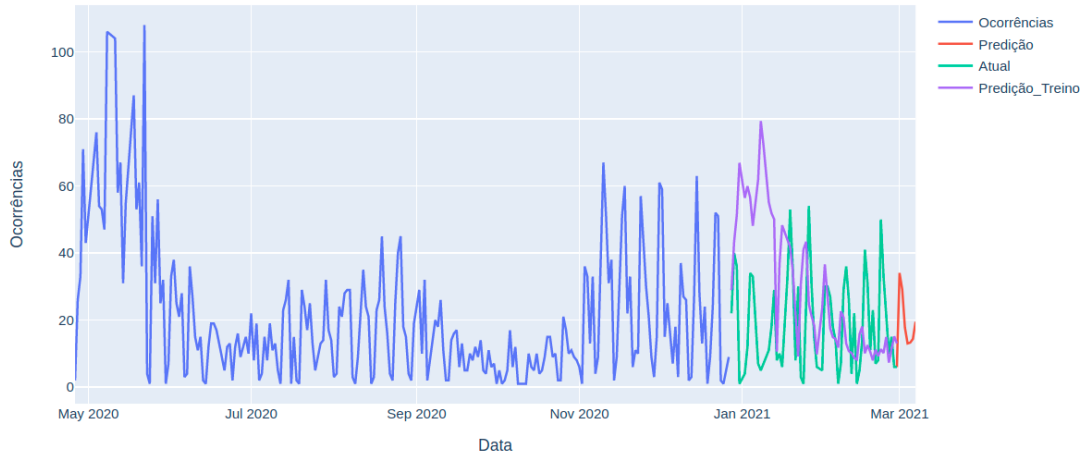
Fonte: Elaborada pelo autor (2021)

Figura 41 – LSTM - CID J11: Análise do desempenho das ocorrências da predição vs atual



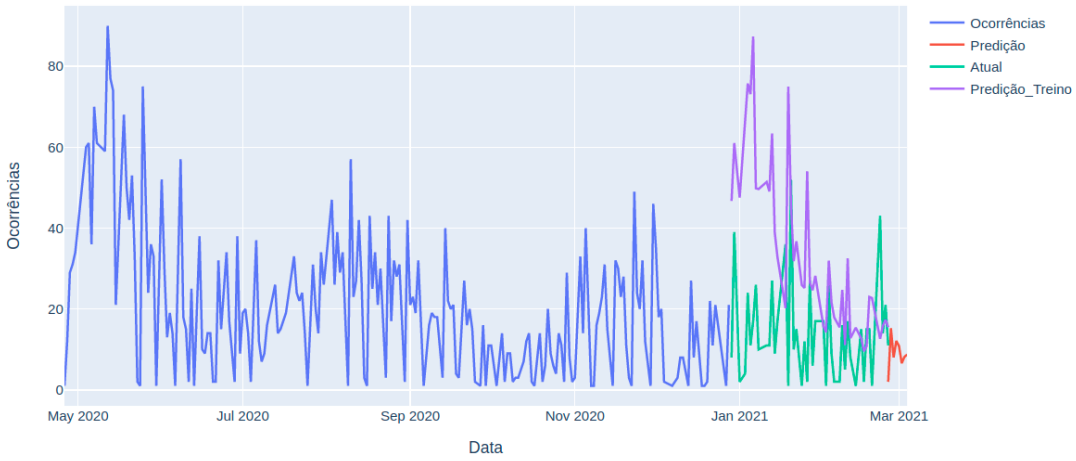
Fonte: Elaborada pelo autor (2021)

Figura 42 – LSTM - CIAP R80: Análise do desempenho das ocorrências da predição vs atual



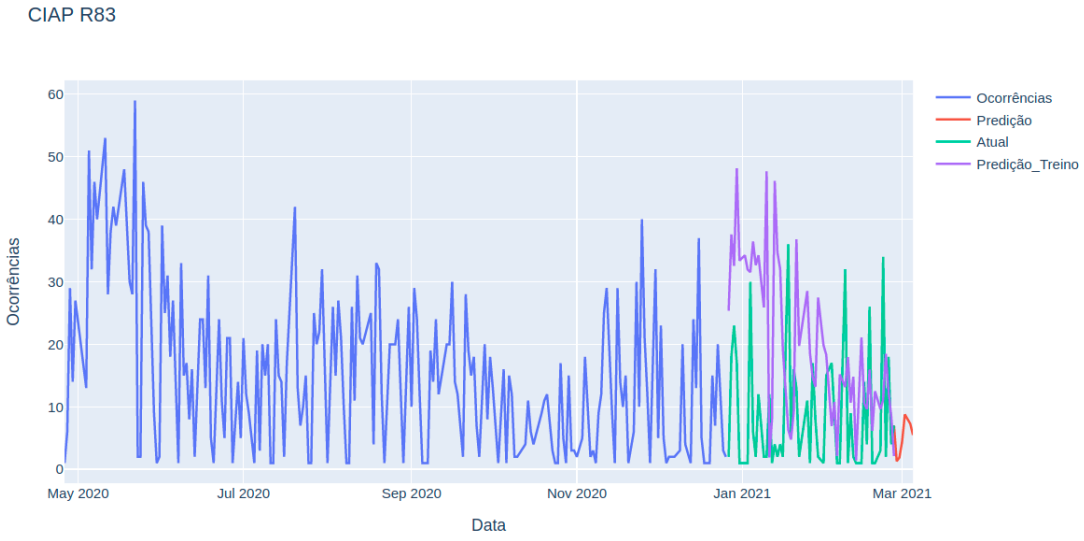
Fonte: Elaborada pelo autor (2021)

Figura 43 – LSTM - CIAP R74: Análise do desempenho das ocorrências da predição vs atual



Fonte: Elaborada pelo autor (2021)

Figura 44 – LSTM - CIAP R83: Análise do desempenho das ocorrências da predição vs atual



Fonte: Elaborada pelo autor (2021)

6.5 *PROPHET*: EXECUÇÃO E RESULTADOS

Na análise exploratória dos dados, foi possível identificar que as séries continham *Missing not at Random* (MNAR) (RUBIN, 1976), uma vez que, no início do uso do sistema atende APS, não havia atendimentos aos finais de semanas e feriados. Para mitigar este problema e obtermos melhores resultados no *Prophet*, calculamos a média geral da semana. Este é um método de imputação muito básico, mas eficiente, visto que as características das séries e a relação entre as variáveis não seriam alteradas.

O comportamento do *Prophet* em relação aos hiperparâmetros é um pouco diferente dos demais algoritmos utilizados neste artigo: não houve a necessidade de criarmos '*loops*' ou '*ranges*' de valores em um *Grid Search*, ou *Random Search* para otimizá-los devido à sua objetividade. Abaixo segue a lista de hiperparâmetros do *Prophet*, divididos em dois grupos: O grupo 1, composto por hiperparâmetros que poderiam ser ajustados, e o grupo 2, contendo os que não precisam ser ajustados.

Grupo 1:

- ***changepoint_prior_scale***: Existe para indicar o quão flexíveis os pontos de mudança podem ser (i.e, mudanças abruptas de tendência na série). Em outras palavras, quanto os pontos de mudança podem caber nos dados. Se você aumentar, será mais flexível, mas você pode acabar superdimensionando. Se a tendência for pequena, será insuficiente e a variância que deveria ter sido modelada, com mudanças de tendência, acabará sendo tratada com o termo de ruído; se for muito grande, a tendência se ajustará demais e, no caso mais extremo, você pode acabar com a tendência capturando a sazonalidade anual;

- ***seasonality_prior_scale***: Este parâmetro controla a flexibilidade da sazonalidade. Da mesma forma, um valor grande permite que a sazonalidade se ajuste a grandes flutuações, um valor pequeno reduz a magnitude da sazonalidade;

- ***holidays_prior_scale***: Controla a flexibilidade para ajustar os efeitos dos feriados;

- seasonality_mode: Este parâmetro indica como seus componentes de sazonalidade devem ser integrados às previsões, podendo ser Aditivo ou Multiplicativo.

Grupo 2:

- ***growth***: Possui as opções 'lineares' e 'logísticas'. Este hiperparâmetro é selecionado conforme o comportamento da série; se houver um ponto de saturação conhecido e crescimento em direção a esse ponto, ele será incluído e será utilizada a tendência logística, caso contrário, será linear.

- ***changepoints***: este hiperparâmetro é utilizado para especificar manualmente os locais dos pontos de mudança.

- ***n_changepoints***: este é o número de pontos de mudança colocados automaticamente.

- ***yearly_seasonality***: por padrão ('automático'), isso ativará a sazonalidade anual se houver um ano de dados e desativará, no caso contrário. As opções são [Auto, Verdadeiro e Falso].

- ***weekly_seasonality***: O mesmo que para *yearly_seasonality*.

- ***daily_seasonality***: O mesmo que para *yearly_seasonality*.

- ***holidays***: isso é para passar em *um data frame* de feriados especificados. Os efeitos do feriado seriam ajustados no *holidays_prior_scale*.

- ***mcmc_samples***: este parâmetro determina se o modelo usa estimativa a posteriori máxima (MAP) ou uma inferência Bayesiana completa, com o número especificado de amostras Monte Carlo de Cadeia de Markov (MCMC), para treinar e prever. Se o MCMC é usado ou não, provavelmente será determinado por fatores como a duração da série temporal e a importância da incerteza do parâmetro.

- ***interval_width***: a predição do *Prophet* retorna intervalos de incerteza para cada componente, como *yhat_lowere* e *yhat_upper* para a previsão *yhat*. Estes são calculados como quantis da distribuição preditiva posterior e o *interval_width* especifica quais quantis usar. O padrão de 0,8 fornece um intervalo de predição de 80%. A mudança deste hiperparâmetro afetará apenas o intervalo de incerteza e não mudará a previsão, portanto, não precisa ser ajustado.

- ***uncertainty_samples***: os intervalos de incerteza são calculados como quantis do intervalo preditivo posterior, e o intervalo preditivo posterior é estimado com amostragem Monte Carlo. Este parâmetro é o número de amostras a serem usadas (o padrão é 1000). O tempo de execução para previsão será linear neste número. Torná-lo menor aumentará,

a variância (erro de Monte Carlo) do intervalo de incerteza, e torná-lo maior reduzirá essa variância. Portanto, se as estimativas de incerteza parecem irregulares, isso pode ser aumentado para suavizá-las ainda mais, mas provavelmente não precisará ser alterado. Tal como acontece com *interval_width*, este parâmetro afeta apenas os intervalos de incerteza e alterá-lo não afetará a previsão.

Mediante tais explicações acima, o algoritmo *Prophet* adicionou os hiperparâmetros padrões, como mostra a Tabela 18 abaixo:

Tabela 18 – *Prophet* - Hiperparâmetros utilizados no modelo

Hiperparâmetros	**Valor**
n_changepoints	25
growth	Linear
changepoint.range	0.8
yearly.seasonality	*TRUE*
weekly.seasonality	*auto*
daily.seasonality	*TRUE*
holidays	*NULL*
seasonality.mode	*Additive*
seasonality.prior.scale	10
changepoint.prior.scale	0.05
holidays.prior.scale	10
mcmc.samples	0
interval.width	0.95
uncertainty.samples	1000
specified.changepoints	*FALSE*

Fonte: Elaborada pelo autor (2021)

Esses hiperparâmetros foram utilizados para todos os CIDs. Como é possível observar, aumentamos o valor do *interval.width* para 95%, aumentando a margem de erro e consequentemente aumentando intervalo.

Na Seção 5.4, analisamos cada um dos conjuntos de dados e foi possível observar que os CIDs U07.2, J11 e R80 são não-sazonais; logo, para obtermos valores significativos para a sua predição, precisaríamos utilizar a diferenciação ou alguma função não-linear no hiperparâmetro *growth*. O *Prophet* permite que você realize previsões usando um modelo de tendência de crescimento logístico, mas se faz necessário definir uma capacidade de carga específica. Com isso, deveríamos adicionar, em nosso conjunto de dados, uma coluna "cap", que receberia a

capacidade de suporte para cada linha. Como há uma variabilidade significativa de aumento ou redução de um dado número de casos para uma doença infecciosa, que está ligado à taxa de transmissibilidade R0, decidimos apenas utilizar o crescimento linear, para não gerar valores de capacidade de suporte sem fundamentação acadêmica.

Após gerarmos a previsão para o número cumulativo de casos para os próximos 7 dias, observou-se que alguns resultados se apresentaram negativos e os resultados negativos não se adequam à nossa proposta. Para mitigarmos tal fato, usamos a abordagem *trend model*, devido ao uso do crescimento linear no nosso modelo. No algoritmo *Prophet* não há nada que impeça que a tendência se torne negativa: para tal, fixá-la em zero elimina toda a incerteza da tendência futura. Segundo (TAYLOR; LETHAM, 2017) a incerteza de tendência é estimada com a amostragem de Monte Carlo, amostrando tendências futuras com a seguinte simulação: em cada momento futuro, verifique se haverá ou não uma mudança de tendência de uma distribuição de Poisson (cuja taxa é estimada durante o ajuste do modelo), se houver uma mudança de tendência, faça uma amostra da magnitude da mudança de tendência de uma distribuição de *Laplace* (cuja escala é estimada durante o ajuste do modelo), atualize a tendência com essa mudança e continue avançando no tempo.

Desta forma, modificamos os modelos para não permitir mudanças de tendência que as tornem negativas. Para melhor compreensão, imagine simular uma tendência futura, quando ela atingir 0 e começar a ficar negativa, adicione uma nova mudança de tendência que a mantenha em 0. Com isso, as mudanças de tendência futuras positivas ainda conseguirão tornar-se uma tendência positiva novamente, nunca se deslocando para abaixo de 0.

Após os ajustes acima citados, obtivemos os seguintes resultados para os modelos, como mostra a Tabela 19:

Tabela 19 – *Prophet*: Média das métricas da previsão do modelo dos sete dias

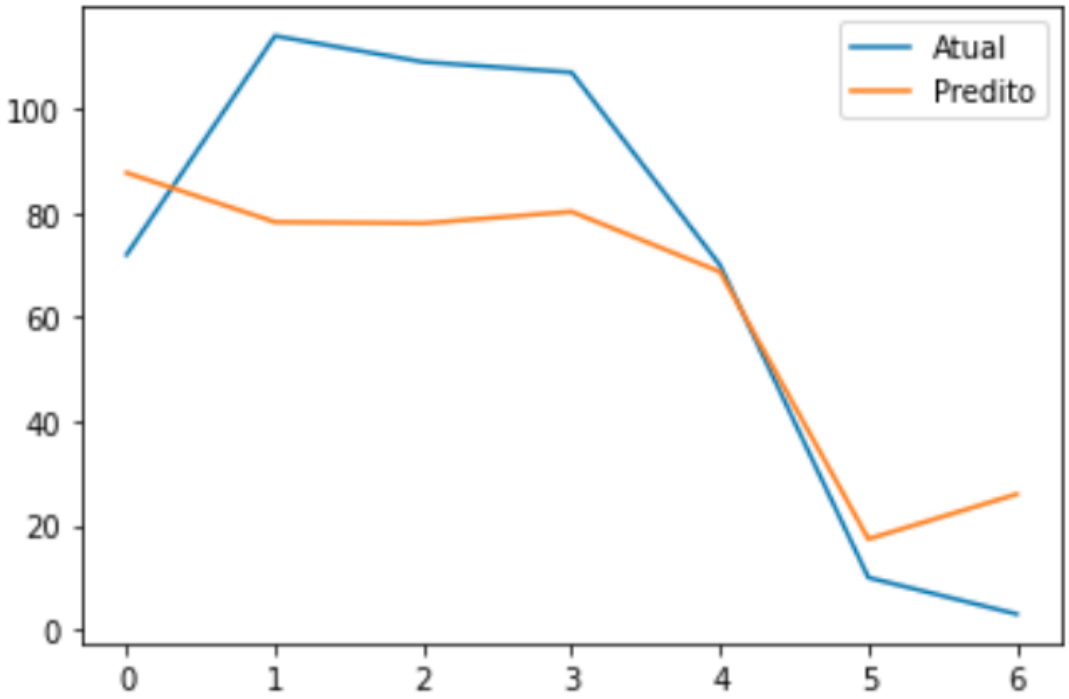
CID/CIAP	MSE	RMSE	MAE	MAPE	MAPE
CID U07.2	1755,31	40,89	28,89	4,32	1,05
CID J11	185,11	13,14	10,66	1,79	0,60
CIAP R80	174,24	12,75	10,28	1,82	0,53
CIAP R74	116,81	10,51	8,74	2,06	0,68
CIAP R83	57,28	6,82	5,72	1,94	1,09

Fonte: Elaborada pelo autor (2021)

Para a obtenção dos resultados, utilizamos a validação cruzada para o intervalo de sete dias e ao final calculamos a média para cada métrica de erro individualmente por CIDs/CIAPs.

Vemos que o CIAP R83 obteve os menores erros *MSE*, *RMSE* e MAE, sendo que CID 07.5 acumulou o maior RMSE, ou seja, o U07.2 apresentou, considerando o acumulado, maior discrepância entre a observação obtida em relação aos dados reais. As Figuras 45, 46, 47, 48 e 49, apresentam gráficos que correlacionam os números de casos preditos em relação aos números de casos atuais.

Figura 45 – *Prophet* - CID U07.2: Análise do desempenho das ocorrências da predição vs atual

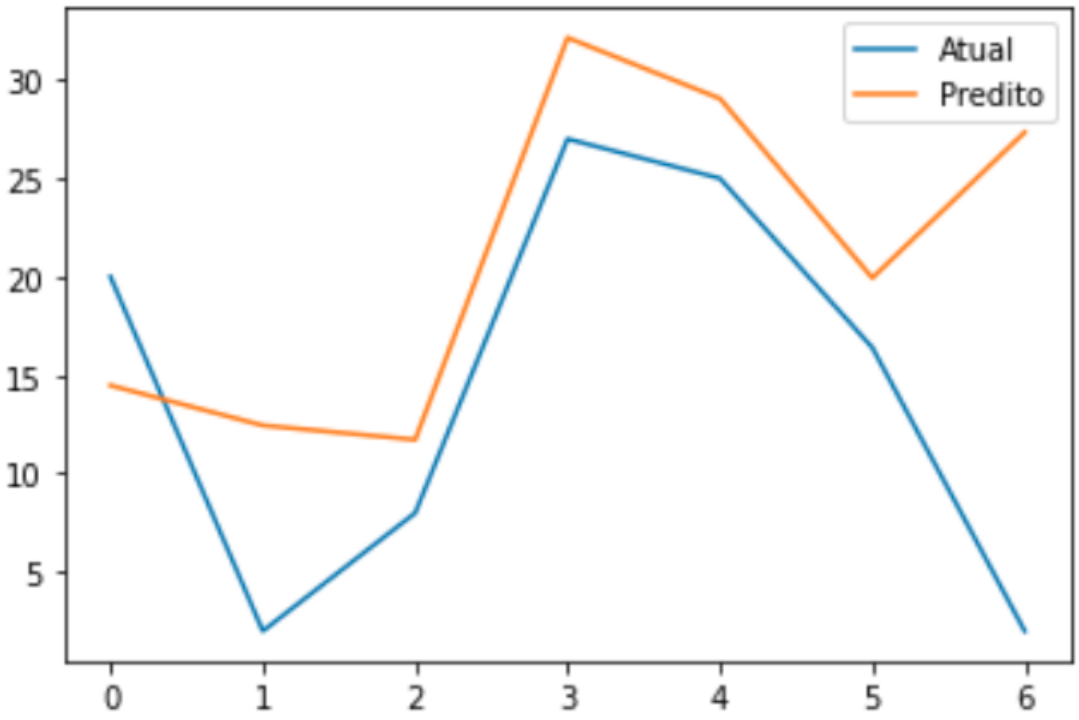


Fonte: Elaborada pelo autor (2021)

Conforme as Figuras 45, 46, 47, 48, 49 é possível observar que no CID U07.2 e no CIAP R80, a relação da previsão com os números observados possui a mesma tendência de direção, tornando essas previsões satisfatórias. Na Tabela 20, apresentamos o somatório dos resultados para a previsão de sete dias, como nos algoritmos anteriores.

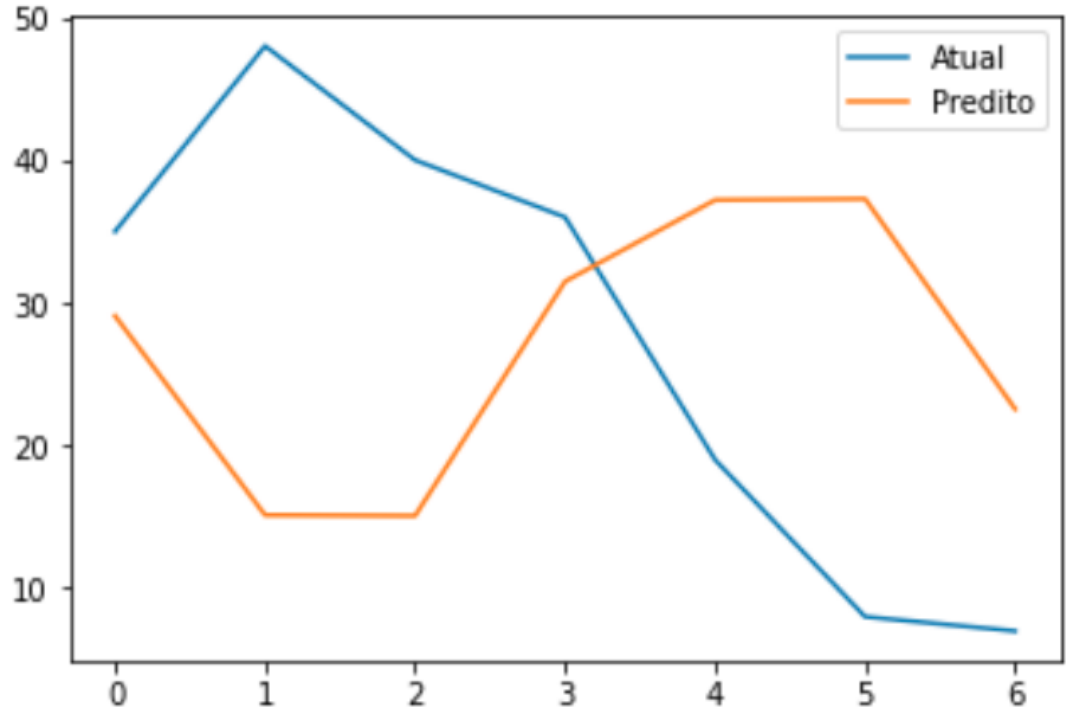
Com esses resultados de somatório, as características do algoritmo *Prophet* se mostraram ser de fácil adaptação para o problema proposto, conseguindo, sem muitas alterações nos seus hiperparâmetros, alcançar resultados próximos à realidade.

Figura 46 – *Prophet* - CID J11: Análise do desempenho das ocorrências da predição vs atual



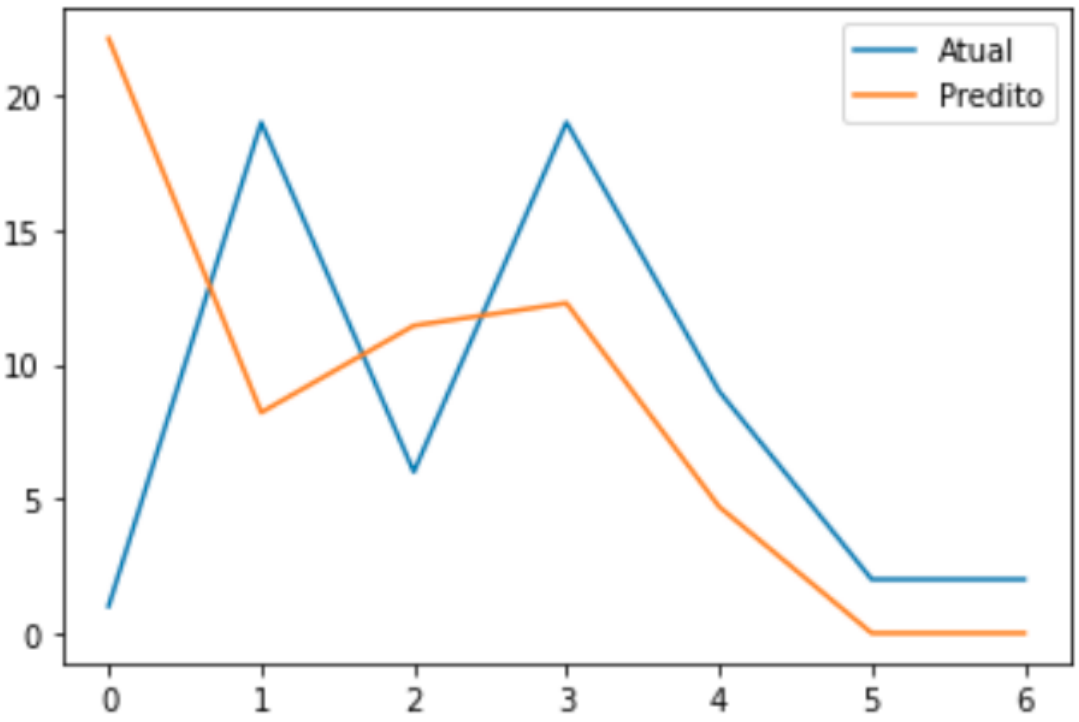
Fonte: Elaborada pelo autor (2021)

Figura 47 – *Prophet* - CIAP R80: Análise do desempenho das ocorrências da predição vs atual



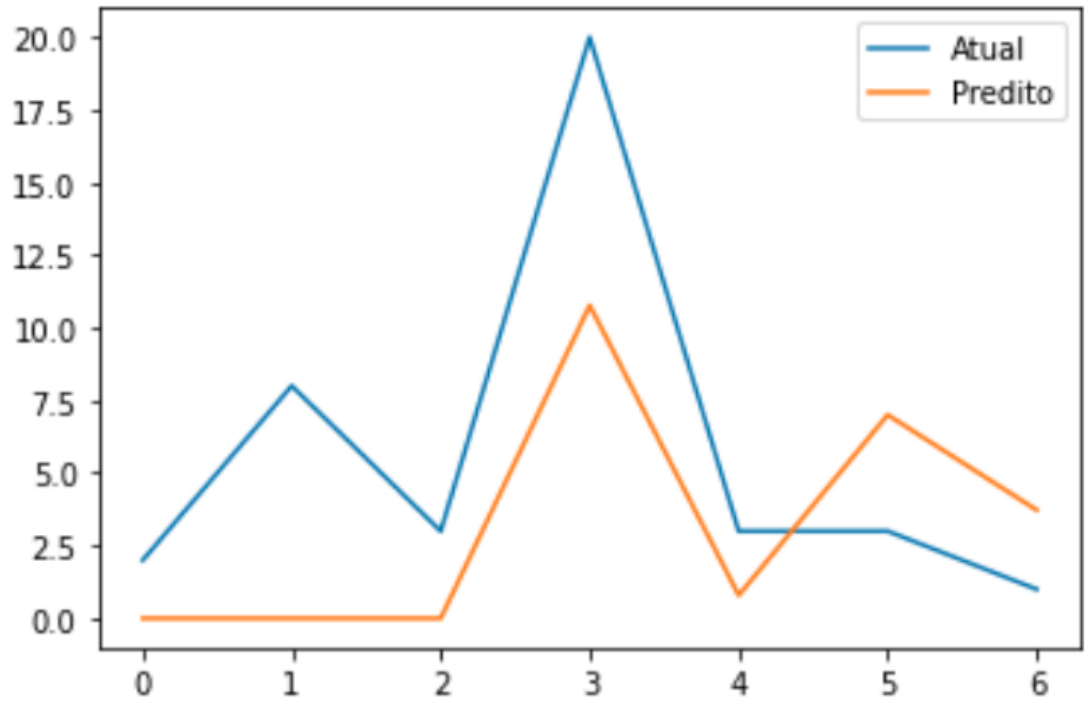
Fonte: Elaborada pelo autor (2021)

Figura 48 – *Prophet* - CIAP R74: Análise do desempenho das ocorrências da predição vs atual



Fonte: Elaborada pelo autor (2021)

Figura 49 – *Prophet* - CIAP R83: Análise do desempenho das ocorrências da predição vs atual



Fonte: Elaborada pelo autor (2021)

Tabela 20 – *Prophet* - Número de casos acumulados para sete dias

CID/CIAP	Ocorrências
CID U07.2	436
CID J11	127
CIAP R80	146
CIAP R74	63
CIAP R83	43

Fonte: Elaborada pelo autor (2021)

6.6 RESULTADOS GERAIS

Nas seções anteriores descrevemos os resultados acumulados alcançados individualmente em cada CID e CIAP, nos algoritmos testados. As Tabelas 21 e 22 com os resultados de previsão para cada um dos CIDs e CIAPs individualmente por técnica utilizada, e os valores observados.

Observando os totais acumulados para cada CID e CIAP e os resultados dos algoritmos, vemos que o ARIMA obteve resultados muito próximos do observado. Analisando com mais

Tabela 21 – Resultados acumulados utilizando as técnicas

CID/CIAP	ARIMA	LSTM	*Prophet*
CID U 07.2	446	570	436
CID J11	85	90	127
CIAP R80	172	214	146
CIAP R74	54	54	63
CIAP R83	49	74	43

Fonte: Elaborada pelo autor (2021)

Tabela 22 – Total acumulado de ocorrências observadas para cada CID/CIAP

CID/CIAP	Total acumulado observado
CID U 07.2	485
CID J11	83
CIAP R80	193
CIAP R74	58
CIAP R83	40

Fonte: Elaborada pelo autor (2021)

cuidado, é possível observar que para o CIAP R83, o *Prophet* obteve resultado mais próximo, já no CIAP R80 o ARIMA e o LSTM obtiveram resultados próximos, com diferenças iguais de 21 casos, um para menos e outro para mais. No CIAP R74, o ARIMA e o LSTM obtiveram os mesmos resultados. Nos demais, o ARIMA obteve um melhor desempenho.

Portanto, pudemos observar que cada algoritmo possui sua particularidade e que o uso de séries temporais para prever o comportamento de doenças infecciosas, para um pequeno intervalo de tempo, possui um ótimo desempenho. Para validar a nossa proposta de apresentarmos aos gestores de saúde os resultados obtidos pelos 3 algoritmos, submetemos os resultados a uma análise par a par, por meio do teste estatístico de Mann-Whitney, indicado quando se deseja comparar se dois grupos, não pareados, pertencem ou não à mesma população. Em resumo, verificamos se as evidências comprovam diferença estatística entre os resultados apresentados pelos algoritmos, em relação aos dados observados. A Tabela 23, contém os valores dos p-values para cada um dos algoritmos.

Tabela 23 – Resultados do algoritmo Mann-Whitney

Modelo	**p-value**
ARIMA	0,50
LSTM	0,33
Prophet	0,50

Fonte: Elaborada pelo autor (2021)

Como é possível observar, não há diferença estatística entre os resultados dos modelos em comparação com os valores observados, logo, todos os algoritmos podem ser utilizados para agregar valor nas tomadas de decisão dos gestores de saúde pública, provendo assim, visões diferentes para o enriquecimento do conhecimento.

7 CONCLUSÃO

Este capítulo trata do encerramento deste trabalho. Nele, os objetivos serão revistos, para discutirmos se foram devidamente alcançados, e as considerações finais com conclusões, tanto do ponto de vista prático, quanto do metodológico serão abordados. Além disso, as oportunidades identificadas para embasar novas pesquisas e trabalhos serão apresentadas.

7.1 CONSIDERAÇÕES FINAIS

Conforme mencionado no Capítulo 1, o objetivo central deste trabalho foi analisar o crescimento no número de casos, considerando os CIDs/CIAPs identificados como anômalos e, com isso, gerar visões inteligentes, disponibilizadas por meio de dashboards, para ajudar na tomada de decisão, de forma proativa, dos gestores de saúde pública. O nosso intuito não foi construir um processo engessado, que fosse útil apenas para os tipos de CIDs/CIAPs disponíveis para a construção desta pesquisa, mas sim abrangente, capaz de ser utilizado para qualquer doença que possua um alto grau de transmissibilidade, como as apresentadas no Capítulo 2.

Para isso ser possível, nos apegamos na seguinte frase, "nem todos os problemas de predição podem ser resolvidos com os mesmos procedimentos" (TAYLOR; LETHAM, 2017). Para tal, fomos em busca de algoritmos de séries temporais que estivessem consolidadas na academia e obtivessem aplicações em estudos epidemiológicos. Com isso, encontramos os algoritmos ARIMA, LSTM e Prophet. O uso desses algoritmos, em nosso estudo, não foi para escolher os que obtivessem resultados mais próximos dos dados observados, e em seguida descartarmos os restantes, mas sim, oferecer diferentes visões aos gestores, desde que os algoritmos obtivessem relações estatisticamente comprovadas.

Para chegarmos até os resultados, a presente pesquisa contextualizou toda a problemática envolvida no tocante às doenças infectocontagiosas (ver Capítulo 2). A relação entre os fatores sociais e as dificuldades encontradas para se implementar medidas de prevenção e controle e, os motivos pelos quais as decisões dos gestores, de forma proativa, são importantes no combate a essas doenças. Com isso, também mostramos em qual esfera de desafios estamos inseridos, mostrando alguns indicadores da cidade do Recife, foco do nosso estudo, que comprovam o grau de necessidade de metodologias, como as que foram apresentadas nesta pesquisa.

Para alcançarmos os nossos objetivos, foi necessária a construção de um processo deno-

minado MAPDI. Este processo permite que os dados advindos das unidades de saúde possam ser coletados de forma automática, e com isso, identificar os CIDs/CIAPs que apresentem comportamento anômalos, reajustando assim, os modelos gerados pelos algoritmos de séries temporais, citados anteriormente, para prever o comportamento dessa doença. Com os dados das previsões alcançadas, é realizado um processo de indexação dos resultados em uma ferramenta de recuperação de informação e apresentados aos gestores, por meio de uma *interface* inteligente já utilizada na saúde do Recife.

Voltando um pouco para as análises dos resultados dos algoritmos, conseguimos demonstrar que as previsões realizadas não possuem diferenças estatísticas em relação aos dados observados. Com isso, é correto afirmar que, as previsões geradas pelos 3 algoritmos possuem resultados muito próximos dos dados observados. Já na análise individual dos algoritmos, o ARIMA obteve um resultado um pouco melhor do que os demais para alguns dos CIDs/CIAPS, para esse conjunto de dados em específico. Desta forma, comprovamos a nossa hipótese inicial, que o uso de mais de um algoritmo de série temporal para análise de dados dos CIDs/CIAPs produz melhores *insights* do que o uso de apenas uma técnica.

7.2 SUGESTÕES PARA PESQUISAS FUTURAS

Demonstramos que os dados apresentados na presente pesquisa são de suma importância para rastreamento de uma dada doença infectocontagiosa, com isso conseguindo prover informações que ajudarão a suprimir a transmissão. Consequentemente, equipes da saúde, equipes de desenvolvimento de diagnósticos e a população, necessitam de mecanismos capazes de estruturar outros dados e informações. Visando rastrear os contaminados para, com isso, prover serviços de análise estatística e o acompanhamento da evolução do quadro geral, nas diferentes regiões de uma determinada região, esses serviços enriquecerão ainda mais a tomada de decisão e as ações estratégicas em diferentes cenários.

Para tal, vemos a necessidade da elaboração de um protocolo de ações que devem ser tomadas para construção de um processo robusto para o combate de novas doenças com alto grau de transmissão. Assim, sugerimos, primeiramente, a elaboração das seguintes abordagens:

Modelos epidemiológicos adaptados em comparação aos modelos de séries temporais — Atualmente, há uma quantidade significativa de modelos epidemiológicos compartimentados na literatura (IVORRA et al., 2020) que vêm sendo fortemente usados para prever o número de casos, óbitos e hospitalizações por Covid-19. Vários desses modelos sofreram adaptações para

que os seus hiperparâmetros fossem mais significativos para o modus operandi da Covid-19. Com isso, sugerimos a comparação dos resultados da presente pesquisa com os resultados advindos de um modelo adaptado que considere a taxa de subnotificação, como o percentual de casos de pessoas assintomáticas.

Análise da transmissibilidade por meio de comunidades de multiagentes — Para este tema, iniciamos com alguns questionamentos: as doenças infectocontagiosas atingem mais as pessoas de grupos sociais vulneráveis? Os modelos compartimentados, comentados anteriormente, consideram que o crescimento da taxa de transmissão pode acontecer devido à falta de saneamento básico, fornecimento de água encanada, etc.? Quais são os estratos sociais de uma dado município e como ocorre o acesso à saúde pública? No processo de isolamento, quais são os períodos de maior desrespeito às medidas e por qual motivo? As medidas de isolamento social, horizontais e verticais, trazem efeitos benéficos em comunidades carentes? Os mais testados são os estratos mais vulneráveis ou os mais abastados? Quando um vírus de alta transmissibilidade surgir em um estrato menos vulnerável, como ele se propaga até os vulneráveis? Como definir a rota do vírus? Todos esses questionamentos, poderão ser respondidos por meio da análise do comportamento social para a sua modelagem computacional, baseada em agentes computacionais, utilizando modelos epidemiológicos compartimentados adaptados.

Rastreabilidade — O uso de aplicações móveis para rastreamento e monitoramento de atividades sociais tem sido bastante utilizado pelos grandes players como Google, Facebook e Apple. Essas empresas forneceram dados do comportamento social de pessoas em todo o mundo, visto que o aumento da mobilidade trouxe desafios para os estudos de segregação social e, em especial, no contexto de confinamento em residências e cuidados gerais com a saúde. Com isso, faz-se necessária a construção de mecanismos capazes de identificar as interações sociais de pessoas infectadas com as suscetíveis. Essas ações podem acontecer por meio de georreferenciamento dos indivíduos testados ou os que apresentarem sintomas, permitindo uma análise de interação social para, com isso, conseguirmos mapear zonas de risco, indivíduos assintomáticos e auxiliar no planejamento das ações estratégicas para possíveis isolamentos sociais.

Todas as propostas de pesquisas, transcritas acima, ajudarão a compor um meio eficiente para mitigar, de forma proativa e mais eficiente, as novas pandemias que poderão surgir fazendo assim com que os gestores possam visualizar e entender os riscos de forma rápida e precisa e, consequentemente, salvar mais vidas.

REFERÊNCIAS

ABATH, M. *Programa Sanar – Doenças Negligenciadas*. 2013. Disponível em: <http://portal.saude.pe.gov.br/programa/secretaria-executiva-de-vigilancia-em-saude/programa-sanar-doencas-negligenciadas>.

ABBOTT, S.; HELLEWELL, J.; THOMPSON, R. N.; SHERRATT, K.; GIBBS, H. P.; BOSSE, N. I.; MUNDAY, J. D.; MEAKIN, S.; DOUGHTY, E. L.; CHUN, J. Y.; CHAN, Y.-W. D.; FINGER, F.; CAMPBELL, P.; ENDO, A.; PEARSON, C. A. B.; GIMMA, A.; RUSSELL, T.; CMMID COVID modelling group; FLASCHE, S.; KUCHARSKI, A. J.; EGGO, R. M.; FUNK, S. Estimating the time-varying reproduction number of SARS-CoV-2 using national and subnational case counts. v. 5, p. 112, 2020. ISSN 2398-502X. Disponível em: <https://wellcomeopenresearch.org/articles/5-112/v2>.

ADHIKARI, R.; AGRAWAL, R. K. An introductory study on time series modeling and forecasting. 2013. Disponível em: <https://arxiv.org/abs/1302.6613v1>.

AKAIKE, H. A new look at the statistical model identification. v. 19, n. 6, p. 716–723, 1974. ISSN 1558-2523. Conference Name: IEEE Transactions on Automatic Control.

ANDERSON, R. M.; ANDERSON, B.; MAY, R. M. *Infectious Diseases of Humans: Dynamics and Control*. [S.l.]: OUP Oxford, 1992. ISBN 978-0-19-854040-3.

ANDERSON, R. M.; GRENFELL, B. T.; MAY, R. M. Oscillatory fluctuations in the incidence of infectious disease and the impact of vaccination: time series analysis. v. 93, n. 3, p. 587–608, 1984. ISSN 0022-1724. Publisher: Cambridge University Press. Disponível em: <https://www.cambridge.org/core/journals/epidemiology-and-infection/article/oscillatory-fluctuations-in-the-incidence-of-infectious-disease-and-the-impact-of-vaccination-time-series-a 4254E738C1D5B883B86DC148DBBEB82D>.

ANDO, S.; KURAKAMI, H. Decomposition of generalized asymmetry model for square contingency tables. v. 6, n. 3, p. 405–411, 2016. Number: 3 Publisher: Scientific Research Publishing. Disponível em: <http://www.scirp.org/Journal/Paperabs.aspx?paperid=67316>.

ARRUDA, K. G. d. *Avaliação dos custos do tratamento de tuberculose em município de médio porte do Nordeste brasileiro*. masterThesis, 2014. Accepted: 2015-03-09T18:11:34Z Publisher: Universidade Federal de Pernambuco. Disponível em: <https://repositorio.ufpe.br/handle/123456789/11539>.

AVELLEIRA, J. C. R.; BOTTINO, G. Sífilis: diagnóstico, tratamento e controle. v. 81, n. 2, p. 111–126, 2006. ISSN 0365-0596. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0365-05962006000200002&lng=pt&tlng=pt>.

AZEEZ, A.; OBAROMI, D.; ODEYEMI, A.; NDEGE, J.; MUNTABAYI, R. Seasonality and trend forecasting of tuberculosis prevalence data in eastern cape, south africa, using a hybrid model. v. 13, n. 8, p. 757, 2016. Number: 8 Publisher: Multidisciplinary Digital Publishing Institute. Disponível em: <https://www.mdpi.com/1660-4601/13/8/757>.

BAEZA-YATES, R.; RIBEIRO-NETO, B. *Modern Information Retrieval: The Concepts and Technology Behind Search*. 2nd ed. edição. ed. [S.l.]: Addison-Wesley Professional, 2011. ISBN 978-0-321-41691-9.

BASTOS, L. S.; ECONOMOU, T.; GOMES, M. F. C.; VILLELA, D. A. M.; COELHO, F. C.; CRUZ, O. G.; STONER, O.; BAILEY, T.; CODEçO, C. T. A modelling approach for correcting reporting delays in disease surveillance data. v. 38, n. 22, p. 4363–4377, 2019. ISSN 1097-0258. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.8303. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.8303>.

BATTINENI, G.; CHINTALAPUDI, N.; AMENTA, F. Forecasting of COVID-19 epidemic size in four high hitting nations (USA, brazil, india and russia) by fb-prophet machine learning model. ahead-of-print, 2020. ISSN 2210-8327. Disponível em: <https://doi.org/10.1108/ACI-09-2020-0059>.

BELLAN, S. E.; PULLIAM, J. R. C.; SCOTT, J. C.; DUSHOFF, J.; the MMED Organizing Committee. How to make epidemiological training infectious. v. 10, n. 4, p. e1001295, 2012. ISSN 1545-7885. Disponível em: <https://dx.plos.org/10.1371/journal.pbio.1001295>.

BENGIO, Y.; SIMARD, P.; FRASCONI, P. Learning long-term dependencies with gradient descent is difficult. v. 5, n. 2, p. 157–166, 1994. ISSN 1941-0093. Conference Name: IEEE Transactions on Neural Networks.

BETTENCOURT, L. M. A.; RIBEIRO, R. M. Real time bayesian estimation of the epidemic potential of emerging infectious diseases. v. 3, n. 5, p. e2185, 2008. ISSN 1932-6203. Publisher: Public Library of Science. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0002185>.

BEZERRA, A. *Biblioteca Virtual em Saúde Doenças infecciosas e Parasitárias chega às redes sociais*. 2016. Disponível em: <https://portal.fiocruz.br/noticia/biblioteca-virtual-em-saude-doencas-infecciosas-e-parasitarias-chega-redes-sociais>.

BINDá, J. M.; BRANDT, M. A. G.; PIEDADE, M. P. Análise da aplicação de sistemas de recuperação de informação usando android numa base bíblica. p. 10, 2013.

BITOUN, J.; DUARTE, C. C.; BEZERRA, A. C. V.; FERNANDES, A. C. d. A.; SANTOS, L. S. Novo coronavírus, velhas desigualdades: distribuição dos casos, óbitos e letalidade por SRAG decorrentes da covid-19 na cidade do recife. n. 48, 2020. ISSN 1958-9212. Number: 48 Publisher: Théry, Hervé. Disponível em: <http://journals.openedition.org/confins/34667>.

BOX, G. E. P.; JENKINS, G. M.; REINSEL, G. C.; LJUNG, G. M. *Time Series Analysis: Forecasting and Control*. [S.l.]: John Wiley & Sons, 2015. Google-Books-ID: rNt5CgAAQBAJ. ISBN 978-1-118-67492-5.

BOX, G. E. P.; PIERCE, D. A. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. v. 65, n. 332, p. 1509–1526, 1970. ISSN 0162-1459. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. Disponível em: <https://www.jstor.org/stable/2284333>.

BRAS, A. L.; GOMES, D.; FILIPE, P. A.; SOUSA, B. de; NUNES, C. Trends, seasonality and forecasts of pulmonary tuberculosis in portugal. v. 18, n. 10, p. 1202–1210, 2014.

BROOKS, L. C.; FARROW, D. C.; HYUN, S.; TIBSHIRANI, R. J.; ROSENFELD, R. Nonmechanistic forecasts of seasonal influenza with iterative one-week-ahead distributions. v. 14, n. 6, p. e1006134, 2018. ISSN 1553-7358. Publisher: Public Library of Science. Disponível em: <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006134>.

BUCKEE, C. O.; CARDENAS, M. I. E.; CORPUZ, J.; GHOSH, A.; HAQUE, F.; KARIM, J.; MAHMUD, A. S.; MAUDE, R. J.; MENSAH, K.; MOTAZE, N. V.; NABAGGALA, M.; METCALF, C. J. E.; MIORAMALALA, S. A.; MUBIRU, F.; PEAK, C. M.; PRAMANIK, S.; RAKOTONDRAMANGA, J. M.; REMERA, E.; SINHA, I.; SOVANNAROTH, S.; TATEM, A. J.; ZAW, W. Productive disruption: opportunities and challenges for innovation in infectious disease surveillance. v. 3, n. 1, p. e000538, 2018. ISSN 2059-7908. Publisher: BMJ Specialist Journals Section: Commentary. Disponível em: <https://gh.bmj.com/content/3/1/e000538>.

CANOVA, F.; HANSEN, B. E. Are seasonal patterns constant over time? a test for seasonal stability. v. 13, n. 3, p. 237–252, 1995. ISSN 0735-0015. Publisher: [American Statistical Association, Taylor & Francis, Ltd.]. Disponível em: <https://www.jstor.org/stable/1392184>.

CAO, L.; GU, Q. Dynamic support vector machines for non-stationary time series forecasting. v. 6, n. 1, p. 67–83, 2002. ISSN 1088-467X. Publisher: IOS Press. Disponível em: <https://content.iospress.com/articles/intelligent-data-analysis/ida00079>.

CHEN, P.-H. C.; LIU, Y.; PENG, L. How to develop machine learning models for healthcare. v. 18, n. 5, p. 410–414, 2019. ISSN 1476-4660. Number: 5 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/s41563-019-0345-0>.

CHENG, C.; SA-NGASOONGSONG, A.; BEYCA, O.; LE, T.; YANG, H.; KONG, Z. J.; BUKKAPATNAM, S. T. S. Time series forecasting for nonlinear and non-stationary processes: a review and comparative study. v. 47, n. 10, p. 1053–1071, 2015. ISSN 0740-817X. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/0740817X.2014.999180. Disponível em: <https://doi.org/10.1080/0740817X.2014.999180>.

CHOISY, M.; GUGAN, J.-F.; ROHANI, P. Mathematical modeling of infectious diseases dynamics. In: TIBAYRENC, M. (Ed.). *Encyclopedia of Infectious Diseases*. John Wiley & Sons, Inc., 2007. p. 379–404. ISBN 978-0-470-11420-9 978-0-471-65732-3. Disponível em: <http://doi.wiley.com/10.1002/9780470114209.ch22>.

CMMID COVID-19 working group; DAVIES, N. G.; KLEPAC, P.; LIU, Y.; PREM, K.; JIT, M.; EGGO, R. M. Age-dependent effects in the transmission and control of COVID-19 epidemics. v. 26, n. 8, p. 1205–1211, 2020. ISSN 1078-8956, 1546-170X. Disponível em: <http://www.nature.com/articles/s41591-020-0962-9>.

CORREIA, M. F. B. (Ed.). *Recuperação de informação*. [S.l.]: Informática aplicada: ITI 4301, 2018.

DJERBOUAI, S.; SOUAG-GAMANE, D. Drought forecasting using neural networks, wavelet neural networks, and stochastic models: Case of the algerois basin in north algeria. v. 30, n. 7, p. 2445–2464, 2016. ISSN 1573-1650. Disponível em: <https://doi.org/10.1007/s11269-016-1298-6>.

EDWARDS, A. L. *An Introduction to Linear Regression and Correlation*. 1st edition. ed. [S.l.]: W. H. Freeman, 1976. ISBN 978-0-7167-0561-1.

ELASTIC. Learn/Docs/Elasticsearch/Reference/7.14, *Data in: documents and indices*. 2021. Disponível em: <https://www.elastic.co/guide/en/elasticsearch/reference/current/documents-indices.html>.

EVOY, D. M.; McAloon, C. G.; COLLINS, \. B.; HUNT, K.; BUTLER, F.; BYRNE, A. W.; CASEY, M.; BARBER, A.; GRIFFIN, J.; LANE, E. A.; WALL, P.; MORE, S. J. preprint, *The relative infectiousness of asymptomatic SARS-CoV-2 infected persons compared with symptomatic individuals: A rapid scoping review*. 2020. Disponível em: <http://medrxiv.org/lookup/doi/10.1101/2020.07.30.20165084>.

FAKHFAKH, M.; BOUAZIZ, B.; FAIEZ, G.; LOTFI, C. ProgNet: COVID-19 prognosis using recurrent and convolutional neural networks. v. 12, p. 11–12, 2020. Disponível em: <https://openmedicalimagingjournal.com/VOLUME/12/PAGE/11/>.

FARRINGTON, C. P.; ANDREWS, N. J.; BEALE, A. D.; CATCHPOLE, M. A. A statistical algorithm for the early detection of outbreaks of infectious disease. v. 159, n. 3, p. 547–563, 1996. ISSN 1467-985X. _eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.2307/2983331. Disponível em: <https://rss.onlinelibrary.wiley.com/doi/abs/10.2307/2983331>.

FERREIRA, R. A. *Modelagem e predição espaço-temporal dos casos de dengue utilizando processo pontual de Cox log-Gaussiano*. phdthesis, 2017. Accepted: 2017-05-10T18:34:04Z Publisher: Universidade Federal de Lavras. Disponível em: <http://repositorio.ufla.br/jspui/handle/1/12942>.

FIOCRUZ. *Monitoramento de casos de síndrome respiratória aguda grave (SRAG) notificados no SIVEP-Gripe*. 2020. Disponível em: <http://info.gripe.fiocruz.br/>.

GONZALEZ, M.; LIMA, V. L. S. de. Recuperação de informação e processamento da linguagem natural. p. 49, 2002.

GRASSLY, N. C.; FRASER, C. Mathematical models of infectious disease transmission. v. 6, n. 6, p. 477–487, 2008. ISSN 1740-1534. Number: 6 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/nrmicro1845>.

GU, J.; LIANG, L.; SONG, H.; KONG, Y.; MA, R.; HOU, Y.; ZHAO, J.; LIU, J.; HE, N.; ZHANG, Y. A method for hand-foot-mouth disease prediction using GeoDetector and LSTM model in guangxi, china. v. 9, n. 1, p. 17928, 2019. ISSN 2045-2322. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Risk factors;Viral infection Subject_term_id: risk-factors;viral-infection. Disponível em: <https://www.nature.com/articles/s41598-019-54495-2>.

HE, F.; HU, Z.-j.; ZHANG, W.-c.; CAI, L.; CAI, G.-x.; AOYAGI, K. Construction and evaluation of two computational models for predicting the incidence of influenza in nagasaki prefecture, japan. v. 7, n. 1, p. 7192, 2017. ISSN 2045-2322. Number: 1 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/s41598-017-07475-3>.

HE, Z.; TAO, H. Epidemiology and ARIMA model of positive-rate of influenza viruses among children in wuhan, china: A nine-year retrospective study. v. 74, p. 61–70, 2018. ISSN 1201-9712. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1201971218344618>.

HEALTH, G. *The top 10 causes of death*. 2020. Disponível em: <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death>.

HEESTERBEEK, H.; ANDERSON, R. M.; ANDREASEN, V.; BANSAL, S.; ANGELIS, D. D.; DYE, C.; EAMES, K. T. D.; EDMUNDS, W. J.; FROST, S. D. W.; FUNK, S.; HOLLINGSWORTH, T. D.; HOUSE, T.; ISHAM, V.; KLEPAC, P.; LESSLER, J.; LLOYD-SMITH, J. O.; METCALF, C. J. E.; MOLLISON, D.; PELLIS, L.; PULLIAM, J. R. C.; ROBERTS, M. G.; VIBOUD, C.; Isaac Newton Institute IDD Collaboration. Modeling infectious disease dynamics in the complex landscape of global health. v. 347, n. 6227, p. aaa4339, 2015. ISSN 1095-9203.

HELD, L.; HöHLE, M.; HOFMANN, M. A statistical framework for the analysis of multivariate infectious disease surveillance counts. v. 5, n. 3, p. 187–199, 2005. ISSN 1471-082X. Publisher: SAGE Publications India. Disponível em: <https://doi.org/10.1191/1471082X05st098oa>.

HELD, L.; MEYER, S.; BRACHER, J. Probabilistic forecasting in infectious disease epidemiology: the 13th armitage lecture. v. 36, n. 22, p. 3443–3460, 2017. ISSN 1097-0258. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sim.7363. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.7363>.

HOCHREITER, S.; SCHMIDHUBER, J. Long short-term memory. v. 9, n. 8, p. 1735–1780, 1997. ISSN 0899-7667. Disponível em: <https://doi.org/10.1162/neco.1997.9.8.1735>.

HU, Z.; GE, Q.; LI, S.; JIN, L.; XIONG, M. Artificial intelligence forecasting of covid-19 in china. 2020. Disponível em: <http://arxiv.org/abs/2002.07112>.

HYNDMAN, R. J.; ATHANASOPOULOS, G. *Forecasting: Principles and Practice (2nd ed)*. [s.n.], 2018. v. 2. Disponível em: <https://Otexts.com/fpp2/>.

HYNDMAN, R. J.; KHANDAKAR, Y. Automatic time series forecasting: The forecast package for r. v. 27, n. 1, p. 1–22, 2008. ISSN 1548-7660. Number: 1. Disponível em: <https://www.jstatsoft.org/index.php/jss/article/view/v027i03>.

IVORRA, B.; FERRáNDEZ, M. R.; VELA-PéREZ, M.; RAMOS, A. M. Mathematical modeling of the spread of the coronavirus disease 2019 (COVID-19) taking into account the undetected infections. the case of china. v. 88, p. 105303, 2020. ISSN 1007-5704. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1007570420301350>.

J., T. S.; LETHAM, B. *Prophet: forecasting at scale*. 2017. Disponível em: <https://research.fb.com/blog/2017/02/prophet-forecasting-at-scale/>.

JONES, K. E.; PATEL, N. G.; LEVY, M. A.; STOREYGARD, A.; BALK, D.; GITTLEMAN, J. L.; DASZAK, P. Global trends in emerging infectious diseases. v. 451, n. 7181, p. 990–993, 2008. ISSN 1476-4687. Number: 7181 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/nature06536>.

JONES, K. S.; WILLETT, P. (Ed.). *Readings in Information Retrieval*. [S.l.]: Morgan Kaufmann Publishers, Inc., 1997.

KHAIR, U.; FAHMI, H.; HAKIM, S. A.; RAHIM, R. Forecasting error calculation with mean absolute deviation and mean absolute percentage error. v. 930, p. 012002, 2017. ISSN 1742-6596. Publisher: IOP Publishing. Disponível em: <https://doi.org/10.1088/1742-6596/930/1/012002>.

KOLEN, J. F.; KREMER, S. C. Gradient flow in recurrent nets: The difficulty of learning LongTerm dependencies. In: *A Field Guide to Dynamical Recurrent Networks*. IEEE, 2009. ISBN 978-0-470-54403-7. Disponível em: <http://ieeexplore.ieee.org/search/srchabstract.jsp?arnumber=5264952>.

KOOP, G.; DIJK, H. K. V. Testing for integration using evolving trend and seasonals models: A bayesian approach. v. 97, n. 2, p. 261–291, 2000. ISSN 0304-4076. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0304407699000718>.

KöPPEN, W. Das geographische system der klimate (1936). p. 44, 1936.

LAMPOS, V.; MILLER, A. C.; CROSSAN, S.; STEFANSEN, C. Advances in nowcasting influenza-like illness rates using search query logs. v. 5, n. 1, p. 12760, 2015. ISSN 2045-2322. Number: 1 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/srep12760>.

LEE, N.; HUI, D.; WU, A.; CHAN, P.; CAMERON, P.; JOYNT, G. M.; AHUJA, A.; YUNG, M. Y.; LEUNG, C.; TO, K.; LUI, S.; SZETO, C.; CHUNG, S.; SUNG, J. J. A major outbreak of severe acute respiratory syndrome in hong kong. v. 348, n. 20, p. 1986–1994, 2003. ISSN 0028-4793. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa030685. Disponível em: <https://doi.org/10.1056/NEJMoa030685>.

LEMOS, F. d. O. Metodologia para seleção de métodos de previsão de demanda. 2006. Accepted: 2007-06-06T18:51:20Z. Disponível em: <https://lume.ufrgs.br/handle/10183/5949>.

LEVNTAL, Z. [history of tuberculosis]. v. 11, n. 3, p. 111–114, 1957. ISSN 0375-9342.

LIN, Y.; JIANG, D.; LIU, T. Nontrivial periodic solution of a stochastic epidemic model with seasonal variation. v. 45, p. 103–107, 2015. ISSN 0893-9659. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0893965915000397>.

LIU, H.; TIAN, H.-q.; LI, Y.-f. Comparison of two new ARIMA-ANN and ARIMA-kalman hybrid methods for wind speed prediction. v. 98, p. 415–424, 2012. Publisher: Elsevier. Disponível em: <https://ideas.repec.org/a/eee/appene/v98y2012icp415-424.html>.

LIU, Q.; JIANG, D. Stationary distribution and extinction of a stochastic SIR model with nonlinear perturbation. v. 73, p. 8–15, 2017. ISSN 0893-9659. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0893965917301349>.

LIU, Q.; LI, Z.; JI, Y.; MARTINEZ, L.; ZIA, U. H.; JAVAID, A.; LU, W.; WANG, J. Forecasting the seasonality and trend of pulmonary tuberculosis in jiangsu province of china using advanced statistical time-series analyses. v. 12, p. 2311–2322, 2019. ISSN 1178-6973. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6666376/>.

LJUNG, G. M.; BOX, G. E. P. On a measure of lack of fit in time series models. v. 65, n. 2, p. 297–303, 1978. ISSN 0006-3444. Disponível em: <https://doi.org/10.1093/biomet/65.2.297>.

MACKWAY-JONES, K.; MARSDEN, J.; WINDLE, J. *Emergency Triage: Manchester Triage Group, Third Edition*. John Wiley & Sons, Ltd, 2013. Section: 3 _eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118299029.ch3. ISBN 978-1-118-29902-9. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118299029.ch3>.

MAGLEBY, R.; WESTBLADE, L. F.; TRZEBUCKI, A.; SIMON, M. S.; RAJAN, M.; PARK, J.; GOYAL, P.; SAFFORD, M. M.; SATLIN, M. J. Impact of severe acute respiratory syndrome coronavirus 2 viral load on risk of intubation and mortality among hospitalized patients with coronavirus disease 2019. 2020. ISSN 1058-4838. Disponível em: <https://doi.org/10.1093/cid/ciaa851>.

MAKRIDAKIS, S.; WHEELWRIGHT, S. C.; HYNDMAN, R. J. *Forecasting: Methods and Applications*. 3ª edição. ed. [S.l.]: Wiley, 1997. ISBN 978-0-471-53233-0.

MARSDEN-HAUG, N.; FOSTER, V. B.; GOULD, P. L.; ELBERT, E.; WANG, H.; PAVLIN, J. A. Code-based syndromic surveillance for influenzalike illness by international classification of diseases, ninth revision. v. 13, n. 2, p. 207–216, 2007. ISSN 1080-6040. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2725845/>.

MARTINEZ, E. Z.; SILVA, E. A. S. d.; FABBRO, A. L. D. A SARIMA forecasting model to predict the number of cases of dengue in campinas, state of são paulo, brazil. v. 44, n. 4, p. 436–440, 2011. ISSN 1678-9849.

MASSAD, E.; BURATTINI, M. N.; LOPEZ, L. F.; COUTINHO, F. A. B. Forecasting versus projection models in epidemiology: The case of the SARS epidemics. v. 65, n. 1, p. 17–22, 2005. ISSN 0306-9877. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306987705000800>.

MEIRELES, M. *Dois primeiros casos do novo coronavírus são confirmados em PE, diz Secretaria de Saúde | Pernambuco | G1*. 2020. Disponível em: <https://g1.globo.com/pe/pernambuco/noticia/2020/03/12/primeiros-casos-de-coronavirus-sao-confirmados-pela-secretaria-de-saude-de-pernambuco.ghtml>.

MELLO, D. *Risco de morrer por coronavírus varia até 10 vezes entre bairros de SP*. 2020. Disponível em: <https://agenciabrasil.ebc.com.br/saude/noticia/2020-05/risco-de-morrer-por-coronavirus-varia-ate-10-vezes-entre-bairros-de-sp>.

MENDES, E.; TEIXEIRA, C. *Distrito sanitário: o processo social de mudança das práticas sanitárias do Sistema Unico de Saúde*. Hucitec, 1993. (Saúde em debate). Disponível em: <https://books.google.com.br/books?id=GxNgAAAAMAAJ>.

MOOSAZADEH, M.; KHANJANI, N.; BAHRAMPOUR, A. Seasonality and temporal variations of tuberculosis in the north of iran. v. 12, n. 4, p. 35–41, 2013. ISSN 1735-0344.

MORENS, D. M.; FAUCI, A. S. Emerging infectious diseases: Threats to human health and global stability. v. 9, n. 7, p. e1003467, 2013. ISSN 1553-7374. Publisher: Public Library of Science. Disponível em: <https://journals.plos.org/plospathogens/article?id=10.1371/journal.ppat.1003467>.

MORENS, D. M.; FOLKERS, G. K.; FAUCI, A. S. The challenge of emerging and re-emerging infectious diseases. v. 430, n. 6996, p. 242–249, 2004. ISSN 1476-4687. Number: 6996 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/nature02759>.

MORETTIN, P. A.; TOLOI, C. Maria de C. *Previsão de Séries Temporais*. 2. ed. [S.l.]: Atual Editora, 1985. v. 2.

MUELLNER, U.; FOURNIé, G.; MUELLNER, P.; AHLSTROM, C.; PFEIFFER, D. U. epidemix—an interactive multi-model application for teaching and visualizing infectious disease transmission. v. 23, p. 49–54, 2018. ISSN 1755-4365. Disponível em: <http://www.sciencedirect.com/science/article/pii/S1755436517300270>.

MUNICIPAL, R. G.; RECIFE, S. de Saúde do; COORDENAçãO, S. E. de; Gerência Geral de Planejamento. *Plano Municipal de Saúde 2014 - 2017*. Secretaria de Saúde do Recife, 2014. 1ª. Ed. - Secretaria de Saúde do Recife, 2014. Disponível em: <http://www2.recife.pe.gov.br/sites/default/files/plano_municipal_de_saude_2015_revisado_menor.pdf>.

MUNICIPAL, R. G.; RECIFE, S. de Saúde do; GERAL, S. E. de C. *Plano Municipal de Saúde 2018 - 2021*. Secretaria de Saúde do Recife, 2018. 1ª. Ed. - Secretaria de Saúde do Recife, 2018.xxx. Disponível em: <http://www2.recife.pe.gov.br/sites/default/files/plano_municipal_de_saude_2018_2021_vf.pdf>.

MURRAY, C. J.; LOPEZ, A. D.; CHIN, B.; FEEHAN, D.; HILL, K. H. Estimation of potential global pandemic influenza mortality on the basis of vital registry data from the 1918–20 pandemic: a quantitative analysis. v. 368, n. 9554, p. 2211–2218, 2006. ISSN 0140-6736. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0140673606698954>.

NAJAFABADI, M. M.; VILLANUSTRE, F.; KHOSHGOFTAAR, T. M.; SELIYA, N.; WALD, R.; MUHAREMAGIC, E. Deep learning applications and challenges in big data analytics. v. 2, n. 1, p. 1, 2015. ISSN 2196-1115. Disponível em: <https://doi.org/10.1186/s40537-014-0007-7>.

NAU, R. *Statistical forecasting: notes on regression and time series analysis*. 2020. Disponível em: <https://people.duke.edu/~rnau/411arim.htm>.

OLSAVSZKY, V.; DOSIUS, M.; VLADESCU, C.; BENECKE, J. Time series analysis and forecasting with automated machine learning on a national ICD-10 database. v. 17, n. 14, p. 4979, 2020. Number: 14 Publisher: Multidisciplinary Digital Publishing Institute. Disponível em: <https://www.mdpi.com/1660-4601/17/14/4979>.

ORGANIZATION, W. H. *WHO issues a global alert about cases of atypical pneumonia*. 2003. Publisher: World Health Organization. Disponível em: <https://www.who.int/csr/sars/archive/2003_03_12/en/>.

ORGANIZATION, W. H. *Severe Acute Respiratory Syndrome (SARS)*. 2021. Disponível em: <https://www.who.int/health-topics/severe-acute-respiratory-syndrome>.

PEAT, J.; BARTON, B. *Medical Statistics: A Guide to Data Analysis and Critical Appraisal*. [S.l.]: John Wiley & Sons, 2008. Google-Books-ID: NHiDnKiDajEC. ISBN 978-0-470-75520-4.

PULVER, A.; LYU, S. LSTM with working memory. In: *2017 International Joint Conference on Neural Networks (IJCNN)*. [S.l.: s.n.], 2017. p. 845–851. ISSN: 2161-4407.

RIBEIRO, M. H. D. M.; SILVA, R. G. da; MARIANI, V. C.; COELHO, L. d. S. Short-term forecasting COVID-19 cumulative confirmed cases: Perspectives for brazil. v. 135, p. 109853, 2020. ISSN 0960-0779. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960077920302538>.

RILEY, S. Large-scale spatial-transmission models of infectious disease. v. 316, n. 5829, p. 1298–1301, 2007. ISSN 0036-8075, 1095-9203. Publisher: American Association for the Advancement of Science Section: Review. Disponível em: <https://science.sciencemag.org/content/316/5829/1298>.

RITCHIE, H.; ROSER, M. Causes of death. 2018. Disponível em: <https://ourworldindata.org/causes-of-death>.

ROOSA, K.; CHOWELL, G. Assessing parameter identifiability in compartmental dynamic models using a computational approach: application to infectious disease transmission models. v. 16, n. 1, p. 1, 2019. ISSN 1742-4682. Disponível em: <https://tbiomed.biomedcentral.com/articles/10.1186/s12976-018-0097-6>.

RUBIN, D. B. Inference and missing data. v. 63, n. 3, p. 581–592, 1976. ISSN 0006-3444. Publisher: [Oxford University Press, Biometrika Trust]. Disponível em: <https://www.jstor.org/stable/2335739>.

SAGHEER, A.; KOTB, M. Unsupervised pre-training of a deep LSTM-based stacked autoencoder for multivariate time series forecasting problems. v. 9, n. 1, p. 19038, 2019. ISSN 2045-2322. Disponível em: <http://www.nature.com/articles/s41598-019-55320-6>.

SAÚDE, M. da. *Doenças Infecciosas e Parasitárias: Guia de bolso*. 2010. 444 p. Doenças Infecciosas e Parasitárias: Guia de bolso. Disponível em: <http://bvsms.saude.gov.br/bvs/publicacoes/doencas_infecciosas_parasitaria_guia_bolso.pdf>.

SAúDE, M. da. *Brasil avança no enfrentamento à sífilis*. 2020. Disponível em: <https://www.gov.br/saude/pt-br/assuntos/noticias/brasil-avanca-no-enfrentamento-a-sifilis>.

SCHULLER, B. W.; SCHULLER, D. M.; QIAN, K.; LIU, J.; ZHENG, H.; LI, X. COVID-19 and computer audition: An overview on what speech & sound analysis could contribute in the SARS-CoV-2 corona crisis. 2020. Disponível em: <http://arxiv.org/abs/2003.11117>.

SCHWARZ, G. Estimating the dimension of a model. v. 6, p. 461–464, 1978. Disponível em: <http://adsabs.harvard.edu/abs/1978AnSta...6..461S>.

SERVICES INC., A. W. *Amazon Forecast - Guia do desenvolvedor*. Amazon Web Services, Inc., 2020. Disponível em: <https://docs.aws.amazon.com/pt_br/forecast/latest/dg/forecast.dg.pdf#aws-forecast-recipe-prophet>.

SHAHID, F.; ZAMEER, A.; MUNEEB, M. Predictions for COVID-19 with deep learning models of LSTM, GRU and bi-LSTM. v. 140, p. 110212, 2020. ISSN 0960-0779. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0960077920306081>.

SHASTRI, S.; SINGH, K.; KUMAR, S.; KOUR, P.; MANSOTRA, V. Time series forecasting of covid-19 using deep learning models: India-USA comparative case study. v. 140, p. 110227, 2020. ISSN 0960-0779. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7440083/>.

SHERSTINSKY, A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. v. 404, p. 132306, 2020. ISSN 0167-2789. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167278919305974>.

SINGH, J.; JAIN, A.; DHYANI, T.; AGRAWAL, P. *Using Deep Learning for Forecasting Tuberculosis Morbidity Rate*. [S.l.: s.n.], 2018.

SINGH, K.; SHASTRI, S.; BHADWAL, A. S.; KOUR, P.; KUMARI, M.; SHARMA, D. A.; MANSOTRA, V. Implementation of exponential smoothing for forecasting time series data. v. 8, n. 1, p. 55–71, 2019. ISSN 2319-1953.

SMIESZEK, T.; FIEBIG, L.; SCHOLZ, R. W. Models of epidemics: when contact repetition and clustering should be included. v. 6, n. 1, p. 11, 2009. ISSN 1742-4682. Disponível em: <https://doi.org/10.1186/1742-4682-6-11>.

SONG, Y.; WANG, F.; WANG, B.; TAO, S.; ZHANG, H.; LIU, S.; RAMIREZ, O.; ZENG, Q. Time series analyses of hand, foot and mouth disease integrating weather variables. v. 10, n. 3, p. e0117296, 2015. ISSN 1932-6203. Publisher: Public Library of Science. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117296>.

SOUZA, H. P. de; OLIVEIRA, W. T. G. H. de; SANTOS, J. P. C. dos; TOLEDO, J. P.; FERREIRA, I. P. S.; ESASHIKA, S. N. G. de S.; LIMA, T. F. P. de; DELáCIO, A. de S. Doenças infecciosas e parasitárias no brasil de 2010 a 2017: aspectos para vigilância em saúde. v. 44, p. 1, 2020. ISSN 1020-4989, 1680-5348. Disponível em: <http://iris.paho.org/xmlui/handle/123456789/51858>.

SOUZA, R. C. T. D. *Previsão De Séries Temporais Utilizando Rede Neural Treinada Por Filtro De Kalman E Evolução Diferencial*. phdthesis, 2008. Disponível em: <https://docs.ufpr.br/~thom/papers/dissertation.pdf>.

STEINSKOG, D. J.; TJøSTHEIM, D. B.; KVAMSTø, N. G. A cautionary note on the use of the kolmogorov–smirnov test for normality. v. 135, n. 3, p. 1151–1157, 2007. ISSN 1520-0493, 0027-0644. Publisher: American Meteorological Society Section: Monthly Weather Review. Disponível em: <https://journals.ametsoc.org/view/journals/mwre/135/3/mwr3326.1.xml>.

STOCK, J.; WATSON, M. *Introduction to Econometrics*. 4ª edição. ed. [S.l.]: Pearson, 2018. ISBN 978-0-13-446199-1.

STONER, O.; ECONOMOU, T.; SILVA, G. D. M. d. A hierarchical framework for correcting under-reporting in count data. v. 114, n. 528, p. 1481–1492, 2019. ISSN 0162-1459. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/01621459.2019.1573732. Disponível em: <https://doi.org/10.1080/01621459.2019.1573732>.

SYCZEWSKA, E. M. Warsaw school of economics. p. 21, 1997.

TANDON, H.; RANJAN, P.; CHAKRABORTY, T.; SUHAG, V. Coronavirus (COVID-19): ARIMA based time-series analysis to forecast near future. 2020. Disponível em: <http://arxiv.org/abs/2004.07859>.

TAYLOR, S. *The Psychology of Pandemics: Preparing for the Next Global Outbreak of Infectious Disease*. [S.l.]: Cambridge Scholars Publishing, 2019. Google-Books-ID: 8mq1DwAAQBAJ. ISBN 978-1-5275-4118-4.

TAYLOR, S. J.; LETHAM, B. *Forecasting at scale*. 2017. ISSN: 2167-9843. Disponível em: <https://peerj.com/preprints/3190>.

TAYLOR, S. J.; LETHAM, B. Forecasting at scale. v. 72, n. 1, p. 37–45, 2018. ISSN 0003-1305. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2017.1380080. Disponível em: <https://doi.org/10.1080/00031305.2017.1380080>.

TEALAB, A. Time series forecasting using artificial neural networks methodologies: A systematic review. v. 3, n. 2, p. 334–340, 2018. ISSN 2314-7288. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2314728817300715>.

THIN, G. *Leprosy*. [S.l.]: Percival, 1891. Google-Books-ID: SuFNAQAAMAAJ.

THIYAGARAJAN, K.; KODAGODA, S.; ULAPANE, N.; PRASAD, M. A temporal forecasting driven approach using facebook's prophet method for anomaly detection in sewer air temperature sensor system. In: *2020 15th IEEE Conference on Industrial Electronics and Applications (ICIEA)*. [S.l.: s.n.], 2020. p. 25–30. ISSN: 2158-2297.

THODE, H. C. *Testing For Normality*. [S.l.]: CRC Press, 2002. Google-Books-ID: gbegXB4SdosC. ISBN 978-0-203-91089-4.

TIAN, C. W.; WANG, H.; LUO, X. M. Time-series modelling and forecasting of hand, foot and mouth disease cases in china from 2008 to 2018. v. 147, p. e82, 2019. ISSN 1469-4409.

TUITE, A. R.; TIEN, J.; EISENBERG, M.; EARN, D. J.; MA, J.; FISMAN, D. N. Cholera epidemic in haiti, 2010: Using a transmission model to explain spatial spread of disease and identify optimal control interventions. v. 154, n. 9, p. 593–601, 2011. ISSN 0003-4819. Publisher: American College of Physicians. Disponível em: <https://www.acpjournals.org/doi/10.7326/0003-4819-154-9-201105030-00334>.

VOLKOVA, S.; AYTON, E.; PORTERFIELD, K.; CORLEY, C. D. Forecasting influenza-like illness dynamics for military populations using neural networks and social media. v. 12, n. 12, p. e0188941, 2017. ISSN 1932-6203. Publisher: Public Library of Science. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0188941>.

WANG, H.; TIAN, C. W.; WANG, W. M.; LUO, X. M. Time-series analysis of tuberculosis from 2005 to 2017 in china. v. 146, n. 8, p. 935–939, 2018. ISSN 0950-2688, 1469-4409. Publisher: Cambridge University Press. Disponível em: <https://www.cambridge.org/core/journals/epidemiology-and-infection/article/timeseries-analysis-of-tuberculosis-from-2005-to-2017-in-china/53C9D44E40C9A9F01A16B2ADB359814B>.

WANG, X.; SMITH, K.; HYNDMAN, R. Characteristic-based clustering for time series data. v. 13, n. 3, p. 335–364, 2006. ISSN 1573-756X. Disponível em: <https://doi.org/10.1007/s10618-005-0039-x>.

WEISS, R. A.; McMichael, A. J. Social and environmental risk factors in the emergence of infectious diseases. v. 10, n. 12, p. S70–S76, 2004. ISSN 1546-170X. Number: 12 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/nm1150>.

WILLIS, M. D.; WINSTON, C. A.; HEILIG, C. M.; CAIN, K. P.; WALTER, N. D.; KENZIE, W. R. M. Seasonality of tuberculosis in the united states, 1993–2008. v. 54, n. 11, p. 1553–1560, 2012. ISSN 1058-4838. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4867465/>.

WILLMOTT, C. J.; MATSUURA, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. v. 30, n. 1, p. 79–82, 2005. ISSN 0936-577X. Publisher: Inter-Research Science Center. Disponível em: <https://www.jstor.org/stable/24869236>.

WONCA. *ICPC-2-R: International Classification of Primary Care*. 2ª edição. ed. [S.l.]: OUP Oxford, 2005. ISBN 978-0-19-856857-5.

WOOLHOUSE, M. E. J.; DYE, C.; TAYLOR, L. H.; LATHAM, S. M.; WOOLHOUSE, M. E. Risk factors for human disease emergence. v. 356, n. 1411, p. 983–989, 2001. Publisher: Royal Society. Disponível em: <https://royalsocietypublishing.org/doi/10.1098/rstb.2001.0888>.

XIE, G. A novel monte carlo simulation procedure for modelling COVID-19 spread over time. v. 10, n. 1, p. 13120, 2020. ISSN 2045-2322. Number: 1 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/s41598-020-70091-1>.

XU, J.; XU, K.; LI, Z.; MENG, F.; TU, T.; XU, L.; LIU, Q. Forecast of dengue cases in 20 chinese cities based on the deep learning method. v. 17, n. 2, p. 453, 2020. Number: 2 Publisher: Multidisciplinary Digital Publishing Institute. Disponível em: <https://www.mdpi.com/1660-4601/17/2/453>.

XU, Q.; GEL, Y. R.; RAMIREZ, L. L. R.; NEZAFATI, K.; ZHANG, Q.; TSUI, K.-L. Forecasting influenza in hong kong with google search queries and statistical model fusion. v. 12, n. 5, p. e0176690, 2017. ISSN 1932-6203. Publisher: Public Library of Science. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0176690>.

XU, R.-H.; HE, J.-F.; EVANS, M. R.; PENG, G.-W.; FIELD, H. E.; YU, D.-W.; LEE, C.-K.; LUO, H.-M.; LIN, W.-S.; LIN, P.; LI, L.-H.; LIANG, W.-J.; LIN, J.-Y.; SCHNUR, A. Epidemiologic clues to SARS origin in china. v. 10, n. 6, p. 1030–1037, 2004. ISSN 1080-6040. Disponível em: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3323155/>.

YOLCU, U.; EGRIOGLU, E.; ALADAG, C. H. A new linear & nonlinear artificial neural network model for time series forecasting. v. 54, n. 3, p. 1340–1347, 2013. ISSN 0167-9236. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167923612003557>.

ZHAN, Z.; DONG, W.; LU, Y.; YANG, P.; WANG, Q.; JIA, P. Real-time forecasting of hand-foot-and-mouth disease outbreaks using the integrating compartment model and assimilation filtering. v. 9, n. 1, p. 2661, 2019. ISSN 2045-2322. Number: 1 Publisher: Nature Publishing Group. Disponível em: <https://www.nature.com/articles/s41598-019-38930-y>.

ZHANG, Q.; ZHOU, K. Stationary distribution and extinction of a stochastic SIQR model with saturated incidence rate. v. 2019, p. e3575410, 2019. ISSN 1024-123X. Publisher: Hindawi. Disponível em: <https://www.hindawi.com/journals/mpe/2019/3575410/>.

ZHI, Z. L. X. B. X. Z. The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in china. v. 41, n. 2, p. 145–151, 2020. ISSN 0254-6450.

ZHU, Y.-G.; GILLINGS, M.; PENUELAS, J. Integrating biomedical, ecological, and sustainability sciences to manage emerging infectious diseases. v. 3, n. 1, p. 23–26, 2020. ISSN 2590-3322. Disponível em: <http://www.sciencedirect.com/science/article/pii/S2590332220302931>.